Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# deepsing: Generating sentiment-aware visual stories using cross-modal music translation

Nikolaos Passalis [a,*,1], Stavros Doropoulos [b,1]

[a] Aristotle University of Thessaloniki, Greece
[b] DataScouting, Greece

## ARTICLE INFO

*Keywords:*
Music-to-image translation
Visual stories
Generative Adversarial Networks
Neural style transfer

## ABSTRACT

In this paper we propose a deep learning method for performing attributed-based music-to-image translation. The proposed method is applied for synthesizing visual stories according to the sentiment expressed by songs. The generated images aim to induce the same feelings to the viewers, as the original song does, reinforcing the primary aim of music, i.e., communicating feelings. The process of music-to-image translation poses unique challenges, mainly due to the unstable mapping between the different modalities involved in this process. In this paper, we employ a trainable cross-modal translation method to overcome this limitation, leading to the first, to the best of our knowledge, deep learning method for generating sentiment-aware visual stories. The proposed method was evaluated both quantitatively and qualitatively using a collection of songs that belong to 10 different genres, demonstrating that it is indeed possible to generate visual content that can match the sentiment expressed in songs. A user study was also conducted further validating the ability of the proposed method to provide sentiment-enriched visualizations.

## 1. Introduction

Music is closely tied to human evolution, with various musical instruments, such as flutes, dating back at least 40,000 years (Wade et al., 1979), while music itself can be traced back even before the Paleolithic era (Morley, 2003). Compared written and spoken language, music cannot (and does not aim to) accurately and concisely transfer semantic and quantitative information. In that sense, it seems like it has no functional role in our life, since it cannot be used to communicate for practical matters. Despite this, it excels at performing another function: *conveying emotions*. In fact, psychologists and neuroscientists suggest that music had a critical role in developing human societies, since it significantly assisted the process of socialization (Cross, 2010). Indeed, music consists an indispensable part of our life in modern societies, with the typical listener spending more than 30 h per week listening to music (Nielsen Report, 2017).

Our ability to perceive music is often reinforced by visual stimuli. For example, even in ancient Greek tragedies, music and dance performances were combined in the so-called *stasima*, which were interludes, often emotionally-charged, between the main episodes, explaining and/or commenting on the episodes (Taplin, 2003). The advent of digital technology provided further opportunities toward integrating audio and visual content in novel ways. Most songs are now accompanied by videos that further reinforce their sentiment, while many music players are capable of generating visual patterns that are synchronized with each song, e.g., based on concurrent tones (Ciuha, Klemenc, & Solina, 2010), harmonic structures and other acoustic features (Malandrino, Pirozzi, Zaccagnino, & Zaccagnino, 2015; Uehara & Itoh, 0000), or by detecting the sentiment of music (Chen, Weng, Jeng, & Chuang, 2008; Grekow, 2011).

The success of deep learning (DL) in various content generation and stylization tasks provides a powerful tool for tackling the aforementioned tasks. For example, Generative Adversarial Networks (GANs) (Brock, Donahue, & Simonyan, 2018; Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, et al., 2014; Karras, Aila, Laine, & Lehtinen, 2017) are capable of synthesizing highly realistic visual content that has not been encountered during the training process, neural style transfer methods can re-paint images to match the style of reference images (Luan, Paris, Shechtman, & Bala, 2017), or even to follow the style of well-known artists (Gatys, Ecker, & Bethge, 2015; Wang, Oxholm, Zhang, & Wang, 2017), while deep dreaming methods have demonstrated that neural networks can exhibit a behavior known as *pareidolia* in humans, i.e., recognize and synthesize patterns on seemingly random data (Mordvintsev, Olah, & Tyka, 2015). Despite the

[*] Corresponding author.
*E-mail addresses:* passalis@csd.auth.gr (N. Passalis), sdoropoulos@gmail.com (S. Doropoulos).
[1] Equal contribution.

https://doi.org/10.1016/j.eswa.2020.114059
Received 12 May 2020; Received in revised form 31 August 2020; Accepted 24 September 2020
Available online 25 September 2020

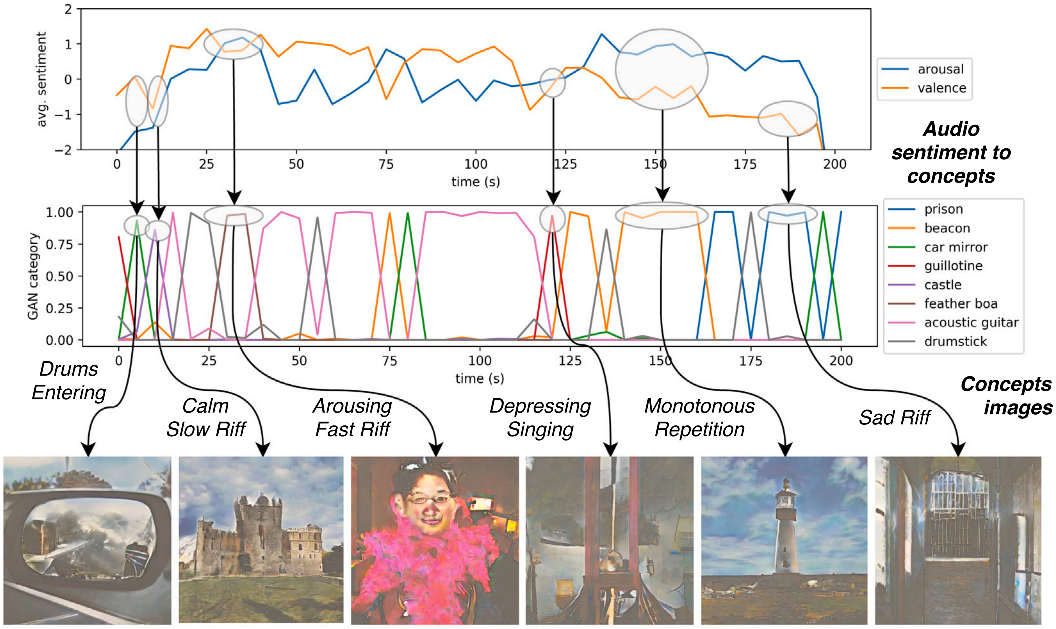


**Fig. 1.** Overview of the proposed method and a visualization example: music is translated into a visual story that express the same sentiment.

advanced capabilities provided by these approaches there has been no attempt, to the best of our knowledge, to employ DL methods for generating emotionally-rich visual content that can match the sentiment expressed in various music compositions.

