Student Conference Proceedings
Vol 1, No 1, 2025, Article ID 1993, 4 Pages

Student Conference Lübeck

Proceedings Article

Disentanglement Learning of Facial Expression and Appearance

Rafael Hoock[a,c,*] · Nele Sophie Brügge[c] · Heinz Handels[b,c]

[a]Student of Medical Informatics, Universität zu Lübeck, Lübeck, Germany
[b]Institute of Medical Informatics, Universität zu Lübeck, Lübeck, Germany
[c]German Research Center for Artificial Intelligence, Lübeck, Germany
[*]Corresponding author, email: rafael.hoock@student.uni-luebeck.de; heinz.handels@uni-luebeck.de

Received 11 February 2025; Accepted 26 November 2025; Published online 01 December 2025

Abstract

Facial expression analysis holds great promise beyond conventional emotion recognition, particularly in medical diagnostics and emotional well-being. However, disentangling tightly intertwined factors, such as facial expression and appearance, remains a significant challenge. This study presents a framework utilizing StyleGAN inversion and SimCLR to focus on expression-specific attributes while systematically minimizing appearance-related factors. Despite the implementation of tailored augmentations like Style-Mixing and Latent Space Slider, the disentanglement of expression from appearance was not fully successful. Residual appearance information persisted in the learned representations, as shown by clustering dominated by appearance rather than expression in t-SNE visualizations. However, the accuracy of emotion classification reached 95.82 % with augmented CNNs, demonstrating the potential of this approach. These findings highlight the limitations of current disentanglement techniques and the need for further refinement to achieve robust separation of expression and appearance. Advancing this work could enhance applications in emotion recognition and privacy-preserving medical diagnostics.

I. Introduction

Facial expression analysis has immense potential beyond conventional emotion recognition, extending to critical medical applications. For example, understanding and disentangling expressions can aid in diagnosing neurological disorders, monitoring emotional well-being, and creating personalized treatments. In this study, we used baseline datasets, such as CK+, to develop reproducible methodologies. For facial analysis tasks, we use StyleGAN2 latent vectors that contain all the information necessary to achieve this facial expression analysis. This is because latent spaces encapsulate all the information necessary for accurate image synthesis. However, these spaces also include superfluous appearance-related features, which could lead to model learning subject-related characteristics instead of expression-related ones. This issue is particularly pronounced for small datasets, such as those in medical applications. We use disentanglement to eliminate such irrelevant features and focus solely on the expressive aspects.

Related Work Disentanglement of latent representations in generative models has gained significant interest due to its diverse applications in emotion recognition, medical diagnostics, and image synthesis. StyleGAN [1] has been instrumental in advancing disentanglement tasks by providing a semantically rich latent space for high-quality image generation. Techniques like StyleGAN inversion [2] map real-world images into this latent space, enabling detailed attribute editing. However, separating closely intertwined factors such as facial expression and appearance remains a challenge.

Supervised approaches have tackled disentangle-

ment in specific contexts. Peng *et al.* [3] proposed separating "identity" and "non-identity" features using samples generated by 3D morphable models, while Nitzan *et al.* [4] trained a deep CNN to disentangle identity from other facial attributes, such as pose and illumination. Unlike these methods, our study operates directly in the StyleGAN latent space and focuses specifically on disentangling appearance from expression.

Contrastive learning has proven effective for disentanglement by structuring latent spaces through positive and negative pairs. Frameworks such as SimCLR [5] leverage data augmentation to identify meaningful similarities. Duarte et al. [6] adapted this approach to the StyleGAN latent space by generating synthetic videos to define contrastive pairs, effectively disentangling features related to identity and expression. Expanding on these concepts, our study employs tailored latent space augmentations, including style mixing and semantic slider adjustments.

II. Methods and materials

Building upon the approaches outlined in the related work section, our methodology aims to disentangle facial expressions from appearance using a combination of StyleGAN inversion and contrastive learning. To achieve this, we leverage datasets that provide diverse facial expressions, employ specialized augmentations to refine latent space representations, and implement SimCLR for structured contrastive learning. The following subsections detail the datasets used, the feature extraction process through StyleGAN inversion, and the application of SimCLR to enhance disentanglement.

