Disentangling Embedding Vectors for Controllable Facial Video Generation

Anonymous Author(s)

ABSTRACT

The task of editing video aims to control the content whilst generating realistic and coherent videos. The embedding vectors of an encoding-decoding architecture can be manipulated to create novel videos with certain characteristics, but they are typically entangled, making editing difficult and generalization weak. In this paper, we propose a novel vision transformer architecture and contrastive training regime for facial video generation and editing. Our model is able to disentangle embedding vectors, which yields embeddings with semantic interpretations. This allows for manipulation of videos in a direct and intuitive manner. We show that our model is effective for facial video editing. This has many potential applications in the animation, gaming, and video editing industries.

KEYWORDS

Disentangled vectors, video editing, video generation

ACM Reference Format:

Anonymous Author(s). 2023. Disentangling Embedding Vectors for Controllable Facial Video Generation. In *Proceedings of The 15th Annual ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23).* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Video content editing is a challenging task, with many applications in the animation and gaming industries. One approach is to encode each frame, manipulate the encoding and then decode the frame to produce novel videos. However, the embedding vectors (or latent codes) are typically entangled, meaning that the axes of the vectors do not have clear semantic interpretations. This entanglement makes editing difficult and generalization to new scenarios weaker.

This paper proposes a novel contrastive training approach to disentangle embedding vectors, and demonstrates the effectiveness of this approach when applied to facial video editing.

2 METHODOLOGY

2.1 Model Architecture

In recent years, vision transformers (ViTs), [Dosovitskiy et al. 2020] have achieved state-of-the-art results in many image processing and vision tasks, such as image classification, object detection, and image segmentation. ViTs are particularly well-suited for video generation, as they can learn long-range dependencies and spatial-temporal contexts in videos. In this paper, we use a vision transformer architecture for video generation. Our architecture is based on the following key components:

- A spatial transformer encoder to learn spatial features from each frame in the video.
- A spatial transformer decoder to generate the output video frame by frame.
- A temporal transformer to learn temporal dynamics.
- A latent code editor to edit the embedding representation of the video before decoding.

Our spatial transformer encoder is a standard ViT encoder, which divides each frame into $32 \times 32 \times 3$ patches mapped onto 128-dimensional vectors and then uses self-attention to learn spatial features from these patches. Specifically, we have a 10-layer 8-headed transformer to perform encoding, and each of the patches uses additive position encoding. We adopt layer normalisation and use a 0.1 dropout during training to improve generalisation and help training.

Our decoder has the same ViT architecture as the spatial encoder with a 2-layer feed-forward network to cast the latent vector into image space. Note that as we want each dimension of our embedding vector to carry independent semantic information, we do not use dropout on the first layer of this decoder, as it encourages redundancy in the representation.

The temporal transformer works on the embedding vectors produced by the spatial encoder, generating vectors in the same space. As the present work focuses on the disentangling and is orthogonal to the temporal prediction, we will not detail the temporal transformer further.

The latent code editor simply maps each user action onto one specific dimension of the embedding vector before being decoded, and again is not the focus of this current work.

2.2 Training

The combined architecture is trained to minimize the reconstruction error between the generated video and the ground-truth video. Specifically, we use a SSIM similarity metric [Wang et al. 2004] to measure the reconstruction error, modified to emphasize the luma:

$$\mathrm{SSIM}(\mathrm{image}_1^{rgb}, \mathrm{image}_2^{rgb}) + \mathrm{SSIM}(\mathrm{image}_1^{gray}, \mathrm{image}_2^{gray})$$

A proprietary database of around 400 hours of video footage is used, with both natural conversational recordings as well as recordings of isolated facial movements. Heads are tracked and faces are extracted and normalized.

During training, contrastive learning [Chen et al. 2020] is applied. The recordings of isolated facial movements are used to disentangle the embeddings by constraining the model to associate only a single dimension of the embedding vector with each facial movement. Specifically, pairs of images from the same motion are presented

MIG '23, November 15—17, 2023, Rennes, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

to the encoder. The embedding vectors e_1 and e_2 are averaged in all but a single dimension i to form $\hat{e}_1$ and $\hat{e}_2$.[1] The $\hat{e}_1$ is presented to the decoder and the reconstruction error to the first image is minimised and similarly with $\hat{e}_2$ and the second image. The mean-squared error between e_1 and e_2 in all dimensions except the ith dimension is added to the loss function.[2] In this way, the decoder is encouraged to associate only the ith dimension with the associated motion, and the encoder is encouraged to create a representation where only the ith dimension varies.