In this paper we provide the first, to the best of our knowledge, DL attribute-based music-to-image translation method, called *deepsing*, that is capable of generating novel sentiment-aware visual content to accompany songs. That is, the synthesized visual stories are expected to induce the same sentiment to the viewers as the music does, aiding in this way the primary purpose of music, i.e., communicating feelings. It is worth noting that even though we used sentiment as the basis of deepsing, the proposed method can be directly used to transfer any attribute across two different domains, providing a generic cross-modal translation methodology. The proposed method works as follows: first the sentiment is extracted from a music track. Then, the sentiment is appropriately *translated* and mapped into a space capable of generating visual content. Finally, a generator model, e.g., a GAN, is employed to generate the final content. This process, along with a few illustrative examples are provided in Fig. 1. It is worth noting that significant challenges are faced during this process, as we will further discuss through this paper, with the most significant one being the existence of multiple, equally viable, mappings between the sentiment space and the generator space. This often leads to unstable mappings, prohibiting us from training models for performing cross-modal translation. In this paper, we provide a novel way to translate the audio sentiment into images with similar sentiment, effectively overcoming this limitation. Compared with existing literature on music visualization (Chen et al., 2008; Grekow, 2011; Malandrino et al., 2015; Uehara & Itoh, 0000), the proposed method employs a DL music-to-image translation approach for generating visual content that matches the sentiment of corresponding music segments. Note that, to the best of our knowledge, none of the existing methods is capable of generating meaningful visual content that can match the sentiment of songs, with most of them being limited to either simplistic animations, often with limited artistic value, e.g., Malandrino et al. (2015) and Uehara and Itoh (0000), or using collections of photos to generate slide-shows Chen et al. (2008). The proposed method does not only provide a novel way to generate sentiment-enriched visual stories, but opens a whole new research area for music-to-image translation with several high-value applications in many different domains. To aid research on this domain, we provide an open-source implementation of the method proposed in this paper at https://github.com/deepsing-ai/deepsing.

The rest of the paper is structured as follows. First, related work is briefly introduced and discussed in Section 2, while the problem of music-to-image translation is formally defined and the proposed method is derived in Section 3. Then, the experimental evaluation of the proposed method is provided in Section 4. Finally, conclusions are drawn and future research directions are discussed in Section 5.

## 2. Related work

There is a vast literature on music visualization approaches that have been proposed, highlighting the importance of visualizing various aspects of music and providing tools for various tasks, ranging from analyzing the structure of music (Wattenberg, 2002) to creating emotion-based slideshows (Chen et al., 2008).

Most of the existing works are devoted to provide tools for visualizing specific aspects of music, allowing for better understanding various of its qualities (Khulusi, Kusnick, Meinecke, Gillmann, & Jänicke, 2020). For example, arc diagrams have been proposed to provide simple visualizations of repetitions of musical phrases in songs (Wattenberg, 2002), while other approaches attempted to visualize the structure in music by capturing the chord progression (Bergstrom, Karahalios, & Hart, 2007), concurrent tones (Ciuha et al., 2010), as well as melodic and harmonic patterns (Snydal & Hearst, 2005). While all these approaches were based on 2D visualizations, other methods employed more complex 3D graphics (Fonteles, Rodrigues, & Basso, 2013; Miyazaki, Fujishiro, & Hiraga, 2003). More recent works also provided tools for better understanding the harmonic structures in music (Malandrino et al., 2015), the semantic structure in classical music works (Chan, Qu, & Mak, 2009) or even improving the understanding of music compositions (De Prisco, Malandrino, Pirozzi, Zaccagnino, & Zaccagnino, 2017), allowing for assisting the learning process (Malandrino, Pirozzi, & Zaccagnino, 2019). These methods indeed provided powerful tools for analyzing various musical compositions. However, most of these methods typically target experts in the field, aiming to produce simple and clear visualizations that will allow for better understanding various music qualities. On the other hand, the method proposed in this paper aims to perform *sentiment*-based visualization by generating realistic images instead of merely using vector-based or color-based visualizations. Therefore, the proposed method mainly targets non-experts, aiming at generating *visual stories* that can accompany the process of listening to music, similar to the way music players' visualizations are typically used. The interested reader is refereed to Khulusi et al. (2020) for an in depth review of such approaches.

Perhaps the most closely related method to ours is provided by Chen et al. (2008), where the sentiment expressed in musical compositions is visualized by matching it to the most appropriate photo. This method indeed allowed for enriching the listening experience of users, as the conducted study suggested (Chen et al., 2008). However, this method was limited to a small number of pre-selected photos. In other words, it was only able to generate slideshows that matched, to some extent, the emotion expressed in songs. A similar approach was also used in Sra, Maes, Vijayaraghavan, and Roy (2017), but employed virtual worlds to express the sentiment of songs instead. Our work go beyond existing approaches by exploiting the power of Generative Adversarial Networks (GANs) to generate unconstrained visualizations (Goodfellow et al., 2014). Instead of directly training GANs for music visualization, which would be very challenging given the lack of appropriate datasets, we propose employing an efficient cross-modal translation approach that allows for translating the *audio* sentiment space into the *visual* sentiment space, through any pre-trained GAN model. In this way, the proposed method can create a practically unlimited number of visualizations, instead of being limited to a small set of pre-selected photos (Chen et al., 2008), or requiring the use of tools for generating virtual reality spaces, as in Sra et al. (2017). To the best of our knowledge, this is the first work that proposed and evaluated an efficient cross-modal translation approach for sentiment-aware music visualization using GANs.

## 3. Proposed method

In this Section we formally define the problem of music-to-image translation, provide the proposed pipeline for this process and demonstrate how this approach can be used for generating sentiment-aware visual stories. Then, we analytically derive the employed neural attribute translation method, which lies at the heart of the proposed pipeline. Finally, we introduce an attribute-based neural stylization approach, that can further improve the relevance of the generated images with the given attributes, while also allowing the users to further adapt this process to their preferences.

### 3.1. Music-to-image translation

Let $\mathbf{x} \in \mathbb{R}^{N_s}$ be an audio segment with $N_s$ samples. Also, let $f_a(\mathbf{x}) \in \mathbb{R}^{N_a}$ be an *audio attribute estimator*, where $N_a$ is the dimensionality of the attribute space. The attribute space describes a specific property of the music that we want to maintain in the visual domain. For example, $f_a(\cdot)$ can extract information regarding the sentiment of the music, e.g., its valence and arousal, or information regarding the semantics of the lyrics. We also introduce an attribute extractor for the visual domain $g(\mathbf{y}) \in \mathbb{R}^{N_a}$, where $\mathbf{y} \in \mathbb{R}^{W \times H \times C}$ is a $C$-channel image of dimensions $W \times H$. The*visual attribute estimator* extracts the same attributes as the audio attribute estimator, but operates on images instead of audio. For example, it can extract information regarding the sentiment that an image induces to its viewers, or attributes regarding the semantic content of images, e.g., categories of the objects that appear in an image. These two attribute estimators should *aligned*, i.e., extract the same attributes and operate on the same output space. We also define a divergence metric $\mathcal{D}(\mathbf{t}_1, \mathbf{t}_2)$ for measuring the dissimilarity between two attribute vectors $\mathbf{t}_1, \mathbf{t}_2 \in \mathbb{R}^{N_a}$. For example, $\mathcal{D}(\cdot)$ can be defined using $l^2$ norm, i.e., $\mathcal{D}(\mathbf{t}_1, \mathbf{t}_2) = \|\mathbf{t}_1 - \mathbf{t}_2\|_2$.

The problem of music-to-image translation can be then defined as:

**Definition 1** (*Music-to-Image Translation*)**.** Given an audio segment $\mathbf{x}$, generate an appropriate image $\mathbf{y}$ so as:

$$\mathbf{y} = \arg_{\mathbf{y}} \min \mathcal{D}\left(f_a(\mathbf{x}), g(\mathbf{y})\right), \tag{1}$$

for an appropriately defined divergence metric $\mathcal{D}(\cdot)$.