II.I. Datasets

CK+ The CK+ (Extended Cohn-Kanade) dataset [7] consists of 593 videos from 123 subjects. Each video begins with a neutral facial expression and progresses to one of the basic emotions as defined by Paul Ekman [8]: anger, contempt, disgust, fear, happy, sadness, and surprise. The preprocessing steps ensure consistent analysis, including cropping to isolate facial regions.

Mead The MEAD dataset [9] was used for training because the CK+ dataset was too small for effective contrastive learning. MEAD consists of videos of 60 different actors reading sentences while expressing the basic emotions defined by Paul Ekman [8], each at three different intensity levels. Due to its massive size, we used approximately 800,000 images from 10 selected actors. The preprocessing included background removal and cropping to ensure uniformity across images.

CelebA The CelebA dataset [10] contains images of celebrities and is widely recognized for its extensive annotations and utility in attribute analysis. For this study, we used a subset of 50,000 images, for style mixing to support augmentation.

II.II. Generating Features Through StyleGAN Inversion

As mentioned in I, we require an effective method to map real images from datasets into the StyleGAN latent space. This mapping process, referred to as StyleGAN inversion, serves as the foundation for fine-grained semantic editing and disentanglement tasks.

However, while synthetic images generated by StyleGAN can often be mapped back into the original latent space W with satisfying results, the same is not true for real images. Due to the gap between the real and synthetic data distributions, real images cannot be directly mapped into W. Instead, they are mapped into an extended latent space W^+. In W^+, w is represented as a concatenation of N latent blocks $\{w^1, w^2, \ldots, w^N\}$, where each block controls a specific convolutional layer in the generator. This extension provides the additional flexibility needed to account for the greater complexity of real-world data, enabling a more accurate inversion [5].

The Feature-Style encoder proposed by Yao *et al.* [2] is an approach designed for StyleGAN2 inversion. It achieves this by encoding images into two complementary codes:

- **Latent code** ($w \in W^+$)**:** Capture global attributes such as style and overall appearance, enabling broad semantic edits.
- **Feature code** ($f \in F$)**:** Encode high-resolution spatial details crucial for precise reconstructions and targeted modifications.

This dual encoding mechanism enables the manipulation of specific attributes in the latent space while preserving critical details in the feature space. Using the Feature-Style encoder, we can generate realistic variations in appearance while maintaining the core emotional expression of the original image. This capability is pivotal for creating tailored positive pairs and ensuring robust disentanglement between appearance and expression.

II.III. SimCLR on Latent Space Features

SimCLR [5], or Simple Framework for Contrastive Learning of Visual Representations, is a self-supervised learning framework that trains models to recognize meaningful similarities in data by contrasting positive and negative pairs. The key idea is to bring similar data (positive pairs) points closer together in the representation space

while pushing dissimilar ones (negative pairs) further apart.

Traditionally, SimCLR relies on augmentations such as cropping, flipping, and color jittering to create transformations of the same image to form positive pairs. Negative pairs, on the other hand, are formed from representations of different data points. However, since we are working in the latent space of StyleGAN, these common augmentations are not applicable. Instead, we leverage specialized augmentations tailored to the latent codes: **Style-Mixing** and **Latent Space Slider**.

Style-Mixing This method replaces the last 10 style layers (e.g., layers 8–18) in the W^+ latent space of StyleGAN2 with the corresponding style layers from a randomly selected image in the CelebA dataset. This technique introduces significant variations in attributes such as facial structure and features [6].

Latent Space Slider This technique manipulates specific dimensions in the latent space that correspond to semantic features, including: Eyeglasses, Bags under eyes, Blurriness, Age, High cheekbones and Skin tone [6]

In Figure 1, it is shown how these augmentations affect the reconstructed images generated from edited latent codes.

