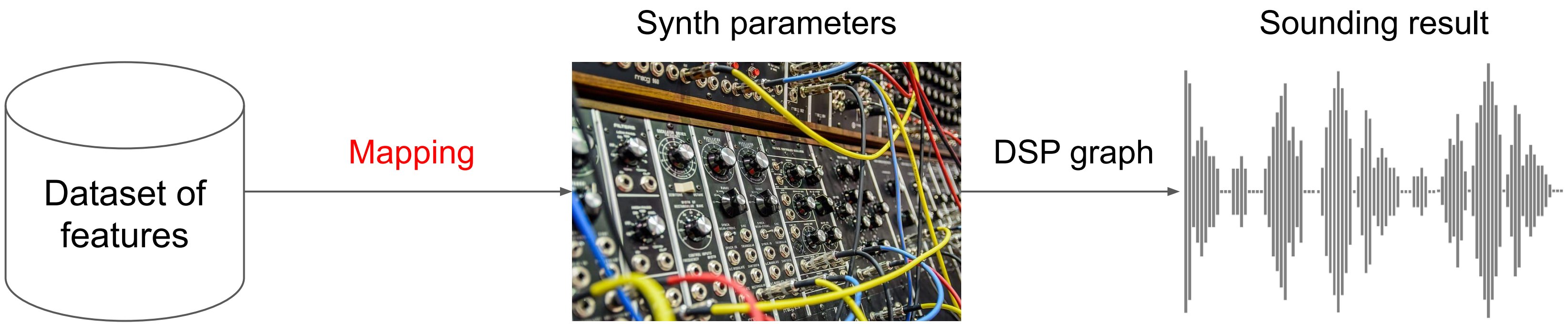
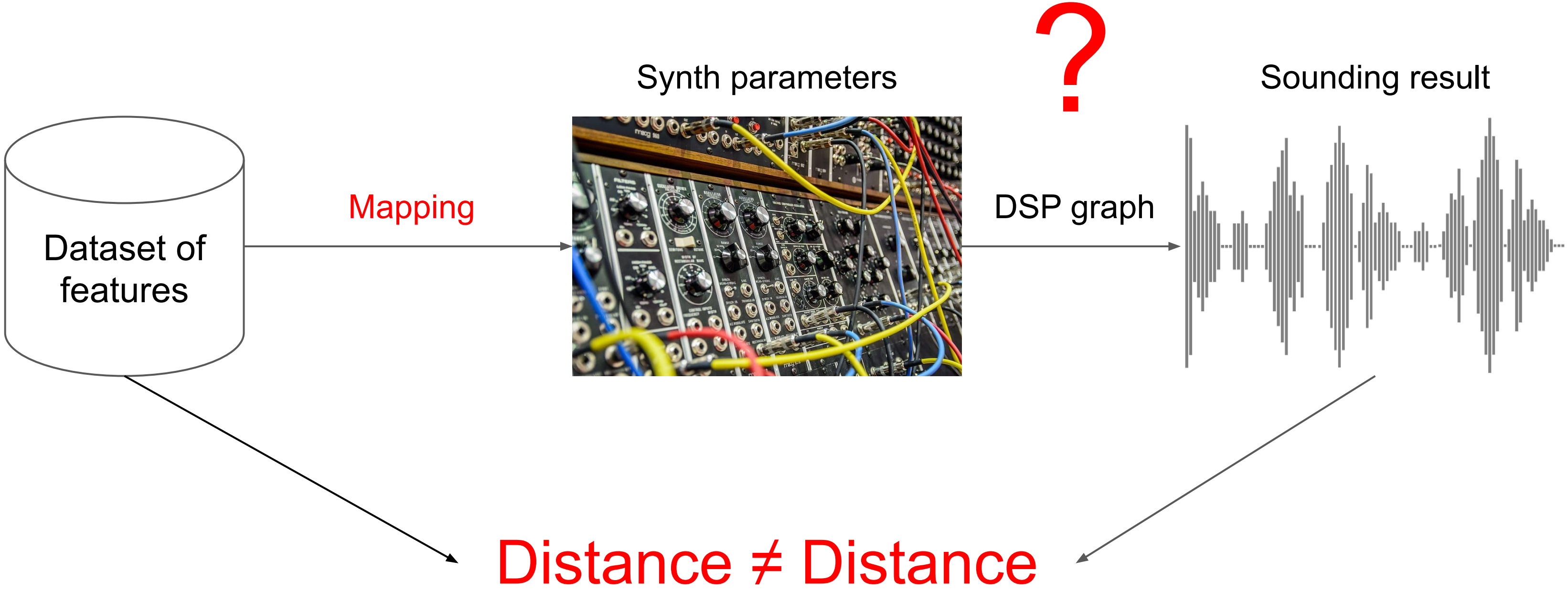
Image Sonification as Unsupervised Cross-Modal Domain Transfer (W.i.P)

