Synesthetic Variational Autoencoders

Translating Visual Works of Art into Music

Maximilian Müller-Eberstein
edu@personads.me
11740590

Universiteit van Amsterdam
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Instituut voor Informatica

MSc Artificial Intelligence
August 2019

36 EC
January 2019 - August 2019

Universiteit
van Amsterdam

Supervisors
dr. Nanne van Noord
Marco Federici, MSc

Assessor
prof. dr. Marcel Worring

ABSTRACT

Translating visual art into music using machine learning models would be desirable in order to make large museum collections accessible to the visually impaired. However, generative methods so far are either unable to work across more than one sensory modality [24, 42] or require paired audio-visual datasets [3, 18, 19, 47]. The Synesthetic Variational Autoencoder (SynVAE) introduced in this research is able to learn a consistent mapping between sensory modalities in absence of any such paired datasets by exploiting a common prior latent space distribution for otherwise independent image and music generation models.

Evaluation on the common MNIST [28] and CIFAR-10 [25] datasets as well as on the Behance Artistic Media dataset (BAM) [46] shows that SynVAE is capable of retaining sufficient information content during the translation process while maintaining cross-modal latent space consistency. Using information theoretic metrics such as MINE [2], information content and consistency can further be quantified effectively. These quantitative metrics were then used to create informed qualitative evaluation tasks in which human evaluators matched musical samples with the images which generated them. Accuracies of up to 73% in these trials confirm a high degree of naturally perceived audio-visual consistency.

1 INTRODUCTION

Art is experienced as a flow of information between an artist and an observer. Should the latter be impaired in the principal sense which the artwork is aimed at however, a barrier appears. Such is the case for visually impaired people and paintings, for instance. One way to overcome this obstacle might be to translate the artwork from an inaccessible sensory modality into an accessible one.

Considering this problem a cross-modal transformation of information representations, we find similarities to the methods applied in machine learning: complex information such as images or language are transformed into latent representations which the models are then able to process. The research question of this thesis will therefore be to examine how we can leverage such models to create representations in one sensory modality to encode the information of another. Specifically, to make visual art accessible by translating it into music.

Our research builds upon single-modality generative models for images [14–16, 24] and music [33, 42] as well as multi-modal models which leverage corresponding audio-visual data in order to learn more stable information representations [18] or make visual information more accessible [3, 47]. Furthermore, generative models have also been used to measure the expressiveness of image-based audio generation tools for the visually impaired [19] and as such they offer a solid basis for our approach.

After outlining the task and its associated challenges in this section, we will go over the theoretical background in Section 2 and compare them with related work in this field in Section 3. In Section 4, we introduce our cross-modal translation model, the Synesthetic Variational Autoencoder (SynVAE). The associated unsupervised training methodology as well as the layout of our experiments and evaluation procedures are provided in Section 5. The results thereof are presented in Section 6 and discussed in Section 7 before applications and future areas of research are identified in Section 8.

1.1 Motivation

Having access to the information content of a work of art as well as to its place in the context of its contemporaries is a prerequisite for participating in cultural settings such as art exhibitions. In absence of the original visual information conveyed by the artist, alternative channels of relaying the information represented in the artwork are required. However, these may be insufficient in several ways.

A verbal, auditive description of a painting may provide a solid idea of what it depicts and place it within a richer context. Describing the exact positioning and style of each element within it for instance, is however more informative than enjoyable. Modern art with its high levels of abstraction may in addition be hard or impossible to describe using this method. A tactile approach such as allowing someone to touch the canvas or sculpture is more engaging, but may be insufficient to convey higher level information. Implementation is also difficult or impossible if the painting is flat or if touching the artwork would irreparably damage it.

A musical approach seems fitting for this task since the medium has a dense way of encoding artistic information and can also provide a greater level of engagement due to its intuitive nature. Similarities between different musical representations might be recognised much quicker than descriptions of similar contents across multiple paintings.

Indeed, projects such as "Eyes-Free Art" [40] in which a small number of visual artworks were accompanied by specifically composed music as well as verbal descriptions have yielded positive feedback from visually impaired participants. Having an artist compose a new piece of fitting music for each artwork which shares consistent characteristics within the larger context of the museum's overall collection may however not be possible due to cost concerns. As such, an artificial composer could be used to fill that gap.

Generative machine learning models have already shown promise in generating realistic music (see Section 3.2) and recent research on cross-modal transformations between the visual and auditive domains have shown that images can be used as a basis to produce related audio (see Section 3.3). The areas of application for these models range from improving the accessibility of information for the visually impaired [47] to enabling the evaluation of assistive sensory devices using reproducible quantitative metrics [19].

We aim to extend this field of research by not only working with more abstract visual information in the form of art, but also by enforcing a consistency criterion between audio-visual pairs such that a listener may be able to infer properties and similarities of images by hearing their corresponding musical pieces alone.

1.2 Challenges

Automatically composing music which is conditioned on a visual prior and retains shared properties between similar items comes with many challenges, starting with finding audio-visual datasets, ensuring consistency between the two sensory modalities during translation and finally, evaluating the results in a quantitative manner while staying close to how humans actually perceive them.

Training models to generate realistic audio is difficult in itself, mainly due to the sensitivity of human listeners to the realism of a composition, but also due to the sparse availability of

structured musical input data. Although some public datasets of music are available, they unsurprisingly do not have any annotations relating to related artwork [35, 39]. Image data for artworks is more readily available, but even with well annotated sources, related music is not amongst the metadata. Constructing a model proficient with both visual art and music will therefore require it to learn relationships between those two domains on its own, ideally in an unsupervised manner.

Since determining which music best represents a certain piece of visual art will always be a subjective endeavour, we will focus on the information content across representations and on whether our model is able to correlate certain aspects of music (e.g. a major or minor scale) with certain types of visuals (e.g. light or dark images). Since we are not building upon any human intuitions, the model might however also produce results such as mapping brighter colours to deeper notes although our experience might predict the contrary. The key is consistency such that similar images produce similar music.

This exposes the most crucial challenge of the process: consistency within the latent space. Established models for learning such latent representations from complex single-modality input do exist in the form of Variational Autoencoders [24] (see Chapter 2), but because we are simultaneously working across the image and audio domains, latent consistency must also hold across modalities. This means that similar images must not only be embedded close to each other, but must also produce similar audio.

Measuring correlation in these domains quantitatively is possible using information theoretic metrics, but it will nonetheless be necessary to also measure whether audio-visual similarity is perceived by humans consistently. Performing qualitative analyses with human evaluators runs into the issue that the number of samples and evaluators required to produce a generalisable analysis might be implausible given the diversity of both visual art and music, especially in the absence of paired ground truths. Pre-selecting samples for human evaluation manually would however inevitably introduce bias of the curator and make results impossible to replicate elsewhere. It will therefore be necessary to first define quantitative metrics for measuring latent space consistency and then use these metrics to make an informed and reproducible selection of representative samples to use when testing whether the model's translations line up with the intuitions of human evaluators.

1.3 Contributions

In an effort to make visual art accessible to the visually impaired, we attempt to find suitable auditory representations of artworks by leveraging the strengths of autoencoding generative models. The cross-modal synesthetic model introduced in this thesis encodes an image into joint audio-visual latent space, generates a melody based on the embedding and estimates the quality of this mapping by trying to reconstruct the original image based on the re-encoded musical latent representation (see Section 4). This process allows for a fully unsupervised pipeline without the need for paired audio-visual training data which are not available in the fields of auditive and visual art.

Within the audio latent space, melodic samples which are recognized as music are relatively rare compared to the amount of audio noise. Therefore, we leverage a pre-trained model specifically designed for music, fittingly called MusicVAE [42] (see Section 2.2). By utilising its encoder and decoder within our synesthetic model, we are able to bypass the complex issue of music generation and can have a relatively high certainty that the audio output will be a valid melody.

In order to verify the validity of this approach, the model is first applied to the simple MNIST digit dataset [28], continuing on to the more complex CIFAR-10 dataset [25] and finally an extensive collection of contemporary works of art called the Behance Artistic Media dataset (BAM) [46]. At each step our method is evaluated by measuring the shared information content in both the visual as well as the auditive representations using correlation metrics, the agreement of separately trained classifiers and an estimation of mutual information using MINE [2] (see Section 5.2). As a final test for the perceived quality of the generated audio-visual pairs, a study with human evaluators will be conducted according to a quantitatively informed and reproducible method (see Section 5.3).

To summarize, our main contributions in this research are as follows:

- With the Synesthetic Variational Autoencoder (SynVAE), we introduce an unsupervised cross-modal architecture for translating data from one sensory modality into another consistently without the need for subjectively paired ground truth datasets (see Section 4).
- In a series of experiments on generating music from increasingly complex visual datasets (MNIST [28], CIFAR-10 [25], BAM [46]), we leverage and compare a variety of mutual information metrics in order to establish a quantitative basis for evaluating such cross-modal models (see Sections 5.2 and 6).
- In a qualitative study based on these quantitative metrics, we evaluate the human perception of the cross-modal translation consistency and lay out a framework for avoiding subjective biases within this process (see Sections 5.3 and 6).

2 BACKGROUND

In translating between visual and auditive domains, we rely on an unsupervised generative model which learns consistent latent representations. This section will therefore provide the necessary background on our architecture of choice, the Variational Autoencoder (VAE), as well as on methods used to improve its performance (see Section 2.1). Furthermore, we explain how such an autoencoder can be applied to music in Section 2.2 by taking a closer look at the MusicVAE architecture [42] which is employed in the auditive components of our final model.

2.1 Variational Autoencoders

Since machine learning models typically transform raw input such as image pixels or sparse bag-of-word vectors into more information dense representations, we can find an abundance of encoder models which can perform this transformation task. Actively enforcing meaningful consistencies between latent representations is however more difficult. Much research has gone into generative models which use meaningful latent representations, but the sensitivity of human observers to smallest aberrations in the output makes it exceptionally challenging to produce realistic text, images or audio.

Variational Autoencoders (VAEs) [24] (see Figure 1) have a specific focus on finding meaningful representations for the data they are presented with in an unsupervised manner and

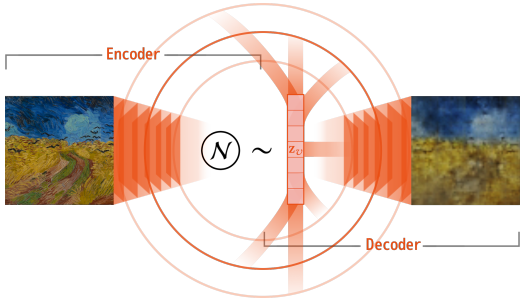

Figure 1: Visual VAE Architecture. An image is encoded to parametrize a Normal distribution from which the latent vector z_v is sampled. Adherence to the spherical prior constrains its values to a consistent range. The output is reconstructed based on z_v.

are therefore well suited for our task. Using a generative component called a decoder $\boldsymbol{x}' \sim p_{\text{dec}}(\boldsymbol{x}'|\boldsymbol{z})$ which, given the latent representation $\boldsymbol{z}$ of a data point $\boldsymbol{x}$, attempts to reproduce the underlying data realistically as $\boldsymbol{x}'$. The quality of said reconstruction is measured by the expectation of the original $\boldsymbol{x}$ being generated by the decoder distribution given $\boldsymbol{z}$.

$$\mathbb{E}_{\boldsymbol{z}\sim p_{\text{enc}}(\boldsymbol{z}|\boldsymbol{x})}[\ln p_{\text{dec}}(\boldsymbol{x}|\boldsymbol{z})] \tag{1}$$

Typically, the expectation is approximated by reconstruction measures such as the Mean Squared Error (MSE) or the L_1 norm of the difference between $\boldsymbol{x}$ and $\boldsymbol{x}'$. More importantly however, we see that $\boldsymbol{z}$ is sampled from yet another distribution. This encoder $\boldsymbol{z} \sim p_{\text{enc}}(\boldsymbol{z}|\boldsymbol{x})$ is typically modelled as a multivariate Gaussian with diagonal covariance of which the mean $\boldsymbol{\mu}$ and scale $\boldsymbol{\sigma}$ are determined using a model $f^{\theta}_{\text{enc}}(\boldsymbol{x})$ with learned parameters θ (see Equation 2). The encoder distribution is therefore parametrized differently depending on the original input $\boldsymbol{x}$, resulting in a distinct latent representation $\boldsymbol{z}$.

