Cross-category Video Highlight Detection via Set-based Learning

Minghao Xu[1] Hang Wang[1] Bingbing Ni[1]* Riheng Zhu[2] Zhenbang Sun[2] Changhu Wang[2]
[1]Shanghai Jiao Tong University, Shanghai 200240, China [2]ByteDance AI Lab
{xuminghao118, wang–hang, nibingbing}@sjtu.edu.cn
{zhuriheng, sunzhenbang, wangchanghu}@bytedance.com

Abstract

Autonomous highlight detection is crucial for enhancing the efficiency of video browsing on social media platforms. To attain this goal in a data-driven way, one may often face the situation where highlight annotations are not available on the target video category used in practice, while the supervision on another video category (named as source video category) is achievable. In such a situation, one can derive an effective highlight detector on target video category by transferring the highlight knowledge acquired from source video category to the target one. We call this problem cross-category video highlight detection, which has been rarely studied in previous works. For tackling such practical problem, we propose a ***D****ual-****L****earner-based* ***V****ideo* ***H****ighlight* ***D****etection (DL-VHD) framework. Under this framework, we first design a* ***S****et-based* ***L****earning module (SL-module) to improve the conventional pair-based learning by assessing the highlight extent of a video segment under a broader context. Based on such learning manner, we introduce two different learners to acquire the basic distinction of target category videos and the characteristics of highlight moments on source video category, respectively. These two types of highlight knowledge are further consolidated via knowledge distillation. Extensive experiments on three benchmark datasets demonstrate the superiority of the proposed SL-module, and the DL-VHD method outperforms five typical Unsupervised Domain Adaptation (UDA) algorithms on various cross-category highlight detection tasks. Our code is available at* `https://github.com/ChrisAllenMing/Cross_Category_Video_Highlight`.

1. Introduction

In current days, people show growing interests in sharing the videos recording their daily life on the social media platforms like *YouTube* and *Instagram*. Among all these videos, the well-edited ones that summarize the highlights of specific events are apparently more attractive to the audience. However, in most cases, the original video of a real-world event contains many contents unrelated to its gist, and it is an onerous and time-consuming task to pick out the highlight parts of the video manually. Therefore, in order to enhance the efficiency of video content refinement, it is desirable to develop a machine learning model for autonomous video highlight detection.

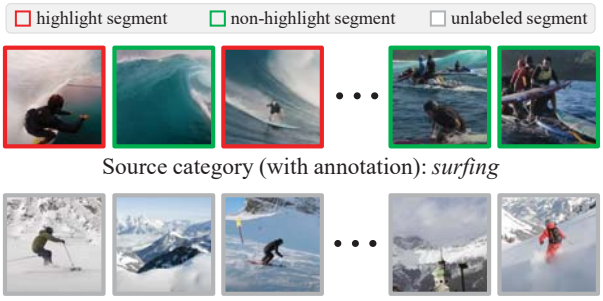

Figure 1. The situation in which the target video category used in practice lacks supervision, while another video category, *i.e.* the source one, possesses annotation.

To endow a model with the capability of identifying the highlight segments within a video, existing works have explored various ways of supervision, including the explicit highlight annotations [10, 49, 15], the frequent occurrence of specific video segments [23, 48, 21], the duration of a video [41], *etc.* These approaches generally focused on training a highlight detector for a specific video category (*e.g.* surfing, skiing, parkour, *etc.*), while the transferability of a highlight detection model across different video categories has been less studied in previous works.

As a matter of fact, in practical applications, one can face the situation where supervisory signal is lacked on the target video category intended to be used in practice, while the supervision on another video category is available, just as shown in Fig. 1. Under such situation, we consider the problem of *Cross-category Video Highlight Detection*. The setting of this problem is analogous to that of Unsupervised Domain Adaptation (UDA) [20] in which one seeks to adapt

*Corresponding author: Bingbing Ni.

the knowledge learned from the labeled source domain (the source video category with supervision) to the unlabeled target domain (the unsupervised target video category).

In addition, for optimizing the highlight detector, most of existing methods [10, 49, 15, 23, 41, 14] followed the philosophy of pair-based learning, *i.e.* comparing a positive sample (*e.g.* a highlight video segment or a segment bag containing highlights) with a negative one, and, after training, the former is expected to rank higher than the latter. Nevertheless, such learning manner might not fully exploit the contextual information spanning among different video segments. For example, in a soccer match, the moment of a player's dribbling the ball is more attractive than the one of the players' entering the pitch, and both of them are less exciting than the moment of a goal. These relationships can hardly be captured by a single segment pair, which makes the highlight prediction of a pair-learning-based model potentially imprecise in the span of a whole video.

Motivated by the facts above, in this work, we propose a **D**ual-**L**earner-based **V**ideo **H**ighlight **D**etection (*DL-VHD*) framework to address the Cross-category Video Highlight Detection problem. Under this framework, we first devise a **S**et-based **L**earning module (SL-module) to improve the conventional pair-based learning manner for highlight detection. In a nutshell, this module learns to regress the highlight score distribution over a set of segments from the same video, in which a Transformer encoder [37] is employed to model the interrelationship among various video segments. Based on this learning mechanism, we further introduce two different learners to capture two types of knowledge about highlight moments. In specific, the *coarse-grained learner* gains the basic concepts about what distinguishes the videos of target category from other ones, and the *fine-grained learner* acquires the precise highlight notions on source videos. These two kinds of knowledge are further integrated by distilling each of them into the other learner, and such integrated knowledge forms the more complete concepts about the highlight moments on target video category. In practice, the SL-module can be individually applied to derive an effective highlight detector when the segment-level annotation is available on the target video category, while, when such annotation is unobtainable, we can resort to the DL-VHD method for highlight knowledge transfer.