Bálint Laczkó
Main Supervisor: Alexander R. Jensenius

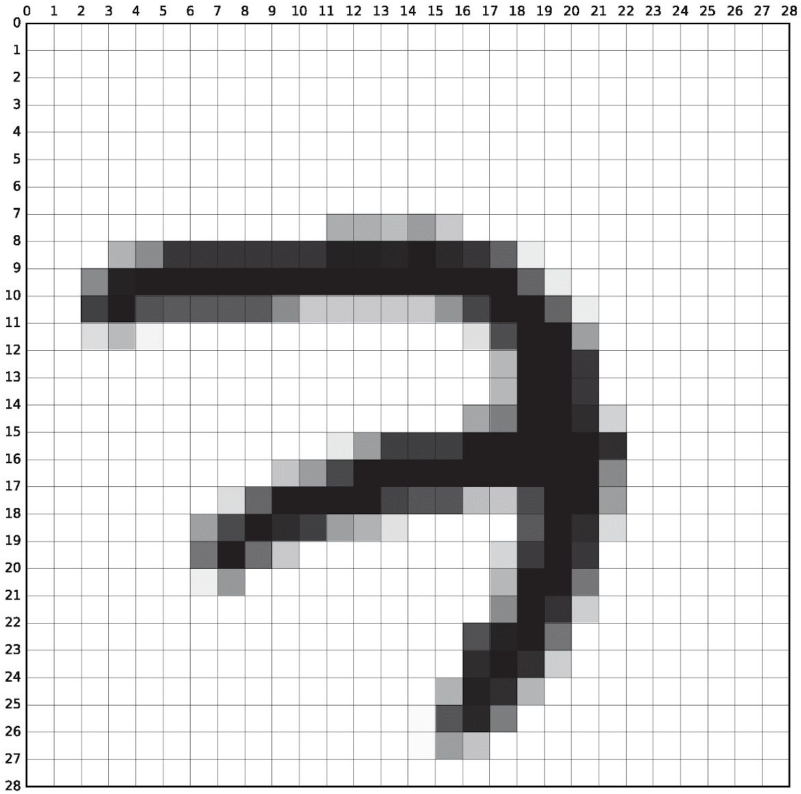
Parameter Sonification



The Limits of Parameter Sonification



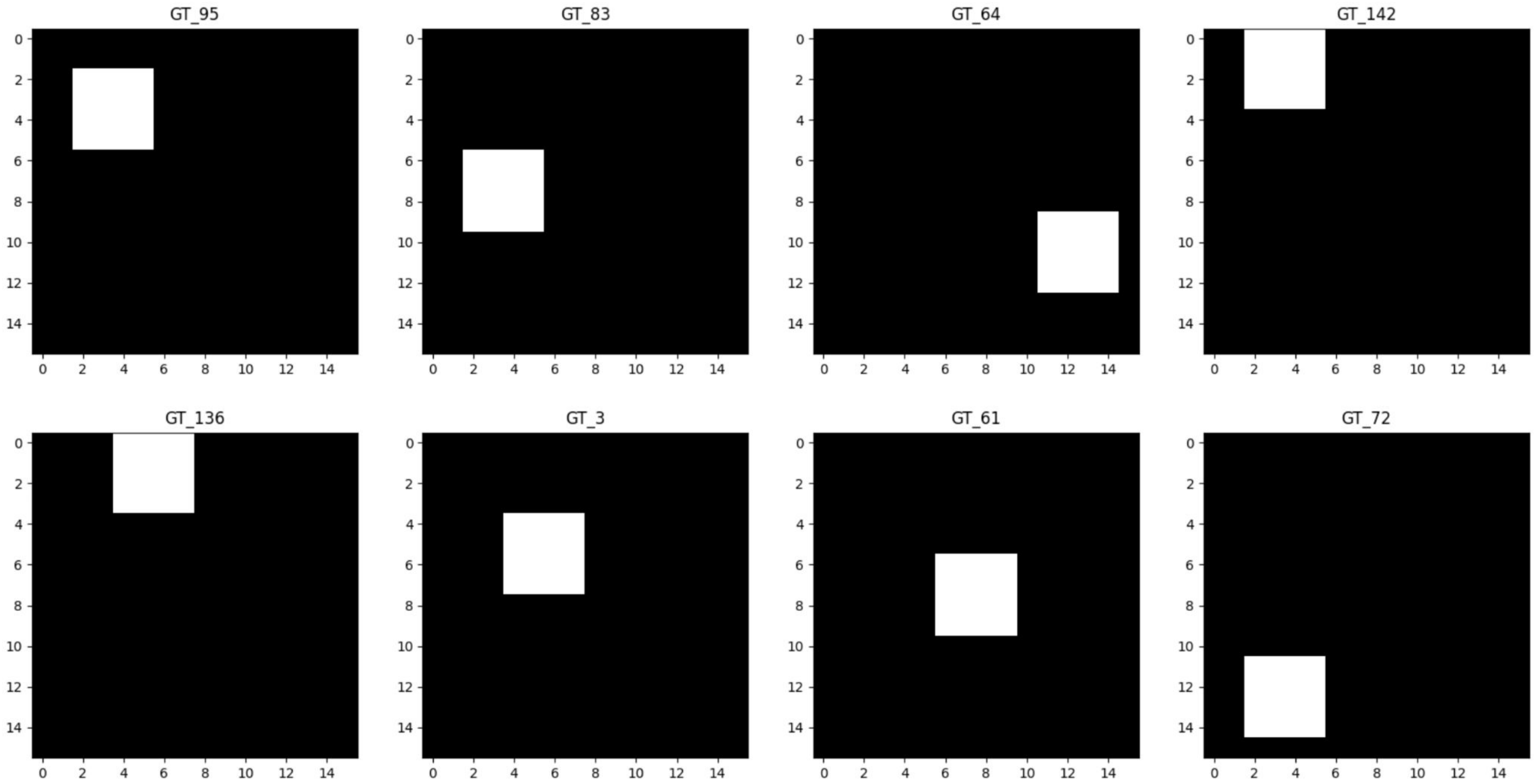
The latent meaning

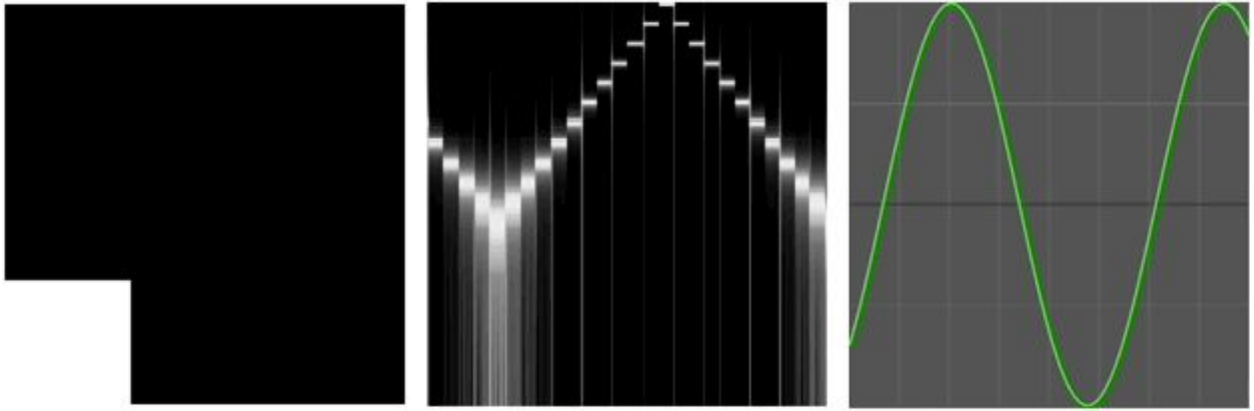


(**a**) MNIST sample belonging to the digit '7'.

(**b**) 100 samples from the MNIST training set.

Image from: [Baldominos, A., Saez, Y., & Isasi, P. (2019). A survey of handwritten character recognition with mnist and emnist. Applied Sciences, 9(15), 3169.]


GT_95
GT_83
GT_64
GT_142
GT_136
GT_3
GT_61
GT_72



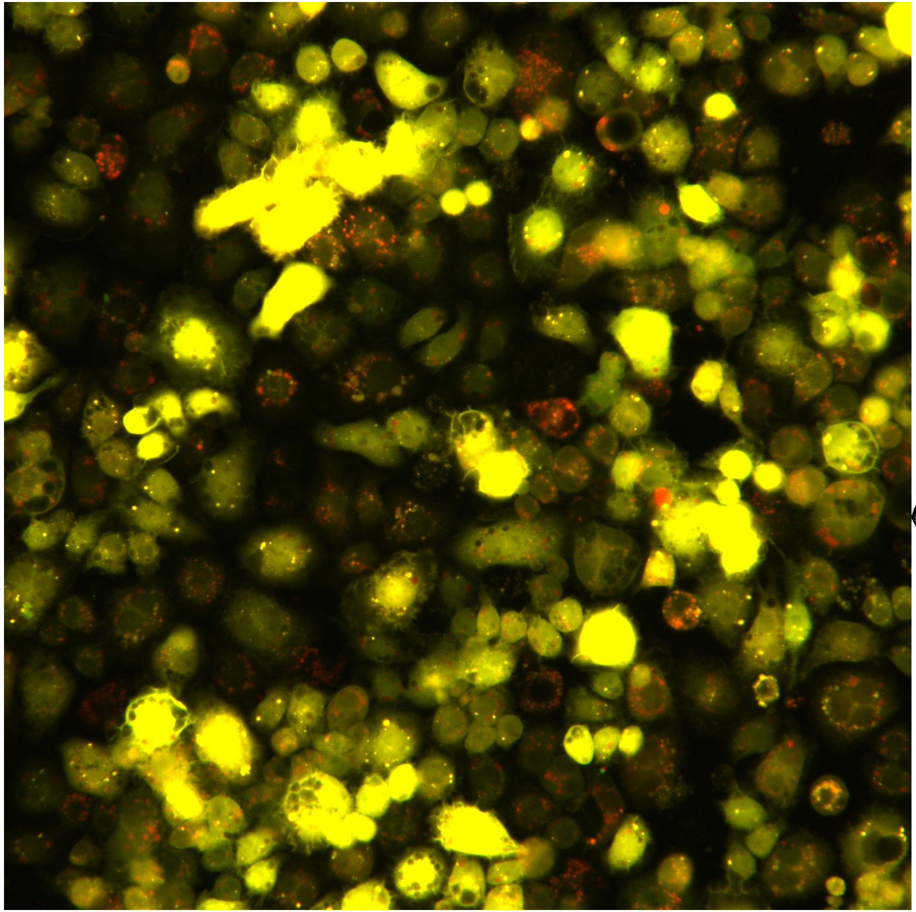
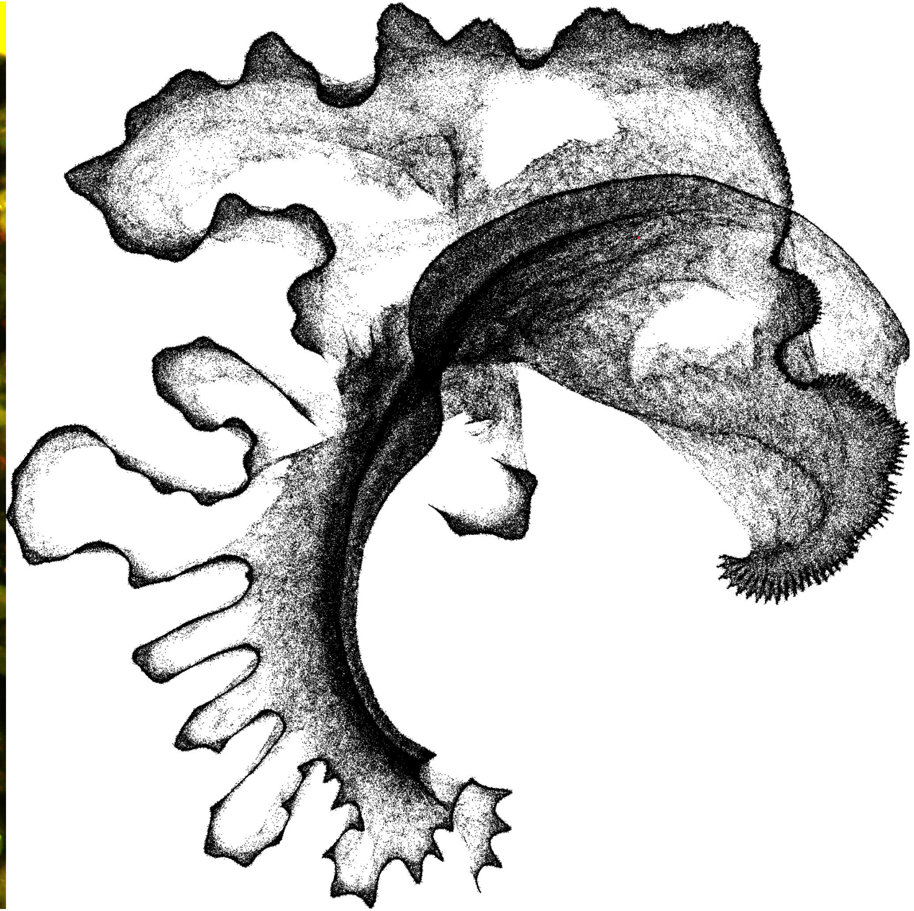


