Promotional videos are rapidly becoming a popular medium for persuading people to change their behaviours in many settings (e.g., online shopping, social enterprise initiatives). Today, such videos are often produced by professionals, which is a time-, labour- and cost-intensive undertaking. In order to produce such contents to support large applications (e.g., e-commerce), the field of artificial intelligence (AI)-empowered persuasive video generation (AIPVG) has gained traction in recent years. This field is interdisciplinary in nature, which makes it challenging for new researchers to grasp. Currently, there is no comprehensive survey of AIPVG available. In this paper, we bridge this gap by reviewing key AI techniques that can be utilized to automatically generate persuasive videos. We offer a first-of-its-kind taxonomy which divides AIPVG into three major steps: (1) visual material understanding, which extracts information from the visual materials (VMs) relevant to the target of promotion; (2) visual storyline generation, which shortlists and arranges high-quality VMs into a sequence in order to compose a storyline with persuasive power; and (3) post-production, which involves background music generation and still image animation to enhance viewing experience. We also introduce the evaluation metrics and datasets commonly adopted in the field of AIPVG. We analyze the advantages and disadvantages of the existing works belonging to the above-mentioned steps, and discuss interesting potential future research directions.

CCS Concepts: • Applied computing → Online shopping; • Computing methodologies → Computer vision;

Additional Key Words and Phrases: Artificial intelligence, video generation, storyline generation

ACM Reference format:
Chang Liu and Han Yu. 2023. AI-Empowered Persuasive Video Generation: A Survey. ACM Comput. Surv. 55, 13s, Article 285 (July 2023), 31 pages.
https://doi.org/10.1145/3588764

This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-019); the Joint SDU-NTU Centre for AI Research (C-FAIR); the RIE 2020 Advanced Manufacturing and Engineering (AME) Programmatic Fund (No. A20G8b0102), Singapore; the Nanyang Assistant Professorship (NAP); and Future Communications Research & Development Programme (FCP-NTU-RG-2021-014).

Authors’ address: C. Liu and H. Yu, School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore, 639798, emails: chang015@e.ntu.edu.sg, han.yu@ntu.edu.sg.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
0360-0300/2023/07-ART285 $15.00
https://doi.org/10.1145/3588764
1 INTRODUCTION

In e-commerce platforms (e.g., Taobao, Lazada, JD), promotional videos are rapidly becoming a popular form of product advertising. According to reports, 80% of video marketers stated that videos had helped increase their sales. The production of promotional videos has hitherto been performed mainly by professional teams. For individual or micro e-commerce sellers, hiring a professional video production team is unaffordable. Besides, manual video production is time-consuming, which hinders large-scale adoption of video-based product promotion campaigns.

To address this issue, artificial intelligence (AI) techniques for generating videos have been developed [16, 21, 54, 85, 86]. This field can be referred to as AI-empowered persuasive video generation (AIPVG). AIPVG systems usually take visual materials (VMs) (e.g., images and video clips) provided by sellers as the input, and generate videos as the output. One of the main design objectives of AIPVG is to motivate the viewers to change their behaviours (e.g., purchase products from an e-commerce platform, adopt an active and healthy lifestyle). Thus, generating videos that are persuasive is a major goal of AIPVG.

Persuasion has been a subject of study in social sciences for decades [5, 118]. Perloff [118] defines persuasion as a symbolic process in which communicators try to convince other people to change their attitudes or behaviours regarding an issue through the transmission of a message, in an atmosphere of free choice. In e-commerce, the transmission of messages can be carried out via images, texts, and promotional videos. The goal is to influence potential buyers’ purchasing decisions. Here, we provide a brief overview of persuasion models commonly adopted by AIPVG literature.

Elaboration Likelihood Model (ELM) [119] divides the process of persuasion into two routes: (1) the central route, and (2) the peripheral route. Persuasion through the central route requires the persuader to carefully analyse the facts surrounding the subject of persuasion in order to formulate a persuasive message. On the other hand, the peripheral route relies on auxiliary cues of the persuasion context (e.g., aesthetics, building relationships with the persuadees). In [119], the authors proposed to use peripheral routes when the persuadee has no motivation to analyze, cannot process information, or holds a neutral attitude towards the issue. Note that if a persuadee holds unfavourable thoughts, the peripheral routes may not be able to change his/her mind.

Focusing more on the knowledge of both the persuadee and persuader, the Persuasion Knowledge Model (PKM) [39] models the persuasion process in an adversarial form. The persuader performs persuasion based on his/her knowledge, while the persuadee copes with the persuasion attempt. Three types of knowledge are considered in the PKM: (1) topic knowledge, which refers to the knowledge related to the subject of persuasion (e.g., a product); (2) persuasion knowledge, which is about how to successfully persuade people (for persuaders), or how to cope with such persuasion (for persuadees); and (3) persuadee/persuader knowledge, which is the information that the persuadee or the persuader know about each other.

Applying AI techniques to videos generation has been a long-running research field. Many survey papers have been published on this broad topic. For example, video summarization techniques have been surveyed in [6, 55, 131]. Though video summarization and AIPVG generate videos that can tell stories, the design objectives and the inputs of these techniques are very different. Specifically, the video summarization trims a long video into a short version while aiming to preserve the original information. AIPVG, on the other hand, takes a set of VMs as the input to generate videos that aim to persuade viewers to change certain aspects of their behaviour. Several other papers have surveyed the field of video synthesis with generative models [17, 59]. Such techniques

1https://www.wyzowl.com/sovm-results-2020/.
do not require a set of VMs as the input to generate the videos. However, they can only generate videos without sophisticated storylines (e.g., talking-head videos [17] or virtual try-on videos [59]). AIPVG generates videos with storylines designed for persuasion by selecting and sequencing input VMs. This line of research work is distinct from the topics the previous survey papers covered. Moreover, some papers survey a part of the AIPVG techniques, e.g., deep metric learning [63, 95], CNN architectures [64], and recommender systems [25, 196]. However, they mainly focus on the specific techniques instead of their functionality, strength, weakness, and future works in the context of AIPVG. Currently, to the best of our knowledge, there is no comprehensive survey on the topic of AIPVG.

The techniques involved in AIPVG are interdisciplinary in nature, involving computer science, social sciences, and film theory. Thus, it can be challenging for new-comers to grasp the important developments in this field. The lack of a comprehensive survey on this topic further exacerbates this challenge. In this paper, we bridge this gap by providing a comprehensive review of the existing literature on AIPVG. This paper contributes to the AI literature in the following ways:

- Firstly, we provide an overview of the persuasion theory which plays an important role in AIPVG approach designs. This can be helpful for researchers to gain useful background knowledge on the ways through which technologies can influence users’ behaviours.
- Secondly, we propose a unique taxonomy which divides existing works in AIPVG into three main categories: (1) visual material understanding, (2) visual storyline generation, and (3) post-production. For each of these categories, we provide an overview of the main research challenges, discuss the key ideas of notable works, and highlight potential areas for improvement. We also introduce how to evaluate the generated videos.
- Finally, we outline ten promising future research directions towards building practical and efficient AIPVG approaches.

2 THE PROPOSED AIPVG TAXONOMY

The proposed AIPVG taxonomy divides existing works in this field into three main categories, as shown in Figure 1. Each of these categories correspond to one important step in the process of generating promotional videos with AI:

1. **Visual Material Understanding**: Works in this category aim to enable AI to extract useful information from images or videos provided by the sellers relevant to the target of promotion. They generally serve to enhance the informativeness and attractiveness of the generated promotional videos, which can be utilized by the subsequent steps.

2. **Visual Storyline Generation**: Works in this category focus on selecting and ordering the visual materials to compose a persuasiveness sequence. In this step, the informativeness and attractiveness of the visual materials are often taken into account to facilitate the selection and sequencing operations.

3. **Post-Production**: Works in this category aim to improve the viewing experience for the users. They include background music generation, image animation, and video rendering with special effects. This step can add more style and attractiveness to the final video output.

In the following sections, we discuss the main research challenges, ideas of notable works, and highlight potential areas for improvement in each of the aforementioned categories.

3 VISUAL MATERIAL UNDERSTANDING

Visual Material Understanding techniques that are useful for video generation can be divided into two parts: (1) Visual Material Representation, which takes a VM as the input and extracts a feature
from the VM. A good representation preserves information about the target object depicted in the image. It serves as a low-level understanding of the VMs for video generation and can be helpful to facilitate informativeness based central route of persuasion \[119\]; and (2) Cognitive Property Assessment (CPA), which estimates scores of cognitive properties for given VMs. It aims to predict the intensity of stimulation when viewers watch the VMs. Thus, it is considered as a high-level understanding approach. The cognitive properties help identify high quality VMs for the peripheral route of persuasion \[119\]. Commonly adopted cognitive properties in visual material understanding include emotions, aesthetics, and memorability.

3.1 Visual Material Representation

To generate promotional videos by leveraging AI techniques, the first step is to represent the visual materials (VMs). The VM representations can be used to perform tasks important for VM sequence generation and post-production (e.g., clustering, similarity calculation, and saliency map generation). A high quality VM representations generation model needs to satisfy the following requirements: (1) it shall generate embeddings that can be easily processed by AI models (e.g., 1-D vectors); and (2) it shall generate embeddings which preserve the rich product information, which is crucial in enhancing the informativeness of the resulting visual storylines, thus improving their persuasiveness \[8\].

3.1.1 Feature Extraction. Feature extraction approaches for visual material understanding have evolved over the years from traditional methods, to deep learning classification methods, and more recently to deep metric learning-based methods.

**Traditional Methods:** Before the emergence of Convolutional Neural Networks, feature engineering was generally performed manually by experts. Scale-invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF) \[10, 94\] have been commonly used to extract image features. The SIFT \[94\] uses local features near detected key points of an image as the representation.
The local features are calculated based on the gradients of the pixels in each sub-region. Rotation and scale normalization are then performed on the features. SURF further leverages an integral image to accelerate SIFT [10]. These traditional image representation methods are generally fast. However, the major drawback is that they cannot learn what features are to be extracted from the input images. It heavily relies on manual feature engineering, which is hard to scale up. The model performance is also affected by the feature engineers’ expertise levels.

Classification-based Deep Learning Methods: Classification-based deep learning VM representation generation methods use the middle layer outputs of a trained classification Deep Neural Networks (DNN) as the feature. Commonly adopted classification DNNs for this task include convolutional neural networks (CNNs) such as AlexNet [73], ResNet [48], and VGG [142]. Recent research also suggests that the transformer networks [33] can achieve superior performance in classification. 3D-CNNs [37] can be used for video classification and feature extraction. Classification DNNs are trained by minimizing the classification loss via gradient descent optimization (e.g., stochastic gradient descent (SGD), Nesterov SGD [150], and Adam [68]). Commonly used loss functions are the cross-entropy loss and focal loss [82]. In practice, to extract the features in general, it is common to use the DNNs which have been pre-trained on well-known large-scale datasets (e.g., ImageNet [27] for images, and Kinetics-400 [62] for videos) as the feature extractors. Since such pre-trained DNNs are often available online, the training process can be skipped.