Therefore, music-to-image translation aims to generate images that will carry the same attributes as the corresponding audio segments do, given the estimators $f_a(\cdot)$ and $g(\cdot)$. These estimators can be trivially fitted using annotated datasets for the corresponding domains. As already discussed, in this paper we focus on sentiment-oriented music-to-image translation, aiming to generate images that will induce the same sentiment to a viewer as the sentiment of the music to be translated. Therefore, the audio attribute estimator extracts the sentiment of the music, while the visual attribute estimator extracts the sentiment of the corresponding images. Please refer to Section 4 for more details on the datasets and models used for training these estimators. It is also worth noting that there is a "1-to-N" correspondence between each sentiment and the possible visual outputs that can induce this sentiment, i.e., there are many different visual stimuli that can lead to the same sentiment. Therefore, it is often critical to restrict the search space according to preferences of users, as we also discuss later in this Section, in order to produce meaningful and consistent results.

The proposed pipeline for music-to-image translation is shown in Fig. 2. First, the audio attribute extractor is employed to extract the attribute vector $\mathbf{t}^{(a)} = f_a(\mathbf{x})$ for a given audio segment $\mathbf{x}$. The proposed method aims to *translate* this vector into an appropriate *generator vector* $\mathbf{t}^{(t)}$, that can be fed into a generator model $\mathbf{y} = f_g(\mathbf{t}^{(t)})$, in order to acquire an image with the same attributes as the corresponding audio segment. In this work, we employ a GAN for generating the final images from the intermediate generator vectors. The gist of the proposed method is to learn an appropriate translation model $\mathbf{t}^{(t)} = f_t(\mathbf{t}^{(a)})$ that will map the attribute vectors to the appropriate generator vectors, ensuring that the divergence between the audio attributes and the visual attributes of the generated image will be minimized, as required by (1). In other words, the translation model must learn how to appropriately "control" the generator in order to generate images with the attributes provided by the audio attribute estimator.

### 3.2. Neural attribute translation

In this work, we propose to learn how to translate the attribute vectors to generator vectors by learning how to inverse the visual attribute estimator $g(\cdot)$. Therefore, the optimization problem given in (1) can be reduced to:

$$\mathbf{y} = f_g\left(\mathbf{t}^{(t)}\right), \text{ where } \mathbf{t}^{(t)} = f_t\left(f_a(\mathbf{x})\right), \tag{2}$$

and

$$f_t = \arg_{f_t} \min \mathcal{D}\left(f_a(\mathbf{x}), g\left(f_g\left(\mathbf{t}^{(t)}\right)\right)\right). \tag{3}$$

That is, the image $\mathbf{y}$ can be trivially generated after fitting an appropriate translation model $f_t(\cdot)$ that minimizes the attribute divergence between audio attributes and visual attributes. To efficiently learn the translation model $f_t(\cdot)$ we propose sampling the generator space in order to collect training data in the form of pairs $(\mathbf{t}_i^{(t)}, \tilde{\mathbf{t}}_i^{(a)})$, where $\tilde{\mathbf{t}}^{(a)}$ is the estimated attribute vector of the generated image, as calculated using the visual attribute estimator:

$$\tilde{\mathbf{t}}_i^{(a)} = g(f_g(\mathbf{t}_i^{(t)})). \tag{4}$$

It is worth noting that for the case of GANs, sampling the generator space is easy, since GANs are typical trained to generate images from a Gaussian distribution (Brock et al., 2018). Then, the translation model can be trivially learned to minimize the following loss:

$$\mathcal{L} = \sum_{i=1}^{N} \|f_t(\tilde{\mathbf{t}}_i^{(a)}) - \mathbf{t}_i^{(t)}\|_2, \tag{5}$$

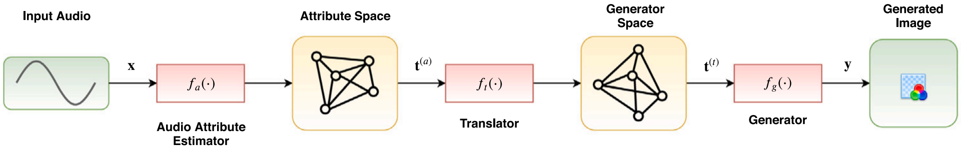


**Fig. 2.** Proposed audio-to-image translation pipeline.

for $N$ pairs of generator-attribute vectors. That is, the translation model learns how to map the attribute vectors, as induced by the visual modality, to the generator vectors that should be used to generate the images with the corresponding attributes. Gradient descent can be used to fit the translation model, i.e., $\Delta\mathbf{W} = -\eta\frac{\partial\mathcal{L}}{\partial\mathbf{W}}$, where $\eta$ is the learning rate and the notation $\mathbf{W}$ is used to refer to the parameters of the translation model, which is typically a deep neural network. It is trivial to verify that learning the translation model in this way ensures that (3) is indeed minimized, since

$$g\left(f_g\left(f_t(\mathbf{t}^{(a)})\right)\right) = \mathbf{t}^{(a)}, \tag{6}$$

by (4) and assuming that $f_t = \arg_t\min\mathcal{L}$. Therefore, the divergence given in (3) reduces to $\mathcal{D}\left(\mathbf{t}^{(a)}, g\left(f_g\left(f_t(\mathbf{x}^{(a)})\right)\right)\right) = \mathcal{D}\left(\mathbf{t}^{(a)}, \mathbf{t}^{(a)}\right) = 0$.

The proposed method for learning the translation model is summarized in Fig. 3(a). A sampler is employed to generate multiple generator vectors by drawing from an appropriate distribution, typically a Gaussian one. Then, these vectors are used to generate images, from which we extract visual attributes with the visual attribute estimator. The visual attribute estimator is also called *supervisor* in Fig. 3(a), since it supervises the training of the translation model. After gathering enough pairs of sampled generator vectors and the corresponding attribute vectors, the translator is fitted.

The aforementioned process theoretically guarantees that the optimal translator will be obtained, if enough samples are acquired and an appropriate translation model, with respect to its learning capacity, is used. However, in practice we observed that it was very difficult to fit such translation models, especially if GANs with very complex generator spaces are used, e.g., GANs capable generating 1000 classes (Brock et al., 2018). To understand why this happens, we have to consider the mapping between the (sub)-classes produced through the GAN and the attribute space. For example, consider the case of an attribute space that describes the sentiment of an image and a generator space from which multiple classes will be generated, as shown in Fig. 3(b). Note that for very small variations in the sentiment (denoted by $\epsilon$ in Fig. 3(b)) very large changes can occur in the mapping with the generator space, i.e., small variations of the negative sentiment can lead to many significantly different, yet equally negative, sub-classes. It is worth noting that this is the typical behavior of a chaotic system (Tsuda, 2001), explaining the difficulties in fitting such translation models. At the same time, also note that large variations in the sentiment might lead to very slight variations in the generator space, e.g., happy/sad dogs are closer compared to happy dogs and happy faces. This behavior was also experimentally confirmed for the GAN used in the experiments conducted in this paper, as shown in Fig. 3(c). The plot shown in Fig. 3(c) was generated by clustering the sound attribute space and then measuring the number of GAN classes mapped in each cluster. Note that a very large number of classes (often more than the half, i.e., more than 500) are mapped in most of the clusters, rending the mapping especially unstable.