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}I) \quad \text{with} \quad \boldsymbol{\mu}, \boldsymbol{\sigma} = f^{\theta}_{\text{enc}}(\boldsymbol{x}) \tag{2}$$

Since a sampling operation is non-differentiable, reparametrization is used to produce an accurate sample while keeping the chain of gradients intact. This is achieved by simply drawing a noise vector $\boldsymbol{\epsilon}$ from an independent Gaussian distribution and scaling the parameters produced by the encoder such that the sampling operation becomes Equation 3. In effect, the encoder is forced to distribute the area occupied by latent representations of $\boldsymbol{x}$ continuously around $\boldsymbol{\mu}$ since it is not guaranteed a direct mapping of $\boldsymbol{x}$ to $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}$. At the same time, the decoder has to learn that latent vectors in a certain range encode the same or at least very similar images.

$$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} * \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I) \tag{3}$$

In order to produce distinct latent vectors, $f^{\theta}_{\text{enc}}(x)$ might simply learn to produce means which are as far apart as possible, leading to an undesirably uneven latent space. Therefore the Kullback-Leibler divergence (KL divergence) between the encoder distribution and a canonical prior of the same distribution (e.g. $p_{\text{prior}}(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, I)$) is added as a regularizing loss term (see Equation 4). This ensures that whatever mean and scale values may be produced, they should lie close enough to the prior such that differences in the data should only be represented within the context of the latent space which is specified by the prior distribution.

$$\text{KL}(p_{\text{enc}}(\boldsymbol{z}|\boldsymbol{x}) \parallel p_{\text{prior}}(\boldsymbol{z})) = \int_{\boldsymbol{z}\in Z} p_{\text{enc}}(\boldsymbol{z}|\boldsymbol{x}) \ln \frac{p_{\text{enc}}(\boldsymbol{z}|\boldsymbol{x})}{p_{\text{prior}}(\boldsymbol{z})} \tag{4}$$

All of this leads us to the so-called Evidence Lower Bound (ELBO) loss $\mathcal{L}_{\text{ELBO}}$ which encompasses both the reconstruction quality of the encoder-decoder pair $\mathcal{L}_{\text{rec}}$ as well as a constraint on how far the encoder may stray from the prior distribution $\mathcal{L}_{\text{lat}}$:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} &= \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{lat}} \\ \mathcal{L}_{\text{rec}} &= -\mathbb{E}_{\boldsymbol{z}\sim p_{\text{enc}}(\boldsymbol{z}|\boldsymbol{x})}[\ln p_{\text{dec}}(\boldsymbol{x}|\boldsymbol{z})] \\ \mathcal{L}_{\text{lat}} &= \text{KL}(p_{\text{enc}}(\boldsymbol{z}|\boldsymbol{x}) \parallel p_{\text{prior}}(\boldsymbol{z})) \end{aligned} \tag{5}$$

This has several important implications: For one, sampling from a latent embedding which is close to another should produce similar outputs due to the fact that $p_{\text{dec}}(\boldsymbol{x}'|\boldsymbol{z})$ assigns it a higher probability than if they were farther apart. The encoder's sampling operation as well as $\mathcal{L}_{\text{lat}}$ further ensure that given enough training time and data, the latent embeddings should flow into each other such that the latent space becomes continuous around the prior distribution. In theory, sampling an embedding from anywhere within that space and passing it to the decoder should therefore yield a sensible generative output.

Additionally, sampling from in between embedding vectors by means of interpolation should produce output which is a coherent mix of all original input data points. This also allows for the calculation of attribute vectors for certain properties by subtracting the embeddings of inputs which do not have that property from those which do. This vector can then be applied to data points to either include (addition) or exclude (subtract) that property.

For our purposes, the continuous latent spaces learned by VAEs are especially valuable. They provide us with a degree of certainty that the latent representations encode some consistent meaning and that intermediate points in the latent space characterise meaning differences continuously. Since we are furthermore trying to work with disparate spaces with respect to modality, it becomes even more important that if one space is mapped into another, meaning differences in one translate to consistent meaning differences in the other.

Although we prioritise the continuity of the latent space, this comes at the cost of reconstruction quality as outputs produced by the VAE's decoder tend to be blurrier. This is because $\mathcal{L}_{\text{lat}}$ can be minimized by mapping all data as close as possible to the prior and by compensating for the less distinct reconstructions through ambiguity (i.e. collapsing all reconstructions to their average). Additionally, constraints on the latent space may need to be relaxed if data is more complex than can be modelled by the spherical Gaussian prior. The β-VAE architecture [16] therefore incorporates an upper bound τ on the $\mathcal{L}_{\text{lat}} < \tau$ term, resulting in the reformulated loss:

$$\mathcal{L}_{\beta} = \mathcal{L}_{\text{rec}} + \beta\mathcal{L}_{\text{lat}} \tag{6}$$

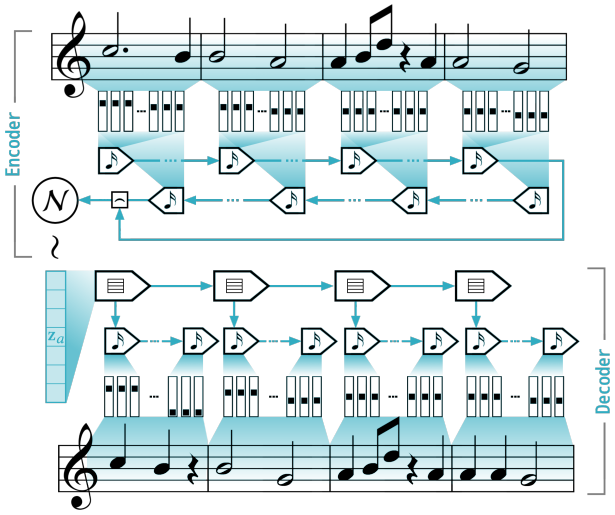

Figure 2: MusicVAE Architecture. 1/16th notes in a one-hot piano-roll are encoded using a bidirectional RNN to parametrize a Normal distribution. The latent vector z_a sampled from the encoder initializes the hierarchical decoder's conductor RNN which in turn initializes the output generator RNN.

in which the β hyperparameter is tuned according to the complexity of the task at hand. A $\beta > 1$ would encourage stronger adherence to the prior while values < 1 allow for more flexibility. Since it is difficult to encode pixel-level detail into our short audio representations, we are more interested in higher level features of the input images. Blurrier reconstructions therefore represent an acceptable trade-off for consistent latent spaces.

2.2 MusicVAE

Although VAEs are typically used in a visual setting, recent work has found them to be applicable to music as well. Focusing on both the quality of the output as well as the consistency of the latent space, MusicVAE [42] provides another indispensable piece of groundwork for our task. Due to the availability of its pre-trained latent space and decoder, it alleviates us from the need to train and tune our own musical generative model from scratch.

The VAE architecture behind MusicVAE (see Figure 2) uses Recurrent Neural Networks (RNNs) with Long Short-Term Memory cells (LSTM) [17] to encode and decode MIDI music representations. A "generator" RNN sequentially produces 1×90 note distributions at each step until one bar of music consisting of 16 notes has been generated. The melodies produced are monophonic and each one-hot output over all N notes is produced by sampling from a softmax distribution $\sigma(\boldsymbol{o}, \tau)$. It is parametrised by the note probability distribution output of the RNN $\boldsymbol{o}$ and a temperature parameter $\tau \in (0, 1]$.

$$\sigma(o_n, \tau) = \frac{\exp(\frac{o_n}{\tau})}{\sum_{i=0}^{N} \exp(\frac{o_i}{\tau})} \tag{7}$$

Here τ controls the degree to which the distribution approaches a one-hot categorical, a value of 1 preserving the original distribution while a smaller value shifts the distribution even more towards higher probability notes. A value of $\tau = 0.5$ as used in the original paper therefore produces note sequences which follow a more coherent pattern which is desirable for producing the final output melodies. During training however, the default value of 1 is retained in order to optimise the actually generated distribution in a smooth and proportional manner.

Once the output note sequence has been sampled, each of these one-hot vectors is converted into the corresponding Musical Instrument Digital Interface (MIDI) format. This involves setting the pitch of the corresponding 1/16 note as well as controlling for continuous note presses as opposed to repeated attacks and adding pauses where specified.

Each full bar of music is generated sequentially using a generator network, but the initialising input which it receives at the beginning of each bar is produced by a higher level RNN called the "conductor". This network takes the original music's encoded latent embedding as its initial input and produces initial states for the lower note-level generator RNN after each consequent bar. This hierarchical approach was found to preserve longer term dependencies much better than a flat architecture using only the lower-level generator. This enables MusicVAE to produce music with up to 16 bars (i.e. 256 note slots across approximately 30 seconds) using piano melodies, bass, drums and combinations thereof.

Although hierarchical embedding architectures are also technically possible, most publicly available pre-trained models make use of a flat bidirectional RNN architecture. Its final forward-directional output is concatenated with the final backward-directional output to produce the mean and scale of a multivariate Normal distribution from which the latent embedding with a consistent dimensionality of 512 across all model types is sampled. The VAE-specific properties related to latent space consistency appear to hold such that interpolation between the embeddings of two separate pieces of music produces new musical output which realistically lies between them. This consistency in addition to the availability of pre-trained models allows us to focus on the actual task of audio-visual translation and makes MusicVAE an ideal candidate for our experiments.

3 RELATED WORK

To the best of our knowledge, no previous work attempts to generate music based on visual art automatically. The methods used in different parts of this task are nonetheless strongly tied to ongoing research. For example, the need to learn a latent space for visual art based on generated reconstructions relates closely to the field of image generation (see Section 3.1). Similarly, research on music generation corresponds to our goal of generating realistic music from a consistent latent space (see Section 3.2). Finally, we also take a look at efforts in the area of cross-modal models which aim to convey visual information in an auditory manner (see Section 3.3).

3.1 Image Generation

Tasks involving the generation of realistic images have been an active field of research, partially due to the wide availability of well curated datasets. In addition to common candidates such as MNIST [28], CIFAR [25] or ImageNet [5], famous paintings from before the 20th century have already largely come into the public domain. Permissibly licensed contemporary artworks are harder to come by, but a highly curated set of approximately 2 million artworks which are consistently annotated with content,

artwork type and simple sentiment is available as the Behance Artistic Media dataset (BAM) [46] and will be used for our final experiments involving visual art.

The most common tools used to process such data are Convolutional Neural Networks (CNNs) [27] which represent the state-of-the-art in creating latent representations from image data (e.g. AlexNet [26], VGG [43], Inception [44]). By inverting their convolutional operations, latent representations can also be iteratively up-sampled to 2D RGB images, thereby making them indispensable for image generation.

Generative Adversarial Networks (GANs) [12] have been especially successful in generating high-resolution images of remarkable quality [12] even when conditioned to produce conditional output not seen during training [13]. Using deep CNN architectures alone, it has even been shown to be possible to transfer an artistic style from one image to another [6, 11]. In this setting, a painting or similarly distinct style template is used to compute a representative latent vector. It is then applied to an unrelated image such that the output bears resemblance both to the style as well as the original image. This shows that strong artistic constraints can be realistically applied within one modality. While the quality of productions from these kinds of models is high, they lack explicit embeddings. Input vectors can of course be trained to represent good initialisations for certain kinds of output, but additional constraints are required in order to guarantee a consistent latent space.

VAEs have also enjoyed success in this field, but it has been notoriously difficult to tune them in order to produce similarly detailed results. At the cost of increased complexity, there have been attempts to find loss formulations which allow for higher quality reconstructions while maintaining the desirable quality of latent space continuity. Besides the aforefemntioned β-VAE architecture [16], PixelVAE [15] for instance has shown to be able to produce realistic reconstructions of faces using the CelebA dataset [30] while maintaining a continuous latent space.

Another approach using the so-called DRAW architecture [14] has shown how VAEs combined with RNNs can amplify each other's strengths by sequentially creating an image with selective attention. This process has been shown to work for both the MNIST [28] and StreetView house number dataset [37]. Due to the increased complexity of both PixelVAE and DRAW however, we decided to use a simpler β-VAE architecture as a starting point for image generation.

3.2 Music Generation

Generating music synthetically requires an equal amount of realism compared to image generation and also imposes its own additional challenges. While it is inherently a sequential task, the scale-differences between short-term and long-term dependencies can be enormous. Individual time-steps in an audio waveform may be minuscule and range in the milliseconds while the full musical composition they are a part of can last several seconds, minutes or hours. As such, typical sequential models such as RNNs have trouble maintaining consistency across both scales.

Nonetheless, direct waveform synthesis of 16 kHz audio samples with highly rated quality has been accomplished using WaveNet-like architectures [45]. Using deep CNNs at different timescales, they manage to maintain consistency at both the 1 millisecond and multi-second time scale. While focused mainly on speech synthesis, generating realistic single note waveforms using this architecture has also been accomplished with NSynth [8]. Unfortunately, this method comes with high computational complexity due to the high number of sampling operations over time. Tackling this problem in particular, GANSynth [7] has been introduced to reliably generate full waveforms in a single step by using a waveform representation which is less prone to shifts across convolutions and by using a progressive training technique [22].

Although research in the field of direct waveform synthesis is progressing quickly, the ability to generate spectrograms of full musical compositions is still a work in progress. On these longer scales, the current focus lies on generating not the waveforms themselves, but the notes which represent them. This allows for each individual the time step to be much larger (e.g. 1/16 notes). Using the MIDI piano-roll format, output representations can thus also be modelled with one-hot encodings for each instrument and pitch. A model using this representation can generate one note of music at a time until the full musical composition is completed.

Building on a Biaxial LSTM architecture [21], the DeepJ model [33] consists of an RNN architecture which is capable of composing MIDI music conditioned on certain composers and their respective styles. The synthesized reconstructions were rated highly by human listeners and the conditioned style was also confirmed by music experts. Since the model was able to produce polyphonic melodies over long stretches of time, RNNs seem appropriate for this type of music generation. However this method alone does not suffice for our purposes, since the conditioning of this synthesis process is not precise enough to model complex information such as images. Similarly, the recently introduced MuseNet [38] which uses sparse transformers [4] to generate highly realistic, multi-minute polyphonic MIDI output, can be conditioned on certain styles, composers and combinations thereof, but also lacks consistent embeddings for the music itself.

