Robust Discrimination and Generation of Faces using Compact, Disentangled Embeddings

Björn Browatzki
Korea University
Seoul
browatbn@korea.ac.org

Christian Wallraven
Korea University
Seoul
wallraven@korea.ac.kr

1. Summary of supplementary materials

In the following experiments, we present additional results that validate the autoencoder part of our architecture (Sec. 2,3), present examples of how disentangling and joint training of the full pipeline improves the reconstructions (Sec. 4), visualize the compact feature embeddings (Sec. 5), provide additional results for attribute interpolation of images (Sec. 6), and finally list results of how the different loss terms of the pipeline affect generative and discriminative performance (Sec. 7).

2. Matches in training set

In Fig.1 we show reconstructed images from our autoencoder part that were previously unseen as well as the most similar images contained in our training set. The first row shows input images from the CelebA test set. Below are reconstructed images created by the autoencoder trained on the VGGFace2(1M) training set (note, that the autoencoder therefore has never seen any image from either training or test sets of CelebA). The remaining images are faces with low MMSIM distance [3] queried from the training set. Reconstructions of test images are very accurate but do not hallucinate previously seen training images indicating that our autoencoder can generate novel faces.

3. Random samples

To further illustrate the generative capabilities of our autoencoder, Fig.2 shows randomly generated images. We show results from two recent methods that combine generative networks with inference capabilities [2,4] as qualitative comparison (note, that due to the intended use for face analysis tasks, we operate on tighter cropped faces). Although artefacts can occur in the background of faces with extreme poses, our results are comparable if not better qualitatively to the state-of-the-art.

4. Reconstructions with disentanglement

The disentanglement process increases the fidelity of face reconstructions as can be seen in Fig.3. Each of the four blocks shows (unseen) input images in the top row. The following two rows show reconstructions *before* disentanglement training and *after*. Note, how subtle facial attributes are much better preserved in the reconstructions after disentanglement training: male faces are often given a female bias, which is gone after disentanglement training (cf. the second and eighth face in the first block). Similarly, expressions are reproduced much more faithfully following training with the whole pipeline.

5. Learned embedding spaces

Fig.4 shows t-SNE [1] visualizations for the expression attribute f_e and its complement $f_{\neg e}$. Colors represent ground truth annotations - note, how the space groups together similar colors and how the space overall also captures valence (top left = positive, bottom right - negative) and arousal aspects (top - right = weak, bottom = strong). Importantly, the embedding space for $f_{\neg e}$ loses this information and instead groups faces based on on identity and style similarity.

The embeddings space for attribute f_{id} is visualized in Fig.5, presenting several clear clusters of the colorcoded CelebA identities - again, this clustering is broken up when using the complement $f_{\neg id}$.

6. Attribute interpolations

Attribute feature vector interpolations are visualized in Fig.6. Attributes are interpolated between the source images (left column) and the target images (right column). Columns inbetween show reconstructions of interpolated attribute vectors with 10 steps from 0% to 100% (hence, the second and the second-to-last picture represent the reconstructions of the source and target faces, respectively). We demonstrate interpolation of pose, identity, expression,

and style in the first four rows and interpolation of all information *except* for pose, identity, expression, and style in rows 5-8. Finally, interpolation of the entire feature vector is shown in the last row.

Note, how the interpolations follow the given attribute well, keeping all other facial attributes constant throughout the process.

7. Ablation studies

We evaluate the impact of individual loss terms in Tab.1 for the autoencoder part of the pipeline and in Tab.2 for the disentanglement part of the pipeline.

In Tab.1 we can observe the trade off between high reconstruction accuracy on the training set (as measured by RMSE) and generalization performance to unseen samples (as measured by FID), ranging from (A) to (C). The adversarial autoencoder produces blurry images, which exhibit low RMSE values but high FID scores as they are dissimilar from real images. Importantly, expression and identity recognition performance does not yet benefit from more realistic reconstructions without the additional disentanglement step as joint training results ("J") are much better overall.