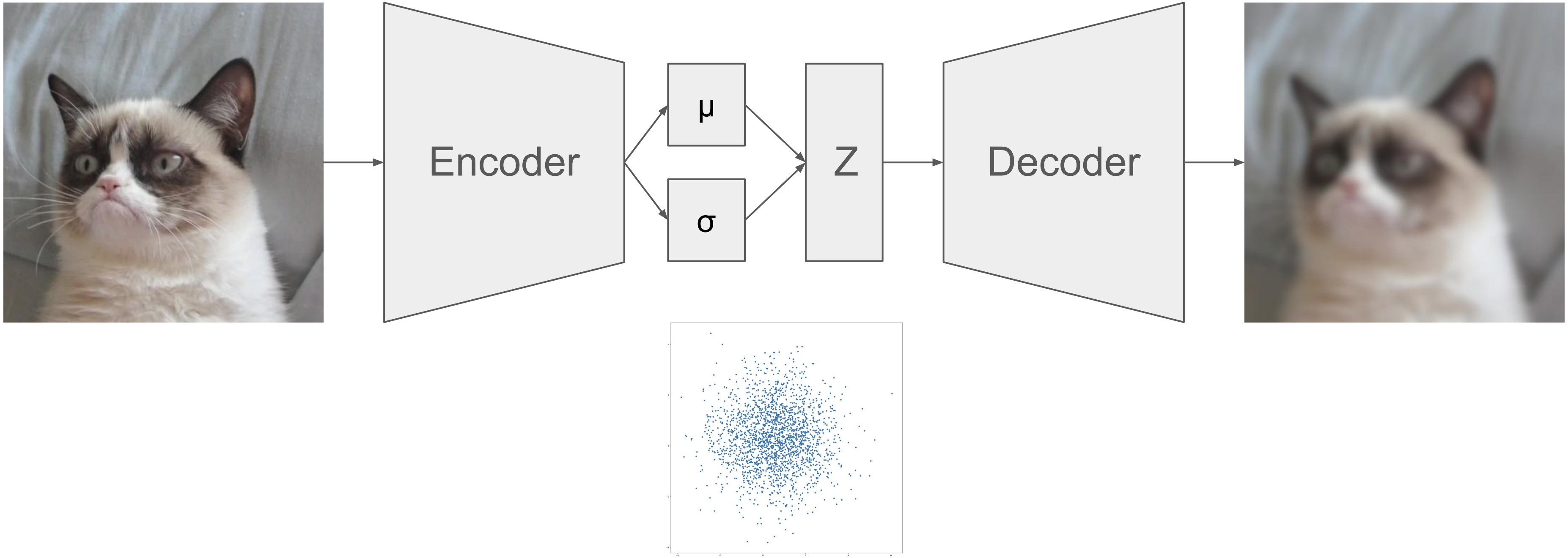
Image from: Al Outa, A., Hicks, S., Thambawita, V., Andresen, S., Enserink, J. M., Halvorsen, P., ... & Knævelsrud, H. (2023). Cellular, a cell autophagy imaging dataset. Scientific data, 10(1), 806.



1M FM synth sounds represented by 200 Mel-bands, clustered by UMAP

The potential of representation learning and unsupervised domain transfer for image sonification

Variational Auto-Encoders



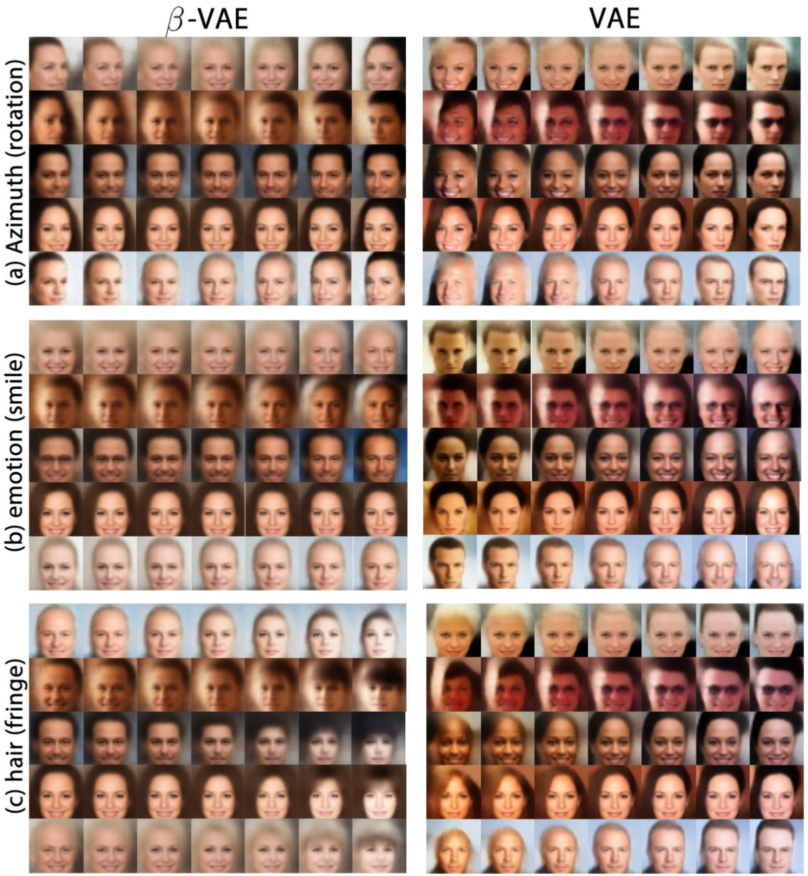
In search of disentangled representations

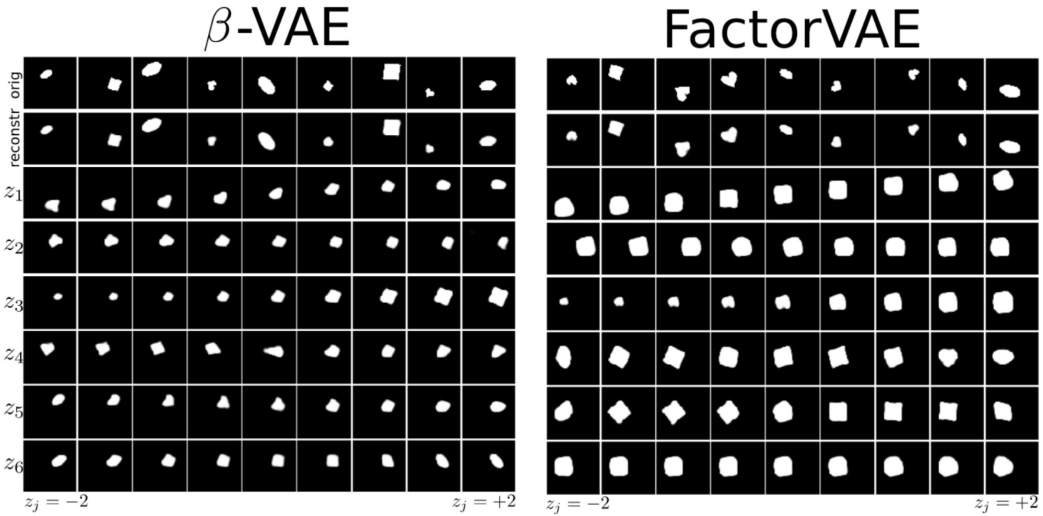
β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, Alexander Lerchner
Google DeepMind
{irinah,lmatthey,arkap,cpburgess,glorotx,
botvinick,shakir,lerchner}@google.com

Abstract

Learning an interpretable factorised representation of the independent data generative factors of the world without supervision is an important precursor for the development of artificial intelligence that is able to learn and reason in the same way that humans do. We introduce β-VAE, a new state-of-the-art framework for automated discovery of interpretable factorised latent representations from raw image data in a completely unsupervised manner. Our approach is a modification of the variational autoencoder (VAE) framework. We introduce an adjustable hyperparameter β that balances latent channel capacity and independence constraints with reconstruction accuracy. We demonstrate that β-VAE with appropriately tuned $\beta > 1$ qualitatively outperforms VAE ($\beta = 1$), as well as state of the art unsupervised (InfoGAN) and semi-supervised (DC-IGN) approaches to disentangled factor learning on a variety of datasets (*celebA*, *faces* and *chairs*). Furthermore, we devise a protocol to quantitatively compare the degree of disentanglement learnt by different models, and show that our approach also significantly outperforms all baselines quantitatively. Unlike InfoGAN, β-VAE is stable to train, makes few assumptions about the data and relies on tuning a single hyperparameter β, which can be directly optimised through a hyperparameter search using weakly labelled data or through heuristic visual inspection for purely unsupervised data.