Finally but crucially, sufficient amounts of musical data unfortunately very often remain copyrighted such that associated research only publishes meta-data instead of the original training files. Notable exceptions to this rule include the Lankh MIDI Dataset (LMD) [39] with about 180,000 songs and the Saarland Music Data [35] which includes 50 piano pieces donated by music students. To avoid the music data acquisition problem entirely, we will instead make use of the knowledge stored in the pre-trained MusicVAE (see Section 2.2) which has already learned to generate and encode music based on a proprietary dataset consisting of ~1.5 million songs and provides us with the best balance between performance and latent space consistency.

3.3 Audio-Visual Models

Prior research on audio-visual models focuses on improving model performance through multi-modal information as well as on improving the accessibility of visual information through audio representations.

Exploring the capability of generative models to translate from one sensory modality into another, attempts have been made to translate videos of instruments being played into the corresponding sound and vice versa [3] as well as to generate realistic background audio for visual scenes (e.g. rustling leaves

for a forest video) in order to provide visually impaired people with relevant semantic information for a given image [47]. Although some artefacts remained in the generated images and audio, the correct audio-visual correspondence was typically maintained when evaluated by human annotators. This confirms that cross-modal transformations in the audio-visual domain are indeed possible. However, these methods did not produce longer pieces of music and did not enforce latent space consistency in any way.

Focusing on cross-modal latent spaces in particular, the Partitioned Variational Autoencoder (PVAE) [18] has been successfully used to encode MNIST digits paired with audio of their spoken counterparts. This model uses one latent space per modality and joins them during decoding. It was shown that the multi-modal information aided in making the embeddings of different digits more distinct while also maintaining continuity of the latent space and allowing for interpolation, even between visual and auditive information. Models with VAEs at their core are therefore able to work with multi-modal information simultaneously.

The aforementioned methods lend credibility to the fact that generative models can be used in a cross-modal fashion in the audio-visual domains. However, these approaches have one commonality which does not apply to paintings and music: paired datasets. Additionally, the final quality of the model can only be measured with the aid of human evaluators. A representative number of participants and data points must therefore be available in order for generalizable conclusions to be drawn from the results.

Faced with a similar problem Hu et al. (2019) [19] therefore propose an automated cross-modal evaluation procedure for measuring the expressiveness of assistive devices which represent visual information in audible form. While the so-called vOICe device [34] used in their research uses a direct mapping of positional pixel values to high/low and loud/quiet audio instead of synthetic melodic output, their evaluation methodology involves cross-modal GANs which attempt to reconstruct images based on their automatically generated auditive representations. The reconstructions are then classified using pre-trained models such that their accuracy can be used as an indicator for the retained cross-modal information content. In their additional human evaluations, these metrics correlate strongly with human scores. This shows that even without existing paired datasets, it is possible to evaluate cross-modal models in a quantitative and meaningful manner.

Given the existing research, we can conclude that audio-visual translation using cross-modal generative models is a viable endeavour and that quantitative and qualitative evaluation metrics can be used in conjunction to draw representative conclusions. Our model will attempt to improve upon these findings by learning to produce longer, melodic representations of complex visual information in an unsupervised manner while maintaining a consistent latent space.

4 SYNESTHETIC VAE

Translating information across the audio-visual modal boundary requires a synesthetic approach. In the following, we introduce the SynVAE architecture which is capable of maintaining cross-modal information consistency and circumvents the lack of paired image-music datasets by learning latent representations in an unsupervised manner. SynVAE as described in this section is therefore applicable to any cross-modal transformation task.

By initially treating each modality separately and by using single-modality models which have unrestricted access to abundant high quality data in their respective domains, we are able to remove the need for subjective correlations of images and music and are able to follow a fully unsupervised approach. Equipped with models from both the audio and the visual domain, we can construct the pipeline of our fully synesthetic model SynVAE which is outlined in Figure 3.

Initially, the visual encoder $p_{\text{venc}}(\boldsymbol{z}_v|\boldsymbol{x})$ creates a 512 dimensional latent representation $\boldsymbol{z}_v$ from the original image $\boldsymbol{x}$ using a CNN architecture based on the single-modality VisVAE. As is the case with all VAE-based models, $\boldsymbol{z}_v$ is not computed directly, but rather sampled from a distribution which is parametrized by the visual encoder network. In this case, this distribution is a multivariate Normal $\mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\sigma}_v I)$ with mean $\boldsymbol{\mu}_v$ and diagonal covariance $\boldsymbol{\sigma}_v$. Akin to the single-modality case, the sampling operation of $\boldsymbol{z}_v$ itself is performed through reparametrization (see Equation 3) in order to maintain differentiability.

Since this vector has the same dimensionality as the MusicVAE latent space, it provides the initial state for the pre-trained music decoder $p_{\text{adec}}(\boldsymbol{a}|\boldsymbol{z}_v)$ which then produces a melodic output $\boldsymbol{a}$ using its hierarchical conductor-generator architecture. These two components make up the overall synesthetic encoder, the input of which is the original image $\boldsymbol{x}$ and the output of which corresponds to its audio representation $\boldsymbol{a}$. During inference, this stand-alone encoder $p_{\text{senc}}(\boldsymbol{a}|\boldsymbol{x})$ can be used to perform the audio-visual transformation (see Equation 8).

$$\boldsymbol{a} \sim p_{\text{senc}}(\boldsymbol{a}|\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}_v \sim p_{\text{venc}}(\boldsymbol{z}_v|\boldsymbol{x})}[p_{\text{adec}}(\boldsymbol{a}|\boldsymbol{z}_v)] \tag{8}$$

Training the model however requires a differentiable loss formulation which quantifies the difference in information content before and after the latent space transformations. Therefore the audio output is first re-encoded as $\boldsymbol{z}_a$ using the pre-trained bidirectional MusicVAE encoder RNN $p_{\text{aenc}}(\boldsymbol{z}_a|\boldsymbol{a})$ into the 512 dimensional music latent space. Once again, this auditive latent vector is computed by sampling from a multivariate Normal parametrized by the auditive encoder network. It is then passed through VisVAE's CNN decoder $p_{\text{vdec}}(\boldsymbol{x}'|\boldsymbol{z}_a)$ to produce an image reconstruction $\boldsymbol{x}'$. These two components which in conjunction transform audio $\boldsymbol{a}$ into a corresponding image $\boldsymbol{x}'$ make up the synesthetic decoder $p_{\text{sdec}}(\boldsymbol{x}'|\boldsymbol{a})$ in Equation 9.

$$\boldsymbol{x}' \sim p_{\text{sdec}}(\boldsymbol{x}'|\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{z}_a \sim p_{\text{aenc}}(\boldsymbol{z}_a|\boldsymbol{a})}[p_{\text{vdec}}(\boldsymbol{x}'|\boldsymbol{z}_a)] \tag{9}$$

Both latent vectors in this model are sampled from the same type of distribution and are also trained to express differences in the data while staying close to the same canonical prior (i.e. $\mathcal{N}(\mathbf{0}, I)$) through the KL divergence term in the ELBO loss (Equation 5). This actively encourages the latent spaces to follow a similar and consistent shape across modalities when compared to an unregularized training procedure and is therefore at the core of our unsupervised approach.

The architecture requires an expressive musical latent space with high variability. For this reason, the weights of the pre-trained MusicVAE model remain fixed throughout the entire training process as it is assumed that the overall model is able

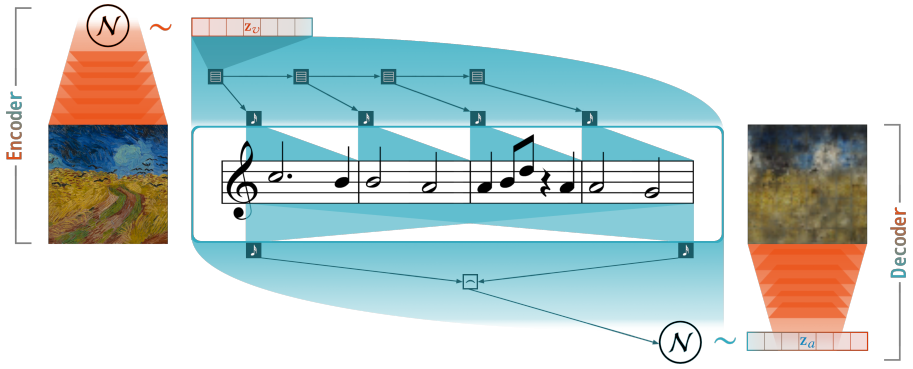

Figure 3: Synesthetic VAE Architecture. An image is first encoded into a latent vector z_v before being decoded into music. The music in turn is re-encoded into z_a and reconstructed into the output image.

to encode different images into areas of the latent space which are distinct enough to enable meaningful reconstructions.

In addition to ensuring stable audio generation, the frozen music components also allow for the total loss formulation to remain almost identical to the visual-only case (Equation 6). This is due to the fact that, firstly, audio reconstruction quality need not and cannot be measured due to the general absence of an audio-visual ground truth. Secondly, since the auditive encoder's parameters are not being updated and it was trained using the same canonical prior as the visual components, no additional regularising KL constraint on its distribution is required. The differentiable basis $\mathcal{L}_{\text{syn}}$ for the optimisation process therefore only consists of a KL constraint on the visual encoder, in addition to the comparison of the synesthetic decoder's reconstruction against the original image:

$$\begin{aligned} \mathcal{L}_{\text{syn}} &= \mathcal{L}_{\text{srec}} + \beta\mathcal{L}_{\text{slat}} \\ \mathcal{L}_{\text{srec}} &= -\mathbb{E}_{\boldsymbol{a}\sim p_{\text{senc}}(\boldsymbol{a}|\boldsymbol{x})}[\ln p_{\text{sdec}}(\boldsymbol{x}|\boldsymbol{a})] \\ \mathcal{L}_{\text{slat}} &= \text{KL}(p_{\text{venc}}(\boldsymbol{z}_v|\boldsymbol{x}) \parallel p_{\text{prior}}(\boldsymbol{z}_v)) \end{aligned} \tag{10}$$

The β parameter once again controls the balance between reconstruction quality and adherence to the canonical prior. In the synesthetic case this carries additional importance since the actively trained visual components cannot stray too far from the prior without risking to land in undefined music space. Within this pipeline, the auditive components therefore act as strong regularizers on the overall model since they are kept fixed during training and therefore pre-determine the expressiveness of the audio-visual latent space. The remaining trainable parameters of the model's visual components can then either be learned from scratch or be initialised using a pre-trained visual VAE which is known to produce good image reconstructions. Since only the parameters of one modality are trained, we are able to uncouple the simultaneous use of paired image and audio data and are also able to maintain a simple optimisation target.

Furthermore, by replacing the visual or auditive components of this model, it would be possible to extend this approach to different modalities as well. For instance, switching out the auditive and visual components of this model results in a SynVAE which encodes musical data into corresponding visual representations. Regardless of the modality pair, as long as the encoders are regularized with matching prior distributions and the central generative component (i.e. the music decoder in our experiments) is capable of producing realistic results across its latent space, the cross-modal translation is likely to succeed. Since advances in single-modality models are being made constantly, this architecture could keep on improving in parallel with these higher quality generative models given that these offer consistent latent spaces.

5 METHODOLOGY

Considering the lack of established methodology for the entirety our task, we approach each partial problem with well proven methods before combining them into our final solution. In Section 5.1 we therefore first describe the implementation and training procedures behind both the single-modality β-VAE architectures for encoding images (VisVAE) and the pre-trained MusicVAE before finally combining these visual and auditive models into a single synesthetic pipeline. Since novel methods of evaluating this approach are required, we lay out both the quantitative and qualitative procedures in Sections 5.2 and 5.3 respectively.

5.1 Implementation

As described in Section 4, SynVAE is constructed from initially independent single-modality components which are combined into the final model. For the task at hand, this means that it is necessary to first construct appropriate VisVAEs and then combine them with the pre-trained MusicVAE models.

Learning a latent space for simple images such as from MNIST and CIFAR-10 is a common baseline task for VAEs (e.g. [14, 15, 24, 41]). As such, we also begin by training a purely visual VAE and by tuning the β hyperparameter in order to obtain the appropriate trade-off between reconstruction quality and KL divergence. The encoder follows a typical CNN architecture with a different number of convolutions and filters depending on the complexity of the task at hand. The final CNN output is passed into two separate fully connected layers which independently produce the mean and scale vectors which parametrize the encoder's multivariate Gaussian distribution.

Using reparametrization (see Equation 3), the latent vector is sampled from said distribution and passed to the decoder. The embedding then passes through a fully connected layer followed by deconvolutional layers which mirror the encoder's architecture until the original image dimensions are restored in the final layer. Each pixel in the final layer is normalized

in $[0, 1]$ using a sigmoid activation, thereby representing the decoder's output distribution.

