Use of Affective Visual Information for Summarization of Human-Centric Videos

Berkay Köprü and Engin Erzin, *Senior Member, IEEE*

Abstract—The increasing volume of user-generated human-centric video content and its applications, such as video retrieval and browsing, require compact representations addressed by the video summarization literature. Current supervised studies formulate video summarization as a sequence-to-sequence learning problem, and the existing solutions often neglect the surge of the human-centric view, which inherently contains affective content. In this study, we investigate the affective-information enriched supervised video summarization task for human-centric videos. First, we train a visual input-driven state-of-the-art continuous emotion recognition model (CER-NET) on the RECOLA dataset to estimate activation and valence attributes. Then, we integrate the estimated emotional attributes and their high-level embeddings from the CER-NET with the visual information to define the proposed affective video summarization (AVSUM) architectures. In addition, we investigate the use of attention to improve the AVSUM architectures and propose two new architectures based on temporal attention (TA-AVSUM-GRU) and spatial attention (SA-AVSUM-GRU). We conduct video summarization experiments on the TvSum and COGNIMUSE datasets. The proposed temporal attention-based TA-AVSUM architecture attains competitive video summarization performances with strong improvements for the human-centric videos compared to the state-of-the-art in terms of F-score, self-defined face recall, and rank correlation metrics.

Index Terms—Affective computing, video summarization, continuous emotion recognition, neural networks

1 Introduction

The internet infrastructure and services have recently matured to the point that any user can record and share unedited videos, resulting in over 500 hours of footage being uploaded per minute to video streaming platforms like YouTube and Twitch. The vast majority of these uploads tend to be user-generated human-centric videos focusing on human activities, interactions, and their emotional and physical attributes [1]. Video summarization, which preserves human-centric representations of emotional and visual elements, looks to be a key challenge for such a massive collection of unedited human-centric videos [2], [3], [4].

Summarization of human-centric videos should address the problem of selecting salient video segments or key frames and constructing an informative, interesting, and short summary of the human-centric actions in the video. Emotional changes are frequently associated with notable human-centric actions in the videos. The continuous emotion recognition (CER) problem has been extensively studied in recent literature for tracking the emotional changes as the prediction of affect attributes from audio and visual modalities [5], [6], [7], [8]. In affective and emotional settings, high activation and extreme, low or high, valence attributes are expected to appear with informative and interesting human-centric actions. This research aims to look into the relationship between affective information and human-centric actions to improve the human-centric video summarization task.

The coupling of affective computing and video summarization has received little attention in the literature. Two related tracks of study for this coupling appear as: (i) studies modeling the human actions in videos [3], [9]; (ii) studies modeling the emotional responses of the viewers [10], [11].

In a recent work for the first track, Bhattacharya et al. use human-centric modalities such as body poses and faces, and model the inter- and intra-relations between humans with graphs on the human-centric videos [9]. In [3], Sun et al. address the summarization of the unconstrained videos by first finding salient people in the video and then creating montages capturing the salient actions of these people. Although both studies utilize human actions, which are correlated with the emotional state of humans to an extent, they do not focus on the emotional content or saliency in the emotional domain.

In the second track, video summarization selects key frames based on the physiological responses of the viewers that are expected to be highly correlated with the viewers' emotional states [10], [11]. In [10], the viewers' facial activity is tracked, and then the frames are ranked to extract personal highlights from the videos by using heuristics. Whereas in the other study, the viewers' physiological responses, such as heart rate and electrodermal response, are tracked to analyze sub-segments of the watched videos [11]. These studies leverage physiological information that is highly related to humans' emotional states to generate personal highlights/summaries. On the other hand, these approaches require at least one subject as a video

- *Berkay Köprü is with KUIS AI Lab and Electrical & Electronics Engineering Department, Koç University, 34450 Istanbul, Turkey. E-mail: bkopru17@ku.edu.tr.*
- *Engin Erzin is with KUIS AI Lab, Computer Engineering Department and Electrical & Electronics Engineering Department, Koç University, 34450 Istanbul, Turkey. E-mail: eerzin@ku.edu.tr.*

Manuscript received 7 July 2021; revised 14 October 2022; accepted 10 November 2022. Date of publication 17 November 2022; date of current version 29 November 2023.
(Corresponding author: Berkay Köprü.)
Recommended for acceptance by F. Ringeval.
Digital Object Identifier no. 10.1109/TAFFC.2022.3222882

viewer and do not provide feasible and time-efficient solutions for automatic video summarization.

In this paper, we study the use of affective information extracted from visual data to summarize human-centric videos, which we call affective video summarization. We model affective states from the video itself, unlike studies that evaluate viewers' perspectives [10], [11]. Our model also targets the summarization of human-centric videos, but not specifically over visual human-centric modalities as in [9], but using the affective information extracted from the visual data.

For affective video summarization, we adopt a two-step approach. First, a convolutional recurrent neural network-based affective analysis model has been constructed with the *GoogLeNet* driven with visual features [12]. Then, we explore affective feature fusion over the affect attributes and affect embeddings together with the temporal and spatial attention mechanisms to define the proposed fully convolutional network-based affective video summarization models. To the best of our knowledge, this is the first work in literature that combines affective analysis with video summarization for the affective summarization of human-centric videos.

The visual affective analysis model has been trained over the REmote COLlaborative and Affective (RECOLA) dataset [13]. To show the benefit of our proposed affective video summarization model, we evaluated it on the TvSum [14], and COGNIMUSE [4] datasets using F-score, face recall, cumulative Kendall's rank correlation coefficient, and Spearman's rank correlation metrics in comparison with four video summarization baselines from the recent literature. The proposed temporal attention-based affective video summarization model is observed to sustain higher performance than the baselines, with consistent improvements for all performance metrics.

To summarize, the main contributions of this study are as follows:

- We formulate a novel end-to-end learning problem for affective-information enriched video summarization targeting human-centric videos.
- A CER model extracts affective information in terms of emotional attributes and learned embeddings from the visual inputs.
- We investigate the use of attention mechanisms in video summarization for human-centric videos and develop temporal and spatial attention-based frameworks.
- We conduct affective video summarization evaluations on the state-of-the-art using both standard and self-defined evaluation metrics.

The rest of this paper is organized as follows. Section 2 reviews related work, and Section 3 describes the main building blocks of the proposed framework. Section 4 presents the experiments conducted together with the performance evaluations. Finally, the conclusion is presented in Section 5.

2 Related Work

In this paper, we are investigating the video summarization problem for human-centric videos with an attention to emotional changes in the video content. This brings a novel perspective to the affective video summarization task. In the recent literature, several lines of work are related to this task.

While video summarization is often defined as a supervised sequence-to-sequence mapping problem over encoder-decoder architectures [15], [16], [17], [18], it has also been formulated under unsupervised settings to maximize a reward function for diversity and representativeness of summaries [19] or to minimize distribution differences between videos and their summarizations [20], [21]. In the supervised approaches, while LSTM models are used to capture long-range dependencies in [16], [17], [18], Rochan et al. differ from them by processing all frames of subsampled video through a fully convolutional neural network (SUM-FCN) [15]. On the other hand, Zhou et al. formulate the summarization task as sequential decision-making using an end-to-end reinforcement learning-based framework (DR-DSN) [19]. They utilize a reward function that jointly accounts for the diversity and representativeness of the generated summaries in an unsupervised setting. Mahasseni et al. propose a deep LSTM-based summarizer and a discriminator network (SUM-GAN) to minimize the distribution of keyframes and the original video in an unsupervised manner [20]. The summarizer selects keyframes and then reconstructs the input video, while the discriminator distinguishes the original video from its reconstruction resulting in an adversarial learning framework. Later, Apostolidis et al. introduce an attention mechanism to SUM-GAN by integrating a deterministic attention auto-encoder (SUM-GAN-AAE) [21].

Several recent works in video summarization adapt attention to enhance summarization performance [16], [21], [22], [23]. Early work by Fajtl et al. employs a sequence-to-sequence transformation for video summarization using a soft self-attention mechanism by defining a frame-level meta-representation as a weighted combination of features from all time frames [22]. However, the order of information across frames is lost during this process, which could be essential for video summarization. Jung et al. extend the use of self-attention to capture the positional information by introducing global-and-local relative position embeddings [23]. In another study, Ji et al. encode visual features with a Bidirectional LSTM network [16]. Encoded vectors combine Bahdanau attention [24] and then feed into an LSTM based decoder. Although our formulation of video summarization follows a sequence-to-sequence mapping with attention mechanisms, our study differs from these studies in the exploitation of affective information for the video summarization task.

Affective information processing attempts to define and extract attributes of emotion. *Emotions* are defined as mental states that are triggered by neurophysiological changes and are related to feelings, and behavioral responses [25]. Behavioral responses co-articulate with speech, bodily gestures, and facial expressions. Hence human-centric videos are expected to carry emotional cues over the behavioral responses [4]. Emotion recognition from audio-visual signals has been extensively studied in the literature [5], [6], [8]. Emotions are represented both in discrete and continuous domains. Discrete categorical emotions, such as happiness and sadness, can also be represented in the 3-dimensional continuous affect space of activation, valence, and dominance, which are the indicators of activenesspassiveness, positiveness-negativeness, and dominance-submissiveness, respectively [26], [27].

