TLDW: Extreme Multimodal Summarization of News Videos

Peggy Tang, Graduate Student Member, IEEE, Kun Hu, Member, IEEE, Lei Zhang, Fellow, IEEE, Jiebo Luo, Fellow, IEEE, and Zhiyong Wang, Member, IEEE

Abstract—Multimodal summarisation with multimodal output is drawing increasing attention due to the rapid growth of multimedia data. While several methods have been proposed to summarise visual-text contents, their multimodal outputs are not succinct enough at an extreme level to address the information overload issue. To the end of extreme multimodal summarisation, we introduce a new task, Extreme Multimodal Summarisation with Multimodal Output (XMMSO) for the scenario of TLDW: Too Long; Didn’t Watch, akin to TL;DR. XMMSO aims to summarise a video-document pair into a summary with an extremely short length, which consists of one cover frame as the visual summary and one sentence as the textual summary. We propose a novel unsupervised Hierarchical Optimal Transport Network (HOTNet) consisting of three components: hierarchical multimodal encoder, hierarchical multimodal fusion decoder, and optimal transport solver. Our method is trained, without using reference summaries, by optimising the visual and textual coverage from the perspectives of the distance between the semantic distributions under optimal transport plans. To facilitate the study on this task, we constructed a large-scale dataset, XMMSO-News, by harvesting 4,891 video-document pairs. The experimental results show that our method achieves promising performance in terms of ROUGE and IoU metrics. Our dataset and source code will be publicly available in GitHub.

Index Terms—Extreme multimodal summarization, multimedia, unsupervised learning.

I. INTRODUCTION

SUMMARISATION aims to condense a given piece of information into a short and succinct summary that best covers its semantics with the least redundancy. This helps users quickly browse and understand long content by focusing on the most important ideas [27]. Summarisation on a single modality, such as video summarisation [25], [46] and text summarisation [16], [20], [30], [39], has been actively studied for decades.

Video summarisation aims to summarise a video into keyframes or key segments [11], [14], [19], [23], [24], [26], [44], [46], [47] that provide a compact yet informative representation of the video. The majority of existing methods focus on modelling the temporal dependency and spatio structure among frames [1]. To address information overload issues, extreme video summarisation has been proposed as a sub-task of video summarisation [2], [10], [38], [41], [48], which aims to summarise a video into a cover frame, preview frame, or thumbnail frame. It involves high source compression and allows users to quickly discern the essence of a video and decide whether it is worth watching or not.

Text summarisation aims to condense a given document into a short and succinct summary that best covers the document’s semantics. The majority of existing methods are either extractive or abstractive. Extractive methods [20], [32], [51], [53] select salient sentences from a document to form its summary. Abstractive methods [16], [34], [39], [49] involve natural language generation to generate a summary for a given document. To further condense the text and address information overload issues, extreme text summarisation has been proposed as a sub-task of text summarisation. Extreme text summarisation [6], [22], [31], [42] aims to summarise a document into a one-sentence summary. It helps users quickly comprehend the main information of a document.

While unimodal summarisation has been investigated for decades, with the rapid growth of multimedia data [57], there is an emerging interest on Multimodal Summarisation with Multimodal Output (MSMO) [17], [55], [56]. MSMO aims to summarise a video or a set of images and its companion document into a visual-textual summary which exploits the complementary nature of image and text, since images help users to grasp events while texts provide more details related to the events. This help users to better obtain a more informative and richer understanding of events. Moreover, this can help to make the piece of information more accessible to users with various individual needs. For example, users with a reading disorder may refer more to the visual summary, and users with vision impairments may find the textual summary more accessible by using a screen reader. However, most of the existing MSMO methods are designed for short visual inputs, such as short videos and multiple images [17], [55], [56], without considering the summary length. According to the statistics of YouTube,¹ one of the most popular online video-sharing platforms, in 2023, more than 500 hours of video contents are uploaded every minute; and more than a billion logged-in users visit YouTube each month and watch over a billion

¹https://blog.youtube/press/
hours of video every day. In particular, YouTube channels of news organisations are a significant part of the content, which often have millions of subscribers and views. For example, as of May 2023, the British Broadcasting Corporation News YouTube channel had over 14 million subscribers and had accumulated over 4 billion views. Given the increasing pace of producing multimedia data and the subsequent challenge in keeping up with the explosive growth of such rich content, these existing methods may be sub-optimal to address the imminent issue of information overload of multimedia data.

In this paper, we introduce a new task, eXtreme Multimodal Summarisation with Multiple Output (XMSMO), for the scenario TLDW which stands for Too Long; Didn’t Watch. As shown in Figure 1, XMSMO aims to summarise a pair of a video and its corresponding document into a multimodal summary with an extremely short length. That is, an extreme multimodal summary consists of one cover frame as the visual summary and one sentence as the textual summary.

Specifically, the hierarchical visual encoder formulates the representations of a video from three levels including frame-level, scene-level and video-level. The hierarchical textual encoder formulates the representations of a document from three-levels as well: word-level, sentence-level and document-level. Then, the hierarchical decoder formulates the cross-modal representations in a local-global manner and evaluates candidate cover frames and candidate words, which are used to form a visual summary and a compressive textual summary, respectively. Note that a compressive textual summary offers a balance between the conciseness issue of extractive summarisation and the factual hallucination issue of abstractive summarisation. Finally, our optimal transport-based unsupervised training strategy is devised to mimic human judgment on the quality of an extreme multimodal summary in terms of visual and textual coverage. The coverage is measured by a Wasserstein distance with an optimal transport plan measuring the distance between the semantic distributions of the summary and the original content. Wasserstein distance is used to take the semantic relationship between tokens into account and bring interpretability into the summarisation process. In addition, textual fluency and cross-modal similarity are further considered, which can be important to obtain a high-quality multimodal summary.

Additionally, to facilitate the study on this new task XMSMO and evaluate our proposed HOT-Net, we constructed the first dataset of such kind, namely XMSMO-News, by harvesting 4,891 video-document pairs as input and cover frame-title pairs as multimodal summary output from the British Broadcasting Corporation (BBC) News YouTube channel from Year 2013 to Year 2021.

In summary, the key contributions of this paper are:

- We introduce a new task, eXtreme Multimodal Summarisation with Multiple Output (XMSMO) as TLDW, which stands for Too Long; Didn’t Watch. It aims to summarise a video-document pair into an extreme multimodal summary (i.e., one cover frame as the visual summary and one sentence as the textual summary).
- We propose a novel unsupervised Hierarchical Optimal Transport Network (HOT-Net). The hierarchical encoding and decoding are conducted across both the visual and textual modalities, which improve abstraction at multiple levels and multiple modalities and integration across these levels and modalities. Optimal transport solvers are introduced to guide the summaries to maximise their semantic coverage.
- We devise a new unsupervised training strategy that mimics the human judgment of a multimodal summary’s quality by minimising the quartet loss of visual coverage, textual coverage, textual fluency, and cross-modal consistency.
- We constructed a new large-scale dataset, XMSMO-News, for the research community to facilitate research in this new direction. Experimental results on this dataset demonstrate that our method outperforms other baselines in terms of ROUGE and IoU metrics.