Our contributions can be summarized as follows:

- To the best of our knowledge, this work is the first attempt at cross-category video highlight detection, in which we utilize a dual-learner-based scheme to transfer the concepts about highlight moments across different video categories.
- We propose a novel set-based learning mechanism which is able to identify whether a video segment is highlight or not under a broader context.
- Under the category-specific setting, we verify the superior performance of the SL-module over previous methods. For cross-category highlight detection, the DL-VHD model substantially surpasses existing UDA algorithms and performs comparably with the supervised model trained on target video category.

2. Related Work

Video Highlight Detection. This task aims at assigning each video segment a score of its worthiness as highlight. In recent years, the videos studied for this task extend from sport videos [24, 42, 33] to general videos from social media [32] or first-person camera shooting [49]. According to the manner of supervision, the existing works on this topic can be generally divided into two classes. For the supervised methods [10, 49, 15, 32], the highlight annotations of all segments in a video are given. For the weakly-supervised approaches [23, 48, 21, 41, 14], various weak supervisory signals have been exploited to define highlights, including the frequent occurrence of specific segments within a video category [23, 48, 21], the duration of a video [41] and the information from segment bags [14]. For model optimization, most of these methods [10, 49, 15, 32, 41, 14] followed the philosophy of pair-based learning, *i.e.* comparing between a positive sample and a negative one.

Improvements over existing methods. In this work, we novelly explore the cross-category video highlight detection problem through learning two types of knowledge about highlight moments and integrating them on target video category. In addition, a set-based learning mechanism is proposed to improve the pair-based learning by performing highlight prediction on a set of video segments, such that the highlight extent of each segment can be judged more precisely with rich contextual information.

Unsupervised Domain Adaptation (UDA). UDA focuses on generalizing a model learned from the labeled source domain to another unlabeled target domain. To pursue this goal, a commonly used strategy is to minimize a specific metric for measuring domain shift [2, 20], *e.g.* Maximum Mean Discrepancy (MMD) [8, 36], Multi-Kernel MMD [17], Weighted MMD [47], Wasserstein Distance [28, 16] and the difference of feature covariance [31] or feature norm [46]. On another line of research, adversarial learning is employed to facilitate domain-invariance on either pixel level [3, 27, 13] or feature level [6, 35, 18, 45]. In order to introduce the discriminative information on target domain, recent works [40, 5, 25, 44, 38, 43] utilized the pseudo labels of target samples for category-level domain alignment. This work explores cross-category video highlight detection, a similar problem as UDA, in which one intends to transfer the highlight knowledge acquired from the source video category to the target one.

3. Method

In the cross-category video highlight detection problem, a set of videos containing the highlight moments of source video category, *i.e.* $D_{\mathcal{S}} = \{v_k^{\mathcal{S}}\}_{k=1}^{|D_{\mathcal{S}}|}$, are given, and each video $v \in D_{\mathcal{S}}$ is divided into N_v segments $\{(s_i, y_i)\}_{i=1}^{N_v}$ with similar duration, where y_i denotes the ground-truth highlight label for segment s_i. In addition, we have another set of videos including the highlight moments of target video category, *i.e.* $D_{\mathcal{T}} = \{v_k^{\mathcal{T}}\}_{k=1}^{|D_{\mathcal{T}}|}$, while the segment-level highlight annotations of target category are not available on these videos. Under such condition, the main objective is to derive an effective highlight detector on target video category through fully exploiting the labeled source videos and the unlabeled target ones.

3.1. Motivation and Overview

Cross-category Video Highlight Detection. In real-world applications, the segment-level highlight annotations may not be available for the target video category that the model is applied to, while one can obtain the supervision on another video category (named as source video category). Therefore, in such a situation, a natural question to ask is how to transfer the knowledge about highlight moments on source video category to the target one, *i.e.* performing *cross-category video highlight detection*. A straightforward answer is to leverage the existing Unsupervised Domain Adaptation (UDA) techniques for feature distribution alignment between two distinct video categories. However, such distribution alignment is hard, if not ill-posed, for the highlight detection problem, since the highlight segments for the target category may be nuisance for the source one, and vice versa, which is experimentally illustrated in Sec. 4.3.

To acquire the exact highlight concepts for target video category using the data from both categories, we propose a **D**ual-**L**earner-based **V**ideo **H**ighlight **D**etection (*DL-VHD*) framework. Under this framework, the model learns two kinds of knowledge about highlight moments, *i.e.* the distinction of target category videos with other ones and the characteristics of highlights on source category. These two types of knowledge are further merged to form the more complete highlight concepts about target video category.

Set-based Learning. Previous works [10, 49, 32, 41, 14] commonly trained the highlight detection model by contrasting a highlight segment s_+ with a non-highlight segment s_-, which seeks to model the conditional distribution $p(y_+, y_-|s_+, s_-)$. However, such pair-based learning may fail to discover the more complex highlight relations among more than two segments. For example, the excitement level of a soccer match differs from moment to moment, and the relative highlight extent of these moments cannot be sufficiently captured by pairs of video segments.

Motivated by such limitation, we propose a **S**et-based

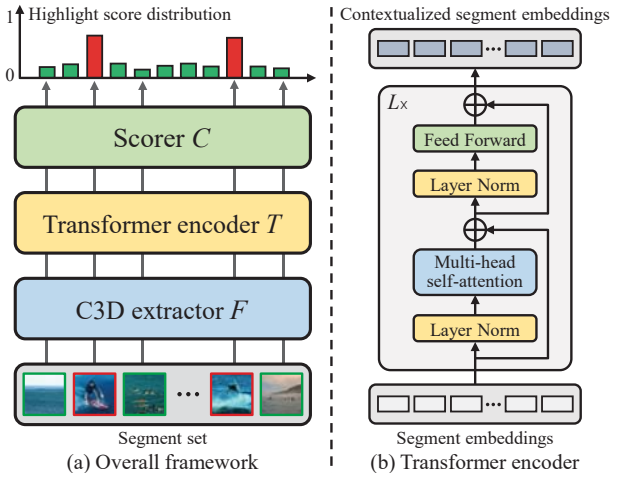

Figure 2. (a) The overall framework of SL-module. (b) The architecture of the Transformer encoder used in this module.