The advantage of DNNs over traditional methods is that they directly take the RGB images as inputs, and learn to find important regions and features by fitting the training data. In this way, manual feature engineering is no longer required, which enhances the scalability of such approaches. However, the classification tasks do not force models to encode rich product information commonly found in e-commerce applications. The resulting models can achieve good classification performance by only extracting information relevant to identifying pre-defined labels. One may propose to solve this issue by using finer-grained labels (e.g., defining each unique product as a class). However, this leads to a huge classification layer, which can incur high memory and computation costs during training.

Deep Metric Learning-based Methods: Compared to classification-based DNNs, deep metric learning (DML) can better preserve the rich product information in VMs when trained with large-scale fine-grained labels. This is because DML trains a CNN that directly outputs an embedding. The goal of DML is to make the data points within a class to be close together, and those from different classes far apart. The closeness measures can be defined in different ways. Since DML directly optimizes the data representations, the classification layer is not required in the training process. This reduces memory and computation costs, making training on large-scale fine-grained labels viable. The DNNs used for DML are the same as those under the classification-based deep learning methods, with the classification layer replaced by a fully connected layer to output the VM representations with a desired number of dimensions. To train such a model, a variety of loss functions have been proposed.

Proxy-based DML: Proxy-based loss, which represents a group of data with a proxy, has been proposed for DML. The proxy is usually a trainable 1-D vector, which is randomly initialized at first. The loss tries to minimize the distances between the proxy and corresponding data embedding extracted by the model and maximize the distances between proxies and data points that do not match. In the process of minimizing the loss, the proxy can be jointly trained with the model through backpropagation. For example, ProxyNCA loss [106] is defined as the logarithm of the Softmax function. The logits of the Softmax are the euclidean distance between the features of data \(x_i\) and the trainable proxies of \(x_i\).

Many previous works assign one proxy for each class (i.e., one-to-one mapping between data class and proxy), [28, 67, 91, 92, 164]. Nevertheless, approaches that assign multiple proxies for
each class have emerged [106, 121]. Movshovitz-Attias et al. propose randomly assigning multiple classes to a proxy so that the number of proxies can be smaller than the number of classes [106]. This strategy helps to tackle the high costs incurred by having a large number of classes, albeit at the cost of some performance degradation. Besides, [106] also proposed a dynamic proxy assignment, which assigns each data point to its closest proxy. The advantage of dynamic proxy assignment is that it does not require ground truth labels to train the model. On the other hand, static proxy assignment achieves better accuracy under supervised learning settings.

**Pair-based DML:** Pair-based loss is computed based on the similarities between pairs of data points. Two types of data pairs are usually considered in the loss: (1) positive pairs, in which the two data share the same ground truth labels; (2) negative loss, in which the two labels are different. The data pairs are usually formed within minibatch data. The most straightforward pair-based loss is the contrastive loss [22]. It is the sum of the euclidean distance of positive pairs, minus the sum of distance of selected negative pairs. The selection of negative pairs is based on their distance. If the distance is smaller than $\lambda$, then it will be selected. $\lambda$ is a predefined hyperparameter. It helps the loss function to focus on the hard negative examples which appear close together. By minimizing the contrastive loss, the similarities between negative data pairs are minimized, and those between positive pairs are maximized. Another commonly used pair-based loss is the triplet loss [128]. Instead of optimizing the similarity between each pair of data points, the triplet loss considers the relative difference between the distance among positive pairs and that among negative pairs.

Existing research works such as [144, 148, 165, 186] attempt to assign high learning weights to hard examples (i.e., negative data pairs that have high similarity values, or positive data pairs that have low similarity values) so that the model can learn to represent hard examples better. Wang et al. [165] found that most pair-based models weight the negative data pairs based on either self similarities, positive relative similarities, or negative relative similarities compared with other pairs. Specifically, the three types of similarities of negative data pairs are defined as follows:

- **Self similarity**, which is the cosine similarity between the negative sample and the anchor. This is used in contrastive loss [22], binomial deviance loss [186], and MS loss [165].
- **Positive relative similarity**, which is defined as the relative difference between the similarity of a negative data pair and that of positive data pairs. This is used in triplet loss [128], histogram loss [160], NCA loss [127], MS loss [165], and circle loss [148].
- **Negative relative similarity**, which represents the relative difference between the similarity of a negative data pair and that of other negative data pairs. Examples include N-pairs loss [144], lifted structure loss [113], NCA loss [127], and MS loss [165].

The similarities of positive data pairs can be defined in a similar fashion.

Pair-based loss functions enable training VM representation models without a classification layer. However, the accuracy heavily depends on training batch size, because the pairs are built from data within a minibatch. More data pairs are available if the batch size is large, which can better reflect the underlying data distribution. To reduce the dependency on the batch size settings, Wang et al. [167] proposed a memory bank which saves the data features obtained from previous iterations in a first-in-first-out queue. In each iteration, the features of current minibatch data is enqueued into the memory bank, while the oldest features are dequeued if the number of features reaches its maximum limit. By comparing the minibatch data with the memory bank data, the number of compared data pairs can be significantly increased. Thus, applying the memory bank with pair-based loss improves the accuracy of DML.

### 3.1.2 Feature Aggregation

Feature aggregation uses multiple features of the local patches or frames in VMs to generate a compact VM representation. Typically, the output representation has a fixed dimension, while the number of input features is not fixed. Feature aggregation can be...
leveraged to obtain a global feature for an image (when inputs are local patch features) or a video (when inputs are frame features).

**Bag of Visual Words (BoVW)** [24, 143] performs K-means on the features of local patches on all images in a dataset. Each local feature is then denoted by the ID of the cluster it belongs to, which is generated by K-means. Finally, BoVW represents each image as the histogram of local feature IDs. **Vector of Local Aggregated Descriptors (VLAD)** [57] also applies K-means in the same way as BoVW. The sum of the residuals between the local features and its cluster centre is used as the feature of image \(i\) belonging to the cluster \(k\) (denoted as \(V(i, k)\)). Finally, \([V(i, 1), V(i, 2), \ldots, V(i, K)]\) is used as the global representation of the \(i\)-th VM.

To further improve VLAD, NetVLAD [7] has been proposed to express \(V(i, k)\) in a differentiable form so that it can be incorporated into DNNs. The clustering process is also skipped by replacing the cluster centre \(c_k\) with a learnable parameter. NeXtVLAD [81] has been proposed to further reduce the number of parameters compared to VLAD while maintaining model accuracy.

### 3.2 Cognitive Property Assessment

Cognitive properties [43] (e.g., memorability, aesthetics, perceived emotion) are crucial for AIPVG. They represent a high-level understanding of the visual materials which is required for storyline generation. Accurately measuring the cognitive properties of VMs helps identify memorable and attractive VMs, which in turn, increase the persuasiveness of the generated videos [114]. Cognitive properties have been studied in-depth in fields such as psychology, sociology, and cognition science [126, 170]. There has been no mathematical formulation proposed for them, which makes it challenging to estimate the cognitive properties of images and videos using algorithms with hand-crafted features. In recent years, researchers start to build deep learning models to address this problem [76]. Many techniques are proposed for CPA tasks, especially image-based emotion estimation and aesthetics estimation. Table 1 reports the emotion classification accuracy on the FI dataset [191]. Table 2 reports binary classification accuracy on the AVA dataset of existing methods.

### Table 1. Comparative Analysis of Image-based Emotion Assessment Techniques

| Authors            | Classification Accuracy | Venue          | Year |
|--------------------|-------------------------|----------------|------|
| Borth et al. [12]  | 49.23%                  | ACM MM         | 2013 |
| Zhao et al. [198]  | 46.13%                  | ACM MM         | 2014 |
| Chen et al. [19]   | 51.54%                  | Arxiv          | 2014 |
| Zhu et al. [202]   | 73.03%                  | IJCAI          | 2017 |
| Yang et al. [181]  | 67.48%                  | IJCAI          | 2017 |
| Yang et al. [180]  | 67.64%                  | AAAI           | 2018 |
| Yang et al. [179]  | 70.07%                  | CVPR           | 2018 |
| Rao et al. [123]   | 75.46%                  | Neurocomputing | 2019 |
| Zhan et al. [193]  | 68.87%                  | ICCV           | 2019 |
| He et al. [49]     | 68.13%                  | IJCNN          | 2019 |
| Zhang et al. [195] | 71.77%                  | TMM            | 2019 |
| Lin et al. [78]    | 75.24%                  | AAAI           | 2020 |
| Rao et al. [122]   | 67.75%                  | Neural Processing Letters | 2020 |
| Zhang et al. [194] | 69.32%                  | Knowledge-Based Systems | 2020 |
| Yao et al. [185]   | 68.37%                  | TMM            | 2021 |
| Yang et al. [178]  | 72.42%                  | TIP            | 2021 |

Eight-class classification accuracy on FI dataset [191] is reported. Works are sorted by chronological order.
Table 2. Comparative Analysis of Image-based Aesthetics Estimation Techniques

| Authors              | Classification Accuracy | Venue    | Year   |
|----------------------|-------------------------|----------|--------|
| Lu et al. [96]       | 75.4%                   | ICCV     | 2015   |
| Mai et al. [102]     | 77.1%                   | CVPR     | 2016   |
| Kong et al. [72]     | 77.3%                   | ECCV     | 2016   |
| Kao et al. [61]      | 79.1%                   | TIP      | 2017   |
| Murray and Gordo [108]| 80.3%                   | Arxiv    | 2017   |
| Ma et al. [100]      | 82.5%                   | CVPR     | 2017   |
| Schwarz et al. [130] | 75.8%                   | WACV     | 2018   |
| Talebi and Milanfar [151]| 81.5%               | TIP      | 2018   |
| Ko et al. [70]       | 82.2%                   | ICIP     | 2018   |
| Sheng et al. [135]   | 83.0%                   | MM       | 2018   |
| Zeng et al. [192]    | 80.8%                   | TIP      | 2019   |
| Hosu et al. [50]     | 81.7%                   | CVPR     | 2019   |
| Shu et al. [136]     | 65.8%                   | Neurocomputing | 2020 |
| Hou et al. [51]      | 81.7%                   | ACM MM   | 2020   |
| Chen et al. [18]     | 83.2%                   | CVPR     | 2020   |
| Liu et al. [90]      | 83.6%                   | WACV     | 2020   |
| Li et al. [77]       | 83.7%                   | TIP      | 2020   |
| Jang and Lee [56]    | 69.5%                   | IEEE Access | 2021 |
| Hou et al. [52]      | 81.9%                   | Arxiv    | 2021   |
| She et al. [134]     | 84.6%                   | CVPR     | 2021   |
| Lv et al. [99]       | 85.1%                   | TMM      | 2021   |

Binary classification accuracy on the AVA dataset [109] is reported. Works are sorted by chronological order.

3.2.1 CNN-based CPA. The baseline models to assess cognitive image properties are CNNs (e.g., AlexNet [73], ResNet [48], and VGG [142]) which have been pre-trained on ImageNet [27] and fine-tuned on CPA datasets. A CNN takes an image as the input, and outputs a score that measures each type of cognitive property. Existing research works [56, 175] show that features learned from image classification tasks benefit cognitive image property assessment. To improve the traditional CNN architecture for CPA tasks, a variety of model components have been proposed.

