Emotion-Aware Music Driven Movie Montage

Wuqin Liu
School of AI, UCAS & NLPR, CASIA
liuwuqin21@mails.ucas.ac.cn

Minxuan Lin
Kuaishou Technology
linminxuan@kuaishou.com

Haibin Huang
Kuaishou Technology
jackiehuanghaibin@gmail.com

Chongyang Ma
Kuaishou Technology
chongyangm@gmail.com

Yu Song
School of ME, USTB
yusong@ustb.edu.cn

Weiming Dong
NLPR, CASIA
weiming.dong@ia.ac.cn

Changsheng Xu
NLPR, CASIA
csxu@nlpr.ia.ac.cn

Abstract

In this paper, we present emotion-aware music driven movie montage, a novel paradigm for the challenging task of generating movie montages. Specifically, given a movie and a piece of music as the guidance, our method aims to generate a montage out of the movie that is emotionally consistent with the music. Unlike previous work such as video summarization, this task requires not only video content understanding, but also emotion analysis of both the input movie and music. To this end, we propose a two-stage framework, including a learning based module for prediction of emotion similarity and an optimization based module for selection and composition of candidate movie shots. The core of our method is to align and estimate emotional similarity between music clips and movie shots in a multi-modal latent space via contrastive learning. Subsequently, the montage generation is modeled as a joint optimization of emotion similarity and additional constraints such as scene-level story completeness and shot-level rhythm synchronization. We conduct both qualitative and quantitative evaluations to demonstrate that our method can generate emotionally consistent montages and outperforms alternative baselines.

***Keywords:** Movie montage, emotion analysis, audio-visual modalities, contrastive learning.*

1. Introduction

In recent years, with the rapid growth of social network and mobile applications, it has become increasingly popular and important to create high-quality short videos and montages. As one of the best resources for montages, movies are often cut and composed into shorter versions accompanied by a piece of background music, to obtain the trailers, previews and/or highlights of the original ones. However, existing montage editing tools typically rely on the users to manually pick shots from the movie and align with the music, which is tedious and time consuming. It remains difficult for non-professional users to generate a movie montage of satisfactory quality to match the rhythm and emotion of the music, with the additional constraint that the selected shots provide a reasonable and comprehensible summary of the original content or story.

As machine learning technologies emerge and advance, several methods have been proposed in the past few years for automatic generation of montages, ranging from video summarization [22] to emotion-oriented music video generation [14, 16]. However, the former mainly focuses on the content of the video itself, ignoring the correlation with any input music, while the latter has difficulty in understanding and handling long videos.

Walter Scott Murch, one of the most famous movie editors, has summarized *the Rule of Six* for film editing, including emotion, story, rhythm, eye trace, 2D plane of screen, and 3D space of action [21], which have different values in terms of importance for the final cut. Among these six elements, emotion is the most important one and has an importance factor of 51%, while story and rhythm correspond to a factor of 23% and 10%, respectively. Inspired by Murch's Rule of Six, we propose emotion-aware music driven movie montage, a method to automatically generate a montage from an input movie with a piece of user-specified music as the guidance. Specifically, we compose the output montage by taking the most important three elements for film editing into account to meet the following requirements:

Emotional consistency. The shots that are used to com-

Figure 1. Emotion score based music-driven movie montage. When editing the same input movie with different background music, the corresponding emotion scores are completely different, and thus the final movie montages guided by different music are also distinct.

pose the output montage are emotionally consistent with the input music.

Story completeness. The montage needs to present a story that is relatively complete and comprehensible.

Rhythm synchronization. The visual and audio content of the montage should have synchronized rhythms.

To achieve the above goals, we adopt a two-stage framework. In the first stage, we build a network to align multimodal signals of music, text, and image in the emotion space based on CLIP and contrastive learning [26]. In the second stage, we formulate the task of composing montages as an optimization problem and generate the output using a knapsack based solver. Specifically, we divide the input movie at both the scene level and the shot level. The output montage is generated by maximizing the emotional similarity between scenes/shots and the input music. We ensure that the story in the montage is comprehensible by adding constraints on the number of selected scenes. Furthermore, we align selected shots with bars of the input music using quantified duration so that the rhythm of both the visual and audio signals is synchronized.

As illustrated in Figure 1, given a movie as a candidate, we can choose different shot combinations to form a montage result according to the user-supplied emotional music. The changing emotion score in the movie will be used as a significant indicator to select the target shots.

In summary, our main contributions are as follows:

- We present a novel method for montage generation from an input movie and a user-specified music clip based on well-established rules for film editing.
- We propose a two-stage framework to generate output movie montages, by formulating the generation task as a constrained optimization problem.
- We conduct qualitative and quantitative evaluations to demonstrate that our method leads to high-quality emotionally consistent montages and outperforms alternative baselines.

2. Related Work

Music-driven video generation. The purpose of music video generation is to combine music and video to enhance entertainment quality and emotional resonance. Most previous methods [36, 12] only considered the relationship between low-level acoustic features and visual features while ignoring semantic constraints. Liao et al. [13] cut the input video to synchronize the music rhythm and generated audio-visually consistent results. To narrow the semantic gap between low-level acoustic features and human perception, some methods [33, 14, 16] tried to map the two into the emotion space and made the audience have a good match in their emotional perception when watching the generated music video. Lin et al. [15] proposed an emotion-based pseudo-song prediction and matching framework. And Lin et al. [17] considered the continuity of video content while matching music and video. Gross et al. [4] generated music videos by using features of video color histogram, key changes in music and genre. However, these methods are rarely concerned with long sequence videos, so when feeding videos that are much longer than the audio time, those methods ignore the relevance of video content while ensuring emotional consistency, so the generated results often do not have any storyline. To address this issue, we propose an algorithm to select shots that enhance the storytelling of videos while maintaining emotional consistency.

