


PAC-Net: Highlight Your Video via History Preference Modeling

Hang Wang¹, Penghao Zhou², Chong Zhou³, Zhao Zhang⁴, and Xing Sun^{5*}

¹Huawei ²ByteDance ³Nanyang Technological University ⁴NanKai University ⁵Shopee
{francis970625, patrick.phzhou, winfred.sun}@gmail.com
chong033@ntu.edu.sg zzhang@mail.nankai.edu.cn

Abstract. Autonomous highlight detection is crucial for video editing and video browsing on social media platforms. General video highlight detection aims at extracting the most interesting segments from the entire video. However, interest is subjective among different users. A naive solution is to train a model for each user but it is not practical due to the huge training expense. In this work, we propose a *Preference-Adaptive Classification* (PAC-Net) framework, which can model users’ personalized preferences from their user history. Specifically, we design a *Decision Boundary Customizer* (DBC) module to dynamically generate the user-adaptive highlight classifier from the preference-related user history. In addition, we introduce *Mini-History* (Mi-Hi) mechanism to capture more fine-grained user-specific preferences. The final highlight prediction is jointly decided by the user’s multiple preferences. Extensive experiments demonstrate that PAC-Net achieves state-of-the-art performance on the public benchmark dataset, whilst using substantially smaller networks.

Keywords: Personalized Video Highlight Detection, User-adaptive Learning, Decision Boundary, User Preference Modeling

1 Introduction

Nowadays, people show growing interest in short videos to record and share their daily life. However, it is a laborious task to manually pick out the more attractive highlight parts from the long video to get a well-edited one. Therefore, autonomous video highlight detection arouses great attention in the vision community, and many efforts have been devoted into the study of general video highlight detection (VHD) [9,13,50,46,4].

In real-world applications, a more practical task is personalized video highlight detection (P-VHD), which aims at extracting user-adaptive highlight predictions guided by annotated user history. Since user preferences vary a lot, it is very subjective when it comes to determining how interesting a video segment, *e.g.*, for a sports competition video, some prefer the scoring moments while some enjoy more teamwork plays. General VHD algorithms may not perform well due to the neglect of user’s personal information. Such an issue calls for efficient and specialized techniques for P-VHD problem.

* Corresponding author

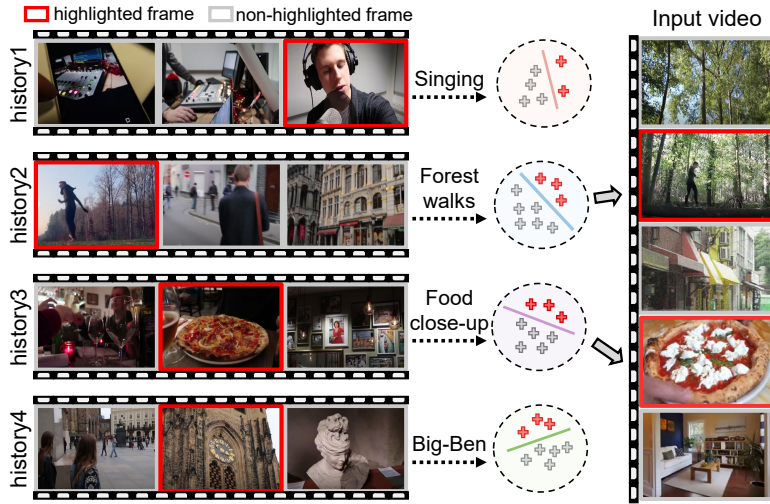


Fig. 1: **Motivation of PAC-Net.** A user has diverse preferences, and these preferences can be collected from user history and used as guidance for personalized prediction. In this example, the user is interested in singing, forest walks, food close-up, and Big Ben. In PAC-Net, we first summarize each preference with a classifier that can classify highlight parts and non-highlight parts for the corresponding preference. Then, multiple preference-specific classifiers jointly result in the final prediction for the input video.

PHD-GIFs [24] is the first attempt and it proposes a large-scale dataset as well as a baseline model. PHD-GIFs concatenates the historical highlight video segments with the input video so that the input contains the personalized information. Recently, Adaptive-H-FCSN [29] proposes to use adaptive instance normalization [11] that conditions on the user history along the temporal dimension, which utilizes the user history more effectively than direct concatenation.

In this paper, we propose a new perspective to better address the P-VHD problem. Imagine you are required to pick out several interesting segments from a new video for a customer, and you are also given several videos that have annotations indicating which frames the customer prefers, one intuitive solution is to first summarize the labeled history videos into assessment standards then apply these standards to the unlabeled video to filter out the desired contents. See Figure 1 for an example. To mimic what a person would do, we propose a novel framework called *Preference-Adaptive Classification* (PAC-Net), which consists of two key components, *Decision Boundary Customizer* (DBC) module and *Mini-History* (Mi-Hi) mechanism.

Specifically, DBC is responsible for extracting the aforementioned assessment standards from the user history. In a nutshell, we represent each standard with a highlight classifier that draws a decision boundary between the highlighted frames and non-highlighted frames. DBC module is designed to dynamically

generate such preference-specific classifier based on labeled user history. In fact, such process is a feature transformation from user history to the weights of the preference-related highlight classifier, and the generated classifier is responsible for the highlight prediction. During inference, by plugging in the history of a specific user, DBC could generate personalized classifier without any retraining. In order to encourage DBC to generate diverse highlight classifiers, we equip DBC with a regularization module to prevent the personalized highlight classifier from degenerating into a generic highlight classifier. In addition, we also find that previous methods discard the non-highlighted frames of the user history and only utilize the highlighted ones. However, the non-highlights could also be beneficial for the P-VHD task, *e.g.*, helping to eliminate the false positives. Therefore, we further enable DBC to take the non-highlighted user history as part of its inputs to generate a more precise highlight classifier.