To overcome this limitation, we propose creating a stable *attribute view* of the mapping between these two spaces. Therefore, instead of trying to match each attribute with every possible generator vector that can induce this attribute, we propose using only a part of the generator space. This limits the number of images that can be generated by $f_g(\cdot)$, effectively providing a specific view on the given attributes. Multiple views that can lead to the same attributes can be acquired by small perturbations of the attribute space.

**Algorithm 1** Calculating Stable Attribute Views

```
1:  procedure AttributeView($N_K$, $N_S$)
2:      Sample the generator space and generate $N$ pairs ($\mathbf{t}_i^{(t)}$,$\tilde{\mathbf{t}}_i^{(a)}$)
3:      Cluster attribute pairs into $N_K$ clusters according to $\tilde{\mathbf{t}}_i^{(a)}$
4:      Start with any empty training set $\mathcal{X}_{train}$ = []
5:      for each cluster do
6:          Sample one GAN category $c$ from each cluster according to
            the probability of observing each class in the generator space
7:          Add every instance ($\mathbf{t}_i^{(t)}$,$\tilde{\mathbf{t}}_i^{(a)}$) of category $c$ to $\mathcal{X}_{train}$
8:      for each category $c$ in $\mathcal{X}_{train}$ do
9:          Cluster the attribute vectors $\tilde{\mathbf{t}}_i^{(a)}$ of each category $c$ into $N_S$
    sub-categories
10:         Let $\mathcal{X}_i$ contain every pair ($\mathbf{t}_i^{(t)}$,$\tilde{\mathbf{t}}_i^{(a)}$) that belong to the $i$-th
    sub-category
11:         for each sub-category $i$ do
12:             Calculate a smooth attribute vector $\mathbf{t}_{target}^{(a)}$ =
    $\frac{1}{|\mathcal{X}_i|}\sum_{(\mathbf{t}^{(t)},\mathbf{t}^{(a)})\in\mathcal{X}_i}\mathbf{t}^{(a)}$
13:             Calculate a smooth generator vector $\mathbf{t}_{target}^{(t)}$ =
    $\frac{1}{|\mathcal{X}_i|}\sum_{(\mathbf{t}^{(t)},\mathbf{t}^{(a)})\in\mathcal{X}_i}\mathbf{t}^{(t)}$
14:             Replace each ($\mathbf{t}_i^{(t)}$,$\tilde{\mathbf{t}}_i^{(a)}$) pair in $\mathcal{X}_i$ by its corresponding
                smoothed target ($\mathbf{t}_{target}^{(t)}$,$\mathbf{t}_{target}^{(a)}$)
```

The algorithm employed for acquiring a view of the generator space is provided in Algorithm 1. In this work, we assume that a class-based GAN is employed. However, the proposed method can be also extended to handle any kind of GANs. First, the generator space is sampled and the visual attribute vectors are extracted (line 2). Then, these vectors are clustered into $N_K$ categories (line 3). For each cluster, we sample a GAN category with probability proportional to its cardinality in the cluster. Then, the latent vectors for all other categories in this cluster are discarded (lines 4–7). This process allows to effectively keep only one GAN category per cluster, leading to a smoother and much stabler matching, since every remaining attribute vector in each cluster is mapped to the same class. Note that, as demonstrated in Fig. 3(c), for a GAN capable of generating images belonging to 1,000 different categories, each initial cluster could be mapped to more than 500 different categories. Therefore it is expected that this process will improve the stability of the mapping, as we indeed experimentally demonstrate in Section 4.

After selecting the categories to be used, then we further cluster each category to detect sub-cluster that express different attributes (lines 8–14), e.g., different sentiments are detected in the case of our application. To this end, we cluster the attribute vectors for each category and then we calculate the centroid of each sub-cluster and each generator vector (lines 12–14), further smoothing the mapping. This process effectively allows to keep only the most prominent mappings between attribute and generator vectors. This process can seemingly reduce the variation during the content generation process. However, this is not expected to be a significant issue since (a) a large number of sub-clusters is typically used (i.e., $N_S > 10$), while (b) during the content generation and after calculating the vector $\mathbf{t}^{(t)}$ for a given class, a small (and easily controllable compared to the translation mapping)

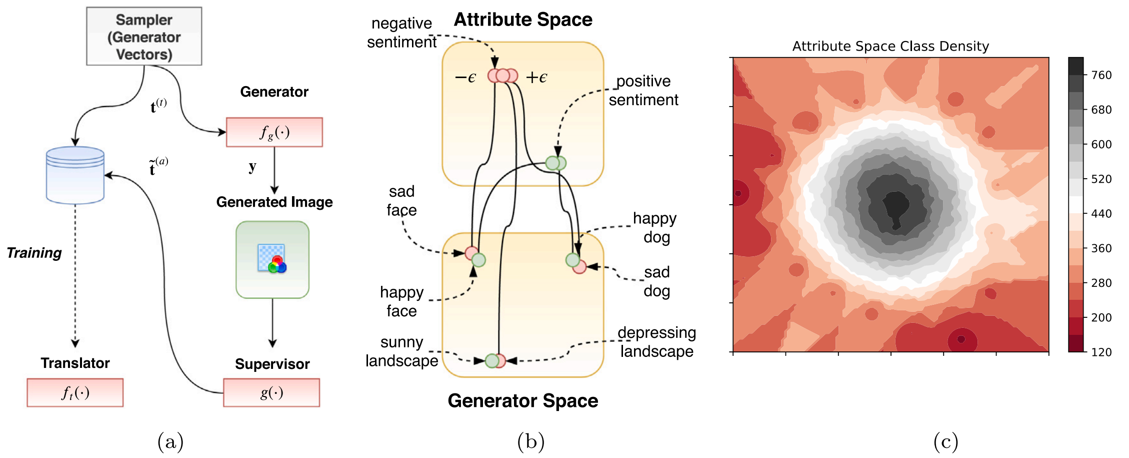


**Fig. 3.** Proposed method for fitting the translation model (a), an unstable mapping can occur between the attribute and generator spaces (b), and a toy example demonstrating the unstable mapping for a GAN with 1000 classes (c).

amount of noise can be optionally used to increase the variation of the generated content, if needed. Finally, note that users can also supply a number of categories that wish to use during the content generation. In this case, these categories can be directly used, skipping the first clustering step (lines 3–7) in Algorithm 1.

### 3.3. *Hyper-stylization*

Even though GANs can offer a satisfactory degree of variation for the generated content, they currently mostly fail to also simultaneously stylize the generated images according to the requirements of the users. Therefore, to further improve the style variation of the generated content, we propose using a few user-supplied styles to stylize the generate images. This process allows the users to adapt the generated visual stories to their preferences. In this work we opt for employing a universal style transfer approach, e.g., Li, Fang, Yang, Wang, Lu, and Yang (2017), allowing for using just one image per style. Therefore, after the user supplies the style images, these styles can be directly mapped to the attribute space using the visual attribute estimator $g(\cdot)$. Then, the style image that is closer to the current attributes can be employed to stylize the generated content. Examples of this process are also provided in Section 4. Finally, note that for human-friendly attributes, such as sentiment-based ones, e.g., arousal, valence, etc., the users can also manually set the thresholds that should be used for this process.