Prediction of activation, valence and dominance attributes from behavioral responses are categorized as continuous emotion recognition (CER) task [7], [8], [28]. In the recent literature, deep neural network (DNN) based approaches have been extensively used for the CER task. Schmitt et al. investigate arousal and valence prediction from the low-level descriptors (LLDs) of speech signals using RNNs with the CCC loss function [28]. Tzirakis et al. design an end-to-end network utilizing raw video, audio, and text modalities [7]. They construct a multimodal network having an attention layer to fuse deep embeddings of the text, audio, and visual modalities and an LSTM layer that predicts the activation and valence attributes. In another study, we performed a feature-level audio-visual information fusion to train a CRNN model with multi-task learning for the prediction of affect attributes [8].

Although the intersection of affective computing and video summarization is narrow, several studies focus on emotion recognition while either relating their findings to video summarization [29], [30] or utilizing similar ideas like shot segmentation and keyframe selection for emotion recognition [31]. Xu et al. investigate information transfer from the image and textual data of videos for emotion recognition, emotion attribution, and emotion-oriented summarization [29]. Using zero-shot learning to estimate categorical emotions, they learn video representations over image transfer-encoding and textual representations. Later, they perform emotion attribution to identify the contribution of each frame to the video's overall emotion. Finally, video summarization is formulated as a selection of keyframes by maximizing an emotion attribute-based score function. In the second related study, Tu et al. train a joint model to capture emotion attribution and recognition using multitask learning [30]. Later, similar to the previous study [29], the video summarization task is formulated as a post-processing optimization problem and solved using MINMAX dynamic programming. Wang et al. segment the videos into shots using audio-visual information and represent each shot with three keyframes [31]. Then, they process these keyframes with an action detector, face detector, and person detector while extracting low-level descriptors and high-level features from audio by OpenSmile tool [32] and a pre-trained VGGish network. Concatenation of the outputs from these detectors and extractors is then fed into an LSTM with temporal attention to estimate arousal and valence. Note that both [29] and [30] formulate the summarization task as an optimization problem, which is solved in post-processing. Hence their video summarization frameworks are not learning-based, and they do not explore how affective information alters the behavior of the proposed summarization architecture. Furthermore, the proposed solutions are not evaluated on the state-of-the-art video summarization datasets. On the other hand, although [31] utilizes shot segmentation and key frame selection, they do not include these processes in training; hence these are not updated regarding any saliency in the affective attribute space.

There are also video summarization studies where affect-related attributes or modalities are utilized to generate human-centric video summaries [3], [9], [33]. Koutras et al. extract audio-visual and affective text features to classify each frame as salient or not using a k-nearest neighbors classifier [33]. Affective text features are generated from their distance to selected seed words with known affective attributes. Sun et al. perform a human pose detection to locate people and their trajectories and detect salient human-centric segments using a random forest classifier from the pose, motion, and self-defined camera-centric features [3]. Then they generate salient montages representing the video using the human trajectories and saliency scores. In another study utilizing human poses, Bhattacharya et al. represent humans as graphs whose nodes are either 2 d face landmarks or 3 d body joints of humans [9]. Then using an auto-encoder type of architecture based on a spatial-temporal graph convolutional network, they estimate the importance scores of frames. Although these studies utilize human-centric cues, our study differs from them as we focus directly on extracting affective attributes and high-level emotion embeddings from visual data to locate salient affective regions with attention mechanisms.

3 Methodology

In this paper, we investigate the use of affective information for enriching video summarization by capturing emotionally salient regions of human-centric videos. First, we state and formulate the video summarization problem and define the visual feature extraction for both the CER and video summarization tasks. Then, an end-to-end framework for the CER is presented. Emotional attributes and high-level embeddings from the CER framework are later used as affective representations by the video summarization framework. Finally, we introduce the proposed affective video summarization architectures by first defining a video summarization baseline and then enriching this baseline with the fusion of affective information.

3.1 Problem Statement

Video summarization is widely formulated as either a binary classification or a frame-level regression task. In the binary classification task, summarization outputs are either key-frames [15], [34] or key-shots [14], [34] from the video. On the other hand, frame-level importance scores are extracted in the regression task [16], [34].

In this study, we formulate video summarization as a binary classification task where the positive labels correspond to the selected key-frames. The summarization network receives a visual feature matrix, $\mathbf{v} \in \mathbb{R}^{N \times D}$, and emits an output matrix, $\mathbf{s} \in \mathbb{R}^{N \times C}$, where N is the number of frames in the video, D is the dimensionality of the frame-level visual feature, and C is the number of classes. We take $C = 2$ representing the positive and negative classes for the key-frame selection, and these two nodes output class probability values at the output of the network. Then the positive class for key-frame selection or the negative class for frame-skip is set by picking the node with a higher probability. Eventually, the video summary is constructed from the key-frames that are labeled as positive.

In this study, we extract the visual information using the *GoogLeNet* [12]. Output of the *pool5* layer of the pre-trained

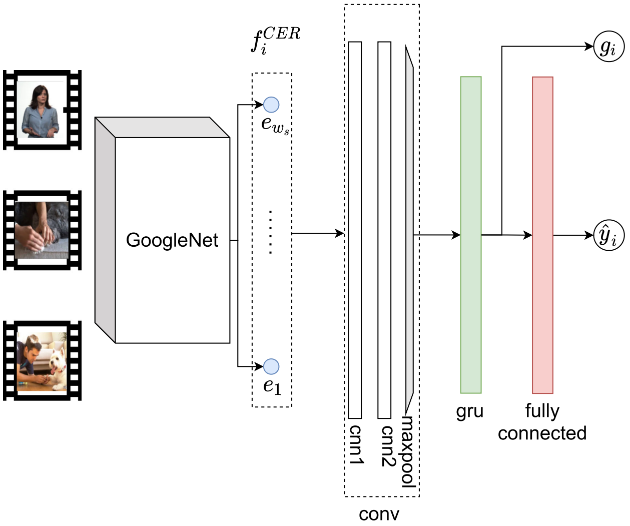


Fig. 1. CER-NET: The continuous emotion recognition network.

GoogLeNet is used as the visual feature and represented as $\mathbf{e}_i \in \mathbb{R}^D$ at frame i with dimension $D = 1024$.

3.2 CER Network (CER-NET)

The continuous emotion recognition problem is set as the continuous regression of an emotional attribute from the temporal visual features. For this purpose, we construct a CER network (CER-NET), which consists of two back-to-back convolutional layers, a max pool layer in the temporal domain, a Gated Recurrent Unit (GRU) layer, and a fully connected layer as shown in Fig. 1. We use the CER-NET to train two separate networks to estimate the activation and valence (AV) attributes separately. A temporal visual feature matrix is defined to be the input of the CER-NET. For this purpose, visual features around the ith frame are cascaded to define the temporal visual feature as

$$\mathbf{f}_i^{\mathrm{CER}} = [\mathbf{e}_{i-\Delta+1}, \ldots, \mathbf{e}_{i-1}, \mathbf{e}_i, \mathbf{e}_{i+1}, \ldots, \mathbf{e}_{i+\Delta}] \in \mathbb{R}^{D\times T}, \tag{1}$$

at frame i, where T is the temporal window size and it is set as $T = 2\Delta = 20$ frames.

We refer to the group of back-to-back two convolutional and max-pool layers as the *conv* layer for the sake of simplicity. In CER-NET, the *conv* layer models the spatial relations and provides a compact representation of the temporal window. In this manner, both temporally related features are highlighted, and the dimensionality of the representation is reduced. Dimensionality reduction is the key to complexity reduction, and it also prevents overfitting. The compact representation from the *conv* layer is fed into GRU to model long-term temporal relations.

At the inference phase, CER-NET receives $\mathbf{f}_i^{\mathrm{CER}}$ and provides the GRU layer outputs as $\mathbf{g}_i \in \mathbb{R}^G$ and the fully connected layer outputs as $\hat{y}_i \in \mathbb{R}$. We define the GRU layer output, $\mathbf{g}_i$, as an affective embedding, which can carry affective information to the video summarization model. On the other hand, the fully connected layer output, $\hat{y}_i$, represents the estimated emotional attribute (A or V) at frame i and delivers another source of affective information.

3.3 Affective Video Summarization

The proposed affective video summarization (AVSUM) architecture is based on a fully convolutional neural network (FCN) for semantic segmentation [35], which is adapted by [15] into video summarization (SUM-FCN). In the following, we briefly describe SUM-FCN and then present two fusion architectures to combine affective information with video summarization, which are later extended by temporal and spatial attention mechanisms.

3.3.1 SUM-FCN

SUM-FCN adopts an encoder-decoder architecture where the encoder is based on fully convolutional layers, and the decoder is based on deconvolutions. Fig. 2 includes the architecture of the SUM-FCN compromising encoder-bottleneck-decoder layers driven by the visual input $\mathbf{v}$. The encoder compresses the temporal information while increasing spatially, and the bottleneck passes it to the decoder to reconstruct necessary information from this compact representation. SUM-FCN receives the visual features for the whole video at once and provides the summarization outputs at once. The video frame rate is typically down-sampled before the summarization. Let $\mathbf{v}$ be the input of SUM-FCN, then $\mathbf{v} = [\mathbf{e}_1, \ldots, \mathbf{e}_N]' \in \mathbb{R}^{N\times D}$ where N is number of frames in the down-sampled stream. Given $\mathbf{v}$, SUM-FCN emits $\mathbf{s} \in \mathbb{R}^{N\times C}$, where C is the dimension of the summarization annotations and set as $C = 2$ as defined in Section 3.1.

