Learning Disentangled Representations via Independent Subspaces

Maren Awiszus, Hanno Ackermann, Bodo Rosenhahn
Leibniz University Hannover
Hanover, Germany
{awiszus, ackermann, rosenhahn}@tnt.uni-hannover.de

Abstract

Image generating neural networks are mostly viewed as black boxes, where any change in the input can have a number of globally effective changes on the output. In this work, we propose a method for learning disentangled representations to allow for localized image manipulations. We use face images as our example of choice. Depending on the image region, identity and other facial attributes can be modified. The proposed network can transfer parts of a face such as shape and color of eyes, hair, mouth, etc. directly between persons while all other parts of the face remain unchanged. The network allows to generate modified images which appear like realistic images. Our model learns disentangled representations by weak supervision. We propose a localized resnet autoencoder optimized using several loss functions including a loss based on the semantic segmentation, which we interpret as masks, and a loss which enforces disentanglement by decomposition of the latent space into statistically independent subspaces. We evaluate the proposed solution w.r.t. disentanglement and generated image quality. Convincing results are demonstrated using the CelebA dataset [24].

1. Introduction

Neural networks (NNs) are the current algorithm of choice for many different applications. NNs show impressive results on various tasks but a disadvantage is their function as black box, i.e. it is very difficult to retrace and understand the decisions made within. Our method is based on a special form of neural networks called autoencoders. These are network architectures with a bottleneck layer [14] which enforces that a low-dimensional representation of data is learned. The activations of the bottleneck layer define what is called a latent space. In this work, we present an approach which allows, to some extent, a semantic interpretability and control of the latent space of such an autoencoder.

Traditional autoencoders usually have no incentive to construct the latent space in an interpretable manner. This means that samples which are perceived to be similar in the original space do not have to be close in the latent space. A standard approach to handle this problem is to apply a variational autoencoder (VAE) [20], where a normal distribution is enforced on the latent space. Another, more recent method to enforce a distribution are generative adversarial networks (GAN) [9], which do so by adding a network called a discriminator. With the enforcement of a distribution, samples that are close in the latent space are also similar in the reconstructed space. However, it is still unknown

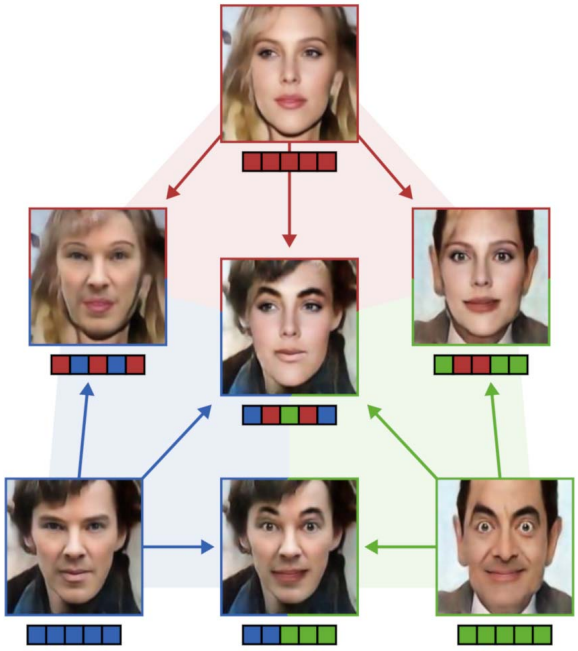

Figure 1. Autoencoded images and combinations of thereof. The corner images are directly autoencoded. The images in between are new combinations generated with our proposed network. The center image is a combination of all three corner images. The color code below the image corresponds to which parts of the face are taken from which image (from left to right): 1. Background with hair, 2. Face with nose, 3. Eyebrows, 4. Eyes and 5. Mouth.

which changes in the latent space affect which parts of the original data and might even change it in its entirety. Affecting only certain parts of the data (e.g. only image regions) requires a disentanglement, which would allow different dimensions of the latent space to change known, possibly independent parts of the original data.

Face images are a prime example for data that can be disentangled. For example attributes such as gender or hair color can be changed [13], but usually the entire image is affected from these modifications. In this work, however, we want to disentangle the face regions such as the face skin, the eyes and the mouth, which allows us to change one while we keeping the rest unchanged. The authors of [11] approach this problem by using a combined network with an adversarial autoencoder for each face region with an additional network for the background. In contrast to this work, our goal is to create one end to end trainable neural network with semantic interpretability and control of the latent space. To achieve this, we propose a NN architecture which decomposes the latent space into subspaces which are mutually statistically independent, much akin to independent subspace analysis (ISA). It allows for correlation between signals of the same subspace whereas signals of different subspaces are statistically independent. This is motivated by the fact that pixels which are close often correlate, whereas pixels from semantically different parts of the face can be quite different. The network is trained with two specially designed losses: a mask loss and an entropy loss. As consequence, our method does not rely on a semantic segmentation of the input data. Instead, it is implicitly learned by our network.