Learning module (SL-module). Its core idea is to train the model to predict the highlight score distribution over a set of video segments, and the prediction of a single segment is depended on all the other segments in the set, which models $p(y_1, y_2, \cdots, y_N|s_1, s_2, \cdots, s_N)$ (N denotes the set size). By including such contextual information spanning among different video segments, it is expected that the model can assign more accurate highlight score to each segment.

Cross-category Video Highlight Detection via Set-based Learning. In order to bridge the highlight patterns of two distinct video categories, it is essential to explore the interrelationship among the video segments within the same category and also across different categories. Such complex relational patterns can be better captured under the rich context provided by segment sets. Based on such motivation, in DL-VHD, we employ SL-module as the basic learning module to acquire more precise highlight knowledge.

3.2. Set-based Learning Module

The SL-module models the interdependency among the video segments in a set and predicts the highlight score of each segment under such set-determined context, as shown in Fig. 2(a). Next, we introduce the detailed learning and inference schemes of this module.

Learning scheme. In each learning step, a set of N annotated segments randomly sampled from the same video, *i.e.* $x = \{(s_j, y_j)\}_{j=1}^{N}$, is given, and a pre-trained C3D [34] model F extracts the feature embedding of each segment, *i.e.* $z = \{z_j\}_{j=1}^{N} = \{F(s_j)\}_{j=1}^{N}$ (F is fixed in the learning phase). On these segment embeddings, a Transformer encoder [37] T models the interrelationship among different segments and outputs the contextualized segment embeddings, *i.e.* $\tilde{z} = \{\tilde{z}_j\}_{j=1}^{N} = T(z)$. The Transformer encoder used in our method basically follows the original design in [37] which stacks L layers of multi-head self-attention

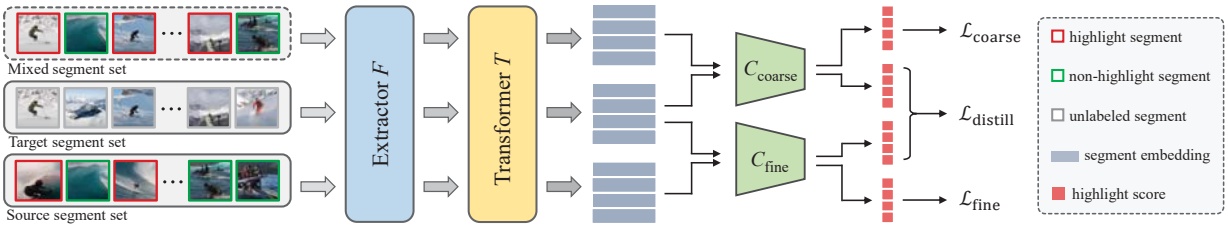

Figure 3. **Illustration of DL-VHD.** Three kinds of segment sets are constructed with a labeled source video and an unlabeled target video, and the segment embeddings are derived by a C3D extractor and a Transformer encoder. A coarse-grained and a fine-grained learner are supervised by the mixed and source segment set, respectively, and their knowledge is further consolidated by knowledge distillation.

and feed forward network. In contrast, we remove the positional encoding module for the permutation-invariance of set learning, and, as suggested in [39], Layer Normalization (LN) [1] is applied before each self-attention and feed forward module. The architecture of the Transformer encoder is shown in Fig. 2(b). We refer readers to the original literature [37] for more details.

Upon the contextualized segment embeddings, a scoring model C predicts the highlight score of each video segment, *i.e.* $\hat{y} = \{\hat{y}_j\}_{j=1}^N = \{C(\tilde{z}_j)\}_{j=1}^N$. Now that the highlight prediction on a segment set is obtained, we define the learning objective. During the learning phase, the basic desiderata is to match the highlight score distribution on set x predicted by the model with the ground-truth distribution. To attain this goal, we define the learning objective as follows:

$$\min_{T,C} \mathcal{L}_{\text{pred}}, \tag{1}$$

$$\mathcal{L}_{\text{pred}} = D_{\text{KL}}\Big(\sigma\big(\{y_j\}_{j=1}^N\big), \sigma\big(\{\hat{y}_j\}_{j=1}^N\big)\Big), \tag{2}$$

where $\sigma(\cdot)$ denotes the softmax function producing the predicted and ground-truth highlight score distributions, and $D_{\text{KL}}(\cdot,\cdot)$ stands for the Kullback–Leibler divergence.

Inference scheme. To infer the highlight score of a segment s from a test video, we first construct a segment set containing s and other $N-1$ context segments adjacent to s in the video, denoted as $x = \{s_j\}_{j=1}^N$. Half of these context segments are right before s in the video, and the other half are right after. Segment duplication is conducted if the segments around s cannot fill the set. Here, we do not use the set composed of random segments as in the learning phase for the sake of suppressing the variance of prediction. We then infer the highlight score of all segments in the set, *i.e.* $\hat{y} = \{\hat{y}_j\}_{j=1}^N$, by feeding them into the C3D feature extractor, the Transformer encoder and the scoring model successively. Finally, we pick out $\hat{y}_{\text{id}(s)}$ ($\text{id}(s)$ stands for the index of s in x) as the highlight score of the video segment to be evaluated.