Fusing Features of Different Layers: It has been found that combining features of different layers in a CNN can improve emotion classification performance [122, 180, 202]. This is because different layers refer to different parts of an image [202]. To fuse the features of different layers, a layer fusion component is needed in the existing CNN architectures. Zhu et al. [202] designed this component as a bi-directional GRU, which is a variant of RNN [20]. Other works such as [122, 180] designed it as a sub-network consisting of convolutions and fully connected layers. These components can either be trained separately [202] or together with the CNN backbones [122, 180].

Avoiding Re-scaling: The resolution and aspect ratio of an image are important when assessing its quality and aesthetics. Thus, it would be desirable if these two pieces of information can be preserved in CPA tasks. There are two main approaches for achieving this goal: (1) using pooling layers, or (2) using feature extractors.

Existing works such as [18, 70, 108] leverage spatial pyramid pooling (SPP) [47] to avoid altering the resolution and aspect ratio of an input image. The SPP layer is able to pool arbitrarily-sized features x and generate fixed-length outputs. It can be viewed as a layer composed of three adaptive pooling operations. Each adaptive operation takes x as the input and outputs a fixed-sized
matrix. Here, “adaptive” means that the kernel size \((w_{\text{kernel}}, h_{\text{kernel}})\) and stride \((s_{\text{width}}, s_{\text{height}})\) of the pooling depends on the input size and desired output size (i.e., \(w_{\text{kernel}} = \lceil \frac{w_{\text{input}}}{w_{\text{output}}} \rceil \) where \(w_{\text{input}}\) and \(w_{\text{output}}\) are the widths of the input and the output, respectively). Similarly, the height of the kernel can be defined. Finally the three pooled features are flattened and concatenated to become the output of the SPP layer.

Besides, Chen et al. [18] proposed a component that linearly combines features extracted via two dilated convolution operations at different dilation rates so as to preserve information related to aspect ratios. The feature combination can be viewed as a variant of (weighted) average pooling, where the weights are calculated based on the aspect ratio and dilation rate. Ma et al. [100] proposed A-Lamp, which selects multiple patches using a patch selection sub-module, VGG16 is used to extract features from the unscaled selected patches. The resulting features are the concatenation of max-pooled and average-pooled features of multiple patches.

Another line of research first extracts features by using feature extractors, and then uses a regressor to predict the final score. The extractor can either be expert-designed [156, 158] or pretrained neural networks [50, 188]. The extractors are not fine-tuned on the target dataset to reduce time and memory costs. After obtaining the features, a regressor which takes the features as the inputs and outputs the score is trained. Ying et al. [188] proposed to use 2D frame-level features and 3D video-level temporal features to extract video features. The 2D features are extracted by a image-quality-aware CNN called PaQ-2-PiQ [189], while the 3D features are extracted with a 3D ResNet-18 [46] pre-trained on the Kinetics dataset. InceptionTime [36] is utilized as the regressor network, which is a CNN that is designed for time series regression. A follow-up work [157] uses the features extracted by multiple video quality assessment networks that are publicly available, and then uses the SVM as the regressor.

Both approaches successfully avoid re-scaling in CPA tasks, albeit with reduced model throughput or accuracy. For most pooling-based techniques, models without re-scaling can only be trained on a batch size of 1, thereby consuming a huge amount of time. For models that first extract features with a pretrained model, the training is not in an end-to-end fashion. The feature extractor is not fine-tuned on the target dataset, which might not always achieve acceptable performance. Alternatively, an input image can be directly split into grids and take each grid cell as a region [90, 134].

To obtain the final predictions from the regions, [90, 134] proposed to use graph convolutional operations to reason on the graph of ROIs. Each node in the graph represents an ROI, while each edge represents the similarity between two ROIs. Yang et al. [182] proposed to linearly combine global and local predictions as the final results. Rao et al. [123] proposed to concatenate the ROIs and global features and send them to a classifier to predict the results. Wei et al. [169] first performs prediction on each local region, then aggregates the predictions by combining both the results of score fusion and Top-K decision fusion.

Another line of research employs the attention mechanism [162] to implicitly mine local regions. The attention mechanism allows the inputs to interact with each other, and allocates its attention.
to the inputs [162]. The outputs are aggregates of these interactions and attention scores. The attention mechanism has been used in existing CPA tasks to identify and highlight important local regions of a given VM. Depending on the dimensions involved, the attention mechanisms for CPA models can be divided into spatial attention [23, 135, 179, 199] and channel-wise attention [199]. The spatial attention calculates an importance score for each vector in the spatial dimension, while the channel-wise attention generates a score for each channel.

3.2.2 CPA Model Optimization. Apart from the model architecture, model optimization methods are also crucial to CPA. The conventional optimization approach is to minimize the mean squared error (MSE) loss [31, 72, 136, 177] for regression tasks. To improve the MSE loss, Murray and Gordo [108] proposed the Huber loss to improve model robustness to outliers. Zhao et al. [199] proposed the polarity-consistent regression (PCR) loss for emotion regression. It assigns a penalty to the predictions of a sample that has opposite polarity to the ground truth. The penalty is applied to the loss value of each sample in the form of MSE. The cross-entropy loss is commonly adopted by works that treat CPA as a classification task [90, 100, 135]. Talebi and Milanfar [151] proposed the Earth mover’s distance (EMD) loss to improve classification performance. Compared to the cross-entropy loss, EMD accounts for the distance between categories. EMD is defined as the minimum cost to move the mass of one distribution to another [151]. In addition, instead of improving the MSE or the cross-entropy, an alternative is to use pairwise-based loss (e.g., triplet loss, ranking loss) to optimize the model [30, 72, 130].

Moreover, multi-task learning is adopted by [61, 72, 136] in which auxiliary training tasks are incorporated to provide extra supervision to the models, thereby enhancing CPA performance. Models reported in [61, 72, 136] are trained on the AADB dataset that provides scores of aesthetics-related attributes (e.g., colour harmony, vivid colour, good lighting) [72, 136]. The model jointly outputs the prediction of the aesthetics score as well as the attribute score.

3.3 Discussion
In this section, we introduced VM understanding methods for visual storytelling, including VM representation and cognitive property assessment (CPA). On one hand, well-learnt representations can enhance the central route of persuasion [119] by improving the informativeness of storytelling. On the other hand, an accurate CPA model can improve the peripheral route of persuasion [119]. In the next section, we further elaborate how these considerations make an impact on visual storyline generation.

4 VISUAL STORYLINE GENERATION
Visual Storyline Generation (VSG) approaches take the information of the VMs as inputs in order to optimize the selection and sequencing of the VMs to produce visual storylines. Based on the criteria adopted for VM selection and sequencing, we divided the existing VSG methods into three categories: (1) Spatial-Temporal VSG, which use the time and location information as criteria to perform VSG; (2) Data-Driven VSG, which attempt to train a VM selection or sequencing model based on data; and (3) Knowledge-Based VSG, which perform selection and sequencing of VMs based on relevant domain knowledge (e.g., plot analysis of written stories or film making principles).

4.1 Spatial-Temporal VSG
The timestamp information and location information related to VMs have been leveraged to provide intuitive storytelling structures by existing VSG approaches [16, 54, 173]. Spatial-Temporal
VSG approaches are well-suited to photos which have been taken by users while traveling. These approaches commonly involve two major steps: VM selection and VM sequencing.

4.1.1 VM Selection. The aim of the VM Selection step is to identify and discard low quality and duplicate VMs. Earlier works define low quality images as underexposed, overexposed or blurry photos [16]. Hua et al. [54] further considers the homogeneity among candidate VMs for selection. Based on [54], Wu et al. [173] further added sharpness and features of important patches as selection criteria. All three lines of works use a wavelet-based blur detection method reported in [155] to measure the degree of blur in VMs. In [54], underexposure, overexposure, and homogeneity are measured by color entropy (i.e., the Shannon entropy of normalized HSV color histogram). Nevertheless, these approaches can be computationally costly. Thus, more straightforward methods to measure the underexposure and overexposure by counting the number of dark and bright pixels in a photo have been proposed [16, 173]. If the intensity value of a pixel is higher than a designated threshold, it is considered as bright; if the value is lower than a predefined threshold, it is considered as dark.

Existing works have leveraged the above-mentioned evaluation metrics to design policies for filtering VMs. In [16], a threshold is set for each of these metrics. If any photo scores below any threshold, it is discarded. Hua et al. [54] adds two more metrics (i.e., color entropy and the degree of blur) into consideration. Photos with final scores below a predefined threshold are considered as low quality and removed. Wu et al. [173] treats these metrics as features, and uses an SVM to determine VM quality. The SVM is trained on a large image set with each image manually labeled as “good” or “bad” quality. In addition, [16, 54] use the temporal information to group the VMs. Then, for each time segment, a content-based clustering algorithm is used to generate clusters of VMs. The top K clusters are selected as the candidate VMs for generating the storyline. The importance of each cluster is also calculated by jointly considering the number of photos in the cluster and the within-cluster distance.

Another important consideration during VM selection is duplication filtering. For temporally adjacent photos, [16, 54] calculate the pairwise euclidean distance on the down-sampled images to determine if they are duplicates. Wu et al. [173] uses a feature extraction method [172] to obtain features and calculates pairwise similarity. If a pair has a small distance, the image with a lower quality score will be removed [16, 54, 173]. Wang et al. [163] proposed a graph-based solution to select representative images for building a visual storyline. A graph is first built in which each node represents an image with a caption, and each edge represents the similarity between a pair of images considering both visual and textual components. The VM selection task is then performed by minimum-weight dominating set finding in the graph.

4.1.2 VM Sequencing. Spatial-Temporal VSG approaches generally sequence the selected images according to their timestamps, while placing photos that are taken in close spatial proximity together [16, 54, 163]. Chen et al. [16] presents the photos in each cluster as a tiling slideshow. Each frame is a tiling slide defined by a template that contains several cells, with each cell containing one photo. The method dynamically determines the template for each cluster. The photo importance (PI) score and the template importance (TI) score have been proposed to provide more fine-grained control over the sequencing step. PI is calculated by considering the face region ratio and an attention value [101] near the center of an image. TI is defined as the ratio between the area of a cell over the entire area. The PI values for all photos in a cluster are sorted in descending order, and thus becomes a cluster feature. The template feature is obtained in a similar fashion. The template is eventually determined by finding the one that maximizes the cosine similarity between the cluster feature and template feature. After the template is selected, photos are placed...
in cells according to their importance scores (i.e., a photo that ranks $i$-th by its importance score is placed into the $i$-th cell).

4.2 Data-Driven VSG

Another line of research performs visual storyline generation using machine learning models trained on large-scale datasets.

4.2.1 Supervised Learning. Agrawal et al. [1] proposed a supervised learning solution to sort image-caption pairs into storylines. The proposed model takes a permutation of $n$ image-caption pairs as the input and outputs the appropriateness of the permutation. Specifically, a VGG19 model [142] is used to embed images and a GRU [20] trained on the BookCorpus dataset [69] is used to embed the captions. Both embeddings are concatenated and fed into an MLP. During training, only the parameters of the MLP layers are updated. The best permutation is found with the Hungarian algorithm [107].