3.3.2 Affective Feature Fusion for AVSUM

Affective information is extracted from the two CER-NET models, which are trained to estimate the activation and valence (AV) attributes separately. Two types of affective information cues are extracted from the CER-NET models: (i) the estimated AV attributes ($\mathbf{f}^{\mathrm{AV}}$), and (ii) the learned high-level CER-NET representations based on the GRU embeddings ($\mathbf{g}^{\mathrm{AV}}$). The estimated AV attribute vector $\mathbf{f}_j^{\mathrm{AV}}$ is constructed as a column vector from the estimated activation and valence attributes as $\mathbf{f}_j^{AV} = [\hat{y}_j^A, \hat{y}_j^V]' \in \mathbb{R}^2$ at frame j in the down-sampled stream. Similarly, $\mathbf{g}_j^{\mathrm{AV}}$ is constructed by concatenating the outputs of the GRU layers from the two CER-NET models as $\mathbf{g}_j^{\mathrm{AV}} = [\mathbf{g}_j^{\mathrm{A}\prime}, \mathbf{g}_j^{\mathrm{V}\prime}]' \in \mathbb{R}^{2\,G}$. Then, the emotional attribute $\mathbf{f}^{\mathrm{AV}}$ and the affect embedding $\mathbf{f}^{\mathrm{GRU}}$ representations of the video are defined as

$$\mathbf{f}^{\mathrm{AV}} = [\mathbf{f}_1^{\mathrm{AV}}, \ldots, \mathbf{f}_N^{\mathrm{AV}}]' \in \mathbb{R}^{N\times 2} \tag{2}$$

$$\mathbf{f}^{\mathrm{GRU}} = [\mathbf{g}_1^{\mathrm{AV}}, \ldots, \mathbf{g}_N^{\mathrm{AV}}]' \in \mathbb{R}^{N\times 2\,G}. \tag{3}$$

The first proposed AVSUM architecture, referred to as AVSUM-GRU, combines the affect embedding $\mathbf{f}^{\mathrm{GRU}}$ with the visual input $\mathbf{v}$. Fig. 2 presents the AVSUM-GRU architecture receiving the $\mathbf{v}_{\mathrm{GRU}}$ input as

$$\mathbf{v}_{\mathrm{GRU}} = \mathbf{v} \oplus \mathbf{f}^{\mathrm{GRU}} \in \mathbb{R}^{N\times(D+2\,G)}, \tag{4}$$

where the $\oplus$ operator is representing the feature vector combining over the whole video.

Alternatively, we define the AVSUM-SCAV architecture, as illustrated in Fig. 2, by combining the emotional attribute $\mathbf{f}^{\mathrm{AV}}$ with the visual input $\mathbf{v}$ to receive $\mathbf{v}_{AV}$ as

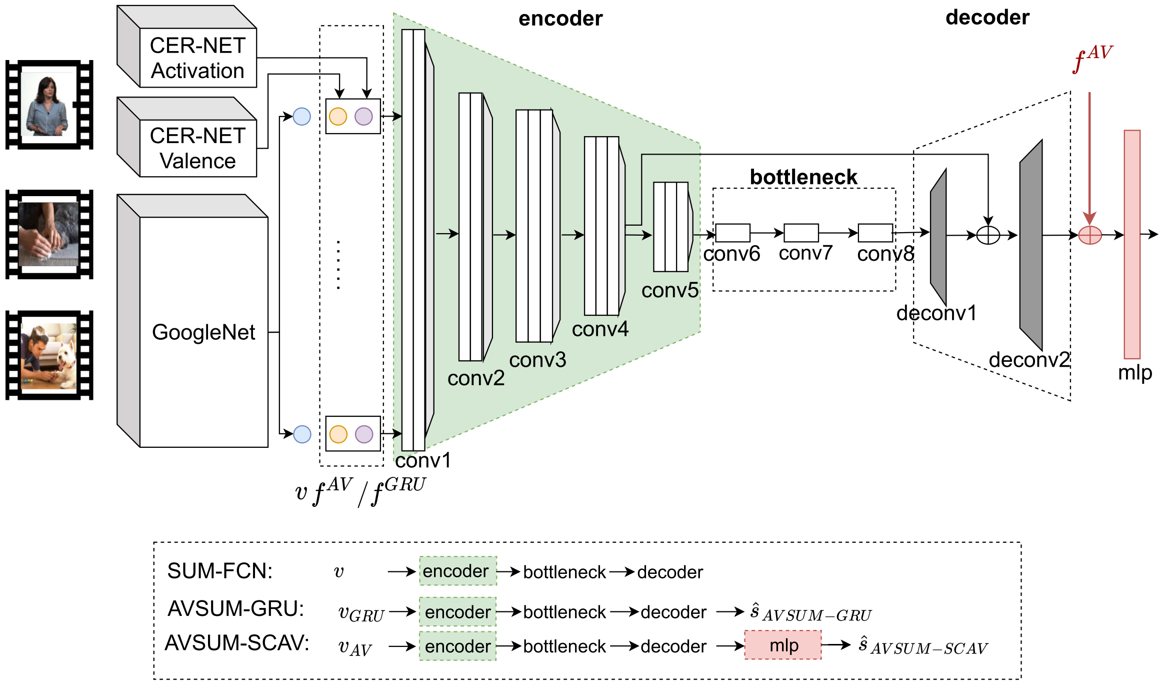

encoder
decoder
f^{AV}
CER-NET Activation
CER-NET Valence
GoogleNet
conv1
conv2
conv3
conv4
conv5
bottleneck
conv6
conv7
conv8
deconv1
deconv2
mlp
$v\ f^{AV} / f^{GRU}$
SUM-FCN: v → encoder → bottleneck → decoder
AVSUM-GRU: v_{GRU} → encoder → bottleneck → decoder → $\hat{s}_{AVSUM-GRU}$
AVSUM-SCAV: v_{AV} → encoder → bottleneck → decoder → mlp → $\hat{s}_{AVSUM-SCAV}$

Fig. 2. An overview of the proposed AVSUM-GRU and AVSUM-SCAV architectures. AVSUM-GRU processes v_{GRU} which combines the high-level affective information f_{GRU} from the CER-NET estimators with the GoogleNet drived visual features. AVSUM-SCAV processes v_{AV} as the input and defines a skip-connection for the emotional attributes f_{AV} as well as combining them with the GoogleNet drived visual features for the video summarization.

$$\mathbf{v}_{\text{AV}} = \mathbf{v} \oplus \mathbf{f}^{\text{AV}} \in \mathbb{R}^{N\times(D+2)}, \tag{5}$$

and incorporating a long skip-connection by concatenating the emotional attribute $\mathbf{f}^{\text{AV}}$ to the final layer of the summarization network. AVSUM-SCAV inserts a skip-connection to the output of *deconv2* layer and has a final fully connected *mlp* layer to reduce the dimension from $C+2$ to C.

3.3.3 Temporal Attention for AVSUM

We adapt the multi-headed attention (MHA) mechanism into the AVSUM architecture to efficiently model temporal dependencies across the video frames and refer to as TA-AVSUM-GRU. Similar to AVSUM-GRU, TA-AVSUM-GRU receives affective information enriched features $\mathbf{v}_{\text{GRU}}$ as the input, and in addition to AVSUM-GRU it employs attention to the temporally compressed sequence. Fig. 3 depicts the MHA based temporal attention structure, where MHA is placed to the output of *conv4* layer, which emits $\mathbf{X} \in \mathbb{R}^{M\times S}$. Here, M and S are, respectively, the temporal and spatial dimensions of $\mathbf{X}$.

MHA receives three inputs as Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$), then outputs a weighted summation of the rows of $\mathbf{V}$. The weights are calculated from the similarity between the $\mathbf{Q}$ and $\mathbf{K}$. In this context, the MHA is defined as

$$\text{head}_h = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{W}_h^Q(\mathbf{K}\mathbf{W}_h^K)^T}{\sqrt{M}}\right)\mathbf{V}\mathbf{W}_h^V \tag{6}$$

$$\text{MHA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\text{head}_1,\ldots,\text{head}_H)\mathbf{W}^O, \tag{7}$$

where $\mathbf{W}_h^Q$, $\mathbf{W}_h^K$, $\mathbf{W}_h^V$ and $\mathbf{W}^O$ are the learned linear projections, and H is the number of heads. We employ multi-headed self-attention by setting $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ to $\mathbf{X}$. Hence, the learned projections matrices are formed as $\mathbf{W}_h^Q$, $\mathbf{W}_h^K$, $\mathbf{W}_h^Q \in \mathbb{R}^{S\times\frac{S}{H}}$, and $\mathbf{W}^O \in \mathbb{R}^{M\times M}$.

3.3.4 Spatial Attention for AVSUM

Motivated by the Squeeze and Excitation Networks [36], which attend to different channels of an image, we propose the fourth AVSUM network with spatial domain attention and refer to it as SA-AVSUM-GRU. Fig. 3 depicts the SA-AVSUM-GRU structure where we employ attention to the output of the *deconv2* layer. Unlike TA-AVSUM-GRU, we adopt single-headed attention, which is formulated in (6). Then, $\mathbf{K}$ and $\mathbf{V}$ are set to transpose of $\hat{\mathbf{s}}_{\text{AVSUM}}$ and $\mathbf{Q}$ is set from the affective embedding $\mathbf{f}^{\text{GRU}}$.