Fig. 1 shows results from our network. From the autoencoded images in the corners, it is possible to combine parts in the latent space to create new, mixed images. The mixed images recognizably show the parts they are made of, while still showing a coherent face image.

In summary, our **contributions** are:

- A novel approach for semantic disentanglement of (image) data with autoencoders.
- A network architecture that allows for decomposition into statistically independent subspaces.
- An effective method for training such a model by using a mask and entropy loss.
- An evaluation and comparison of our method to existing other works. In contrast to other works, we achieve disentanglement of subspaces without the need of a semantic segmentation during testing.

1.1. Related Work

Subspace Analysis on Faces Traditional and well explored approaches for subspace analysis are for example principal component analysis (PCA) [18] and linear discriminant analysis (LDA) [25]. Independent component analysis (ICA) [15] and, as an extension of that, independent subspace analysis (ISA) which enforce low mutual information between their components/subspaces are further tools to achieve a more sophisticated disentanglement. An example for a very recent work that uses ISA for uncalibrated non-rigid factorization can be found in [3].

Several works exist to represent and model human faces as combination of subspaces, for example [2] proposes a 3D morphable model that can be used to disentangle shape from expression which in turn allows for applications such as facial reenactment [31]. Another way to parameterize faces is to use a multilinear tensor based approach as done in [32, 4]. Using a similar approach, the authors in [10] show that found subspaces can hold important information, such as an apathy mode.

Generation and Analysis of Faces with NNs Trying to generate faces with neural networks is a well known topic which gained increasing attention in the last years. One reason for using neural networks is to enhance the traditional models with more advanced non-linear embeddings. Taking an autoencoder as a basis, variational autoencoders (VAE) [20] enforce a normal distribution on their latent space by splitting the latent space in mean and variance layers and adding a distribution loss. However, as a black box approach, it is not possible to explicitly control which transformations in the latent space can correspond to meaningful transformations in the image space. Most current works regarding face image generation rely on the similar concept of adversarial learning. First introduced in [9], these generative adversarial networks (GANs) allow for the enforcement of a distribution via a discriminator. Similar to VAE, they can shape data in the latent space, but are not guaranteed to do so in a meaningful way. An extension of this method is InfoGAN [8] in which an additional information loss is used to semantically meaningful directions in the dimensions of the latent space. This method is akin to a learned ICA, where each component or dimension is trained to have barely any mutual information with the other dimensions. Further works use latent spaces: A CycleGAN [36] makes it possible to transfer an image from one domain into another. Such a CycleGAN is trained on two sets of unpaired data, for example real photos and drawings. Shape and identity of an image can be disentangled by the FusionGAN framework [19] to allow for the fusion of a different shape and identity. The AttGAN [13] is an adversarial autoencoder which makes it possible to change given attributes of faces. In contrasts to [13], our work does not rely on manually annotated attributes, but on masks which can be obtained in an unsupervised manner.

Image Generation and Editing with Masks Using masks or semantic segmentations supports the generation and editing of images. There are many works that use deep neural networks to generate realistic images from masks, for example [7, 6] and the more widely known pix2pix [16] and pix2pixHD [34]. There are also non-parametric methods which generally do not generate images themselves but fuse existing images or parts thereof. Examples include [5, 12, 22]. Another approach is sketchGAN [17], where masks are manually drawn onto the images to allow for easy and fast editing.

The recent work of Gu *et al*. [11] is very similar to our work. The authors propose multiple separate autoencoders with an additional network for the background. Compared to their work, we only train one network. This is done because we want to learn and later analyze the structures in one representative latent space, for which multiple networks are not feasible. Additionally, our method does not require a semantic segmentation of the face which is needed to generate separate input masks in [11]. Insteadt, we only rely on the semantic segmentation during the training phase but not during testing. Thus, the semantic segmentation is implicitly learned from our network.

2. Method

In the following section, we describe the proposed method. First, we describe the architecture with emphasis on our contribution, the architecture that allows decomposition into independent subspaces. Afterwards, we define the losses for training, in particular the mask loss and the entropy loss.

2.1. Autoencoder

For the architecture depicted in the first row of Fig. 2, we start with a Resnet encoder Q to extract features from the image data. These features are projected onto a latent space of d dimensions using fully connected layers. We indicate samples from this as space z-samples. A Resnet decoder P reconstructs the images. The full Resnet autoencoder architecture is based on the architecture found in [36]. In contrast to [36], we use a fully connected layer after the last convolutional layer of the encoder to obtain a latent space of pre-determined size. The transformation can be formulated as $z = Q(I_{in})$ and $I_{out} = P(z)$. where $z_{enc} = z_{dec}$.

For training, the first loss L_a is the standard autoencoder loss: a mean squared error between input image I_{in} and output image I_{out}. We also add a gradient loss L_g between the input and output image to encourage sharper images

$$L_g = \frac{1}{p}\|\nabla I_{in} - \nabla I_{out}\|_F^2 \tag{1}$$

with p being the number of pixels.