In Tab.2 we study the effect of different training configurations on discriminative face analysis performance. A purely discriminative setup (A) serves as baseline - by removing the reconstruction term, however, we also lose all generative capabilities of the system. These capabilities are restored with the loss terms in (B) and (C). Overall, both loss types manage to maintain attribute information in their attribute vectors, however, by means of the disentanglement process, the attribute information content in the *complementary parts of the attribute vector is reduced*, as indicated by lower $f_{\neg e}$ and $f_{\neg id}$ scores. Finally, best overall performance for the full pipeline is achieved with the augmentations loss L_{aug} in (F) with joint training through the full pipeline delivering best results on both expression and identity recognition tasks. Rows (D) and (E) demonstrate that it is possible to improve the results even further when training with one modality only. Although this setup loses the multi-attribute capabilities central to our framework, these results show that some discriminative information is lost through multi-attribute disentanglement as compared to single-attribute disentanglement - an issue that remains to be improved for future work.

References

[1] Laurens van der Maaten and G. Hinton. Visualizing Data using t-SNE Laurens. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 1, 6, 7

[2] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed. Variational Approaches for Auto-Encoding Generative Adversarial Networks. 2017. 1, 4

[3] E. P. Simoncelli, H. R. Sheikh, A. C. Bovik, and Z. Wang. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1, 3

[4] D. Ulyanov, A. Vedaldi, and V. Lempitsky. It Takes (Only) Two : Adversarial Generator-Encoder Networks. pages 1250–1257, 2016. 1, 4

