MVVA-net: a video aesthetic quality assessment network with cognitive fusion of multi-type feature–based strong generalization.

LI, M., WANG, Z., REN, J. and SUN, M.

2022

MVVA-Net: A Video Aesthetic Quality Assessment Network With Strong Generalization Based On Multi-type Features

Min Li · Zheng Wang* · Jinchang Ren · Meijun Sun

* **Corresponding author : Email: wzheng@tju.edu.com**

Received: date/ Accepted

Structural abstract

Background

Although short videos have been widely popular on various social-media platforms, it is still a challenging research issue to evaluate the aesthetic quality of these videos.

Methods

In this paper, we construct a large-scale and properly annotated short video aesthetics (SVA) dataset. We further propose a multi-type feature fusion network (MVVA-Net) for video aesthetic quality assessment. MVVA-Net consists of two branches: intra-frame aesthetics branch and inter-frame aesthetics branch. These two branches take different types of video frames as input. The inter-frame aesthetic branch extracts the inter-frame aesthetic features based on the sequential frames extracted at fixed intervals, and the intra-frame aesthetic branch extracts the intra-frame aesthetic features based on the key frames extracted by the inter-frame difference method. Through the adaptive fusion of inter-frame aesthetic features and intra-frame aesthetic features, the video aesthetic quality can be effectively evaluated. At the same time, MVVA-Net has no fixed number of input frames, which greatly enhances the generalization ability of the model.

Results

We performed quantitative comparison and ablation studies. The experimental results show that the two branches of MVVA-Net can effectively extract the intra-frame aesthetic features and inter-frame aesthetic features of different videos. Through the adaptive fusion of intra-frame aesthetic features and inter-frame aesthetic features for video aesthetic quality Assessment, MVVA-Net achieves better classification performance and stronger generalization ability than other methods.

Conclusions

In this paper, we construct a dataset of 6900 video shots, and propose a video aesthetic quality Assessment method based on non-fixed model input strategy and multi-type features. Experimental results show that the model has a strong generalization ability and achieved a good performance on different datasets.

Min Li · Zheng Wang ·Meijun Sun

College of Intelligence and Computing, Tianjin University, Tianjin Key Lab of Machine Learning, Tianjin, China

* Zheng Wang is the corresponding author. Email: wzheng@tju.edu.com

Jinchang Ren

Department of Electronic and Electrical Engineering, University of Strathclyde

1 Introduction

Video aesthetic quality assessment aims to predict the aesthetic score of the video. Nowadays, short video is widely popular on various platforms. It is very important to evaluate the aesthetic quality of video for video classification, video recommendation, auxiliary video editing [1] and video generation. For example, Kuang et al. [6] proposed a multi-stream framework for video grading and professional segment detection.

Previous studies have made some progress in the field of video aesthetic quality assessment, but there are still several problems in this field. The first is the lack of a sufficient large and properly annotated dataset, and the second existing models and algorithms still shows limited efficacy for video aesthetics quality assessment. These are discussed in detail as follows.

Although several previous studies on video aesthetic quality evaluation have established some datasets for video aesthetic quality evaluation, most of the datasets are not publicly. At the same time, some datasets directly label high-rated videos downloaded from websites or videos taken by professionals as high aesthetic quality, and low-rated videos downloaded from websites or videos taken by amateurs as low aesthetic quality. These video sites allow users to rate videos, but users may not rate videos based on their aesthetic quality. The professionalism of the video photographer does not represent the aesthetic quality of the video. Therefore, this way of labeling is unscientific.

According to the principle of persistence of vision, it can be known that when the continuous image changes more than 24 frames per second, the human eye cannot distinguish a single static image. Such a continuously changing image is called a video. The beauty of the frame in the video is called intra-frame aesthetics, and the visual effects of continuous changes in the video are called inter-frame aesthetics. Evaluating the aesthetic quality of the video is to evaluate the aesthetic quality of intra-frame and inter-frame. Most of the existing methods only consider the intra-frame aesthetic features [2,6,7] or inter-frame aesthetic features [5] unilaterally and evaluate the aesthetic quality of video by using the intra-frame aesthetic features or inter-frame aesthetic features. These methods do not comprehensively take into account the intra-frame aesthetic quality and the inter-frame aesthetic quality, so they do not achieve good results.