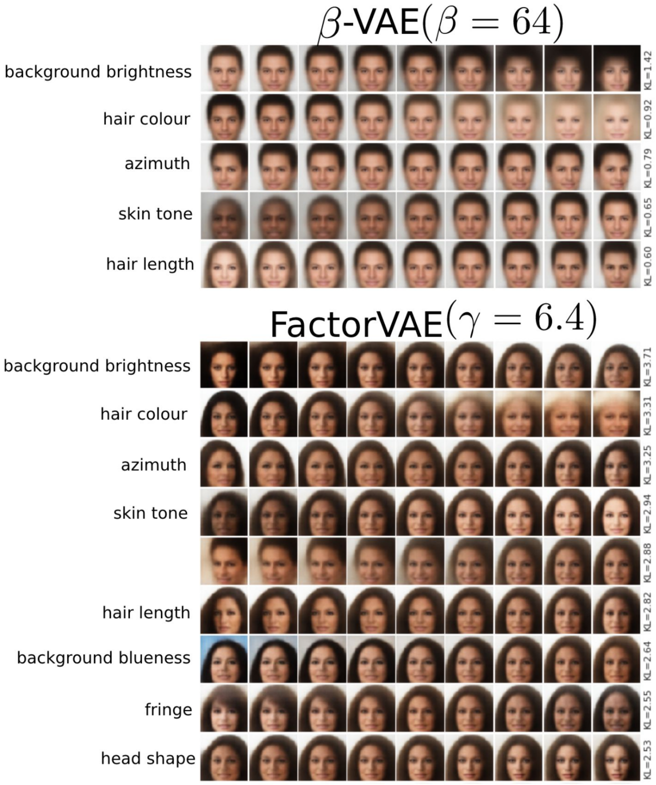


Figure 6. First row: originals. Second row: reconstructions. Remaining rows: reconstructions of latent traversals across each latent dimension sorted by $KL(q(z_j|x)||p(z_j))$, for the best scoring models on our disentanglement metric. Left: β-VAE, score: 0.814, $\beta = 4$. Right: FactorVAE, score: 0.889, $\gamma = 35$.

Fader Networks:
Manipulating Images by Sliding Attributes

Figure 1: Interpolation between different attributes (Zoom in for better resolution). Each line shows reconstructions of the same face with different attribute values, where each attribute is controlled as a continuous variable. It is then possible to make an old person look older or younger, a man look more manly or to imagine his female version. Left images are the originals.

Unsupervised cross-modal domain transfer

Cross-modal Variational Alignment of Latent Spaces

Thomas Theodoridis Theocharis Chatzis Vassilios Solachidis Kosmas Dimitropoulos
Petros Daras
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece
{tomastheod, hatzis, vsol, dimitrop, daras}@iti.gr

Abstract

In this paper, we propose a novel cross-modal variational alignment method in order to process and relate information across different modalities. The proposed approach consists of two variational autoencoder (VAE) networks which generate and model the latent space of each modality. The first network is a multi-modal variational autoencoder that maps directly one modality to the other, while the second one is a single-modal variational autoencoder. In order to associate the two spaces, we apply variational alignment, which acts as a translation mechanism that projects the latent space of the first VAE onto the one of the single-modal VAE through an intermediate distribution. Experimental results on four well-known datasets, covering two different application domains (food image analysis and 3D hand pose estimation), show the generality of the proposed method and its superiority against a number of state-of-the-art approaches.

the cross-modal objective, they are categorized as discriminative and generative. Approaches that fall into the first category model the probability of an outcome conditioned on the given observation. Generative approaches, on the other hand, model the underlying distribution of the observed variables, thus obtaining valuable information regarding their origin.

Most recent approaches have adopted deep generative models, such as VAEs, GANs or a combination of them, to encode cross-modal data into a shared latent space [30, 34]. However, the main problem in these approaches is the fact that each modality has completely different characteristics from the others and, as a result, it is difficult to efficiently model the heterogeneous modalities (like image, speech or text) into a shared latent space. To address the problem of learning meaningful mappings among embedding spaces, we propose a novel variational alignment framework of latent spaces, which performs the mapping of the latent space of one modality onto the one of another modality. More

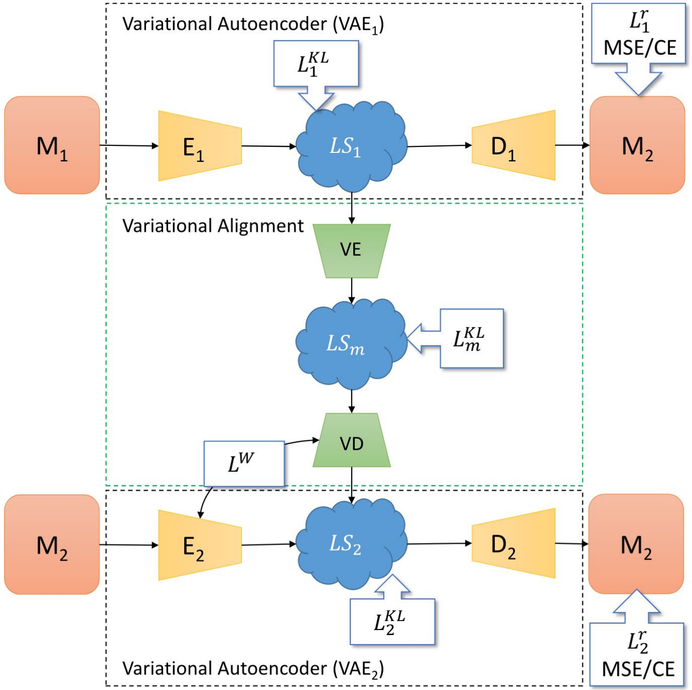


Figure 1. The proposed variational alignment architecture. The upper branch transitions from modality M_1 to M_2 using encoder E_1 and decoder D_1. The lower branch autoencodes M_2 through encoder E_2 and decoder D_2. The middle branch aligns the distribution produced by E_1 to the one produced by E_2 using the variational encoder (*VE*) and decoder (*VD*), which map to and sample from an intermediate distribution.

Latent Translation: Crossing Modalities by Bridging Generative Models

Yingtao Tian[1] **Jesse Engel**[2]

Abstract

End-to-end optimization has achieved state-of-the-art performance on many specific problems, but there is no straight-forward way to combine pretrained models for new problems. Here, we explore improving modularity by learning a post-hoc interface between two existing models to solve a new task. Specifically, we take inspiration from neural machine translation, and cast the challenging problem of cross-modal domain transfer as unsupervised translation between the latent spaces of pretrained deep generative models. By abstracting away the data representation, we demonstrate that it is possible to transfer across different modalities (e.g., image-to-audio) and even different types of generative models (e.g., VAE-to-GAN). We compare to state-of-the-art techniques and find that a straight-forward variational autoencoder is able to best bridge the two generative models through

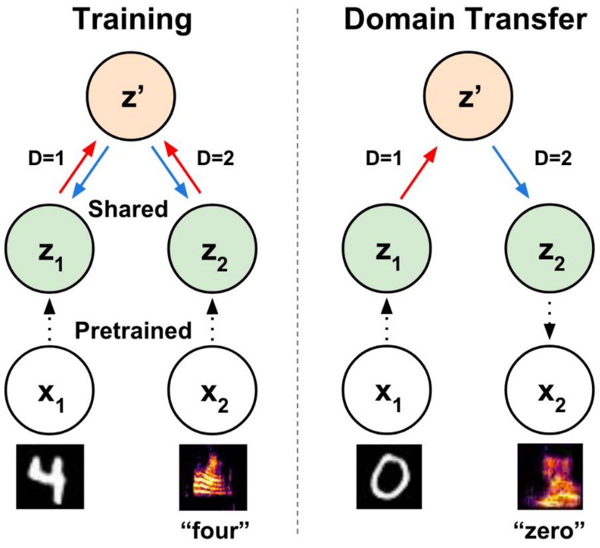
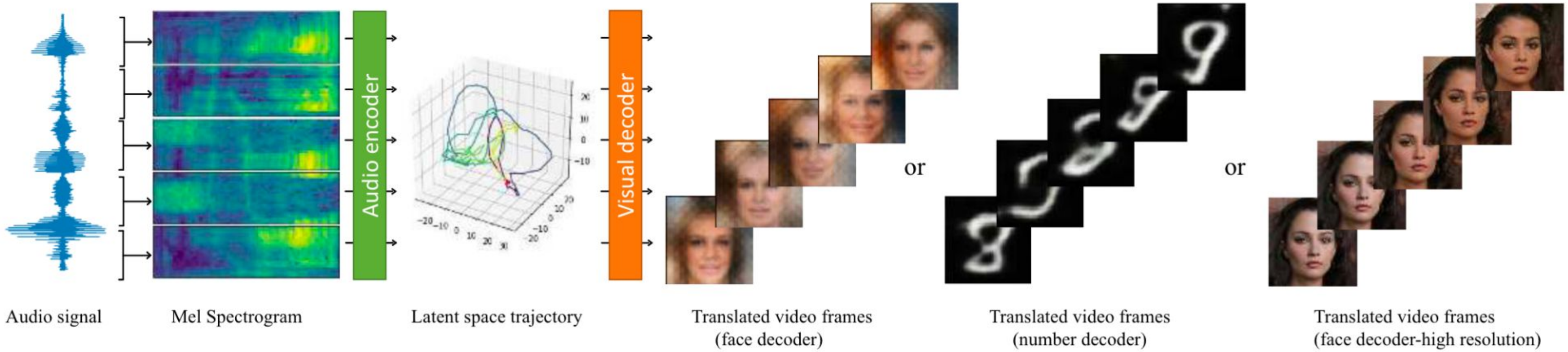