The Mini-History (Mi-Hi) mechanism is proposed to make the assessment standards more fine-grained. Previous P-VHD methods view the user preferences at a user level, that is, concatenating features of all the highlighted frames into one feature vector thus each user only has one preference representation. However, as a matter of fact, one user could have diverse preferences, let alone one video could contain multiple topics. Therefore, we argue that user history could be utilized at a more fine-grained level. Mi-Hi mechanism makes it possible by extracting all highlight segments from user history and converting each preference-related highlight segment into a preference-specific highlight classifier by DBC. The final highlight prediction is jointly decided by all the classifiers.

To prove the effectiveness of PAC-Net, we conduct extensive experiments on PHD-GIFs dataset [24], the only related large-scale dataset for this task. The results show that PAC-Net outperforms state-of-the-art methods, with a relative improvement of 9% over Adaptive-H-FCSN. Comprehensive ablation experiments validate the effectiveness of our method, and we also provide visualizations for better understanding.

Our contributions are summarized as follows:

1. We propose PAC-Net, a novel framework for personalized video highlight detection, which achieves state-of-the-art performance on a large-scale benchmark dataset whilst using substantially smaller networks.
2. We design the Decision Boundary Customizer (DBC) module to dynamically generate the preference-related highlight classifier conditioned on labeled user history, which makes the highlight classifier user-adaptive.
3. To capture more fine-grained user preferences, we introduce the Mini-History (Mi-Hi) mechanism, where each historical highlight segment is first extracted and then converted into a highlight classifier by DBC, and the final prediction is jointly decided by user’s multiple highlight classifiers.

2 Related Work

Personalized Video Highlight Detection. Video highlight detection aims at detecting the attractive clips from the whole video, which capture the impor-

tant information in the video. Early works mainly focus on sports videos [32,44,38], and in recent years, the research area has been extended to general videos like social media videos [36,24] or first-person videos [48]. Most video highlight detectors do not consider user preference [9,13,49,50,10,3,46,42] and these general detectors are mostly ranking models, whose key idea is to rank the highlighted segments higher than the non-highlighted ones. For example, Video2GIF [9] proposes a deep model to generate a ranked list of segments according to their suitability as GIF. Recently, several works are proposed for P-VHD task. PHD-GIFs [24] is the first works for personalized video highlight detection and it also creates a large-scale dataset. PHD-GIFs is a ranking model as well but it concatenates the historical highlight segments into the input to capture the user’s interests. Adaptive-H-FCSN [29] proposes T-AIN, which extends the adaptive instance normalization [11] to the time dimension. The T-AIN layer is conditioned on user history and is more effective than concatenating the input feature with the history feature. In this work, we tackle the P-VHD problem from a new perspective for better personalization.

Personalized Video Summarization. Video highlight detection is evolved from the video summarization task. The major difference is that video summarization targets on providing a concise overview of the entire video, while video highlight detection requires extracting the most interesting segments and does not enforce integrity. Early video summarization systems are based on heuristic rules and are fully unsupervised [40,15,16,53,20,22,25,26,35,56]. Later, as deep learning develops rapidly, several data-driven supervised approaches outperform the hand-crafted rules [5,6,8,18,21,28,45,51,52,7,55]. However, all the aforementioned video summarization methods are for general video summarization. To adapt a general model to be user-specific, one naive solution is to retrain the whole model or part of the model for each user, which, however, is usually infeasible in practice due to that retraining takes much more time than feed-forward computation and there might not be sufficient training data for just one user. Another line of work performs personalization with extra annotations, such as metadata (*e.g.*, user profiles) [1,2,12,37] and user interaction (*e.g.*, text queries) [23,19,33,34,41,47,54]. Note that, our method does not rely on either metadata or user interaction.

3 Methodology

The proposed PAC-Net consists of three components namely encoder, Decision Boundary Customizer (DBC), and Mini-History (Mi-Hi) mechanism. In this section, we first formally describe the personalized video highlight detection problem. Then we briefly go over the overview of the proposed PAC-Net framework and describe the details of the feature encoder. Finally, we elaborate on how the DBC enables personalization and how to utilize user history more precisely and flexibly with the Mi-Hi mechanism.

3.1 Problem Definition

Personalized video highlight detection (P-VHD) is a sub-problem of video highlight detection (VHD), but it is more practical and challenging than VHD. Thus, we first introduce the VHD then P-VHD for better understanding. The goal of VHD is to, for each frame in the input video, correctly predict the likelihood of its being highlight. We denote the input as $V_{in} \in \mathbb{R}^N$, where N denotes the number of frames. Each frame is annotated with 1 if is highlight, otherwise 0. The VHD problem is essentially a binary classification problem, that is, given a entire video, classifying each frame into either highlight or non-highlight category:

$$\text{VHD}(v_i | V_{in}) = \begin{cases} 1, & \text{if } v_i \in \text{highlight} \\ 0, & \text{otherwise} \end{cases}, \quad (1)$$

where v_i denotes the feature of the i -th frame.