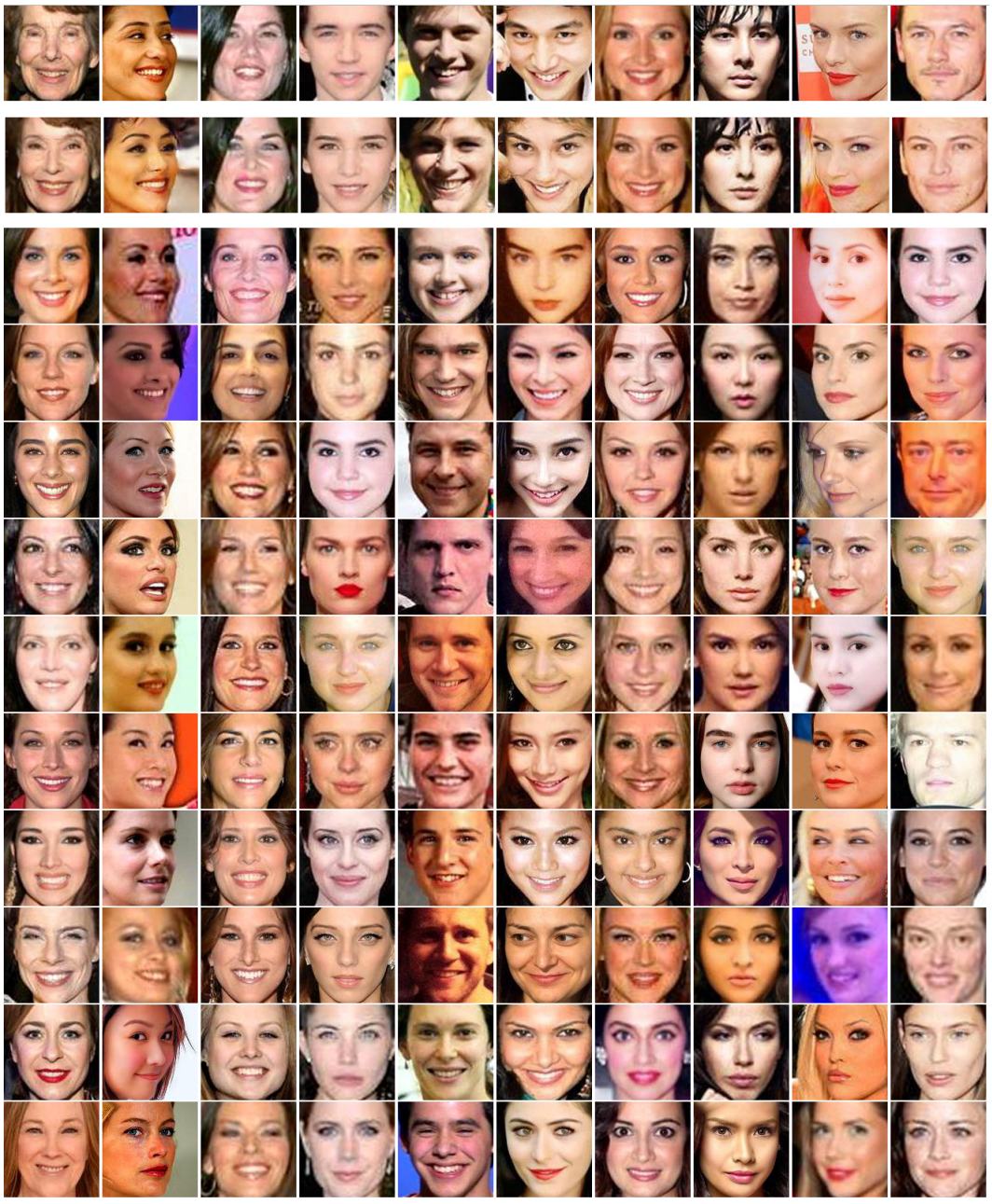
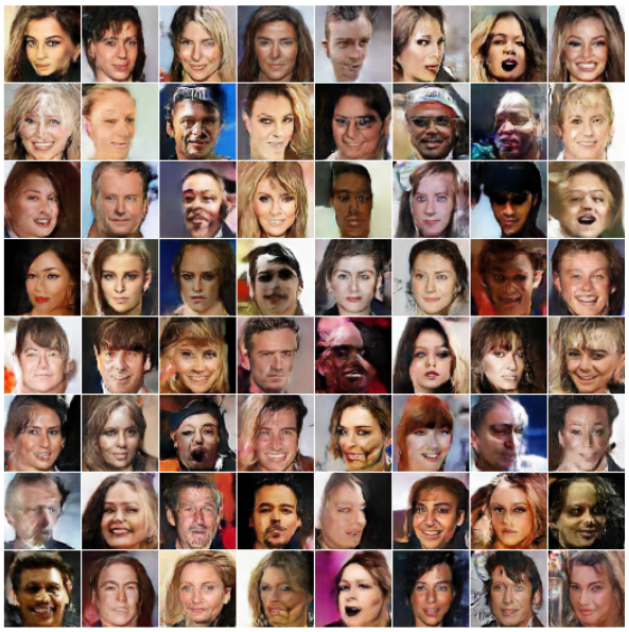
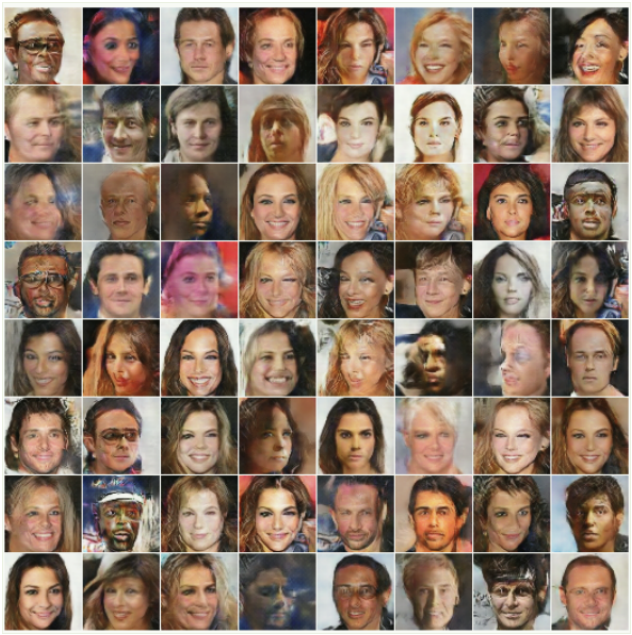