Figure 1. Latent translation with a shared autoencoder. Pretrained generative models provide embeddings (z_1, z_2) for data in two different domains (x_1, x_2), here shown as written digits and (spec-



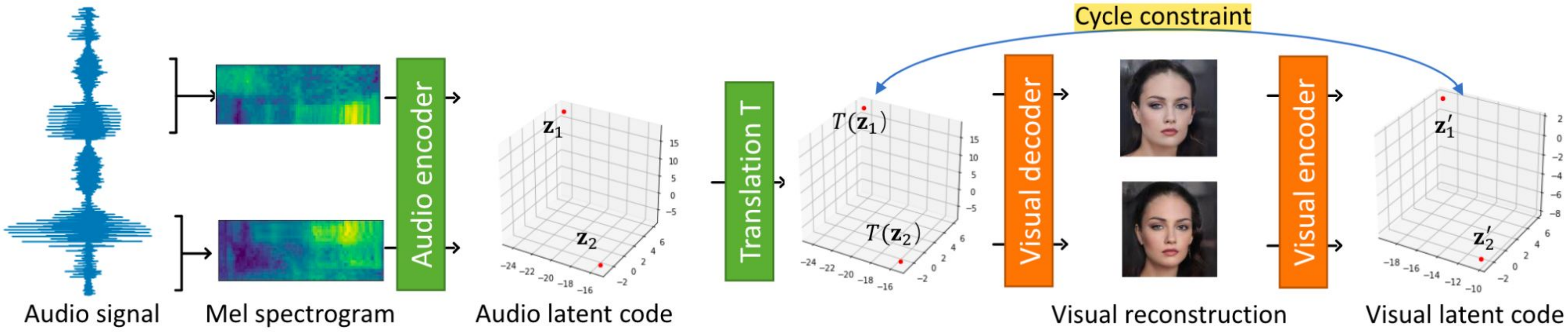
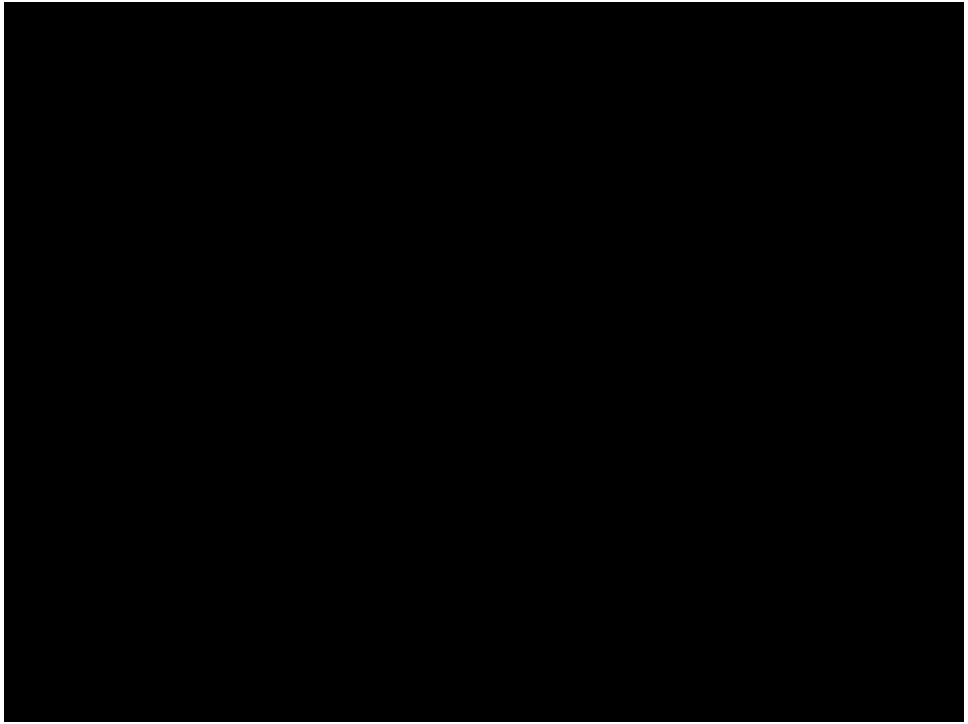


Figure 3. **Cycle constraint.** We apply a cycle constraint to ensures that the signal is preserved through video decoding and encoding.

AudioViewer demo

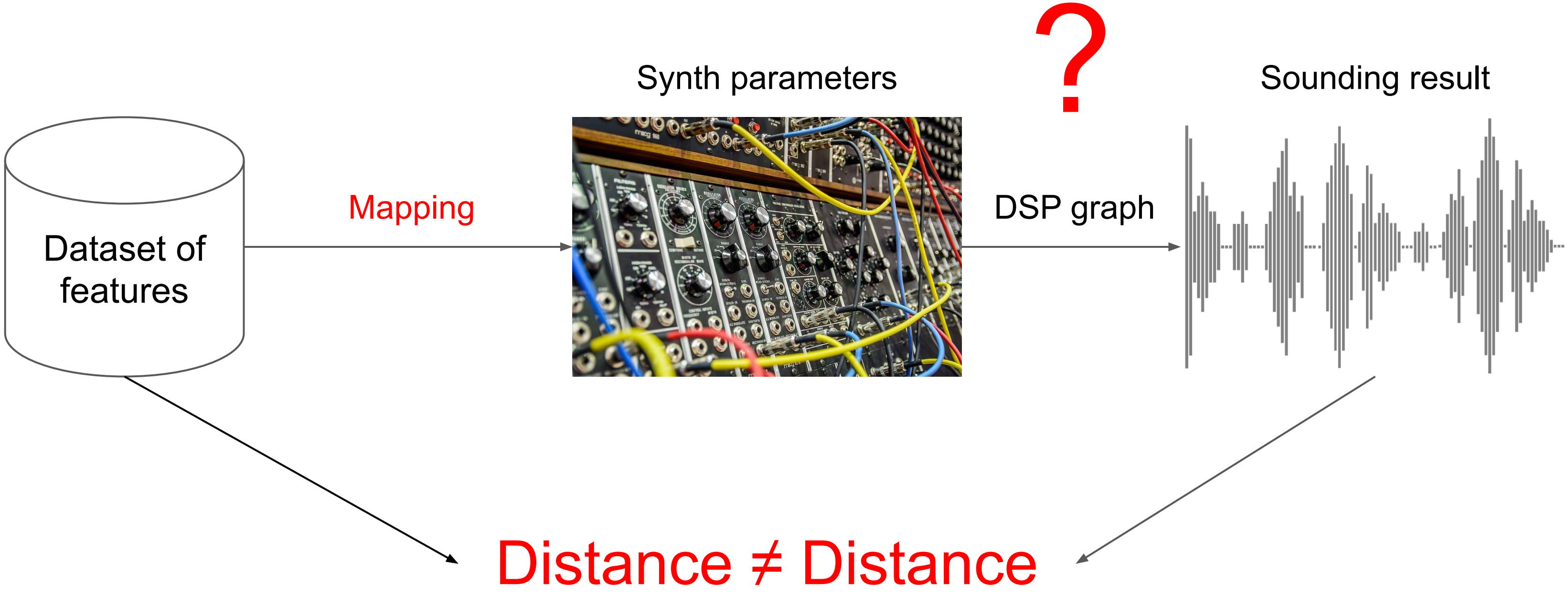


Potential benefits

- Learn “best” features
- (challenge existing bias in image analysis)
- Able to capture abstract features like “age” or “gender”
- Fit the mapping to the data
- Generalize better across similar datasets
- Builds upon pre-trained models (can swap models & retrain mapping)

Image Sonification as Unsupervised Cross-Modal Domain Transfer (W.i.P)

“Problem”



Dataset of features
Mapping
Synth parameters
?
DSP graph
Sounding result
Distance ≠ Distance

“Solution”