Due to the aforementioned unavailability of ground truth paired music data for the visual datasets, as well as to avoid training a complex generative model for music from scratch, we rely on pre-trained models from the MusicVAE project [42]. Their general architecture is described in Section 2.2. Although we found their complex 16-bar melodic model to fit well within our architecture, our final experiments were run using the large 2-bar melodic architecture because of the latter's leaner architecture and the fact that its shorter outputs are easier to digest for our human evaluators. Due to the brevity of the output, this model neither uses a hierarchical encoder, nor decoder and instead uses a flat RNN structure. The encoder hereof uses the aforementioned bidirectional LSTM structure with a 2×2048 output length vector. Passing this vector through two separate fully connected layers produces the mean and scale parameters of a Gaussian from which a latent representation of size 512 is sampled. This embedding is decoded using 3 stacked LSTM cells with 2048 units each, until a sequence of 32 notes (i.e. 2 bars of music) have been generated. This yields a final melodic audio with a length of approximately 4 seconds.

After tuning the single-modality models to generate results of sufficient quality in their respective domains, they are transferred into SynVAE. Their encoder and decoder architectures (i.e. the number of weights and how these are connected) are kept fixed and for the auditive components, the values of these weights are frozen as well. For the visual components which continue to be optimized during SynVAE training, we found that using the VisVAE's weights as initialization slightly improved the convergence time for more complex datasets, but that this was not strictly necessary in order for it to produce meaningful reconstructions.

The software implementation itself is built upon the TensorFlow framework [1] and is written in Python 3. The full source code for SynVAE is available at https://personads.me/x/synvae-code.

5.2 Quantitative Evaluation

The metrics by which the synesthetic model's performance is measured should ideally reflect how strong correspondences between similar images and their generated music are. As with any generative model, measuring this quantitatively is quite difficult. Our gold standard is therefore set by the ability of human evaluators to extract correspondence information between similar images and audio. Nonetheless, we also use quantitative metrics to measure overall correlative effects and in order to pre-select representative samples for the final human evaluation.

One principal quantitative metric is already present in the loss formulation $\mathcal{L}_{\text{syn}}$, namely the reconstruction error term $\mathcal{L}_{\text{srec}}$ (see Equation 10). Although the pixel-by-pixel difference measured by MSE is by no means an ideal way to measure image similarity, differences between models nevertheless indicate the degree to which information may be lost within the latent embedding space.

Furthermore, using the labels available in the visual datasets, we are able to measure how well the latent representations encode semantic similarity by calculating the nearest Euclidean neighbours of each encoded data point in both the visual as well as the auditive latent spaces and by evaluating the number of neighbours with the same label in the closest 10, 5, and 1 results (i.e. precision at rank n). Although our models are not trained on these labels, images of the same class nonetheless share visual similarities. This is especially the case for simpler datasets such as MNIST. Higher precision may therefore also indicate more expressive latent representations. By measuring the difference in precision before and after the audio transformations, we also gain an additional quantifiable insight into the amount of information encoded in the generated music.

This method assumes a correlation between labels and MSE pixel similarity since this is the main metric which the model is optimizing for. For more complex datasets and even CIFAR-10, this may be insufficient since, for instance, not all cars have the same colour. Therefore we attempt to bridge this gap using the metric of reconstruction classification accuracy. Using the same network architectures as for the visual encoders, we exchange the final layer with a simple softmax output which corresponds to the class probabilities assigned to the input image. This classification network is then trained and tested on the original dataset to produce a baseline and then re-initialized, retrained and evaluated using the reconstructed dataset produced by VisVAE and SynVAE. The resulting classification accuracy can then also be used as a quantitative metric for the retained information content.

Finally, we are strongly interested in the degree to which the visual and auditive latent spaces within the synesthetic model overlap. As a prior, they both share a zero-centred multivariate Gaussian with unit variance and as such they should not lie to far away from its spherical shape. In principle, images lying closer together in the visual latent space should also lie closer to each other in the auditive latent space forming similar clusters. Since we do not have paired images and audio however, there is no immediate way to measure this.

Instead we make use of Mutual Information Neural Estimation (MINE) [2] in order to approximate a lower bound on the mutual information $I(\mathbf{Z}_v; Z_a)$ of corresponding visual and auditive latent vectors, $\boldsymbol{z}_v \in Z_v$ and $\boldsymbol{z}_a \in Z_a$. This is achieved by maximizing the score of true audio-visual latent pairs (i.e. $p(Z_v, Z_a)$) while minimizing the same value for non-matching latent vectors (i.e. $p(Z_v)p(Z_a)$). The estimation of the values themselves can be performed using any function $T_\theta(\boldsymbol{z}_v; \boldsymbol{z}_a)$ and is typically chosen to be a neural network with weights θ. In our experiments, it is a small estimator network with one 128 unit and one 64 unit layer. While the original MINE objective uses N latent vector pairs as negative samples, we opt for the more data efficient and less biased formulation of DEMINE [29] which uses N^2 negative pairs (see Equation 11).

$$
\begin{aligned}
I(Z_v; Z_a) \geq & \frac{1}{N} \sum_{i=1}^{N} [T_\theta(\boldsymbol{z}_{v,i}, \boldsymbol{z}_{a,i})] \\
& - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} [e^{T_\theta(\boldsymbol{z}_{v,i}, \boldsymbol{z}_{v,j})}] + 1
\end{aligned}
\tag{11}
$$

Since we do not use this objective for training SynVAE or tuning its hyperparameters, we can optimize on the entire dataset and use T_θ's estimation as the actual mutual information's lower bound. Higher values would indicate that true audio-visual latent vector pairs share more information than unrelated pairs. This allows us to calculate a quantitative metric based on solid information theoretical principles which can then be

used to compare synesthetic models amongst each other with regard to their cross-modal latent space consistency.

5.3 Qualitative Evaluation

All of the aforementioned quantitative metrics are useful indicators of our model's performance. In the end however, we are most interested in whether human listeners can correlate similar images and audio in the same way as SynVAE would predict. For instance, given that the model embeds two images close to each other in latent space, the audio generated from them should also be perceptibly similar. Evaluating the entire corpus covering many thousands of images is however not feasible and as such, a pre-selection of representative samples is a practical necessity. Determining representative data points manually would also involve processing the dataset in its entirety and would additionally introduce the selector's bias into the evaluation. We therefore propose a more consistent and reproducible approach.

For labelled image data, we are able to compute the mean latent vector per class by averaging the embedded latent representations of all in-class data. By measuring the distances between class means we can then determine which of them are most distinct from each other according to the model. Choosing very disparate classes should yield cases in which data is easier to distinguish from each other than if their means were located closely together. Given this information, we can further differentiate between samples which lie closer to the mean of their class than others (i.e. avoid outliers).

The human evaluators can therefore now be presented with a smaller subset of representative image-audio pairs which lie relatively close to their respective class means to get an idea of what certain classes sound like. After an initial training phase using such example data points, they can then be tasked to match an audio to the image which generated it by choosing from a list of options. The difficulty of this task can further be varied by choosing options which are embedded closer or farther away from each other in the latent space. Using this qualitative evaluation methodology, it is possible to determine whether the model's latent space predictions line up with human notions of similarity in a more reproducible manner.

In our application of this approach, the 3 class means with the highest cumulative distance are used. Based on them, the 24 closest data points each are collected. From these points, 4 examples per class are randomly selected and presented to the evaluators at the beginning of the evaluation for a total of 12 example image-audio pairs (note that evaluators do not have access to these examples after this initial stage). The other 60 samples are presented in 20 tasks with three distinct options each. The corresponding audio of one of these three options is randomly chosen as the truth value and evaluators are tasked to identify the image from which this particular audio was generated. To avoid any further bias, the options' ordering is shuffled randomly across participants as well. During this process, no direct feedback is given regarding whether a choice was correct or not and thus each person has to rely on their musical intuition and short-term memory to identify the correct answer. The more distinct and intuitive the classes sound and the more internally coherent they are, the better the performance should be.

Finally, the accuracy with which the evaluators were able to identify the true audio-visual pairs generated by SynVAE is measured. Higher values should mean that people were able to correctly discern differences between the three presented classes by hearing alone. Lower values would mean that the model's cross-modal translation is not consistent enough for human listeners to accurately perceive. To surface further patterns in the annotations, we employ Fleiss' kappa [9] and measure whether certain types of errors occur more often than others.

6 EXPERIMENTS

Our experiments are aimed at increasingly complex visual datasets, beginning with the simple, but interpretable MNIST dataset [28] (see Section 6.2), continuing with the slightly more complex CIFAR-10 dataset [25] (see Section 6.3) and finally concluding with the highly diverse BAM dataset [46] (see Section 6.4). Each task follows a similar set-up which is described in Section 6.1 according to the previously outlined methodology (see Section 5). In an ablation study, we further analyse whether generated audio representations fall into a qualitatively acceptable music space and how this relates to hyperparameter settings and optimization objectives (see Section 6.5). Additionally, we highly recommend listening to selected audio-visual examples from each dataset on https://personads.me/x/synvae.

6.1 Set-up

All experiments follow a similar pipeline: First, visual β-VAEs (VisVAEs) are trained on the respective dataset, tuned on the validation split and evaluated on the held-out test set. Next, they can be used to initialise the synesthetic VAE which is trained using the methodology explained in Section 4. It too is trained, validated and tested on the respective splits of the same data. For each task, β-values in $[0.1, 2.0]$ were tested in a grid search pattern.

The quantitative evaluation involves measuring the reconstruction error term (MSE), KL divergence from the canonical prior in addition to metrics of the latent spaces and image reconstructions. Using the methods outlined in Section 5.2, the class consistency within the n nearest neighbours of each data point's latent representation is measured. Overall mutual information between latent spaces is estimated using MINE [2] and the quality of the reconstructed images is measured through the accuracy with which they can be classified.

After running the quantitative evaluation and determining which model parameters are best suited for the task at hand, a qualitative evaluation task is generated based on the similarity metrics of the visual model. This task is then loaded into a web-based evaluation tool for the final evaluation by human annotators. The code for the tool itself is available at https://personads.me/x/syneval-code.

Unless otherwise noted, all models (including MINE and reconstruction classifiers) are trained for 100 epochs using early stopping on the validation set. This takes approximately 8 hours for VisVAEs, 30 hours for SynVAEs, 4 hours for the classifiers and 1 hour for MINE on a single GPU instance.

6.2 MNIST

MNIST [28] is a typical dataset used for baseline evaluation which consist of 70,000 monochromatic 28×28 images of individual, handwritten digits in a 6:1 training and testing split. By splitting the training data into 50,000 and 10,000 images, we create an additional validation set for tuning. Although the original images are provided in grayscale, we make use of the

Model	MSE	KL	P@10	Acc	MINE
VisVAE ($\beta = 0.1$)	5.46	74.69	0.47	0.99	-
VisVAE ($\beta = 0.5$)	16.43	24.38	0.31	0.99	-
VisVAE ($\beta = 1.0$)	22.48	14.79	0.23	0.99	-
SynVAE ($\beta = 0.1$)	29.31	46.33	0.27	0.98	5.02
SynVAE ($\beta = 0.5$)	36.66	15.63	0.28	0.96	5.03
SynVAE ($\beta = 1.0$)	42.73	8.69	0.26	0.94	5.02

Table 1: MSE, KL divergence, precision at rank 10 (P@10) and classification accuracy (Acc) for VisVAE and SynVAE models on MNIST test set given different β values. Additionally, estimated mutual information (MINE) between SynVAE's visual and auditive latent spaces.

commonly used additional binary simplification of the data to pixels with intensity 1 or 0.

Due to the small size of the images, some models take flattened versions of the images as input, but in order to emulate the more complex models used later on, we use a convolutional architecture (see Appendix A.3 for details). Latent vectors are sampled from the encoder-specified distribution using reparametrisation (see Equation 3) and have a dimensionality of 50 (or 512 in the synesthetic case). Using the decoder's output, MSE reconstruction loss is computed and added to the KL divergence term to produce the total loss $\mathcal{L}_{\text{syn}}$ as defined in Equation 10. Optimisation is performed using Adam [23] and a learning rate of 10^{-3}.

6.2.1 Quantitative Evaluation. Quantitative metrics for both the single modality and the synesthetic VAE architectures with varying β values are provided in Table 1 with additional results in Appendix A.2, Table 4. We report MSE, KL divergence, nearest neighbour precision at rank 10 (for SynVAE, embeddings are drawn from the final auditive latent space) and classification accuracy on the reconstructed images. First and foremost, the β parameter's effect on reconstruction quality is clearly recognisable for all model types. Lower values result in lower MSE at the cost of higher KL divergence. In general, synesthetic models with corresponding β have higher reconstruction error rates than their VisVAE counterparts and higher MSE overall. The transformation of the images into music space therefore results in a definite reduction of visual fidelity. Looking at reconstructions from either model type (see Appendix A.1, Figure 10), it is indeed the case that SynVAE's digits are blurrier than in the visual-only case.

KL divergence however is lower for corresponding synesthetic models than for visual ones (in fact, up to a factor of 61%). We attribute this to the fact that the fixed auditive component of these models acts as a strong regularizer on the remaining weights such that even without strongly weighting the KL loss term, adherence to the prior distribution can be more easily enforced.

The effect which the regularization of the latent distribution has on embeddings and reconstructions is slightly more difficult to interpret. For the visual models reconstruction quality seems to be stable across β as classifiers trained and evaluated on reconstructed digits consistently achieve 0.99 classification accuracy. This score is identical to the baseline classifier using original images which also achieves an accuracy of 0.99. Nearest neighbour precision however shows more variance between models, the highest score being obtained by the visual model with the lowest $\beta = 0.1$.