When the existing deep learning method [12-17] processes video, it mostly uses a fixed number of frames extracted from the video to replace the video. Since the range of video duration that the dataset can cover is limited, training the model with a fixed number of frames will make the model tend to fit the data of the current duration interval. When it is generalized to other time interval data, using a fixed number of frames can not effectively extract the characteristics of the video, that is, using a fixed number of frames to train the model in a certain time interval is difficult to generalize to other time intervals.

To solve the above problems, we build a video aesthetics dataset with 6900 videos, which is large-scale and scientifically labeled. According to the aesthetic quality score given by the viewer, the video is labeled as high aesthetic quality and low aesthetic quality. Based on this dataset, a multi-type video aesthetics network (MVVA-Net) is proposed to evaluate the video aesthetics quality more effectively, considering the video intra-frame aesthetics and inter-frame aesthetics comprehensively. Two branches are designed in MVVA-Net, namely the intra-frame aesthetic branch and the inter-frame aesthetic branch. These two branches take different types of video frames, namely key frames and sequence frames, as input to extract the intra-frame and inter-frame aesthetic features of the video. Through adaptively fusing different types of features, high-quality and low-quality videos can be effectively distinguished. At the same time, our network does not limit the number of input frames, allowing different videos to have a different number of frames as input, so that our model can effectively extract the features of different videos, greatly enhancing the generalization ability of the model.

The experimental results show that our method is superior to other methods on the open aesthetics dataset. In addition, we conducted a series of ablation experiments, and the results showed that each part of our model has a positive effect on the final experimental results.

The main contributions of this work can be summarized as follows:

- We constructed a scientifically labeled video aesthetic quality assessment dataset containing 6900 videos. As far as we know, this is the largest dataset in the field of video aesthetics research, and the annotation process is more scientific than other datasets.
- An MVVA-Net including an intra-frame aesthetic branch and an inter-frame aesthetic branch is proposed to evaluate the aesthetic quality of video by fusing video multi-type features.
- The MVVA-Net has good generalization capability and performs well on different videos.

2 Related work

We divide the discussions of related works into the following two subsections.

2.1 Video Aesthetic Quality Assessment

Traditional video aesthetic quality assessment methods use hand-made aesthetic features to distinguish the aesthetic quality of the video. Moorthy et al. [3] proposed and evaluated a set of low-level features, selected the most discriminative seven features, and successfully classified high aesthetic quality video and low aesthetic quality video. In [4], Yang et al. compared the evaluation accuracy between two different semantic types of features and found that the accuracy of features not related to semantics is more reliable. [5] explores a method to evaluate the aesthetic quality of video by analyzing key motion features. Niu et al. [6] studied the general aesthetic features for still photos and extended them to video. The author of [9] combined photo-based and motion-based visual cues and proposed a time-sequential perception framework that integrates frame-based features to further improve the evaluation accuracy by considering time-varying attributes. Tzelepis et al. [10] used the information obtained from the low-level and high-level analysis of the video layout to evaluate the aesthetic quality of the video.

Different from the traditional methods, [11] proposed a deep multimodal learning method for video aesthetic quality assessment by using deep learning to automatically extract aesthetic features. This method uses the aesthetic attributes of multiple modes to evaluate the aesthetic quality of videos and has achieved good results.

However, the above method does not take into account intra-frame aesthetics and inter-frame aesthetics, and cannot adapt to videos of different lengths. In contrast, our method considers the relationship between intra-frame aesthetics, inter-frame aesthetics and video aesthetics, and does not limit the total number of input frames, which greatly enhances the generalization ability of the model.

2.2 Dataset

Moorthy et al. [3] established a Telefonica dataset containing 160 short consumer videos. Each video in this dataset is 11 to 60 seconds. Among them, 80 videos are of high aesthetic quality, and the other 80 are of low aesthetic quality. The authors in [6] established a dataset, including 1000 professional videos collected from 16 feature films and 34 commercial TV programs, and 1000 amateur videos shot by 23 amateur users. [8] builds an ADCCV dataset to enhance the Telefonica dataset by adding more positive examples. The dataset of [9] consists of 1000 professionally produced video clips, each of which is about 1 minute. The CERTH-ITI-VAQ700 [10] dataset consists of 700 videos with a duration of 1 to 6 minutes. Each video is classified into high aesthetic quality or low aesthetic quality according to the aesthetic quality scores given by multiple annotators. The AVAQ6000 [11] dataset established contains 6000 videos, each of which is less than 1 minute. The videos are divided into professional videos and amateur videos according to the source.