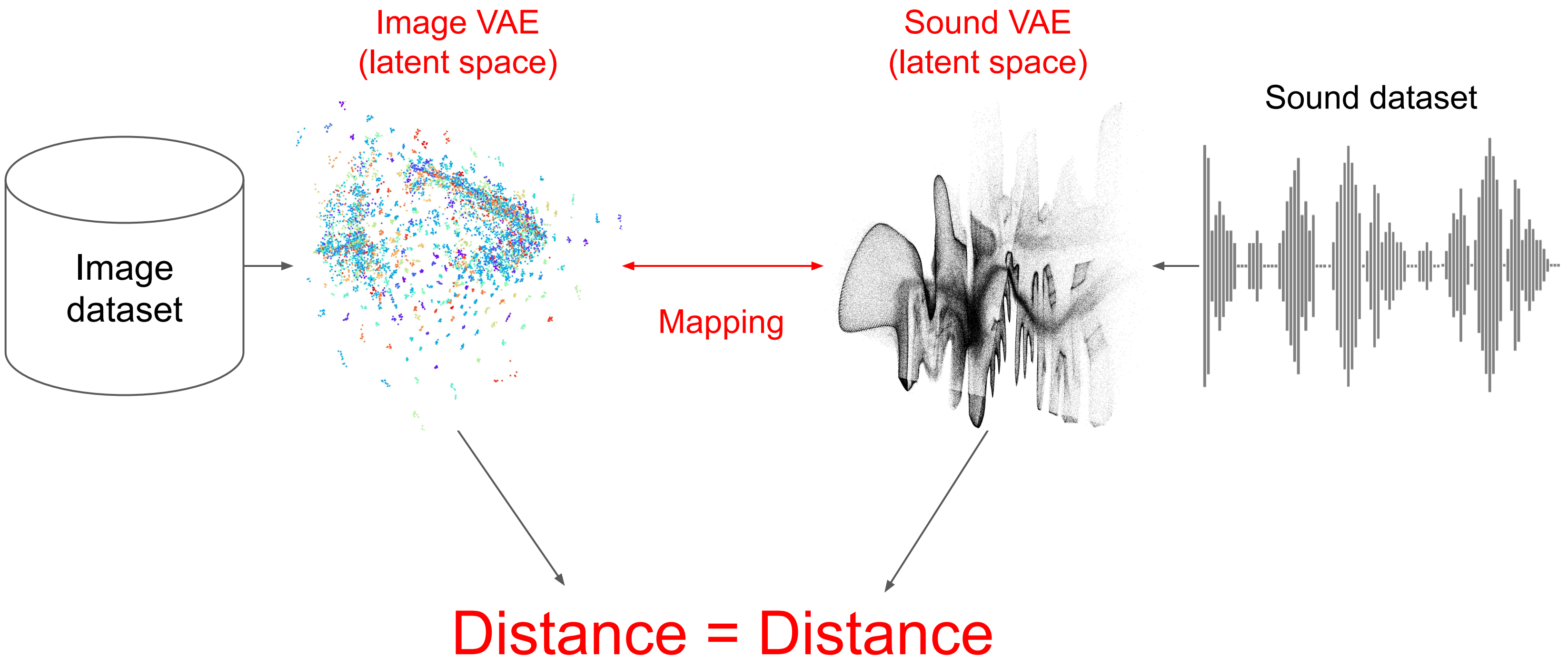
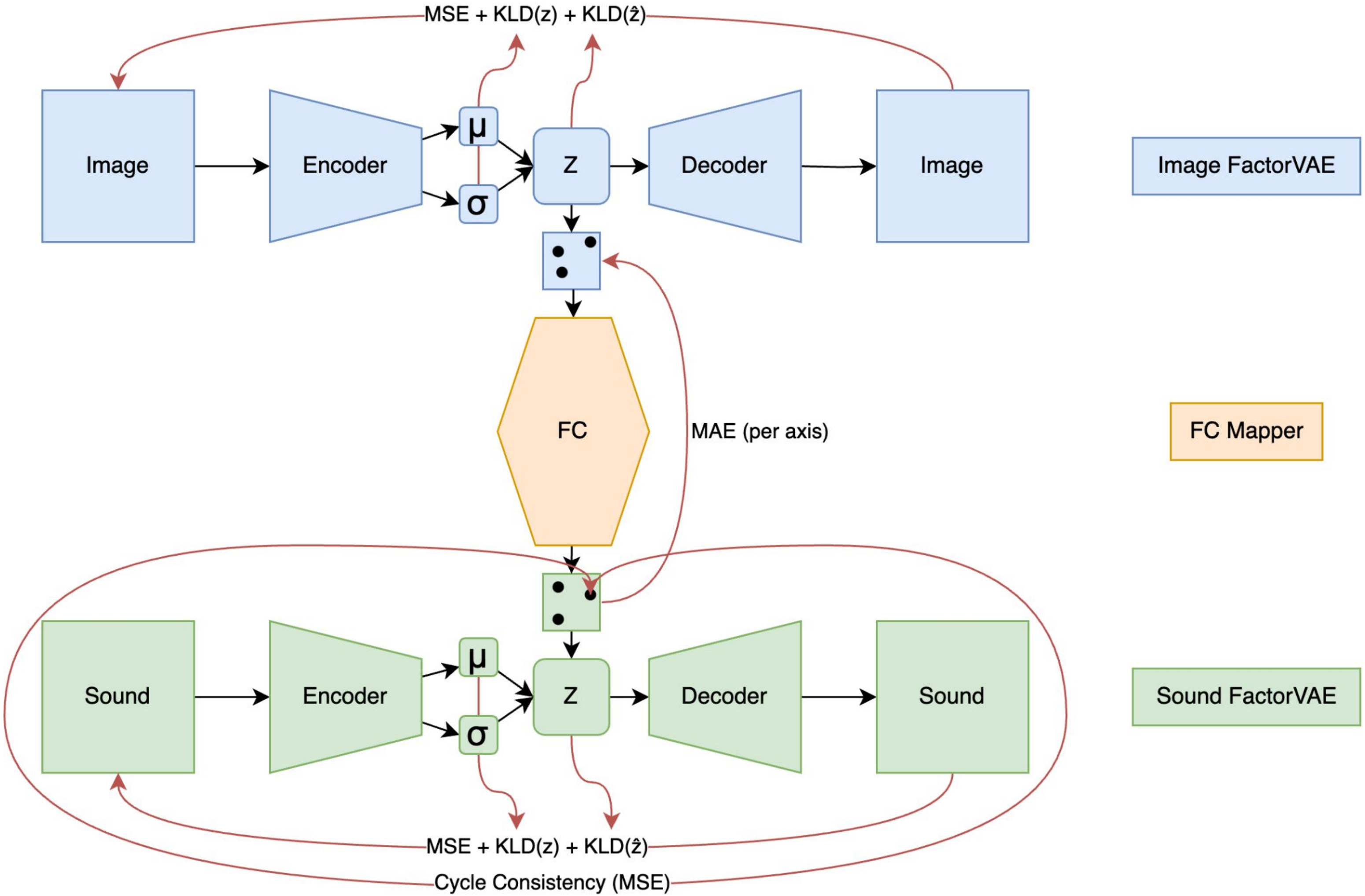


Image VAE
(latent space)
Sound VAE
(latent space)
Sound dataset
Image
dataset
Mapping
Distance = Distance


MSE + KLD(z) + KLD(ẑ)
Image
Encoder
μ
σ
Z
Decoder
Image
Image FactorVAE
FC
MAE (per axis)
FC Mapper
Sound
Encoder
μ
σ
Z
Decoder
Sound
Sound FactorVAE
MSE + KLD(z) + KLD(ẑ)
Cycle Consistency (MSE)

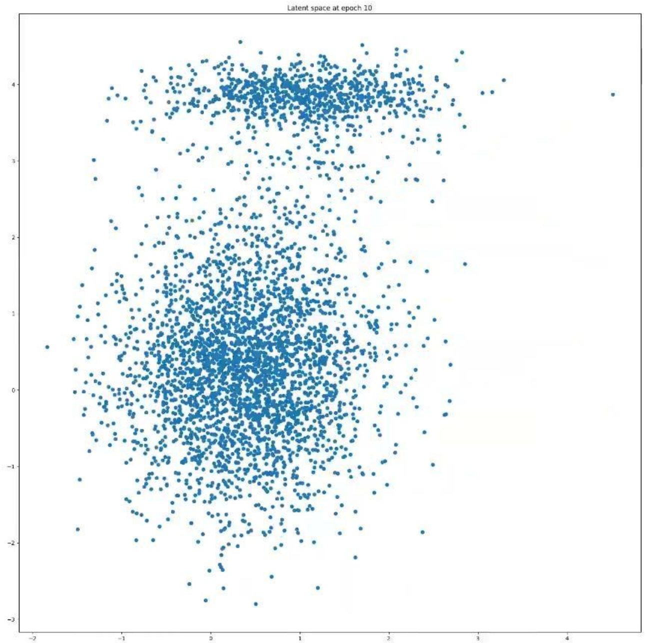
Experiment 1: synthetic datasets, map 2 factors

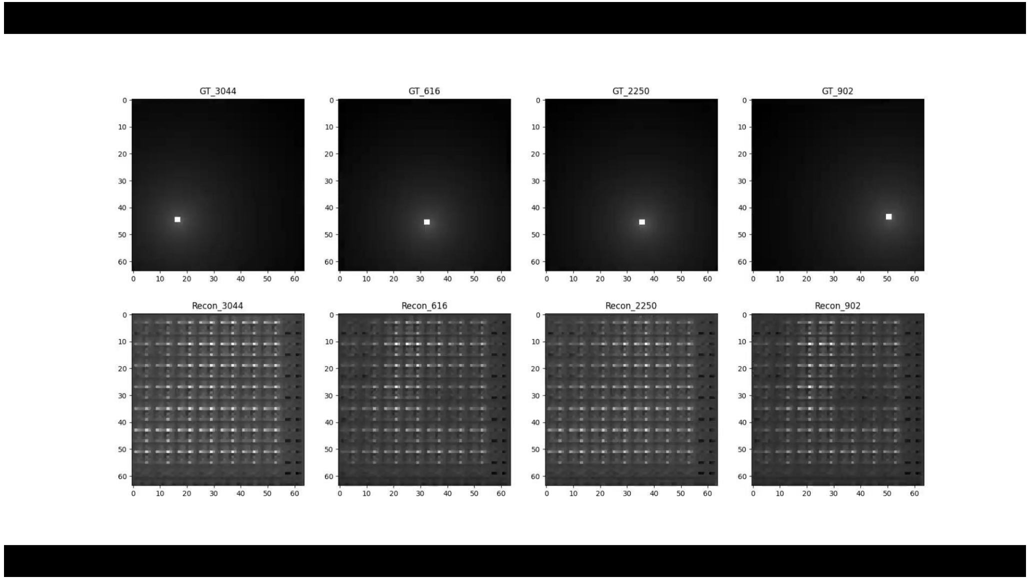
- Create a scenario that's easy to verify
- Create image & sound datasets with two independent varying factors
- Test if the system can:
  - Recognise the factors in both datasets and create disentangled representations of them
  - Find the best fitting mapping between latent spaces
- The system is only told that there are 2 factors