In order to provide further insights into this phenomenon, it helps to compare results from the synesthetic case. Nearest neighbour precision for the SynVAE models lies around 0.27 which is mostly lower than for their VisVAE counterparts, but classification accuracy is rather comparable at 0.94 to 0.98, meaning that digit classes in the reconstructions are still identifiable. Comparing reconstructions, we can observe reasons for why these effects may occur: while the VisVAEs are not only able to reproduce clearer images than the SynVAEs, they also seem to embed more digit details such as rotation and writing style. This means that although both reconstructions are recognizable as the correct digit, SynVAE generates more general representations.

MINE values are consistent across SynVAEs at around 5.02, clearly indicating shared information across the visual and auditive latent spaces. Since a mutual information of $\ln(10) \approx 2.3$ would at least be sufficient to encode MNIST's class information over 10 digits, the additionally available nats are likely to encode more detailed style information of the images. Although this metric is consistent across βs, the differences in MSE, P@10 and Acc would nonetheless indicate that some fidelity is lost. Most importantly however, cross-modal latent representations seem to share a higher level of consistency.

6.2.2 Qualitative Evaluation. In our qualitative experiments, we examined whether the quantitative indications of audio-visual consistency are perceived by humans as well. By embedding the test set into the VisVAE latent space, we first identify the digits which are most visually distinct. For MNIST, all VisVAE agree upon the same three digits: "0", "1" and "4". The $\beta = 0.1$ distinguishes these classes most strongly by embedding them at a cumulative Euclidean distance of 12.16. This in addition to this model's high nearest neighbour P@10 points towards a clearer separation of digit clusters which is why we chose it to generate the evaluation task.

For the choice of SynVAE we especially focused on a balance between reconstruction quality and KL divergence. During initial listening tests, we found that if the SynVAE's KL divergence is too high, the music it generates becomes more erratic and amelodic, thereby making it more difficult for listeners to discover coherent patterns within the data. We examine this phenomenon in further detail in Section 6.5. In this task, the $\beta = 0.5$ SynVAE strikes this balance best as its reconstruction quality is still met with 96% classification accuracy while its KL divergence is a third of the $\beta = 0.1$ model. Although the $\beta = 1.0$ model in turn has approximately half the KL divergence of the $\beta = 0.5$ model, this comes at the detriment of all other metrics. As such, the latter model seems to be best suited for the task at hand.

Following the methodology outlined in Section 5.3 we can now further select the most representative samples of the evaluation classes "0", "1" and "4". In Figure 4 these images in addition to 1000 randomly selected data points are shown in the tSNE [32] projections within VisVAE and SynVAE auditive latent space respectively. From the highlighted means, we can see that the largest cumulative distance of the class triplet is maintained for both the synesthetic and single-modality case.

Comparing these two projections, we also see more of the quantitative metrics reflected. For one, the clusters formed by the VisVAE embeddings show clearer separations while the

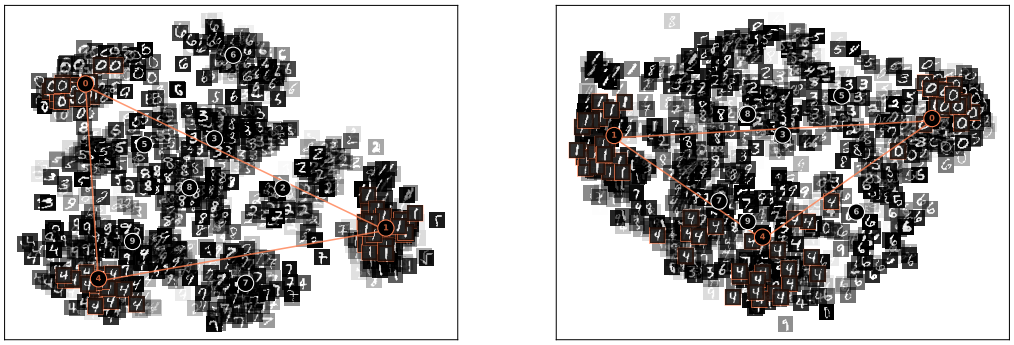

Figure 4: TSNE plots of MNIST VisVAE latent space with $\beta = 0.1$ (left) and SynVAE auditive latent space with $\beta = 0.5$ (right) with highlighted evaluation samples.

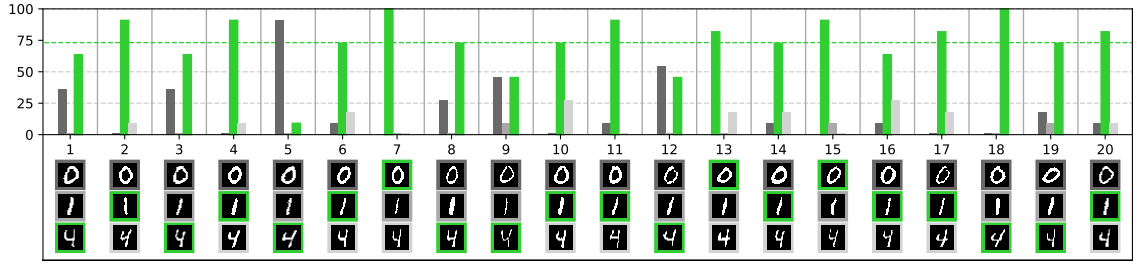

Figure 5: Percentages of evaluator choices per class ("0", "1", "4") and task on the qualitative MNIST evaluation. Correct options highlighted in green. Average accuracy marked at 73%.

SynVAE clusters are less clear cut and have some samples appearing far away from their class means. Some class-internal style characteristics are also represented within the clusters. For example, within the "1" cluster, the skew of the vertical line making up the digit increases when moving up and away from the mean. In the synesthetic space, these diagonal "1"s also transition and mix with diagonal "8"s. Additionally, digits such as "4" and "9" which are already visually similar on a macro-level are embedded much closer to each other than in the VisVAE space. This once again highlights how SynVAE is able to maintain higher-level characteristics of the images, but lacks fidelity when it comes to finer details within the classes as corroborated by both the P@10 and Acc metrics.

To check whether human evaluators could nonetheless discern classes from each other given the auditive representation of images produced by SynVAE, we conducted a classification study with 11 participants (using the method described in Section 5.3). After being presented with 12 audio-visual example pairs, they were presented with a single audio and were tasked to identify which of the three images was used to generate it.

Results from this task are presented in Figure 5. The participants achieved a mean accuracy of 0.73 across 20 tasks with a maximum evaluator score of 0.95 and a standard deviation of 0.22. For two of the tasks all evaluators unanimously paired up the correct image and audio. The overall inter-annotator agreement as measured by Fleiss' kappa is also consistent at around 0.48.

Subjective feedback by the annotators showed that differences in pitch between classes were most easily identified. Further deductions were based on major/minor chord separation and the tempo of a composition as well as the overall number of notes. In general, images of the digit "1" were identified with the least ambiguity and an in-class precision of 0.96 since they seem to consistently follow a major chord with lower pitch. "0"s are of a higher pitch and tempo while "4"s seem to be mid-range in pitch and tend to follow a minor chord. The discrimination between the latter two classes seems to have been the most difficult as reflected in their lower in-class precisions of 0.43 for "0"s and 0.75 for "1"s. Indeed, incorrect choices typically involved a switch-up between these two classes as seen in task 9 and 12 for instance. At an extremum, 91% of evaluators mistook a "4" for a "0" in task 5.

A closer look at per class quantitative metrics lines up with these observations. "1"s achieve the highest P@10 scores for both VisVAE (P@10: 0.55) and SynVAE (P@10: 0.51). "0" images trail slightly behind with 0.53 P@10 for VisVAE and 0.46 P@10 for SynVAE. Meanwhile, "4"s are at the lower end with 0.43 P@10 for VisVAE and 0.16 P@10 for SynVAE.

Nonetheless, we still find that some "0"s and "4"s can be distinguished with high accuracy, as can be seen in tasks 7 and 18 in which all evaluators unanimously made the correct connection. Tasks 13 and 15 also show above average accuracy with regard to "0" classification. Considering that evaluators only got to see 4 audio-visual example pairs per class at the very beginning of the task and nonetheless achieved a relatively high accuracy shows that the cross-modal translations of SynVAE are consistent and intuitive enough for human listeners to understand and keep in their memory.

Model	MSE	KL	P@10	Acc	MINE
VisVAE ($\beta = 0.1$)	19.75	122.70	0.12	0.91	-
VisVAE ($\beta = 0.3$)	31.77	58.01	0.11	0.90	-
VisVAE ($\beta = 1.0$)	52.70	22.30	0.11	0.89	-
SynVAE ($\beta = 0.1$)	74.05	69.81	0.13	0.87	5.17
SynVAE ($\beta = 0.3$)	80.55	30.90	0.13	0.87	5.17
SynVAE ($\beta = 1.0$)	88.76	14.14	0.13	0.86	5.14

Table 2: MSE, KL divergence, precision at rank 10 (P@10) and classification accuracy (Acc) for VisVAE and SynVAE models on CIFAR-10 test set given different β values. Additionally, estimated mutual information (MINE) between SynVAE's visual and auditive latent spaces.

Figure 6: TSNE plot of CIFAR-10 SynVAE auditive latent space with $\beta = 0.3$.

6.3 CIFAR-10

In order to increase complexity compared to the MNIST task while retaining a smaller number of distinct classes, our second set of experiments focuses on the CIFAR-10 image dataset [25] consisting of $32 \times 32 \times 3$ RGB images which are labelled with 10 mutually distinct classes. It provides a training set of 50,000 images which we further separate into 40,000 training and 10,000 validation images. A 10,000 image test set is additionally used for analyses and evaluation. All pixel values are scaled from $[0, 255]$ to $[0, 1]$ for easier processing.

Similarly to the MNIST visual model, we employ a symmetric CNN encoder decoder architecture, the details of which are provided in Appendix A.3. While the final latent dimensionality of 512 remains the same to ensure compatibility with the auditive components of SynVAE, the remaining networks have more weights and are also deeper than for MNIST to account for the increased image complexity. Loss is calculated based on Equation 10 as before and optimised using Adam [23] with a learning rate of 10^{-3}.

Since our focus lies on the musical interpretation of visual art, we used the experiments conducted on CIFAR-10 as an intermediate stepping stone between the simple, yet not very artistic, MNIST digits and the artistic, but far more diverse and complex, BAM artworks. As such, the effort required for the qualitative evaluation involving human participants was skipped for CIFAR-10 in favour of the final BAM data and we will concentrate on the quantitative metrics alone, the most interesting of which are presented in Table 2 with additional results in Appendix A.2, Table 5.

With a higher overall MSE, reconstructing CIFAR-10 already presents itself as a visually more complex task than the monochrome MNIST digits. The β hyperparameter once again adequately controls the reconstruction quality to prior adherence trade-off with lower values producing clearer reconstructions at the cost of higher KL divergence. As with the MNIST experiments, the visual-only VAE architectures achieve crisper reconstructions in general, the highest MSE amongst them still being almost 30% lower than for the best performing SynVAE (see reconstruction in Appendix A.1, Figure 11). It can therefore be surmised that a certain loss of image fidelity is unavoidable when passing the same information through music space.

Nonetheless, that same fixed, auditive latent space provides its own regularisation on the distribution from which the intermediate latent embeddings are sampled. Even with $\beta = 0.1$ which permits a KL divergence of 122.70 for the VisVAE model, the corresponding SynVAE remains solidly bounded at 69.81. All other β settings also show that the synesthetic model is able to maintain KL values close to half of that of their single-modality counterparts.

Precision at rank 10 as averaged over the 10 classes in the dataset is unfortunately less expressive than for the previous MNIST case. Regardless of the type of model or its hyperparameters, P@10 ranges from around 0.11 (VisVAE) to 0.13 (SynVAE), indicating performance similar to what a random baseline would achieve. We attribute this difference to the fact that while digits are fairly reliably identifiable by means of pairwise MSE alone, the more complex CIFAR images are not. Since this reconstruction error is the only metric which our models use to approximate visual similarity, it stands to argue that semantic similarity as defined by the CIFAR-10 class labels does not correlate strongly with pixel value differences. Looking closer at P@10 per class, this seems further indicated by the fact that images of classes with common visual attributes achieve higher scores than those with more variance: air planes for instance are typically visible on the backdrop of a blue sky and ships are found on bodies of water. Corresponding to the lower class-internal MSE, both classes have slightly higher P@10 scores of up to 0.22.

Although the embeddings of our models do not seem to conform to their semantic labels, identifying class characteristics are maintained in the reconstructions as evidenced by the relatively high classification accuracies between 0.86 and 0.91 (0.93 for the baseline classifier). Even when the baseline classifier which was trained on the original images is tested on reconstructions from the VAEs accuracy, differences remain within a -0.07 bound. The β parameter does not seem to have a very strong effect on these scores regardless of model type, but a slight drop-off of up to 0.04 in accuracy can be observed from visual to synesthetic models. As corroborated by the higher MSE of SynVAE models and the previous observations for MNIST, we can therefore conclude that while some amount of information is lost in the audio-visual translation, consistency can be maintained to such a degree that semantic classes can still be identified fairly accurately.