However, some of the above datasets are based on the professionalism of video shooting and the ratings of video website users as the evaluation criteria of the aesthetic quality of videos, which is unscientific. Due to the lack of a controlled environment when scoring, users will to a large extent rate videos based on factors other than aesthetic factors. At the same time, the professionalism of video shooting does not represent the aesthetic quality of the video. In contrast, the public dataset we provide is fully labeled based on the annotator's ratings of the aesthetic quality of the video. This controlled situation ensures the scientific annotation of our dataset.

3 Data Collection

Table 1. Datasets for video aesthetic quality assessment.

Dataset	Labeling basis	Number of videos	Public situation	Duration
Telefonica [3]	Aesthetic quality	160	Private	11s-1m
Niu [6]	Professionalism of the shooting or website rating	2000	Private	—
ADCCV [8]	Professionalism of the shooting or website rating	200	Private	11s-1m
NHK [9]	Professionalism of the shooting or website rating	1000	Private	1m
CERTH-ITI-VAQ700 [10]	Aesthetic quality	700	Public	1m-6m
AVAQ6000 [11]	Professionalism of the shooting or website rating	6000	Public	<1m

Table 1 shows the statistics of existing video aesthetic quality assessment datasets. It can be seen that there are three problems with the existing dataset.

- Private: Most of the existing video aesthetic quality assessment datasets (1st row-4th row) are not public.
- Insufficient data: Some datasets (1st row-5th row) are not large enough, which makes the trained depth model perform poorly.
- The annotation method is unscientific: Some datasets (2nd row-4th row, 6th row) are based on the professionalism of video shooting or the ratings of video website users as the evaluation criteria for the evaluation of video aesthetic quality.

To solve these problems, we build a large-scale short video aesthetics (SVA) dataset with scientific annotation methods. SVA includes 6900 edited videos from YouTube and AVAQ6000, each lasting 10 to 30 seconds. Some examples of datasets are shown in Figure 1.

The labeling process of SVA is detailed in Algorithm 1. The labeling process involves 15 viewers of different genders and different ages. Before labeling, each viewer will watch some indicative videos with high and low aesthetic quality in advance. When labeling, the viewer watches the video and assigns an aesthetic quality score of 1 to 10 points to the watched video, of which 1 to 5 points are assigned to videos with a low aesthetic quality, and 6 to 10 points are assigned to high aesthetics quality. After labeling, the final decimal aesthetic score of each video is the average score after removing the highest and lowest scores. If the decimal aesthetic score of a video is greater than σ, the video is considered to be of high aesthetic quality, otherwise, the video is considered to be of low aesthetic quality. In this paper, we set σ to 5.

Algorithm 1 Labeling process

Input：Video

Output：Binary aesthetic score and decimal aesthetic score

1：Viewers watch the video and get an aesthetic score of 1-10 points $S_1, S_2, S_3, \ldots S_{15}$

2: $S = 0$

3: for $i = 1{:}n$

$S \mathrel{+}= S_i$

4: $S = S - S_{max} - S_{min}$

5: $S_D = S//(n-2)$

6：

$$S_B = \begin{cases} 1 & S_D > \sigma \\ 0 & S_D \le \sigma \end{cases}$$

(a)

(b)

Figure 1. The first frame of partial video in SVA. (a) It is of low aesthetic quality. (b) It is of high aesthetic quality.