Video summarization. Video summarization refers to the task of generating summaries by stitching together important contents of a video. Early approaches mainly used

unsupervised methods [18, 20] to generate video summaries due to lack of useful datasets. After the creation of some manually collected datasets [31, 6], several supervised methods have emerged [39]. However, when users browse videos, they always try to find something specific. Therefore, Sharghi et al. [29] proposed the Query-Focused Video Summarization (QFVS) dataset, allowing video summaries to find specific shots through a query to generate results, making the results more user-friendly. After the introduction of CLIP [26], Narasimhan et al. [22] proposed a single framework for solving general and query-focused video summarization in both unsupervised and supervised methods by combining CLIP and video summarization. Movie trailer generation is one of the main applications of video summarization work, which attracted the attention of many researchers. Existing methods usually exploit shallow audio-visual features [8, 35, 25, 30], but these methods usually only focus on information about the movie itself. However, music is an integral part of video editing, that can improve the viewing experience of the final result. Thus, we use music as guidance to generate an emotion-aware movie montage.

Emotion analysis of music and videos. The emotions associated with music and video have been well-studied. It has been suggested that emotions are one of the main reasons why people engage in music [9], and psychological research has shown that people also have emotional responses to visual stimuli [3]. Therefore, it is a very natural way to connect video and music through emotion. Categorical and dimensional representations have been used to represent emotion in music [10]. Discrete categorical labels include terms like excited, relaxation, sad, etc. One study found that the number of emotion categories did not reflect the richness of emotion that humans perceive, or that the taxonomy is inherently ambiguous [9]. Therefore, some other works used dimensional labels in the two-dimensional plane of valence and arousal to represent music [28]. This continuous representation has no classification problems, but it is difficult to distinguish some mental and emotional concepts. Similar to music, emotions associated with images and videos are also represented by categories [38] and dimensions [19]. Baveye et al. [1] expressed the features of movie scenes in the valence-arousal space. Hanjalic et al. [7] introduced dominance as an additional dimension to characterize the emotion of videos.

3. Method

In this section, we formally introduce our pipeline for emotion-aware movie montage generation. We first revisit the general setting of montage generation and then extend it into an emotion-aware constrained optimization problem. As demonstrated in Figure 2, there are two key components in our framework, including (1) multi-modal emotion latent space alignment, and (2) emotion score based shots selection.

3.1. Problem Statement

Our goal is to generate a montage given the user-specified music $\mathbf{x^m}$ and a long movie $\mathbf{x^v}$. Following the common practice for montage generation, we divide the movie $\mathbf{x^v}$ into a set of scenes $\mathcal{E} = \{e_1, e_2, ..., e_m\}$ and each scene can be split into multiple shots. We denote all the shots as a shot set $\mathcal{S} = \{s_1, s_2, ..., s_n\}$ and use a mapping function $\tau(s_i) = e_j$ to record the relationship between scenes and shots. Similarly, the input music $\mathbf{x^m}$ is split into a series of bars $\mathcal{B} = \{b_1, b_2, ...b_l\}$. Then the goal of montage generation is to select a subset of shots $\mathcal{R}$ from $\mathcal{S}$ and associate each bar to a movie shot. In other words, montage generation requires (1) a shot indicator function $\mathbb{1}_s(s_i)$, determining which shots are selected and (2) a mapping function $\phi(b_k) = s_i$ to present the relationship between shots and bars. This task is in general an under-constrained problem, and hence additional constrains $\mathbb{C} = \{c_1, c_2, ..., c_\alpha\}$ are required to limit the feasible solutions. Valid constraints include the total number of selected scenes and rhythm synchronization between shots and bars.

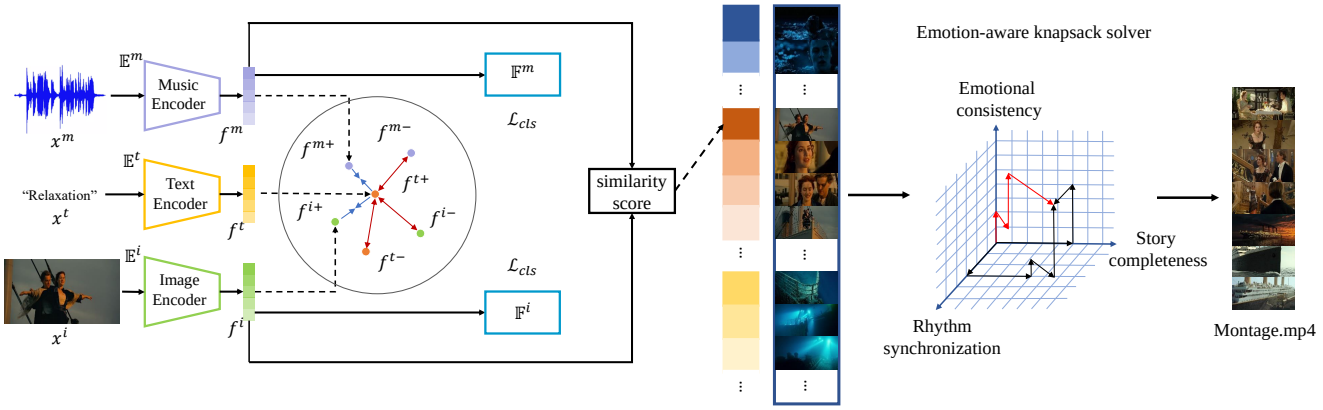
In this work, we add emotion-aware constraints for the shot selection task. Our key insight is to introduce an emotion measurement function $\mathbb{M}(s_i, \mathbf{x^m})$, which can be used to evaluate the consistency between each shot and the whole music. With $\mathbb{M}$, we can formulate the optimization target such that the selected subset of shots $\mathcal{R}$ can construct a montage by maximizing the emotional consistency between the audio and visual signals, as shown below:

$$\mathcal{R} = \underset{i}{\operatorname{argmax}} \sum_{i=1}^{n} \mathbb{M}(s_i, \mathbf{x^m}) \mathbb{1}_s(s_i), \text{ s.t. } \mathbb{C}. \quad (1)$$

To solve the proposed optimization problem, we further develop a two-stage paradigm to learn the required functions. Specifically, we adopt a CLIP-based multi-modal alignment approach for emotion latent representation learning and use it as $\mathbb{M}(s_i, \mathbf{x^m})$. The optimization of scenes and shots selection can be modeled as a knapsack problem given the constrains $\mathbb{C}$. The shot indicator function $\mathbb{1}_s(s_i)$ and shot-bar mapping function $\phi(b_k) = s_i$ can be obtained via a deterministic knapsack solver. We will provide details in following sections.