Compared to the general VHD, P-VHD task further requires the highlight detector to be adaptive to the different users' preferences, and such preferences are given by a set of history videos $\{H_1, H_2, \dots, H_M\}$ with highlight annotations of each user :

$$\text{P-VHD}(v_i | V_{in}, \{H_1 \dots H_M\}) = \begin{cases} 1, & \text{if } v_i \in \text{highlight} \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

where M denotes the number of history videos.

3.2 Overview of PAC-Net

General VHD approaches do not consider the user history so they tend to highlight the common flavors of all the users. To achieve the goal of personalization, we propose the PAC-Net, which consists of encoder \mathcal{F} , Decision Boundary Customizer (DBC) module, and Mini-History (Mi-Hi) mechanism. The key idea of PAC-Net is that the standards for determining the highlights can be extracted from the user history and the extracted standards can be used to pick out the highlighted frames in the input query video. We represent each preference standard with a classifier that draws a decision boundary between the highlighted frames and non-highlighted frames. Such process is conducted by the DBC module, which is designed to convert user history into the parameters of the preference-specific highlight classifier.

In addition, prior P-VHD works view the history as holistic information by summarizing all user history into one preference vector, which is oversimplified and sub-optimal since one user could have diverse preferences let alone the segments in the same video could have multiple topics. To make the preferences more fine-grained, we propose the Mini-History (Mi-Hi) mechanism, where each highlight segment in history is picked out and converted to a preference-dependent highlight classifier by DBC, and the final highlight prediction for the input video is jointly decided by multiple preference-specific classifiers.

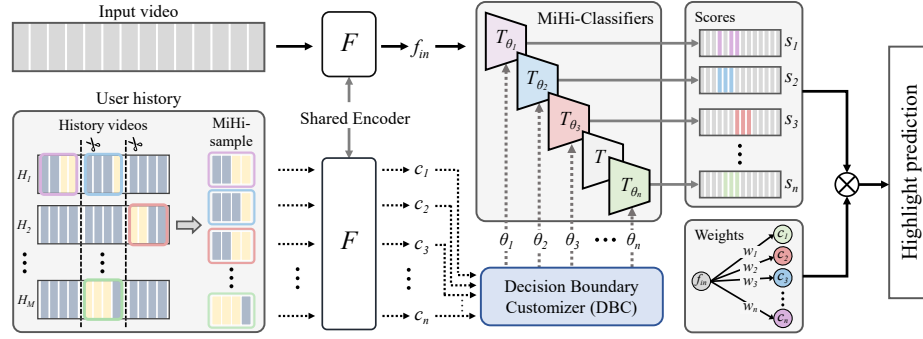


Fig. 2: **PAC-Net architecture.** PAC-Net consists of the encoder \mathcal{F} , DBC module, and Mi-Hi mechanism (highlighted with the gray background). In particular, DBC converts the preference-related history features c into the parameters θ of the highlight classifiers T_θ . And Mi-Hi mechanism corresponds to (1) splitting user history into multiple Mi-Hi samples, (2) multiple Mi-Hi highlight classifiers generated by DBC, and (3) results fusion to get final prediction.

The whole pipeline of the PAC-Net framework is shown in Figure 2. First, we split the history videos into equal-length segments and select the ones contained both highlighted frames and non-highlighted frames as Mi-Hi samples. Then, the input video and these Mi-Hi samples are fed into the shared encoder \mathcal{F} and become input feature f_{in} and Mi-Hi features c_i . Next, DBC converts each Mi-Hi feature c_i into the parameters θ_i of a Mi-Hi highlight classifier T_{θ_i} . We apply all the Mi-Hi classifiers to input feature f_{in} and each one will predict a set of frame-level highlight scores. Finally, the final highlight result is produced by the weighted sum of predictions from all Mi-Hi classifiers.

3.3 Encoder

Following previous work [29], we use the fixed C3D [39] features for both training and testing to reduce the computational burden. The features are extracted by a C3D model pre-trained on the Sports-1M [14] dataset. Since the C3D [39] features are fixed, we further introduce an encode \mathcal{F} to map the fixed input feature into a new feature space to facilitate later learning. Specifically, a standard U-Net [31] without the last fully connected layer serves as the encoder. The U-Net consists of 4 downsampling blocks and 4 upsampling blocks, and each block is composed of 2 temporal convolution layers.

For an input video $V_{in} \in \mathbb{R}^N$ with N frames, we first extract its C3D features $f_{C3D} \in \mathbb{R}^{N \times d}$, where d is the feature dimension. The encode performs temporal 1D convolution along time dimension to propagate temporal information. We denote the mapped feature as f_{in} , thus

$$f_{in} = \mathcal{F}(f_{C3D}) \in \mathbb{R}^{N \times d}. \quad (3)$$

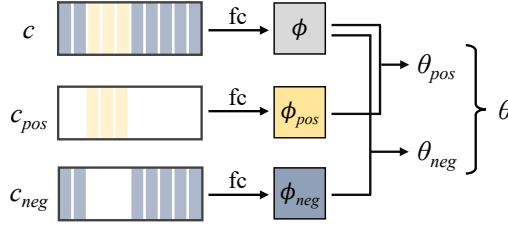


Fig. 3: **DBC architecture.** DBC takes three kinds of inputs and each one goes through a single non-shared fully connected layer. See Equation 5 for details about how the output θ is constructed.

