Aesthetics-Oriented Video Generation and Editing

Martin Nicolas Everaert
IVRL, IC, EPFL

***Abstract*—The average time spent watching online videos increases every year, across all demographics. Videos are more engaging and are shared twice as much as other types of media. However, making or editing such videos can be expensive and time-consuming. Our research goal is to propose solutions based on machine learning and computational aesthetics to automate steps in the creation and editing of videos that are appealing and of interest for the viewer.**

In this proposal, we discuss three existing works and how they relate to our research. We first examine how generative adversarial networks (GANs) can be used to generate videos and what are their limitations. Then, we take a look at an example of data collection and annotation process, allowing training of models for video aesthetics and message understanding. Finally, we discuss a framework to navigate GANs' latent space to improve aesthetics.

***Index Terms*—Computational aesthetics, generative models, generative adversarial networks (GANs), video generation, video understanding, aesthetics assessment, latent space navigation.**

I. Introduction

A. Motivations

WHY are we interested to automatically generate or edit videos that are appealing to the viewer? Recent surveys show that people spend on average 19 hours per week watching online videos [94], which account for 82% of all Internet traffic [88]. This duration appears to increase substantially and regularly every year [92, 94], including across older demographics [92]. Some marketing study states that videos are shared twice as much as any other type of media [94]. Today, when looking at the top-10 smartphone apps sorted by their estimated revenue, 4 apps directly relate to videos (YouTube, TikTok, HBO Max, and Disney+) [89]. In the last months, we have seen companies modifying their app to display videos more often, such as Instagram and Facebook. We have also seen companies making announcements of new video-related products, such as Google launching their Ads Creative Studio.

B. Research area

Motivated by those facts and the recent research progress in video generation [3]-[12, 33, 65, 81] and text-to-image generation [19, 20, 58, 59, 60, 61, 83]-[86, 87], we are interested in the topic of aesthetics-oriented video generation and editing. This topic is part of the field of computational aesthetics and borrows methods from many other fields, including generative models and automatic aesthetics assessments.

C. Organization of this write-up

To provide detailed background on the topic, we discuss below three existing works. Each of these works illustrates one main challenge of our research area, namely the challenge of automatic generation of videos (section II), the challenge of collecting aesthetics measurements and predicting them (section III), and finally the challenge of combining video generation models with aesthetics assessment models (section IV). The first work, by Munoz *et al.* [3], proposes a GAN architecture that can be used to generate videos, focusing on temporal coherency between consecutive frames. The second work, by Hussain *et al.* [2], introduces a dataset of video advertisements with aesthetics annotations, allowing experiments in training models for automatic message understanding in advertisements and computational aesthetics. The third work, by Goetschalckx *et al.* [1] proposes a framework to find, inside the latent space of generative models, directions that correspond to modifications of some aesthetics measures. Finally, in section V, we expose possible research directions for our current and future works.

II. Generative models for videos

In this section, we discuss the subject of video generation with deep learning techniques. We illustrate it with a work by Munoz *et al.*: temporal shift GAN (TSGAN) [3]. This work proposes an improvement on video-specific architectures of GANs. They cleverly integrate into their method the latest progress in image generation achieved by BigGAN [11] and confirm that a carefully-design treatment of the temporal dimension is required for temporal consistency in the generated video.

A. Problem statement and background

Overview of main architectures of generative models. Various architectures of artificial neural networks can be used for generative tasks. Those models can often be seen, in probabilistic terms, as learning the distribution of the data and providing a way to sample new data points from this distribution. We identify five main classes of architectures of generative models that are often used in the literature: synthesis through optimization, variational auto-encoders (VAEs), denoising diffusion probabilistic models (DDPMs), autoregressive models (AMs), and generative adversarial networks (GANs).

Synthesis through optimization consists in directly optimizing the image to maximize some objective, such as being classified as *a banana* by a pre-trained classifier. We want to mention here DeepDream [93], that popularized synthesis through optimization in 2015. Auto-encoders (AEs) are a class of architectures made of an encoder and a decoder, converting a data point (e.g. an image) into a latent code and *vice-versa*. The encoder and the decoder are trained jointly to minimize the reconstruction error. VAEs regularize the latent space of AEs, which allows generation using the decoder [44]. DDPMs consist of models that, in simple terms, are trained to gradually denoise data, allowing image generation by gradually denoising an image made of random noise [32, 66]. AMs are models that predict the next value of sequential data. They are widely used in natural language processing (NLP), predicting the next word/token of a text. AMs can be used iteratively for generative tasks, predicting words one after the other for text generation, or predicting the values of pixels one after the other for image generation [75]. GANs are made of a discriminator and a generator, which are jointly trained to distinguish real data from generated data, and to generate data that cannot be distinguished from real data, respectively [29].

Each of those five classes of architectures has its own advantages (e.g. realism, no training required, tractable likelihood computation) and drawbacks (e.g. slow sampling, difficult training). Those architectures can always be modified to obtain conditional generation, e.g. conditioning the generation to some prior constraints or conditioning to a specific class [51].

GANs for videos. Until recently [18], GANs were the state-of-the-art architecture for image generation, as seen in the breakthrough models such as Progressive GAN [36], StyleGANs [37, 38, 39, 40] and BigGAN [11]. Hence most progress in video generation was also made using GANs.