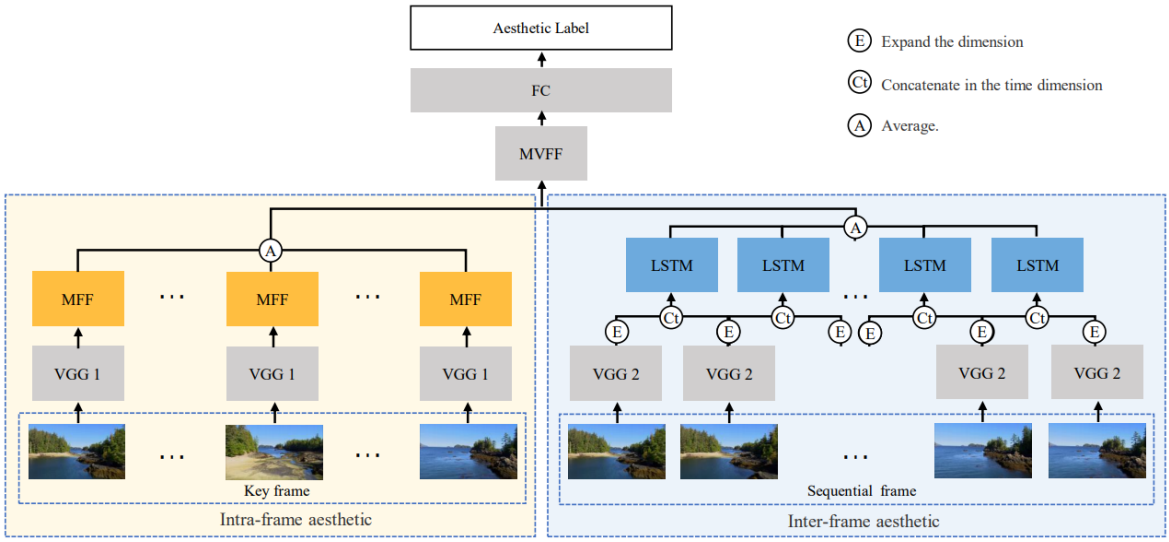


Figure 3. Illustration of the proposed framework. MVVA-Net contains two branches, the intra-frame aesthetic branch and the inter-frame aesthetic branch. The intra-frame aesthetics branch extracts the intra-frame aesthetic features of a single frame through the VGG-16 [18] convolution structure and MFF and merges the intra-frame aesthetic features extracted from all frames; the inter-frame aesthetics branch uses the VGG-16 convolution structure and LSTM extracts the inter-frame aesthetic features between every two frames and merges the inter-frame aesthetic features extracted from all frames. MVFF adaptively fuses the intra-frame aesthetic features and inter-frame aesthetic features, and the fused features are mapped to one dimension through the full connection layer to represent the aesthetic quality of the video.

In SVA, 3735 videos are labeled as high aesthetic quality and 3165 videos are labeled as low aesthetic quality. The labeling basis and scale of the SVA and the existing video aesthetic quality assessment dataset are listed in Figure 2.

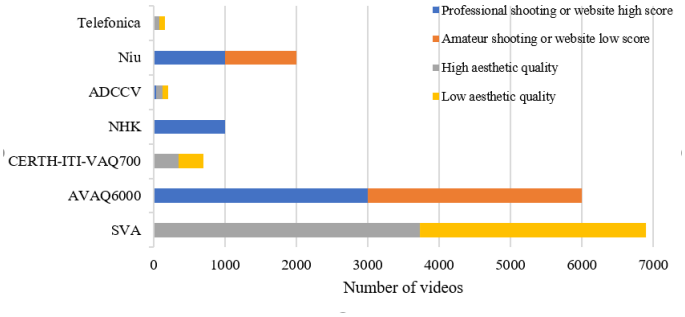


Figure 2. The labeling basis and scale of the SVA and the existing video aesthetic quality assessment dataset.

As can be seen from Figure 2, compared with other existing videos aesthetic quality assessment datasets, SVA data volume is larger and the labeling basis is more reasonable.

4 Methods

MVVA-Net is shown in Figure 3. The intra-frame aesthetic branch takes key frames as input to extract intra-frame aesthetic features; the inter-frame aesthetic branch takes sequential frames as input to extract inter-frame aesthetic features. We adaptively fuse the multi-type features extracted from the two branches to evaluate the aesthetic quality of the video. At the same time, both the intra-frame aesthetic branch and the inter-frame aesthetic branch support videos of different durations with different frame numbers as input.

In this chapter, we will introduce the acquisition of multiple types of frames, the intra-frame aesthetic branch, the inter-frame aesthetic branch, and the adaptive fusion of the two branches.

4.1 Multi-type frame

Our method designs two branches to extract intra-frame aesthetic features and inter-frame aesthetic features of the video. These two branches take different types of video frames, namely key frames and sequence frames, as input.The sequential frame is a frame extracted from video based on a fixed interval, which contains the changing relationship between frames in a video, so it is used as the input of inter-frame aesthetic branch to extract inter-frame aesthetic features; the key frame is obtained by frame difference method, which can represent different pictures in the video, so it is used as the input of intra-frame aesthetic branch to extract intra-frame aesthetic features.