II. RELATED WORK

In this section, we first review existing deep learning-based extreme unimodal summarisation methods in two categories, video-based and text-based, since they are closely related to our study. We then review existing multimodal summarisation with multimodal output methods which share similar input and output modalities with our study. We further review existing video captioning methods, since video captioning and multimodal summarisation can be seen as relevant tasks which involve the generation of natural language descriptions for videos.
A. Extreme Video Summarisation

Extreme video summarisation methods can be conceptualized as a frame ranking task, which scores the frames in a video as the output. A deep learning method based on a CNN-based autoencoder architecture was first proposed [10], in which the training is unsupervised and the goal is to minimise a reconstruction loss considering the representativeness and aesthetic quality of the selected frames. The performance of different CNNs was compared and the ResNet-50 CNN outperformed the other CNNs, as studied by [35]. The scoring was improved by [38] by incorporating additional CNNs to consider the quality of faces. It utilised a Siamese CNN architecture, which was optimized by a piece-wise ranking loss using pairs of frames. [2] proposed a generative adversarial network that introduced a reinforcement learning scheme by rewarding the representativeness and aesthetic quality. Note that most of these methods encode a video as a sequence of frames directly, whilst the hierarchical semantic structure of a video has not been adequately explored.

B. Extreme Text Summarisation

The extreme text summarisation task was first explored by [31] who formulated the task as a sequence-to-sequence learning problem, where the input was a source document and the output was an extreme summary. A supervised encoder-decoder framework was studied and a topic model was incorporated as an additional input to involve the document-level semantic information and guide the summary to be consistent with the document theme. Reference [6] introduced multi-task learning and incorporated the title generation as a scaffold task to improve the learning ability regarding the salient information in the document. These methods relied on integrating the knowledge from pre-trained embedding models to generate abstractive summaries. As a result, these generative models are highly prone to external hallucination and it is possible to generate contents unfaithful to the original document, which was shown by [28].

C. Multimodal Summarisation With Multimodal Output

Multimodal summarisation with multimodal output task was first studied in [55], which took a document and an image set as the input. A supervised attention-based encoder-decoder framework was devised. For encoding, a textual encoder and a visual encoder formulate the document and visual representations, respectively. For decoding, a textual decoder generates a textual summary, and a visual decoder selects the most representative image as a visual summary. Additionally, a multimodal attention layer was incorporated to fuse the textual and visual context information. To alleviate the modality-bias issue, a multitask learning was applied to jointly consider the two MSMO subtasks: summary generation and text-image relation recognition [56]. A hierarchical intra- and inter-modality correlation between the image and text inputs was studied to enhance the multimodal context representation [50]. Reference [17] extended visual inputs to short videos, and introduced self-attentions to improve the multimodal context representation. Nonetheless, most of these methods encode the video and document inputs directly without considering their semantic hierarchical structure. Moreover, these existing methods have been mainly studied in a supervised manner. To the best of our knowledge, our work is the first unsupervised method for MSMO.

D. Video Captioning

The task of video captioning is to generate cohesive and informative text that describes a video. It is usually formulated as a sequence-to-sequence learning problem, where the input is the frames of the video and the output is a number of sentences. Existing methods often follow a supervised encoder-decoder framework [5], [33], which utilises CNNs to embed the video frames, and RNN-based networks to encode the sequence of frames and then decode and generate the output text. The efficiency of the network was improved by [9], which explored the use of informative frames for video captioning. The accuracy of generated captions was
improved by [12], which explored the joint optimisation of the two sub-tasks of video captioning: syntagm representation learning and visual cues translation. To address the issue of long-term dependencies of a RNN-based network, CNN-based [8] and transformer-based [54] encoder-decoder frameworks were introduced. To improve the fluency and cohesion of generated captions, [52] explored a multi-modal dependency tree construction method that better incorporates information from multiple modalities.

III. METHODOLOGY

As shown in Figure 2, our proposed eXtreme Multimodal Summarisation method, namely unsupervised Hierarchical Optimal Transport Network (HOT-Net), consists of three components, hierarchical multimodal encoder including a hierarchical visual encoder and a hierarchical textual encoder, hierarchical multimodal (fusion-based) decoder and optimal transport solver. Specifically, the hierarchical visual encoder formulates frame-level, scene-level and video-level representations of a video V. The hierarchical textual encoder formulates word-level, sentence-level and document-level representations of a document D. Then, the hierarchical visual decoder selects an optimal frame \( f^* \) as an extreme visual summary, and the hierarchical textual decoder produces an extreme textual summary \( s^* \) based on the cross-modal guidance. Finally, the optimal transport solver conducts unsupervised learning to optimise the encoders and the decoders in pursuit of the best semantic coverage of the obtained summaries.

A. Hierarchical Multimodal Encoder

1) Hierarchical Visual Encoder: Given an input video \( V \), it can be represented as a sequence of \( T \) frames, i.e. \( V = \{ x^f_{i|j} \}_{i=1}^{T} \). By grouping the consecutive frames with similar semantics, this input video \( V \) can be segmented into a sequence of \( T' \) scenes, i.e. \( V = \{ x^s_{j|j} \}_{j=1}^{T'} \), where \( x^s_{j|j} \) consists of the video frames from the \( i_{j_0} \)-th to the \( i_{j_1} \)-th frame, where \( j_0 \) indicates the start index of the frame and \( j_1 \) indicates the end index of the frame for the \( j \)-th frame in the video. The hierarchical visual encoder learns the scene-level and video-level representations based on \( x^f_{i|j} \) and \( x^s_{j} \), respectively.

To characterize a video frame \( x^f_{i|j} \), a pre-trained neural network can be introduced. The CLIP model [37] is adopted in this study since it is the state-of-the-art multi-modal embedding model. For the sake of convenience, we use the symbol \( x^f_{i|j} \) to represent this pre-trained feature of the \( i \)-th frame. To further model the scene-level features, a pooling method is introduced, which is denoted as a function \( g^s \). In detail, for the \( j \)-th scene, its representation \( x^s_j \) can be obtained by observing its associated frame-level features \( x^f_{i|j} \), \( i = i_{j_0}, \ldots, i_{j_1} \) as:

\[
x^s_j = g^s(x^f_{i_{j_0}}, \ldots, x^f_{i_{j_1}}).
\]  

(1)

Particularly, a generalized pooling operator (GPO) [7] is adopted as the pooling method in this study, since it is shown to be an effective and efficient pooling strategy for different features. With the scene-level features, a pooled global (i.e., video-level) representation can be derived as:

\[
x^v = g^v(x^s_{1}, \ldots, x^s_{T'}),
\]  

(2)

where \( g^v \) is a video-level pooling function based on a GPO operator.