In contrast, [141] proposed a variant of RNN, namely skip RNN, to construct a visual storyline in a bottom-up manner by iteratively predicting the following image to be placed in a storyline. To perform image selection, the skip RNN model takes a series of photos as the input. At step $t$, the features of the currently selected images are taken as the input, and the model outputs a feature $y$. Softmax operation is performed on the dot product of $y$ and features of subsequent images $[x_{t+1}, x_{t+2}, \ldots, x_n]$ to calculate the probability of the image being $x_j, j \in [t+1, n]$ to be the following image. In addition, the skip RNN can also be used in the sequencing step by predicting the next image that has not yet been selected in the output sequence. The model is supervised by the orders given by a set of Flickr Albums.

4.2.2 Unsupervised Learning. Unsupervised learning-based models have been proposed to perform the VM sequencing steps. They generally learn to order images or video frames by learning from many videos in which temporal orders are given naturally. The unsupervised models usually take a pair of elements $(i, j)$ as the input and predict whether $i$ should be placed before $j$. By utilizing the pairwise comparison information, a global order can be determined, and thus the sequence can be generated.

Existing works such as [53, 137, 176, 203] utilize a large-scale video dataset and divide each video into several clips. The task is to sequence the video clips by inferring their temporal dependency. Xu et al. [176] proposed a solution that first uniformly samples clips from the video. Then, a 3D CNN is used to extract the features of each clip. Finally, the ordering of the video clips is established by an MLP layer. The MLP layer takes the concatenation of embeddings of the video clip pairs as the input, and outputs features for each possible pair. The features are further concatenated as the final representation to produce the probability distribution over all possible sequences. The training loss is the cross-entropy, in which each possible permutation is considered a class.

In [176], the number of clips is a predefined hyperparameter. The approach can only support between two to five video clips, because a large number of clips leads to a huge number of permutations, thus making the classification very difficult. Siar et al. [137] proposed to solve this issue by grouping the frames. A total of $\binom{N!}{2}$ groups can be formed. During training, the classification layer is only required to choose one of the two possible permutations.

El-Nouby et al. [35] noticed that for actions like doing pull-ups, the reverse order of the pulling up frames in a video is also temporally plausible. To address the problem of label ambiguity caused by this problem, they proposed a solution called skip-clip to take contextual information into consideration to improve order prediction. The context is defined as a clip that is $r$ frames before the first target clip. During training, instead of using classification, a hinge rank loss is used, which
aims to ensure that the target clips that are temporally close to the context clip have high cosine similarity values.

### 4.3 Knowledge-Based VSG

A major limitation of data-driven VSG approaches is the heavy reliance on relevant data, which might not be readily available. Therefore, most existing data-driven VSG works leverage temporal visual storylines as the training data, which can be obtained relatively easily by cutting up existing videos into clips. For other VSG tasks that consider more than temporal relationships (e.g., visual coherency, persuasiveness, viewing experience), there is no publicly available dataset to support data-driven VSG approaches. Collecting such a dataset is difficult and expensive because the criteria involved tend to be subjective.

To address this issue, some works incorporate knowledge from other domains to perform visual storyline generation. In this section, we describe the goals, domain knowledge selected, objective function definitions based on such knowledge, and the optimization approaches of these knowledge-based VSG approaches.

#### 4.3.1 Domain Knowledge from Storytelling Research

Storytelling plot analysis has been incorporated into personal video clip sequencing [21, 201]. They propose that visual storylines should follow a common plot like other written stories. Such a plot template starts with exposition, then moves on to rising action followed by a climax, and finally reaches a resolution. The authors define this plot using the metric of dynamicity. At the beginning of storylines, the content dynamicity is relatively low. Then, it gradually increases until the climax. Finally, it decreases but still remains higher than the dynamicity at the beginning. The dynamicity is calculated as the average magnitude of the dense optical flow features [84]. In the context of persuasion, the dynamicity is useful when applying the peripheral route persuasion under the Elaborate Likelihood Model (ELM) [119].

In addition, Choi et al. [21] proposed to consider coherence during the sequencing step. The dissimilarity between two adjacent video clips is utilized to calculate coherence for a possible sequence. The final objective function is the linear combination of coherence and dynamicity. Zhong et al. [201] calculates coherence using two RNNs. The RNNs are trained in an unsupervised manner, which aims to select the next video clip from the set of clips that are not yet selected. The video clips are obtained in a similar way as the temporal-based data-driven VSG [176]. One of the RNNs takes the spatial pyramid pooling on the histogram of dense optical flow (SPP-HOOF) as the input. The other RNN takes the C3D feature of current video clips as the input. They each output a probability distribution for the next clip. Then, the two distributions are added together to form the final coherence vector.

#### 4.3.2 Domain Knowledge from Persuasive Theory

Liu et al. [83] attempts to generate visual storylines that can maximize persuasiveness. According to the research on persuasive advertising [8], informativeness, attractiveness, and emotion are key stimuli for achieving persuasion. In the ELM persuasion model [119], informativeness measures whether the provided information is sufficient for the central route of persuasion, while attractiveness and emotion are important for the peripheral route of persuasion. The authors propose to use dissimilarities between adjacent visual materials, aesthetics [151], and arousal scores to compute informativeness, attractiveness, and emotion, respectively. Both aesthetics and arousal scores have been described in Section 3.2.

A Learnable Wundt Curve (LWC), which is inspired by the Wundt Curve [11], has been proposed to fuse the scores and produce a final persuasive score. The Wundt Curve is a bell-shaped curve in which increases in persuasion-related stimuli first cause the persuasiveness of the videos to increase. Once these stimuli increase past certain values, further increasing them causes the
persuasiveness of the videos to decrease rapidly. Such a theoretical curve indicates the effect of over stimulation on viewers’ experience. Liu et al. [83] uses two transformed Sigmoid functions to translate the Wundt Curve into a machine learning model, in which the key parameters are learned via backpropagation. The LWC is also used in [86] to calculate persuasion scores.

4.3.3 Domain Knowledge from Film-making Principles. Liu et al. [86] incorporates film-making principles with the aim of enhancing the viewing experience of the generated videos. This can, in turn, improve the persuasiveness of the videos under the peripheral route of persuasion. Three principles are considered: (1) the visual storyline should start with a wide shot and gradually narrow to a close-up shot; (2) the visual storyline should have a logic, and thus be easy for viewers to follow; and (3) there should be graphic discontinuities between scenes to keep viewers excited. To implement these principles algorithmically, the VMs are first clustered. Each cluster is treated as a shot in the video. In this way, a preliminary logic flow for the visual storyline can be created. The objective function for the sequencing step is a linear combination of distance penalty and graphic discontinuity. The distance penalty considers two distance functions: (1) the Salient Region Ratio (SRR), which measures the distance in terms of attention, which is calculated by the attention map generated by the CNNs and SIFT; and (2) the semantic distance (SED), which measures the distance between the shots and the product which is being promoted in the video. It is measured by the text embedding distance between the predicted classes and the text description of the product. The distance penalty function produces a penalty when the SRR decreases and the SED increases over time. By minimizing the penalty, Principle 1 can be achieved. The graphic discontinuity is represented by the cosine dissimilarities between features of adjacent shots.

Finding a sequence that minimizes the objective function in VSG is an NP-hard problem. In [21], a branch-and-bound searching approach is proposed, which reduces the search space by updating the lower and upper bound scores for each subspace, and removes all subdivided sub-spaces that the lower bound is larger than the upper bound. Zhong et al. [201] proposes a sub-modular ranking method, which iteratively selects the next video clip that shares high coherence with other clips while increasing the dynamics. Liu et al. [83] introduces a dynamic programming method, which formulates the searching process as a group backpacking problem. Each group is considered as a cluster of VMs, and the items in a group are a sub-sequence with a specific length. The goal is to maximize the persuasiveness within a predefined video length limit.

4.4 Discussion

Visual Storyline Generation selects and sequences the visual materials to form the storyline of the video, which is the most important step in AIPVG. The Spatial-Temporal VSG approaches mainly use meta-data of visual materials to select and order them. In the context of E-commerce product promotion, it is more reasonable to use visual content instead of the time and location information in the images or video clips to form the storylines, as the main purpose of the generated videos is to persuade viewers. In E-commerce, the sellers always edit, crop, and concatenate photos before putting them online. E-commerce platforms also delete the EXIF data to minimize the size of files. These factors make the date and location information of the VM unusable and render the Spatial-Temporal VSG approaches ineffective. Data-driven VSG builds DNN to select and order the storyline based on the contents of visual materials. The DNNs are learned from data by backpropagation. The data-driven solution can be a promising direction. However, current data-driven works focus more on inferring temporal relations between video clips. No dataset about persuasive storylines is publicly available, which hinders the research and deployment of data-driven VSG for persuasion. The knowledge-based VSG approaches use heuristics (i.e., domain knowledge created by humans), which can be more readily deployed in
real-world systems. However, the performance of such approaches may not be as good as the data-driven VSG methods. In addition, they are not well suited for personalizing storylines due to limited learning capability. An experimental comparison between different methods can be found in [88].

5 POST-PRODUCTION

The post-production phase often follows the visual storyline generation phase with the aim to further enhance the overall viewing experience of the resulting videos. In this phase, edits on the selected visual materials, including generating visual effects and background music, are often performed. Finally, the videos are rendered using software such as MoviePy\(^2\) and VidGear.\(^3\)

5.1 Music Recommendation & Generation

Music recommendation approaches aim to recommend background music from existing music repositories based on video contents. This is generally performed as a feature extraction and matching task between videos and music. The approach in [74] first extracts features from videos and music (e.g., Saturation Proportion and Contrast for videos, Beat Histogram and Loudness for music). Then, it performs K-means to discretize the features and mine the correlations between discretized video and music features using a modified version of Multiple-type Latent Semantic Analysis (M-LSA)\(^{[166]}\). The recommendations are generated by finding the nearest neighbours of the projected video features from the database of projected music features. Liu and Chen\(^{[89]}\) proposed to extract low-level features for both videos and musics, paying special attention to the emotional undertones of the video contents and the musics. The relevance scores are then calculated for recommendation, based on a custom designed emotion-aware scoring function. In\(^{[79, 80]}\), visual features are extracted by averaging the frame features produced by a pretrained VGG16 model. A multi-task learning method is proposed to train a network that jointly generates the music features and analyzes the emotions and styles of the musics. Music retrieval is then performed by a regression model, which computes the similarities between the generated music features and the musics in the repository. The Cross-modal Variational Auto-encoder (CMVAE)\(^{[187]}\) constrains the latent variables of the musics and videos while jointly generating these latent variables. An embedding for each video or music is generated by the encoder of CMVAE. Finally, the model recommends musics by calculating the similarity scores between the music and video embeddings. Suris et al.\(^{[149]}\) proposed to model the music recommendation as a video-to-audio retrieval problem. Music and video embeddings are extracted separately by modality-specific Transformers that are self-supervised by InfoNCE, then the Top-K retrieved audio is recommended to users. Recommending musics from repositories risks infringing music copyrights. Thus, another line of research attempts to address this issue by training DNNs to generate musics based on given video contents. In\(^{[40, 146, 147]}\), approaches have been proposed to generate musics for silent videos containing a musician playing a musical instrument. However, the application scenarios of such techniques are not well aligned with promotion video generation in e-commerce. Di et al.\(^{[29]}\) proposed to generate background music based on given video contents. During training, a transformer is used to analyze the attributes of Musical Instrument Digital Interface (MIDI) musics. The attributes are defined by compound words. The transformer decoder uses the sequence of words, along with beat timing information, to learning to re-generate the input MIDI musics. During inference, the attributes of input videos (e.g., motion speed, motion saliency, and timing) are also denoted by the compound words heuristically. Then, the trained decoder outputs the music.