For the extraction of sequential frames, in all the video frames $V = \{I_1, I_2, \dots, I_n\}$, the frames $I_{ms+1}(m \in (0,1,2,\dots))$ are extracted every fixed interval S to get the

(a)

(b)

Figure 4. Sequential frame example. (a) and (b) are examples of sequential frames of two different videos.

video sequence frames. Figure 4 (a) and (b) are examples of sequential frames of two videos in SVA.

For the extraction of key frames, we propose a frame difference method. The key frames extracted based on this method represent different pictures in the video. (a) and (b) in Figure 5 are examples of key frames of two videos in SVA.

We add the first frame I_1 in the video $V = \{I_1, I_2, \dots, I_{\mathrm{n}}\}$ to the key frame set $K = \{k_1, k_2, \dots, k_{\mathrm{m}}\}$. The sum of the difference graph d_i of each frame o_i in the remaining frame set $O = \{o_1, o_2, \dots, o_{\mathrm{n-m}}\}$ and all frames in the key frame set is calculated. The expression is as follows:

$$d_i = \sum_{j=1}^{m}(o_i - k_j) \quad (1)$$

Then, we calculate the average pixel value a_i of the sum of the difference graph, that is, the difference between the remaining frames and all the current key frames.

$$a_i = AVG(d_i) \quad (2)$$

where AVG represents the sum of pixels and then divide the number of pixels. We select the frame o with the largest difference value a among all the current remaining frames.

$$a = \max(a_i) \rightarrow o \quad (3)$$

where max represents the maximum value. If a is greater than the preset threshold T, o is added to the key frame set K. Cycle the above steps until all key frames are selected.

4.2 Intra-frame aesthetic branch

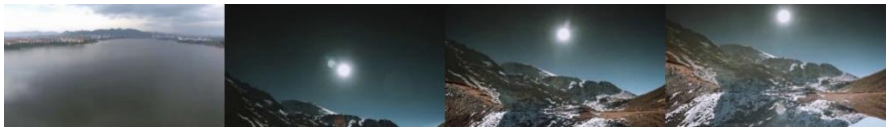
The intra-frame aesthetics branch takes the key frames extracted based on the frame difference method as input, extracts the intra-frame aesthetic features of each frame of the video, and merges the intra-frame aesthetic features of each frame to obtain the intra-frame aesthetic features of the video.

Suppose input $I = \{k_1, k_2, \dots, k_{\mathrm{n}}, \dots k_{\mathrm{N}}\} (k \in R^{C\times H\times W})$, N is the number of key frames extracted from a video. The number of key frames extracted from different videos is different. The intra-frame aesthetics branch uses the VGG-16 convolution structure of N shared parameters and the multi-receptive field fusion module(MFF) of N shared parameters to extract the intra-frame aesthetic features $f_k^n \in R^{C\times H\times W} (n \in N)$ of N frames. The final intra-frame aesthetic features $f_{\mathrm{K}} \in R^{C\times H\times W}$ of the video are obtained by fusing the intra-frame aesthetic features of the extracted N frames.

Evaluating the aesthetic quality of video frames requires comprehensive consideration of the details and the whole of the frame. In this regard, we designed MFF to extract the aesthetic features of the frame. MFF structure is shown in Figure 6.

MFF contains four branches, which can extract different scale aesthetic features of the frame respectively, and fuse the four features to obtain multi-scale intra-frame aesthetic features f_k. Specifically, for the input $f \in R^{C\times H\times W}$, we respectively extract intra-frame aesthetic features of different scales $f_{MFF}^{\mathrm{i}} (i \in (1,2,3,4))$ expression is as follows:

$$f_{MFF}^{i} = conv^{i}(f) \quad (4)$$



(a) (b)

Figure 5. Key frame example. (a) (b) correspond to the key frames of (a) and (b) in Figure 4.

where $conv^i(i \in (1,2,3,4))$ represent four convolution operations with convolution kernel sizes of 1 × 1, 3 × 3, 5 × 5 and 7 × 7, and groups of 1, 4, 8 and 16 respectively. The intra-frame aesthetic features of different scales f_{MFF}^i are concatenated to obtain the intra-frame aesthetic features of a single frame f_k.