## 4. Experimental evaluation

### 4.1. *Experimental setup*

In this work, we focused on generating sentiment-oriented visual stories from music. To this end, we trained both the audio and visual attribute estimators to regress the valence and arousal of audio and visual stimuli respectively. It is worth noting that this choice was largely dictated by the current availability of open datasets for training sentiment-related attribute extractors. Any kind of attributes/features, can be also used to this end, e.g., sentiment embedding can be used instead of the employed valence-arousal features (Tang, Wei, Qin, Yang, Liu, & Zhou, 2015).

For developing the audio attribute estimator we extracted (a) 40 mel frequency cepstral coefficients (MFCC) (Logan et al., 2000), (b) chroma energy normalized (CENS) features (Müller & Ewert, 2011), and (c) tempogram features (Grosche, Müller, & Kurth, 2010). For all these features, we used the default parameter/extraction setup provided by the librosa library (McFee, Lostanlen, McVicar, Metsai, Balke, Thomé, et al., 2019). One feature vector containing the concatenation of these three different features was extracted every 500 ms. The extracted features were then fed into a neural regressor consisting of one hidden layer with 256 neurons (using sigmoid activations, which led to consistently better regression performance) and one output regression layer with 2 neurons. The regressor was trained to predict the valence and arousal of music segments of 500 ms using the Database for Emotional Analysis of Music (DEAM) (Alajanki, Yang, & Soleymani, 2016). For training the visual attribute extractor we employed a MobileNetV2 model (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018), trained to regress the valence and arousal using the Open Affective Standardized Image Set (OASIS) (Kurdi, Lozano, & Banaji, 2017). The sound attribute estimator was trained for 30+20 training epochs using a learning rate of $10^{-4}/10^{-5}$. The visual attribute estimator was pretrained on the Imagenet dataset and then fine-tuned (after replacing the last regression layer) for 50+10 epochs using the same learning rates ($10^{-4}/10^{-5}$). The Adam method was used for optimizing the models (Kingma & Ba, 2015). The visual attribute space and the audio attribute space were aligned by performing z-score normalization. For the visual space we employed the statistics of the corresponding training dataset, while for the audio space we use song-level z-score normalization, except from the third song used for the conducted qualitative experiments, where using the statistics of the whole dataset led to a wider variety of generated concepts.

A pretrained BigGAN (Brock et al., 2018) model was used for generating images of 512 × 512 pixels. The translation model consists of two hidden layers with 64 and 256 neurons respectively which branches out into two streams: (a) a 1000 classification (softmax) layer used for predicting the category that will be used by the GAN, and (b) a 128 fully connected layer with no activation function used for predicting the latent vector to be fed to the GAN. The translation model was trained for 200 epochs with a learning rate of 0.001. Gaussian noise with $\sigma = 0.1$ was added to the translated latent vectors, to increase the variation of the generated content. Finally, for performing stylizations we employed the Whitening Coloring Transform (WCT)-based stylization method (Li et al., 2017). The first four encoder–decoders were used for the stylization process, while the stylization blending factor was set to 0.1.

For the three song used for qualitatively studying the behavior of the proposed method, the concept classes were manually selected, and the algorithm presented in Section 3 was used to automatically match the sentiment of each song to the available concepts. It is worth noting that no human intervention was allowed during the process of generating the visual story. Therefore, the matching between the available concepts and classes was performed fully autonomously.

**Table 1**
Evaluating the matching between the target sentiment and the generated sentiment using two different metrics (mean absolute error (MAE) and mean precision on the valence and arousal).

| Genre | Control | | Proposed | |
|---|---|---|---|---|
| | MAE | Precision | MAE | Precision |
| Blues | 0.787 | 50.38 | 0.581 | 73.33 |
| Classical | 1.138 | 49.95 | 0.734 | 89.25 |
| Country | 0.707 | 49.54 | 0.561 | 68.17 |
| Disco | 0.921 | 50.13 | 0.741 | 58.75 |
| Hiphop | 0.760 | 49.65 | 0.697 | 56.33 |
| Jazz | 0.722 | 49.30 | 0.529 | 78.63 |
| Metal | 0.918 | **50.38** | 0.893 | 47.04 |
| Pop | 0.885 | 48.54 | 0.669 | 62.50 |
| Reggae | 0.738 | 50.25 | 0.524 | 67.92 |
| Rock | 0.752 | 50.75 | 0.652 | 59.12 |
| Average | 0.833 | 49.89 | 0.658 | 66.10 |

### *4.2. Experimental results*

First, we provide quantitative results using the GTZAN dataset (Sturm, 2012), along with the results of the conducted user study. Then, we provide three examples of visual stories generated using the proposed method to further qualitatively evaluate the proposed method.

The results of the conducted quantitative study are reported in Table 1. The GTZAN dataset, which contains music segments of 10 different music genres was used. For each genre 100 different music segments, each with a duration of 30 s, is provided. The proposed method was applied on all data provided by the GTZAN dataset and the agreement between the audio sentiment and the sentiment expressed by the generated images was measured. The sentiment in both cases was measured using the valence-arousal model employing the DL models used for training the cross-modal translator. The agreement was measured using two different measures: (a) mean absolute error (MAE) and (b) mean precision on two class (positive/negative) valence and arousal. The proposed method was compared to a control model that employed a random translator, i.e., a translator that randomly selects a class to generate the corresponding images, instead of taking into account the sentiment of the corresponding audio, following the setup used in Chen et al. (2008). Note that the proposed methods works significantly better than the control method, increasing the average precision from about 50% (random choice) to over 66%, while reducing the mean absolute error from about 0.83 to 0.65. Note that the reported differences are statistically significant ($a = 0.01$), as verified by applying the Wilcoxon signed rank test ($p = 0.007$) (Woolson, 2007). Another quite interesting observation is that the effectiveness of the proposed method varies for different music genres, e.g., classical music leads to the best results, followed by jazz. On the other hand metal, disco and rock consistently led to the lowest precision compared to the rest of the evaluated methods.