\(^2\)zulko.github.io/moviepy.
\(^3\)abhitronix.github.io/vidgear.
### Table 3. Comparative Analysis of Video-based Pose Transfer

| Authors          | Tai-Chi | Tai-Chi-HD | Venue   | Year |
|------------------|---------|------------|---------|------|
| Tulyakov et al.  | 0.27    | -          | CVPR    | 2018 |
| Siarohin et al.  | 0.17    | 0.228      | CVPR    | 2018 |
| Wiles et al.     | -       | 0.272      | ECCV    | 2018 |
| Siarohin et al.  | -       | 0.179      | NeurIPS | 2019 |
| Siarohin et al.  | -       | 0.152      | CVPR    | 2021 |
| Tao et al.       | -       | 0.148      | ECCV    | 2022 |
| Tao et al.       | -       | 0.146      | CVPR    | 2022 |

Average Euclidean Distances (AED) (i.e., the mean value of euclidean distances between the ground truth and generated frame features. Features are extracted by [138]) [139] on Tai-Chi [138] and Tai-Chi-HD [139] dataset are reported. Works are sorted by chronological order.

### 5.2 Still Image Animation

Still image animation aims to create a short video clip with realistic motions based on given image(s). Mathieu et al. [103] proposed to predict future images from a video sequence. A multi-scale architecture is designed, in which the images are fed into a network with size $k/2$. Then, the outputs are aggregated with the result produced by the network with size $k$. Zhao et al. [197] extended this idea and proposes a two-stage generation framework in which videos are generated from structures and then refined by temporal signals. Shaham et al. [132] proposed SinGAN, which performs single image animation using a pyramid of GANs. Each level takes the output of the previous level, as well as a random noise as the input to produce refined images with higher resolutions.

Other works focus on animating images involving human bodies. Table 3 gives a performance overview of existing works on video-based pose transfer on Tai-Chi and Tai-Chi-HD datasets. Chan et al. [14] proposed a model for video-based pose transfer. It leverages a reference video that shows a person performing an action, and a photo with the target person. The model transfers the reference moves to the target person to generate animations. The model extracts the pose from the reference video as a sequence of skeleton images, then computes the pose-to-appearance mappings to generate the frames of the target person for each skeleton image. Temporal coherence is considered when training the generator. Siarohin et al. [138, 139] achieves video-based pose transfer by the proposed Monkey Net. It comprises a keypoint detector for generating sparse keypoints, a dense motion network that generates optical flow and an encoder-decoder style motion generator that uses the keypoints and optical flow to generate the synthesized frames. The output video clip is the sequence of the generated frames. Siarohin et al. [140] propose to detect regions instead of keypoints in the reference videos to achieve a more accurate detection. Furthermore, a method for decoupling the foreground from the background is proposed to remove the foreground introduced by a region-based solution. Yoon et al. [190] further proposed to improve pose-based human image animation by reducing the visual artifact of the output videos. Sub-networks are trained to explicitly predict silhouettes, garment labels, and textures with synthetic data. To reduce artifacts in the generated videos, [153] proposed to use the transformer model to explicitly mine the relation between different frames in the reference video. The enhanced motion information is then sent to an animation generator, combining the occlusion map and optical flow to generate the output frames.

### 5.3 Discussion

The post-production step improves the attractiveness of the AI generated promotional videos. Music generation and recommendation enhances the sense of immersion, while still image animation
enhances the dynamism of the videos. Together, they serve to enhance the peripheral route of persuasion under ELM.

However, excessive use of post-production techniques might negatively affect the persuasiveness of the generated videos under the central route of persuasion under ELM. Improper still image animation might result in incorrect information being presented to the viewers as the animation transformation is learned from the training data. For example, for a new product, some information might not be present in the input image, but is required to generate informative video clips. This makes still image animation ambiguous, reducing the informativeness of the video. An alternative solution that can enable the post-production step to enhance the central route of persuasion is to extract selling points from the VMs or descriptions by summarization or image/video captioning techniques, then explicitly present this information as voice-over or captions in the generated videos.

6 EVALUATION

In this section, we discuss the techniques and methods for evaluating AIPVG approaches. Specifically, we introduce the evaluation metrics and datasets for representation learning and cognitive property assessment. We also summarize online testing methods for evaluating the generated storylines, the post-production approaches, and the final generated videos.

6.1 Evaluating VM Representations

The quality of the VM representations can be assessed individually using a labelled test set. The evaluation is typically conducted in a retrieval fashion. In this section, we first introduce the publicly available datasets for VM representation learning, then discuss the evaluation metrics.

6.1.1 Datasets. The representation learning datasets for e-commerce product VMs mainly contain furniture and clothing. Table 4 summarizes the publicly available datasets for evaluating VM representation models in the context of E-commerce. The Stanford Online Products dataset [113] contains images of furniture items on eBay. Others [41, 45, 93] provide clothing photos taken by consumers and sellers. Generally the photos from sellers are of better quality than those from consumers. Thus, the datasets can be used to assess the VM representations under different image qualities. Street2Shop dataset [45] contains photos from ModCloth,[4] while DeepFashion [93] contains images collected from Forever21[5] and Mogujie.[6] All of them are E-commerce platforms selling clothing. MovingFashion [42] is a publicly available video dataset for clothing items. Each video is associated with a distinct image of a shop, and is retrieved from Net-A-Porter,[7] Instagram, or Tik Tok.

Table 4. Publicly Available Products Datasets for Evaluating VM Representations

| Dataset                  | Instance       | # image/video | # class |
|--------------------------|----------------|---------------|---------|
| Stanford Online Products | Furniture      | 59,551        | 11,318  |
| DeepFashion [93]         | Clothing       | 239,557       | 33,881  |
| DeepFashion2 [41]        | Clothing       | 224,114       | 45,417  |
| Street2Shop [45]         | Clothing       | >420,357      | 204,795 |
| MovingFashion [42]       | Clothing       | ~15,045       | 15,045  |

[4]https://modcloth.com/
[5]https://www.forever21.com/
[6]https://www.mogujie.com/
[7]https://www.net-a-porter.com/
6.1.2 Metrics. Features of each VM in the test set are extracted using the learned DNNs. Then, for each query VM in the test set, the $k$ nearest neighbours (NNs) are retrieved. The score is then calculated using the labels of $k$ NNs and that of the query VM. For instance, Recall@$k$ is calculated as the ratio of query VMs such that at least one of their $k$ NNs has the same label as that of the query VM.

Recall@$k$ is commonly adopted by VM representation research works as the evaluation metric \[87, 110, 121, 148, 165, 167\]. However, analysis has revealed the weaknesses of the Recall@$k$ \[110\]. Improper embedding space designs can lead to artificially high recall@$k$ scores. To address this issue, Musgrave et al. \[110\] has proposed R-Precision and Mean Average Precision at R (MAP@R) as alternative evaluation metrics. Denoting the total number of VMs belonging to the class of the query VM $y_q$ as $R$, the $R$ nearest neighbours can be retrieved. Suppose $r$ is the number of nearest neighbours that belongs to $y_q$. R-Precision is calculated as $\frac{r}{R}$. MAP@R considers the ranking of the correct retrievals. It is defined as mean average precision of the $R$ nearest neighbours for each query.

Another popular evaluation metric is Normalized Mutual Information (NMI) \[112, 113, 144\]. NMI, which is defined as the ratio between the mutual information and the square root of the product of entropy values between the ground truth labels and clustering results, is originally designed to assess clustering quality \[129\]. As DML-based VM representation learning approaches carry out clustering on the representations extracted by the learned CNN models, NMI has been adopted to evaluate the performance of such approaches.

Besides, \[110\] proposed a VM representation evaluation pipeline to fairly compare the DML losses and models employing existing datasets and metrics. A performance comparison of existing DML methods can be found in \[110\].

6.2 Evaluating Cognitive Property Assessment Models

The evaluation metrics of CPA models depends on the type of labels provided in the dataset. When the model is trained via classification, accuracy is commonly used as the main evaluation metric. While Spearman correlation coefficient (SPCC), mean absolute error (MAE), or mean squared error (MSE) are commonly adopted for evaluating the performance of regression-based CPA models. In this section, we focus on discussing the publicly available datasets for CPA evaluation.

There are a large number of publicly available datasets related to aesthetics, emotion, and memorability which can be used for CPA evaluation. Here, we only list large-scale datasets containing more than 10,000 images (Table 5) or more than 1,000 videos (Table 6).

6.2.1 Emotions. Many large-scale datasets are built for emotion assessment that aims to detect the induced emotion given an image/video, i.e., what emotion will be aroused after the viewer sees the image. Based on how the dataset defines emotions, the emotion assessment can either be a classification task or regression task.

Many datasets \[12, 60, 116, 161, 183, 184, 191\] contain emotion labels for each image/video. The emotion category system varies across datasets (e.g., Plutchik’s Wheel of Emotions \[120\] used by \[12, 60\], Ekman’s six emotions \[34\] in \[174, 184\], eight emotions defined by \[104\] are used for \[9, 58, 183, 191\], and Parrott’s wheel of emotions \[117\] in \[116\]). In terms of the annotation method, the majority uses crowdsourcing \[13\] to label the images \[9, 183, 191\]. Other datasets are built by extracting emotions from the comments/text/tags for the images \[12, 161, 184, 200\]. There are also works which skip annotation by directly using the emotion categories as keywords when querying the image search engine \[60, 116\]. In addition, Panda et al. \[116\] found that existing datasets contain significant biases. Thus, they have collected a large-scale web image dataset to
Table 5. Image-based Cognitive Property Assessment Dataset

| Task | # image | Source | # category | Annotation |
|------|---------|--------|------------|-----------|
| Emotion | The Flickr CC [12] | classification | 487,227 | Flickr | 24 | Extracted from image tag |
| Emotion | [184] | classification | 554,192 | Flickr | 36 | Emotion and text comments |
| Emotion | [200] | classification | 90,000 | Flicker and Instagram | 8 | Crowdsourcing |
| Emotion | T4SA [183] | regression | 1,012,901 | Flickr | N.A. | Extracted from text |
| Emotion | LDL [183] | classification | 20,745 | Flicker and Twitter | 8 | Crowdsourcing |
| Emotion | Cgna [66] | regression | 10,766 | Flickr | N.A. | Crowdsourcing |
| Emotion | LUCFER [9] | classification | 3,605,101 | Bing Search | 8 | Crowdsourcing |
| Emotion | WEBEmo [116] | classification | 268,000 | Multiple Websites | 354 | Queries |
| Emotion | T4SA [181] | classification | 1,473,394 | Twitter | 3 | Extracted from comments |
| Emotion | MVSO [60] | classification | 7,365 | Flickr | 8 | Queries + Crowdsourcing |
| Aesthetic | AROD [130] | regression | 380K | Flickr | N.A. | Crowdsourcing |
| Aesthetic | AADB [72] | regression | 10,000 | Flickr | N.A. | Crowdsourcing |
| Aesthetic | AWA [109] | regression | 255,530 | DPChallenge.com | N.A. | Multiple Experts rating |
| Aesthetic | CUHK-PQ [152] | classification | 17,690 | professional photography websites | 2 | Crowdsourcing |
| Aesthetic | Flickr-AES [124] | regression | 40,000 | Flickr | N.A. | Crowdsourcing |
| Aesthetic | L2PA [30] | regression | 72,3M | Instagram | N.A. | Number of likes |
| Memorability | LaMem [63] | regression | 60,000 | various public datasets | N.A. | Crowdsourcing |

Table 6. Video-based Cognitive Property Assessment Datasets

| Dataset | Task | Avg. Length (Sec) | # video | Source | # category | Annotation |
|---------|------|------------------|---------|--------|------------|-----------|
| Emotion | VideoEmotion-8 [58] | classification | 107 | 1,101 | YouTube and Flickr | 8 & 24 | Crowdsourcing |
| Emotion | YF-Ed [174] | classification | 112 | 1,837 | YouTube and Flickr | 24 | Crowdsourcing |
| Memorability | Memento [111] | regression | 3 | 10,000 | Internet | 8 | N.A. | Crowdsourcing |
| Memorability | VideoMem [25] | regression | 7 | 10,000 | TRECVID and Hollywood-like movies | N.A. | Crowdsourcing |

minimize the effect of dataset bias. Jou et al. [60] attempts to understand the relationships between emotions and languages. Twelve languages are considered when collecting images on Twitter so that culture-specific and inherent linguistic contexts can be taken into account by the dataset.