Measuring audio-visual mutual information more directly using MINE, we observe values which are comparable and even slightly higher than for MNIST. Similar to the classification accuracy which measures the inner-class consistency of reconstructions, measuring the consistency of individual pairs of auditive and visual latent representations shows relatively high mutual information content across β. The $\beta = 1.0$ SynVAE

achieves a slightly lower MINE score of 5.14 when compared to the other two which indicates that it may become more difficult to encode audio-visual pairs more distinctly, the closer the latent space is pushed towards the canonical prior. Considering that there is more information to encode within CIFAR-10 more distinct latent representations are preferable. At the same time excessively high KL divergence is detrimental to the quality of the generated music (see Section 6.5). It may therefore be even more important to find a balance between the model's expressiveness and its adherence to a joint prior distribution than for simpler and more distinct data such as MNIST.

Finally, a look at the embedded position of test images within the final auditive latent space of the $\beta = 0.3$ SynVAE in the tSNE plot of Figure 6 can provide an intuition about the information actually retained across modalities. The loss in visual fidelity which results as a consequence of passing data through music space inhibits the model's ability to encode object details for distinguishing between a deer in front of green foliage from a car in front of a forest or a black cat on a bright background from a car on a white background, the overall colour differences are nonetheless maintained across the entire space. Even in this two-dimensional representation, we can therefore still observe how bright images with dark objects, objects on blue skies, predominantly green nature images and similar colour-based groups are clustered together and transition smoothly into each other. The fact that these visually similar groups do not strongly overlap with the semantic classes certainly marks a loss in information content, but the audio-visual pairs generated by the model should nonetheless provide valuable higher-level information. For visually complex datasets such as this one, the best strategy for tuning SynVAE may therefore be to focus on improving higher-level consistency while finding a compromise with joint latent space consistency as measured by the KL divergence.

6.4 BAM

The Behance Artistic Media dataset (BAM) [46] contains ~2.3m annotated contemporary works of visual art and allows us to evaluate SynVAE on a very complex and diverse dataset. Annotations concern 9 content classes, 4 emotions and 7 media types and were added automatically based on a model trained on ~400k crowd-sourced ground truth labels. These follow a system of "positive", "unsure" and "negative" certainties for which the authors cite 90% accuracy [46].

To simulate the environment of a painting exhibition, we filter the original set down to oil paint and watercolour art which leaves us with around 200k images of which ca. 180k original images were still available to be retrieved from the Behance online platform [20]. These images were then split into a 115k training set, 29k validation set and a 36k held-out test set. In order to speed up training and since detailed reconstruction quality is of a lesser concern to us, we scaled and cropped the images to $64 \times 64 \times 3$ with randomly positioned crops for the training data and consistent centred crops for validation and test splits. As with the previous experiments, pixel values were scaled from $[0, 255]$ to $[0, 1]$. The mirrored CNN encoder-decoder architecture is slightly larger than for MNIST and CIFAR-10 (see Appendix A.3 for details) and is once again optimised using Adam [23] with a learning rate of 10^{-3}.

For the evaluation, we focus on the provided emotion labels "happy", "scary", "gloomy" and "peaceful". Since one image can be labelled with multiple emotions, each with its own certainty, and filtering data down to "positive" labels alone leaves us with only a fifth of the data in very unbalanced splits (90% "peaceful" and 2% "scary"), there is a need to employ slightly different pre-processing steps for each type of quantitative evaluation.

Model	MSE	KL	P@10	Acc	MINE
VisVAE ($\beta = 0.1$)	142.54	426.12	0.24	0.82	-
VisVAE ($\beta = 0.3$)	188.30	189.29	0.24	0.82	-
VisVAE ($\beta = 1.0$)	257.14	69.48	0.23	0.80	-
VisVAE ($\beta = 1.3$)	273.08	56.34	0.23	0.80	-
VisVAE ($\beta = 1.7$)	291.17	45.17	0.22	0.79	-
SynVAE ($\beta = 0.1$)	397.06	171.69	0.23	0.78	5.22
SynVAE ($\beta = 0.3$)	426.45	73.11	0.23	0.77	5.16
SynVAE ($\beta = 1.0$)	452.47	30.65	0.25	0.77	5.15
SynVAE ($\beta = 1.3$)	455.16	27.89	0.25	0.77	5.16
SynVAE ($\beta = 1.7$)	461.32	23.36	0.24	0.77	5.16

Table 3: MSE, KL divergence, precision at rank 10 (P@10) and classification accuracy (Acc) for VisVAE and SynVAE models on BAM test set given different β values. Additionally, estimated mutual information (MINE) between SynVAE's visual and auditive latent spaces.

Reconstruction classification is treated as a multi-label problem with "uncertain" labels rounded up to "positive" in order to ensure sufficient data quantity. Meanwhile, nearest neighbour precision is treated as a multi-class problem. This means that images which are both gloomy and scary are treated as one class "gloomy+scary" and only images that are exclusively scary are marked as "scary". For this metric we also round "uncertain" values to "positive" which results in a total of 16 distinct emotion classes. Including uncertain labels however comes with the drawback of less distinct image clusters since annotations for highly subjective properties such as emotion are bound to conflict each other. For generating qualitative evaluation tasks, we therefore only use images which are exclusively labelled as "positive" in order to filter out potentially ambiguous data.

6.4.1 Quantitative Evaluation. The quantitative metrics presented in Table 3 (additional results in Appendix A.2, Table 6) reflect the increased complexity of BAM compared to both MNIST and CIFAR-10, in that MSE and KL divergence are higher across the board. Trends within the BAM models are however consistent with previous observations. Adjusting the hyperparameter β in order to weight the adherence to the standard Normal prior distribution behaves as expected and can therefore be used to find a suitable model for further qualitative evaluation.

As before, the regularization by the auditive components of SynVAE remain in effect such that the KL divergence can reach values which are up to 61% lower than for VisVAEs with equal β values. The larger MSE term nonetheless requires β to be set slightly higher than in the previous experiments in order to balance out its stronger influence. We find that a $\beta \geq 1.0$ is necessary to constrain SynVAE to a KL divergence which approaches the values which were found to work well for CIFAR-10 and MNIST while not compromising too much on the amount of high-level information that is retained.

Visual information which conveys emotion typically seems to be overall colour (e.g. scary dark images) and content (e.g.

peaceful landscapes). Based on the observations of CIFAR-10, one would expect that SynVAEs would tend to encode the colour information, but would lose detailed content, while VisVAEs should be able to retain a certain level of both colour and content details. The latter models' higher reconstruction classification accuracy seems to concur with this hypothesis since the lower β VisVAEs consistently achieve accuracies of 0.80 to 0.82 (closest to the 0.84 baseline) while the best SynVAE is slightly behind with 0.78. These gains of the single-modality models could be explained by the fact that atypical cases, such as a bright scary image with the thin outline of a skull, would retain the emotion-distinguishing content while a SynVAE would likely loose this information and reconstruct a predominantly white image without the outline. Although this increased level of detail is beneficial to classification performance, the difference its absence makes is acceptable. Furthermore, the higher-level features of the SynVAE reconstructions still allow for a consistent classification with around 0.77 across all examined β (see Appendix A.1, Figure 12 for a visual comparison).

Within latent space, this difference is even less pronounced, as P@10 lies between 0.22 and 0.25 for all models. Looking at the tSNE plots of both VisVAE and SynVAE latent space in Figure 7 shows why this might be the case. As with CIFAR-10, the strongest measure of visual similarity seems to be overall colour. Semantic classes can therefore only be encoded as far as they are correlated with said high-level information. Although emotions are connected with an image's colour to some degree, this alone is not sufficient to guarantee that it is embedded close to images of the same class. The precision with which such an embedding is possible seems to average out at the aforementioned ~0.25 range since this number is consistent across both the visual and synesthetic VAEs. This, in addition to the comparatively high classification accuracy, reaffirms what has been observed before: SynVAE loses low-level details during the translation into music space, but is able to encode high-level information to a comparable degree as its single modality counterpart.

Across modalities, MINE confirms that latent representations share strong consistencies amongst each other. Given the higher degree of flexibility the $\beta = 0.1$ model has in making more distinct representations, its estimated mutual information is also higher at 5.22. The remaining SynVAEs share similar MINE scores of ~5.16, pointing towards a stable lower bound for mutual information within the evaluated range of β values. This provides a high degree of certainty regarding the ability of SynVAE to encode both differences and similarities in the visual data consistently in the auditive space.

6.4.2 Qualitative Evaluation. Although the emotion classes of BAM are not as easily identified as the digits of MNIST, the VisVAEs nonetheless agree upon a triplet of class means which are especially distinctive: "happy" ("h"), "scary" ("s") and "happy +peaceful" ("hp"). Although the joint class "gloomy+scary" is actually more distinct than "scary" alone, it only contains 12 data points with "positive" certainty. In order to generate a task with a sufficient number of data points, we therefore opt for the slightly less homogeneous, but next best, "scary" class.

The images selected by visual models for the evaluation task strongly overlap across different β values and share almost all data points (maximum difference of 3 images). Out of these models, the $\beta = 0.1$ VisVAE was used to generate the final task, since it embeds the class means furthest apart with a cumulative distance of 15.89 and also comes with the highest reconstruction classification accuracy and precision.

The choice of SynVAE is slightly more difficult to make, especially since the quantitative metrics lie very close to each other. Choosing a model with low KL divergence is nonetheless paramount since the generated music might be amelodic to human evaluators if it is too high (see Section 6.5). This leaves models with $\beta \geq 1.0$ since they have KL divergences which are closest to the MNIST SynVAE used for qualitative evaluation. While one could simply choose the model with the lowest KL divergence, we opted for the intermediate $\beta = 1.3$ since it achieves minimally higher P@10 as well as higher accuracy and precision scores when evaluated with the baseline classification model.

Making use of the $\beta = 0.1$ VisVAE and $\beta = 1.3$ SynVAE, the qualitative evaluation is performed using an identical set-up and procedure to the MNIST evaluation. 21 participants with low overlap to those in the MNIST trial were presented with 4 audio-visual examples per class. It is important to note that although the examples are ordered by their class, it is not explicitly stated which class is which. For MNIST digits, their class adherence is simple to recognize, but for the more abstract classes in BAM this may be more difficult. Furthermore, during the testing phase, the image options are shuffled each time such that class adherence cannot be inferred from their presentation.

Results for the 20 tasks in this trial are presented in Figure 8. Although the visual complexity of the task has risen significantly from the MNIST task, the evaluators are still able to achieve an average accuracy of 71% with a standard deviation of 0.13 and a maximum individual score of 95%. Fleiss' kappa is very close to that of MNIST with 0.46 and an actually greater number of unanimous agreements. In 5 tasks, all evaluators paired up the correct image to its musical representation.

A closer look reveals that 4 of these unanimous choices were made for "scary" images with predominantly dark colour and musical tones. This reflects the constellation of the VisVAE and SynVAE latent spaces shown in Figure 7 in which this class is furthest apart from the rest. The only other task in which the "scary" image is the correct choice is number 19 with a nonetheless high accuracy of 61%. Its correct audio-visual pair has a dark red hue which results in music which is not as fast or low as in the mostly black cases. Tasks 10 and 11 in which a "happy+peaceful" and a "happy" image were correctly identified by 90% and 100% of evaluators respectively, in addition to tasks 1, 2, 4 and 7 which also feature images from these classes and have accuracies $\geq$ 71%, further show that high accuracy is not necessarily limited to the most distinct "scary" class.

Larger errors typically occurred when the classes of two visually similar images were mistaken for each other. Tasks 5 and 6 for instance both feature audio generated from the "happy" image, however the numer of votes for the "happy+peaceful" image are almost equal in both cases. Looking at the options available, task 5 features images with largely white backgrounds and task 6 has more colourful images with blue skies for the two aforementioned classes. This higher visual similarity is bound to produce similar sounding audio which is more difficult to distinguish as is evident from the scores for these two tasks. The fact that none of the evaluators chose the darker "scary" image in both cases further shows that the degree to which the options differ visually, strongly affects the performance of evaluators when distinguishing between them by ear alone.

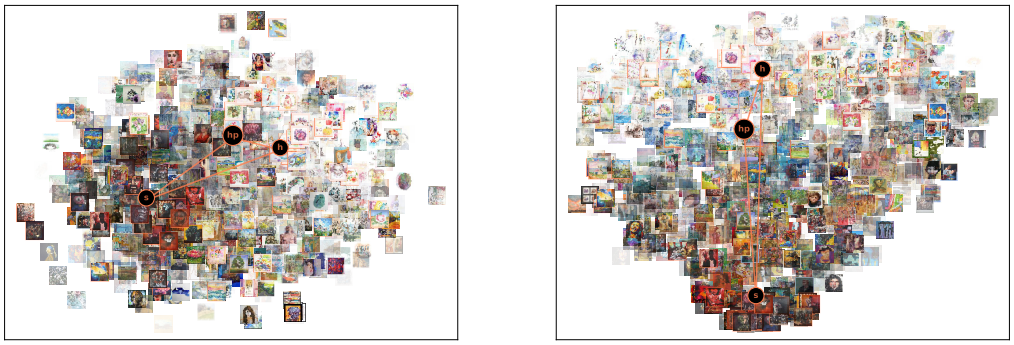

Figure 7: TSNE plots of BAM VisVAE latent space with $\beta = 0.1$ (left) and SynVAE auditive latent space with $\beta = 1.3$ (right) with highlighted evaluation samples.