In 2016, Vondrick *et al.* [76] showed that 2D convolutions of image GANs can be replaced by 3D convolutions for video generation, resulting in a video GAN (VGAN). More precisely, VGAN uses a 2D convolutional neural network (a 2D CNN) to generate the background of a video, which is assumed to be constant in this work, and a 3D CNN to generate the time-dependant foreground video. Temporal GAN (TGAN) [62] separates the temporal generation from the frames generation inside the generator network. The video generator (3D CNN) is replaced by a sequence generator (a 1D CNN) followed by an image generator (a 2D CNN). The sequence generator produces a latent code for each frame of the video, and those latent codes are decoded independently by the image generator to produce the frames. In 2017, motion-content GAN (MoCoGAN) [72] added an image discriminator (a 2D CNN) in addition to the video discriminator (a 3D CNN). Through this image discriminator, one specifically aims at making frames extracted from generated videos look realistic. Their generator also includes a sequence generator (a recurrent neural network RNN) followed by an image generator (a 2D CNN), which they respectively interpret as motion generation and content generation. On another note, in 2018, progressive VGAN [4] showed that progressive training of GANs [36] is also effective for video generation. Starting from generation of 4-frame videos of 4×4 pixels, network blocks are gradually added to the generator and the discriminator during the training procedure of progressive VGAN, to generate 32-frame videos of 256×256 pixels at the end of the training.

Using datasets from action recognition for video generation. Since the distribution of natural videos in the video space is extremely complex, the existing works aim to generate data from the distribution of specific datasets. The training and evaluation are usually done on relatively small recognition datasets, on specific domains such as human actions, sports, or facial expression. For instance, the Weizmann [30] and UCF101 [67] datasets are human action and sport recognition datasets, with 93 videos and 9 classes for Weizmann, and 13K videos and 101 classes for UCF101.

Evaluation of generative models. To evaluate the quality of generative models, one assesses the quality of the generated samples. For image GANs, it is common practice to use the inception score (IS) [63] and the Fréchet inception distance (FID) [31]. Both metrics use a pretrained image classifier (Inception v3 [68, 69, 70]). In 2017, the TGAN paper [62] introduced the video inception score by replacing the Inception model by a pretrained video classifier. Similarly, in 2018, Unterthiner *et al.* [73] introduced the Fréchet video distance (FVD). We note that the pretrained video classifiers used for video IS and FVD are 3D CNNs trained for action recognition.

B. *Contributions of the paper*

The main contribution of the paper is TSGAN, which is an architecture improvement of existing GANs for video generation. Additionally, the authors of TSGAN also propose a new metric to evaluate generative models, as well as a new dataset allowing a semantic-oriented evaluation of video generation models.

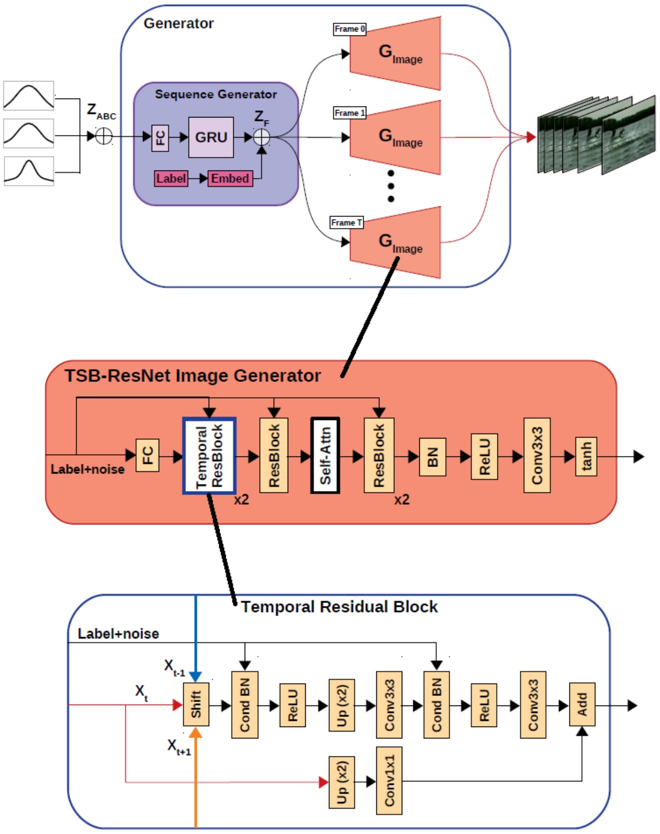

Fig. 1. Schema of the TSGAN architecture. *Image source from the original paper with slight modifications.*

TSGAN. The authors motivate their work by stating that 3D CNNs (such as VGAN and ProgressiveVGAN) treat temporal dimension same as spatial dimensions, require fixed-length videos, and are computationally and data expensive. Therefore, in the authors' work, the video generator is composed of a sequence generator followed by an image generator, in line with TGAN and MoCoGAN. However, in TGAN and MoCoGAN, the image generator processes frames completely independently, which induces low temporal consistency between frames.

Hence, the image generator should decode frames by sharing some intermediate representations between consecutive frames. The image generator of TSGAN is a BigGAN, initialized with pretrained weights, therefore including in TSGAN the latest progress in image generation. However, in two locations inside the BigGAN architecture, the intermediate representation of a frame t is replaced by a concatenation of the intermediate representations of the frames $t-1$, t and $t+1$. The resulting architecture is depicted in figure 1.