3.3. Dual-Learner-based Video Highlight Detection

On the basis of SL-module, we now explore the cross-category video highlight detection problem. Its main objective is to derive an effective highlight detector on target video category by fully exploiting the labeled source videos $D_{\mathcal{S}}$ and the unlabeled target ones $D_{\mathcal{T}}$. To pursue such goal, we seek to capture the highlight concepts about target video category from two aspects. On one hand, there exists some obvious features that distinguish the target category videos from the ones of other topics, *e.g.* the surfboard in surfing videos, the ski pole in skiing videos, *etc.* The perception of such features endows a model with the basic capability of picking out the segments of the target category from a video mixing different contents. On the other hand, there are some common characteristics of highlight moments sharing between distinct video categories. For instance, the moments with a standing person moving on some surface of the scene can be the highlights of both surfing and skiing videos. Such generic knowledge can be employed to identify the highlight moments for the target video category. However, neither of these two types of concepts alone can sufficiently define the highlights on the target category, which calls for a scheme that integrates different knowledge.

Following the above intuitions, we design a dual-learner-based framework, in which two kinds of highlight knowledge are learned by a *coarse-grained learner* and a *fine-grained learner* respectively, and they are further integrated by a knowledge distillation [12] scheme. The graphical illustration of this framework is shown in Fig. 3. We state the detailed learning and inference schemes as follows.

Learning scheme. In each learning step, we use a set of labeled segments randomly sampled from a video in $D_{\mathcal{S}}$, denoted as $x_{\mathcal{S}} = \{(s_j^{\mathcal{S}}, y_j^{\mathcal{S}})\}_{j=1}^N$, and a set of unlabeled segments randomly sampled from a video in $D_{\mathcal{T}}$, denoted as $x_{\mathcal{T}} = \{s_j^{\mathcal{T}}\}_{j=1}^N$. Based on these two sets, we further construct a mixed set with the segments from two video categories, denoted as $x_{\mathcal{M}} = \{(s_j^{\mathcal{M}}, y_j^{\mathcal{M}})\}_{j=1}^N$ ($y_j^{\mathcal{M}}$ equals to 1 if $s_j^{\mathcal{M}}$ is a target category segment and is 0 otherwise), in which half of the segments are randomly sampled from $x_{\mathcal{S}}$, and the other half are from $x_{\mathcal{T}}$. Using the C3D feature extractor and Transformer encoder, we respectively derive the contextualized segment embeddings for these three sets, *i.e.* $\tilde{z}_{\mathcal{S}} = \{\tilde{z}_j^{\mathcal{S}}\}_{j=1}^N$, $\tilde{z}_{\mathcal{T}} = \{\tilde{z}_j^{\mathcal{T}}\}_{j=1}^N$ and $\tilde{z}_{\mathcal{M}} = \{\tilde{z}_j^{\mathcal{M}}\}_{j=1}^N$.

Upon these segment embeddings, on one hand, we introduce a *coarse-grained learner* C_{coarse} to learn the basic dis-

tinction of target video segments with the source ones. This is achieved by matching the highlight prediction of C_{coarse} on mixed set, *i.e.* $\hat{y}_{\mathcal{M}} = \{\hat{y}_j^{\mathcal{M}}\}_{j=1}^N = \{C_{\text{coarse}}(\tilde{z}_j^{\mathcal{M}})\}_{j=1}^N$, with the ground-truth highlight distribution on that set, which defines the coarse-grained highlight prediction loss:

$$\mathcal{L}_{\text{coarse}} = D_{\text{KL}}\Big(\sigma\big(\{y_j^{\mathcal{M}}\}_{j=1}^N\big), \sigma\big(\{\hat{y}_j^{\mathcal{M}}\}_{j=1}^N\big)\Big). \tag{3}$$

On the other hand, a *fine-grained learner* C_{fine} is introduced to acquire the knowledge about highlight moments on the source video category. This is attained through the supervised learning on set $x_{\mathcal{S}}$, in which the prediction of C_{fine}, *i.e.* $\hat{y}_{\mathcal{S}} = \{\hat{y}_j^{\mathcal{S}}\}_{j=1}^N = \{C_{\text{fine}}(\tilde{z}_j^{\mathcal{S}})\}_{j=1}^N$, is aligned with the ground-truth highlight score distribution:

$$\mathcal{L}_{\text{fine}} = D_{\text{KL}}\Big(\sigma\big(\{y_j^{\mathcal{S}}\}_{j=1}^N\big), \sigma\big(\{\hat{y}_j^{\mathcal{S}}\}_{j=1}^N\big)\Big). \tag{4}$$

Now that two types of knowledge about highlight moments are acquired by two different learners, we aim to integrate them on the target video category. Inspired by the idea of knowledge distillation [12], we would like to distill the knowledge of each learner into the other learner without impairing its original knowledge. Specifically, the coarse-grained and fine-grained learner are both utilized to predict the highlight scores of the segments in set $x_{\mathcal{T}}$, which gives out $\hat{y}_{\mathcal{T},\text{coarse}} = \{\hat{y}_j^{\mathcal{T},\text{coarse}}\}_{j=1}^N = \{C_{\text{coarse}}(\tilde{z}_j^{\mathcal{T}})\}_{j=1}^N$ and $\hat{y}_{\mathcal{T},\text{fine}} = \{\hat{y}_j^{\mathcal{T},\text{fine}}\}_{j=1}^N = \{C_{\text{fine}}(\tilde{z}_j^{\mathcal{T}})\}_{j=1}^N$. We then generate the prediction reflecting both kinds of highlight knowledge by averaging $\hat{y}_{\mathcal{T},\text{coarse}}$ and $\hat{y}_{\mathcal{T},\text{fine}}$, which produces $\hat{y}_{\mathcal{T},\text{avg}} = \{\hat{y}_j^{\mathcal{T},\text{avg}}\}_{j=1}^N = \{(\hat{y}_j^{\mathcal{T},\text{coarse}} + \hat{y}_j^{\mathcal{T},\text{fine}})/2\}_{j=1}^N$. In order to perform knowledge distillation between two learners, we constrain the individual prediction from either the coarse-grained or fine-grained learner to approach the average prediction, which defines the distillation loss as below:

$$\begin{aligned}\mathcal{L}_{\text{distill}} = \frac{1}{2}\Bigg(& D_{\text{KL}}\Big(\sigma\big(\{\hat{y}_j^{\mathcal{T},\text{avg}}\}_{j=1}^N\big), \sigma\big(\{\hat{y}_j^{\mathcal{T},\text{coarse}}\}_{j=1}^N\big)\Big) \\ &+ D_{\text{KL}}\Big(\sigma\big(\{\hat{y}_j^{\mathcal{T},\text{avg}}\}_{j=1}^N\big), \sigma\big(\{\hat{y}_j^{\mathcal{T},\text{fine}}\}_{j=1}^N\big)\Big)\Bigg).\end{aligned} \tag{5}$$

The overall learning objective can be summarized as:

$$\min_{T, C_{\text{coarse}}, C_{\text{fine}}} \mathcal{L}_{\text{coarse}} + \mathcal{L}_{\text{fine}} + \lambda \mathcal{L}_{\text{distill}}, \tag{6}$$

where λ is the trade-off parameter balancing between highlight prediction and knowledge distillation losses.

Inference scheme. During inference, given a segment s from a target category video, we first extend it into a set with other $N-1$ context segments adjacent to s in the same video, and the set is denoted as $x_{\mathcal{T}} = \{s_j^{\mathcal{T}}\}_{j=1}^N$. The selection of these context segments follows the scheme depicted in the inference part of Sec. 3.2. The highlight score of each segment in $x_{\mathcal{T}}$ is respectively inferred by the coarse-grained and fine-grained learner, which derives the highlight prediction $\hat{y}_{\mathcal{T},\text{coarse}} = \{\hat{y}_j^{\mathcal{T},\text{coarse}}\}_{j=1}^N$ and $\hat{y}_{\mathcal{T},\text{fine}} = \{\hat{y}_j^{\mathcal{T},\text{fine}}\}_{j=1}^N$. These two kinds of predictions are further averaged to produce $\hat{y}_{\mathcal{T},\text{avg}} = \{\hat{y}_j^{\mathcal{T},\text{avg}}\}_{j=1}^N$. Finally, we pick out $\hat{y}_{\text{id}(s)}^{\mathcal{T},\text{avg}}$ ($\text{id}(s)$ denotes the index of s in $x_{\mathcal{T}}$) as the highlight score of segment s.

4. Experiments

In this section, we compare the proposed *SL-module* and the *DL-VHD* method with existing video highlight detection approaches under the category-specific and cross-category setting, respectively.

4.1. Experimental Setup

Model details. Following [10, 41], a C3D model [34] pre-trained on the UCF101 dataset [30] serves as the backbone for feature extraction, and its parameters are fixed during training. The Transformer encoder is constructed with 5 layers of self-attention and feed forward block, and each multi-head self-attention module is equipped with 8 attention heads. The scoring model C, coarse-grained learner C_{coarse} and fine-grained learner C_{fine} are all instantiated as a multi-layer perceptron with architecture FC(4096,1024) $\rightarrow$ ReLU $\rightarrow$ FC(1024,256) $\rightarrow$ ReLU $\rightarrow$ FC(256,1), where FC is short for fully-connected layer.

Training details. In all experiments, an SGD optimizer (initial learning rate: 0.001, momentum: 0.9, weight decay: 5×10^{-4}) is employed to train the model for 50 epochs, and the learning rate is multiplied by 0.1 every 20 epochs. For each video segment, 16 frames are sampled from it with the same interval. Without otherwise specified, the set size N is set as 20, and the trade-off parameter λ is set as 1.0 (parameter sensitivity is analyzed in Sec. 5.2). We use an NVIDIA Tesla V100 GPU for training. Our method is implemented with the PyTorch [22] deep learning framework, and the source code will be released for reproducibility.

Performance comparison. Under the category-specific setting, six supervised video highlight detection (or video summarization) methods, *i.e.* Video2GIF [10], LSVM [32], KVS [23], DPP [7], vsLSTM [50] and SM [9], and six weakly-supervised approaches, *i.e.* RRAE [48], SG [19], DSN [21], VESD [4], LIM-s [41] and MINI-Net [14], are introduced for comparison. For the cross-category setting, the SL-module trained on the source/target video category serves as the lower/upper bound of model performance. For the sake of fair comparison, five UDA algorithms, *i.e.* DAN [17], DeepCORAL [31], RevGrad [6], MCD [26] and AFN [46], are combined with SL-module to compare with the proposed DL-VHD method, and the detailed combination schemes are provided in the supplementary material.

Table 1. Highlight detection results (mAP) of weakly-supervised and supervised methods on the YouTube Highlights dataset.

Category	Weakly-supervised Methods			Supervised Methods			
	RRAE [48]	LIM-s [41]	MINI-Net [14]	Video2GIF [10]	LSVM [32]	SL-module (w/o T)	SL-module
dog	0.49	0.579	0.537	0.308	0.60	0.690	**0.708**
gymnastics	0.35	0.417	0.528	0.335	0.41	0.506	**0.532**
parkour	0.50	0.670	0.689	0.540	0.61	0.690	**0.772**
skating	0.25	0.578	0.709	0.554	0.62	0.687	**0.725**
skiing	0.22	0.486	0.583	0.328	0.36	0.636	**0.661**
surfing	0.49	0.651	0.638	0.541	0.61	0.695	**0.762**
Average	0.383	0.564	0.614	0.464	0.536	0.651	**0.693**

Table 2. Highlight detection results (top-5 mAP score) of weakly-supervised and supervised methods on the TVSum dataset.