2) Hierarchical Textual Encoder: An input document \( D \) can be viewed as a sequence consisting of \( U \) words as \( x^w_{m|n} = \{ x^w_{m|n} \}_{n=1}^{U} \}, or a sequence of \( U' \) sentences as \( x^s_{n|m} = \{ x^s_{n|m} \}_{m=1}^{U'} \}. The \( n \)-th sentence consists of consecutive words in \( D \) from the \( m_{n_0} \)-th to the \( m_{n_1} \)-th word. Similar to the visual encoder, a hierarchical textual encoder is introduced to learn the sentence-level and the document-level representation.

A pre-trained CLIP model is introduced to formulate the word-level features, which is denoted as \( x^w_{m|n} \) for the \( m \)-th word. Next, a pooling mechanism \( g^s \) is adopted to formulate the sentence-level features. In detail, the \( n \)-th sentence-level features can be computed as:

\[
x^s_n = g^s(x^w_{m_{n_0}|n}, \ldots, x^w_{m_{n_1}|n}).
\]  

(3)

Finally, the global representation of the document \( D \) can be derived based on the sentence-level features:

\[
x^d = g^d(x^s_1, \ldots, x^s_{U'}).
\]  

(4)

where \( g^d \) is a document-level pooling function based on GPO.

B. Hierarchical Multimodal Fusion

To attend and fuse the representations from the visual and textual modalities, we adopt a graph-based attention mechanism (GAT) [43]. This multimodal fusion formulation helps easily extend the attention layer to future additional modalities, such as an audio modality. Each modality feature can be treated as a vertex feature of a graph. The relationships between modalities are formulated by graph convolution to attend over the other modalities, which then updates the representations of each modality. Particularly, a hierarchical local, which focuses between scene and sentence levels, and global, which focuses between video and document levels, observations are introduced by a graph fusion strategy.

For local multimodal fusion, the representations of the scenes \( x^s = \{ x^s_1, \ldots, x^s_{T'} \} \) and sentences \( x^s = \{ x^s_1, \ldots, x^s_{U'} \} \) are fed into graph fusion modules \( f^s_{local} \) and \( f^s_{local} \). The resulted representation, which can be viewed as an information exchange between modalities, are fed into an average pooling operator \( g^s \) to obtain the local multimodal context representations \( x^s \) and \( x^s \):

\[
x^s = g^s(f^s_{local}(x^s_1, \ldots, x^s_{T'}), \ldots, f^s_{local}(x^s_1, \ldots, x^s_{T'})),
\]  

(5)

\[
x^s = g^s(f^s_{local}(x^s_1, \ldots, x^s_{U'}), \ldots, f^s_{local}(x^s_1, \ldots, x^s_{U'})).
\]  

(6)

For global multimodal fusion, the global representations of the document \( x^d \) and video \( x^v \) are fed into a graph fusion module \( f^s_{global} \):

\[
\hat{x} = g^s(f^s_{global}(x^v, x^d)).
\]  

(7)
C. Hierarchical Multimodal Decoder

1) Visual Decoder: Our visual decoder consists of three stages: 1) scene-guided frame decoding, 2) video-guided frame decoding, and 3) cross-modality-guided frame decoding. It aims to evaluate the probability of a particular frame being a cover frame.

To produce a scene-aware decoding outcome of evaluating each frame, a scene-guided visual decoder $h_{\text{scene}}^{n}$ derives a latent decoding $y_{n}^{\text{scene}}$ for frames from $i_{j0}$ to $i_{j1}$, $j = 1, \ldots, T'$, as follows:

$$y_{j}^{\text{scene}} = \{y_{j0}^{\text{scene-frame}}, \ldots, y_{j1}^{\text{scene-frame}}\} = h_{\text{scene}}^{n}(x_{j0}^{\text{frame}}, \ldots, x_{j1}^{\text{frame}}|\tilde{x}_{n}^{\text{scene}}),$$  

(8)

where $h_{\text{scene}}^{n}$ is a bi-directional GRU [4] and $\tilde{x}_{n}^{\text{scene}}$ is a multimodal scene guidance, which can be viewed as a prior knowledge. Next, to produce a video-guided frame decoding outcome, we have:

$$y^{\text{video}} = \{y_{1}^{\text{video-frame}}, \ldots, y_{T}^{\text{video-frame}}\} = h^{\text{video}}(x_{1}^{\text{frame}}, \ldots, x_{T}^{\text{frame}}|\tilde{x}^{\text{video}}),$$  

(9)

where $h^{\text{video}}$ is a bi-directional GRU and $\tilde{x}^{\text{video}}$ is a unimodal video guidance as a prior knowledge. Finally, to produce a global multimodal context-aware decoding, we adopt a Bi-GRU decoder $\hat{h}$ with the guidance of the cross-modal embedding $\hat{x}$:

$$\hat{y}^{\text{video}} = \hat{y}_{1}^{\text{video-frame}}, \ldots, \hat{y}_{T}^{\text{video-frame}} = h^{\text{video}}(x_{1}^{\text{frame}}, \ldots, x_{T}^{\text{frame}}|\hat{x}).$$  

(10)

To this end, the optimal frame $f^{*}$ is obtained with a frame-wise linear layer activated with a softmax function:

$$f^{*} = \arg \max_{f} (\text{Linear}(\hat{y}^{\text{video}})).$$  

(11)

2) Textual Decoder: Similar to the visual decoder, the textual decoder also consists of three stages: 1) sentence-guided word decoding, 2) document-guided word decoding, and 3) cross-modality-guided word decoding. It aims to evaluate the probability of a word being selected in a compressive summary.