By evaluating their method, the authors showed that they boost video generation performance compared to the previous methods, and that this improvement is coming both from the integration of the recent progress of image generation (BigGAN backbone) and from the share of intermediate representations between consecutive frames.

Symmetric Similarity Score. The authors of TSGAN also noticed that using IS as metric for evaluation is problematic since significant changes in the IS did not match qualitative evaluation. Moreover, the IS fails to capture the intra-class diversity of the generated samples.

To address IS shortcomings, a new metric to evaluate generative models is introduced. This metric relies on training and evaluating a downstream model (e.g. a classifier) on a real dataset and a dataset generated by the generative model. It is based on two observations. Firstly, the performance on the generated dataset of a model trained on the real dataset is related to the realism of the generated dataset, but unrelated to the diversity of the generated samples. Indeed, if the evaluation samples (generated samples) are out-of-domain of the training data (i.e. nonrealistic), the performance will be low. Secondly, the performance on the real dataset of a model trained on the generated dataset is related to both the realism and the diversity of the generated samples. Indeed, if the generated samples lack diversity, the properties of some real samples will be underrepresented in the training data. Contrary to the IS, this accounts for both inter-class and intra-class diversity. Those two performances are normalized using the performance on the real dataset of a model trained on the generated dataset. They are then combined into a single metric. They found that this metric better matches their qualitative evaluation and is more reliable than IS, e.g. to identify per-class mode collapse.

MaisToys. This work also introduces a new dataset, MaisToys, which contains basic clay figure videos, combining 5 shapes, 4 colors and 4 motions. The size, nature and balance of the dataset is well suited for a semantic-oriented evaluation of an architecture, e.g. assessing whether a model trained on 3 of the 4 colors generalizes well to the 4^{th} color.

C. Discussion

Remaining challenges in video generation. The state of the research on video generative models is not yet suited for downstream applications such as generating a movie. Mostly, current models, including TSGAN, lack the ability to learn the full, large and complex, domain of natural videos. While TSGAN is designed to improve the temporal consistency of consecutive frames, the long and very long temporal consistency is not taken into consideration, in TSGAN as well as other models. Most existing works only produce low-resolution videos with a low framerate. Yet, we noticed very recent works showing promising results in those three directions. Last December, StyleGAN-V [65] showed major improvements on resolution, framerate, long-term coherency and quality of generated videos. Last May, CogVideo [33] showed the possibility to generate videos in the natural domain through large-scale training. Last June, a work by Brooks *et al.* [12] shows the possibility to generate videos with long temporal consistency.

Possible directions for progress in video generation. Among the subsequent works on video generation, we also want to mention here other directions of progress in designing generative models for videos. One direction is to include more priors into the generative models, leading for instance to "3D-aware" models [8] or "dynamics-aware" models [84]. Furthermore, we have seen recently significant progress in image generation and text-to-image generation. Many recent models/techniques do not rely solely on GANs to generate images: VQGAN [24], Craiyon [86] and Google Parti [83] mix AMs and GANs to generate images, VQVAE [74] and OpenAI DALL-E [59] mix AMs and VAEs, CLIPDraw [25] uses synthesis through optimization, CLIP-GLaSS [27] mixes synthesis through optimization and GANs, OpenAI DALL-E 2 [58], Google Imagen [61] and Stable Diffusion [60] use DDPMs, *etc*. The use of generative models that do not solely rely on GANs for generation also started to be used for video generation. For instance, VideoGPT [81] and CogVideo [33] mix AMs and VAEs to generate videos.

III. Aesthetics of videos: Data collection and prediction

In this section, we discuss the subject of collecting and predicting the aesthetics of videos.

We illustrate the subject with a work by Hussain *et al.* [2]. This work introduces two datasets of image and video advertisements as well as corresponding annotations for advertisements understanding, hence, in broader terms, for computational aesthetics.

A. Problem statement and background

This work relates to our research area for several reasons. First, the advertisement industry is one possible downstream application of the fields of computational aesthetics and video generation/editing. Secondly, it shows how complex assessment of appeal/aesthetics/effectiveness of images and videos might be. Indeed, it requires non-trivial understanding of which objects/styles are present, but also how they are shown and combined in order to transmit which message/emotion to the viewer. Thirdly, it gives an example of the data collection and human annotation procedures, which are necessary in many research works.

Hussain *et al.* do not mention the word *aesthetics* in their work. They described their work as a new problem of understanding the messages in ads and decoding their meaning. However, we use the term computational aesthetics in a broader manner. Indeed, message understanding is necessary to achieve high performance of assessing the appeal of videos (and images), to understand what provoke interest and how we can use it in videos, to understand how the viewer is engaged by the video. The fields of message understanding and computational aesthetics are really imbricated with each other. Understanding the meaning of videos rely on many intermediate steps including understanding sentiments and emotions that are induced, understanding humor. Therefore we highlight some background works of computational aesthetics in the next paragraphs.

Emotion models for image aesthetics. The aesthetics of an image correspond to the emotional/affective state experienced by someone visualizing that image. Hence, it can be described using emotion models. According to the literature, there are two main types of emotion models [85]. They consist in either a list of supposedly distinct basic emotions (categorical emotion models) [17, 21, 71] or a dimensional space where dimensions describe the characteristics of the emotions (dimensional emotion models) [10, 34, 56, 78]. Paul Ekman's model is one of the most used categorical emotion models. It lists six basic emotions: anger, surprise/shock, disgust, joy, fear, sadness/loneliness [21]. Most dimensional models use a *valence* dimension and an *arousal* dimension [56]. Valence describes how pleasant the emotion is and arousal describes how aroused/stimulated/awake one feels with this emotion. Dominance and memorability are also some common dimensions in the literature.