2.2. Group Independence

Simply training the network as depicted up to this point would lead to an entangled latent space, i.e. changes to a single neuron of the latent vector lead to global changes in the decoded image. While maximizing statistical independence between all 1-dimensional directions is a viable option as [8] demonstrates, we aim to infer directions which correspond to different parts of faces. While modifying coordinates of points in latent space along 1-dimensional directions can be reasonable expected to model for instance color differences of skin and hair, variations of mouth, eyes, etc. , are unlikely to be as much compressible. It is thus much more reasonable to expect *multi-dimensional* subspaces to correspond to such parts of faces. Since different parts of faces should have little in common, subspaces should be as different as possible. Similar to [8], we employ statistical independence as measure of dissimilarity. In contrast to [8], we also allow for correlations within the same subspace, motivated by the fact that variations in the same semantic region of a face are often highly correlated. The difference between the model used in [8] and the one proposed here is similar to the difference between independent component analysis (ICA) and independent subspace analysis (ISA).

To factorize the latent space into mutually independent subspaces, we define a non-singular matrix A such that source signals $s \in \mathcal{S}$ can be obtained from the encoder outputs z_{enc} by

$$s = A^{-1} \cdot z_{enc} \tag{2}$$

and the inputs to the decoder z_{dec} by

$$z_{dec} = A \cdot s. \tag{3}$$

The matrix A is equivalent to the product between the mixing and the permutation matrices used in classical ISA.

The layers used for the decomposition into independent subspaces should not influence the reconstruction loss. Due to the requirement that $z_{dec} = z_{enc}$, we may therefore *skip* the layers for independent subspaces decomposition during backpropagation of the reconstruction loss. These layers are instead trained by two different losses: a mask loss L_m and an entropy loss L_e. Both will be explained in the following sections.

2.3. Mask Loss

We need to infer a latent space for which it is known which of its variables change which parts of the image. The mapping between image area and particular variables in latent space is enforced by a mask loss L_m. It is calculated as follows: Two images, input I_{in} and target I_t, are mapped to $s_{in} = Q(I_{in})$ and $s_t = Q(I_t)$ respectively. An interpolate can be defined by

$$s_{mix} = D_{-m} \cdot s_{in} + D_m \cdot s_t. \tag{4}$$

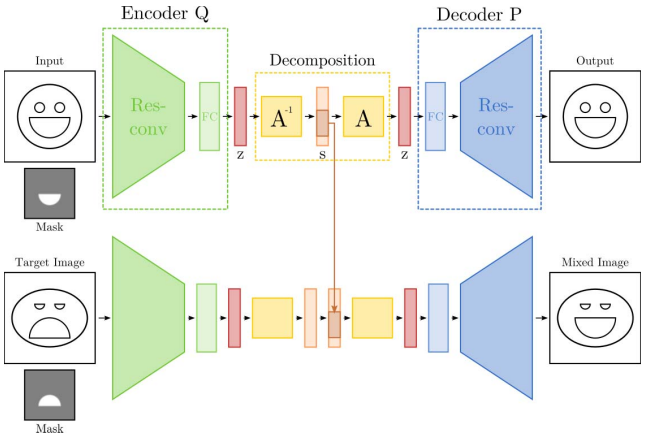

Figure 2. Architecture of the resnet autoencoder. The Res-conv parts of the network are motivated from [36]. Additionally we propose a intermediate decomposition into independent subspaces connecting both blocks. The proposed network projects the original latent space sample z into a disentangled latent sample s.

where m is the the index of the currently selected mask, D_m is a diagonal matrix whose entries corresponding to mask m equal 1 whereas all others equal 0, and D_{-m} is a diagonal matrix whose diagonal entries not corresponding to mask m equal 1 and all others equal 0. We decode s_{mix} to obtain the mixed image I_{mix}. By $M_{i,in}$ and $M_{i,t}$, respectively, we indicate the areas of input and target images corresponding to the ith mask. Whenever a variable associated with a particular mask is changed, regions of I_{mix} that are outside of $M_{i,in}$ and $M_{i,t}$ should be identical to I_{in}. On the other hand, regions of I_{mix} that are inside, need be identical to I_t. This can be formulated as

$$\begin{aligned} L_m = (I_{mix} - I_{in}) \cdot (1 - \max(M_{i,in}, M_{i,t})) \\ + (I_{mix} - I_t) \cdot \min(M_{i,in}, M_{i,t}). \end{aligned} \tag{5}$$

The process is visualized in Fig. 2.

2.4. Entropy Loss

InfoGAN [8] aims to disentangle data by learning what amounts to an independent component analysis (ICA). It maximizes statistical independence between each dimension of the latent space by formulating it as a classification task where each dimension is interpreted as a class to be separated. In many data, for instance face images, however, the complexity of the set of all possible configurations of, e.g., a mouth prohibits using a single vector only. To account for that shape complexity, we propose to allow for correlations between particular groups of variables but mutual statistical independence between the groups. In the following, it will be explained how this prior can be enforced by a neural network.