To produce a sentence-aware decoding outcome, a sentence decoder $h_{\text{sentence}}^{n}$ derives a latent decoding $y_{n}^{\text{sentence}}$ for words from $m_{n0}$ to $m_{n1}$, $n = 1, \ldots, U'$, where $n_{0}$ indicates the start index of the word and $n_{1}$ indicates the end index of the word for the $n$-th sentence in the document, as follows:

$$y_{n}^{\text{sentence}} = \{y_{m_{n0}}^{\text{sentence-word}}, \ldots, y_{m_{n1}}^{\text{sentence-word}}\} = h_{\text{sentence}}^{n}(x_{m_{n0}}^{\text{word}}, \ldots, x_{m_{n1}}^{\text{word}}|\hat{x}_{n}^{\text{sentence}}),$$  

(12)

where $h_{\text{sentence}}^{n}$ is a bi-directional GRU and $\hat{x}_{n}^{\text{sentence}}$ is used as a prior knowledge for the multimodal sentence guidance. Then, to produce a document-level textual decoding, we have:

$$y_{\text{document}} = \{y_{1}^{\text{document-word}}, \ldots, y_{U}^{\text{document-word}}\} = h_{\text{document}}^{n}(x_{1}^{\text{word}}, \ldots, x_{U}^{\text{word}}|\hat{x}^{\text{document}}),$$  

(13)

where $h_{\text{document}}$ is a bi-directional GRU and $\hat{x}^{\text{document}}$ is a unimodal document guidance. Finally, to produce a global cross-modal context-aware decoding for each word, a Bi-GRU decoder $\hat{h}^{\text{document}}$ is adopted with the guidance of the global multimodal embedding $\hat{x}$:

$$\hat{y}_{\text{document}} = \hat{y}_{1}^{\text{document-word}}, \ldots, \hat{y}_{U}^{\text{document-word}} = h_{\text{document}}^{n}(y_{1}^{\text{document-word}}, \ldots, y_{U}^{\text{document-word}}|\hat{x}).$$  

(14)

As a result, the optimal compressive summary $s^{*}$ with length $k$ is obtained by:

$$s^{*} = \text{topk}(\text{Linear}(\hat{y}_{\text{document}})).$$  

(15)

Note that the selected $k$ words are ranked in line with their scores obtained from the linear layer with a softmax activation. Thus, the sentence $s^{*}$ can be constructed with these words and their orders.

D. Optimal Transport-Guided Semantic Coverage

Our method is trained without reference summaries by mimicking the human judgment on the quality of a multimodal summary, which minimises a quartet loss of optimal transport-guided semantic coverage between the document and one-sentence summary and between the video and the cover frame, textual fluency, and cross-modal similarity.

Fig. 3. Optimal transport solver of HOT-Net for our unsupervised training strategy. Our method is trained without reference summaries by mimicking the human judgment on the quality of a multimodal summary, which minimises a quartet loss of optimal transport-guided semantic coverage between the document and one-sentence summary and between the video and the cover frame, textual fluency, and cross-modal similarity.

1) Optimal Transport-Guided Document Coverage: Intuitively, a high-quality summary is supposed to be close to the original document regarding their semantic distributions. Wasserstein distance is used to take the semantic relationship between tokens into account and bring interpretability into the summarisation process. We measure the Wasserstein distance [15] $L_{\text{document}}$ between the document $D$ and the selected sentence $s^{*}$. It is the minimal cost required to transport the semantics from $s^{*}$ to $D$, measuring the semantic coverage of $s^{*}$ on $D$.

Given a dictionary, the number of the $\alpha$-th token (i.e., a word in a dictionary) occurred in $D$ can be counted as $P_{D}(\alpha)$. As a result, the semantic distribution $TF_{D}$ of the document $D$ can be defined with the normalized term frequency of each token. In detail, for the $\alpha$-th element of $TF_{D}$, we have:

$$TF_{D}(\alpha) = \frac{P_{D}(\alpha)}{\sum_{\alpha'} P_{D}(\alpha')}.$$  

(16)

The semantic distribution $TF_{s^{*}}$ of the selected sentence $s^{*}$ can be derived in a similar manner. The normalized term frequency...
of the $\alpha$-th token in $s^*$ is:

$$\text{TF}_s^*(\alpha) = \frac{P_s^*(\alpha)}{\sum_{\alpha'} P_{s}^*(\alpha')}.$$  \hspace{1cm} (17)

Note that $\text{TF}_D$ and $\text{TF}_s^*$ have an equal total token quantities of 1 and can be completely transported from one to the other mathematically.

A transportation cost matrix $C = (c_{aa'})$ is introduced to measure the semantic similarity between the tokens. Given a pre-trained tokenisation and token embedding model, define $u_\alpha$ to represent the feature embedding of the $\alpha$-th token. The transport cost $c_{aa'}$ from the $\alpha$-th token to the $\alpha'$-th one is computed based on the cosine similarity:

$$c_{aa'} = 1 - \frac{\langle u_\alpha, u_{a'} \rangle}{\|u_\alpha\|_2 \|u_{a'}\|_2}. \hspace{1cm} (18)$$

Note that the method to obtain token representations $u_\alpha$ follows the same method that we formulate for word representations $x_i$ by a pre-trained model.

Then, an optimal transport plan matrix $T^*(D, s^*) = (t_{aa'}^*(D, s^*))$ in pursuit of minimizing the transportation cost can be obtained by solving the following optimization problem:

$$T^*(D, s^*) = \arg\min_{T(D, s^*)} \sum_{a, a'} t_{aa'}(D, s^*) c_{aa'},$$

$$\text{s.t.} \sum_{a'} t_{aa'}(D, s^*) = \text{TF}_D(\alpha),$$

$$\sum_{a=1}^{L} t_{aa'}(D, s^*) = \text{TF}_s^*(\alpha'),$$

$$t_{aa'}(D, s^*) \geq 0,$$

$$\forall \alpha, \alpha' \hspace{1cm} (19)$$

To this end, the Wasserstein distance can be defined as:

$$L_{\text{document}} = \sum_{a, a'} t_{aa'}^*(D, s^*) c_{aa'}, \hspace{1cm} (20)$$

which is associated with the optimal transport plan. By minimizing $L_{\text{document}}$, a high-quality summary sentence is expected to be obtained.

2) Optimal Transport-Guided Video Coverage: In parallel, a good cover frame is supposed to be close to the original video regarding their perceptual similarity. We measure the loss of visual coverage by computing the Wasserstein distance $L_{\text{video}}$ between the corresponding colour signatures of the mean of video frames in $V$ and the cover frame $f^*$. It can be viewed as the minimal cost required to transport the semantics from $f^*$ to $V$.

By defining $\tilde{f}$ as the mean of the video frames in $V$, we define $\tilde{f}$ and $r^*$ as the colour signatures of $\tilde{f}$ and $f^*$, respectively. In detail, we have:

$$\tilde{f} = \{ (\tilde{\mu}_1, \tilde{\tau}_1), \ldots, (\tilde{\mu}_n, \tilde{\tau}_n) \},$$

$$r^* = \{ (\mu_1^*, \tau_1^*), \ldots, (\mu_m^*, \tau_m^*) \}, \hspace{1cm} (21)$$

where $\tilde{\mu}_i$ and $\mu_i^*$ are the points in the colour space, and $\tilde{\tau}_i$ and $\tau_i^*$ are the corresponding weights of the points.