Image Dataset: white squares over black bg

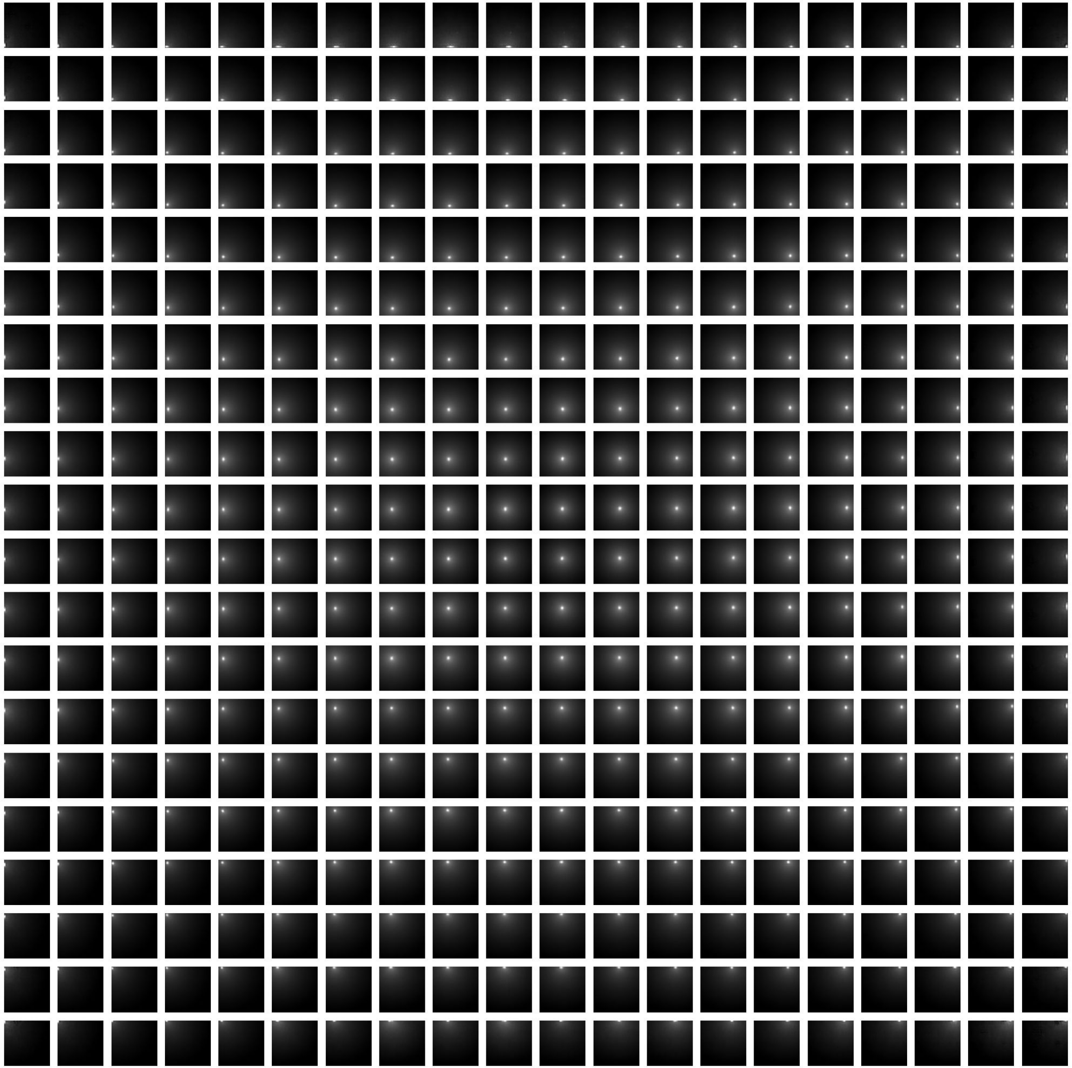
- Only varying factors: x & y coordinates
- Use a FactorVAE with a 2D latent space
- (use falloff “light” to combat sparse image)