3.2. Multi-Modal Emotion Latent Space Alignment

The first stage of our pipeline is to learn an emotion measurement function $\mathbb{M}$ between movie shots and music. It requires embedding and alignment of signs from different modalities in the emotion space. Inspired by AudioCLIP [5], we train three encoders ($\mathbb{E^m}$, $\mathbb{E^t}$, $\mathbb{E^i}$) of different



(a) Multi-modal emotion latent space alignment

(b) Emotion score based shots selection

Figure 2. Illustration of our framework. (a) In the first stage, we construct an emotion space by aligning the latent representation of multiple modalities. (b) In the second stage, we select and compose several emotion-related shots from the candidates using our emotion-aware knapsack based optimization solver.

modalities (music, text and image) to produce matched representations, as shown in Figure 2(a). Specifically, given a tuple of music, text, and image ($\mathbf{x^m}$, $\mathbf{x^t}$, $\mathbf{x^i}$) as the input, we use the three encoders to obtain a set of latent representations ($\mathbf{f^m}$, $\mathbf{f^t}$, $\mathbf{f^i}$). The purpose of introducing text modality is to use it as an anchor to improve classification accuracy. We initialize the encoders with the pretrained AudioCLIP model and further optimize joint audio-text-visual representations via contrastive learning procedure [26].

Contrastive constraints for multi-modal emotions. The pretrained AudioCLIP model gives a good embedding space for features from different modalities, we further align the feature space and make it emotion-aware. Specifically, for arbitrary feature pair ($\mathbf{f}^a$, $\mathbf{f}^b$), where a, b are from different modalities, we aim to align the distribution of $\mathbf{f}^a$, $\mathbf{f}^b$ if they correspond to the same emotion, and push away otherwise.

Towards this end, we first define an emotion indicator for different modalities. As demonstrated in Pandeya et al. [24], emotions of audio and visual signals can be measured together in the 2D valence-arousal space, which provides a reasonable indicator to compare the differences of both modalities. Thus, we follow the settings in [24] and divide emotion space into six categories to cover the emotion space of daily communications, i.e., *excited*, *fear*, *neutral*, *relaxation*, *sad*, and *tension*.

For each iteration of the network training stage, we build three 6×6 modality constraint matrices. Taking image-music modalities as an example, we denote a music feature and an image feature either as a positive pair ($\mathbf{f^{m+}}$, $\mathbf{f^{i+}}$) which will be placed on the diagonal of the matrix if they have same emotion indicator, or as negative examples $\mathbf{f^{i-}}$ and $\mathbf{f^{m-}}$ which will be placed on the off-diagonal of that.

We use the cross entropy loss $\mathcal{L}_{CE}$ to push the convergence of emotion space between two specific modalities in the matrix diagonal. The detailed image-music contrastive loss is as follows:

$$\begin{aligned}\mathcal{L}_{image\_music} = &\mathcal{L}_{CE}(S^{(\mathbf{i+},\mathbf{m+})}, 1) \\ &+ \mathcal{L}_{CE}(S^{(\mathbf{i+},\mathbf{m-})}, 0) \\ &+ \mathcal{L}_{CE}(S^{(\mathbf{i-},\mathbf{m+})}, 0).\end{aligned} \tag{2}$$

where 1 represents a full one vector, and 0 denotes a zero vector. S is emotional consistency score we use to evaluate the distance between different modalities, defined as follows:

$$S^{(a,b)} = \frac{< \mathbf{f}^a \cdot \mathbf{f}^b >}{\|\mathbf{f}^a\| \cdot \|\mathbf{f}^b\|}. \tag{3}$$

We compute the constraints $\mathcal{L}_{text\_image}$ and $\mathcal{L}_{text\_music}$ for text-image modalities and text-music modalities in the same way.

Emotion classification constraints. In order to further improve discriminability of emotion features, we add a fully connected layer after the image and music encoder to classify the emotion categories, which enhances the linearity of latent emotion space. The text information is used as a tag to influence feature space construction. More concretely, a text prompt feature $\mathbf{f^t}$ will be reshaped to a one-hot vector C_{text} as the target. We denote the image linear classification layer as $\mathbb{F}^{\mathbf{i}}$ and the music linear classification layer as $\mathbb{F}^{\mathbf{m}}$. These classification layers learn to discriminate the emotion categories of image and music by the cross-entropy loss. Therefore, the final classification constraint is formu-

lated as:

$$\mathcal{L}_{cls} = \mathcal{L}_{CE}(\mathbb{F}^{\mathbf{i}}(\mathbf{f}^{\mathbf{i}}), C_{text}) + \mathcal{L}_{CE}(\mathbb{F}^{\mathbf{m}}(\mathbf{f}^{\mathbf{m}}), C_{text}). \tag{4}$$

Total loss. Our full objective loss function can be written as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{image\_music} + \mathcal{L}_{text\_music} + \mathcal{L}_{text\_image} + \alpha\mathcal{L}_{cls}. \tag{5}$$

where α is a parameter to balance different loss terms.

After the training stage, we apply the image encoder $\mathbb{E}^{\mathbf{i}}$ on each shot to get the image feature set $F^{\mathbf{i}} = \{\mathbf{f}_1^{\mathbf{i}}, \mathbf{f}_2^{\mathbf{i}}, ..., \mathbf{f}_n^{\mathbf{i}}\}$. Since each frame in a single shot is similar, we represent the content of a single shot by picking an intermediate frame $\mathbf{x}^{\mathbf{i}}$ in the shot interval. Meanwhile, the trained music encoder $\mathbb{E}^{\mathbf{m}}$ is used to extract features of the input music $\mathbf{x}^{\mathbf{m}}$. We collect all the emotion consistency scores for each shot and the whole music to form a set of emotion scores $\Omega = \{S_1^{(\mathbf{i},\mathbf{m})}, S_2^{(\mathbf{i},\mathbf{m})}, ..., S_n^{(\mathbf{i},\mathbf{m})}\}$ as the original value of $\mathbb{M}(s_i, \mathbf{x}^{\mathbf{m}})$.