Datasets for image aesthetics. Many datasets of images with aesthetics annotations can be found [5, 10, 14, 15, 35, 42, 45, 46, 48, 52, 53, 55]. We note that those works use different taxonomies of annotations, e.g. categorical emotion models, dimensional emotion models, memorability, dynamics of image popularity, *etc*. We want to mention in particular the IAPS (from 1997) [46] and the OASIS [10, 45] datasets using a dimensional emotion model (valence, arousal, and dominance), the Cornell Emotion6 [55] and the ArtEmis [5] datasets using categorical emotion models, and the LaMem dataset [42] for image memorability.

Predicting image aesthetics. Those datasets were used to show that various aesthetics measures can be predicted using hand-crafted features [48], traditional computer vision features [35], low-level and high-level (CNN) features [41, 55], features extracted by visual-textual deep-learning models [43], *etc*.

Engagement of videos. The works focusing on videos usually refer to "engagement" indicating how much the viewer is appealed/interested by the video, which we consider as a dimension of the video aesthetics. The engagement of a video depends on video content factors (duration, style, information, *etc*.) and content-agnostic factors (upload time, popularity of content creator, advertising budget, *etc*.) [79]. Most works only focus on predicting the engagement induced by video content factors. The engagement of videos is usually a short-term engagement, derived using metrics as numbers of likes and views, watch time, *etc*. Concerning advertisement videos in particular, the study is more complex. For example, metrics like the view count are biased by paid advertising on the video hosting platform. Moreover, specific techniques are used in advertisements: humor, surprise, cartoon/animated characters, theme of the brand (colors, jingles, mascot characters), repetition, *etc*.

B. Contributions of the paper

The main contribution of the work from Hussain *et al.* [2] is the collection and annotation of datasets of advertisements. Additionally, they perform some experiments for automatic assessment tasks on those advertisements.

The construction of the advertisement datasets processes as follow. They first manually collected a list of keywords related to advertisements. They then used those keywords on search engines to collect noisy datasets of possible advertisements. They finally used Amazon Mechanical Turk to filter out the images/videos that are not ads and to precisely annotate the cleaned-up advertisements.

Collection of datasets. The authors shared the list of keywords related to advertisements in a supplementary material. It covers very broad topics (e.g. with the keywords "Food" "Electronics" "Publics service announcements"), finer topics (e.g. "Cookies" "Phones" "Domestic violence"), as well as precise brands (e.g. "Oreo" "Nokia"). They then use those keywords on Google Images and on YouTube, collecting 190K images and 5K videos of possible advertisements. They filtered out duplicates and disregarded low-quality data by removing videos with low view and like counts, as well as low-res images and videos. Among those collected images/videos, an important part is not actually advertisements. The clean-up of those non-ads data is done at the same time as the annotation procedure, asking the annotators whether it is an advertisement or not. However, in order to save cost on annotations of images, the authors trained a ads/non-ads classifier using a first batch of annotations and then did not send to the annotators the images classified as non-advertisement.

Annotations for datasets clean-up. The authors used Amazon Mechanical Turk to get their data annotated. They asked the annotators whether the image/video is indeed an advertisement. This allows to clean-up the datasets of possible advertisements, into datasets of confirmed advertisements. The final datasets contain 65K images and 3.5K videos of confirmed advertisements.

Aesthetics annotations. The authors asked the annotators whether the advertisement is effective and if it is funny/exciting (for videos only). Those annotations provide an aesthetics measure of the advertisement (e.g. effectiveness). However, aesthetics is also linked to the sentiments induced in the viewer, which itself also partly relates to the topics of the ads. The annotators had to select, for each advertisement, one topic in a list of 38 topics, and at least one sentiment in