Category	Weakly-supervised Methods					Supervised Methods					
	SG [19]	DSN [21]	VESD [4]	LIM-s [41]	MINI-Net [14]	KVS [23]	DPP [7]	vsLSTM [50]	SM [9]	SL-module (w/o T)	SL-module
VT	0.423	0.373	0.447	0.559	0.803	0.353	0.399	0.411	0.415	0.837	**0.865**
VU	0.472	0.441	0.493	0.429	0.653	0.441	0.453	0.462	0.467	0.663	**0.687**
GA	0.475	0.428	0.496	0.612	**0.754**	0.402	0.457	0.463	0.469	0.724	0.749
MS	0.489	0.436	0.503	0.540	0.813	0.417	0.462	0.477	0.478	0.851	**0.862**
PK	0.456	0.411	0.478	0.604	0.780	0.382	0.437	0.448	0.445	0.767	**0.790**
PR	0.473	0.417	0.485	0.475	0.545	0.403	0.446	0.461	0.458	0.594	**0.632**
FM	0.464	0.412	0.487	0.432	0.559	0.397	0.442	0.452	0.451	0.580	**0.589**
BK	0.417	0.368	0.441	0.663	0.717	0.342	0.395	0.406	0.407	0.708	**0.726**
BT	0.483	0.435	0.492	0.691	0.769	0.419	0.464	0.471	0.473	0.779	**0.789**
DS	0.466	0.416	0.488	0.626	0.591	0.394	0.449	0.455	0.453	0.612	**0.640**
Average	0.462	0.424	0.481	0.563	0.698	0.398	0.447	0.451	0.461	0.712	**0.733**

Table 3. Highlight detection results (mAP) of weakly-supervised and supervised methods on the ActivityNet dataset.

Category	Weakly-supervised		Supervised		
	LIM-s [41]	MINI-Net [14]	LSVM [32]	SL-module (w/o T)	SL-module
eat&drink	0.638	0.702	0.670	0.716	**0.736**
personal care	0.663	0.689	0.657	0.725	**0.744**
household	0.621	0.745	0.707	0.763	**0.787**
sport	0.710	0.794	0.769	0.835	**0.849**
social	0.743	0.760	0.740	0.758	**0.779**
Average	0.675	0.738	0.709	0.759	**0.779**

4.2. Category-specific Video Highlight Detection

When highlight annotations are available on the video category intended to be used, SL-module can be individually applied to train a highlight detector under the category-specific setting. We compare it with existing video highlight detection and video summarization methods in this section.

Datasets. *YouTube Highlights* [32] is composed of six video categories, *i.e.* dog, gymnastics, parkour, skating, skiing and surfing, and each category has approximately 100 videos. Segment-level annotations are provided to indicate whether a segment is a highlight moment or not. We follow the standard training-test split [32] for model evaluation.

TVSum [29] is a video summarization dataset consisting of 10 categories of video events with 5 videos in each category, and frame-level importance score is provided in this dataset. Following previous works [41, 14], we average the frame-level importance scores to achieve the segment-level highlight scores. For each video category, we select the two longest videos (about 10 minutes in total) for training and the rest three ones for test.

ActivityNet [11] is a large-scale database for human activity classification and detection. We employ the data of the temporal action localization track for highlight detection. Specifically, we split the video samples to five categories, *i.e.* eat&drink, personal care, household, sport and social, according to the first-level action label. The temporal Intersection over Union (tIoU) between a video segment and a ground-truth event of a specific category is used as the segment's highlight label for this video category. Totally, we utilize 2520 videos for training and 1260 videos for test, and the detailed dataset statistics for all video categories are provided in the supplementary material.

Results on YouTube Highlights. In Tab. 1, we compare our method with existing approaches on six video categories of YouTube Highlights. It can be observed that the proposed SL-module outperforms previous pair-learning-based algorithms, *i.e.* LIM-s, MINI-Net, Video2GIF and LSVM, on all six categories, and superior average mAP is still obtained when the Transformer encoder T is removed from our model. This phenomenon illustrates the superiority of set-based learning over pair-based methods, in which the broader contextual information within a segment set enables more precise highlight prediction of each video segment.

Results on TVSum. Tab. 2 reports the performance of various video highlight detection and video summarization approaches on TVSum. On nine of ten video categories, the proposed SL-module achieves the best performance, and, when removing the Transformer encoder, it still outperforms the state-of-the-art MINI-Net on seven of ten categories. These results verify the effectiveness of set-based

Table 4. Cross-category highlight detection results (mAP) on the YouTube Highlights dataset. (source video category: surfing; the underlined result surpasses the target-oracle.)

Methods	→dog	→gymnastics	→parkour	→skating	→skiing
Source-only	0.485	0.505	0.547	0.568	0.545
DAN [17]	**0.652**	0.487	0.713	0.638	0.611
DeepCORAL [31]	0.634	0.513	0.732	0.659	0.620
RevGrad [6]	0.628	0.493	0.654	0.640	0.597
MCD [26]	0.567	0.529	0.499	0.642	0.654
AFN [46]	0.625	0.517	0.575	0.653	0.626
DL-VHD ($\mathcal{L}_{\text{coarse}}$ only)	0.574	0.498	0.704	0.635	0.631
DL-VHD ($\mathcal{L}_{\text{fine}}$ only)	0.485	0.505	0.547	0.568	0.545
DL-VHD (w/o $\mathcal{L}_{\text{distill}}$)	0.630	0.529	0.718	0.683	0.658
DL-VHD (full model)	0.649	**0.540**	**0.748**	**0.713**	**0.686**
Target-oracle	0.708	0.532	0.772	0.725	0.661

Table 5. Cross-category highlight detection results (mAP) on the ActivityNet dataset. (source video category: sport; the underlined result surpasses the target-oracle.)