Note that C3D features only encode local motion information, and the encoder \mathcal{F} equips features with contextual semantics.

3.4 Decision Boundary Customizer

Given a video with labels on the highlights, a person can easily summarize a standard to distinguish these highlights. And in PAC-Net, Decision Boundary Customizer (DBC) is designed for the same purpose, that is, building a customized classifier given a history clip. Briefly speaking, DBC learns to convert the feature c of a history clip into the parameters θ of the classifier T_θ ,

$$\text{DBC: } c \rightarrow \theta. \quad (4)$$

The feature c of the historical segment contains a specific user preference, and the generated classifier T_θ represents the corresponding personal preference.

Figure 3 gives a graphical illustration of the proposed DBC. Given a historical highlight segment $c \in \mathbb{R}^{L \times d}$ with L frames, we first pick out the highlighted-frames-only segment c_{pos} and the non-highlighted-frames-only segment c_{neg} based on the user’s label in order to explicitly input the highlight and non-highlight information to DBC. Then we pool all three features along the time dimension and each of them is fed into a non-shared fully connected layer. We denote these intermediate results as ϕ , ϕ_{pos} , and $\phi_{neg} \in \mathbb{R}^d$ respectively. Finally, the parameters $\theta \in \mathbb{R}^{d \times 2}$ of the classifier T_θ is generated by

$$\begin{aligned} \theta &= [\theta_{pos}; \theta_{neg}], \\ \theta_{pos} &= \phi + \phi_{pos}, \\ \theta_{neg} &= \phi + \phi_{neg}. \end{aligned} \quad (5)$$

The generated classifier T_θ is actually a single fully connected layer, which takes $f_{in} \in \mathbb{R}^{N \times d}$ as input and yields the prediction $T_\theta(f_{in}) \in \mathbb{R}^{N \times 2}$ with the second dimension denotes the likelihood of being highlight and non-highlight.

During training, it is not guaranteed that DBC will learn to generate diverse classifiers. In fact, it is easier for the DBC to generate only the classifiers that

capture the general preferences. For instance, assume the input video is about baseball and there are three history videos about baseball, tennis, and golf respectively, the DBC will not receive any punishment by generating three similar classifiers that focus on swinging. This might potentially cause the model to degenerate to a general highlight detector. Thus, to enforce the DBC to focus on more detailed preferences and generate diverse classifiers, we propose a regularization module during training. Specifically, in addition to the input feature, we also apply the generated classifier to the corresponding historical feature that generates this highlight classifier and use cross-entropy loss to provide supervision. And the formula of the regularization loss is

$$\mathcal{L}_{reg} = \text{CE}(T_{\theta}(c), Y_c), \quad (6)$$

where Y_c denotes the set of frame-level labels of c .

3.5 Mini-History

It's common sense that in real world a user could have diverse interests, which can be summarized from user history. Previous methods simply average the history so that one user would only have a holistic user-level history feature. We argue that user-level history feature might not be able to sufficiently represent the user preference since the user history could have diverse topics. Therefore, we introduce the Mini-History (Mi-Hi) mechanism that utilizes the history in a more fine-grained level. There are three steps in the Mi-Hi mechanism: (1) splitting the history videos into Mi-Hi samples, (2) converting each Mi-Hi sample into corresponding preference-specific Mi-Hi classifier, and (3) fusing the predictions of all the Mi-Hi classifiers.

In specific, we split each history video into equal-length segments and keep only the ones that contain both highlighted and non-highlighted frames. We regard each kept segment as a Mi-Hi sample, and other segments are not used for training. Every Mi-Hi sample will get through the encoder \mathcal{F} and get its feature c_i , which is then fed into the DBC to generate the parameters θ_i of a preference-specific highlight classifier T_{θ_i} , called Mi-Hi classifier,

$$\theta_i = \text{DBC}(c_i). \quad (7)$$

Thus, if a user provides multiple Mi-Hi samples, there will be more than one highlight classifiers $\{T_{\theta_1} \dots T_{\theta_n}\}$. n denotes the number of Mi-Hi samples, and the number of Mi-Hi samples is greater or equal to the number of history videos ($n \geq M$) since a history video could contain more than one highlight segment. Each Mi-Hi sample corresponds to one specific Mi-Hi classifier. Applying each Mi-Hi classifier to the input video yields a prediction s_i . We calculate the weighted-sum of all the s_i to get the final highlight prediction,

$$s = \sum_i^n w_i \cdot s_i, \quad s_i = T_{\theta_i}(f_{in}). \quad (8)$$

And the weight w_i is computed by applying the radial basis function (RBF) kernel on the pooled input feature and pooled Mi-Hi feature followed by a SoftMax function:

$$\mathcal{K}(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\tau^2}\right), \quad (9)$$

$$w_i = \sigma\left(\mathcal{K}(\text{Pool}(c_i), \text{Pool}(f_{in}))\right), \quad (10)$$

where $\mathcal{K}(\cdot, \cdot)$ represents the RBF kernel, and τ is the temperature hyper-parameter. $\text{Pool}(\cdot)$ is a global pooling function along the time dimension, which converts a highlight $c \in \mathbb{R}^{L \times d}$ with L frames into $\text{Pool}(c) \in \mathbb{R}^d$, where d is feature dimension. σ denotes the SoftMax function.