Figure 1: Comparison of an original image (left) from the MEAD dataset and two reconstructed images (right) generated using edited latent codes. The edits maintain the angry emotional expression while introducing variations in appearance features.

The objective is to develop representations that focus solely on expression attributes while systematically filtering out unrelated factors such as facial appearance and pose. This is achieved through the application of SimCLR, enhanced by specialized augmentations tailored to the latent space , as demonstrated in Figure 2.

III. Results and discussion

When the approach is successful, the representations of images belonging to the same class should be highly similar. In an t-SNE plot, the data points corresponding to a single expression should ideally form distinct clusters. Figure 3a presents such a plot, where data points are color-coded according to their expressions. However,

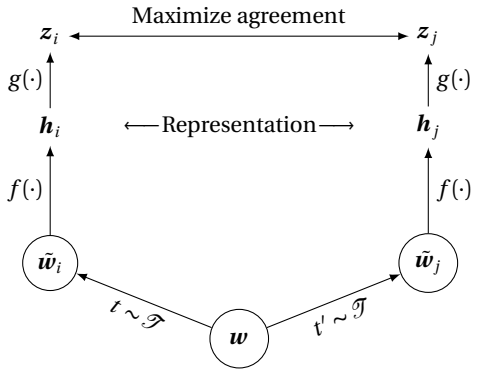


Figure 2: A framework for contrastive learning in the latent space of StyleGAN. Two separate augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) are applied to a latent code $\boldsymbol{w}$ to generate two correlated latent views ($\tilde{\boldsymbol{w}}_i$ and $\tilde{\boldsymbol{w}}_j$). A base encoder $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training, only the encoder $f(\cdot)$ is retained to produce latent representations $\boldsymbol{h}$ for downstream tasks, such as disentangling expression from appearance. Adapted from Chen *et al.* [5]

it can be observed that while some clusters emerge, most consist of data points from multiple expressions. Only a small number of points from the class surprise are closely located.

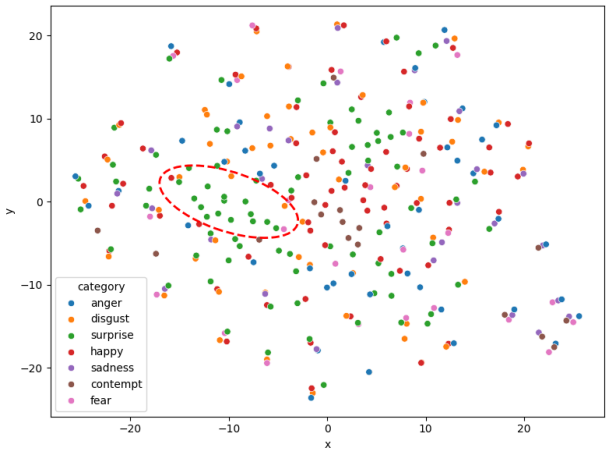
In Figure 3b, the t-SNE plot reveals that these clusters predominantly group images of the same individual rather than those of the same expression. This observation indicates that in the new representation space, appearance-related information remains dominant.

Emotion classification using an SVM and the encoder trained with SimCLR achieved an accuracy of 77%. In contrast, a CNN trained on the latent codes of the Feature-Style encoder achieved an accuracy of 93.16%. Interestingly, this accuracy improved to 95.82% when augmented with transformations similar to those used for SimCLR. These results highlight the need for further disentanglement of expression and appearance information in the learned representations.