In some cases the motion cannot be adequately described using a single attribute. In this case, we repeat the contrastive learning outlined above, but allowing variability on a small number of attributes. Principal components analysis is performed and this process is repeated along each of the principal components.

3 RESULTS AND DISCUSSION

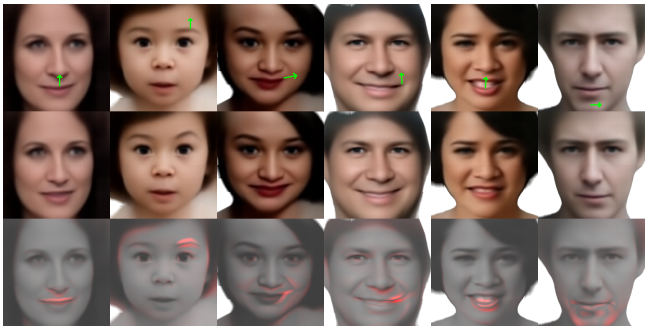

Figure 1: Editing tasks showing isolated yet natural manipulations of facial images.

Figures 1 and 2 show the results of editing tasks using the proposed model. Images in the top row are edited by changing a single dimension to produce the images in the middle row. Green arrows show the direction of operation of the change due in the corresponding dimension. The bottom row shows the top row images in black and white, with the pixel-wise differences between the two images superimposed in red.

Only regions of the image that are directly associated with the corresponding edit are effected: the rest of the face is almost completely unchanged. Nonetheless, the edits are still natural and connected, for example, in Figure 1 note (1) the deepening of the nasolabial fold with the widening of a corner of the mouth, and in Figure 2, note (2) the strengthening of the glabellar lines with an eyebrow constriction and (3) the movement of the nasolablial folds and the moustache with a nostril flare.

Figure 2 depicts our 3 partial failure cases where our proposed model is unable to completely isolate the movement to a single control. Whilst movement is constrained effectively to the appropriate region, it is still bilaterally ambiguous, meaning that for these 3 features, we are not able to be control them individually but only in tandem on the left and right. Applying lateral masks to the regions should be able to overcome this, but this is left for future work.

[1] $\hat{e}_1 = \delta(j \neq i)\frac{1}{2}(e_1 + e_2) + \delta(j = i)e_1$. Similarly for $\hat{e}_2$.

[2] $E = \sum_{j \neq i}(e_1^{(j)} - e_2^{(j)})^2$

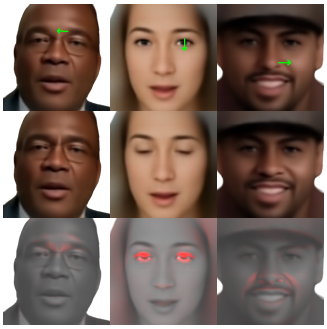

Figure 2: Brow constriction, eyelid closing and nostril flaring are not fully disambiguated: motion is only able to be controlled bilaterally.

3.1 Supplementary Video

Supplementary video material demonstrates that our proposed model is effective for both facial video generation and facial video editing. In particular, it shows that the proposed model:

- can generate coherent facial videos which capture enough of the motion to yield high fidelity movement,
- separates motion information from the subject's appearance
- allows for easily real-time manipulation of individual features in a direct and intuitive manner, and
- generalises feature manipulation over poses and expressions.

4 CONCLUSION

This paper has proposed a novel vision transformer architecture and training regime for facial video generation and editing. Our model is based on a contrastive training approach to disentangle embedding vectors. We have shown that our model yields embeddings with isolated semantic interpretations allowing ease of editing, yet maintaining connected natural motions.

5 FUTURE WORK

These preliminary results focus on low-resolution videos. Future work will focus on extending the model to generate high-resolution and high-quality videos, and bilateral disambiguation along the 3 ambiguous dimensions.

REFERENCES

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. *CoRR* abs/2002.05709 (2020). arXiv:2002.05709 https://arxiv.org/abs/2002.05709

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR* abs/2010.11929 (2020). arXiv:2010.11929 https://arxiv.org/abs/2010.11929

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861