3.3. Emotion Score Based Shot Selection

Given the learned emotion score function $\mathbb{M}$, our next step is to select candidate shots which yield maximum emotion score w.r.t the optimization target in Eq 1. Following Walter Murch's montage criterion [37], we use two constraints as $\mathbb{C}$ to limit the solution space: (1) *scene-based story completeness constraint* to improve the causality of montage; (2) *shot-based audio-visual rhythm synchronization constraint* to guarantee audio-visual harmonious degree of montage. The shot indicator function $\mathbb{1}_s(s_i)$ will be obtained during optimization with these constrains.

Scene-level constraint for story completeness. Our key observation is that the less changes in character and environment, the easier it would be for audiences to understand the storyline. Therefore, a high aggregation degree of scene can provide a better story completeness. Intuitively, we can improve the story completeness by limiting the number of scenes to be involved.

Hence, we define a function $\mathbb{1}_e(e_j)$ to indicate whether a scene is selected. A scene is denoted as selected when one of the shots belonging to it is picked:

$$\mathbb{1}_e(e_j) = \begin{cases} 1, & \text{if } \sum_{i,\tau(s_i)=e_j}^{n} \mathbb{1}_s(s_i) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Furthermore, we denote $\mathbb{N}_e$ as the maximum number of selected scenes, and take it as an upper bound on the sum of $\mathbb{1}_e(e_j)$, formulated as:

$$\sum_{j=1}^{m} \mathbb{1}_e(e_j) \leq \mathbb{N}_e. \tag{7}$$

Shot-level constraint for rhythm synchronization. Empirically, the audiences feel more harmonious if the shot and music rhythm of a montage is changed synchronously. Here the music rhythm is defined as the the duration of bars. We can model such rhythm synchronization constraint by establishing a mapping relationship between shots and bars. Specifically, we require each music bar should correspond to a shot and each shot should contain at least one complete music bar.

To achieve this, we first quantify the duration of both shot and bar. Since the variation of music bar duration is small, we take the average continuous bar duration $t_k^{c,b}$ as the unit of discrete time, noted as $t_k^{d,b}$. Then for each shot of movie, we obtain the discrete shot duration $t_i^{d,s}$ by exactly dividing the continuous shot duration $t_i^{c,s}$ with $t_k^{d,b}$. We further require that the sum of discrete selected shot duration is equal to the sum of all discrete bar duration $\mathbb{N}_b$, which is formulated as follows:

$$\sum_{i=1}^{n} t_i^{d,s} \mathbb{1}_s(s_i) = \sum_{k=1}^{l} t_k^{d,b} = \mathbb{N}_b. \tag{8}$$

Final optimization formula. We define the complete optimization problem as:

$$\begin{aligned} \mathcal{R} = \underset{i}{\text{argmax}} & \sum_{i=1}^{n} \mathbb{M}(s_i, \mathbf{x}^{\mathbf{m}}) \mathbb{1}_s(s_i), \\ \text{s.t.} & \sum_{i=1}^{n} t_i^{d,s} \mathbb{1}_s(s_i) = \mathbb{N}_b, \\ & \sum_{j=1}^{m} \mathbb{1}_e(e_j) \leq \mathbb{N}_e. \end{aligned} \tag{9}$$

3.4. Emotion-Aware Knapsack Solver

To tackle the above optimization problem, we design an emotion-aware multi-dimensional knapsack solver with the proposed constraints. Specifically, we define three attributes belonging to shot s_i according to the optimization formula. The first one is a weighted emotion score p_i. To further enhance the importance of scene, for each $S_i^{(\mathbf{i},\mathbf{m})}$, we adjust the value by adding the average emotion score of the scene which the shot belongs to and form the weighted emotion score set $P = \{p_1, p_2, ...p_n\}$. Each item in P is formulated as:

$$p_i = S_i^{(\mathbf{i},\mathbf{m})} + \frac{1}{\sum_{i=1}^{n} \mathbb{1}(\tau(s_i) = e_j)} \sum_{i,\tau(s_i)=e_j}^{n} S_i^{(\mathbf{i},\mathbf{m})}, \tag{10}$$

where $\mathbb{1}(\cdot)$ is a boolean indicator function, when the condition $(\cdot)$ holds, it returns 1, and 0 otherwise. The second

attribute is the discrete shot length $t_i^{d,s}$, used to ensure the visual-audio rhythm synchronization constraint in Eq 8. After traversing each item, we obtain the discrete shot duration set $T^{d,s} = \{t_1^{d,s}, t_2^{d,s}, ..., t_n^{d,s}\}$. The third attribute is the scene number constraint score q_i corresponding to Eq 7. We define a step function relying on the subscripts of scene to which the shot belongs, used to classify different scene categories:

$$q_i = j,\ e_j = \tau(s_i). \tag{11}$$

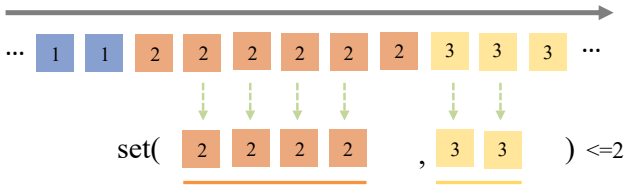
We denote $Q = \{q_1, q_2, ..., q_n\}$ as the set of scene number constraint scores. The three attribute sets will be respectively regarded as individual factors in the knapsack solver.

Algorithm 1 Hard scene constraint knapsack solver

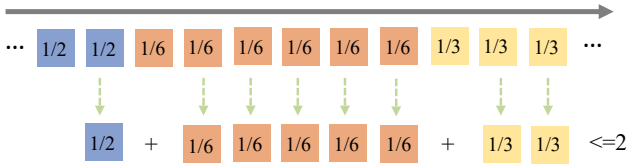
Input: $P, Q, T^{d,s}$ set, and $n, \mathbb{N}_b, \mathbb{N}_e$ as capacity.
Output: The maximum emotion score $\mathbb{P}$, the picked shot index set R.