Finally, the cross entropy between the prediction s and labels Y serves as the classification loss

$$\mathcal{L}_{cls} = \text{CE}(s, Y). \quad (11)$$

Besides, since there are multiple Mi-Hi features that input into the DBC, the regularization loss in Eq. 6 now becomes

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_i^N \text{CE}(T_\theta(c_i), Y_{c_i}). \quad (12)$$

The overall training loss is the sum of \mathcal{L}_{cls} and \mathcal{L}_{reg} ,

$$\min \mathcal{L}_{cls} + \mathcal{L}_{reg}. \quad (13)$$

4 Experiments

In this section, we conduct extensive experiments on the largest benchmark dataset to verify the effectiveness of the proposed PAC-Net.

4.1 Experimental Setup

Dataset. All the experiments are conducted on the PHD-GIFs [24] dataset, which is by far the largest public highlight detection dataset with personalized highlight annotation, *i.e.*, user history. The original dataset contains 119,938 YouTube videos with 222,015 annotations from a total of 13,822 users, among which, 850 users are selected for testing. There are at least five history videos for each user, where the last video is reserved for the test set. Testing videos are clamped to a proper length to avoid extreme scenarios.

PHD-GIFs dataset only provides a list of YouTube URLs. Since some videos are no longer available on YouTube, we can not obtain the whole dataset. This problem has also been reported in Adaptive-H-FCSN [29]. This causes that the previous two works and ours conduct experiments on three different versions of the PHD-GIFs dataset. Thus, we collect several statistics of the dataset of each

Table 1: **Statistics of three versions of the PHD-GIFs dataset.** The dataset used in PHD-GIFs [24], Adaptive-H-FCSN[29], and ours is denoted as D_{v1} , D_{v2} , and D_{v3} respectively. The difference is caused by that the invalid YouTube URLs increase gradually.

Dataset	Users _(train+val)	Users _{test}	Total videos	Videos per user
D_{v1} [24]	12972	850	119938	9.25
D_{v2} [29]	7818	727	104828	12.27
D_{v3} (Ours)	10146	675	95250	8.80

Table 2: **Comparisons with state-of-the-art methods.** We report the performance on three versions of PHD-GIFs dataset separately. On D_{v1} and D_{v2} , we report the results in their papers. On D_{v3} , we train the models marked with * with their official open-source code.

Methods	Params(M)	Dataset	mAP(%)
Video2GIF [9]	2.23	D_{v1}	16.68
PHD-GIFs [24]	-		16.68
Video2GIF [9]	2.23	D_{v2}	14.75
FCSN [30]	23.90		15.22
Adaptive-H-FCSN [29]	197.36		16.73
Video2GIF	2.23	D_{v3}	13.82
FCSN*	23.90		15.09
Adaptive-H-FCSN*	197.36		16.04
PAC-Net	5.89		17.51

paper for clear comparison as shown in Table 1. We denote the dataset in PHD-GIFs [24], Adaptive-H-FCSN [29], and ours as D_{v1} , D_{v2} , and D_{v3} respectively. Note that our version has the smallest video-user ratio.

Evaluation Metric. Following the former works [9,24,29] in video highlight detection, we also use the mean Average Precision (mAP) as the evaluation metric. The mAP summarizes the precision-recall curve of the detection results and has been extensively used in object detection and retrieval tasks to measure the accuracy of the model. We report the mAP on the test set. And the mAP is first calculated separately for each testing video and finally averaged.

Implementation Details. In all experiments, temperature parameter τ in RBF kernel and length of the Mi-Hi sample are set to 0.05 and 256 respectively. We train the model for 20 epochs with the Adam [17] optimizer (learning rate: 0.001, weight decay: 1×10^{-4}). The learning rate and weight decay are set to 0.001 and 0.0001 separately. And the training usually takes 8 to 9 hours on 8 Nvidia Tesla V100 GPUs. Our method is implemented with the PyTorch [27] deep learning framework, and the source code will be released for reproducibility.

Table 3: **Detailed improvement of the PAC-Net over baseline.**

Gains (mAP)	≤ 0	$0.0 \sim 2.0$	$2.0 \sim 5.0$	≥ 5.0
Proportion (%)	2.3	48.9	35.2	13.6

Table 4: **Ablation study experiments.** PAC-Net(*) denotes the different modification based on our method. PR-Net_{full} is our full method.

Methods	mAP(%)
PAC-Net _G	15.28
PAC-Net _{Agg}	15.67
PAC-Net _{DBC}	15.89
PAC-Net _{MiHi}	16.23
PAC-Net _H	15.66
PAC-Net _{full}	17.51

4.2 Results

We compare our method with several previous state-of-the-art approaches including Video2GIF [9], FCSN [30], PHD-GIFs [24], and Adaptive-H-FCSN [29]. The first two are general video highlight detectors while the rest are personalized detectors. As we mentioned before, the datasets used in different papers are not exactly the same due to some YouTube URLs gradually becoming invalid. Therefore, for fair comparisons, we re-train previous methods with their official open-source code on our dataset.

As shown in Table 2, our method achieves state-of-the-art results and surpasses the second-best by 1.47 mAP, with a relative improvement of 9% over Adaptive-H-FCSN. Besides, Adaptive-H-FCSN performs worse on our dataset, and only obtains +0.95 mAP gain over FCSN. Such undesirable performance degradation implies that our dataset is more difficult and more challenging. Note that, despite our dataset having the smallest video-user ratio, we still get the highest mAP. We also report the number of parameters of each method. Our model has around thirty times fewer parameters than Adaptive-H-FCSN, but achieves the best performance, which illustrates the superiority of PAC-Net. Finally, In Table 3, We analyze the detailed performance improvement ratios of different users. From the table, we can see the vast majority of users have an obvious gain over baseline, which proves the effectiveness of our method.