3.4 Model Training

Training of the CER-NET and video summarization models are executed in two phases. First, the CER-NET models are trained, later they are fixed and integrated for the affective video summarization to train the AVSUM-GRU, AVSUM-SCAV, TA-AVSUM-GRU, and SA-AVSUM-GRU models.

3.4.1 CER-NET

The CER-NET models are trained separately to estimate the activation and valence attributes using the concordance correlation coefficient (CCC) based loss function. The loss function of the CER-NET is defined as the negated CCC value,

$$L_{\text{CER}} = -\frac{2\sigma_{y\hat{y}}^2}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}, \tag{8}$$

where y is the ground truth and $\hat{y}$ is the estimated attribute.

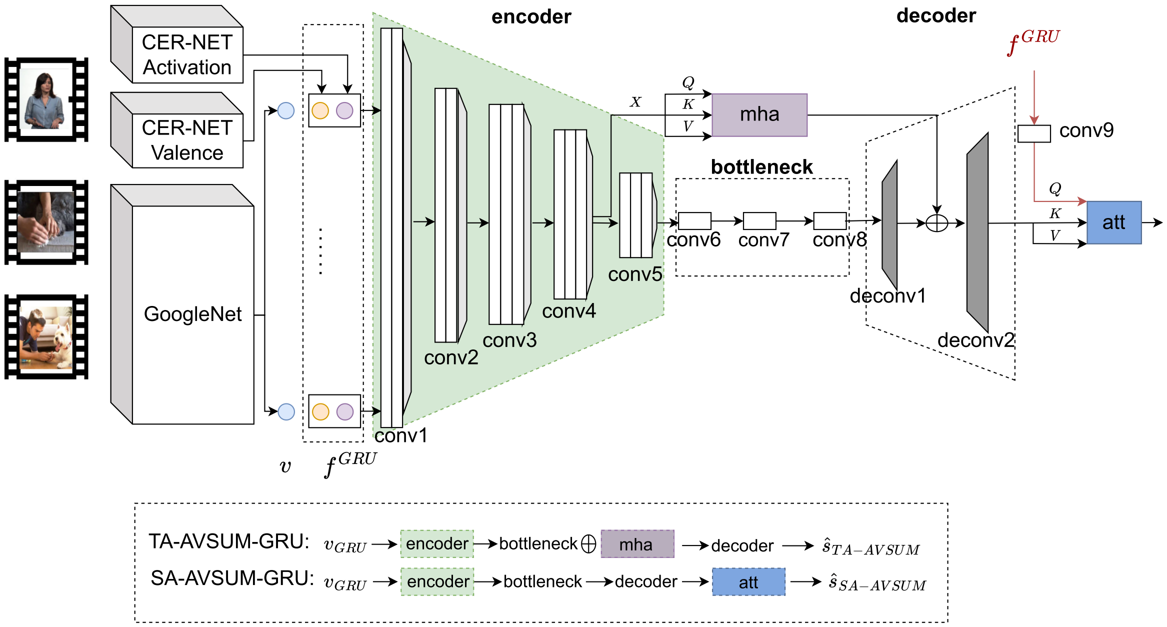


Fig. 3. An overview of the proposed TA-AVSUM-GRU and SA-AVSUM-GRU architectures. TA-AVSUM-GRU applies multi-head-self attention to the output of *conv4* layer (X), on top of AVSUM architecture. On the other hand, SA-AVSUM-GRU modifies the long skip connection of affective features by applying spatial attention.

3.4.2 Video Summarization Networks

The key frame selection problem has an imbalanced nature since only a small number of frames are selected for the summary [15]. In order to overcome the imbalance problem, a weighted binary cross entropy loss is defined as

$$L_{\text{SUM}} = -\frac{1}{N}\sum_{j=1}^{N} w_{z_j}(z_j \log \hat{z}_j + (1-z_j)\log(1-\hat{z}_j)), \tag{9}$$

where $z_j \in 0, 1$ is the binary ground truth, $\hat{z}$ is the predicted score and w_{z_j} is the weight of the j^{th} frame. The weights for the binary target are defined as

$$w_0 = \frac{1}{N}\sum_{j=1}^{N} z_j \quad \text{and} \quad w_1 = 1 - w_0. \tag{10}$$

4 Experiments

Experimental evaluations of the proposed models and comparisons with the state-of-the-art are performed using two datasets. In this section, we first introduce the datasets and evaluation metrics, then explain implementation details. Finally, experimental results are presented and discussed.

4.1 Datasets

We train and evaluate the CER-NET on the RECOLA dataset, which is a popular multi-modal dataset for emotion recognition [13]. The RECOLA dataset is composed of multi-modal recordings of dyadic conversations from 27 French speakers. Of these 27 recordings, 18 of them are annotated, and the rest of the records are used for testing. The annotations are at the rate of 40 msec and from 6 different annotators. In total, we used 90 minutes of recordings from the RECOLA dataset in this study.

Experimental evaluations on the proposed AVSUM architectures are executed on the frequently used TvSum [14], and COGNIMUSE [4] datasets. Note that there is no available video summarization dataset containing only human-centric videos in the literature. The TvSum dataset contains 50 user-generated videos from 10 different categories, such as vehicle tire changing, sandwich making, grooming an animal, etc. The demographics of the dataset in terms of the face-including frames in the summary and the full video clip are presented in Fig. 4. We categorize videos into human-centric and rest regarding the number of faces in summary, where we set the threshold to 80 frames labeling 15 videos as human-centric. The frame-level importance scores are provided as ground truths for each video from 20 raters. We follow the approach in [15], [34] to convert the frame-level importance scores into key shot-based summaries.

The COGNIMUSE dataset contains half-hour continuous segments from seven Hollywood movies and is annotated with sensory and semantic saliency, audio and visual events, cross-media relations as well as emotion [4]. Emotions are annotated at the frame level for arousal and valence as continuous values ranging between -1 and 1. Saliency annotations are generated from audio, video,

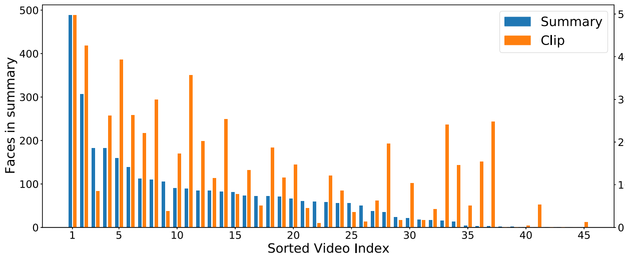


Fig. 4. Number of face including frames in the summary and total video for TvSum dataset where the videos are sorted regarding the number of faces in the summary.

semantics, audio-visual, and audio-visual semantics as binary labels. Saliency ground truths are provided as a single stream generated from agreed regions of 3 annotators. In this study, we utilize ground truths from the visual-only medium.

4.2 Evaluation Metrics

Emotional attribute estimation with the CER-NET is evaluated using the CCC metric following [8], which is the negative of the L_{CER} loss in (8).

For the video summarization task, following [15], [16], [34], F-score is used as an evaluation metric. For the kth video, let the ground truth binary summarization vector be $\mathbf{s}^k$ and estimated binary summarization vector to be $\hat{\mathbf{s}}^k$ where $\mathbf{s}^k, \hat{\mathbf{s}}^k \in \mathbf{R}^N$. Then precision P^k and recall R^k for the kth video are calculated as

$$P^k = \frac{\mathbf{s}^k \cdot \hat{\mathbf{s}}^k}{\sum_j^N \hat{s}_j^k} \quad \text{and} \quad R^k = \frac{\mathbf{s}^k \cdot \hat{\mathbf{s}}^k}{\sum_j^N s_j^k}, \tag{11}$$

where j runs over the frames. Video level F-score is defined as the harmonic mean of the precision and recall,

$$F1^k = \frac{2P^k R^k}{P^k + R^k}. \tag{12}$$

Then, the final F-score metric $F1$ is defined as the unweighted average of the video level F-score values as

$$F1 = \frac{1}{K}\sum_{k=1}^{K} F1^k, \tag{13}$$

where K is the number of videos in the dataset.

In addition, we used Area-Under Curve (AUC) score as an additional metric which is the integral of the receiver operating characteristic (ROC) curve. ROC curve is calculated by calculating True-Positive and False-Positive rates for each threshold value between 0 and 1, where the threshold value is used to assign predicted score $\hat{z}_j$ to a binary value $\hat{s}_j$.

Since the affective information is learned from a dataset of human-centric videos, the capability of capturing affective frames of the video during the summarization is expected to be better with human-centric videos. By highlighting this fact, we define two new metrics for the evaluation of the affective video summarization. The first metric computes normalized F1 score differences with the baseline over the videos with the highest number of face appearances. To define this metric, let us first define the normalized F1 score difference for the kth video with respect to the SUM-FCN baseline model as

$$\Delta F1^k = \frac{F1^k - F1^k_{SUM-FCN}}{F1^k_{SUM-FCN}}, \tag{14}$$

where $F1^k$ refers to the F1 score of the model in evaluation. Also, assume that all the videos in the dataset are sorted with the highest number of face appearances in descending order and are indexed with k_l for $l = 1, \ldots, K$. Then, the cumulative F1 score difference metric for the *Top-L* human-centric videos is defined as

$$\Delta F1_L = \sum_{l=1}^{L} \Delta F1^{k_l}. \tag{15}$$

Associated with the $\Delta F1_L$, we also compute the F1 score of the *Top-L* human-centric videos and refer to it as $F1_L$.