for $i : 1 \rightarrow n$ **do**
  for $j : 1 \rightarrow \mathbb{N}_b$ **do**
    for $k : 1 \rightarrow \mathbb{N}_e$ **do**
      if $q_i \neq q_{i-1}$ **then**
        $(i, j, k, 1) \leftarrow \max((i-1, j-t_i^{d,s}, k-1, 0) + p_i, (i-1, j-t_i^{d,s}, k-1, 1) + p_i)$
        $(i, j, k, 0) \leftarrow \max((i-1, j, k, 0), (i-1, j, k, 1))$
      else
        $(i, j, k, 1) \leftarrow \max((i-1, j-t_i^{d,s}, k-1, 0) + p_i, (i-1, j-t_i^{d,s}, k, 1) + p_i, (i-1, j, k, 1))$
        $(i, j, k, 0) \leftarrow (i-1, j, k, 0)$
      end if
    end for
  end for
end for
$\mathbb{P} \leftarrow \max((n, \mathbb{N}_b, \mathbb{N}_e, 1), (n, \mathbb{N}_b, \mathbb{N}_e, 0))$
$R \leftarrow Backtrack(\mathbb{P})$
return $\mathbb{P}, R$

Hard scene constraint knapsack solver. We denote (i, j, k, z) as the basic state, which represents the maximum emotion score when exactly k scenes are selected, the sum of the discrete duration of shot is j and the first i shots are iterated over. z means whether the scene to which the i-th shot belongs is selected. We display the detailed state transition equation in Algorithm 1. Considering whether the current state is on the boundary of the scene ($q_i \neq q_{i-1}$), four possible state transition paths need to be discussed separately. When the user queries a specific upper bound on the number of scenarios $\mathbb{N}_e$, the maximum emotion score $\mathbb{P}$ can be quickly looked up. Meanwhile, $Backtrack(\cdot)$ method, as the shot indicator function $\mathbb{1}_s(s_i)$, will trace a legal path in inverse order and return a possible index set of shots R. Then, we can obtain the mapping function $\phi(b_k)$ by matching selected shots and bars of music in chronological order.



(a) The implementation of hard scene constraint.



(b) The implementation of soft scene constraint.

Figure 3. The hard and soft scene constraints. (a) The sum of weights of all picked shots is limited. (b) The number of scenes that picked shots belong to is constrained.

The hard scene constraint requires that the total number of selected scenes is less than an upper bound. As illustrated in Figure 3(a), we constrain the capacity of the set of scenes to which selected shots belong. The Algorithm 1 in the main paper displays the details of hard scene constraint knapsack. We iterate through all possible states with a triple loop which contains three core factors. In each state transition, the current state will obtain the maximum emotion score from some legal substates. Specifically, four different state transition cases need to be discussed:

1. For q_i and q_{i-1} belong to the different scenes, and choose the scene of i-th shot belong to ($z = 1$). The current state locates at a scene boundary, and the substate must reduce one scene number. Two valid substates whether to select the previous scene need to be considered.

2. For q_i and q_{i-1} belong to the different scenes, but not choose the scene of i-th shot belong to ($z = 0$). The scene number will not reduced in the substate. We also need to consider two possible sub-states, whether to choose the previous scene.

3. For q_i and q_{i-1} belong to the same scene, and choose the scene of i-th shot belong to ($z = 1$). If pick the i-th shot, there are two possible states, the first one $(i-1, j-t_i^{d,s}, k-1, 0)$ means that a new scene will be added in the i-th position, another one $(i-1, j-t_i^{d,s}, k, 1)$ means that new scenes will not be added. On

the contrary, if the i-th shot is not selected, the scene of i-th shot belongs must be chosen before this state, so the $(i-1, j, k, 1)$ state is the only choice.

4. For q_i and q_{i-1} belong to the same scene, but not choose the scene of i-th shot belong to ($z = 0$). The scene of the i-th shot belong to cannot be selected in the substate.

Soft scene constraint knapsack solver. As illustrated in Figure 3(b), soft scene constraint knapsack solver assigns corresponding scene constraint weight for each shot and limits the sum of weights for all selected shots. Before starting optimization, we multiply all scene constraint weights by a magnification constant and round them down to ensure each weight is an integer.

Fixing the number of scenes may result in the failure to obtaining the highest sum of emotion scores. Thus, we loosen the restriction in Eq 7. Instead of limiting the upper bound of the sum of selected scenes, we constrain that the sum of the inverse of the number of shots in the scene to which the selected shots belongs is less than $\mathbb{N}_e$:

$$\sum_{i=1}^{n} \frac{1}{\sum_{\alpha=1}^{n} \mathbb{1}(\tau(s_\alpha) = \tau(s_i))} \mathbb{1}_s(s_i) \leq \mathbb{N}_e. \tag{12}$$

where $\mathbb{1}(\cdot)$ is a standard indicator, when the equation is established, the function value is 1, otherwise it is 0.

Then, we reconstruct the soft scene number constraint score set as $\tilde{Q} = \{\tilde{q_1}, \tilde{q_2}, ..., \tilde{q_n}\}$, where each item of that set is formulated as:

$$q_i = \frac{1}{\sum_{\alpha=1}^{n} \mathbb{1}(\tau(s_\alpha) = e_j)}, \; e_j = \tau(s_i). \tag{13}$$

In this condition, Algorithm 1 will degenerate into a vanilla three-dimensional knapsack solver. We assume a basic state (i, j, k), which stores the maximum emotion score when traversing to i-th shot constrained by the sum of picked scene weight j and the sum of picked discrete shot duration k. During optimization, the state (i, j, k) will visit all $(i-1, j-t_i^{d,s}, k-q_i)$ states, and pick the maximum value to transfer. We get the same results as above.

At last, when the function of searching best solution $\mathbb{1}_s(s_i)$ has been obtained, we discard the part where the shot is longer than the bar to align duration, concatenate all selected shots in chronological order and append the given music according to the $\phi(b_k)$ to get the final montage. In general, we provide two knapsack-based deterministic optimization schemes to select the shot with high emotion relevance from abundant candidate shots.

4. Experiments

4.1. Dataset

The music video dataset [24] is used to train our model. This dataset focuses on multimodal emotion classification task, utilizing audio and visual information to discriminate the category of music video. In training stage, 4788 samples are used, including videos of 843 excited, 828 fear, 678 neutral, 1057 relaxation, 730 sad, and 652 tension emotions. In each batch of training, we randomly pick a frame from video as the input of image encoder, and use the full music to encode the audio feature. For text modality, we use fixed six text prompts. Finally, we test the generated results on a test set of 300 samples, where each emotion category contains 50 videos.