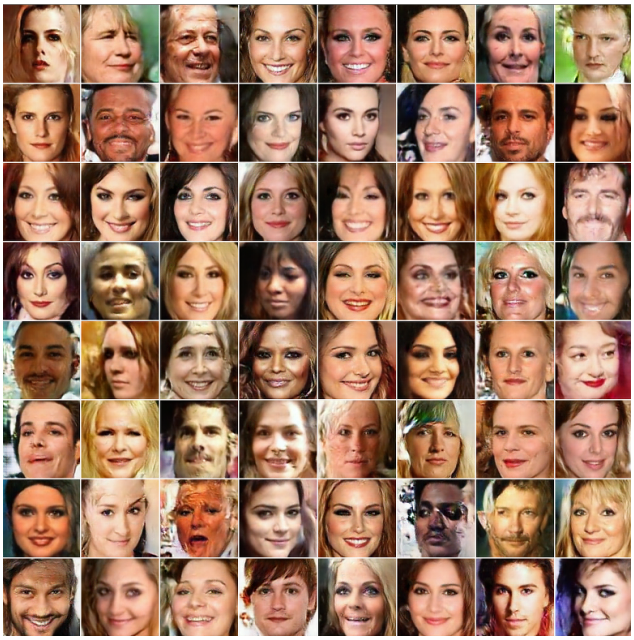
Figure 1: Top: Input images from CelebA test set. Second: Reconstucted images. Next 10 rows: Similar images in the training set (VGGFace2 1M) to the reconstructed images. Similarity measured using MSSIM [3] evaluated on the shown crops.



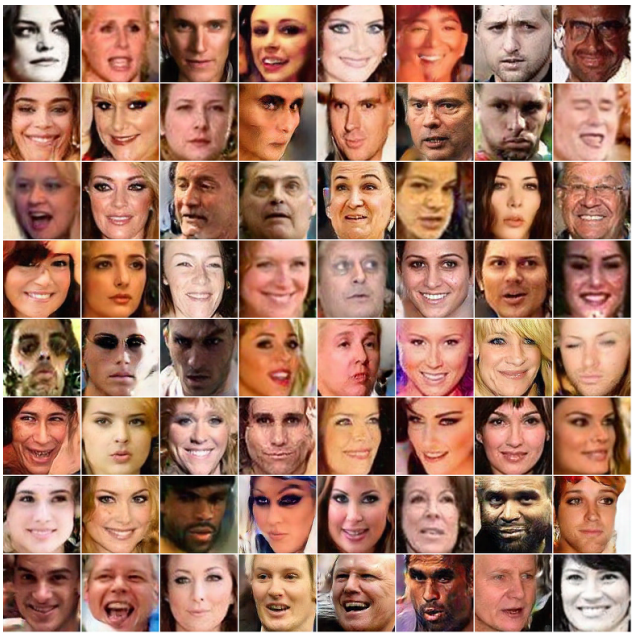
(a) AGE [4] (results from [2])



(b) α-GAN [2] (results from [2])



(c) Ours (trained on CelebA)



(d) Ours (trained on VGGFace2)

Figure 2: Random faces produced by generative networks.