Methods	→eat&drink	→personal care	→household	→social
Source-only	0.674	0.667	0.707	0.722
DAN [17]	0.656	0.678	0.694	0.735
DeepCORAL [31]	0.708	0.705	0.765	0.744
RevGrad [6]	0.687	0.701	0.722	0.731
MCD [26]	0.712	0.713	0.761	0.756
AFN [46]	0.718	0.704	0.750	0.749
DL-VHD ($\mathcal{L}_{\text{coarse}}$ only)	0.689	0.694	0.742	0.741
DL-VHD ($\mathcal{L}_{\text{fine}}$ only)	0.674	0.667	0.707	0.722
DL-VHD (w/o $\mathcal{L}_{\text{distill}}$)	0.713	0.715	0.778	0.754
DL-VHD (full model)	**0.730**	**0.728**	**0.793**	**0.766**
Target-oracle	0.736	0.744	0.787	0.779

learning under the circumstances with limited training data, *i.e.* only two videos per category for training.

Results on ActivityNet. In Tab. 3, we evaluate the performance of three existing methods and two configurations of the proposed model. Since the experiments on ActivityNet dataset were not commonly included in previous works, we examine these works by the released source code (for MINI-Net and LSVM) or our re-implementation (for LIM-s). The experimental results on this large-scale dataset further verify the superiority of the proposed set-learning method (*i.e.* obtaining the highest test mAP on all five video categories) when the training data is abundant.

4.3. Cross-category Video Highlight Detection

Under the cross-category highlight detection setting, we evaluate the effectiveness of DL-VHD and various UDA algorithms on transferring the highlight knowledge from source video category to the target one. In all experiments, the videos of source category possess segment-level annotations, while the videos of target category are unannotated.

Tasks. *YouTube Highlights* consists of six video categories, and we employ *surfing* as the source category and evaluate each of the cases that one of the other five categories serves as the target one. Also, we consider a more difficult setting where *dog* is used as the source category (*i.e.* adapting from dog activities to human ones), and the results of this setting are in the supplementary material.

ActivityNet contains five categories of human activities, and we utilize *sport* as the source category and aim at transferring the knowledge of sport highlights to other four video categories. The adaptation towards each target video category is separately examined.

Cross-category results on YouTube Highlights. Tab. 4 reports the performance of various approaches on five cross-category highlight detection tasks, in which *surfing* serves as the source category. Source-only (target-oracle) method represents the SL-module trained on the source (target) video category in a supervised fashion, where an obvious performance gap exists between them. We can observe that the full model of DL-VHD surpasses five existing UDA algorithms on four of five tasks, and it surprisingly outperforms the target-oracle model on two tasks, *i.e.* surfing → gymnastics and surfing → skiing. Such results illustrate that cross-category video highlight detection cannot be easily deemed as a variant of UDA problem, and more dedicated techniques (*e.g.* the proposed dual-learner and knowledge distillation schemes) can better discover the transferrable highlight patterns across different video categories.

Cross-category results on ActivityNet. In Tab. 5, we compare the proposed DL-VHD model with five UDA methods on the cross-category highlight detection tasks of ActivityNet, and *sport* is utilized as the source category in all these tasks. The full model of DL-VHD achieves higher mAP than the UDA algorithms on all four tasks, and it even outperforms the target-oracle model on the sport → household task. These empirical results verify that the DL-VHD model succeeds in capturing the human-related action patterns on the target video category under the guidance of labeled source videos and unlabeled target videos.

5. Analysis

In this section, we conduct more in-depth analysis of our approach to evaluate the effectiveness of major model components both quantitatively and qualitatively.

5.1. Ablation Study

Effect of Transformer encoder. In all three video highlight detection datasets, we compare the performance of the SL-module with and without Transformer encoder T, as shown in Tabs. 1, 2 and 3. It can be observed that, after applying the Transformer encoder, the proposed set-based learning method obtains a clear performance gain on all tasks, which demonstrates the importance of interrelationship modeling when learning from a set of video segments.

Effect of dual learners and knowledge distillation. In Tabs. 4 and 5, we investigate the main components of DL-VHD through three additional model configurations: (1) $\mathcal{L}_{\text{coarse}}$ only: only the coarse-grained learner is utilized

Figure 4. Highlight predictions of three methods on the *surfing* → *skiing* task. (Each video segment is denoted by its first and last frames.)

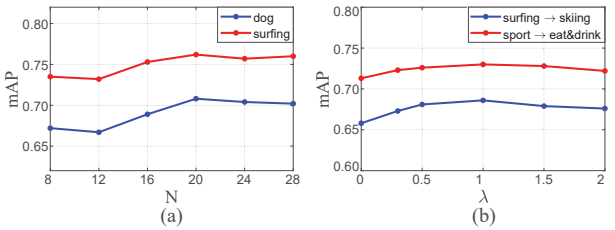

Figure 5. Sensitivity analysis of set size N (left) and trade-off parameter λ (right).

to predict the highlight extent of a segment with respect to the target video category; (2) $\mathcal{L}_{\text{fine}}$ only: only the fine-grained learner is employed for highlight prediction on target category (this configuration is equivalent to the source-only baseline); (3) w/o $\mathcal{L}_{\text{distill}}$: both the coarse- and fine-grained learners are trained, while their knowledge is not integrated by the knowledge distillation loss. When the two learners are individually applied, the coarse-grained learner outperforms the fine-grained one, which, we think, is because the supervision for coarse-grained learner is more relevant to the highlight patterns on target video category than the supervision applied to fine-grained learner. In the full model, the knowledge distillation scheme is able to further promote model's performance upon configuration (3) by integrating the knowledge of two learners.