To make our intra-frame aesthetics branch effectively extract the intra-frame aesthetics features of video without limiting the number of input frames, we average the N intra-frame aesthetics features extracted from N key frames to get the intra-frame aesthetics features of the video. The expression is as follows.

$$f_{\mathrm{K}} = \frac{\sum_{n=1}^{N} f_k^n}{N} \tag{5}$$

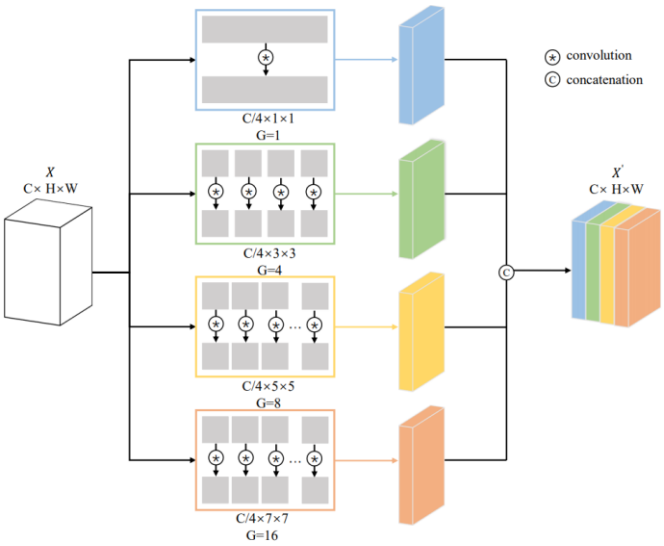


Figure 6.MFF uses grouped convolution of multiple different receptive fields to extract multi-scale features.

4.3 Inter-frame aesthetic branch

The inter-frame aesthetics branch uses the sequential frame mentioned above as input to extract the inter-frame aesthetic features of two consecutive frames of the video, and merge the extracted multiple inter-frame aesthetic features to obtain the inter-frame aesthetic features of the video.

Suppose input $I = \{t_1, t_2, \dots, t_{\mathrm{n}}, \dots t_{\mathrm{N}}\}(t \in R^{C\times H\times W})$, M is the number of sequential frames extracted from a video. The number of sequential frames extracted from different videos is different. The inter-frame aesthetics branch uses the VGG-16 convolution structure of M shared parameters and the LSTM of $M-1$ shared parameters to extract the inter-frame aesthetic features $f_t^m \in R^{C\times H\times W}(m \in M)$ of M frames. We average the $M-1$ inter-frame aesthetic features extracted from M frames to get the inter-frame aesthetic features of the video $f_{\mathrm{T}} \in R^{C\times H\times W}$.

$$f_{\mathrm{T}} = \frac{\sum_{m=1}^{M-1} f_t^m}{M-1} \tag{6}$$

This method enables the inter-frame aesthetic branch to effectively extract the inter-frame aesthetic features of the video without limiting the number of input frames.

4.4 Multi-type feature fusion

Considering that the fusion of intra-frame aesthetic features and inter-frame aesthetic features will inevitably introduce irrelevant noise, we design a multi-tupe feature fusion module (MVFF) to adaptively fuse the intra-frame aesthetic features and inter-frame aesthetic features of video for video aesthetic quality assessment. MVFF structure is shown in Figure 7.

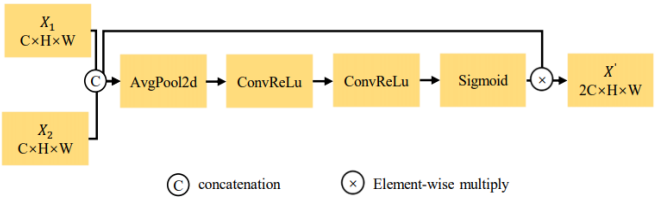


Figure 7. Detailed structure of MVFF. X_1 and X_2 are the intra-frame aesthetic features and inter-frame aesthetic features extracted from the intra-frame aesthetic and inter-frame aesthetic branches respectively.