Human-centric nature of the videos can be associated with the face appearances in the video frames. Motivated by this fact, we set a second metric for evaluation of the affective video summarization as the recall rate of face-appearing frames in the extracted summary. Let $\mathbf{d}_F^k$ be the binary vector representing whether a frame includes a face appearance or not for the kth video. Binary face appearance vectors are extracted by the histogram of oriented gradients (HOG) based face detector [37]. Then, the recall rate of face-appearing frames, let's refer to it as face recall, R is defined as

$$R = \frac{1}{K}\sum_{k=1}^{K} \frac{\hat{\mathbf{s}}^k \cdot (\mathbf{d}_F^k \odot \mathbf{s}^k)}{\sum_j^N s_j^k}, \tag{16}$$

where $(\mathbf{d}_F^k \odot \mathbf{s}^k)$ is the Hadamard product of $\mathbf{d}_F^k$ and $\mathbf{s}^k$.

Following [38], to evaluate the predicted importance scores ($\hat{z}$), we utilize Kendall's rank correlation τ and Spearman's rank correlation coefficient (ρ). Let ζ^k be the reference importance scores, and z^k be the predicted scores for the kth video, then $(\zeta_i^k, \hat{z}_i^k)$ and $(\zeta_j^k, \hat{z}_j^k)$ are concordant pairs if either both $\zeta_i^k > \zeta_j^k$ and $\hat{z}_i^k > \hat{z}_j^k$ holds or both $\zeta_i^k < \zeta_j^k$ and $\hat{z}_i^k < \hat{z}_j^k$, else $(\zeta_i^k, \hat{z}_i^k)$ and $(\zeta_j^k, \hat{z}_j^k)$ are considered as discordant pairs. Based on the concordant and discordant pairs, Kendall's τ is defined as

$$\tau = \frac{1}{N}\sum_{k=1}^{K} \frac{N_c^k - N_d^k}{\sqrt{(N_c^k + N_d^k + T^k)(N_c^k + N_d^k + U^k)}}, \tag{17}$$

where N_c^k, N_d^k, T^k and U^k are the number of concordant pairs, discordant pairs, ties in ζ^k, and ties in $\hat{z}^k$ for the kth video.

Spearman's rank correlation coefficient is defined as

$$\rho = \frac{1}{N}\sum_{k=1}^{K} \frac{cov(R_{\zeta^k}, R_{\hat{z}^k})}{\sigma_{R_{\zeta^k}} \sigma_{R_{\hat{z}^k}}} \tag{18}$$

where R_{ζ^k} and $R_{\hat{z}^k}$ are the ranks, $cov(R_{\zeta^k}, R_{\hat{z}^k})$ is the covariance between R_{ζ^k} and $R_{\hat{z}^k}$, and $\sigma_{R_{\zeta^k}}$ and $\sigma_{R_{\hat{z}^k}}$ are the standard deviations of R_{ζ^k} and $R_{\hat{z}^k}$ respectively for the kth video.

Similar to $F1_L$, we compute the rank correlations of *Top-L* human-centric videos and refer them as τ_L and ρ_L.

We also study statistical differences of a given affective feature dimension, f, across face-appearing and not-appearing frames. For this purpose, the Kullback-Leibler (KL) divergence of f in these two classes is defined as

$$D(P_f || Q_f) = \sum_{x \in \mathcal{X}} P_f(x) \log\left(\frac{P_f(x)}{Q_f(x)}\right), \tag{19}$$

where P_f and Q_f are, respectively, probability distributions of the affective feature dimension f across face-appearing and face-not-appearing frames. Note that the affective feature dimension f is driven from the affective feature set as $f \in \{f^A, f^V, g_1^{AV}, g_2^{AV}, \ldots, g_{2G}^{AV}\}$.

4.3 Implementation Details

4.3.1 CER-NET

The window size T is selected as 20 frames. The *cnn1* and *cnn2* layers have 10 filters, max-pooling layer reduces the temporal dimension from 20 to 5, *gru* layer has 10 cells and *FCN* layer has 1 node. Hence, the dimensionality of $\mathbf{f}^{\mathrm{AV}}$ is 2 and $\mathbf{g}_{\mathrm{AV}}$ is 20 ($G = 10$).

Annotated recordings of the RECOLA dataset are used during the training of the CER-NET. RECOLA recordings are divided as 10% for the test, 10% for the validation, and 80% for the training. We applied the Adam optimizer with a learning rate of 10^{-4} and with a batch size of 256.

4.3.2 Video Summarization Networks

Following [15], [34], TvSum videos are downsampled to 2 fps, and frames are fed into GoogLeNet. For the AVSUM-SCAV, the input feature dimension is set as 1026. The input dimension of the AVSUM-GRU, TA-AVSUM-GRU, and SA-AVSUM-GRU models became 1044 with the GRU embeddings. *conv1* layer has For the TA-AVSUM-GRU, we set the number of heads $H = 4$.

Two convolutional layers in *conv1* and *conv2* have 1044 filters and a kernel size of 3. The three convolution layers in *conv3*, *conv4*, and *conv5* have a filter size of 3 and a number of filters of 1044, 2088, and 2088 respectively. The *conv6*, *conv7*, *conv8*, and *conv9* have a filter size of 1 and the number of filters of 4196, 4196, 2, and 2 respectively. All the max-pool layers have a kernel size of 2 and a stride length of 2. Finally, the *deconv1* block upsamples the input by 4, while the *deconv2* block upsamples the input by 16.

We adopted the leave one-group-out cross-validation technique to compare the performances. Nine videos are selected from the TvSUM dataset for each fold, and the rest are used for training. This procedure results in five folds where none-of-these-test sets intersect; hence each video is included once at the test set. One video is selected for the cross-validation evaluation on the COGNIMUSE, and the rest is used for training, resulting in seven folds. Results of the state-of-art models are also generated following the described cross-validation procedures.

We mimic fixed size cropping in semantic segmentation by uniformly sampling the video frames and using $N = 320$ for TvSUM [15] and $N = 1280$ for COGNIMUSE dataset. The training is held for 50 epochs for both datasets with a batch size of 5 and 2 videos for the TvSUM and COGNIMUSE datasets, respectively. During the training phase, Adam optimizer with the learning rate of 10^{-3} is used. For each training fold, a model achieving the highest F-score ($F1$) and a model achieving the highest face recall (R) are selected for the performance evaluations.

4.4 CER-NET Performance

TABLE 1
CCC Performance of the CER-NET Emotion Recognition Model on the RECOLA Dataset

Model	Activation	Valence
CER-NET	0.40	0.17
AVEC-2016 Video-geometric [39]	0.38	0.61
CER-MTL Facial [8]	0.15	0.06
End2You [40]	0.47	0.58
End-to-end Visual Network [6]	0.36	0.48

Table 1 presents the CCC performance of the CER-NET emotion recognition model driven by the visual modality on the RECOLA dataset. The proposed architecture performs better at estimating the activation than the valence. The end-to-end CER-NET architecture outperforms CER-MTL Facial model [8], AVEC-2016 baseline [39] and End-to-end visual network [6] in estimating activation by achieving CCC of 0.40. The end-to-end visual networks [6], [40] and AVEC-2016 baseline [39] perform strongly for the valence estimation and fairly close to the CER-NET for the activation estimation. Different from CER-NET, [8] and [39] receive visual information as facial activation units and optical flow vectors. Due to the input dimension, CER-NET has more trainable coefficients leading to a more complex structure than the CER-MTL Facial. Also note that End2You [40] and AVEC-2016 baseline [39] results are reported after post-processing of the affect attributes with median filtering, centering, scaling, and time-shifting, which are not used in the CER-NET evaluations.

In comparison with these baseline visual models [6], [8], [39], [40], CER-NET performs competitively in representing affective information on the visual channel.

4.5 Cumulative AVSUM Performance

This section presents performance evaluations of the proposed affective video summarization models in terms of cumulative F-score, face recall, and rank correlation metrics. Then, the video-level performances are investigated to better highlight characteristics of videos that have improved summarization performance with the affective cues.

Tables 2 and 3 present cumulative F-score ($F1$), face recall (R), and AUC score together with the *Top-15* F-score ($F1_{15}$), AUC score, and face recall (R_{15}) performances for the proposed AVSUM and the baseline models, respectively over the TvSUM and COGNIMUSE datasets. Furthermore, Table 3 presents the number of trainable parameters for model complexity on the last column. In each performance score column, the top-two scoring performances are highlighted in bold. Recall that we apply two model selection criteria based on F-score and face recall maximization. Each model selection criterion is observed to favor its related performance metric in the evaluations. That is, maximization of $F1$ (Max $F1$) yields a higher F-score while maximization of R (Max R) yields higher face recall R.