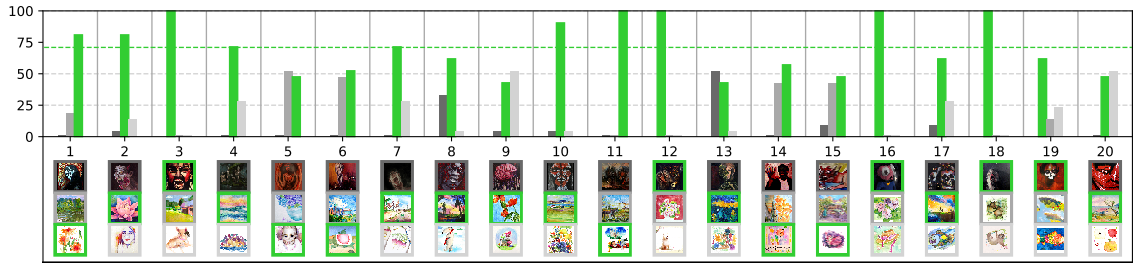

Figure 8: Percentages of evaluator choices per class ("scary", "happy+peaceful", "happy") and task on the qualitative BAM evaluation. Correct options highlighted in green. Average accuracy marked at 71%.

Another task in which this phenomenon becomes evident is number 13. Here, the correct choice is a "happy+peaceful" image with dark blue skies and an equally dark foreground and an orange area in-between. It was most often confused with the "scary" image which also features a dark background and foreground matter in a slightly orange tint. While these two images were confused frequently, this was not the case for the "happy" image which was drawn on a white background.

Considering the complexity of the BAM task and the coherency of the choices made by the human evaluators with respect to audio-visual consistency, it can be surmised that SynVAE is very much able to translate the visual artwork in this dataset into the musical domain with highly perceivable accuracy.

6.5 Music Space

In our experiments in Sections 6.2, 6.3 and 6.4, the qualitative metrics reflected a definite reduction in reconstruction quality when passing image information through the musical auditive latent space when compared to the single-modality VisVAE architecture. One might therefore propose setting $\beta \rightarrow 0$ in order to further prioritize MSE. Indeed, it seems plausible that the strong regularization imposed by MusicVAE might suffice to improve the reconstruction loss term while keeping KL divergence low enough to produce coherent audio-visual pairs. Unfortunately, we did not find this to be the case.

When lacking a strong enough KL constraint, SynVAE tends to push latent vectors far outside of the canonical prior. Although the KL term might not be as large as for a VisVAE with the same β-value, after a certain threshold, these vectors stray into undefined music space. To illustrate this effect, we sampled 100 latent vectors from a standard Normal distribution and used the single-modality MusicVAE to generate music based on them. Furthermore we scaled each of these vectors by factors 0.1, 0.5, 2.0, 5.0 and 10.0 to simulate a stronger or weaker adherence to the canonical prior. Due to the difficulty of measuring the "realness" of the generated music, we used the surface level note attack frequency feature to provide a rough estimation. The results using the 2-bar melodic model (i.e. 32 notes maximum) are shown in Figure 9.

The unscaled latent vectors produce a relatively evenly distributed set of melodies with around 3 to 15 notes (8% and 5% respectively) centred around a 12% peak at 7 notes and with the remaining percentages lying approximately in between. This seems reasonable for a total 2-bars of music since the length of notes may vary to fill out the full timeline. Scaling the vectors by a factor of 2 or 0.5 seems to roughly maintain this unaltered frequency distribution albeit not as smoothly. Moving further towards the 0 centred mean with factors 0.5 and 0.1, we see however that variance decreases until we arrive at a canonical melody of 7 notes with 99% probability.

Moving out further outside of the canonical space, as is the case when no strong KL bound is enforced, we observe for the factors 5 and 10 that the generated music spreads out to either a single note or two with around 25% and 14% probability or to the other extreme with a full set of 32 notes at around 10% - 14%. These melodies do not sound very realistic, since they either have only one note onset in the entire piece or are very fast without any pauses.

These measurements line up with observations from low β SynVAE and become even more pronounced as the KL loss term is omitted entirely. With such objectives, music becomes very

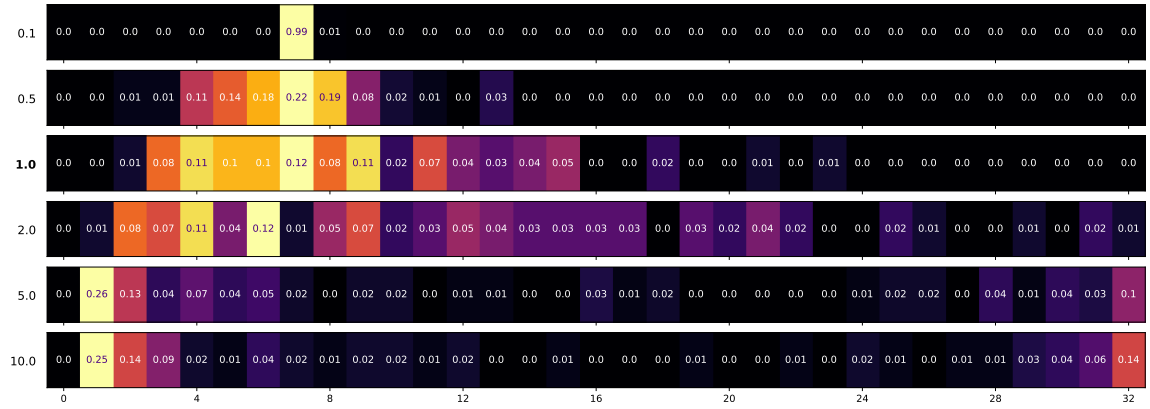

Figure 9: Percentages of note attack frequencies in 2-bar music generated from 100 latent vectors sampled from a standard Normal distribution and scaled by the factor on the y-axis.

fast and dissonant. Similarly, enforcing a KL constraint on the auditive components of the model in addition to the visual ones quickly lead to latent collapse with very non-distinct music unless learning rates and β were kept especially low.

When working across modalities it therefore seems necessary to balance reconstruction quality metrics sufficiently with the latent space's divergence from the cross-modal prior since it may have a strong effect on the realism of the output in the target modality. In our experiments, this would correspond to finding an optimal MSE-KL trade-off in order to maximize audio-visual correspondence while not compromising the perceived quality of the musical output.

7 DISCUSSION

In the following, we will summarize the results of all three experiments with respect to our goal of audio-visual consistency and assess how the quantitative metrics relate to the humanly perceived consistency measured in our qualitative studies.

Since SynVAE must learn to translate between sensory modalities in the absence of paired ground truth data, audio-visual consistency is its principal focus. Evaluating its performance with regard to this was shown to be difficult using a single quantitative metric alone. Precision metrics based on the nearest neighbours of a data point embedded in the learned latent space were indicative of consistency only if visual similarity was strongly correlated with semantic similarity as defined by the labelled image data. For MNIST this was found to be the case since the monochrome nature of the dataset as well as the relatively low variance between images of the same class allow for MSE to be an appropriate surrogate for the information content of an image.

For more complex datasets such as CIFAR-10, this metric was not as expressive. With an exponentially higher amount of information as well as a higher in-class variance, it becomes more difficult for SynVAE to encode lower-level details. Similarly, BAM increases the amount of pixels and the visual variance once more such that not all visual details are retained when they are passed through SynVAE's latent space. In this case however, the emotion labels tend to share a larger degree of correlation with higher-level image features (e.g. dark images being "scary") such that the precision at rank 10 is almost twice as high as that of CIFAR-10.

Although it is more difficult to interpret latent space consistency of the latter two datasets with nearest neighbour metrics alone, reconstruction classification provides a more flexible way to do so. It relies on semantic labels as well, but has the benefit that it is indicative of whether same-class image data is consistently encoded and decoded. Using the baseline classification model trained on unaltered images, it is further possible to measure how consistent reconstructed images are with respect to the original data. Across all experiments classification accuracy showed that while there is a certain degree of information loss when passing through latent spaces cross-modally, overall visual consistency of semantic classes is maintained.

Independent of labelled data, MINE allows us to quantitatively measure the consistency across latent spaces. The relatively high amount of mutual information between visual and auditive latent spaces show that SynVAE does indeed learn to embed a substantial amount of information consistently across modalities and different types of visual data. Because the estimated mutual information is slightly higher for CIFAR-10 and BAM than for MNIST, this might be pointing towards SynVAE learning to embed the more complex, higher variance datasets more distinctively.

Combining these three quantitative metrics, it becomes apparent that information content is being translated across modalities consistently. In the qualitative evaluations of MNIST and BAM, we have further shown that this theoretical consistency is actively perceived by human evaluators. Through the quantitatively informed pre-selection process, the evaluated data is ensured to be representative of similarity relations within each dataset. The high accuracy with which the evaluators were able to distinguish between the three most distinct classes shows that low-level information can be conveyed audibly for simple data and high-level information can be conveyed for more complex data as well. While there is a cross-modal information bottleneck, mistakes nonetheless line up with visual similarity. This in addition to the fact that evaluators are able to achieve up to 100% accuracy for very distinct data points confirms that audio-visual consistency is not only theoretical, but also very perceivable.

8 CONCLUSION

As shown by our results, it can be concluded with high confidence that SynVAE is able to consistently translate a wide range of images, including art, into the auditory domain of music through unsupervised learning mechanisms. In this final section, we will lay out potential areas of applications for SynVAE (Section 8.1) as well as highlight topics which might be interesting for future research (Section 8.2).

8.1 Applications

SynVAE as implemented here is already applicable in many practical scenarios. First and foremost, abstract visual information such as art can be translated into short musical pieces which convey high-level structures in a fast and intuitive way. Instead of listening to a long list of artwork titles and descriptions for instance, a museum visitor could simply scan through four second bits of music and decide which part of a collection they would like to experience most. This feature could not only be useful to visually impaired visitors, but to sighted ones as well. As a proof of concept, we have made a subset of ~1000 musically translated artworks from the Van Gogh Museum's permanent collection available on the project's website (https://personads.me/x/synvae). With a pre-trained model this translation process takes less than one minute to complete. While the shorter musical compositions generated in the scope of this research may not be enjoyable enough to be played long term at art exhibitions, they may nonetheless be helpful in providing intuitive insights into the nature of the visual artworks which are being encoded.

Depending on the domain of the data which the models were trained on, they can furthermore be used as a sonification tool in many more areas. Especially when abstract forms of visual data are in use, a short and intuitive auditive representation could help translate insights across sensory modalities. So while detailed schematics, graphs or figures would likely not be translated precisely enough, higher-level similarities and differences could still be retained. With better performing generative components which process data from a wider range of modalities, the scope of applications could furthermore be increased extensively.

8.2 Future Work

The possibilities for future research which are opened up by this synesthetic approach are manifold. Using the steadily advancing methods of music generation, the quality of cross-modal translation can be improved further in the future to provide longer, polyphonic pieces which are more distinctive and also more enjoyable to listen to. Such methods would of course still have to maintain latent space consistency, but using more advanced optimization targets based on reconstruction classification accuracy instead of MSE (if labels are available) or mutual information estimation between latent spaces instead of the KL divergence, could provide the necessary components to achieve this.

In the evaluation procedure so far, we have relied on the semantic labels provided in the image data. This can be sufficient if these classes line up strongly with visual features, but may become impractical if they do not or are not available at all. Training SynVAE would of course not be an issue since it does not rely on labelled data, but evaluation becomes difficult nonetheless. For such cases, it could be interesting to evaluate the performance of automatic clustering algorithms such as k-means [31]. By once again calculating these means in the VisVAE latent space, we circumvent potential self-confirmation bias which could arise if clusters are constructed in SynVAE space. In a small number of trial runs, we found that this method may be promising, but is highly sensitive to k as the number of clusters determines their internal variance and cross-cluster distinctiveness. Further investigation is therefore needed in order to achieve a fully unsupervised evaluation pipeline.

In the scope of a larger qualitative study with more participants and a larger number of tasks it would furthermore be interesting to explore more properties of the latent space and how they relate to human perception. By systematically presenting evaluators with samples which are closer or farther away from class means and by presenting options which are either very close to each other or more distant, the correlation between the theoretical difficulty defined by latent space distances and the actually perceived difficulty could be determined. Given a large number of representative tasks, these measurements could be used to further inform the relationship between quantitative and qualitative metrics in a synesthetic environment.

Additionally, this research has not yet made use of the space in-between data points. As VAEs guarantee a degree of latent space consistency and have been shown to interpolate well between data in single-modality settings [14, 15, 24, 42], this should also be explored in a synesthetic environment. The fact that we find consistency between similar data across modalities in our experiments lends strong credibility to this hypothesis.

Finally, the modular nature of SynVAE also allows for this approach to be extended to more modality-pairs. With the set-up used in this research, we could for instance theoretically generate digits or visual art based on musical melodies by exchanging the modalities of the model's encoder and decoder components. Since the architecture does not limit us to audio-visual data, it is furthermore possible to extend it to any modality pair for which high quality datasets in each respective domain exist. We hope that the evaluation methodology outlined in this research will provide a solid basis for measuring cross-modal information consistency, in addition to SynVAE itself enabling better access to that information across sensory boundaries.