The original data in this dataset that we use has consistent and rich emotions. Concretely, the consistency represents the raw materials convey the same emotion signal in the visual and audio modalities. For example, the "excited" contains positive emotions with bright hued scenes and the corresponding music has light rhythm and pleasant chords. Meanwhile, the richness of emotion means that each category in the dataset covers various fine-grained emotions. For example, "excited" includes happy, joy, love, and excited, while "fear" includes scary, disgust, terror, and so on.

4.2. Experimental Setup

For visual modality, we extract the shot of video by TransNet v2 [32] and obtain the scene segmentation boundary by Rao et al. [27]. For audio modality, we split the bar of music by Madmom library [2]. We train our model for 50 epochs with the Adam optimizer [11] on a single Nvidia RTX 3090. The learning rate is 0.0001 and the batch size is 6. Meanwhile, we set the trade-off α as 1. In the optimization stage, we denote Ours(h) as the montage results generated by using hard scene constraint, and Ours(s) as the ones generated by using soft constraint. The scene number constraint for both methods is 5.

To comprehensively compare the differences between various types of movies and music in the montage task, in evaluation phrase, we choose 11 movies covering action (e.g., Léon), love (e.g., Titanic), science fiction (e.g., Inception), comedy (e.g., The Grand Budapest Hotel) and fear (e.g., Train to Busan) emotions. 16 pieces of music with distinct emotions are used as background songs.

Figure 4. Montage results generated using by our framework driven by music with different emotions.

4.3. Qualitative Evaluations

In this section, we show some qualitative results from visual-audio aspect. To directly evaluate the quality of montage, Figure 4 shows some visualized results that frames are picked from the montage generated by our method. Forrest Gum, as an example, is clipped by various music with different emotions. The representative pictures with strong emotions are shown in each row. Apparently, some optimistic scenes (excited or relaxation) with bright light are selected by delighted music. On the other side, the painful scenes (fear or sad) are often accompanied by crying and dark atmosphere. To some extent, the results demonstrate the audio-visual emotion relevance of our framework.

Ablation study. To explore the impact of each component in our solvers on audience preferences, we ablate three key factors in Eq 9, including emotional consistency, story completeness, and rhythm synchronization, to make a 30-second montage with fixed music. For the user study, we select 5 movies and generate 5 montages for each movie by our models and the baseline (ablated) models. We also invite an expert to make a montage for each movie under the same conditions. Finally, we get 30 montages and invite 36 investigators to rate them between 1 to 5, considering four aspects: 1) the degree of audio-visual emotional consistency; 2) the degree of story completeness; 3) the degree of audio-visual rhythm synchronization; 4) the overall quality of the montage.

Table 1 shows the average rating statistics. Apart from the results from expert as upper bound, Ours (h) achieves the highest rating in story completeness, rhythm synchronization and overall evaluation under full constraints. With a movie of about 2 hours as a benchmark, our method only takes 20 minutes to process a montage, but it takes an expert 2-3 days to process a 30-second video. Further, by relaxing the constraint of the scenes, Ours (s) outperforms on emotional consistency than Ours (h) but slightly decreases in other metrics due to the loss of overall coherence. When emotion factor is not considered (w/o emotion), there is a significant drop in all ratings, proving the importance of audio-visual emotional consistency for montages. Similarity, despite the selection of the largest emotion score, the lack of a story completeness constraint (w/o story) will limit the overall quality of montages. Due to people's sensitivity to audio-visual rhythm synchronization, the last factor (w/o rhythm) gets almost the worst score in most aspects.

Table 1. Results of ablation study.

Method	Emotional consistency	Story completeness	Rhythm sync.	Overall
Ours (h)	3.574	**3.624**	**3.616**	**3.783**
Ours (s)	**3.672**	3.148	3.438	3.502
w/o emotion	3.026	3.146	3.412	3.105
w/o story	3.384	3.140	3.460	3.328
w/o rhythm	3.182	3.044	3.124	3.138
Expert	3.938	3.886	3.966	4.037

Table 2. Qualitative comparisons with other methods.

Method	Emotional consistency	Story completeness	Rhythm sync.	Overall
Lin et al. [17]	3.690	3.723	3.178	3.401
Ours	**3.782**	**4.101**	**3.678**	**3.987**
Expert	3.835	3.948	4.024	4.103

Qualitative comparisons with other methods. To our knowledge, the proposed framework is the first to achieve music-driven movie montage, lacking of comparable methods and open source codes. Lin et al. [17] is the most similar work with ours, which firstly recommends a piece of matched music from a fixed music database according the user-supplied video, then obtains the final montage by selecting shots under cost-based constraints. Although the input is not exactly consistent, the output of that whole system is the same as ours, therefore, we set it as a baseline. To further prove the effectiveness of our approach, we invite the expert to clip the montages under the same conditions.

In this study, we make montages of lengths between three and five minutes with a piece of music from the user-specified video about 15 minutes. Finally, we produce 5 montages and invite 40 investigators to participate this experiment. These participants receive the same questions as the user study. They vote each montage between 1 to 5 from emotional consistency, story completeness and rhythm synchronization and overall aspect. Table 2 demonstrate our result, compared with Lin et al. [17], we achieve significant superiority on all evaluation metrics. In particular, since we explicitly consider the influencing factors of film editing, we achieve large improvements in story completeness and