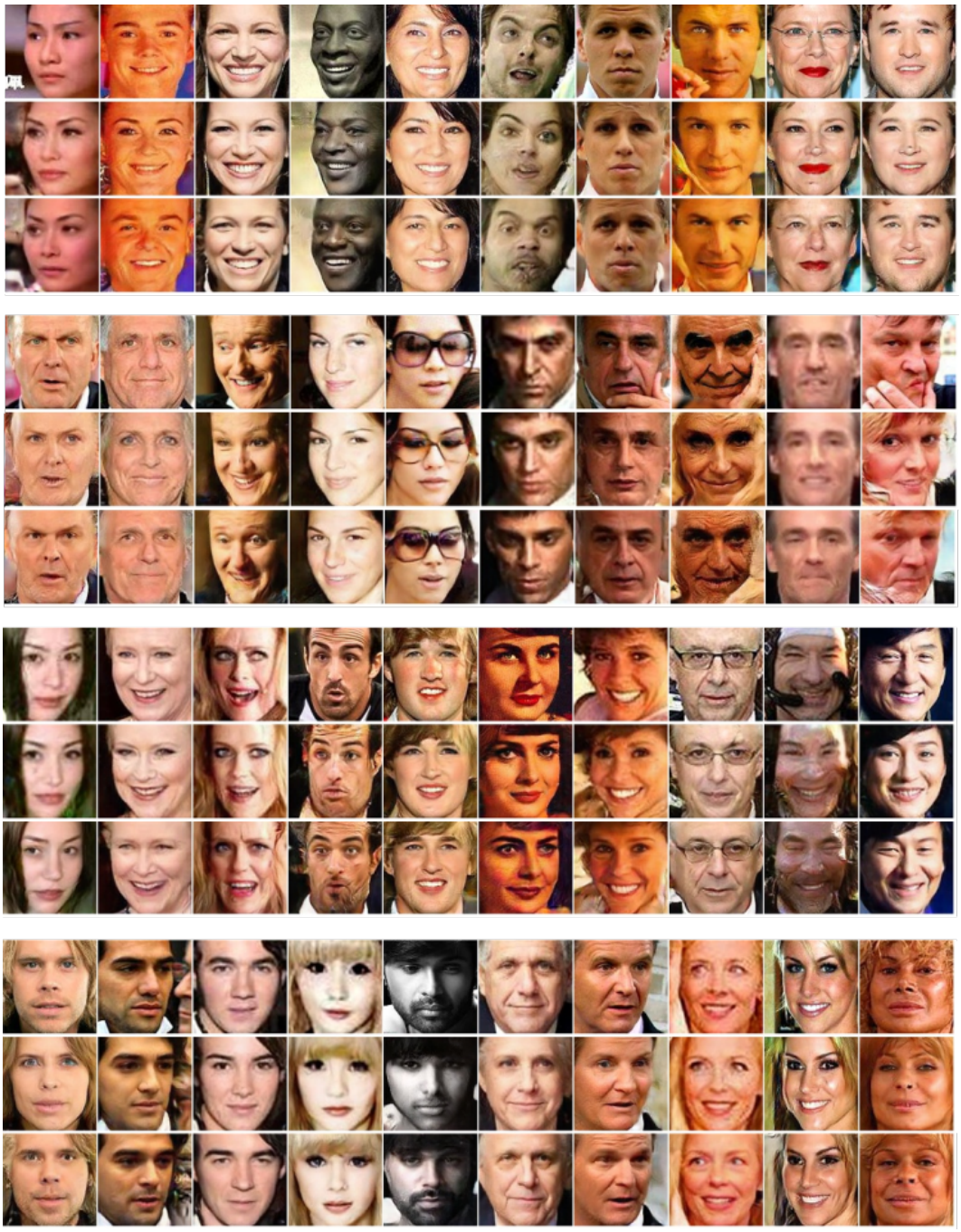
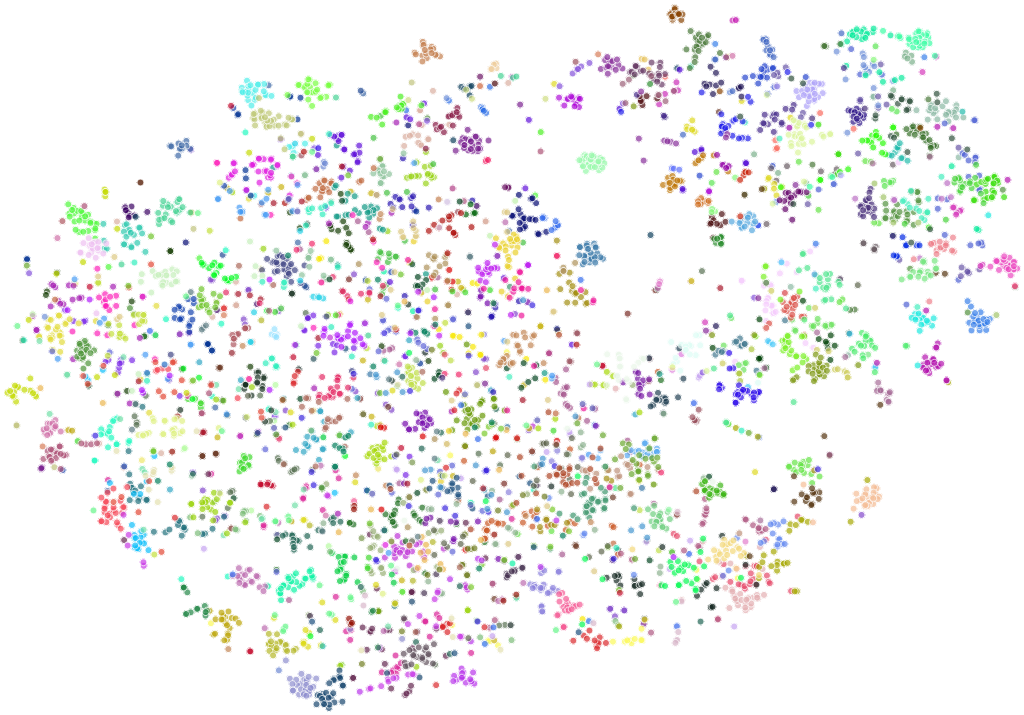