All d_i variables X_i of a batch corresponding to the ith out of C subspaces are selected, and and mapped by a function $F_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{max}}$ with d_{max} being the maximum number of variables associated with any subspace. The matrix $X = \begin{bmatrix} X_1 & \cdots & X_C \end{bmatrix}$ consisting of stacked matrices X_i can now be used to learn a decision problem with C classes. In other words, this technique can be used to learn a decomposition into multi-dimensional subspaces, i.e. a decomposition which is similar to independent subspace analysis (ISA) in classical statistics. The architecture of the proposed model is shown in Fig. 3.

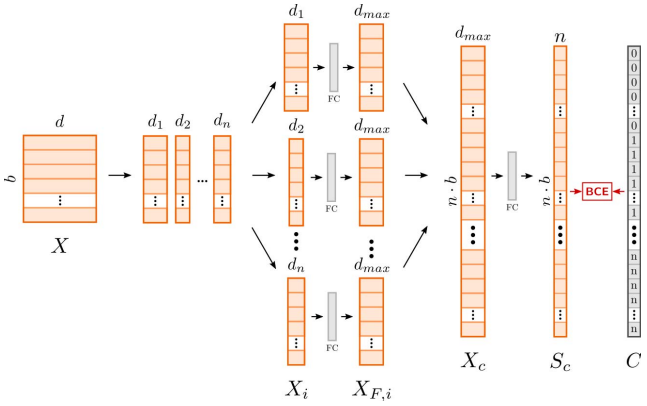

Figure 3. During training, the subspaces are interpreted as different classes so that a binary cross entropy loss can be used as an additional loss of our model.

The function F_i are implemented by a fully-connected layer with ReLU activations. The last classificator is also based on a fully-connected layer with softmax activation. For the loss L_e, we use binary cross entropy.

All losses are combined by

$$L = \lambda_1 \cdot L_a + \lambda_2 \cdot L_g + \lambda_3 \cdot L_m + \lambda_4 \cdot L_e \tag{6}$$

3. Experiments

In this section, after describing implementation details and the dataset, we demonstrate the effect of our contribution on four different experiments. We show qualitative results of swapping face parts with our network, discuss the importance of our independent subspace decomposition as a contribution, analyze the information contained in the subspaces and finally discuss changing attributes of the face with our network in contrast to the state-of-the-art network AttGAN [13].

3.1. Implementation Details

The hyper-parameters in Eq. 6 are set to $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 1$, and $\lambda_4 = 1$. We use five different masks: background and hair (BG+hair), face, eyebrows, eyes and mouth. The dimensions of the corresponding subspaces are $d_1 = 512$, $d_2 = 256$, $d_3 = 128$, $d_4 = 128$, and $d_5 = 128$. For more information on the masks confer to sec. 3.3.

Figure 4. Some random examples of segmentations on the CelebA database [24]. These segmentations are used as masks when training our network (see sec. 2.3) and represent the five subspaces we want to disentangle.

3.2. Databases

We use two face databases, CelebA [24] and color FERET [26, 27]. The faces are cut out such that the mean of the points of the eyes and the mouth coincide. These images are aligned so that the masks overlap as much as possible. A total of 65880 images are extracted from CelebA, and 2225 images from color FERET. Each image is scaled 160×160. CelebA also contains attribute labels used in Secs. 3.6 and 3.7.

3.3. Generating Masks with Semantic Segmentation

To obtain masks, we train a semantic segmentation network on an extended Helen dataset [23, 30]. We simplify the annotations by merging the background and hair, face and nose, the left eye and right eye, the left eyebrow and right eyebrow, and both the lips and inner mouth masks together each. This results in 5 masks: background and hair (BG+hair), face, eyebrows, eyes and mouth.

We used an existing implementation of a fully convolutional VGG-net [29] from GitHub[1]. Some example results for CelebA are shown in Fig. 4. The label probability maps of the segmentation network are used as masks. Please further note that the semantic segmentation is only used during training and is not required in the testing phase.

3.4. Disentangling Face Images

With the combined dataset of the frontal views from CelebA [24], color FERET [26, 27] and the masks resulting from the semantic segmentation, we train the proposed network described in section 2. This network can now disentangle BG+hair, face, eyebrows, eyes and mouth in any image in a way that allows them to be recombined with the parts from any other face image.

In a first experiment, we selected images from the internet which are not part of the training set and encoded them. We then exchange their coordinates on the same subspaces, and then decode the resulting points to images. As can be seen in Fig. 5, the network succeeds to replace particular parts of the faces while keeping the remaining parts almost unchanged.