Observing the cumulative $F1$ performances, the AVSUM-GRU model is competitive for both model selection criteria and observed as the best-performing model on the TvSum dataset and the second-best performing model on the COGNIMUSE dataset with the Max $F1$ criterion. On the other hand, the cumulative R score highlights AVSUM-GRU and the temporal attention-based TA-AVSUM-GRU models with the Max $F1$ criterion on both datasets. On the other hand, it

TABLE 2
Cumulative F-Score ($F1$) and Face Recall (R) Together With the *Top-15* F-Score ($F1_{15}$) and Face Recall (R_{15}) Performances of the AVSUM and the State-of-Art Models With the Maximization of $F1$ and R Model Selection Criteria Over the TvSUM Datasets (Top Two Performances for Each Metric are in Bold)

Model	Max $F1$						Max R					
	$F1$	$F1_{15}$	R	R_{15}	AUC	AUC_{15}	$F1$	$F1_{15}$	R	R_{15}	AUC	AUC_{15}
SUM-FCN [15]	57.46	60.00	53.20	**65.52**	74.88	76.43	54.10	57.40	60.62	60.28	73.68	74.86
DR-DSN [19]	56.46	57.54	48.87	60.56	74.50	73.83	**56.46**	57.54	48.87	60.56	**74.50**	73.83
SUM-GAN [20]	56.74	58.14	50.19	61.78	74.60	75.43	**55.51**	58.06	53.07	68.38	73.89	75.55
SUM-GAN-AAE [21]	57.37	58.81	45.33	59.87	**74.98**	75.78	54.73	57.38	50.74	**69.79**	73.44	75.00
AVSUM-GRU	**57.50**	**60.25**	**54.04**	59.14	**74.95**	**76.50**	54.02	57.40	60.77	66.20	73.60	74.86
AVSUM-SCAV	56.64	**60.60**	52.11	63.80	74.43	**76.75**	53.80	57.42	**62.06**	59.64	73.47	73.36
TA-AVSUM-GRU	57.47	59.95	**53.22**	**65.55**	74.88	76.33	53.79	**59.23**	**65.12**	**70.31**	**73.91**	**76.01**
SA-AVSUM-GRU	55.92	59.98	49.25	62.38	74.03	76.35	52.76	**59.90**	58.85	62.55	72.82	**76.27**

highlights AVSUM-SCAV and TA-AVSUM-GRU models with the Max R criterion on the TvSum dataset. The temporal attention-based TA-AVSUM-GRU model especially performs significantly better with the Max R criterion achieving a 65.12% face recall rate. TA-AVSUM-GRU's high achieving R score performance consolidated on the COGNIMUSE dataset. TA-AVSUM-GRU is observed as the best model on R score with the Max $F1$ criterion and the second-best model with the Max R criterion on the COGNIMUSE dataset.

AUC score performances in Table 2 are aligned with the $F1$ performances, and AVSUM-GRU outperforms the rest for AUC and AUC_{15} with the Max $F1$ criterion. On the other hand, the Max R criterion highlights TA-AVSUM-GRU as it achieves the second best AUC and AUC_{15} consolidating its strong performance at the other classification metrics on the TvSUM dataset. In Table 3, AUC score performances on COGNIMUSE highlight TA-AVSUM-GRU, which outperforms other proposed models and SUM-FCN with both optimization criteria. We should note that SUM-GAN-AAE [21] has a strong AUC performance and attains the highest AUC scores with both optimization criteria.

Table 4 depicts the evaluation of the predicted importance scores with Kendall's τ and Spearman's ρ for Max $F1$ and Max R. Similar to Table 2, the top two performing scores are highlighted on each column. Observing Kendall's τ and Spearman's ρ performances for both of the model selection criteria, TA-AVSUM-GRU sustains a place in the top-two performing scores for all columns. This finding aligns with the TA-AVSUM-GRU's high-performing F-score and face recall scores in Table 2. In addition to TA-AVSUM-GRU, SA-AVSUM-GRU has competitive results with the Max $F1$ criterion, and it outperforms the other architectures with the Max R criterion. Within our fusion strategies, TA-AVSUM-GRU is the only module that operates in the latent space and successfully isolates temporally or emotionally salient regions for video summarization.

Affective information is modeled with the continuous emotion recognition task trained on the RECOLA dataset. Since RECOLA is a human-centric dataset, which includes facial videos, we choose to evaluate the video summarization performance for the *Top-L* human-centric videos, where L is set as 15 with the discussion in Section 4.1. Table 2 presents the *Top-15* F-score $F1_{15}$ and face recall R_{15} performances on the TvSUM dataset. While attention-based SA-AVSUM-GRU and TA-AVSUM-GRU models perform best for the $F1_{15}$ score with the Max R criterion, AVSUM-SCAV and AVSUM-GRU models perform best with the Max $F1$ criterion. Overall, the best performance is 60.60% $F1_{15}$ score with the Max $F1$ criterion for the AVSUM-SCAV model. Observing the *Top-15* face recall R_{15} performances, while TA-AVSUM-GRU model is competitive with the baseline SUM-FCN model with the Max $F1$ criterion, it performs significantly superior with the Max R criterion achieving 70.31% face recall R_{15} rate.

TABLE 3
Cumulative F-Score ($F1$) and Face Recall (R) Performances of the AVSUM and the State-of-Art Models With the Maximization of $F1$ and R Model Selection Criteria Over the COGNIMUSE Datasets (Top Two Performances for Each Metric are in Bold) With the Last Column Presenting Number of Trainable Parameters for Model Complexity Comparison

Model	Max $F1$			Max R			Parameters
	$F1$	R	AUC	$F1$	R	AUC	$(\times 1e6)$
SUM-FCN [15]	47.55	55.21	57.55	46.24	57.73	56.38	116
DR-DSN [19]	46.19	51.81	56.90	46.19	51.81	56.25	2
SUM-GAN [20]	46.09	52.72	57.37	45.35	54.59	**56.74**	25
SUM-GAN-AAE [21]	47.45	53.05	**58.51**	45.32	**64.76**	**57.51**	29
AVSUM-GRU	**47.56**	**55.46**	57.31	**46.58**	57.78	56.40	121
AVSUM-SCAV	**47.59**	55.26	57.36	46.40	57.00	56.48	120
TA-AVSUM-GRU	47.54	**58.16**	**57.55**	**46.41**	**59.00**	56.48	121
SA-AVSUM-GRU	47.01	55.22	56.49	46.34	56.80	56.47	121

TABLE 4
Cumulative Kendall's τ and Spearman's Correlation ρ Together With the *Top-15* τ (τ_{15}) and Spearman's Correlation (ρ_{15}) Performances of the AVSUM and the State-of-Art Models With the Maximization of $F1$ and R Model Selection Criteria Over the TvSUM Dataset (Top Two Performances for Each Metric are in Bold)

Model	TvSUM							
	Max $F1$				Max R			
	τ	τ_{15}	ρ	ρ_{15}	τ	τ_{15}	ρ	ρ_{15}
SUM-FCN [15]	**0.012**	0.012	0.016	0.016	0.012	0.013	0.016	0.017
DR-DSN [19]	-0.001	-0.008	-0.001	-0.011	-0.001	-0.008	-0.001	-0.011
SUM-GAN [20]	-0.024	0.010	-0.032	0.012	-0.042	0.010	-0.055	0.012
SUM-GAN-AEE [21]	0.010	**0.047**	0.013	**0.061**	-0.034	0.001	-0.045	0.001
AVSUM-GRU	0.012	0.013	0.015	0.015	0.010	0.014	0.013	0.018
AVSUM-SCAV	0.011	0.011	0.015	0.014	0.013	0.018	0.17	0.023
TA-AVSUM-GRU	**0.013**	**0.014**	**0.017**	**0.018**	**0.015**	**0.019**	**0.020**	**0.024**
SA-AVSUM-GRU	0.012	0.014	**0.016**	0.018	**0.016**	**0.031**	**0.022**	**0.041**

Table 4 presents Kendall's τ_{15} and Spearman's ρ_{15} performances over the *Top-15* human-centric videos on the TvSUM dataset. Similar to $F1_{15}$ with Max R criterion, attention-based SA-AVSUM-GRU and TA-AVSUM-GRU outperform the rest for τ_{15} and ρ_{15} with Max R criterion. These results are aligned with their $F1_{15}$ performance and show that attention models are successful in fusing the affective information. On the other hand, SUM-GAN-AEE [21] outperforms proposed architectures with the Max-$F1$ criterion, while TA-AVSUM-GRU achieves the second-best results. Interestingly, GAN-based models gain significant performance on human-centric videos, which could be due to their ease at generating and discriminating shared human-centric content across the videos.

Table 2 highlights two runner-up models, AVSUM-GRU and TA-AVSUM-GRU, while Table 4 consolidates the strong performance of TA-AVSUM-GRU. AVSUM-GRU model sustains strong F-score performance with the Max $F1$ criterion, especially for human-centric videos targeted with $F1_{15}$ performance. Alternatively, while the temporal attention based TA-AVSUM-GRU model performs strongly for the face recall with the Max R criterion and attains 70.31% face recall R_{15} rate, it also sustains a competitive performance for the F-score and face recall rates with the Max $F1$ criterion.

Model complexities of the proposed networks have a similar number of trainable parameters and indicate a slight 5% increase over the baseline SUM-FCN architecture as presented in Table 3. However, the SUM-FCN architecture is considerably larger than the other state-of-the-art solutions.

The performance results in Tables 2, 3 and 4 show that the TA-AVSUM-GRU model is a strong runner up, indicating a strong integration of temporal attention with affective information for affective video summarization.

4.6 Explainability

We conduct explainability evaluations to understand better the contributions of the affective information and the proposed model architectures for the affective video summarization. First, we present KL divergence analysis for the affective feature dimensions. Then, we investigate video-level summarization performances of the AVSUM models in terms of the cumulative F-score difference metric $\Delta F1_L$ for the *Top-L* human-centric videos.