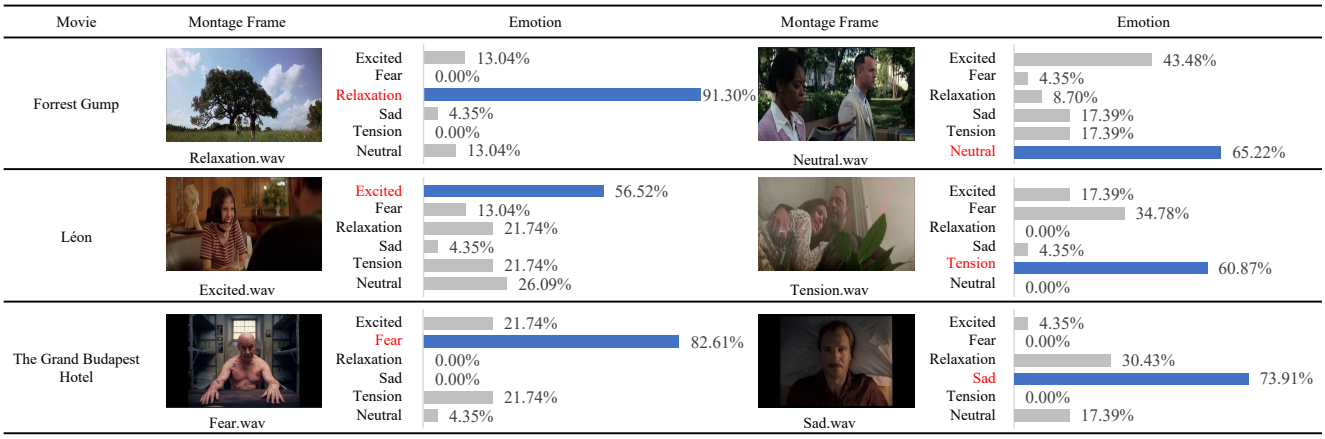


Figure 5. The percentage of most possible emotions for movie montages guided by two random emotional music. Six examples are displayed, which consist of the guided music, the montage frame and the emotion probability map voted by investigators.

rhythm synchronization, and can even achieve scores that are competitive with expert results.

Arbitrary music driven movie montage. To reflect the disparity guided by different music, we conduct a user study to explore the audio-visual emotional consistency of our approach. For comparison, we select different pieces of emotional music to drive movie clips, and each movie will be edited by two random emotions. In this experiment, investigators need to answer two questions: 1) What emotions do you feel from the movie montage? 2) How strong the emotion you feeling? Considering that the proportion of emotions of diverse people is different, we allow the participants to choose multiple emotions for each movie montage and give a score from 1 to 10 about what is the emotion degree in the montage. Finally, we receive a total of 46 valid questionnaires. As shown in Figure 5, in each row, we display the voted percentage of each emotion category for a single movie driven by two pieces of music with different emotions. For each piece of music, we draw the normalized degree of relevance that the investigators voted, the red word means the highest probability category. By observing the results, we achieve the distinct differences in all six emotions. The relaxation is the easiest category to tell due to the beautiful landscape and bright scene are often appeared. The excited and tension become the most confusion category on account of a large amount of similar facial expressions and body movements in both emotions.

4.4. Quantitative Evaluations

The confusion matrix of emotion classification. We apply the music video emotion classification accuracy to assess the performance of our model. We validate on the test set by using confusion matrix as a visual evaluation method, which counts the number of samples in classes that are confused with each other. As shown in Figure 6, our model performs well on categories of "Fear", "Relaxation" and "Excited". However, "Neutral" is highly confused with other classes, because the data of this category are similar to other emotions.

Table 3. Statistics of the accuracy and F1-score.

Method	Acc.(%)		F_1(%)	
	Audio	Visual	Audio	Visual
Pandeya et al. [23]	74.0	**74.0**	73.0	**73.0**
AudioCLIP [5]	18.3	34.0	12.6	32.9
Wav2CLIP [34]	16.3	12.7	4.7	11.5
Ours w cls	**82.0**	69.7	**82.0**	70.0
Ours w/o cls	76.7	67.7	76.9	67.7
Ours w/o text-enc	75.3	63.2	75.1	63.2
Ours w/o pretrain	69.0	56.8	69.2	56.6

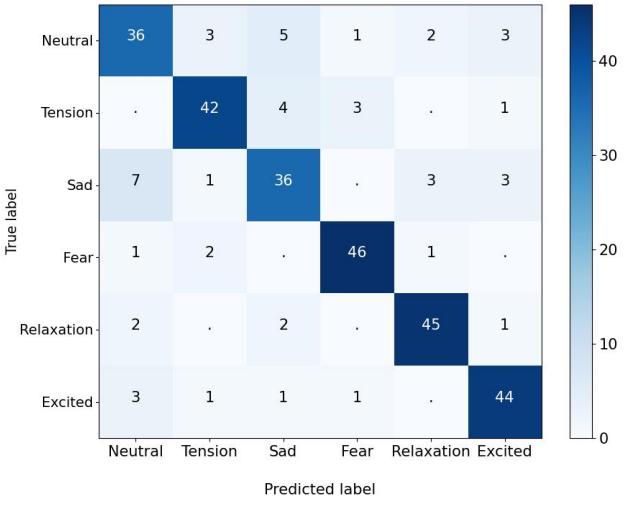


Figure 6. The confusion matrix of emotion classification.

Statistics of the accuracy and F1-score. We compare with other methods in terms of Top-1 accuracy (Acc.) and F1-Score (F_1) to prove the discriminability of our emotion space by feeding signals in different modalities. We select Pandeya et al. [23], vanilla AudioCLIP [5], and Wav2CLIP [34] as baselines. The results are shown in Table 3, where "Ours w cls" and "Ours w/o cls" correspond to training models with and without classifier, respectively. "Ours w/o text-enc" means to remove the text encoder and do not use the text modality to enhance the feature. And, "Ours w/o pretrain" means the encoder trained from scratch. We demonstrate the effectiveness of our full framework by comparing the classification performance of encoders under various conditions. We achieve the best Top-1 accuracy and F1-Score on the emotion classification task of music videos. The AudioCLIP and Wav2CLIP completely lost the ability to classify video emotions due to the constraint of original pretrained dataset. Compared to [23], we also achieve the highest performance in the audio modality.

4.5. Video Demo

To demonstrate the effectiveness of our framework, we provide a video demo that consists of movie montages guided by various pieces of emotional music. Two specific tasks are presented in the video.

The first task is to generate movie montages driven by different pieces of emotional music for a single movie. We list two different movies, including "Forrest Gump" and "Leon". For the same movie, we process it with our pipeline, adding a piece of 30-second emotional music, such as "Forrest Gump" edited by relaxation and neutral music and "Leon" with excited and tension music. Apparently, we can easily observe the difference in the montage results. For example, the beautiful landscape (i.e., sea and forest) frames are mainly picked in the montage when a music with relaxation emotion is used as guidance. On the contrary, with a piece of neutral music, the movie montage often contains static pictures, for example, the expressionless man sitting on the chair. Based on our solver, we successfully select the suitable set of movie shots to create a montage, which leads to the disparity.