Depending on the two mixed images, some combinations are not as aesthetically pleasing as others. For example in the last row of Fig. 5, the mouth shape extracted is quite large, which does not perfectly fit the smaller faces in the set. Furthermore, hairstyles with bangs obscure parts of the face. If a hairstyle without or with different bangs is replaced, the network does not succeed to fill in the generated gap. As can be seen in the examples in the first result row of the figure, it can nonetheless generate pleasing results. Generally, the best results are achieved, when the mixed images have a similar jaw- and hairline.

3.5. Significance of the Entropy Loss

In the following experiment, we show that adding our independent subspace decomposition and entropy loss results in a significant improvement when mixing images. The network is trained twice, once with all losses, and once without the decomposition into independent subspaces.

First, we compare the reconstruction errors within changed masks. The smaller the error, the less influence other subspaces have on the one currently observed. To do this, we combine the encoded vectors s_j of batches of five randomly chosen images I_j in our dataset to create mixed vectors s_{mix}. This is similar to Eq. 4.

$$s_{mix} = \sum_{j=0}^{5} D_j \cdot s_j. \tag{7}$$

This vector is decoded to image I_{mix} and then multiplied with each of the original masks $M_{j,j}$ resulting in 5 masked images $I_{mix,j}$. The same is done for the original images I_j resulting in 5 masked images $I_{j,masked}$.

$$\begin{aligned} I_{mix,j} &= I_{mix} \cdot M_{j,j}, \\ I_{j,masked} &= I_j \cdot M_{j,j}. \end{aligned} \tag{8}$$

The first subscript of M indicates the number of the subspace, whereas the second indicates the number of the image of the current batch.

Next we calculate the difference between each pair of corresponding images, summarize all absolute pixel values

[1] https://github.com/divamgupta/image-segmentation-keras

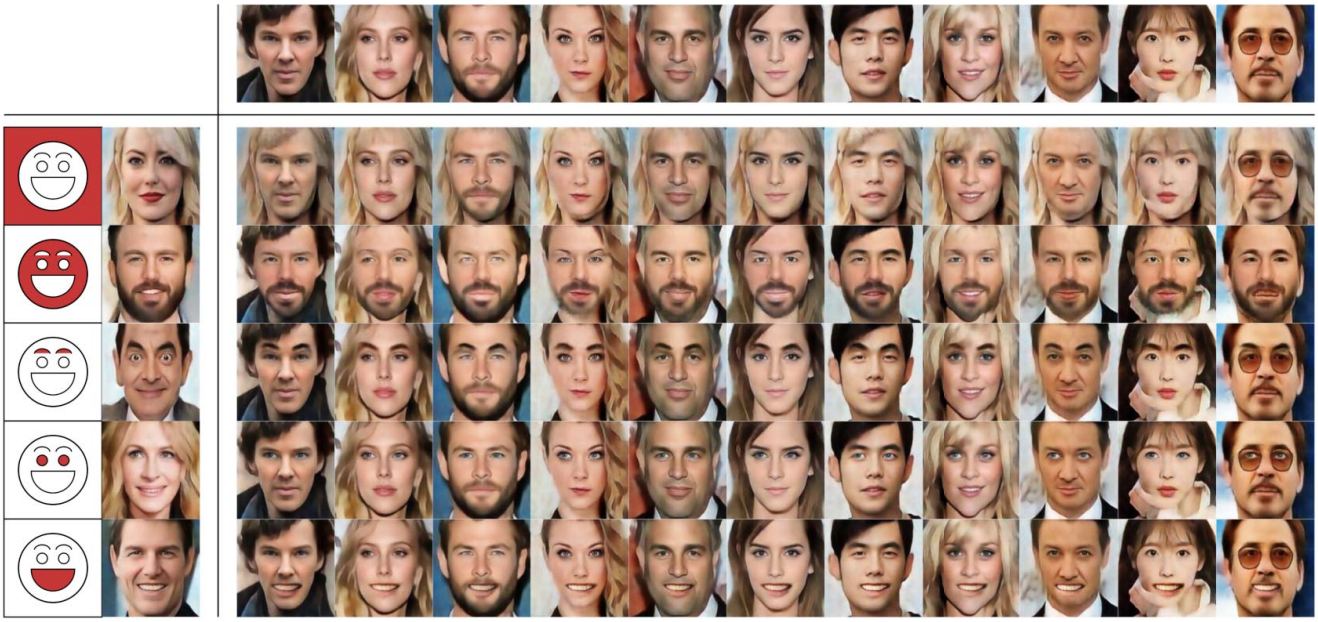

Figure 5. Examples of swapping Attributes from one image to another. The face graphics on the far left indicate, which part of the encoded s-vector of the face to their right was taken and transferred to each of the other images in the top row.