4.6.1 Affective Feature Evaluations

Fig. 5 depicts the KL divergence (KLD) of the affective features across distributions gathered on the frames with faces

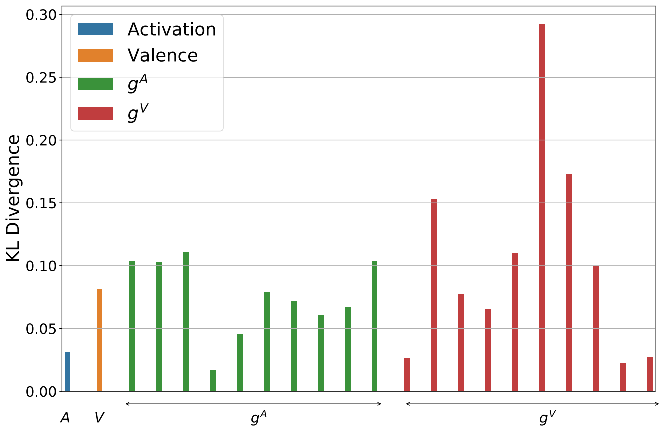


Fig. 5. KL divergence $D(P_f||Q_f)$ of each affective feature dimension.

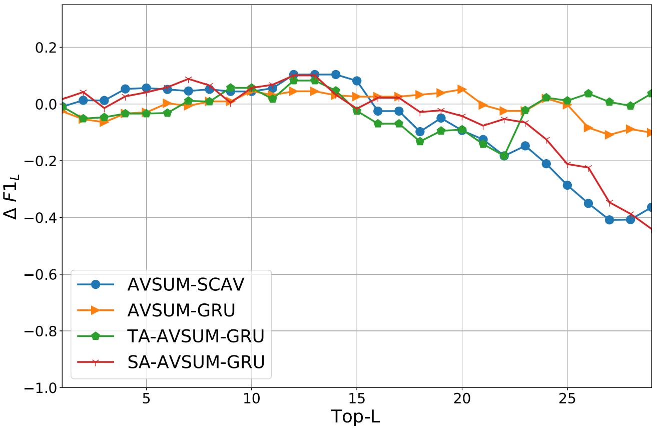
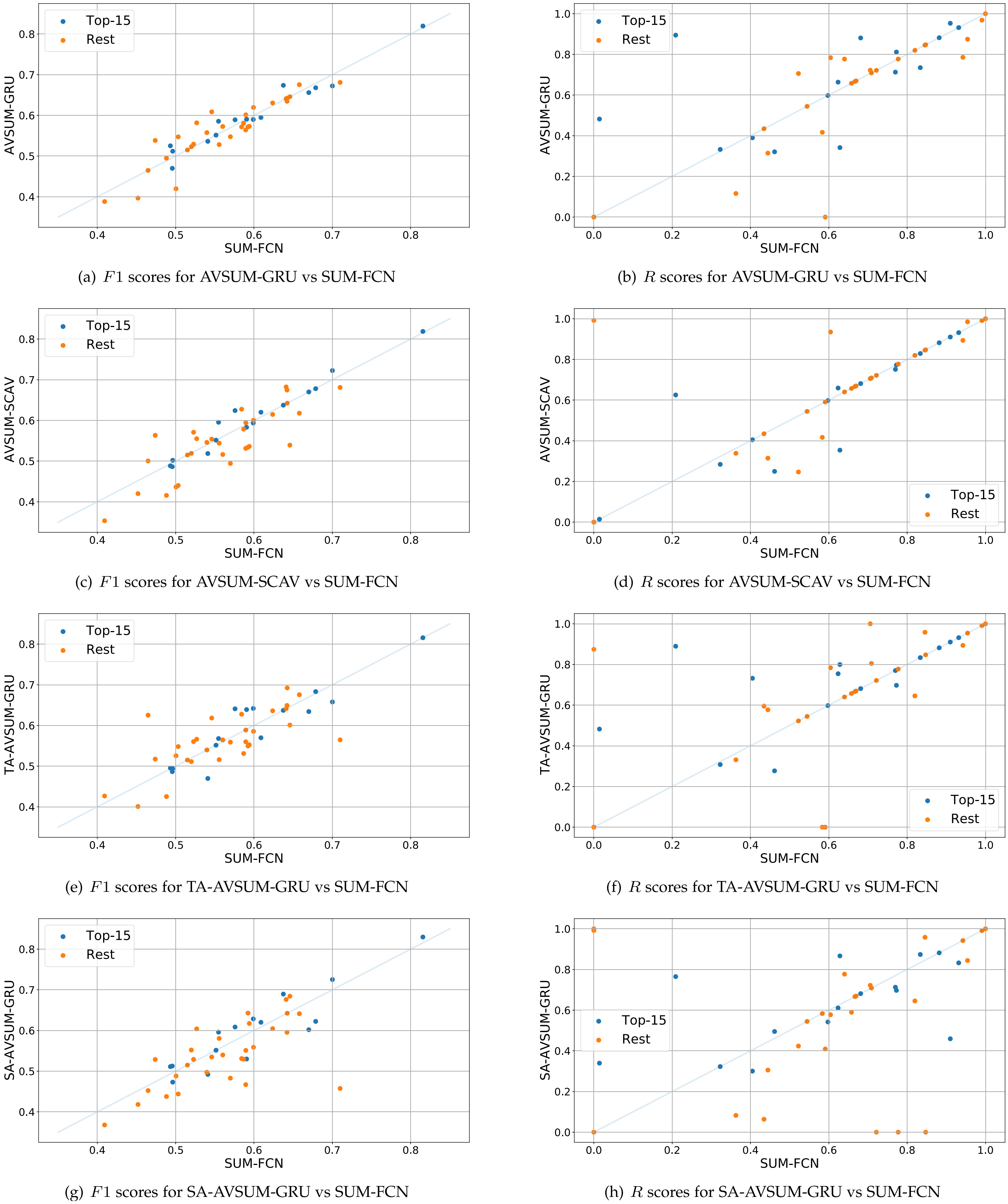


Fig. 6. Performance comparison of the AVSUM models with the $\Delta F1_L$ metric for the *Top-30* human-centric videos at the TvSUM dataset.



(a) $F1$ scores for AVSUM-GRU vs SUM-FCN

(b) R scores for AVSUM-GRU vs SUM-FCN

(c) $F1$ scores for AVSUM-SCAV vs SUM-FCN

(d) R scores for AVSUM-SCAV vs SUM-FCN

(e) $F1$ scores for TA-AVSUM-GRU vs SUM-FCN

(f) R scores for TA-AVSUM-GRU vs SUM-FCN

(g) $F1$ scores for SA-AVSUM-GRU vs SUM-FCN

(h) R scores for SA-AVSUM-GRU vs SUM-FCN

Fig. 7. Video level $F1$ score and R score comparisons of the AVSUM models against the baseline SUM-FCN: $F1$ scores are with Max $F1$ criterion, R scores are with the Max R criterion, and the *Top-15* face including videos are color coded in blue.

and without faces. A higher KL divergence indicates bigger discrimination for the distributions of the feature dimension across the with/without face classes. In Fig. 5, the first two KLD values are for the activation and valence attributes, and the later values are color coded for the dimensions of $\mathbf{g}^A$ and $\mathbf{g}^V$. Note that all the dimensions of the $\mathbf{g}^A$, except the fourth, exhibit higher KLD values than the activation attribute, and at specific dimensions, KLD is almost 4 times higher than the KLD of the activation. A similar trend can be observed for the valence embedding vector $\mathbf{g}^V$, where five dimensions exhibit higher KLD values than the valence attribute, and the largest KLD is extracted for the 6th dimension of the $\mathbf{g}^V$. These higher KLD values for the GRU based feature dimensions can be observed as the discriminative cues for the human-centric videos that also contribute to the performance of the proposed AVSUM architectures.

4.6.2 Video Level Evaluations

Fig. 6 depicts the performance comparison of the AVSUM models with the $\Delta F1_L$ metric for the *Top-30* human-centric videos and yields valuable insights. Note that positive accumulation of $\Delta F1_L$ metric indicates a better performance than the baseline SUM-FCN model. In Fig. 6, the AVSUM-SCAV, TA-AVSUM-GRU, and SA-AVSUM-GRU models perform better than the baseline till the *Top-15* human-centric videos, whereas AVSUM-GRU model sustains a higher performance till the *Top-20* human-centric videos. Similar trends in the AVSUM-SCAV and SA-AVSUM-GRU $\Delta F1_L$ performances can be due to these architectures' common late fusion mechanism. As L increases, we can assume the human-centric characteristic of the videos is getting weaker. Hence, AVSUM-SCAV, SA-AVSUM-GRU, and AVSUM-GRU models perform better for the top human-centric videos, and their performances start degrading as human-centric characteristics get weaker. The performance degradation of AVSUM-SCAV and SA-AVSUM-GRU could be due to the late-fusion employed, as these models are highly conditioned on the affective information. When the saliency of the injected affective information is reduced, for the non-human-centric videos, these features act as noise and reduce the performance. On the other hand, TA-AVSUM-GRU works temporally in the latent domain and can learn temporally salient regions.

We also investigate video-level F-score performances for the proposed AVSUM models. Fig. 7 presents a scatter plot of the video level $F1$ scores on the left column and video level R scores on the right column, where the diagonal in each figure represents similar performances of the compared models. Furthermore, the *Top-15* human-centric videos are color-coded in blue to observe their comparative performances.