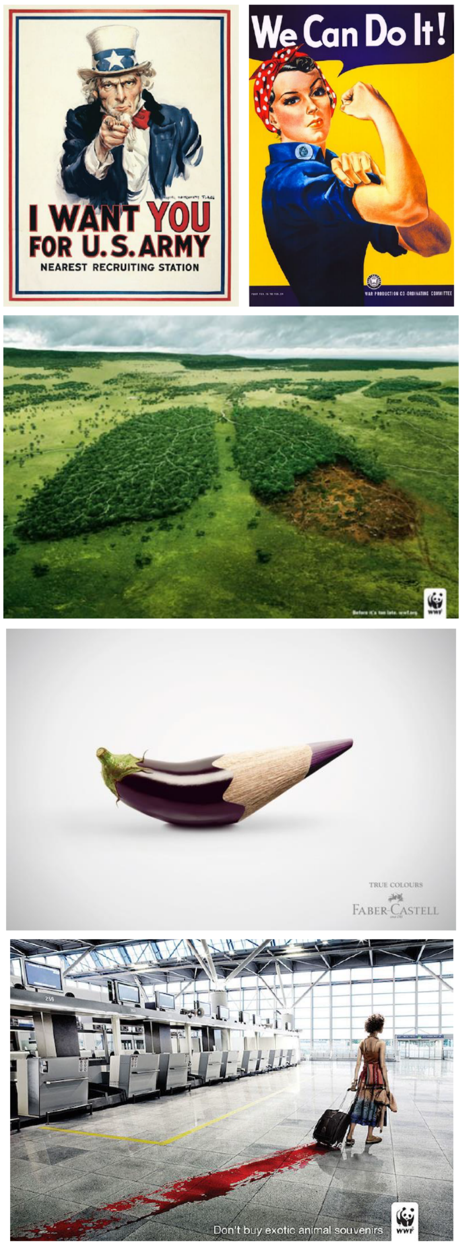

Fig. 2. Five image advertisements. The analysis of those ads is not straightforward as it requires reading the texts "I want you for U.S. army - Nearest recruiting station", "We can do it! - War production coordinating committee", "Before it's too late. - wwf.org", "True colours', "Don't buy exotic animal souvenirs". It also requires understanding the body language in the second ad, understanding that the pencil of the third ad will have an accurate color, a natural color to draw for instance an eggplant, understanding that the shape of lungs in the fourth ad symbolizes life through trees that turn carbon dioxide into oxygen, understanding that the blood in the fifth ads represents danger, injury and death, *etc*. Understanding all those concepts is required to understand the ads and correctly predict the reaction it induced to the viewer. *Image source from the original paper with slight layout modification.*

a list of 30. Those lists were obtained by clustering free-form answers of a first batch of annotations (where annotators had to type raw text instead of selecting among a list). Each image is annotated by several annotators and the authors checked the inter-annotator agreement. Annotations for message understanding were also collected, as motivated in figure 2. The annotators had to say what should they do, according to this ad, and why. For images only, they also had to label whether the understanding of the ad is straightforward or require non-literal interpretation (e.g. a gun might symbolize danger). In the case of non-literal interpretation, annotators were asked to describe the symbolism used in the advertisement (e.g. a gun is represented and symbolizes a danger) and the strategies that are used (e.g., references to cultural knowledge, use of humor, use of surprise).

Assessing aesthetics with machine leaning models. By analyzing the annotations, the authors found interesting relations between the topics of the ads and the induced emotions. For instance they noted that "domestic abuse and human and animal rights ads inspire disturbance and empathy", which makes a lot of sense. They then propose some baseline models for various prediction tasks. First, giving the model the image and the expected action, they aim to predict the answer to the following question: Why should you do [expected action] according to this ad? However, they showed low performance for this task, showing that message understanding is hard and complex. Secondly, they perform experiments on detecting which symbols are used in non-straightforward advertisements, with mitigated performances. Thirdly, they show that detecting topics and sentiments seems easier than the two previous experiments. Finally, in their last experiment, they show that they can reach good performance at predicting if a video advertisement is funny and if it is exciting.

C. Discussion

Alternatives to data annotations. Hussain *et al.* used human annotators to annotate their data. This was necessary as they wanted to focus specifically on message understanding, with some specific annotations such as *What should you do according to the ad and why?* or *Which symbolisms are used in the ad?*. However, there are other ways to collect ground-truth aesthetics annotations for videos. First, several existing studies on video engagement [13, 80] use metrics from video hosting platforms, as likes/dislikes/views count or watch time. This directly provides free annotation of the aesthetics/popularity/appeal of the video. Other works [7, 54] use more advanced tools (logs of HTTP requests and browser extension) to collect data such as the skip duration of the ad, the total session watch time, *etc*. Several works [6, 9, 77] also use electroencephalography (EEG) to study the engagement of video advertisements. Finally, cameras can also be used, for instance with eye tracking, subject tracking, or emotion recognition, to assess how engaged the viewers are and which emotions they feel.

Features to predict video aesthetics. The work from Hussain *et al.* proposes several baseline methods for predictive tasks, including predict how exciting/funny a video advertisement is. In this paragraph we mention some features that relate to the aesthetics/engagement of video. First, the advertisement creativity, originality and relevance impact its success [49, 64]. The video duration and pace have an influence on video engagement [79]-[90, 91]. Existing works use various features for this engagement prediction task: the video title, tags, description, category, and age [23, 47], the video topics [80], and the engagement of existing similar videos [16].

IV. Leveraging generative models for aesthetics

In this section, we discuss the work by Goetschalckx *et al.* [1]. They propose a framework, GANalyze, to leverage generative models and aesthetics assessment models.

A. Problem statement and background

The authors of GANalyze aim to study and represent visually some aesthetics measures (memorability, beauty, valence). Indeed those aesthetics measures do not have concrete and explicit definition. Hence, it is interesting to represent them visually, to get understanding of what those aesthetics measures actually are.

Using generative models for art generation. While the authors describe GANalyze as a framework to study visual definition of various aesthetics image assessments (memorability, beauty, valence), we point out that it can simply be used to make the generative models generate aesthetically pleasant images (and eventually aesthetically pleasant videos). Moreover, when combined with GAN inversion, this framework can lead to aesthetics-oriented image editing, by taking an image and making it more aesthetic.