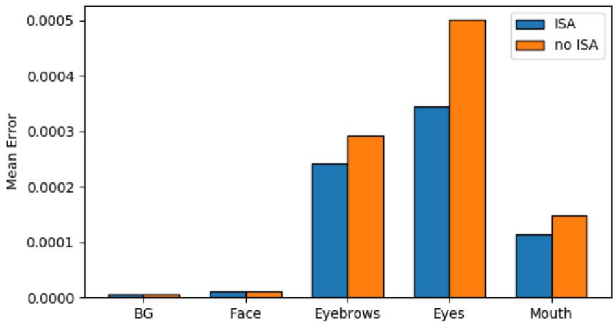

Figure 6. We split all of our data into sets of five and calculate the normalized differences between reconstructed and original image for each mask. The mean over all of those results is shown in the figure.

and divide this sum by the sum of pixels x, y of the corresponding mask.

$$e_j = \frac{\sum_x \sum_y |I_{mix,j}(x,y) - I_{j,masked}(x,y)|}{\sum_x \sum_y M_{j,j}(x,y)} \quad (9)$$

The resulting errors for each subspace are shown in Fig. 6. The overall error is low and also decreases significantly when using our proposed entropy loss. The highest relative error is on the eyes and eyebrows, as these are the smallest areas and therefore more sensitive w.r.t. the pixel-wise normalization.

Some examples of visible differences between the two reconstructed images are shown in Fig. 7. Without the entropy loss, the mask of the background which should perfectly replicate the left image of each pair shows some characteristics of the right image. The boxes highlight some noticeable changes. Especially the yellow box shows that without the entropy loss, the network can create facial parts such as hair where there should be none, in fact.

3.6. Subspace Analysis

We use the class labels found in the CelebA dataset (see sec. 3.2) to demonstrate that the proposed approach separates the latent space into interpretable subspaces. For the experiment, we encode the 65880 frontal faces of CelebA, and split their s-vectors according to their subspaces. The resulting 5 sets each have 65880 samples and dimensions as stated in sec. 3.1. For each of these datasets, we compute a principal component analysis (PCA) to reduce the dimensions to 3.

The labels considered in this experiment can be found in Tab. 1. Fig. 8 shows the first two principal components for both *mouth open* and *male* in all 5 subspaces, where orange points indicate that the attribute is true and blue that it is false. As can be seen, the samples for the attribute of *mouth open* are very mixed in every subspace aside from the mouth, but the samples of attribute *male* form clusters in every subspace.

This result confirms our claim that the subspaces are independent: An attribute that should only affect the mouth area of the image, *mouth open*, only affects the mouth sub-

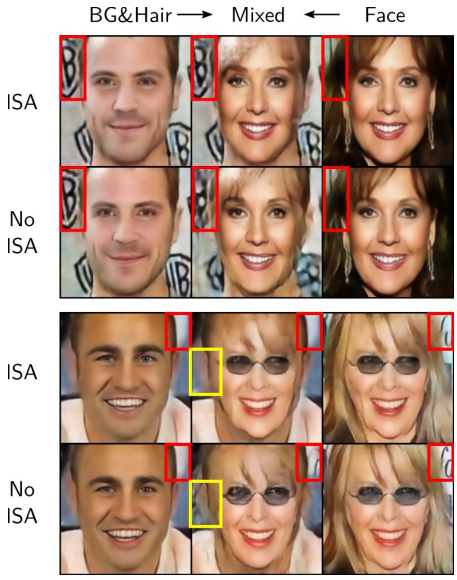

Figure 7. Visualization of the differences between training with or without entropy loss. Images in the upper row are with entropy loss (ISA), lower row are without entropy loss (No ISA). For each pair the center image is a combination of the outer images, with the BG+hair of the left image and the face of the right image. The boxes mark areas where the most notable changes can be seen.

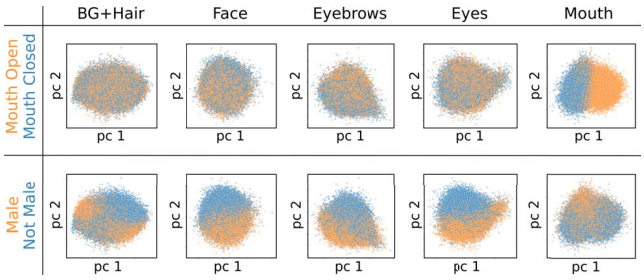

Figure 8. This figure shows the first two principal components for our subspaces, each in one column. In each rows, the points are colored in according to a given attribute of the dataset. The attribute *mouth open* is only easily separable in its associated subspace. This shows that the subspaces are independent. In the second row, we show the attribute *gender* as an example which is spread across all subspaces.

space, and an attribute that can affect the entire image, like the gender, affects all subspaces.

In Tab. 1, we show the distances based on the L_2-norm between the mean 3-dimensional PCA-vectors of all samples in which a label is true, and those in which the label is false. We do that for all previously mentioned labels. The highest value per label is set bold, while the lowest is set italics. It can be seen that most information about the *hair* is found in the BG+hair subspace and for *pale skin* it is in the face. While *Eyeglasses* would be expected to have its highest value in the Eye subspace, it being in the face

Table 1. L_2-norm distances between the center of samples for which an attribute is true and the center of samples for which it is not true (cmp. Fig. 8). The first 3 components of the PCA are used.