ACKNOWLEDGEMENTS

This research would not have been possible without the precise insights and inexhaustible patience of dr. Nanne van Noord as well as the valuable input of Marco Federici and Gjorgji Strezoski. Additional thanks go to the volunteer evaluators for their time and the DAS-4 cluster [10] for providing the necessary computational resources.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R. Devon Hjelm, and Aaron C. Courville. 2018. MINE: Mutual Information Neural Estimation.

CoRR abs/1801.04062 (2018). arXiv:1801.04062 http://arxiv.org/abs/1801.04062

[3] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. 2017. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. ACM, 349–357.

[4] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating Long Sequences with Sparse Transformers. *CoRR* abs/1904.10509 (2019). arXiv:1904.10509 http://arxiv.org/abs/1904.10509

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[6] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2017. A learned representation for artistic style. *Proc. of ICLR* 2 (2017).

[7] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2019. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710* (2019).

[8] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. 2017. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. *CoRR* abs/1704.01279 (2017). arXiv:1704.01279 http://arxiv.org/abs/1704.01279

[9] Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33, 3 (1973), 613–619.

[10] Advanced School for Computing and Imaging. 2019. Distributed ASCI Supercomputer 4. https://www.cs.vu.nl/das4/. Accessed: 19th July, 2019.

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2414–2423.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[14] Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. 2015. DRAW: A Recurrent Neural Network For Image Generation. *CoRR* abs/1502.04623 (2015). arXiv:1502.04623 http://arxiv.org/abs/1502.04623

[15] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. 2016. Pixelvae: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013* (2016).

[16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, Vol. 3.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[18] Wei-Ning Hsu and James Glass. 2018. Disentangling by Partitioning: A Representation Learning Framework for Multimodal Sensory Data. *arXiv preprint arXiv:1805.11264* (2018).

[19] Di Hu, Dong Wang, Xuelong Li, Feiping Nie, and Qi Wang. 2019. Listen to the Image. *arXiv e-prints*, Article arXiv:1904.09115 (Apr 2019), arXiv:1904.09115 pages. arXiv:cs.CV/1904.09115

[20] Adobe Inc. 2019. Behance. https://behance.net/. Accessed: 3rd July, 2019.

[21] Daniel D Johnson. 2017. Generating polyphonic music using tied parallel networks. In *International conference on evolutionary and biologically inspired music and art*. Springer, 128–143.

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *CoRR* abs/1710.10196 (2017). arXiv:1710.10196 http://arxiv.org/abs/1710.10196

[23] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980 http://arxiv.org/abs/1412.6980

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

[25] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning multiple layers of features from tiny images*. Technical Report. Citeseer.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[27] Yann LeCun et al. 2015. LeNet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet* 20 (2015).

[28] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[29] Xiao Lin, Indranil Sur, Samuel A. Nastase, Ajay Divakaran, Uri Hasson, and Mohamed R. Amer. 2019. Data-Efficient Mutual Information Neural Estimator. *CoRR* abs/1905.03319 (2019). arXiv:1905.03319 http://arxiv.org/abs/1905.03319

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[31] Stuart P. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory* 28 (1982), 129–137.

[32] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[33] Huanru Henry Mao, Taylor Shin, and Garrison Cottrell. 2018. DeepJ: Style-specific music generation. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. IEEE, 377–382.

[34] Peter B.L. Meijer. 2019. vOICe - Augmented Reality for the Totally Blind. www.seeingwithsound.com. Accessed: 28th April, 2019.

[35] Meinard Müller, Verena Konz, Wolfgang Bogler, and Vlora Arifi-Müller. 2011. Saarland music data (SMD). In *Proceedings of the international society for music information retrieval conference (ISMIR): late breaking session*.

[36] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

[37] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. (2011).

[38] Christine Payne. 25th April, 2019. MuseNet. https://openai.com/blog/musenet/. Accessed: 3rd July, 2019.

[39] Colin Raffel. 2016. *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. Ph.D. Dissertation. Columbia University.

[40] Kyle Rector, Keith Salmon, Dan Thornton, Neel Joshi, and Meredith Ringel Morris. 2017. Eyes-Free Art: Exploring Proxemic Audio Interfaces For Blind and Low Vision Art Engagement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 93.

[41] Danilo Jimenez Rezende and Fabio Viola. 2018. Taming VAEs. *arXiv preprint arXiv:1810.00597* (2018).

[42] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. 2018. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. *CoRR* abs/1803.05428 (2018). arXiv:1803.05428 http://arxiv.org/abs/1803.05428

[43] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[44] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[45] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. *CoRR* abs/1609.03499 (2016). arXiv:1609.03499 http://arxiv.org/abs/1609.03499

[46] Michael J. Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. 2017. BAM! The Behance Artistic Media Dataset for Recognition Beyond Photography. In *The IEEE International Conference on Computer Vision (ICCV)*.

[47] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. 2018. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3550–3558.

A APPENDICES

A.1 Reconstructions

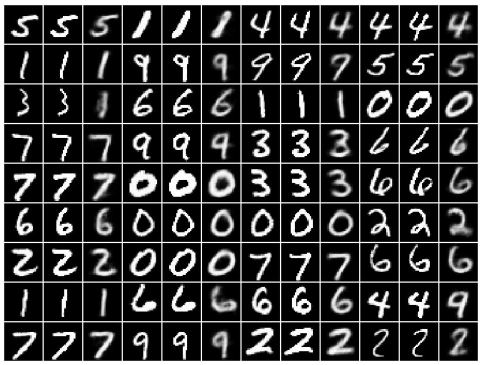

Figure 10: Original images, VisVAE ($\beta = 0.1$) and SynVAE ($\beta = 0.5$) reconstructions of 36 random data points in MNIST test set (left-to-right ordered triplets).

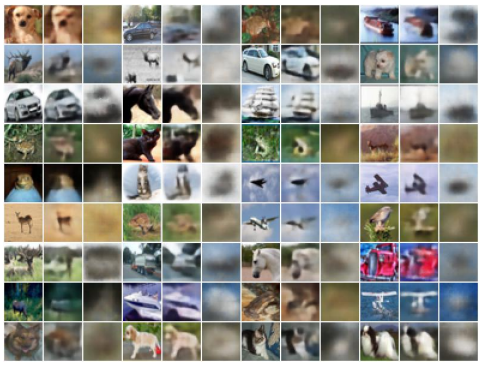

Figure 11: Original images, VisVAE ($\beta = 0.1$) and SynVAE ($\beta = 0.3$) reconstructions of 36 random data points in CIFAR-10 test set (left-to-right ordered triplets).

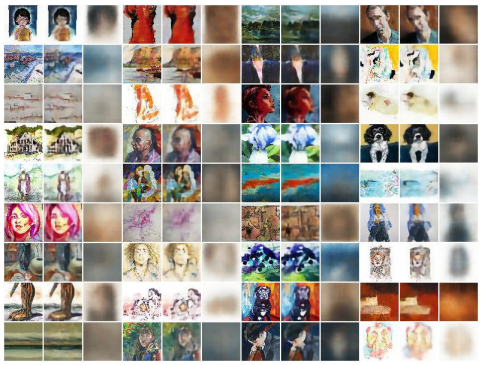

Figure 12: Original images, VisVAE ($\beta = 0.1$) and SynVAE ($\beta = 1.3$) reconstructions of 36 random data points in BAM test set (left-to-right ordered triplets).

A.2 Additional Results

Table 4 contains additional quantitative metrics for the MNIST task described in Section 6.2.

Model	P@1	P@5	ClAcc	ClPr
VisVAE ($\beta = 0.1$)	0.50	0.48	0.99 (0.99)	0.97 (0.97)
VisVAE ($\beta = 0.5$)	0.36	0.33	0.99 (0.99)	0.95 (0.96)
VisVAE ($\beta = 1.0$)	0.27	0.25	0.99 (0.98)	0.93 (0.92)
SynVAE ($\beta = 0.1$)	0.29	0.28	0.98 (0.97)	0.89 (0.88)
SynVAE ($\beta = 0.5$)	0.30	0.29	0.96 (0.95)	0.81 (0.79)
SynVAE ($\beta = 1.0$)	0.27	0.26	0.94 (0.90)	0.71 (0.68)

Table 4: Precision at rank 1 (P@1) and 5 (P@5), classification accuracy (ClAcc) and precision of classifier trained on originals (in brackets) or reconstructions, on MNIST test set.

Table 5 contains additional quantitative metrics for the CIFAR-10 task described in Section 6.3.

Model	P@1	P@5	ClAcc	ClPr
VisVAE ($\beta = 0.1$)	0.12	0.12	0.91 (0.91)	0.57 (0.56)
VisVAE ($\beta = 0.3$)	0.12	0.11	0.90 (0.90)	0.51 (0.47)
VisVAE ($\beta = 1.0$)	0.11	0.11	0.89 (0.87)	0.43 (0.34)
SynVAE ($\beta = 0.1$)	0.13	0.13	0.87 (0.85)	0.36 (0.26)
SynVAE ($\beta = 0.3$)	0.13	0.13	0.87 (0.85)	0.34 (0.24)
SynVAE ($\beta = 1.0$)	0.13	0.13	0.86 (0.85)	0.31 (0.22)

Table 5: Precision at rank 1 (P@1) and 5 (P@5), classification accuracy (ClAcc) and precision of classifier trained on originals (in brackets) or reconstructions, on CIFAR-10 test set.

Table 6 contains additional quantitative metrics for the BAM task described in Section 6.4.

Model	P@1	P@5	ClAcc	ClPr
VisVAE ($\beta = 0.1$)	0.24	0.23	0.82 (0.78)	0.78 (0.77)
VisVAE ($\beta = 0.3$)	0.25	0.24	0.82 (0.75)	0.78 (0.76)
VisVAE ($\beta = 1.0$)	0.23	0.23	0.80 (0.73)	0.76 (0.75)
VisVAE ($\beta = 1.3$)	0.23	0.23	0.80 (0.72)	0.75 (0.75)
VisVAE ($\beta = 1.7$)	0.23	0.22	0.79 (0.71)	0.76 (0.74)
SynVAE ($\beta = 0.1$)	0.23	0.23	0.78 (0.69)	0.72 (0.73)
SynVAE ($\beta = 0.3$)	0.24	0.23	0.77 (0.69)	0.72 (0.73)
SynVAE ($\beta = 1.0$)	0.25	0.25	0.77 (0.68)	0.72 (0.69)
SynVAE ($\beta = 1.3$)	0.25	0.25	0.77 (0.69)	0.72 (0.72)
SynVAE ($\beta = 1.7$)	0.24	0.25	0.77 (0.68)	0.71 (0.72)

Table 6: Precision at rank 1 (P@1) and 5 (P@5), classification accuracy (ClAcc) and precision of classifier trained on originals (in brackets) or reconstructions, on BAM test set.

A.3 Network Architectures

MNIST	CIFAR-10	BAM
$\boldsymbol{x} \in \mathcal{R}^{28\times28}$	$\boldsymbol{x} \in \mathcal{R}^{32\times32\times3}$	$\boldsymbol{x} \in \mathcal{R}^{64\times64\times3}$
RL(Cv(32, 3, 2))	RL(Cv(64, 3, 2))	RL(Cv(64, 3, 2))
RL(Cv(64, 3, 2))	RL(Cv(128, 3, 2))	RL(Cv(128, 3, 2))
FC(512)	RL(Cv(256, 3, 2))	RL(Cv(128, 3, 2))
	FC(4096) $\rightarrow$ FC(512)	RL(Cv(256, 3, 2))
		FC(4096) $\rightarrow$ FC(512)
$\boldsymbol{z} \in \mathcal{R}^{512/50}$	$\boldsymbol{z} \in \mathcal{R}^{512}$	$\boldsymbol{z} \in \mathcal{R}^{512}$
FC($7 \times 7 \times 32$)	FC($4 \times 4 \times 256$)	FC($4 \times 4 \times 256$)
RL(Cv(64, 3, 2))	RL(Cv(256, 3, 2))	RL(Cv(256, 3, 2))
RL(Cv(32, 3, 2))	RL(Cv(128, 3, 2))	RL(Cv(128, 3, 2))
σ(Cv(1, 3, 1))	RL(Cv(64, 3, 2))	RL(Cv(128, 3, 2))
	σ(Cv(3, 1, 1))	RL(Cv(64, 3, 2))
		σ(Cv(3, 1, 1))

Table 7: Network architectures for MNIST [28], CIFAR-10 [25] and BAM [46] VisVAEs. Convolutions denoted as Cv(filters, size, stride), fully-connected layers as FC(outdim), ReLU [36] as RL and sigmoid as σ.

The architectures used for the encoder and decoder networks for the MNIST [28], CIFAR-10 [25] and BAM [46] VisVAEs are shown in Table 7. Set-ups remained identical when they were used in the visual components of the respective SynVAEs. Reconstruction classification was done using models based on the encoder architecture, but with the final FC(512) layer being replaced by FC(N_{classes}) with a softmax activation. Due to the multi-label nature of BAM's classification, its final layer is capped with a sigmoidal activation.