Video level F-score performances of the AVSUM-GRU versus SUM-FCN tend to cluster around the diagonal, which makes these two models be most similar in terms of F-score performance. Furthermore, the majority of the *Top-15* human-centric videos are on or above the diagonal, which indicates a stronger performance for the human-centric videos. Unlike F1-score, the face recall performance comparison depicted by Fig. 7b has a scattered behavior providing improvements to some videos while degrading others. However, it is seen that the majority of the *Top-15* human-centric videos are affected positively. This observation is in line with the $F1_{15}$ performance of AVSUM-GRU.

Video level F-score performances of the AVSUM-SCAV versus SUM-FCN have a higher deviation from the diagonal. However, almost all the *Top-15* human-centric videos are on or above the diagonal. This indicates a strong performance improvement for the human-centric videos. In terms of face recall on Fig. 7d, the majority of the videos are accumulated around the diagonal, indicating that AVSUM-SCAV and SUM-FCN are most similar in terms of face recall.

Video level F-score performances of the TA-AVSUM-GRU and SA-AVSUM-GRU models also have a high deviation from the diagonal. However, like AVSUM-SCAV majority of the *Top-15* human-centric videos are on or above the diagonal for both. Similar to F-score, face recall comparisons of TA-AVSUM-GRU and SA-AVSUM-GRU have a scattered behavior depicted by Figs. 7f, and 7h. However, different from SA-AVSUM-GRU, scattered points accumulated on the positive side for TA-AVSUM-GRU, stating a major performance improvement which is inline with its best achieving R_{15} performance for the Max R criterion.

5 Conclusion

In this study, we proposed a new affective information-enriched end-to-end video summarization framework for human-centric videos. As a first step, we modeled affective information in terms of AV attributes and GRU embeddings, which were extracted from the CER models. The CER-NET, a CER model achieving state-of-the-art CCC performance, was introduced. We explored the use of affective information with the proposed AVSUM-SCAV and AVSUM-GRU fusion models and attention mechanisms based TA-AVSUM-GRU and SA-AVSUM-GRU models. Experimental investigations of the proposed models were conducted on the RECOLA, TvSum and COGNIMUSE datasets.

We observed that with the fusion of affective information, F-score performance of the video summarization on the human-centric videos could be improved. To further analyze the effect of injected features, we defined a face recall (R) metric and showed that AVSUM-GRU and TA-AVSUM-GRU models outperform state-of-art models in face recall on both TvSUM and COGNIMUSE datasets. While AVSUM-GRU model has strong performance improvements for the human-centric videos on $F1$ score and face recall R, TA-AVSUM-GRU has significant performance improvements on face recall and is competitive on $F1$ as well. These two models are the most competitive against the baselines with the Max $F1$ criterion. On the other hand, we observed that attention-enhanced mechanisms exhibit strong performance gains with the Max R criterion for human-centric videos. We should also note that affective GRU embedding features exhibit higher KLD across with-face and without-face frames. Our evaluations with the predicted importance scores further verified the strong performance of the TA-AVSUM-GRU model for the human-centric video summarization task. TA-AVSUM-GRU introduced more than 15% improvement in rank correlation scores against SUM-FCN in both optimization criteria for human-centric videos while sustaining to stay in the top-two performing models with the rank correlation metrics.

Comparing the proposed AVSUM models, temporal attention-based TA-AVSUM-GRU performs competitively with the Max $F1$ criterion and attains substantial improvement with the Max R criterion. Evaluations of the predicted importance scores further stress TA-AVSUM-GRU, as it introduces significant improvements regardless of the model selection criterion. The proposed AVSUM models integrate affective information into the summarization architectures and significantly improve video summarization for human-centric videos.

Our experimental findings suggest that affective states contain highly relevant information for human-centric video summarization and help to detect salient regions. Furthermore, we demonstrate the benefits of using new evaluation metrics, such as face recall R and *Top-L* performances, for human-centric video summarization. The compilation of affective human-centric video datasets for video summarization tasks remains a critical and valuable future study. Investigating multi-modal architectures for human-centric affective video summarization would be another valuable extension of the current work.

References

[1] P. Vicol, M. Tapaswi, L. Castrejon, and S. Fidler, "MovieGraphs: Towards understanding human-centric situations from videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8581–8590.

[2] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," 2021, *arXiv:2101.06072v1*.

[3] M. Sun, A. Farhadi, B. Taskar, and S. Seitz, "Summarizing unconstrained videos using salient montages," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2256–2269, Nov. 2017.

[4] A. Zlatintsi et al., "COGNIMUSE: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, 2017, Art. no. 54.

[5] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

[6] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[7] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Inf. Fusion*, vol. 68, pp. 46–53, Apr. 2021.

[8] B. Köprü and E. Erzin, "Multimodal continuous emotion recognition using deep multi-task learning with correlation loss," 2020, *arXiv:2011.00876*.

[9] U. Bhattacharya, G. Wu, S. Petrangeli, V. Swaminathan, and D. Manocha, "HighlightMe: Detecting highlights from human-centric videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8137–8147.

[10] H. Joho, J. Staiano, N. Sebe, and J. M. Jose, "Looking at the viewer: Analysing facial activity to detect personal highlights of multimedia contents," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 505–523, Jan. 2011.

[11] A. G. Money and H. Agius, "Analysing user physiological responses for affective video summarisation," *Displays*, vol. 30, no. 2, pp. 59–70, Apr. 2009.

[12] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[13] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. IEEE 10th Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

[14] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5179–5187.

[15] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 358–374.

[16] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder–decoder networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1709–1717, Jun. 2020.

[17] Y. Yuan, H. Li, and Q. Wang, "Spatiotemporal modeling for video summarization using convolutional recurrent neural network," *IEEE Access*, vol. 7, pp. 64 676–64 685, 2019.

[18] P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, and L. Shao, "Exploring global diverse attention via pairwise temporal relation for video summarization," *Pattern Recognit.*, vol. 111, 2021, Art. no. 107677.

[19] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, Art. no. 929.

[20] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proc. IEEE 30th Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2982–2991.

[21] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Unsupervised video summarization via attention-driven adversarial learning," in *Proc. Int. Conf. Multimedia Model.*, 2020, pp. 492–504.

[22] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 39–54.

[23] Y. Jung, D. Cho, S. Woo, and I. S. Kweon, "Global-and-local relative position embedding for unsupervised video summarization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 167–183.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[25] P. Ekman and R. J. Davidson, *The Nature of Emotion: Fundamental Questions*. New York, NY, USA: Oxford Univ. Press, 1994.

[26] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, 1977.

[27] H. Schlosberg, "Three dimensions of emotion," *Psychol. Rev.*, vol. 61, no. 2, pp. 81–88, 1954.

[28] M. Schmitt, N. Cummins, and B. W. Schuller, "Continuous emotion recognition in speech – do we need recurrence?," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2808–2812.

[29] B. Xu, Y. Fu, Y. G. Jiang, B. Li, and L. Sigal, "Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 255–270, Second Quarter 2018.

[30] G. Tu, Y. Fu, B. Li, J. Gao, Y. G. Jiang, and X. Xue, "A multi-task neural approach for emotion attribution, classification, and summarization," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 148–159, Jan. 2020.

[31] C. Wang, J. Zhang, W. Jiang, and S. Wang, "A deep multimodal model for predicting affective responses evoked by movies based on shot segmentation," *Secur. Commun. Netw.*, vol. 2021, pp. 1–12, 2021.

[32] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[33] P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, and A. Potamianos, "Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 4361–4365.

[34] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 766–782.

[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[37] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, no. 60, pp. 1755–1758, 2009.

[38] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkila, "Rethinking the evaluation of video summaries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7596–7604.

[39] M. Valstar et al., "AVEC 2016 - depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 3–10. [Online]. Available: http://arxiv.org/abs/1605.01600

[40] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, "End2You–The imperial toolkit for multimodal profiling by end-to-end learning," 2018, *arXiv:1802.01115*.

Berkay Köprü received the BSc and MSc degrees from Bilkent University, Ankara, Turkey, in 2014 and Technical University of Munich, Munich, Germany, in 2017, respectively. He is currently working toward the PhD degree with Koc University, Istanbul, Turkey. His research interest includes human-computer interaction, computer vision, and affective computing.

Engin Erzin (Senior Member, IEEE) received the BSc, MSc, and PhD degrees from Bilkent University, Ankara, Turkey, in 1990, 1992, and 1995, respectively, all in electrical engineering. During 1995-1996, he was a postdoctoral fellow with the Signal Compression Laboratory, University of California, Santa Barbara. He joined Lucent Technologies, in September 1996, and he was with the Consumer Products Group for one year as a member of technical staff of the Global Wireless Products Group. From 1997 to 2001, he was with the Speech and Audio Technology Group of the Network Wireless Systems. Since January 2001, he has been with the Electrical & Electronics Engineering and Computer Engineering Departments of Koc University, Istanbul, Turkey. He is currently a member of the IEEE Speech and Language Processing Technical Committee and associate editor for the *IEEE Transactions on Multimedia*, having previously served as associate editor of the *IEEE Transactions on Audio, Speech & Language Processing* (2010-2014). His research interests include speech-audio-visual signal processing, affective computing, human-computer interaction, and machine learning.