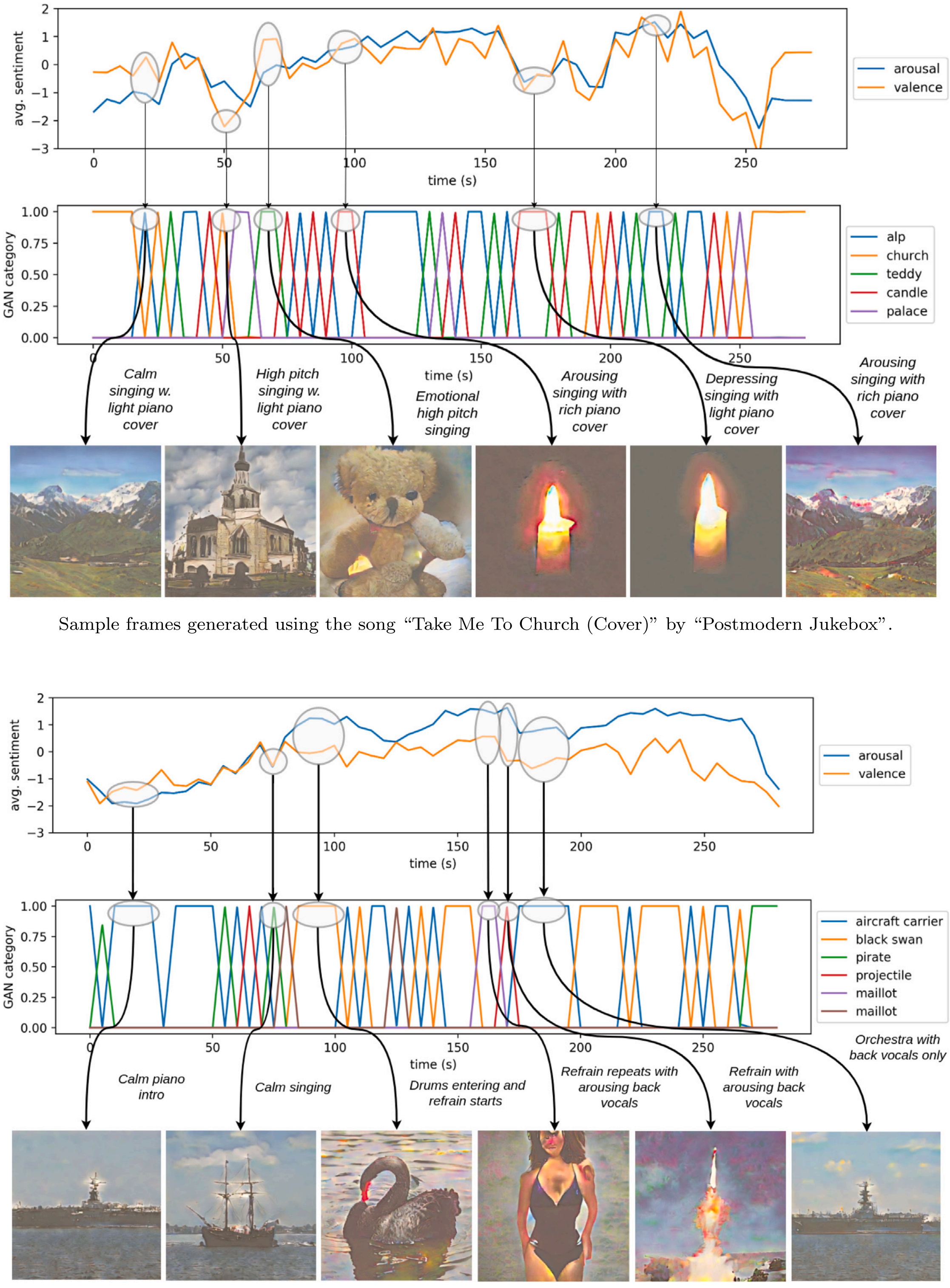
Next, we also designed a simple user-study to examine whether the differences reported on Table 1 are indeed noticeable by humans. To this end, we presented video clips generated using both the proposed and control methods to human subjects and we asked them to select which of two videos (control and proposed) better expresses the sentiment of each song. The question that was posed to the participants after presenting the two videos was: “Which of the two previous videos best expresses the sentiment of the song?”. No hyper-stylization was used for the generated videos in order to avoid any bias that could be induced by this process. A total number of $n = 20$ humans participants took part in the study, while we only collected their preferences for the videos that were presented to them (no other personal data were collected). For 70% of the used genres users agreed that the proposed method leads to a video that better expresses the sentiment of each song. In this case, the average preference on the videos generated by the proposed method was 60.04%. Note that this average agreement is close to the mean precision reported in Table 1, validating the differences between the control videos and the ones generated using the proposed method.

Next, we examine three visual stories generated by the proposed method for three well known songs. The first one was provided in Fig. 1, where the song “Chop Suey!” was used. In the upper plot we provide the average arousal and valence over the music intervals (5 s each) used for the content generation, while in the subsequent plot we provide the activations of the translator model. We examine the behavior of the proposed method at six selected points of interest. For the first two points of interest the arousal is low, while the valence is near to zero. Indeed, these points correspond to a more “relaxed” part of the song, where the drums are just entering and a calm slow riff accompanies the song. The generated images match this sentiment, since they are quite neutral and with low arousal (a car mirror and castle in the country side). On the other hand, the third point of interest corresponds to one of the most arousing parts of the song, where a very fast and exciting riff is played. The proposed method responds to this by generating a bright pink feature boa. Then, the arousal is reduced and a few matching classes are used for the visualization (e.g., acoustic guitar). Then, a guillotine is visualized at a critical transition point of the song, at the fourth point of interest, where both the singing style and the lyrics are less positive (the lyrics actually refer to suicide at this point). Then, the valence is slowly decreasing by a monotonous repetition of the same musical phrase, at the fifth point of interest, which is accompanied by a visualization of an isolated beacon, creating the sense of loneliness. The drop in the arousal, which is especially evident at this point in the song, concludes by generating frames containing a prison, at the final point of interest.

The next song used for evaluating the proposed method is the “Take Me To Church” originally performed by “Hozier”. As before, we visualize some points of interests in the generated visual story in Fig. 4. The song begins with a calm and slow singing style, accompanied with a light piano cover. At this point a low arousal, neutral valence scenery of a mountain is generated. After a few seconds the valence reduces, with the singer increasing the pitch, leading to the generation of a church. The negative valence of the generate image is further reinforced by the cloudy sky. Then, at the third point of interest, the singer still maintains a high pitch, but with a significantly warmer tone. At this point a more positive teddy bear is generated. Then, as the intensity of the song rises, a warmly colored candle is generated. It is worth noting that a similar candle is also used to visualize another point of interest, which has significantly lower valence and arousal. Indeed, this is illustrated in the generated images, since the candle of the fourth point of interest has an overall more positive appearance compared to the one used in the firth point of interest. This behavior confirms the ability of the proposed method to generate images that can highlight the subtle emotional differences between different instances of the same class. This is also confirmed in the last generated instance, where a more positive scenery was generated compared to the one used in the first frame.

The third visual story, depicted in the lower part of Fig. 4, was generated using the song “Skyfall” by “Adele”. Dominant element in this visual story is sea, which exists in two of the most frequently generated classes, i.e., “aircraft carrier” and “pirate ship”. Frames containing these classes are generated mostly during the lower intensity parts of the song, especially in the beginning, where both the valence and the arousal is low. During the arousal build up phase, where the drums enter and the more intense refrain begins, a black swan is generated, while the arousal peaks during the fourth and fifth frame, with more arousing content being generated.

Finally, the effect of the proposed hyper-stylization method is demonstrated in Fig. 5, where images of one class, ranging from the ones with the most negative valence to the ones with the most positive one, were generated. Two classes were used for this experiment: “dog” and “beer”. First, note that even though the employed GAN was



Sample frames generated using the song “Take Me To Church (Cover)” by “Postmodern Jukebox”.