4.3 Analysis

To evaluate the effectiveness of the different components in our proposed model and justify several design choices, comprehensive ablation study experiments are conducted on the PHD-GIFs dataset.

We first explore the usage of user history, and get two variants of PAC-Net named **PAC-Net_G** and **PAC-Net_{Agg}**, where the former works as a baseline

Table 5: **Impact of number of user history.** The number denotes the maximum number of historical videos that each user can use and M means no restriction is imposed.

# of hist.	0	1	2	3	5	M
mAP	15.28	15.53	15.84	16.38	17.02	17.51

Table 6: **F-Score (%) performance comparison on SumMe [7] dataset.** Note that unlike other SOTA methods, H-FCSN and our PAC-Net_G are trained on the PHD-GIFs without fine-tuning on SumMe.

Methods	GAN _{sup} [22]	DR-DSN _{sup} [56]	S ² N[43]	H-FCSN[29]	PAC-Net _G
F-score	41.7	42.1	43.3	44.4	44.6

generic model without the use of user history and the latter directly concatenate input feature with the user-level average history feature. Both PAC-Net_G and PAC-Net_{Agg} are composed of a feature encoder and a generic classifier for highlight prediction; To evaluate the effect of the major components in our method, two variants are studied: using DBC module only (**PAC-Net_{DBC}**) and using Mi-Hi mechanism only (**PAC-Net_{MiHi}**); To evaluate the effect of non-highlighted history, we remove the use of non-highlights of DBC and get the **PAC-Net_H** variant.

Improvement of Individual Module. Our main contributions can be summarized in two folds: DBC module for personalization and Mini-History for fine-grained utilization of user history. Therefore, we show how much each contribution improves the performance in Table 4. PAC-Net_G does not utilize user history and serves as the baseline.

PAC-Net_{MiHi} improves the baseline by 0.95 mAP, where Mi-Hi features and input features are trained with a shared classifier, whose parameters are not adaptive to users. Besides, to prove the performance gain does not all come from incorporating the user history into the model, we conduct another experiment (PAC-Net_{Agg}) where the input feature is concatenated with averaged history feature. PAC-Net_{MiHi} still surpasses PAC-Net_{Agg} by 0.56 mAP, which shows the fine-grained utilization of user history is more effective. In the PAC-Net_{DBC}, all the history videos of a user are concatenated into one giant history video, and the DBC is required to construct only one user-level classifier. Compared with the learned-and-fixed classifier in the baseline, DBC brings 0.61 mAP improvement. Finally, we show the Mini-History and DBC are complementary to each other. Based on Mini-History, DBC could generate various fine-grained Mi-Hi classifiers, which are more flexible and precise, and the full model could provide personalized highlight detection via the integration of multiple preference-specific decision boundaries. As a result, combining DBC with Mini-History significantly improves the baseline by 2.23 mAP, which is even greater than the sum of the improvements of each individual component.

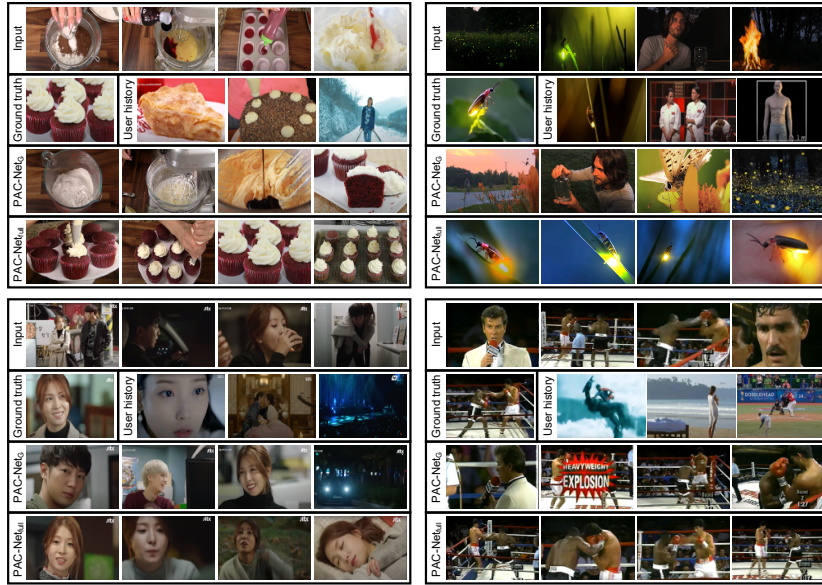


Fig. 4: **Qualitative comparisons between PAC-Net_G and PAC-Net.** We provide a few representative frames from input video and each historical highlight, respectively. We also show the highlighted frames detected by PAC-Net_G and PAC-Net_{full}. These visualizations show that the user-adaptive model well extracts user’s preference from user history, while general model PAC-Net_G fails.

Importance of Non-Highlighted History. Another difference from the previous work is that we are the first to utilize the non-highlighted history. To prove that non-highlighted history is beneficial to P-VHD, we perform an ablation study on whether to utilize the history non-highlights. As shown in Table 4, without non-highlighted history, the performance of PAC-Net_H drops by 1.95 mAP. We think the usage of historical non-highlights plays an important role in defining a more precise decision boundary, which helps to eliminate the false positives and therefore makes the full PAC-Net model achieve better performance.