	BG+Hair	Face	Eyebr.	Eyes	Mouth
Bald	**3.675**	3.000	1.080	1.400	*0.984*
Bangs	3.051	**3.343**	0.997	*0.380*	0.641
Bla. Hair	**3.227**	1.494	0.980	*0.360*	0.488
Blo. Hair	**5.155**	2.390	1.395	*0.585*	0.937
Bro. Hair	**1.576**	1.249	0.423	*0.250*	0.414
B. Eyebr.	**1.605**	0.576	1.086	*0.267*	0.303
Glasses	1.421	**2.422**	1.636	2.388	*0.669*
Male	2.305	**3.052**	1.979	1.877	*0.932*
M. Open	0.416	1.035	0.663	*0.243*	**3.276**
Mustache	1.126	**2.956**	1.495	*1.057*	1.557
No Beard	1.520	**2.549**	1.543	1.144	*1.054*
Pale Skin	1.877	**3.623**	2.000	*1.292*	2.609
Young	**1.927**	1.677	0.990	1.214	*0.386*

subspace makes sense since glasses also cover up part of the face itself. The most surprising result is that the *Bushy Eyebrows* label seems to have the most information in the BG+hair subspace. This is most likely due to the fact, that there is a very high correlation between the bushiness of the eyebrows and the hair.

It should be noted that the actual underlying distributions of the classes might not be the same and therefore the distance between the means is only an indication of the information contained in the subspace.

3.7. Changing an Attribute of the Face

With the mean vectors of the different labels known from the previous section, we can now change images corresponding to them. This is done by subtracting the mean vector of all samples from the mean vectors of the given label and then adding multiples of that to any encoded sample. When decoded, that attribute is enhanced in the resulting image. This allows for a comparison with AttGAN [13].

We retrained AttGAN on the resized frontal images of CelebA also used in our training. AttGAN itself in its default configuration uses skip connections to improve their results, especially in regards to detail. We do not rely on skip connections, as the entire philosophy of trying to disentangle the *whole* image would be undermined by allowing information to skip this disentanglement. Therefore, we also train the AttGAN without the skip connections and reduce their latent space dimensions to the same as ours to allow for a fair comparison. In Fig. 9, results of our method in comparison to [13] can be seen for two examples when tranforming the gender. The images show, as expected, the detail of our reconstructions is lower than AttGANs with skip connections, but comparable without them. Also, disabling

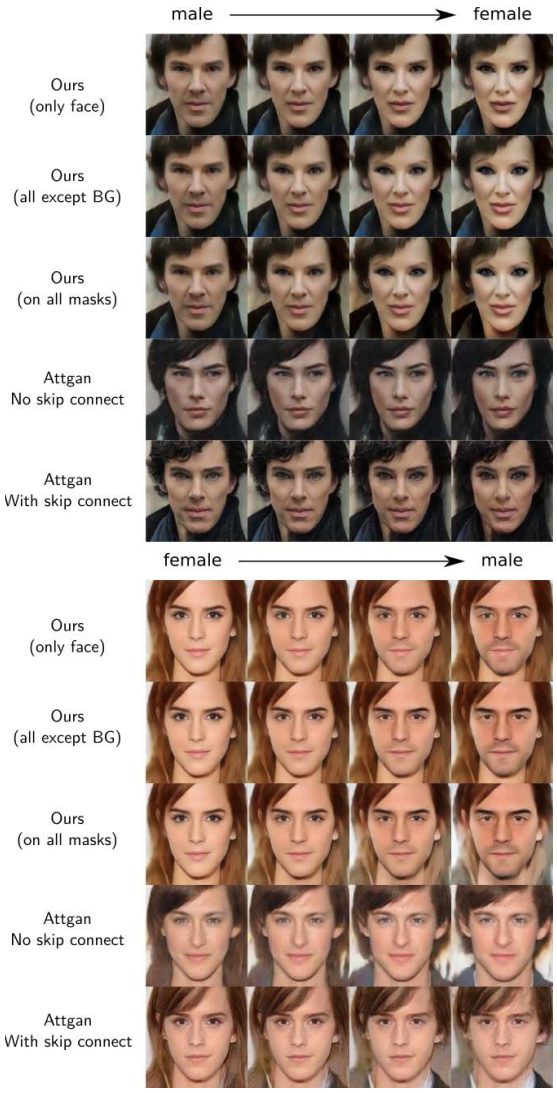

Figure 9. Comparison between our method and AttGAN [13] for transforming an image to the opposite gender. Our method makes it possible to only change parts of the image, while AttGAN always changes the image as a whole.

the skip connections results in a loss of identity in the image for AttGAN, creating a person that looks very different than the original. Most importantly, however, AttGAN always changes the entire image, especially the hair and background changes completely. With our method, we gain full control which part of the image is to be changed and which part is not, making it possible to keep desired features.