An optimal transport plan matrix $T^*(V, f^*) = (t_{bb'}^*(V, f^*)) \in \mathbb{R}^{m \times m^*}$ in pursuit of minimizing the transportation cost between $\tilde{f}$ and $r^*$ can be obtained by solving the following optimization problem:

$$T^*(V, f^*) = \arg\min_{T(V, f^*)} \sum_{b, b'} t_{bb'}(V, f^*) \| \hat{\mu}_b - \mu_{b'}^* \|,$$

$$\text{s.t.} \sum_{b'} t_{bb'}(V, f^*) = \tau_b,$$

$$\sum_{b} t_{bb'}(V, f^*) = \tau_{b'},$$

$$t_{bb'}(V, f^*) \geq 0,$$

$$\forall b, b' \hspace{1cm} (22)$$

where $T(V, f^*)$ is a transport plan. Then, a Wasserstein distance measuring the distance between the two colour signatures can be derived as:

$$L_{\text{video}} = t_{bb'}(V, f^*) \| \hat{\mu}_b - \mu_{b'}^* \|, \hspace{1cm} (23)$$

which is associated with the optimal transport plan. By minimizing $L_{\text{video}}$, a high-quality summary frame is expected to be the cover frame.

3) Textual Fluency: Inspired by [16], we adopt a pre-trained language model $P_{LM}$ to measure the fluency of the textual summary $L_{\text{Fluency}}$. The loss can be defined as:

$$L_{\text{Fluency}} = P_{LM}(s^*), \hspace{1cm} (24)$$

In summary, four losses have been obtained to measure the summarisation quality: $L_{\text{document}}$, $L_{\text{video}}$, $L_{\text{Fluency}}$ and $L_{\text{cross-modal}}$. To this end, a loss function to optimize the proposed architecture can be formulated as follows:

$$L = \lambda_d L_{\text{document}} + \lambda_v L_{\text{video}} + \lambda_f L_{\text{Fluency}} + \lambda_c L_{\text{cross-modal}}, \hspace{1cm} (26)$$

where $\lambda_d$, $\lambda_v$, $\lambda_f$ and $\lambda_c$ are the hyper-parameters controlling the weights of each loss term.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Dataset

To the best of our knowledge, there is no existing large-scale dataset for XMSMO. Hence, we constructed the first large-scale dataset of such kind, XMSMO-News, from the British Broadcasting Corporation (BBC) News Youtube channel.\(^3\) We used the Pytube library to collect 4,891 quartets of video, document, cover frame, and one-sentence summary from Year 2013 to Year 2021. We used the video description as the document and video title as the one-sentence summary.

\(^3\)https://www.youtube.com/c/BBCNews
Fig. 4. Sample quartets (i.e., title, cover frame, description, and video) with their transcripts in our XMSMO-News dataset.

as these visual and textual summaries were professionally created by the BBC.\textsuperscript{4} We then split the quartets randomly into the train, validation, and test sets at a ratio 90:5:5. To facilitate future research that may utilise audio modality, we also collected the transcript of the video, which is automatically generated by Youtube. Six samples from XMSMO-News are shown in Figure 4. It shows that our dataset covers a wide variety of topics.

### Table I: Comparison between XMSMO-News and two existing MSMO datasets

| Dataset          | XMSMO-News | MSMO[5] | VMSMO[17] |
|------------------|------------|---------|-----------|
| Duration (s)     | 345.5      | 346.0   | 346.0     |
| Novel n-grams    | 38.57%     | 17.59%  | 17.59%    |
| Source Documents | 34.4%      | 34.4%   | 34.4%     |

\textsuperscript{4}We removed the trailing promotional text from the video title and video description.
TABLE II
THE PROPORTION OF NOVEL N-GRAMS (%) IN GROUND-TRUTH SUMMARIES IN XMSMO-NEWS AND MSMO DATASETS. RESULTS ARE COMPUTED ON THE TEST SET. WE SHOW THAT OUR XMSMO-NEWS DATASET IS MORE ABSTRACTIVE AND MORE CHALLENGING SINCE THE SUMMARY CONSISTS OF MORE NOVEL WORDS

|              | XMSMO-News | MSMO |
|--------------|------------|------|
| unigrams     | 38.57      | 17.59|
| bigrams      | 78.24      | 52.01|
| trigrams     | 91.25      | 69.49|
| 4-grams      | 95.58      | 77.68|

contain some hallucinated content that cannot be verified from the source video and document.

B. Implementation Details

We used the PyTorch library for the implementation of our method. We set the hidden size of GPO and GRU to 512. For the pre-trained CLIP model and the pre-trained token embedding model BERT (base version) used for computing the loss of textual coverage, we obtained them from HuggingFace. To detect the scenes of a video, we utilised the PySceneDetect library. To compute the Wasserstein distances, we utilised the POT library and the OpenCV library, respectively. For video preprocessing, we extracted one of every 360 frames to obtain 120 frames as candidate frames. All frames were resized to 640 × 360. We trained HOT-Net using AdamW [21] with a learning rate of 0.01 and a batch size of 3 for about 72 hours. All experiments were run on a GeForce GTX 1080Ti GPU card. For evaluation, we obtained our ROUGE scores by using the pyrouge package.

C. Baselines

To evaluate our proposed method HOT-Net, we compared it with the following categories of baseline methods. 1) Extreme multimodal summarisation method PEGASUS-XSUM + CA-SUM [3], [49], which is a combination of the state-of-the-art method of the extreme text summarisation task PEGASUS-XSUM [49] and that of the extreme video summarisation task CA-SUM [3], respectively; 2) Multi-modal summarisation with multimodal output approach includes: VMSMO [17], which is the state-of-the-art multimodal summarisation method utilising video and document as input, and zero-shot CLIP [37] method, which is based on the state-of-the-art multimodal embedding method CLIP with a fully connected layer for classification to perform multimodal summarisation; and 3) Unimodal extreme summarisation methods for reference include: PG [39], ProphetNet [36], and PEGASUS-XSUM [49], which are the state-of-the-art methods of extreme text summarisation, and ARL [2] and CA-SUM [3], which are the state-of-the-art method of extreme video summarisation. The baseline models PEGASUS-XSUM and CLIP were obtained from HuggingFace [45], PG was obtained from the Github, and ProphetNet [36], CA-SUM [3], and VMSMO [17] were obtained from the authors’ implementations. CA-SUM was obtained from the author’s Github; VMSMO was obtained from the author’s Github with modifications on the latest libraries’ update and bug fixing.