Specifically, we splice the intra-frame aesthetic features f_{K} and the inter-frame aesthetic features f_{T} on the channel to obtain the features $f_x \in R^{2C\times H\times W}$ and apply average pooling to them to obtain a feature vector in the channel direction $f_{avgpool} \in R^{2C\times 1\times 1}$. After that, two successive convolutions and ReLu are used to capture the channel dependence and normalize it to get the feature map $f_s \in R^{2C\times 1\times 1}$. Multiply f_s and f_x to get a new feature map $f_x^{'}$.

$$f_x^{'} = f_x \cdot f_s \tag{7}$$

5 Experiments

5.1 Implementation Details

In our experiments, the dataset of Section 3 is randomly split into a training subset (70%) and a test subset (30%). We first train the intra-frame aesthetic branch and the inter-frame aesthetic branch on the training subset of SVA. This training process includes two

TABLE 2. Comparison with other methods

Method	Dataset	Accuracy	F-score	AUC
InceptionV3 + LSTM	SVA	68.4083	70.4659	75.3987
ResNet50 + LSTM	SVA	70.8757	75.3884	77.3917
C3D	SVA	71.5530	74.4348	77.8733
Ours	**SVA**	**75.8104**	**79.0092**	**81.0783**
SVM-based [10]	CERTH-ITI-VAQ700	64.00	-	-
Ours	**CERTH-ITI-VAQ700**	**65.1429**	**64.7399**	**70.1090**

stages: freezing all parameters inherited from the pre-trained network VGG-16 to learn new initialization parameters; fine-tuning the entire network. The SGD optimizer is used to train the network in two learning stages with learning rates of 0.01 and 0.0001, respectively.

After that, we use the above-trained intra-frame aesthetic branch weights and inter-frame aesthetic branch weights to initialize the two branches of MVVA-Net and train our method. The training process of MVVA-Net is similar to the training process for the above two branches. First, we freeze other layers except for MVFF and the full connection layer for training; then, we fine-tune the whole network. Throughout the training phase, we used the same optimizer and learning rate as above. In the whole training process, the minimum batch is set to 1, and all super parameters except the learning rate are set to the default value. The loss of our network is the cross-entropy between the predicted value and the ground truth of each video.

To enhance the robustness of the model, during the training process, we use random rotation and random flip for each key frame, and uniform rotation and flip for all sequential frames of a video. We scale the image size to 224 × 224, set the sequential frame fixed extraction interval S to 15, and set the key frame extraction threshold T to 50. The experiment was performed on a single GTX 2080Ti GPU, and we used PyTorch to implement our model.

5.2 Quantitative Comparison

We compare the MVVA-Net with C3D [19] on SVA. Additionally, we also compare other networks, namely Inception V3 [20] and ResNet50 [21], which perform significantly well on ImageNet [22] for local feature extraction. The top layers of both networks are replaced by the same modified fully connected layers or LSTM layers based on our task. The results are shown in Table 2.

Furthermore, we provide the comparison results between our method and the traditional SVM-based video quality assessment method on CERTH-ITI-VAQ700.

The results show that our method is better than other methods, which also proves the effectiveness of our model.

5.3 Ablation Study

In order to verify the effectiveness of fusing intra-frame aesthetic features and inter-frame aesthetic features for video aesthetic quality assessment and the importance of different modules in our method, we conducted an ablation study.

TABLE 3. Ablation study using different component combinations. In experiments 4 and 5, the intra-frame aesthetic features and the inter-frame aesthetic features are concatenated.

Number	Intraframe aesthetics	Interframe aesthetics	MFF	MVFF	Accuracy
1	√	×	×	×	71.5046
2	√	×	√	×	73.0527
3	×	√	×	×	72.9560
4	√	√	×	×	73.9719
5	√	√	√	×	75.3266
6	√	√	×	√	74.2138
7	√	√	√	√	**75.8104**

As shown in Table 3, the method with all modules achieves the best performance, which shows the effectiveness of each module in our method.

TABLE 4. The MVVA-Net trained by fixed input and unfixed input is compared.

Method	Dataset	Fixed input	Accuracy	F-score	AUC
MVVA-Net	SVA	YES	73.5849	73.5977	81.4194
MVVA-Net	**SVA**	**NO**	**75.8104**	**79.0092**	**81.0783**
MVVA-Net	CERTH-ITI-VAQ700	YES	51.1429	25.9740	47.9637
MVVA-Net	**CERTH-ITI-VAQ700**	**NO**	**65.1429**	**64.7399**	**70.1090**

Comparing experiments 2, 3, and 5, the accuracy of merging the intra-frame aesthetic branch and the inter-frame aesthetic branch to evaluate the video aesthetic quality is 3.1127% and 3.2494% higher than the accuracy of the single branch respectively. Comparing aesthetic quality is 3.1127% and 3.2494% higher than the accuracy of the single branch respectively. Comparing experiments 6 and 7, using MFF increased the accuracy of our method by 2.1562%. By comparing experiments 5 and 7, it is found that using MVFF increases the accuracy of our method from 75.3266 to 75.8104.