4. Summary

In this paper, we propose a novel method of generating disentangled representations of images with autoencoders. This includes a network architecture that allows us to create an independent subspace decomposition inspired by independent subspace analysis (ISA) and two losses, a mask loss and an entropy loss. Training a convolutional resnet autoencoder with frontal face images and their semantic segmentations allows to change each of the face regions background+hair, face, eyebrows, eyes and mouth without affecting the other regions. This enables us to swap face parts between unseen images without the need of a semantic segmentation. In the experiments, we verified the independency of the generated subspaces visually, but also by comparing the cluster center distances. Additionally, a comparison between results from our network and AttGAN [13] is shown with regards to changing an attribute of the face inside one or more regions.

In future work we will further improve the method and apply it to other problems. The method itself is not bound to faces or images at all. We also want to make NNs more interpretable and show the importance of such interpretations by working on both the forced disentanglement presented in this paper and also unsupervised disentanglement as shown with Structuring Autoencoders [28]. We are also interested in making our method non-deterministic, similar to a Markov Chain Neural Network [1] and we want to adapt our method to other data sets in 3D [33]. We further expect that problems, such as vanishing point estimation [21] or semantic image understanding [35] can benefit from our approach.

Acknowledgments

The work is inspired by BIAS ("Bias and Discrimination in Big Data and Algorithmic Processing. Philosophical Assessments, Legal Dimensions, and Technical Solutions"), a project funded by the Volkswagen Foundation within the initiative "AI and the Society of the Future" for which the last author is a Principal Investigator.

References

[1] M. Awiszus and B. Rosenhahn. Markov chain neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2180–2187, 2018. 8

[2] V. Blanz, T. Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999. 2

[3] S. Brandt, H. Ackermann, and S. Graßhof. Uncalibrated nonrigid factorisation by independent subspace analysis. In *The IEEE International Conference on Computer Vision Workshops*, 2019. 2

[4] A. Brunton, T. Bolkart, and S. Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014. 2

[5] P. P. Busto, C. Eisenacher, S. Lefebvre, M. Stamminger, et al. Instant texture synthesis by numbers. In *VMV*, pages 81–85, 2010. 3

[6] A. J. Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*, 2016. 3

[7] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1511–1520, 2017. 3

[8] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016. 2, 3, 4

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2

[10] S. Graßhof, H. Ackermann, S. Brandt, and J. Ostermann. Apathy is the root of all expressions. *12th IEEE Conference on Automatic Face and Gesture Recognition*, 2017. 2

[11] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, and L. Yuan. Mask-guided portrait editing with conditional gans. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 2, 3

[12] J. Hays and A. A. Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics*, 26(3):4, 2007. 3

[13] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want. *arXiv preprint arXiv:1711.10678*, 2017. 2, 4, 7, 8

[14] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 1

[15] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004. 2

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[17] Y. Jo and J. Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. *arXiv preprint arXiv:1902.06838*, 2019. 3

[18] I. Jolliffe. *Principal component analysis*. Springer, 2011. 2

[19] D. Joo, D. Kim, and J. Kim. Generating a fusion image: One's identity and another's shape. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 2

[20] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[21] F. Kluger, H. Ackermann, M. Y. Yang, and B. Rosenhahn. Deep learning for vanishing point detection using an inverse gnomonic projection. In *German Conference on Pattern Recognition*, pages 17–28. Springer, 2017. 8

[22] J.-F. Lalonde, D. Hoiem, A. A. Efros, C. Rother, J. Winn, and A. Criminisi. Photo clip art. In *ACM transactions on graphics*, volume 26, page 3. ACM, 2007. 3

[23] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012. 5

[24] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, Dec. 2015. 1, 5

[25] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pages 41–48. Ieee, 1999. 2

[26] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi. The feret evaluation methodology for face-recognition algorithms. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 137–143. IEEE, 1997. 5

[27] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5):295–306, 1998. 5

[28] M. Rudolph, B. Wandt, and B. Rosenhahn. Structuring autoencoders. In *The IEEE International Conference on Computer Vision Workshops*, 2019. 8

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[30] B. M. Smith, L. Zhang, J. Brandt, Z. Lin, and J. Yang. Exemplar-based face parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3484–3491, 2013. 5

[31] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183–1, 2015. 2

[32] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. In *ACM transactions on graphics*, volume 24, pages 426–433. ACM, 2005. 2

[33] B. Wandt and B. Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Computer Vision and Pattern Recognition*, June 2019. 8

[34] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 3

[35] M. Y. Yang, W. Liao, H. Ackermann, and B. Rosenhahn. On support relations and semantic scene graphs. *ISPRS journal of photogrammetry and remote sensing*, 131:15–25, 2017. 8

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision*, Oct 2017. 2, 3, 4