The second task in our demo is to create the montage using the corresponding theme song of the movie. Two movies, "Mulan" and "The Grand Budapest Hotel", are used as raw materials. In this case, we show our framework can create a montage that fits the overall mood and rhythm of the movie according to the theme song.

5. Conclusion

In this paper, we present a novel emotion-aware music-driven movie montage task, which raises a challenge of retrieving and recombining shots in a long movie based on a user-specified music clip. We formally define it as an optimization problem and propose a two-stage framework which consists of a learning based module for prediction of emotion similarity and an optimization based module for selection and composition of candidate movie shots. In the first stage, we train a three-modality CLIP based model by a contrastive loss to select candidate shots. In the second stage, inspired by Murch's montage criterion [21], we design an emotion-aware optimization solver under a scene-level constraint for story completeness and a shot-level constraint for audio-visual rhythm synchronization to search optional schemes. By qualitative and quantitative evaluations, we demonstrate our method can generate emotionally consistent montages and outperforms alternative baselines.

References

[1] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015. 3

[2] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer. madmom: a new Python Audio and Music Signal Processing Library. In *Proceedings of ACM International Conference on Multimedia*, pages 1174–1178, 10 2016. 7

[3] A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017. 3

[4] S. Gross, X. Wei, and J. Zhu. Automatic realistic music video generation from segments of youtube videos. *arXiv preprint arXiv:1905.12245*, 2019. 2

[5] A. Guzhov, F. Raue, J. Hees, and A. Dengel. AudioCLIP: Extending Clip to Image, Text and Audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980, 2022. 3, 9, 10

[6] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool. Creating summaries from user videos. In *European Conference on Computer Vision*, pages 505–520, 2014. 3

[7] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005. 3

[8] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa. Automatic trailer generation. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 839–842, 2010. 3

[9] P. N. Juslin and P. Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238, 2004. 3

[10] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *Proc. ismir*, volume 86, pages 937–952, 2010. 3

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[12] F.-F. Kuo, M.-K. Shan, and S.-Y. Lee. Background music recommendation for video based on multimodal latent se-

mantic analysis. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2013. 2

[13] Z. Liao, Y. Yu, B. Gong, and L. Cheng. Audeosynth: Music-driven video montage. *ACM Transactions on Graphics*, 34(4):68:1–68:10, jul 2015. 2

[14] J.-C. Lin, W.-L. Wei, and H.-M. Wang. Emv-matchmaker: emotional temporal course modeling and matching for automatic music video generation. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 899–902, 2015. 1, 2

[15] J.-C. Lin, W.-L. Wei, and H.-M. Wang. Automatic music video generation based on emotion-oriented pseudo song prediction and matching. In *Proceedings of ACM International Conference on Multimedia*, pages 372–376, 2016. 2

[16] J.-C. Lin, W.-L. Wei, and H.-M. Wang. Demv-matchmaker: Emotional temporal course representation and deep similarity matching for automatic music video generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2772–2776, 2016. 1, 2

[17] J.-C. Lin, W.-L. Wei, J. Yang, H.-M. Wang, and H.-Y. M. Liao. Automatic music video generation based on simultaneous soundtrack recommendation and video editing. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 519–527, 2017. 2, 8

[18] T. Liu and J. R. Kender. Optimization algorithms for the selection of key frame sequences of variable length. In *European Conference on Computer Vision*, pages 403–417, 2002. 3

[19] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang. On shape and the computability of emotions. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 229–238, 2012. 3

[20] K. M. Mahmoud, N. M. Ghanem, and M. A. Ismail. Unsupervised video summarization via dynamic modeling-based hierarchical clustering. In *International Conference on Machine Learning and Applications*, volume 2, pages 303–308, 2013. 3

[21] W. Murch. *In the Blink of an Eye*, volume 995. Silman-James Press Los Angeles, 2001. 1, 10

[22] M. Narasimhan, A. Rohrbach, and T. Darrell. Clip-it! language-guided video summarization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13988–14000, 2021. 1, 3

[23] Y. R. Pandeya, B. Bhattarai, and J. Lee. Deep-learning-based multimodal emotion classification for music videos. *Sensors*, 21(14):4927, 2021. 9, 10

[24] Y. R. Pandeya and J. Lee. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80(2):2887–2905, 2021. 4, 7

[25] P. Papalampidi, F. Keller, and M. Lapata. Film trailer generation via task decomposition. *arXiv preprint arXiv:2111.08774*, 2021. 3

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4

[27] A. Rao, L. Xu, Y. Xiong, G. Xu, Q. Huang, B. Zhou, and D. Lin. A local-to-global approach to multi-modal movie scene segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10146–10155, 2020. 7

[28] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 3

[29] A. Sharghi, B. Gong, and M. Shah. Query-focused extractive video summarization. In *European Conference on Computer Vision*, pages 3–19, 2016. 3

[30] J. R. Smith, D. Joshi, B. Huet, W. Hsu, and J. Cota. Harnessing ai for augmenting creativity: Application to movie trailer creation. In *Proceedings of ACM International Conference on Multimedia*, pages 1799–1808, 2017. 3

[31] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5179–5187, 2015. 3

[32] T. Souček and J. Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 7

[33] J.-C. Wang, Y.-H. Yang, I.-H. Jhuo, Y.-Y. Lin, and H.-M. Wang. The acousticvisual emotion guassians model for automatic generation of music video. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1379–1380, 2012. 2

[34] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello. Wav2clip: Learning robust audio representations from clip. *arXiv preprint arXiv:2110.11499*, 2021. 9, 10

[35] H. Xu, Y. Zhen, and H. Zha. Trailer generation via a point process-based visual attractiveness model. In *Proceedings of International Conference on Artificial Intelligence*, pages 2198–2204, 2015. 3

[36] J.-C. Yoon, I.-K. Lee, and S. Byun. Automated music video generation using multi-level feature-based segmentation. *Multimedia Tools and Applications*, 41(2):197–214, 2009. 2

[37] X.-J. Yu. A study on the editing frequencies trends for films emotion clips. *International Journal on Organizational Innovation*, 9(3):40A, 2017. 5

[38] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 47–56, 2014. 3

[39] W. Zhu, J. Lu, J. Li, and J. Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962, 2020. 3