Discussion of the Number of User History Videos. In PHD-GIFs dataset, users do not have the same number of history videos and this is usually the case in practice. Hence, to test how much the number of user history affects the performance, we set several fixed upper-bounds of the max allowed history videos per user. Table 5 shows that as the number of allowed history videos increases, the performance gradually increases as expected, which demonstrates the good potential of our method. The trend indicates that as we acquire more history videos, more accurate user preferences can be built, which results in better performance.

Application to Video Summarization. Since highlight detection is highly related to video summarization, we also test our generic variant PAC-Net_G on video summarization task. Following Adaptive-H-FCSN [29], we evaluate PAC-

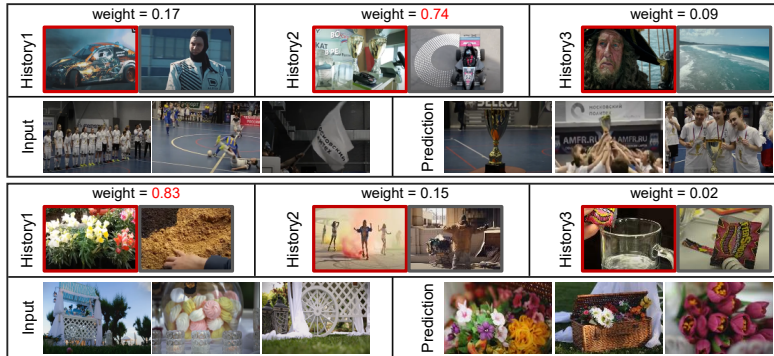


Fig. 5: **Visualizations of the classifier weights during fusion.** On the upper side, we choose one representative sample from each historical highlight. On the lower side, representative frames are selected from the input video and prediction respectively. As expected, the *trophy* and *flower* classifiers play the most importance part in two examples separately.

Net_G (trained on the PHD-GIFs [24] dataset) on *SumMe* [7] dataset without fine-tuning. Table 6 compares the F-Score of our PAC-Net_G and several state-of-the-art video summarization methods. It can be observed that the proposed PAC-Net_G achieves the best performance, verifying its effectiveness.

4.4 Visualizations

In Figure 4, we provide several qualitative results as examples to compare the baseline PAC-Net_G and the full PAC-Net model. In these examples, PAC-Net_{full} obviously outperforms the baseline PAC-Net_G. For example, the user in the second case likes the firefly and PAC-Net_{full} successfully highlights several close-ups of the firefly while the baseline fails. The visualization results illustrate that the fine-grained user preferences contained in user history play an important role for better personalized video highlight predictions.

Moreover, our fusion strategy assigns different weights to different Mi-Hi highlight classifiers. In Figure 5, we visualize the focus of each classifier as well as its associated weight. It clearly shows that the highlight predictions are highly related to the Mi-Hi highlight classifier with the highest score.

5 Conclusion

We propose PAC-Net, a novel personalized video highlight detection framework, which extracts users’ preferences from their history videos and makes highlight predictions based on the extracted preferences. As two core components of PAC-Net, DBC converts the user history into the parameters of the classifier, while Mi-Hi mechanism enables to utilize the user history at a fine-grained level. Extensive experiments and analytical studies verify the superiority of our approach.

References

1. Agnihotri, L., Kender, J., Dimitrova, N., Zimmerman, J.: Framework for personalized multimedia summarization. In: ACM International Conference on Multimedia (MM) (2005)
2. Babaguchi, N., Ohara, K., Ogura, T.: Learning personal preference from viewer's operations for browsing and its application to baseball video retrieval and summarization. *IEEE Transactions on Multimedia* **9**(5), 1016–1025 (2007)
3. Badamdorj, T., Rochan, M., Wang, Y., Cheng, L.: Joint visual and audio learning for video highlight detection. In: IEEE International Conference on Computer Vision (ICCV) (2021)
4. Badamdorj, T., Rochan, M., Wang, Y., Cheng, L.: Contrastive learning for unsupervised video highlight detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
5. Chu, W.S., Song, Y., Jaimes, A.: Video co-summarization: Video summarization by visual co-occurrence. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
6. Gong, B., Chao, W.L., Grauman, K., Sha, F.: Diverse sequential subset selection for supervised video summarization. In: Advances in Neural Information Processing Systems (NeurIPS) (2014)
7. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: European Conference on Computer Vision (ECCV) (2014)
8. Gygli, M., Grabner, H., Van Gool, L.: Video summarization by learning submodular mixtures of objectives. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
9. Gygli, M., Song, Y., Cao, L.: Video2gif: Automatic generation of animated gifs from video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
10. Hong, F., Huang, X., Li, W., Zheng, W.: Mini-net: Multiple instance ranking network for video highlight detection. In: European Conference on Computer Vision (ECCV) (2020)
11. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
12. Jaimes, A., Echigo, T., Teraguchi, M., Satoh, F.: Learning personalized video highlights from detailed mpeg-7 metadata. In: IEEE Conference on Image Processing (ICIP) (2002)
13. Jiao, Y., Yang, X., Zhang, T., Huang, S., Xu, C.: Video highlight detection via deep ranking modeling. In: Pacific-Rim Symposium on Image and Video Technology (PSIVT) (2017)
14. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
15. Khosla, A., Hamid, R., Lin, C.J., Sundaresan, N.: Large-scale video summarization using web-image priors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
16. Kim, G., Xing, E.P.: Reconstructing storyline graphs for image recommendation from web community photos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)