5.2. Sensitivity Analysis

Sensitivity of set size N. In this experiment, we analyze the sensitivity of the proposed SL-module to the set size. Fig. 5(a) shows the model performance on two highlight detection tasks under different set sizes. It can be observed that our set-based learning method can achieve stable performance gain when the size of each segment set is large enough, *i.e.* $N \geqslant 16$.

Sensitivity of trade-off parameter λ. In this part, we discuss the selection of trade-off parameter λ which balances between highlight prediction and knowledge distillation objectives. In Fig. 5(b), we plot the performance of DL-VHD on two cross-category highlight detection tasks using various λ values. The highest mAP on target video category is gained when the value of λ is around 1.0, which indicates that the appropriate balance between two distinct optimization objectives is attained under such condition.

5.3. Visualization

For the cross-category highlight detection task *surfing* → *skiing*, Fig. 4 visualizes the highlight prediction results of three methods, *i.e.* source-only, AFN and DL-VHD, on a target category video. For each method, we select the segments with the closest highlight score to the corresponding coordinate value (0.2, 0.4, 0.6 or 0.8), and each segment is represented by its first and last frames. The source-only model fails to capture the highlight patterns of skiing, and the AFN algorithm performs better but still overvalue a non-highlight segment by a score near 0.6. By comparison, DL-VHD assigns highlight scores to various video segments most appropriately. More visualization results on other tasks can be found in the supplementary material.

6. Conclusions and Future Work

In this research, we novelly explore the cross-category video highlight detection problem with a Dual-Learner-based Video Highlight Detection (DL-VHD) framework. Under this framework, a Set-based Learning module (SL-module) is proposed to improve the commonly employed pair-based learning, and dual-learner and knowledge distillation schemes are further introduced for highlight knowledge transfer. The comprehensive experiments under both the category-specific and cross-category settings verify the exceeding performance of the proposed method.

Our future explorations will involve further improving the algorithm for cross-category highlight detection, applying the proposed approach to more sophisticated real-world applications and studying on the generalization capability of video highlight detection models.

7. Acknowledgement

This work was supported by National Science Foundation of China (U20B2072, 61976137). Authors appreciate the Student Innovation Center of SJTU and the ByteDance AI Lab for providing GPUs. Authors also thank Jie Zhou, Jiawen Li and Xuanyu Zhu for their valuable suggestions.

References

[1] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[4] Sijia Cai, Wangmeng Zuo, Larry S. Davis, and Lei Zhang. Weakly-supervised video summarization using variational encoder-decoder and web prior. In *European Conference on Computer Vision*, 2018.

[5] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[6] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 2015.

[7] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems*, 2014.

[8] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[9] Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[10] Michael Gygli, Yale Song, and Liangliang Cao. Video2gif: Automatic generation of animated gifs from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[12] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, 2018.

[14] Fa-Ting Hong, Xuanteng Huang, Weihong Li, and Wei-Shi Zheng. Mini-net: Multiple instance ranking network for video highlight detection. In *European Conference on Computer Vision*, 2020.

[15] Yifan Jiao, Zhetao Li, Shucheng Huang, Xiaoshan Yang, Bin Liu, and Tianzhu Zhang. Three-dimensional attention-based deep ranking model for video highlight detection. *IEEE Transactions on Multimedia*, 20(10):2693–2705, 2018.

[16] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[17] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 2015.

[18] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, 2018.

[19] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial LSTM networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[20] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[21] Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, and Amit K. Roy-Chowdhury. Weakly supervised summarization of web videos. In *IEEE International Conference on Computer Vision*, 2017.

[22] 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 Workshop*, 2017.

[23] Danila Potapov, Matthijs Douze, Zaïd Harchaoui, and Cordelia Schmid. Category-specific video summarization. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision*, 2014.

[24] Yong Rui, Anoop Gupta, and Alex Acero. Automatically extracting highlights for tv baseball programs. In *ACM International Conference on Multimedia*, 2000.

[25] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *International Conference on Computer Vision*, 2019.

[26] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018.

[27] Swami Sankaranarayanan, Yogesh Balaji, Carlos D. Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[28] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *AAAI Conference on Artificial Intelligence*, 2018.

[29] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[30] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[31] Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In *ECCV Workshop*, 2016.

[32] Min Sun, Ali Farhadi, and Steven M. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European Conference on Computer Vision*, 2014.

[33] Hao Tang, Vivek Kwatra, Mehmet Emre Sargin, and Ullas Gargi. Detecting highlights in sports videos: Cricket as a test case. In *IEEE International Conference on Multimedia and Expo*, 2011.

[34] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision*, 2015.

[35] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[36] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[38] Hang Wang, Minghao Xu, Bingbing Ni, and Wenjun Zhang. Learning to combine: Knowledge aggregation for multi-source domain adaptation. In *European Conference on Computer Vision*, 2020.

[39] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning deep transformer models for machine translation. In *Conference of the Association for Computational Linguistics*, 2019.

[40] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, 2018.

[41] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[42] Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran, and Thomas S. Huang. Highlights extraction from sports video based on an audio-visual marker detection framework. In *IEEE International Conference on Multimedia and Expo*, 2005.

[43] Minghao Xu, Hang Wang, and Bingbing Ni. Graphical modeling for multi-source domain adaptation. *CoRR*, abs/2104.13057, 2021.

[44] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[45] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI Conference on Artificial Intelligence*, 2020.

[46] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *International Conference on Computer Vision*, 2019.

[47] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[48] Huan Yang, Baoyuan Wang, Stephen Lin, David P. Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *IEEE International Conference on Computer Vision*, 2015.

[49] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[50] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European Conference on Computer Vision*, 2016.