The above-mentioned datasets are well-suited for image-based emotion classification. However, to be useful for visual storyline generation, the distinct class labels require further processing (e.g., embedding). Some datasets assign VAD scores for each image [66, 183, 200]. Zhao et al. [200] extracts these scores from the keywords in the texts related to the images (i.e., title, tags and descriptions) using the VAD values of 13,915 English lemma [168]. Extracting VAD values from text is a way to build large-scale datasets in an affordable manner. However, such an approach often sacrifices accuracy to some extent since user-generated texts might not always reflect their real emotions. Other datasets [66, 75] are built by assessing images using Self Assessment Manikin (SAM) [15]. The subjects are either hired offline [75] or employed via online crowdsourcing [66]. The VAD scores are very useful for visual storyline generation algorithms. They can be directly fed into a neural network or linear regression models to predict a final score for each image/video. However, the scales of existing datasets containing VAD information tend to be small.

6.2.2 Aesthetics. Aesthetics scores reflect how beautiful an image is. Murray et al. [109] collected the AVA dataset, in which images and the corresponding aesthetics ratings are from DPChallenge.com. It is a digital photography contest website, where professional photographers provide ratings on user uploaded photos. Each image has 210.13 annotations on average, with a standard deviation of 61.51. Other datasets [72, 152] employ crowdsourcing workers to rate image aesthetics in the datasets. Datasets such as [30, 130] use the like rates (i.e., #likes divided by #views) on Flickr as the aesthetics labels. Schwarz et al. [130] considers the score distribution to ensure that the dataset follows a uniform distribution instead of normal distributions in other datasets [72, 109]. In addition, the Flickr-AES dataset [124] also provides worker IDs for each annotation, which enables the personalized aesthetics score estimation.
6.2.3 Memorability. The memorability score reflects how many viewers can remember an image/video after a certain period of time. A more memorable video can be more helpful in persuading a viewer to purchase a product in time to come [114]. A commonly-used method to collect memorability score is through visual memory games [65]. In such a game, a sequence of images are displayed to crowdsourcing workers, each of which is displayed for a short time (e.g., 500ms), with a certain short gap (800ms [65]) in between images. The task for a viewer is to press a button whenever he/she sees an image at he/she has seen before. Newman et al. [111] considers video memorability at different replay elapse time scales. VideoMem [23] provides short-term and long-term memorability annotations. They are obtained by repeating the videos after a few minutes (for short-term annotations) and 24–72 hours (for long-term annotations).

6.3 Online Testing

The importance of performing online testing on the generated promotional videos is two-fold. Firstly, online testing can directly reflect the impact of the generated videos on viewers’ purchasing behaviours. We refer it this as video-level assessment. Secondly, the e-commerce system administrators might want to compare two sets of videos generated by different algorithms to determine which algorithm to adopt. We refer to this as algorithm-level assessment. Online testing can be used to evaluate storyline generation and post-production.

The most accurate evaluation method is A/B testing in a real-world e-commerce platform [71]. The platform can use a small percentage of the production traffic for testing. Half of the testing traffic goes into bucket A which shows videos generated by algorithm a. The other half goes into bucket B which shows videos generated by algorithm b. Actual business metrics (e.g., add-to-cart click rate, number of orders per view, or GMV per view) of each bucket or video are recorded. By comparing the metrics, both video-level and algorithm-level assessments can be performed. Even though it can reflect the real-world performance of the video generation models, it is an expensive process and researchers may not have access to the A/B testing in real-world e-commerce platforms.

A more accessible way of performing A/B testing for researchers is through crowdsourcing platforms such as Amazon mTurk. This is commonly known as “pairwise comparison” [21, 83, 88, 141, 201]. A micro-task sent to the workers is illustrated in Figure 2. Two videos are presented to each worker. The videos are about the same product, but generated by different algorithms. After watching the videos, workers are asked to provide ratings according to a 5-point Likert scale [3] about whether the left hand side video is better than the right hand side video according to some predefined criteria (e.g., persuasiveness, attractiveness, logic flow). The performance of the video/algorithm is calculated by collecting and averaging the feedback from the workers. The pairwise comparison method might not reflect reality as accurately as A/B testing in e-commerce platforms as it is more subjective in nature [2].

7 PROMISING FUTURE RESEARCH DIRECTIONS

Through discussions in this review, it can be observed that the building blocks for each of the major steps required for AI-empowered automatic generation of promotion videos have been developed. The current state-of-the-art techniques can be used to generate videos that can serve the business objectives of e-commerce platforms to some extent. Nevertheless, research in this interdisciplinary field is still in its early stage. Many research problems remain to be solved. In this section, we highlight promising future research directions for AIPVG.

---

8 mturk.com/
7.1 Visual Material Understanding

7.1.1 Scaling Up Proxy-based Deep Metric Learning. As discussed, a large number of classes in the training dataset leads to a huge proxy matrix. Although it is not used for inference (i.e., representation extraction), it still incurs huge memory and computational costs. The pair-based loss reduces the computational costs by using real data pairs to train the model. However, it suffers from performance loss in some cases or datasets. For example, the proxy-based loss performs better on the Person Re-identification dataset [97, 98] than the pair-based loss. Movshovitz-Attias et al. [106] has enhanced proxy-based DML to only use $\frac{M^2}{2}$ proxies ($M$ is the number of classes), albeit at the cost of a severe drop in model accuracy. Enabling proxy-based DML training with a small number of proxies while maintaining model performance remains an open research problem.

7.1.2 Noise Resistant Deep Metric Learning. For large scale VM datasets, label noises are inevitable. Cleaning up labels manually is an expensive process. Figure 3 illustrates the label noises that may appear in a VM dataset. Within a class, some data can be wrongly labelled. VMs about a same object might also receive different labels. Label noises decrease the performance of DML models. However, existing research works on handling noisy labels are mostly designed for classification tasks [38, 145]. Although efforts to enable DML to work in the presence of label noises have emerged [87], the general approach of removing suspected noisy samples based on analyzing their average distances with clean data points still results in large accuracy drops when the noise ratio is high. DML models which are robust in the presence of label noises are still required in order to enhance the real-world applicability of VM understanding approaches.

7.1.3 Understanding and Extending Pairwise-Comparison Models in CPA. The high accuracy of pairwise comparison models [76] raises some interesting questions. Firstly, pairwise-comparison models are not commonly used in tasks like object recognition, as the state-of-the-art solutions generate predictions without explicit reference points. What is the reason behind the success of the pairwise-based methods in aesthetics estimation tasks? Secondly, given the fact that comparing a target image with all reference images is time-consuming, how to make the evaluation more efficient? Thirdly, existing approaches select the reference images based on human
Fig. 3. An example of label noise in an e-commerce product dataset for DML. Irrelevant images may appear in a class. VMs about a same product can be given different class labels due to mistakes in the data collection process. For example, multiple item IDs can refer to the same product. Thus, using item IDs in e-commerce platforms to label VMs might introduce such noises.

knowledge. Is is possible to perform this action without human intervention? Last but not least, can pairwise comparison models built for aesthetics estimation tasks improve the accuracy of emotion regression and other CPA tasks?

7.2 Visual Storyline Generation

7.2.1 Building Datasets of E-commerce Promotional Videos. Previous methods of visual storyline generation for promotional video production heavily relies on human priors [83, 88]. However, these priors may not always lead to the optimal sequence of VMs as the application scenarios change. Data-driven approaches can be more adaptable for such situations by taking user feedbacks into account. Specifically, user feedbacks can be in the form of user ratings on the storyline, or changes in purchasing behaviours after viewing the generated videos. Such user feedback signals can be leveraged to continually train machine learning models. These models can be generative models for generating visual storylines directly, or storyline rating models for estimating a score for a given visual storyline. In the latter case, a search algorithm is also required to find suitable storylines. All these require the existence of a large-scale visual storyline dataset related to e-commerce promotional videos. Currently, no such dataset is publicly available. In order to address this issue, more in depth open collaboration between e-commerce platform operators and the AI research community is necessary.

7.2.2 Personalized Visual Storyline Generation. As postulated by the Persuasion Knowledge Model (PKM) [39], to maximize persuasiveness, the persuader should adapt the persuasion attempt based on knowledge about the persuadee. In a real-world e-commerce platform, it is natural that different customers have different preferences. Existing AIPVG approaches mostly generate a fixed storyline for a given product for viewers. However, this approach does not account for differences in personal preferences. Personalized visual storyline generation can be a promising research direction to enhance the persuasiveness of the resulting videos. A possible approach is to incorporate user prior information into AIPVG by leveraging personalization techniques in content-based recommender systems [4, 26].

7.2.3 Human Crafted Visual Storyline Templates. Human crafted visual storylines, either by experts and crowdsourcing workers, can be useful for AIPVG. Despite being one of the easiest methods to implement in real-world e-commerce platforms, they have not received much research attention so far. In order for these approaches to be leveraged to enhance AIPVG, some important research problems must be tackled first. They include: (1) how to automatically evaluate the quality of the crowdsourced storylines for given products; (2) how to determine the suitability of the
crafted visual storylines for given products; and (3) how to determine the level of appeal to different viewers of each given visual storyline template?

7.3 Post-Production

7.3.1 End-to-End Background Music Generation. **Background musics (BGMs)** improve the sense of immersion for video viewers. BGM generation which considers video-music consistency is useful for AIPVG but a challenging research problem. One promising approach is to develop an end-to-end model training solution. Achieving this goal requires the following research problems to be tackled: (1) building datasets, which facilitates model training and offline evaluation (videos and soundtracks which hold Creative Common (CC) licenses are a promising source to explore); (2) model architectures, for which both accuracy and model throughput should be considered; and (3) optimization methods, which need to be custom designed to train such models effectively (e.g., adversarial training [115] could be useful).