D. Quantitative Analysis

For the quantitative evaluation of a textual summary, we followed the same evaluation protocol as the baseline methods [17], [49] and adopted the commonly used ROUGE metric [18] for text summarisation. For the visual summary, the commonly used Intersection over Union (IoU) [40] and frame accuracy [29] metrics for video summarisation are adopted.

The ROUGE metric evaluates the content consistency between a generated summary and a reference summary. In detail, the ROUGE-n F-scores calculates the number of overlapping n-grams between a generated summary and a reference summary.

ROUGE-n = \frac{\text{# overlapping n-grams}}{\text{# n-grams in ground-truth summary}}. \tag{27}

IoU metric evaluates the high-level semantic information consistency by counting the number of overlap concepts between the ground-truth cover frame and the generated one:

IoU = \frac{\text{# overlapping concepts}}{\text{# concepts in total}}. \tag{28}

Frame accuracy metric is to compare lower-level visual features, the ground-truth cover frame and generated cover frame are considered to be matched when pixel-level Euclidean distance is smaller than a pre-defined threshold.

Accuracy = \frac{\text{# Matched cover frame}}{\text{# Ground-truth cover frames in total}}. \tag{29}

To evaluate the overall performance on both modalities, we compute the overall evaluation as:

0.5 \times \frac{\text{IoU}}{\text{Best IoU}} + 0.5 \times \frac{\text{ROUGE-L}}{\text{Best ROUGE-L}}. \tag{30}

where the best IoU and the best ROUGE-L are the best scores among all the evaluated methods.

The experimental results of HOT-Net on XMSMO-News are shown in Table III including ROUGE-1, ROUGE-2, and ROUGE-L F-scores, and IoU. Our method outperforms the baseline models in terms of ROUGE-1 and ROUGE-L, which demonstrate the quality of the generated extreme textual summary, and achieves promising results in terms of frame accuracy and IoU, which demonstrate the quality of the generated extreme visual summary. HOT-Net underperforms in terms of ROUGE-2, which may be due to the trade-off between informativeness and fluency. PEGASUS-XSUM was trained on massive text corpora which may help improve the fluency of natural language generation. This trade-off is further discussed in the Qualitative Analysis section. Our work is the

---

5https://huggingface.co
6http://scenedetect.com/en/latest/
7https://pythonot.github.io
8https://pypi.org/project/pyrouge/
9https://github.com/kukrishna/pointer-generator-pytorch-allennlp
10https://github.com/e-apostolidis/CA-SUM
11https://github.com/irisxxy/VMSMO
12We followed [29] to set the predefined threshold to 0.6.
TABLE III

Comparisons Between Our HOT-NET and the State-of-the-Art Summarisation Methods on XMSMO-NEWS. Our Method Outperforms the Baseline Models in Terms of ROUGE-1 and ROUGE-L, Which Demonstrate the Quality of the Generated Extreme Textual Summary, and Achieves Promising Results in Terms of Frame Accuracy and IOU, Which Demonstrate the Quality of the Generated Extreme Visual Summary

| Method                             | Textual Evaluation | Visual Evaluation | Overall Evaluation |
|------------------------------------|--------------------|-------------------|--------------------|
|                                    | ROUGE-1  | ROUGE-2  | ROUGE-L | Frame Accuracy | IOU  |                     |
| Extremal Test Summarisation        |          |          |         |              |      |                     |
| PG [39]                            | 2.43     | 0.08     | 2.25    | -            | -    | -                   |
| ProphetNet [36]                    | 3.77     | 0.09     | 3.56    | -            | -    | -                   |
| PEGASUS-XSUM [49]                  | 4.36     | 0.12     | 4.00    | -            | -    | -                   |
| Extreme Video Summarisation        |          |          |         | 0.59         | 0.68 | -                   |
| ARL [2]                            | -        | -        | -       |              |      | -                   |
| CA-SUM [3]                         | -        | -        | -       |              |      | -                   |
| Multimodal Summarisation with Multimodal Output |         |          |         | 0.57         | 0.69 | 0.49                |
| VMSMO [17]                         | Divergence| Divergence| Divergence| 0.54         | 0.63 | 0.89                |
| CLIP [37]                          | 4.14     | 0.08     | 3.80    |              |      |                     |
| Extreme Multimodal Summarisation   |          |          |         |              |      |                     |
| PEGASUS-XSUM + CA-SUM              | 4.36     | 0.12     | 4.00    | 0.57         | 0.69 | 0.95                |
| HOT-Net (Ours) visual only         |          |          |         |              |      |                     |
| HOT-Net (Ours) textual only        | 3.85     | 0.05     | 3.60    | 0.60         | 0.68 | -                   |
| HOT-Net (Ours) w/o multimodal fusion | 3.99   | 0.05     | 3.73    | 0.56         | 0.70 | 0.93                |
| HOT-Net (Ours) w/o local-level multimodal fusion | 4.45 | 0.06  | 4.16    | 0.59         | 0.70 | 0.98                |
| HOT-Net (Ours) w/o global-level multimodal fusion | 3.65 | 0.06      | 3.45    | 0.58         | 0.68 | 0.88                |
| HOT-Net (Ours) w/o flow loss       | 4.58     | 0.06     | 4.28    | 0.57         | 0.68 | 0.98                |
| HOT-Net (Ours) w/o cross-modal loss | 4.58 | 0.06 | 4.28   | 0.57        | 0.68 | 0.98                |
| HOT-Net (Ours)                      | 4.64     | 0.07     | 4.33    | 0.57         | 0.68 | 0.99                |

Fig. 5. Example summaries generated by baseline methods and HOT-Net on XMSMO-News. The example on the left hand side is about a US congressman who made an unusual appearance and flipped upside down. The example on the right hand side is about US President Donald Trump’s UK visit. HOT-Net produces a factually correct and reasonably fluent extreme textual summary that captures the essence of the document. In comparison, as highlighted in red colour, PEGASUS-XSUM produces a fluent yet unfaithful summary. Most of the methods agree on the choice of the cover frame which may be due to its visual representativeness. For the example on the right hand side, since the aeroplane appears repeatedly and occupies a comparatively large area on the frames, there is room for improvement to learn and identify the information which human considers to be important, such as a frame containing the face of the key human figure.
possible potential in terms of producing a visual summary. Due to that the overall model architecture has achieved its best impact on the results of the visual summary, which may be of the textual summary. However, it does not have much local-and-global hierarchical mechanism improves the results also important to obtain high-quality textual summaries. The to the visual or textual-only method. Our fusion strategy is multimodal learning improves the modelling by comparing it the results can be found in Table III. We first observe that tigate a number of different settings of our HOT-Net and have lower transport cost, and thus achieve a minimum transportation cost in the OT plan.

first study on this new topic and we expect the performance to improve over time.