We trained MVVA-Net on training subset of SVA with fixed input and non-fixed input and tested the trained model on test subset of SVA and CERTH-ITI-VAQ700. The results are shown in Table 4.

The results in the table show that when we fix the model input size, our model can well fit the data in the current time interval. However, due to the limitations of extracting video features with a fixed number of frames, our model cannot be generalized to datasets of other duration intervals. When we do not fix the input of the model, our model can well extract the features of videos with different durations, so that our model can be well generalized to other datasets.

6 Conclusion

In this study, we constructed a dataset containing 6900 video shots. In order to comprehensively consider intra-frame aesthetics and inter-frame aesthetics, and improve the generalization ability of the model, we propose a method of fusing multi-type features for video aesthetic quality assessment based on the strategy of not fixed model input. The experimental results show that our model has shown good performance on different datasets and demonstrated strong generalization ability.

References

[1] Choi JH, Lee JS. Automated video editing for aesthetic quality improvement. Proceedings of the 23rd ACM international conference on Multimedia; 2015.

[2] Luo Y, Tang X. Photo and video quality evaluation: Focusing on the subject. European Conference on Computer Vision; 2008: Springer, Berlin, Heidelberg.

[3] Moorthy AK, Obrador P, Oliver N. Towards computational models of the visual aesthetic appeal of consumer videos. European conference on computer vision; 2010: Springer, Berlin, Heidelberg.

[4] Yang CY, Yeh HH, Chen CS. Video aesthetic quality assessment by combining semantically independent and dependent features. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2011: IEEE.

[5] Chung S, Sammartino J, Bai J, Barsky BA. Can motion features inform video aesthetic preferences. University of California at Berkeley Technical Report No UCB/EECS-2012-172June. 2012;29.

[6] Niu Y, Liu F. What makes a professional video? A computational aesthetics approach. IEEE Transactions on Circuits and Systems for Video Technology. 2012;22(7):1037-49.

[7] Bhattacharya S, Nojavanasghari B, Chen T, Liu D, Chang SF, Shah M. Towards a comprehensive computational model for aesthetic assessment of videos. Proceedings of the 21st ACM international conference on Multimedia; 2013.

[8] Yeh HH, Yang CY, Lee MS, Chen CS. Video aesthetic quality assessment by temporal integration of photo-and motion-based features. IEEE transactions on multimedia. 2013;15(8):1944-57.

[9] NHK:Where is beauty? Grand Challenge at ACM Multimedia Conf (MM'13). 2013.

[10] Tzelepis C, Mavridaki E, Mezaris V, Patras I. Video aesthetic quality assessment using kernel Support Vector Machine with isotropic Gaussian sample uncertainty (KSVM-IGSU). 2016 IEEE International Conference on Image Processing (ICIP); 2016: IEEE.

[11] Kuang Q, Jin X, Zhao Q, Zhou B. Deep multimodality learning for UAV video aesthetic quality assessment. IEEE Transactions on Multimedia. 2019;22(10):2623-34.

[12] Phatak MV, Patwardhan MS, Arya MS. Deep Learning for motion based video aesthetics. 2019 IEEE Bombay Section Signature Conference (IBSSC); 2019: IEEE.

[13] Duta IC, Liu L, Zhu F, Shao L. Pyramidal convolution: rethinking convolutional neural networks for visual recognition. arXiv preprint arXiv:200611538. 2020.

[14] Wang L, Li W, Li W, Van Gool L. Appearance-and-relation networks for video classification. Proceedings of the IEEE conference on computer vision and pattern recognition; 2018.

[15] Feichtenhofer C, Fan H, Malik J, He K. Slowfast networks for video recognition. Proceedings of the

IEEE/CVF International Conference on Computer Vision; 2019.

[16] Zhao T, Wu X. Pyramid feature attention network for saliency detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019.

[17] Li H, Chen G, Li G, Yu Y. Motion guided attention for video salient object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019.

[18] Simonyan K, Zisserman AJapa. Very deep convolutional networks for large-scale image recognition. 2014.

[19] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. Proceedings of the IEEE international conference on computer vision; 2015.

[20] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

[21] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

[22] Krizhevsky A, Sutskever I, Hinton GEJAinips. Imagenet classification with deep convolutional neural networks. 2012;25:1097-105.