Using generative models for art generation is not new however. Indeed in 2015, Google proposed a method, DeepDream [93], to visualize and try to understand the mechanisms and patterns learned by artificial neural network. It was often described as letting the deep neural network "dream", hence the name DeepDream. The images generated by DeepDream have an interesting psychedelic/hallucinogenic appearance and therefore could be considered as AI-generated art images. People wanting to generate art with AI also use style transfer to mix the content of an image with the style of an art piece [28]. Moreover, several works train a GAN on art datasets in order to generate new art. Creative Adversarial Network [22], for instance, did so, but in addition, they added a "stylistic ambiguity" term so that the generated art does not simply copy art, but generate new art that does not match existing art styles.

Semantics in latent spaces. Traditionally, deep neural networks are seen as black boxes, with hidden layers whose exact roles are hard to explain. However, some models such as VAEs are known to have better disentanglement than other models inside their latent space [50]. This makes the latent spaces easier to understand, to interpret, and to manipulate.

Several works, such as semantic face editing [82], aim at manipulating those latent spaces in a semantically-meaningful manner, in order to for instance modify one property of a generated face (e.g. suppressing the facial expression) or interpolate between two images.

B. Contributions of the paper

Proposed framework. The proposed framework, GANalyze, is relatively simple. It is represented schematically in the figure 3.

They learn a direction θ in a latent space. Moving in the direction θ in the latent space should increase the aesthetics measure of the decoded image, proportionally to the distance we moved. Similarly, moving in the opposite direction $-\theta$ should decrease the aesthetics measure.

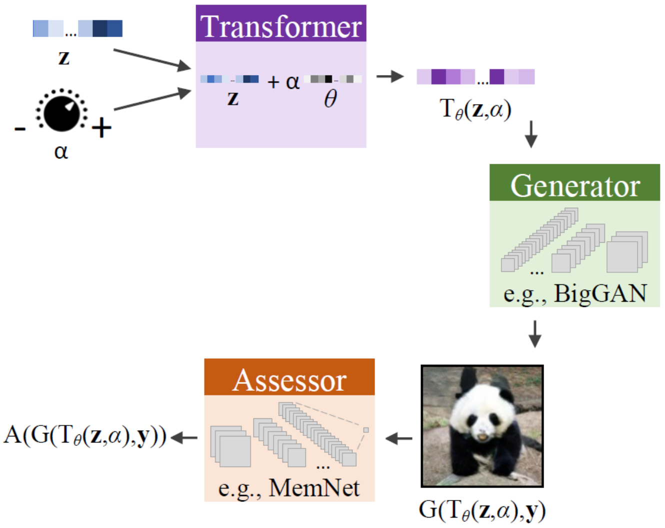
z
α
-
+
Transformer
z + α θ
$T_\theta(\mathbf{z},\alpha)$
Generator
e.g., BigGAN
Assessor
e.g., MemNet
$A(G(T_\theta(\mathbf{z},\alpha),\mathbf{y}))$
$G(T_\theta(\mathbf{z},\alpha),\mathbf{y})$

Fig. 3. Schema of the GANalyze framework proposed by Goetschalckx *et al. Image source from the original paper with slight layout modification.*

Their framework requires two trained models whose weights will not change during the optimization of θ. The first model is a differentiable generative model, which generates images $G(z)$ from points z of a latent space. The second model is a differentiable model that assesses some aesthetics measure $A(i)$ of images i, such as their memorability or their emotion. As the relation from latent space points z to aesthetics measurements $A(G(z))$ is differentiable, the framework can use the standard Adam optimizer.

At each iteration/sample of the training, a point z in the latent space and a small value α are drawn randomly. The MSE cost $((A(G(z+\alpha\theta)) - (A(G(z)) + \alpha))^2$ is used to take an optimization step for θ.

Performance vs realism. The authors of GANalyze showed by some automatic measures and human evaluation that the realness of the generated images is not much impacted by α. In other words, if we use their framework in order to generate only aesthetically pleasing images, then the generated images will still look as realistic as the one coming from the initial image generator.

Similarly, they checked numerically and through a visual memory game experiment that the aesthetics measure indeed increases with α.

Emerging factors. The authors mentioned that object size, subject centeredness, circleness over squareness, redness, brightness, image simplification (low number of objects) and colorfulness are emerging factors that appear when increasing the aesthetics of an image (memorability in this particular case).

C. Discussion

From images to videos. The paper itself focuses on images. However, their framework can be used with any GAN generator and aesthetics assessor that are compatible. Especially, it would be, in theory, possible to apply the GANalyze framework (section IV) to the video generator from TSGAN (section II) and the funny/exciting/sentiments prediction models for videos from Hussain *et al.* (section III), thus linking the three discussed papers together. While this would be computationally very intensive (and not recommended), this is possible from a theoretical perspective and would be one solution to achieve aesthetics-oriented video generation and editing. In section V, we discuss our current plan for aesthetics-oriented video generation and editing.

Leveraging generative models for aesthetics with text-to-image models. Text-driven image generation/editing models [25, 27, 58, 59, 60, 61, 83]-[86] can also be used to generate images with specific aesthetics. Adding words (modifiers) in the text query such as "unreal engine", "hyperrealistic" or "photorealist" appears to produce more pleasant images. Similarly, aesthetics/emotions related modifiers can be used, e.g. "calm", "happy", "angry", "depressed" in AffectGAN [26].

V. Research Proposal

In our first semester project we conducted an investigation into computational aesthetics. Our literature review and experiments covered important related topics such as emotion models, relevant low-level and high-level features, image emotions, aesthetics prediction, aesthetics-oriented image manipulation, and some emotion image datasets suitable for studying computational aesthetics. We have also performed preliminary experiments for a framework that should allow to visualize various emotion models (categorical and/or dimensional) and various emotion datasets in a shared 2D representation, similarly to Plutchik's wheel of emotions.