The ROUGE metric reflects lexical matching only and often overlook the conciseness, linguistic quality and factuality of a summary, which are the key quality of a good summary. Since XMSMO-News extreme summaries are highly abstractive and very short (the summaries of the extreme text summarisation dataset XSum [31] have 23 tokens on average; meanwhile, textual summaries of XMSMO-News have 12 tokens on average), conciseness, linguistic quality and factuality metrics would be essential to provide further insights. However, human evaluation is highly subjective, which is challenging to draw a meaningful comparison and conclusion, as pointed out in [13]. We advocate developing automatic metrics of conciseness, linguistic quality, and factuality for more meaningful evaluations in future research.

E. Ablation Study

To study the effect of the proposed mechanisms, we inves-
tigate a number of different settings of our HOT-Net and the results can be found in Table III. We first observe that multimodal learning improves the modelling by comparing it to the visual or textual-only method. Our fusion strategy is also important to obtain high-quality textual summaries. The local-and-global hierarchical mechanism improves the results of the textual summary. However, it does not have much impact on the results of the visual summary, which may be due to that the overall model architecture has achieved its best possible potential in terms of producing a visual summary. Additionally, the fluency loss and cross-modal loss improve the textual summary as well.

F. Qualitative Analysis

Figure 5 compares the summaries produced by HOT-Net and the baseline methods, and the reference summary of a sample in the XMSMO-News dataset. The example demonstrates that our proposed HOT-Net method produces factually correct and reasonably fluent extreme textual summary that captures the essence of the document even without supervision. In comparison, as highlighted in red colour, PEGASUS-XSUM produces a fluent but unfaithful summary with information that does not occur in the original document. Most of the methods agree on the choice of the cover frame, whilst the results of ours and CA-SUM are closer to the ground truth. For the second example, since the aeroplane appears repeatedly and occupies a comparatively large area on the frames, there is room for improvement to learn and identify the information which human considers to be important, such as a frame containing the face of the key human figure.

G. Interpretable Visualisation of Semantic Coverage

HOT-Net is able to provide an interpretable visualisation of the textual semantic coverage on the summarisation results. Figure 6 illustrates the transport plan heatmap, which indicates the transportation of semantic contents between tokens in the document and its resulting summary. The higher the colour intensity, the more the semantic content of a particular document token is covered by a summary token.

V. CONCLUSION

In this paper, we have introduced a new task to the field of summarisation, eXtreme Multimodal Summarisation with Multimodal Output (XMSMO), which aims to summarise a video-document pair into an extreme multimodal summary, consisting of one cover frame as the visual summary and one sentence as the textual summary. We present a novel unsupervised deep learning architecture, namely, Hierarchical Optimal Transport Network (HOT-Net), which consists of three components: hierarchical multimodal encoder, hierarchical multimodal fusion decoder, and optimal transport solver. To achieve unsupervised learning, besides the optimal transport-based semantic coverage guidance, textual fluency and cross-modal similarity are explored as well. In addition, we constructed a new large-scale dataset, XMSMO-News, to facilitate research in this new direction. Experimental results demonstrate the effectiveness of our method.

According to the data analysis and experiments on our XMSMO-News dataset, our new task XMSMO has the follow-
ing challenges: 1) Difficulty to identify the most salient textual and visual information. Since the summaries are extremely succinct, users would expect the summaries to include only the most important information. 2) Difficulty to ensure factuality and faithfulness. The extreme summaries is expected to provide a only gaze to the full picture of news event, it would be essential for users to trust that the summaries are representative and accurate descriptions of what happened.
without requiring further validation by digesting the details. 3) Difficulty to evaluate the model performance. Since the summaries are extremely short, the commonly used evaluation metrics of textual and visual summarisation, which are mainly designed to longer summaries and usually measure token and object overlapping, may penalise the false tokens and objects heavily and may not be a good fit to reflect the performance of the models. Developing automatic metrics that measure the qualities of the summaries, including informativeness, conciseness, linguistic and image quality, and factuality, would provide more meaningful evaluations.

In the future, we will explore the metric space to measure the optimal transport plan in a more efficient and effective manner. Moreover, we will explore improved ways to learn and identify the information that humans would consider to be important, such as a frame containing the face of a key character.