17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
18. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
19. Liu, W., Mei, T., Zhang, Y., Che, C., Luo, J.: Multi-task deep visual-semantic embedding for video thumbnail selection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
20. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
21. Ma, Y.F., Hua, X.S., Lu, L., Zhang, H.J.: A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia* **7**(5), 907–919 (2005)
22. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
23. Garcia del Molino, A., Boix, X., Lim, J.H., Tan, A.H.: Active video summarization: Customized summaries via on-line interaction with the user. In: Association for the Advancement of Artificial Intelligence (AAAI) (2017)
24. Garcia del Molino, A., Gygli, M.: Phd-gifs: personalized highlight detection for automatic gif creation. In: ACM International Conference on Multimedia (MM) (2018)
25. Ngo, C.W., Ma, Y.F., Zhang, H.J.: Automatic video summarization by graph modeling. In: IEEE International Conference on Computer Vision (ICCV) (2003)
26. Panda, R., Roy-Chowdhury, A.K.: Collaborative summarization of topic-related videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
27. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NeurIPS Workshop (2017)
28. Plummer, B.A., Brown, M., Lazebnik, S.: Enhancing video summarization via vision-language embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
29. Rochan, M., Reddy, M.K.K., Ye, L., Wang, Y.: Adaptive video highlight detection by learning from user history. In: European Conference on Computer Vision (ECCV) (2020)
30. Rochan, M., Ye, L., Wang, Y.: Video summarization using fully convolutional sequence networks. In: European Conference on Computer Vision (ECCV) (2018)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) (2015)
32. Rui, Y., Gupta, A., Acero, A.: Automatically extracting highlights for tv baseball programs. In: ACM International Conference on Multimedia (MM) (2000)
33. Sharghi, A., Gong, B., Shah, M.: Query-focused extractive video summarization. In: European Conference on Computer Vision (ECCV) (2016)
34. Singla, A., Tschitschek, S., Krause, A.: Noisy submodular maximization via adaptive sampling with applications to crowdsourced image collection summarization. In: Association for the Advancement of Artificial Intelligence (AAAI) (2016)
35. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)

36. Sun, M., Farhadi, A., Seitz, S.M.: Ranking domain-specific highlights by analyzing edited videos. In: European Conference on Computer Vision (ECCV) (2014)
37. Takahashi, Y., Nitta, N., Babaguchi, N.: User and device adaptation for sports video content. In: IEEE International Conference on Multimedia and Expo (ICME) (2007)
38. Tang, H., Kwatra, V., Sargin, M.E., Gargi, U.: Detecting highlights in sports videos: Cricket as a test case. In: IEEE International Conference on Multimedia and Expo (ICME) (2011)
39. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: IEEE International Conference on Computer Vision (ICCV) (2015)
40. Truong, B.T., Venkatesh, S.: Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications* **3**(1), 3-es (2007)
41. Vasudevan, A.B., Gygli, M., Volokitin, A., Van Gool, L.: Query-adaptive video summarization via quality-aware relevance estimation. In: ACM International Conference on Multimedia (MM) (2017)
42. Wei, F., Wang, B., Ge, T., Jiang, Y., Li, W., Duan, L.: Learning pixel-level distinctions for video highlight detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
43. Wei, Z., Wang, B., Nguyen, M.H., Zhang, J., Lin, Z.L., Shen, X., Mech, R., Samaras, D.: Sequence-to-segment networks for segment detection. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
44. Xiong, Z., Radhakrishnan, R., Divakaran, A., Huang, T.S.: Highlights extraction from sports video based on an audio-visual marker detection framework. In: IEEE International Conference on Multimedia and Expo (ICME) (2005)
45. Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J.M., Singh, V.: Gaze-enabled ego-centric video summarization via constrained submodular maximization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
46. Xu, M., Wang, H., Ni, B., Zhu, R., Sun, Z., Wang, C.: Cross-category video highlight detection via set-based learning. In: IEEE International Conference on Computer Vision (ICCV) (2021)
47. Yang, H., Chaisorn, L., Zhao, Y., Neo, S.Y., Chua, T.S.: Videoqa: question answering on news video. In: ACM International Conference on Multimedia (MM) (2003)
48. Yao, T., Mei, T., Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
49. Yao, T., Mei, T., Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
50. Yu, Y., Lee, S., Na, J., Kang, J., Kim, G.: A deep ranking model for spatio-temporal highlight detection from a 360o video. In: Association for the Advancement of Artificial Intelligence (AAAI) (2018)
51. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Summary transfer: Exemplar-based subset selection for video summarization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
52. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: European Conference on Computer Vision (ECCV) (2016)
53. Zhang, K., Grauman, K., Sha, F.: Retrospective encoders for video summarization. In: European Conference on Computer Vision (ECCV) (2018)

- 54. Zhang, Y., Kampffmeyer, M., Liang, X., Tan, M., Xing, E.P.: Query-conditioned three-player adversarial network for video summarization. In: British Machine Vision Conference (BMVC) (2018)
- 55. Zhao, B., Li, X., Lu, X.: Hierarchical recurrent neural network for video summarization. In: ACM International Conference on Multimedia (MM) (2017)
- 56. Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: Association for the Advancement of Artificial Intelligence (AAAI) (2018)