7.3.2 Personalization in Subtitle Generation. Informative product descriptions are helpful for the customers to make purchasing decisions. The visual contents in promotional videos may be attractive, but key product information in the images/videos might be overlooked by the viewers when viewing the videos. This could be caused by the low resolution of generated videos, the small size of the devices used to display the videos, or viewers’ attention being diverted by other things when viewing the videos. One way to solve this issue is to explicitly present key information (e.g., product selling points) to the viewers. In this way, such information enhances viewers’ impressions about the product. The selling points can be presented as key words or short phrases in the form of subtitles, or as a sentence with voice-overs. Existing works in image/video captioning [32, 133] and textual selling point mining [44] can be useful starting points.

7.3.3 Personalization in Visual Effect Optimization. **Visual effects (VFX)**, such as zooming-in, zooming-out, and motion blurring, bring more dynamism to videos. A wide variety of visual effects can be implemented with the help of computer vision techniques (e.g., zooming into the product in an image using the product detection and segmentation [105]). However, the most suitable types of visual effects vary across different VMs, products, and viewers. New visual effect optimization approaches that can balance improvements in persuasion power with computational costs incurred are required.

7.4 Evaluation

7.4.1 Offline Video Persuasiveness Assessment. Currently, there is no methodology for evaluating the persuasiveness of videos without involving human viewers. This limits the advancement of video-based persuasion technology since user study-based evaluation is costly and time-consuming. To address this limitation, a video persuasiveness assessment dataset is required. Such a dataset should capture the diverse backgrounds of the viewers and their behaviours when rating the persuasiveness of different types of videos so that useful viewer behaviour phenotypes can be constructed. Based on such a dataset, persuasiveness evaluation metrics which are only related to video features need to be designed so that researchers can circumvent the need for involving human viewers when continually improving the persuasiveness of the generated videos.

8 CONCLUSIONS

In this paper, we provide a comprehensive survey on the artificial intelligence (AI)-empowered persuasive video generation (AIPVG) literature. We first introduce the theoretical foundations for many of such works, which is the persuasion theory. Then, we offer a unique taxonomy of the AIPVG literature which divides it into three steps: (1) visual material understanding; (2) visual
storyline generation; and (3) post-production. We analyse the design rationale, advantages, and limitations of the approaches under each of these three steps. In addition, we also introduce the metrics and datasets for evaluating the performance of AIPVG approaches. Through this survey exercise, we point out promising future research directions that are worth exploring. We hope that this survey can serve as a roadmap for engineers and researchers interested in the field of AIPVG to acquire insights and build useful applications.

REFERENCES

[1] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. 2016. Sort story: Sorting jumbled images and captions into stories. In EMNLP. 925–931.
[2] Herman Aguinis, Isabel Villamar, and Ravi S. Ramani. 2021. MTurk research: Review and recommendations. Journal of Management 47, 4 (2021), 823–837.
[3] Gerald Albaum. 1997. The Likert scale revisited. Market Research Society. Journal. 39, 2 (1997), 1–21.
[4] Poonam Bhatia Anand and Rajender Nath. 2020. Content-based recommender systems. Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries (2020), 165–195.
[5] Kenneth E. Andersen. 1971. Persuasion: Theory and Practice. Allyn and Bacon.
[6] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasilieios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. Proc. IEEE 109, 11 (2021), 1838–1863.
[7] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In CVPR. 5297–5307.
[8] J. Scott Armstrong. 2010. Persuasive Advertising: Evidence-based Principles. Palgrave Macmillan.
[9] Pooyan Balouchian, Marjaneh Safaei, and Hassan Foroosh. 2019. LUCFER: A large-scale context-sensitive image dataset for deep learning of visual emotions. In WACV. 1645–1654.
[10] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. SURF: Speeded up robust features. In ECCV. 404–417.
[11] Daniel E. Berlyne. 1960. Conflict, Arousal, and Curiosity. McGraw-Hill Book Company.
[12] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In ACM Multimedia. 223–232.
[13] Yueting Chai, Chunyuan Xiao, Baowen Sun, Yongqing Zheng, and Qingzhong Li. 2017. Crowd science and engineering: Concept and research framework. International Journal of Crowd Science 1, 1 (2017), 2–8.
[14] Caroline Chan, Shirky Ginosar, Tinghui Zhou, and Alexei A. Efros. 2019. Everybody dance now. In ICCV. 10.
[15] Guillaume Chanel, Julien Kronegg, Didier Grandjean, and Thierry Pun. 2006. Emotion assessment: Arousal evaluation using EEG’s and peripheral physiological signals. In International Workshop on Multimedia Content Representation, Classification and Security. 530–537.
[16] Jun-Cheng Chen, Wei-Ta Chu, Jin-Hau Kuo, Chung-Yi Weng, and Ja-Ling Wu. 2006. Tiling slideshow. In ACM Multimedia. 25–34.
[17] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. 2020. What comprises a good talking-head video generation?: A survey and benchmark. arXiv preprint arXiv:2005.03201 (2020).
[18] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan. 2020. Adaptive fractional dilated convolution network for image aesthetics assessment. In CVPR. 14114–14123.
[19] Junsoo Choi, Tae-Hyun Oh, and In So Kweon. 2016. Video-story composition via plot analysis. In CVPR. 3122–3130.
[20] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In CVPR, Vol. 1. 539–546.
[21] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, and MartinEngilberge. 2019. VideoMem: Constructing, analyzing, predicting short-term and long-term video memorability. In ICCV. 2531–2540.
[22] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. 2004. Visual categorization with bags of keypoints. In ECCV Workshops, Vol. 1. 1–2.
[23] Debashis Das, Laxman Sahoo, and Sujoy Datta. 2017. A survey on recommendation system. International Journal of Computer Applications 160, 7 (2017).
[24] Yashar Deldjoo, Markus Schledl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender systems leveraging multimedia content. Comput. Surveys 53, 5 (2020), 1–38.
AI-Empowered Persuasive Video Generation: A Survey

285:25

Jingwen Hou, Sheng Yang, Weisi Lin, Baoquan Zhao, and Yuming Fang. 2021. Learning image aesthetic assessment.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2017. Learning spatio-temporal features with 3D residual networks for action recognition. In CVPR Workshops.

Marian Friestad and Peter Wright. 1994. The persuasion knowledge model: How people cope with persuasion attempts. Journal of Personality and Social Psychology 21, 1 (1994), 1–31.

Paul Ekman and Wallace V. Friesen. 1971. Constancies across cultures in the face and emotion. Journal of Personality and Social Psychology 17, 2 (1971), 124.

Chuang Gan, Deng Huang, Peihao Chen, Joshua B. Tenenbaum, and Antonio Torralba. 2020. Foley music: Learning to generate music from videos. In ICCV.

Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: A survey. IEEE Transactions on Neural Networks and Learning Systems 25, 5 (2013), 845–869.

Jacques độc và Michel Verleysen. 2013. Classification trong hiện diện của nhiễu nhãn: Một bài nghiên cứu. IEEE Transactions on Neural Networks and Learning Systems 25, 5 (2013), 845–869.
[54] X.-S. Hua, Lie Lu, and H.-J. Zhang. 2006. Photo2Video-A system for automatically converting photographic series into video. *IEEE Transactions on Circuits and Systems for Video Technology* 16, 7 (2006), 803–819.

[55] Tanveer Hussain, Khan Muhammad, Weiping Ding, Jaime Lloret, Sung Wook Baik, and Victor Hugo C. de Albuquerque. 2021. A comprehensive survey of multi-view video summarization. *Pattern Recognition* 109 (2021), 107567.

[56] Hyeongnam Jang and Jong-Seok Lee. 2021. Analysis of deep features for image aesthetic assessment. *IEEE Access* 9 (2021), 29850–29861.

[57] Hervé Jegou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. 2010. Aggregating local descriptors into a compact image representation. In *CVPR*. 3304–3311.

[58] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. 2014. Predicting emotions in user-generated videos. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

[59] Andrew Jong, Melody Moh, and Teng-Sheng Moh. 2020. Virtual try-on with generative adversarial networks: A taxonomical survey. In *Advancements in Computer Vision Applications in Intelligent Systems and Multimedia Technologies*. IGI Global, 76–100.

[60] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. 2015. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *ACM Multimedia*. 159–168.

[61] Yueying Kao, Ran He, and Kaiqi Huang. 2017. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing* 26, 3 (2017), 1482–1495.

[62] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[63] Mahmut Kaya and Hasan Şakir Bilge. 2019. Deep metric learning: A survey. *Symmetry* 11, 9 (2019), 1066.

[64] Asifullah Khan, Anabia Sohail, Umme Zahoor, and Aqsa Saeed Qureshi. 2020. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review* 53, 8 (2020), 5455–5516.

[65] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *ICCV*. 2390–2398.

[66] Hye-Rin Kim, Yeong-Seok Kim, Seon Joo Kim, and In-Kwon Lee. 2018. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia* 20, 11 (Nov. 2018), 2980–2992.

[67] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suhu Kwak. 2020. Proxy anchor loss for deep metric learning. In *CVPR*. 3238–3247.

[68] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

[69] Ryan Kiros, Yukun Zhu, Russ R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *NeurIPS*. 3294–3302.

[70] Keunsoo Ko, Jun-Tae Lee, and Chang-Su Kim. 2018. PAC-Net: Pairwise aesthetic comparison network for image aesthetic assessment. In *ICIP*. 2491–2495.

[71] Ron Kohavi, Randal M. Henne, and Dan Sommerfield. 2007. Practical guide to controlled experiments on the web: Listen to your customers not to the hippo. In *KDD*. 959–967.

[72] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *ECVC*. 662–679.

[73] Alex Krizhevsky. 2014. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997* (2014).

[74] Fang-Fei Ku, Man-Kwan Shan, and Suh-Yin Lee. 2013. Background music recommendation for video based on multimodal latent semantic analysis. In *ICME*. 1–6.

[75] Peter J. Lang, Margaret M. Bradley, Bruce N. Cuthbert, et al. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention* 1, 39–58 (1997), 3.

[76] Jun-Tae Lee and Chang-Su Kim. 2019. Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. In *ICCV*. 1191–1200.

[77] Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. 2020. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE Transactions on Image Processing* 29 (2020), 3898–3910.

[78] Y.-G. Liu, J.-H. Liao, C.-W. Chang, and H.-L. Liu. 2019. Automatic music video generation based on emotion-oriented pseudo song prediction and matching. In *ACM Multimedia*. 372–376.

[79] Leida Li, Haiyang Zhai, Y. Wei, and H.-L. Liu. 2020. Automatic music video generation based on simulated soundtrack recommendation and video editing. In *ACM Multimedia*. 519–527.

[80] Rongcheng Lin, Jing Xiao, and Jianping Fan. 2018. NeXtVLAD: An efficient neural network to aggregate frame-level features for large-scale video classification. In *ECVC Workshops*. 0–0.

ACM Computing Surveys, Vol. 55, No. 13s, Article 285. Publication date: July 2023.
[82] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In ICCV. 2980–2988.

[83] Chang Liu, Yi Dong, Han Yu, Zhiqi Shen, Zhanning Gao, Pan Wang, Changgong Zhang, Peiran Ren, Xuansong Xie, Lizhen Cui, et al. 2019. Generating persuasive visual storylines for promotional videos. In CIRM. 901–910.