REFERENCES

[1] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, “Video summarization using deep neural networks: A survey,” Proc. IEEE, vol. 109, no. 11, pp. 1838–1863, Nov. 2021.
[2] E. Apostolidis, E. Adamantidou, V. Mezaris, and I. Patras, “Combining adversarial and reinforcement learning for video thumbnail selection,” in Proc. Int. Conf. Multimedia Retr., Aug. 2021, pp. 1–9.
[3] E. Apostolidis, G. Balouaras, V. Mezaris, and I. Patras, “Summarizing videos using concentrated attention and considering the uniqueness and diversity of the video frames,” in Proc. Int. Conf. Multimedia Retr., Jun. 2022, pp. 407–415.
[4] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in Proc. Int. Conf. Learn. Represent., 2015.
[5] L. Baraldi, C. Grana, and R. Cucchiara, “Hierarchical boundary-aware neural encoder for video captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3185–3194.
[6] I. Cachola, K. Lo, A. Cohan, and D. Weld, “TLDR: Extreme summarization of scientific documents,” in Proc. Findings Assoc. Comput. Linguistics, EMNLP, 2020, pp. 4766–4777.
[7] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, “Learning the best pooling strategy for visual semantic embedding,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 15784–15793.
[8] J. Chen, Y. Pan, Y. Li, T. Yao, H. Chao, and T. Mei, “Temporal deformable convolutional encoder–decoder networks for video captioning,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 8167–8174.
[9] Y. Chen, S. Wang, W. Zhang, and Q. Huang, “Less is more: Picking informative frames for video captioning,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer-Verlag, 2018, pp. 367–384.
[10] H. Gu and V. Srinivasan, “From thumbnails to summaries—A single deep neural network to rule them all,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2018, pp. 1–6.
[11] G. Guan, Z. Wang, S. Lu, J. D. Deng, and D. D. Feng, “Keypoint-based keyframe selection,” IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 4, pp. 729–734, Apr. 2013.
[12] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, “Joint syntax representation learning and visual cue translation for video captioning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8917–8926.
[13] D. M. Howcroft et al., “Twenty years of confusion in human evaluation: NLI data evaluation sheets and standardised definitions,” in Proc. Int. Conf. Natural Lang. Gener. (INLG), 2020, pp. 169–182.
[14] 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.
[15] M. Kunser, Y. Sun, N. Kolkkin, and K. Weinberger, “From word embeddings to document distances,” in Proc. Int. Conf. Mach. Learn. (ICML), 2015, pp. 967–975.
[16] P. Laban, A. Hsi, J. Canny, and M. A. Hearst, “The summary loop: Learning to write abstractive summaries without examples,” in Proc. 58th Annu. Meet. Assoc. Comput. Linguistics, 2020, pp. 5135–5150.
[17] M. Li, X. Chen, S. Gao, Z. Chan, D. Zhao, and R. Yan, “VYMSMO: Learning to generate multimodal summary for video-based news articles,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 1–6.
[18] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013.
[19] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, “Multi-task deep visual-semantic embedding for video thumbnail selection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3707–3715.
[20] Y. Liu and M. Lapata, “Text summarization with pretrained encoders,” in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), 2019, pp. 3730–3740.
[21] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.
[22] Y. Lu, Y. Dong, and L. Charlin, “Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 1–12.
[23] J. Luo, C. Papin, and K. Costello, “Towards extracting semantically meaningful key frames from personal video clips: From humans to computers,” IEEE Trans. Circuits Syst. Video Technol., vol. 19, no. 2, pp. 298–301, Feb. 2009.
[24] C. Ma, L. Lyu, G. Lu, and C. Lyu, “Adaptive multiview graph difference analysis for video summarization,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 12, pp. 8795–8808, Dec. 2022.
[25] M. Ma, S. Mei, S. Wan, J. Hou, Z. Wang, and D. D. Feng, “Video summarization via block sparse dictionary selection,” Neurocomputing, vol. 378, pp. 197–209, Feb. 2020.
[26] M. Ma, S. Mei, S. Wan, Z. Wang, D. D. Feng, and M. Bennamoun, “Similarity based block sparse subset selection for video summarization,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 10, pp. 3967–3980, Oct. 2021.
[27] I. Mani, Automatic Summarization. Amsterdam, The Netherlands: John Benjamins Publishing, 2001.
[28] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” in Proc. 58th Annu. Meet. Assoc. Comput. Linguistics, 2020, pp. 1–16.
[29] S. Messaoud et al., “DeepQAMVS: Query-awake hierarchical pointer networks for multi-video summarization,” in Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Jul. 2021, pp. 1–4.
[30] R. Mihalcea and P. Tarau, “TextRank: Bringing order into text,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2004, pp. 404–411.
[31] S. Narayan, S. B. Cohen, and M. Lapata, “Don’t give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2018, pp. 1797–1807.
[32] S. Narayan, S. B. Cohen, and M. Lapata, “Ranking sentences for extractive summarization with reinforcement learning,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., 2018, pp. 1–15.
[33] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, “Jointly modeling embedding and translation to bridge video and language,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Los Alamitos, CA, USA, Jun. 2016, pp. 4594–4602.
[34] R. Paulus, C. Xiong, and R. Socher, “A deep reinforced model for abstractive summarization,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.
[35] K. Pretorious and N. Pillay, “A comparative study of classifiers for thumbnail selection,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2020, pp. 1–7.
[36] W. Qi et al., “ProphetNet: Predicting future N-gram for sequence-to-sequence pre-training,” in Proc. Findings Assoc. Comput. Linguistics, EMNLP, 2020, pp. 2401–2410.
[37] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763.
[38] J. Ren, X. Shen, Z. Lin, and R. Mech, “Best frame selection in a short video,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2020, pp. 3201–3210.
[39] A. See, P. J. Liu, and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” in Proc. 55th Annu. Meet. Assoc. Comput. Linguistics, 2017, pp. 1073–1083.
A. Sharghi, J. S. Laurel, and B. Gong, “Query-focused video summarization: Dataset, evaluation, and a memory network based approach,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2127–2136.

Y. Song, M. Redi, J. Vallmitjana, and A. Jaimes, “To click or not to click: Automatic selection of beautiful thumbnails from videos,” in Proc. 25th ACM Int. Conf. Inf. Knowl. Manage., New York, NY, USA, Oct. 2016, pp. 659–668.

S. Sotudeh, H. Deilamsaleh, F. Dernoncourt, and N. Goharian, “TLDR9+: A large scale resource for extreme summarization of social media posts,” in Proc. 3rd Workshop New Frontiers Summarization, 2021, pp. 142–151.

B. Zhang, Z. Wang, D. Tao, X. S. Hua, and D. D. Feng, “Automatic preview frame selection for online videos,” IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 1, pp. 445–456, Jan. 2023.

Y. Yuan and J. Zhang, “Unsupervised video summarization via deep reinforcement learning with shot-level semantics,” IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 1, pp. 445–456, Jan. 2023.

J. Zhang, Y. Zhao, M. Saleh, and P. Liu, “PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization,” in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 11328–11339.

Kun Hu (Member, IEEE) received the B.Sc. and B.Eng. degrees from Shandong University, Jinan, China, in 2013, the M.Sc. degree from Shanghai Jiao Tong University, Shanghai, China, in 2016, and the Ph.D. degree from the School of Computer Science, The University of Sydney, Sydney, NSW, Australia, in 2022. He is a Research Associate with the School of Computer Science, The University of Sydney. His research interests include multimedia computing, computer vision, animation, and pattern recognition.

Lei Zhang (Fellow, IEEE) received the Ph.D. degree in computer science from Tsinghua University in 2001. He is a Chief Scientist of computer vision and robotics with the International Digital Economy Academy (IDEA) and an Adjunct Professor with the Hong Kong University of Science and Technology. Prior to his current post, he was a Principal Researcher and a Research Manager with Microsoft, working on computer vision and machine learning. He is interested in image understanding, visual pattern recognition, and machine learning.

Jiebo Luo (Fellow, IEEE) is currently a Professor of computer science with the University of Rochester, where he joined in 2011, after a prolific career of 15 years with the Kodak Research Laboratories. He has authored over 400 technical articles and holds over 90 U.S. patents. His research interests include computer vision, NLP, machine learning, data mining, computational social science, and digital health. He is a fellow of ACM, AAAI, SPIE, and IAPR. He has been involved in numerous technical conferences, including serving as the Program Co-Chair for ACM Multimedia 2010, IEEE CVPR 2012, ACM ICMR 2016, and IEEE ICIP 2017; and the General Co-Chair for ACM Multimedia 2018. He has served on the Editorial Boards for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, ACM TRANSACTIONS ON INTELLIGENT SYSTEMS AND TECHNOLOGY (TIST), Pattern Recognition, Knowledge and Information Systems (KAIS), Machine Vision and Applications, and Journal of Electronic Imaging. He is currently the Editor-in-Chief of the IEEE TRANSACTIONS ON MULTIMEDIA.