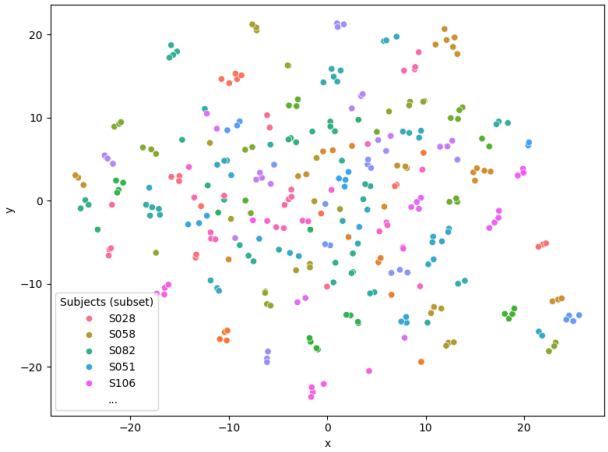
Table 1: Accuracy of emotion classification using SimCLR and CNN-based approaches. Results highlight the improvement achieved with augmentations tailored to disentanglement tasks.

	Accuracy
SimCLR	77 %
CNN	93.16 %
CNN + augmentations	95.82 %

10.18416/SCP.2025.1993



(a) Data points color-coded by emotional expressions reveal partial clustering by expression class, with some overlap.



(b) Data points color-coded by subject identity show dominant grouping based on appearance rather than expression, highlighting residual appearance information in the representation space.

Figure 3: t-SNE visualizations of the CK+ dataset demonstrating the clustering of learned representations.

IV. Conclusion

In this work, we presented a framework leveraging StyleGAN inversion and SimCLR to disentangle facial expressions from appearance attributes. Although the methodology demonstrated promise, particularly through tailored augmentations like Style-Mixing and Latent Space Slider, the results reveal that residual appearance information still influences the learned representations. Improvements in emotion classification accuracy with CNN and augmented datasets indicate potential avenues for refinement.

Future work should focus on further refining disentanglement techniques, exploring additional augmentations, and addressing the limitations highlighted in the t-SNE analysis. By achieving a more robust separation between expression and appearance features, the framework can pave the way for advancements in applications ranging from emotion recognition to privacy-preserving medical diagnostics.

Acknowledgments

The work has been carried out at the German Research Center for Artificial Intelligence (DFKI) in Lübeck and supervised by the Institute of Medical Informatics, Universität zu Lübeck.

I would like to thank the head of the DFKI in Lübeck for providing me with the opportunity and resources to conduct this work at their institute. I am also thankful to my supervisor, for the invaluable guidance and support throughout this project.

Author's statement

Authors state no conflict of interest. Writefull was used for the linguistic fine-tuning of this paper.

References

[1] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, Analyzing and improving the image quality of stylegan, 2019. doi:10.48550/ARXIV.1912.04958.

[2] X. Yao, A. Newson, Y. Gousseau, and P. Hellier, Feature-style encoder for style-based gan inversion, 2022. doi:10.48550/ARXIV.2202.02183.

[3] X. Peng, X. Yu, K. Sohn, D. Metaxas, and M. Chandraker, Reconstruction-based disentanglement for pose-invariant face recognition, 2017. doi:10.48550/ARXIV.1702.03041.

[4] Y. Nitzan, A. Bermano, Y. Li, and D. Cohen-Or, Face identity disentanglement via latent space mapping, 2020. doi:10.48550/ARXIV.2005.07728.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, 2020. doi:10.48550/ARXIV.2002.05709.

[6] K. Duarte, W.-A. Lin, R. Kalarot, J. Lu, E. Shechtman, S. Ghadar, and M. Shah, Contrastive learning on synthetic videos for GAN latent disentangling, in *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022. URL: https://openreview.net/forum?id=B9W7BV6fRC.

[7] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, IEEE, 2010. doi:10.1109/cvprw.2010.5543262.

[8] P. Ekman, Gefühle lesen, Wie Sie Emotionen erkennen und richtig interpretieren, 2. Aufl., unveränd. Nachdruck, S. Kuhlmann-Krieg, Ed. Heidelberg: Spektrum Akademischer Verlag, 2012, 389 pp., ISBN: 9783827425683.

[9] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, Mead: A large-scale audio-visual dataset for emotional talking-face generation, in *ECCV*, 2020.

[10] Z. Liu, P. Luo, X. Wang, and X. Tang, Deep learning face attributes in the wild, in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.