Figure 3: Recontructed images from VGGFace2 test set. First row in each block shows input images. Second row: Reconstruction before training disentanglement. Third row: Reconstruction after joint disentanglement training. Expressions and identity are better preserved in reconstructions after disentanglement training. Note the subtle shifts in gender for reconstructions without disentanglement.

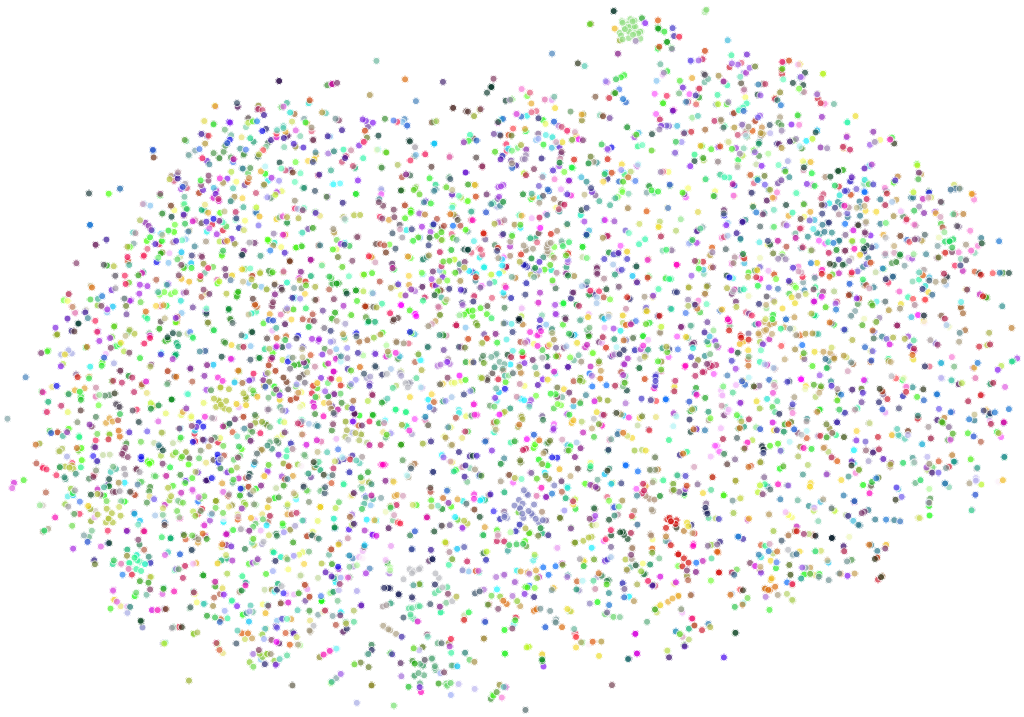
(a) f_e

(b) $f_{\neg e}$

Figure 4: t-SNE [1] visualization of expression embedding space. All 4000 samples from AffectNet test set. Colors indicate ground truth expression categories. Images are overlayed for a random subset of samples to illustrate annotation ambiguities. Frame colors correspond to expression category. Best viewed in color and zoomed in.