[84] Ce Liu, William T. Freeman, Edward H. Adelson, and Yair Weiss. 2008. Human-assisted motion annotation. In CVPR. IEEE, 1–8.

[85] Chang Liu, Han Yu, Yi Dong, Zhiqi Shen, Yingxue Yu, Yan Dixon, Zhanning Gao, Pan Wang, Peiran Ren, Xuansong Xie, et al. 2020. Generating engaging promotional videos for e-commerce platforms (student abstract). In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 13865–13866.

[86] Chang Liu, Han Yu, Boyang Li, Zhiqi Shen, Zhanning Gao, Pan Wang, Peiran Ren, Xuansong Xie, Lizhen Cui, and Chunyan Miao. 2021. Enhancing viewing experience of generated visual storylines for promotional videos. In ICME. IEEE, 1–6.

[87] Chien-Liang Liu and Ying-Chuan Chen. 2018. Background music recommendation based on latent factors and moods. Knowledge-Based Systems 159 (2018), 158–170.

[88] Dong Liu, Rohit Puri, Nagendra Kamath, and Subhabrata Bhattacharya. 2020. Composition-aware image aesthetics assessment. In WACV. 3569–3578.

[89] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. SphereFace: Deep hypersphere embedding for face recognition. In CVPR. 212–220.

[90] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks. In ICCV. Vol. 2. 7.

[91] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2016. Deep Fashion: Powering robust clothes recognition and retrieval with rich annotations. In CVPR. 1096–1104.

[92] David G. Lowe. 1999. Object recognition from local scale-invariant features. In ICCV, Vol. 2. 1150–1157.

[93] Jiwen Lu, Junlin Hu, and Jie Zhou. 2017. Deep metric learning for visual understanding: An overview of recent advances. IEEE Signal Processing Magazine 34, 6 (2017), 76–84.

[94] Xin Lu, Zhe Lin, Xiaoqiang Guan, Radomir Mech, and James Z. Wang. 2015. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In ICCV. 990–998.

[95] Hao Luo, Youzhi Gu, Xingyu Liao, Shengqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[96] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu. 2019. A strong baseline and batch normalization neck for deep person re-identification. In IEEE Transactions on Multimedia (2019), 1–1. https://doi.org/10.1109/TMM.2019.2958756

[97] Pei Lv, Jianqi Fan, Xin Nie, Weimeng Dong, Xiaoheng Jiang, Bing Zhou, Mingliang Xu, and Changsheng Xu. 2021. User-guided personalized image aesthetic assessment based on deep reinforcement learning. IEEE Transactions on Multimedia (2021).

[98] Shuang Ma, Jing Liu, and Chang Wen Chen. 2017. A-Lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In CVPR. 4535–4544.

[99] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. 2002. A user attention model for video summarization. In ACM Multimedia. 533–542.

[100] Long Mai, Hailin Jin, and Feng Liu. 2016. Composition-preserving deep photo aesthetics assessment. In CVPR. 497–506.

[101] Michael Mathieu, Camille Couprie, and Yann LeCun. 2016. Deep multi-scale video prediction beyond mean square error. In ICLR.

[102] Joseph A. Mikels, Barbara L. Fredrickson, Gregory R. Larkin, Casey M. Lindberg, Sam J. Maglio, and Patricia A. Reuter-Lorenz. 2005. Emotional category data on images from the international affective picture system. Behavior Research Methods 37, 4 (2005), 626–630.

[103] Shervin Minaee, Yuri Y. Boykov, Fatih Porikli, Antonio J. Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. 2021. Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[104] Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. 2017. No fuss distance metric learning using proxies. In ICCV. 360–368.

[105] James Munkres. 1957. Algorithms for the assignment and transportation problems. Journal of the Society for Industrial and Applied Mathematics 5, 1 (1957), 32–38.

[106] Naila Murray and Albert Gordo. 2017. A deep architecture for unified aesthetic prediction. arXiv preprint arXiv:1708.04890 (2017).
[109] Naiha Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In CVPR. IEEE, 2408–2415.

[110] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In ECCV. 681–699.

[111] Anéline Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. 2020. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In ECCV. 223–240.

[112] Hyun Oh Song, Stefanie Jegelka, Vivek Rathod, and Kevin Murphy. 2017. Deep metric learning via facility location. In CVPR. 5382–5390.

[113] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savaresi. 2016. Deep metric learning via lifted structured feature embedding. In CVPR. 4004–4012.

[114] Daniel James O’Keefe. 2016. Persuasion: Theory and Research. (2016).

[115] Zhaoping Pan, Weijie Yu, Xiaokai Yi, Asifullah Khan, Feng Yuan, and Yuhui Zheng. 2019. Recent progress on generative adversarial networks (GANs): A survey. IEEE Access 7 (2019), 36322–36333.

[116] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K. Roy-Chowdhury. 2018. Contemplating visual emotions: Understanding and overcoming dataset bias. In ECCV. 579–595.

[117] W. Gerrod Parrott. 2001. Emotions in Social Psychology: Essential Readings. Psychology Press.

[118] Richard M. Perloff. 1980. A general psychoevolutionary theory of emotion. In Theories of Emotion. Elsevier, 3–33.

[119] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. 2019. SoftTriplet loss: Deep metric learning without triplet sampling. In ICCV. 6450–6458.

[120] Tiansrong Rao, Xiaoxu Li, and Min Xu. 2020. Learning multi-level deep representations for image emotion classification. Neural Processing Letters 51, 3 (2020), 2043–2061.

[121] Tiansrong Rao, Xiaoxu Li, Haimin Zhang, and Min Xu. 2019. Multi-level region-based convolutional neural network for image emotion classification. Neurocomputing 333 (2019), 429–439.

[122] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. 2017. Personalized image aesthetics. In ICCV. 638–647.

[123] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In NeurIPS. Vol. 28. 91–99.

[124] Edmund T. Rolls et al. 2005. Emotion Explained. Oxford University Press, USA.

[125] Ruslan Salakhutdinov and Geoff Hinton. 2007. Learning a nonlinear embedding by preserving class neighbourhood structure. In Artificial Intelligence and Statistics. 412–419.

[126] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In CVPR. 815–823.

[127] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In CVPR. 815–823.

[128] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In CVPR. 815–823.

[129] Hinrich Schütze, Christopher D. Manning, and Prabhat K. Raghavan. 2001. Introduction to Information Retrieval. Vol. 39. Cambridge University Press Cambridge.

[130] Katharina Schwarz, Patrick Wieschollek, and Hendrik P. A. Lensch. 2018. Will people like your image? Learning the aesthetic space. In WACV. IEEE, 2048–2057.

[131] Tinumol Sebastian and Jiby J. Pathiyidam. 2015. A survey on video summarization techniques. Int. J. Comput. Appl 132, 13 (2015), 30–32.

[132] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. 2019. SinGAN: Learning a generative model from a single natural image. In ICCV.

[133] Himanshu Sharma, Mammoth Agrahari, Sujeet Kumar Singh, Mohd Firoj, and Ravi Kumar Mishra. 2020. Image captioning: A comprehensive survey. In 2020 International Conference on Power Electronics & Applications (EPA). 325–328.

[134] Dongyu She, Yu-Kun Lai, Gaoxiang Yi, and Kun Xu. 2021. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In CVPR. 8475–8484.

[135] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. 2018. Attention-based multi-patch aggregation for image aesthetic assessment. In ACM Multimedia. 879–886.

[136] Fatemeh Siar, Amin Gheibi, and Ali Mohades. 2020. Unsupervised learning of visual representations by solving shuffled long video-frames temporal order prediction. In ACM SIGGRAPH 2020. 1–2.

[137] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. Animating arbitrary objects via deep motion transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2377–2386.

[138] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. NeurIPS 32 (2019).
[140] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion representations for articulated animation. In CVPR. 13653–13662.

[141] Gunnar A. Sigurdsson, Xinlei Chen, and Abhinav Gupta. 2016. Learning visual storylines with skipping recurrent neural networks. In ECCV. 71–88.

[142] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs] (April 2015).

[143] Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In ICCV, Vol. 3. 1470–1470.

[144] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In NeurIPS. 1857–1865.

[145] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2020. Learning from noisy labels with deep neural networks: A survey. arXiv preprint arXiv:2007.08199 (2020).

[146] Kun Su, Xuilong Liu, and Eli Shlizerman. 2020. Audeo: Audio generation for a silent performance video. NeurIPS 33 (2020), 3325–3337.

[147] Kun Su, Xuilong Liu, and Eli Shlizerman. 2020. Multi-instrumentalist net: Unsupervised generation of music from body movements. arXiv preprint arXiv:2012.03478 (2020).

[148] Yifan Sun, Changmao Cheng, Yuhao Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In CVPR. 6398–6407.

[149] Jiale Tao, Biao Wang, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. 2022. Motion transformer for unsupervised image animation. In ECCV. Springer, 702–719.

[150] Dhuphong Tong, Mingjing Li, Hongjiang Zhang, and Changshui Zhang. 2004. Blur detection for digital images using wavelet transform. In ICME, Vol. 1. IEEE, 17–20.

[151] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell’Orletta, Fabrizio Falchi, and Maurizio Tesconi. 2017. Cross-media learning for image sentiment analysis in the wild. In ICCV Workshops. 308–317.

[152] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NeurIPS. 5998–6008.

[153] Dingding Wang, Tao Li, and Mitsunori Ogihara. 2012. Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. In AAAI. 1006–1013.

[154] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jinchao Zhou, Zhifeng Li, and Wei Liu. 2018. CosFace: Large margin cosine loss for deep face recognition. In CVPR. 5265–5274.

[155] Xu Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In CVPR. 5022–5030.

[156] Xuanhui Wang, Jian-Tao Sun, Zheng Chen, and ChengXiang Zhai. 2006. Latent semantic analysis for multiple-type interrelated data objects. In SIGIR. 236–243.

[157] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods 45, 4 (2013), 1191–1207.
[197] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. 2018. Learning to forecast and refine residual motion for image-to-video generation. In ECCV, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11219. Springer International Publishing, Cham, 403–419.

[198] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. 2014. Exploring principles-of-art features for image emotion recognition. In ACM MM. 47–56.

[199] Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. 2019. PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In ACM Multimedia. 192–201.

[200] Sicheng Zhao, Hongxun Yao, Yue Gao, Guiguang Ding, and Tat-Seng Chua. 2018. Predicting personalized image emotion perceptions in social networks. IEEE Transactions on Affective Computing 9, 4 (Oct. 2018), 526–540.

[201] Guangyu Zhong, Yi-Hsuan Tsai, Sifei Liu, Zhixun Su, and Ming-Hsuan Yang. 2018. Learning video-story composition via recurrent neural network. In WACV. 1727–1735.

[202] Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. 2017. Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition. In IJCAI. 3595–3601.

[203] Dimitri Zhukov, Jean-Baptiste Alayrac, Ivan Laptev, and Josef Sivic. 2020. Learning actionness via long-range temporal order verification. In ECCV. Springer, 470–487.

[204] C. Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In ECCV. 391–405.

Received 17 December 2021; revised 29 January 2023; accepted 10 March 2023