Training…


Latent space at epoch 10


GT_3044
GT_616
GT_2250
GT_902
Recon_3044
Recon_616
Recon_2250
Recon_902

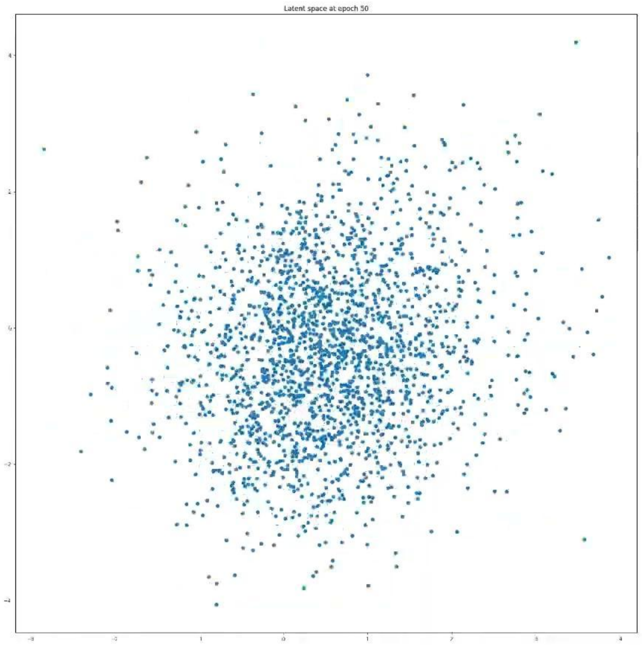
Traverse latent space

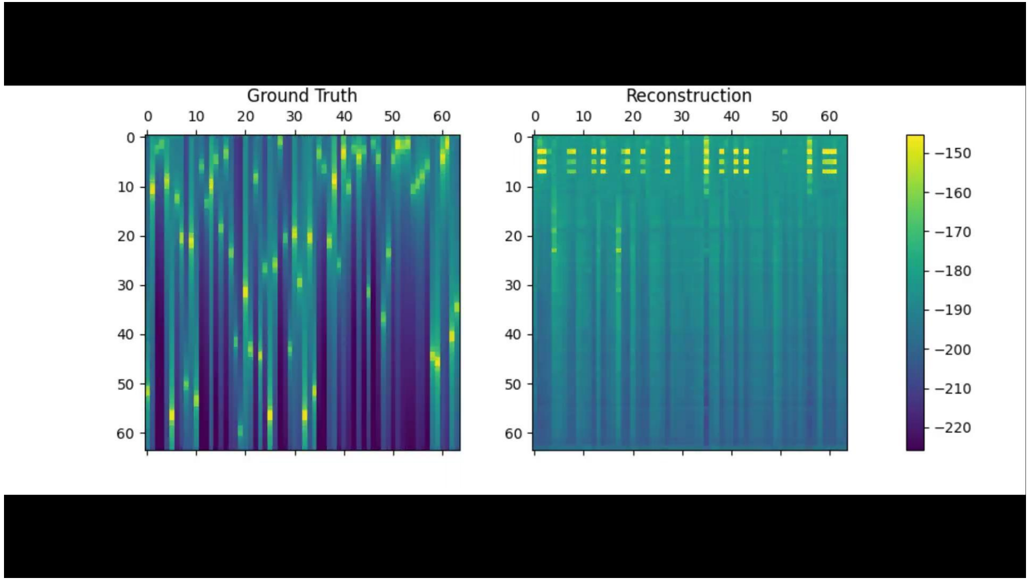


Sound dataset: Sine waves

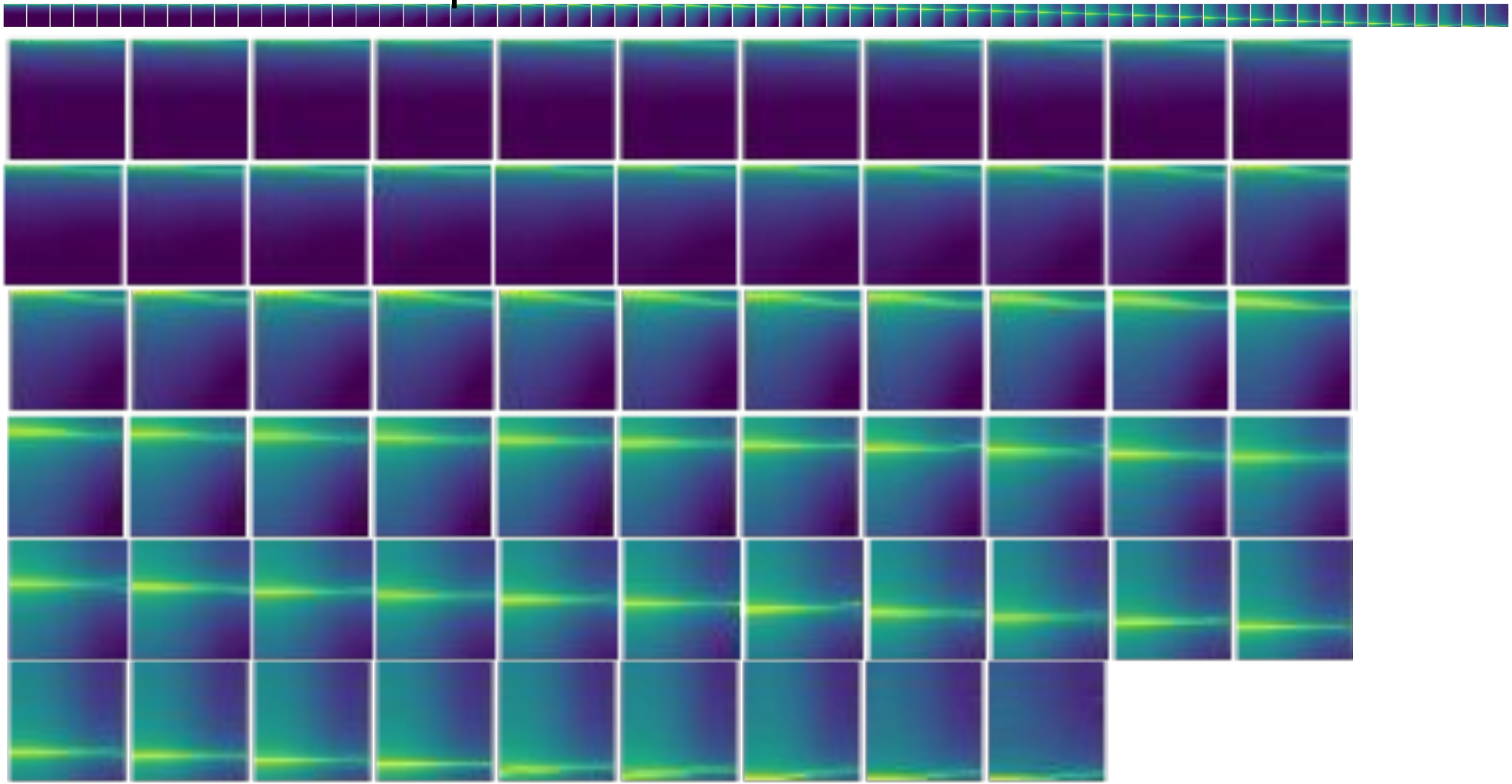
- Only varying factors: pitch & loudness
- Input representation: 64x1 Mel bands averaged over time, dB scaled
- Use a FactorVAE with a 2D latent space

Training…


Latent space at epoch 50

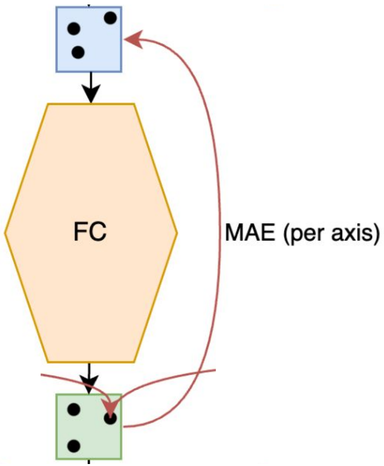

Ground Truth
Reconstruction
0 10 20 30 40 50 60
−150
−160
−170
−180
−190
−200
−210
−220

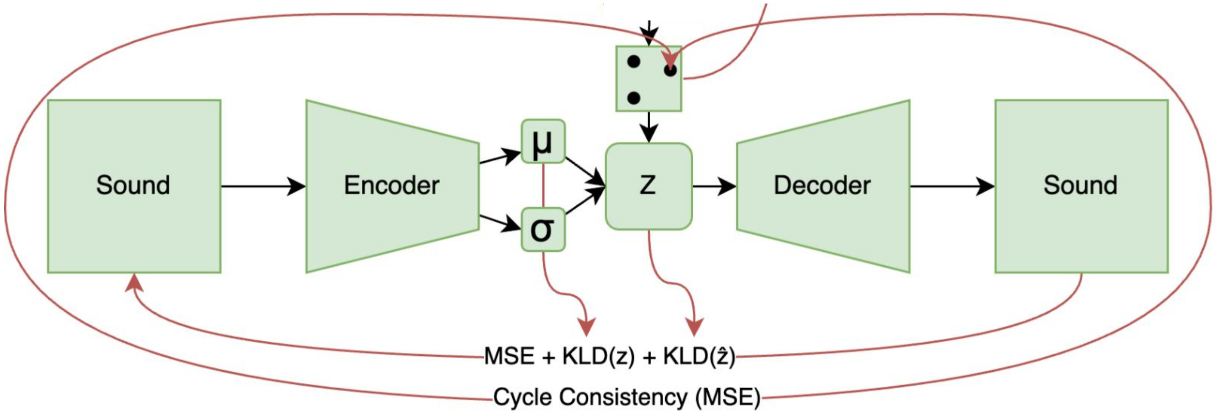
Traverse latent space



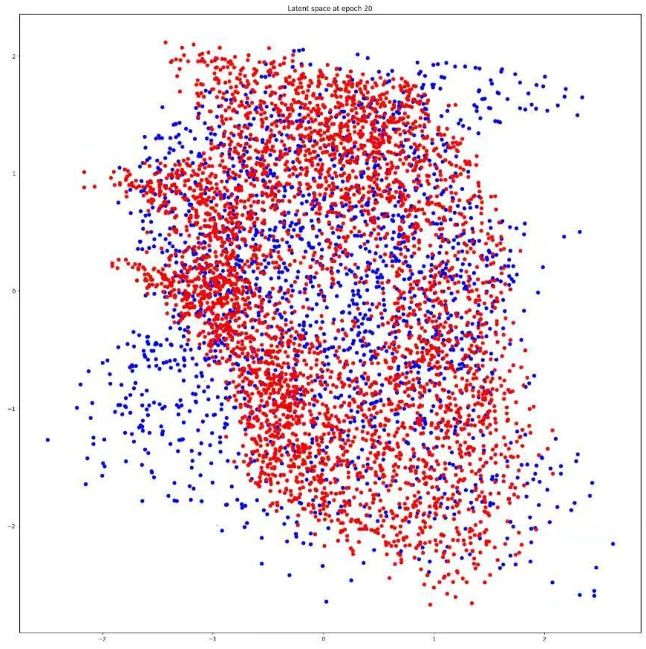
Training the Mapper

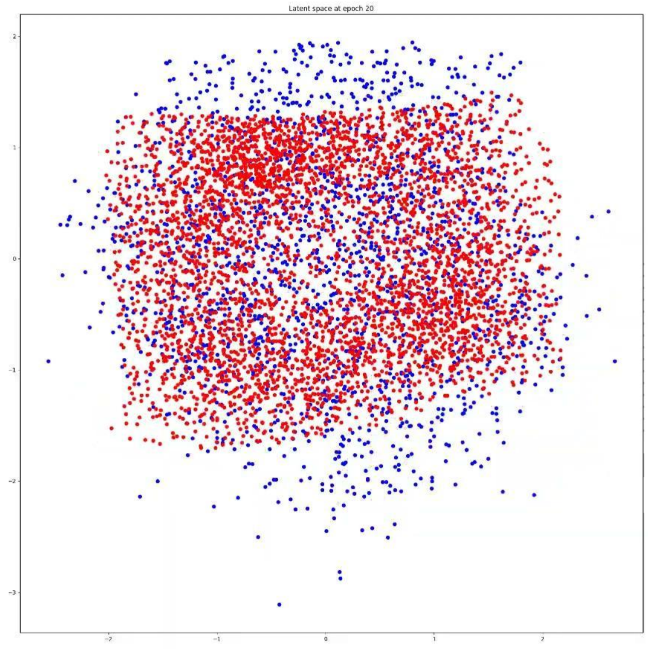
- 1st stage: only use locality loss
  - Since using FactorVAE-s → per axis!
- 2nd stage: ramp up cycle consistency (keep locality)





Training…





Live demo… :)

Discussion

- Reconstructions need to have OK quality in target model
- Assumption 1: latent dimensionality needs to match
- Assumption 2: the extents of latent spaces match
- AudioViewer design vs mine: factorVAE-s → need to preserve the meaning of axes
- Problem with representation: quiet sine waves produce numerically smaller errors?
- Synthetic datasets don't necessarily have gaussian priors
- Mapper training in 2 stages
- Cycle consistency is king

Gollum kitty (thank you :)

Image from: [If 30 Famous Characters Were Kittens, Made By AI Dreams | Bored Panda]