(a) f_{id}



(b) $f_{\neg id}$

Figure 5: t-SNE [1] visualization of identity embedding space for 5000 samples from CelebA test set. Colors indicate ground truth identity labels

(a)

(b)

Figure 6: Attribute interpolations. Source images are depicted in left column, target images in right column. Columns inbetween show reconstructions of interpolated attribute vectors. Different attributes in each row: Pose, identity, expression, and style in rows 1-4 and all information *except* for pose, identity, expression, and style in rows 5-8. Interpolation of entire attribute vector in last row.

	$\mathcal{L}_{rec}$	$\mathcal{L}_{enc}$	$\mathcal{L}_{adv}$	$\mathcal{L}_{gen}$	**Random FID** VGG	**Reconstruction RMSE** train/test VGG	AffectNet	CelebA	**Expr. rec. AUC** AffectNet	**Verific. Acc. (%)** LFW	
(A) Adversarial Autoencoder	✓	✓			65.97	15.04/28.16 16.16/17.71	15.55/30.66 16.15/18.36	14.09/25.75 15.05/15.80	0.7855 0.8414	82.0 82.9	**S** **J**
(B) Adv. AE w/ GAN	✓	✓	✓		24.53	18.01/23.09 19.20/21.42	18.77/24.26 19.24/22.01	16.77/20.91 17.55/19.25	0.8006 0.8420	80.4 83.3	**S** **J**
8908708904854 (X) 195D	✓	✓	✓		14.61	18.01/23.09 19.20/21.42					**S** **J**
(X) 195D	✓	✓	✓		24.38	18.01/23.09 19.20/21.42			 0.8420	 83.3	**S** **J**
(C) Full objective	✓	✓	✓	✓	14.76	18.16/26.58 19.37/21.08	19.01/27.44 19.37/21.85	16.91/23.27 17.71/18.82	0.7951 0.8404	81.3 83.2	**S** **J**

Table 1: Results of autoencoder ablation study. **S** denotes separate training of autoencoder and disentanglement. **J** denotes joint training.

	P	**I**	**E**	**S**	$\mathcal{L}_f^\Phi$	$\mathcal{L}_{rec}^\Phi$	$\mathcal{L}_{cyc}^\Phi$	$\mathcal{L}_{aug}^\Phi$	**AffectNet (AUC)** f_e	$f_{\neg e}$	**LFW Acc. (%)** f_{id}	$f_{\neg id}$	
(A) Discrimin. only	✓	✓	✓	✓	✓				0.8146 0.8426	0.7786 0.7833	82.0 83.3	68.0 67.8	**S** **J**
(B) No disentanglement	✓	✓	✓	✓	✓	✓			0.8036 0.8422	0.7377 0.7488	83.7 83.0	65.9 67.6	**S** **J**
(C) No augmentation	✓	✓	✓	✓	✓	✓	✓		0.8045 0.8413	0.7346 0.7447	82.0 83.3	67.1 67.2	**S** **J**
(D) Identity only		✓			✓	✓	✓	✓	– –	– –	82.6 86.7	64.3 63.7	**S** **J**
(E) Expression only			✓		✓	✓	✓	✓	0.8187 0.8669	0.5880 0.7328	– –	– –	**S** **J**
(F) Full objective	✓	✓	✓	✓	✓	✓	✓	✓	0.8107 **0.8513**	0.7295 0.7578	81.2 **84.4**	67.6 67.8	**S** **J**

Table 2: Results of disentanglement ablation study. **S** denotes separate training of autoencoder and disentanglement. **J** denotes joint training.