Most computational aesthetics works focus until now on the dominant feeling/emotion or on tasks that naturally are group-effect, e.g. popularity of videos assessed with the number of views. There are to our knowledge not many works and datasets targeting the aesthetics assessment of a specific population group or a specific individual, e.g. taking into account the past experiences of one individual to predict its reactions to an image. However it is of interest: especially for advertisements/movies we could focus on the aesthetics tastes of the targeted population. This might be studied in our future research as well.

In our second semester project, we handled the problem of estimating the engagement that advertisements videos reach online. We collect the videos of a Swiss company on YouTube and analyze the engagement (likes and views) of those videos by looking at features such as the duration of the video and the level of drama in the video. We train an engagement-prediction model on this new data and use it to find points in an interpretable feature space corresponding to new possible high-engaging commercials.

As following work of our second semester project, our current research focus on automatic generation of advertisements videos from these points of the interpretable feature space. Instead of directly generating videos, we are working on a two-stage process, firstly generating the corresponding script/screenplay and in a second stage, we will focus more on script-to-video generation. This two-stage process seems simpler and more effective to us. It also allows to have a human in the loop, modifying a few parts of the generated script if required, which is simpler than editing the video afterwise. Especially, we have started to leverage text generation NLP models for generation of scripts of advertisements. Due to the lack of datasets of advertisements scripts, our current experiments aim to fine-tune those text-generation models on both movies plot summaries and voice-overs of advertisements. To generate videos from the generated scripts, our plan is to leverage text-to-image generation models. More precisely, we plan, as a first prototype, to feed a text-to-image model with the sentences of the generated script one after the other.

One other direction we will study shortly is the classification of videos into ads/non-ads. Firstly, it would allow us to take any dataset of videos and filter only the videos that look like advertisements. Secondly, it could also be useful in latter work in order to make a video look more like an advertisement, e.g. in combination of frameworks similar to GANalyze if possible.

Our future research will also be heavily influenced by the advancements in the related challenges described in sections II, III and IV. Especially, many very recent works (CLIP [57], all the text-to-image models, CogVideo [33] for text-to-video) have shown progress that has major impact on our topic of aesthetics-oriented video generation and editing.

The 3 Discussed References

[1] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019. URL: https://arxiv.org/abs/1906.10112.

[2] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715, 2017. URL: https://arxiv.org/abs/1707.03067.

[3] Andres Munoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox. Temporal shift gan for large scale video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3179–3188, 2021. URL: https://arxiv.org/abs/2004.01823.

References

[4] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards high resolution video generation with progressive growing of sliced wasserstein gans. *arXiv preprint arXiv:1810.02419*, 2018.

[5] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. ArtEmis: Affective Language for Visual Art. *arXiv:2101.07396 [cs]*, January 2021. arXiv: 2101.07396.

[6] H Ang, G Sanchez, and J Pascual. Detecting interest in video advertisements using eeg data analysis. *Philipp. Inf. Technol. J*, 7(1):4–12, 2014.

[7] Mariana Arantes, Flavio Figueiredo, and Jussara M Almeida. Towards understanding the consumption of video-ads on youtube. *The Journal of Web Science*, 4, 2018.

[8] Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Hao Tang, Gordon Wetzstein, Leonidas Guibas, Luc Van Gool, and Radu Timofte. 3d-aware video generation. *arXiv preprint arXiv:2206.14797*, 2022.

[9] Sangeetha Balasubramanian, Shruti Shriya Gullapuram, and Abhinav Shukla. Engagement estimation in advertisement videos with eeg. *arXiv preprint arXiv:1812.03364*, 2018.

[10] Aenne A. Brielmann and Denis G. Pelli. Intense Beauty Requires Intense Pleasure. *Frontiers in Psychology*, 10:2420, 2019.

[11] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[12] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A Efros, and Tero Karras. Generating long videos of dynamic scenes. *arXiv preprint arXiv:2206.03429*, 2022.

[13] Jean Burgess and Joshua Green. *YouTube: Online video and participatory culture*. John Wiley & Sons, 2018.

[14] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision Research*, 116:165–178, November 2015.

[15] Ganesh Chandrasekaran, Naaji Antoanela, Gabor Andrei, Ciobanu Monica, and Jude Hemanth. Visual Sentiment Analysis Using Deep Learning Models with Social Media Data. *Applied Sciences*, 12(3):1030, January 2022. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[16] Yen-Liang Chen and Chia-Ling Chang. Early prediction of the future popularity of uploaded videos. *Expert Systems with Applications*, 133:59–74, 2019.

[17] René Descartes. *Les passions de l'âme*. 1728.

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[19] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[20] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

[21] Paul Ekman. Emotions revealed. *BMJ*, 328(Suppl S5):0405184, May 2004.

[22] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017.

[23] Ramalakshmi Eliganti, A Reddy, et al. Youtube data analysis & prediction of views and categories. *Sharvani, YouTube Data Analysis & Prediction of Views and Categories (April 6, 2022)*, 2022.

[24] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[25] Kevin Frans, Lisa B Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*, 2021.

[26] Theodoros Galanos, Antonios Liapis, and Georgios N Yannakakis. Affectgan: Affect-based generative art driven by semantics. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 01–07. IEEE, 2021.