Sample frames generated using the song “Skyfall” by “Adele”.

**Fig. 4.** Two visual stories generated for two different songs.

not trained to generate images with different valance, the proposed translation approach manages to reveal some meaningful valance characteristics. For example, dogs with more positive valence tend to have their mouth open, while beer bottles that co-exist with other items were also classified as more positive. The proposed hyper-stylization approach can further improve the valence of the generated images by employing three user-defined style images (depicted in the lower right part of the images). It is evident that the proposed hyper-stylization process can indeed significantly improve the matching between the desired valence and the valence of the generated images. It is worth noting that the same style images were used for all the conducted experiments in this paper, while the threshold for neutral sentiment

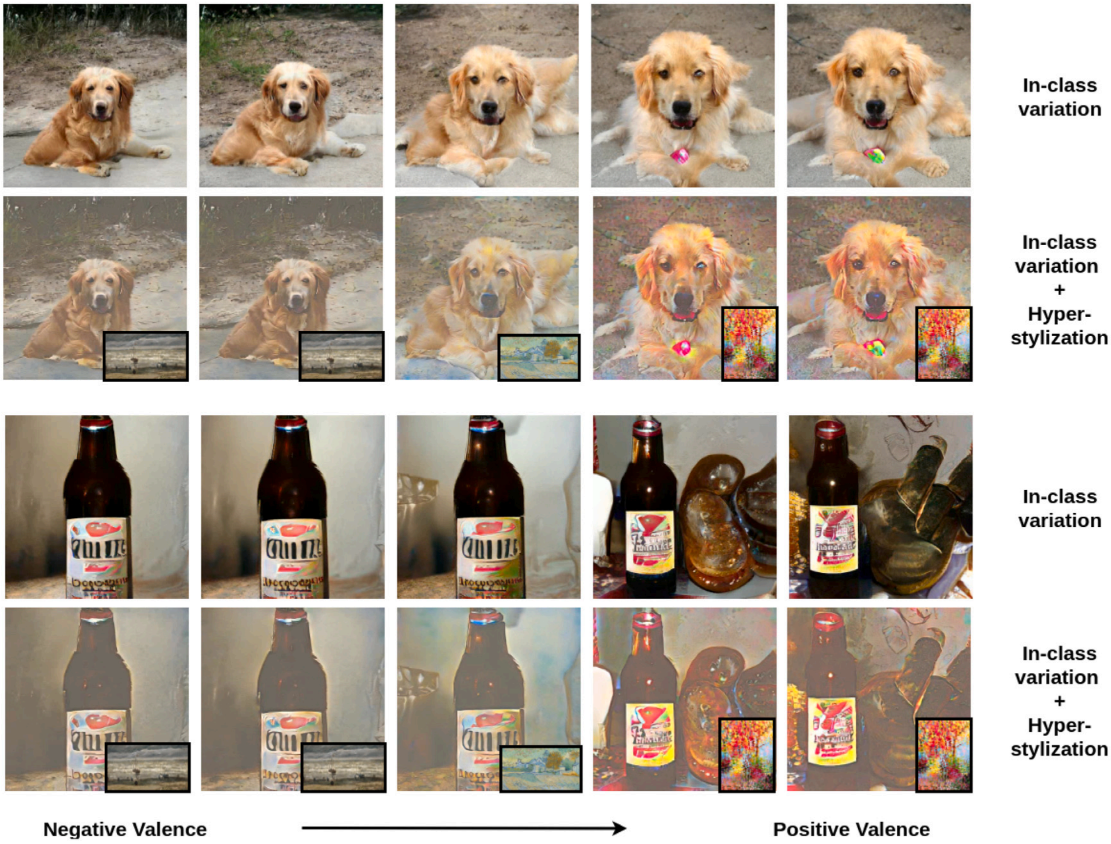

In-class variation
In-class variation + Hyper-stylization
In-class variation
In-class variation + Hyper-stylization
Negative Valence
Positive Valence

**Fig. 5.** Examining the ability of the translator model to discover sentiment-rich regions of the generator space, along with the effect of the proposed hyper-stylization approach.

was set to 0.5: an image with an average sentiment (valence and arousal) lower than −0.5 was considered as negative, while an average sentiment higher than 0.5 was considered positive.

## 5. Conclusions

In this paper we introduced *deepsing*, a deep learning method for generating sentiment-aware visual stories by performing cross-modal translation from the audio domain. The proposed method works by first extracting the sentiment of a music track, which is then appropriately translated into a space, from which a GAN can be employed for generating the frames of the visual story. This process was proven to be especially challenging, since the mapping between the sentiment space and the generator space is unstable, requiring special care to ensure the stability of the matching. To further enhance the quality of the visual stories we employed a hyper-stylization approach, that performs attribute-aware stylization. The ability of the proposed method to produce meaningful visual stories was demonstrated using three well-known modern songs.

The results obtained in the qualitative evaluation were, in some cases, quite spectacular, matching the generated visual content to the audio one in a very satisfactory way. The quantitative evaluation demonstrated that the proposed cross-model translation approach indeed works quite well, leading to generating visual content that matches the audio sentiment, as extracted by the employed DL models. However, the effectiveness of the proposed method seems to be limited by the quality of the employed audio and image sentiment extractors. Indeed, in the conducted user study it was demonstrated that, even though the generated videos matched the sentiment of the audio clips better than the control ones, a significant margin for improvement exist. Among the most important limitations is the lack of large-scale sentiment datasets that can be used for training the aforementioned sentiment extractors. Using EEG-based methods (Gauba, Kumar, Roy, Singh, Dogra, & Raman, 2017) is a very promising future research directions for easily collecting such large sentiment datasets with relatively low effort. Apart from this, for each cluster formed in the attribute space there are many potential candidate classes. Selecting the class to use for the content generation according to a semantic similarity measure with the rest of the selected classes, instead of performing cardinality-based sampling, is expected to improve the semantic consistency of the generated story. Furthermore, lyrics-based methods can be also used to better guide this process, possibly increasing the satisfaction of users (Rospocher, 2020). Finally, employing more advanced methods for aligning the attribute spaces of the audio attribute estimator and visual attribute estimator, e.g., Zhu, Zhuang, and Wang (2019), while at the same time employing diversity metrics, such as entropy, to increase the diversity of the generated content, is expected to further improve the quality of the visual stories.

## CRediT authorship contribution statement

**Nikolaos Passalis:** Methodology, Software, Writing - original draft, Writing - review & editing. **Stavros Doropoulos:** Methodology, Software, Writing - original draft, Writing - review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

Alajanki, Anna, Yang, Yi-Hsuan, & Soleymani, Mohammad (2016). Benchmarking music emotion recognition systems. *PLoS One*, 835–838.

Bergstrom, Tony, Karahalios, Karrie, & Hart, John C. (2007). Isochords: visualizing structure in music. In *Proceedings of graphics interface.* (pp. 297–304).

Brock, Andrew, Donahue, Jeff, & Simonyan, Karen (2018). Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096.

Chan, Wing-Yi, Qu, Huamin, & Mak, Wai-Ho (2009). Visualizing the semantic structure in classical music works. *IEEE Transactions on Visualization and Computer Graphics*, *16*(1), 161–173.