[27] Federico A Galatolo, Mario GCA Cimino, and Gigliola Vaglini. Generating images from caption and vice versa via clip-guided generative latent space search. *arXiv preprint arXiv:2102.01645*, 2021.

[28] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[30] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–2253, 2007.

[31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[33] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[34] Kamil K. Imbir. Affective Norms for 4900 Polish Words Reload (ANPW_r): Assessments for Valence, Arousal, Dominance, Origin, Significance, Concreteness, Imageability and, Age of Acquisition. *Frontiers in Psychology*, 7:1081, 2016.

[35] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1469–1482, 2013. Publisher: IEEE.

[36] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[37] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

[38] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[39] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[40] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[41] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web - WWW '14*, pages 867–876, Seoul, Korea, 2014. ACM Press.

[42] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, pages 2390–2398, 2015.

[43] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. The Hateful Memes Challenge: Competition Report. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, pages 344–360. PMLR, August 2021. ISSN: 2640-3498.

[44] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[45] Benedek Kurdi, Shayn Lozano, and Mahzarin R. Banaji. Introducing the open affective standardized image set (OASIS). *Behavior research methods*, 49(2):457–470, 2017. Publisher: Springer.

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

[47] Yuping Li, Kent Eng, and Liqian Zhang. Youtube videos prediction: Will this video be popular. *URL: http://cs229. stanford. edu/proj2019aut/data/assignment_308832_raw/26*, 647615, 2019.

[48] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia - MM '10*, page 83, Firenze, Italy, 2010. ACM Press.

[49] Andreea-Ioana Maniu and Monica-Maria Zaharie. Advertising creativity–the right balance between surprise, medium and message relevance. *Procedia Economics and Finance*, 15:1165–1172, 2014.

[50] Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412. PMLR, 2019.

[51] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[52] Wyverson Bonasoli de Oliveira, Leyza Baldo Dorini, Rodrigo Minetto, and Thiago H. Silva. OutdoorSent: Sentiment Analysis of Urban Outdoor Images by Using Semantic and Deep Features. *ACM Transactions on Information Systems*, 38(3):1–28, June 2020.

[53] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. Predicting Social Image Popularity Dynamics at Time Zero. *IEEE Access*, 7:171691–171706, 2019. Conference Name: IEEE Access.

[54] Minsu Park, Mor Naaman, and Jonah Berger. A data-driven study of view duration on youtube. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 651–654, 2016.

[55] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion

distributions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868, Boston, MA, USA, June 2015. IEEE.

[56] Jonathan Posner, James A. Russell, and Bradley S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, September 2005. Publisher: Cambridge University Press.

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[58] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[59] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[61] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[62] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, pages 2830–2839, 2017.

[63] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[64] Daniel A Sheinin, Sajeev Varki, and Christy Ashley. The differential effect of ad novelty and message usefulness on brand judgments. *Journal of Advertising*, 40(3):5–18, 2011.

[65] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022.

[66] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[67] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[68] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[69] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[70] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[71] Silvan Tomkins. *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company, 1962.

[72] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

[73] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[74] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[75] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016.

[76] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016.

[77] Regina WY Wang, Yu-Ching Chang, and Shang-Wen Chuang. Eeg spectral dynamics of video commercials: impact of the narrative on the branding product preference. *Scientific reports*, 6(1):1–11, 2016.

[78] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, December 2013.

[79] Dustin J Welbourne and Will J Grant. Science communication on youtube: Factors that affect channel and video popularity. *Public understanding of science*, 25(6):706–718, 2016.

[80] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Beyond views: Measuring and predicting engagement in online videos. In *Twelfth international AAAI conference on web and social media*, 2018.

[81] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[82] Raymond Yeh, Ziwei Liu, Dan B Goldman, and Aseem Agarwala. Semantic facial expression editing using autoencoded flow. *arXiv preprint arXiv:1611.09961*, 2016.

[83] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[84] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022.

Supplementary References

[85] Emotion classification, August 2021. Page Version ID: 1041020718. URL: https://en.wikipedia.org/w/index.php?title=Emotion_classification&oldid=1041020718.

[86] Craiyon AI. Craiyon app. 2022. URL: http://www.craiyon.com/.

[87] MidJourney AI. Midjourney app. 2022. URL: https://www.midjourney.com.

[88] Cisco. Vni complete forecast highlights. 2018. URL: https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_Device_Growth_Traffic_Profiles.pdf.

[89] Data.ai. Top overall apps by ios and google play revenue in the united states as of july 28, 2022. 2022. URL: https://www.data.ai/en/apps/unified-app/top/revenue/united-states-of-america/overall/all-phone/?topchart=last-month.

[90] Ezra Fishman. How long should your next video be?, 2016. URL: https://eddl.tru.ca/wp-content/uploads/2019/08/EDDL5101_W5_Fisherman_2016.pdf.

[91] Paul Grabowicz. Tutorial: The transition to digital journalism, 2014. URL: https://multimedia.journalism.berkeley.edu/tutorials/digital-transform.

[92] Limelight. State of online video 2020. 2020. URL: https://www.limelight.com/resources/market-research/state-of-online-video-2020.

[93] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015. URL: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

[94] Wyzowl. State of video marketing 2022. 2022. URL: https://www.wyzowl.com/video-marketingstatistics.