Chen, Chin-Han, Weng, Ming-Fang, Jeng, Shyh-Kang, & Chuang, Yung-Yu (2008). Emotion-based music visualization using photos. In *Proceedings of the international conference on multimedia modeling.* (pp. 358–368).

Ciuha, Peter, Klemenc, Bojan, & Solina, Franc (2010). Visualization of concurrent tones in music with colours. In *Proceedings of the ACM international conference on multimedia.* (pp. 1677–1680).

Cross, Ian (2010). The evolutionary basis of meaning in music: Some neurological and neuroscientific implications. *The neurology of music*, 1–15.

De Prisco, Roberto, Malandrino, Delfina, Pirozzi, Donato, Zaccagnino, Gianluca, & Zaccagnino, Rocco (2017). Understanding the structure of musical compositions: Is visualization an effective approach? *Information Visualization*, *16*(2), 139–152.

Fonteles, Joyce Horn, Rodrigues, Maria Andréia Formico, & Basso, Victor Emanuel Dias (2013). Creating and evaluating a particle system for music visualization. *Journal of Visual Languages & Computing*, *24*(6), 472–482.

Gatys, Leon A., Ecker, Alexander S., & Bethge, Matthias (2015). A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

Gauba, Himaanshu, Kumar, Pradeep, Roy, Partha Pratim, Singh, Priyanka, Dogra, Debi Prosad, & Raman, Balasubramanian (2017). Prediction of advertisement preference by fusing EEG response and sentiment analysis. *Neural Networks*, *92*, 77–88.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, et al. (2014). Generative adversarial nets. In *Proceedings of the advances in neural information processing systems.* (pp. 2672–2680).

Grekow, Jacek (2011). Emotion based music visualization system. In *Proceedings of international symposium on methodologies for intelligent systems.* (pp. 523–532).

Grosche, Peter, Müller, Meinard, & Kurth, Frank (2010). Cyclic tempogram– A mid-level tempo representation for musicsignals. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing.*(pp. 5522–5525).

Karras, Tero, Aila, Timo, Laine, Samuli, & Lehtinen, Jaakko (2017). Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

Khulusi, R., Kusnick, J., Meinecke, C., Gillmann, J., & Jänicke, S. (2020). A survey on visualizations for musical data. In *Computer Graphics Forum*. Wiley Online Library.

Kingma, Diederik P., & Ba, Jimmy (2015). Adam: A method for stochastic optimization. In *Proceedings of the international conference on learning representations*. arXiv preprint arXiv:1412.6980.

Kurdi, Benedek, Lozano, Shayn, & Banaji, Mahzarin R. (2017). Introducing the open affective standardized image set (OASIS). *Behavior research methods*, *49*(2), 457–470.

Li, Yijun, Fang, Chen, Yang, Jimei, Wang, Zhaowen, Lu, Xin, & Yang, Ming-Hsuan (2017). Universal style transfer via feature transforms. In *Proceedings of the advances in neural information processing systems.* (pp. 386–396).

Logan, Beth, et al. Mel Frequency Cepstral Coefficients for Music Modeling. In *Proceedings of the International Conference on Music Information Retrieva, Vol. 270.* (pp. 1–11).

Luan, Fujun, Paris, Sylvain, Shechtman, Eli, & Bala, Kavita (2017). Deep photo style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* (pp. 4990–4998).

Malandrino, Delfina, Pirozzi, Donato, & Zaccagnino, Rocco (2019). Learning the harmonic analysis: is visualization an effective approach? *Multimedia Tools and Applications*, *78*(23), 32967–32998.

Malandrino, Delfina, Pirozzi, Donato, Zaccagnino, Gianluca, & Zaccagnino, Rocco (2015). A color-based visualization approach to understand harmonic structures of musical compositions. In *Proceedings of the international conference on information visualisation.* (pp. 56–61).

McFee, Brian, Lostanlen, Vincent, McVicar, Matt, Metsai, Alexandros, Balke, Stefan, Thomé, Carl, et al. (2019). Librosa/librosa: 0.7.1. http://dx.doi.org/10.5281/zenodo.3478579, URL https://doi.org/10.5281/zenodo.3478579.

Miyazaki, Reiko, Fujishiro, Issei, & Hiraga, Rumi (2003). Exploring MIDI datasets. In *ACM SIGGRAPH 2003 sketches & applications*. 1–1.

Mordvintsev, Alexander, Olah, Christopher, & Tyka, Mike (2015). Inceptionism: Going deeper into neural networks.

Morley, Iain (2003). *The evolutionary origins and archaeology of music* (Ph.D. thesis), Darwin College, Cambridge University Cambridge.

Müller, Meinard, & Ewert, Sebastian (2011). Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the international conference on music information retrieva*.

Nielsen Report (2017). *MUSIC 360*. https://www.nielsen.com/us/en/insights/report/2017/music-360-2017-highlights.

Rospocher, Marco (2020). Explicit song lyrics detection with subword-enriched word embeddings. *Expert Systems with Applications*, Article 113749.

Sandler, Mark, Howard, Andrew, Zhu, Menglong, Zhmoginov, Andrey, & Chen, Liang-Chieh (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* (pp. 4510–4520).

Snydal, Jon, & Hearst, Marti (2005). ImproViz: visual explorations of jazz improvisations. In *Proceedings of human factors in computing systems.* (pp. 1805–1808).

Sra, Misha, Maes, Pattie, Vijayaraghavan, Prashanth, & Roy, Deb Auris: creating affective virtual spaces from music. In *Proceedings of the ACM symposium on virtual reality software and technology.* (pp. 1–11).

Sturm, Bob L. (2012). An analysis of the GTZAN music genre dataset. In *Proceedings of the international ACM workshop on music information retrieval with user-centered and multimodal strategies.* (pp. 7–12).

Tang, Duyu, Wei, Furu, Qin, Bing, Yang, Nan, Liu, Ting, & Zhou, Ming (2015). Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *28*(2), 496–509.

Taplin, Oliver (2003). *Greek tragedy in action*. Routledge.

Tsuda, Ichiro (2001). Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behavioral and Brain Sciences*, *24*(5), 793–810.

Uehara, Misa, & Itoh, Takayuki Pop music visualization based on acoustic features and chord progression patterns applying dual scatterplots.In *Proceedings of the sound and music computing conference.* (pp. 43–48).

Wade, Bonnie C., et al. (1979). *Music in India: The classical traditions*. Prentice-Hall.

Wang, Xin, Oxholm, Geoffrey, Zhang, Da, & Wang, Yuan-Fang (2017). Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In *Proceedings of the conference on computer vision and pattern recognition.* (pp. 5239–5247).

Wattenberg, Martin Arc diagrams: Visualizing structure in strings. In *Proceedings of the SIEEE symposium on information visualization.* (pp. 110–116).

Woolson, R. F. (2007). Wilcoxon signed-rank test. In *Wiley encyclopedia of clinical trials* (pp. 1–3). Wiley Online Library.

Zhu, Yongchun, Zhuang, Fuzhen, & Wang, Deqing (2019). Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 33.* (pp. 5989–5996).