Facial Appearance Modifications using SKPCA-Derived Features Extracted from Convolutional Autoencoder's Latent Space

Krzysztof Adamiak
Institute of Applied Computer Science
Lodz University of Technology
Lodz, Poland
krzysztof.adam.adamiak@gmail.com

Pawel Kapusta
Institute of Applied Computer Science
Lodz University of Technology
Lodz, Poland
pawel.kapusta@p.lodz.pl

Krzysztof Ślot
Institute of Applied Computer Science
Lodz University of Technology
Lodz, Poland
krzysztof.slot@p.lodz.pl

***Abstract*—The presented paper proposes a method that enables image object appearance editing by modifying its appearance-related high-level attributes. First, attribute-related features get extracted from a latent representation of an image generator and then, their contents gets modified, which results in producing the assumed appearance alterations. Convolutional Autoencoder (CAE) has been adopted as an image manipulation framework and face appearance, characterized by four attributes: age, smile intensity, facial hair intensity and gender, was chosen for modifications. To extract attribute-related features from CAE's latent representation, Supervised Kernel Principal Component Analysis (SKPCA) was used, as this transformation is able to disentangle complex, nonlinear image-to-attribute relationships. The method has been evaluated using large-scale face dataset CelebA. Qualitative results show that realistically-looking appearance modifications can be obtained. To quantify plausibility of introduced modifications, face recognition experiments on altered face images were performed, delivering on average 95% classification accuracy, for twenty-six category dataset.**

***Index Terms*—Autoencoders, kernel methods, image editing**

I. Introduction

Deep neural networks (DNNs) boosted performance of intelligent data analysis and enabled successful implementation of AI algorithms in a wide range of real-world applications. They proved superior not only with respect to other information processing approaches but also with respect to humans in handling several complex tasks, including visual object recognition [1], speech recognition [2], image and video understanding [3], machine translation [4], planning [5] or document analysis [6]. Another domain where DNNs recently proved excellence is content generation. They have been shown to be able to create paintings with a learnable artistic style [7], to synthesize realistic speech [8] or to generate or edit visual image objects [9] [10] [11]. The three concepts that enabled impressive performance in this field are Autoencoders - AE [12], Convolutional Autoencoders (CAE) [13] and their variational extensions, Normalizing Flows [14] [15], as well as Generative Adversarial Networks - GANs [16] [17], which, at the moment, are unbeatable in realistic image synthesis. Despite impressive performance, various approaches adopted for image object editing give little insight and provide little control over this process.

A motivation for the presented research was an attempt to develop visual object editing framework that enables functional control over transformations involved in appearance alterations, where appearance is described using high-level attributes, such as facial expression intensity or age. Such a functional control would enable purposeful and computationally efficient appearance transformations to meet a desired outcome. A core of the proposed idea is to target information extracted in latent image representations, derived by either CAEs or GANs, by means of the proposed, appropriately developed *Attribute Transformation Module* (ATM). Research reported in the presented paper employs Convolutional Autoencoders, trained to reconstruct facial images, as an object appearance editing environment. We propose to extract from latent image representation features, which strongly correlate with considered high-level visual appearance attributes. We adopt Supervised Kernel Principal Component Analysis (SKPCA) [18] as a nonlinear transformation to obtain this objective and perform face appearance modifications in the derived space (this extends our earlier approach, where linear feature extraction was used [19]). Resulting, modified face image representations are projected back to the latent representation to be reconstructed by CAE's decoding module.

A structure of the paper is the following. First, we provide a brief review of related work on realistic image generation and editing. Next, we explain the proposed appearance modification method. Finally, we provide results of experimental evaluation of the procedure (its Python implementation is available at [20]). We show, using examples from a large scale, CelebA face database [21], that the introduced approach enables functional control over appearance modifications. In addition to qualitative assessment, we also quantify the outcome of the procedure by evaluating face classification accuracy on transformed images.

This research was funded by the National Centre for Research and Development under the grant CYBERSECIDENT/382354/II/NCBR/2018.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

II. Related work

Although deep generative models, such as Autoencoders, Normalizing Flows or Generative Adversarial Networks, have been introduced only recently, an impressive amount of successful research on complex visual object generation, mainly attributed to the latter paradigm, has been presented thus far [22] [23] [24] [25] [26]. As adversarial networks are known to be difficult to train, nowadays much of the work is concentrated on improving this process, for example, by mitigating modal collapse [27], stabilizing discriminator training by means of spectral normalization [28], revising the basic architecture and training using large scale datasets [29] or by progressively growing both generator and discriminator [30].

A basic ability to generate realistically-looking visual objects of some specified category was quickly expanded to provide more detailed control over objects' properties. A notable development was the introduction of a conditional GAN (cGAN) concept [31], in which a mechanism for conditioning generator outcome on additional, appearance-related information was proposed. This concept has also quickly quickly evolved. In [32] cGAN-based architecture appended with an additional encoder module for retrieving attribute-related information from input images was proposed. After assembling this information with with latent representation, remarkable appearance modification effects, such as changing hair color, facial expressions or addition of eyeglasses, were reported. Another extension is proposed in [33], where both image generation and image editing is addressed by introducing a 'connection network', which provides a trainable mapping between attributes and the corresponding image space. The resulting framework enables continuous modifications of attribute-expression intensity.

A concept of object generation and editing within a framework of Variational Autoencoders proposed in [34] models an image as a composite of foreground and background with disentangled latent variables, using two encoder-decoder pairs. The former pair is learned using a criterion involving additional attribute information, which is extracted from a description provided in a form of a natural language. The indicated idea of exploiting natural language description for visual object generation through conditioning image formation processes was utilized in [35] [36] [37]. An extension to this concept that provides also capability to modify existing scenes using continuous linguistic feedback, has been presented in [38].

A specific task of generating and modifying realistically looking faces is another challenging, yet highly researched topic. Face image synthesis methodology termed Semi-Latent Facial Attribute Space (SL-FAS), which combines user defined and latent spaces, has been formulated in [39]. Other approaches are based, for example on applying attribute classification constraints [40] or by introducing semantic label masks into a training process in order to achieve interactive face editing and manipulation [41]. Recently, an approach inspired by neural style transfer literature has been proposed, which modifies generator architecture to gain fine-grained control over image synthesis process [7] [42]. Image manipulations are performed at a level of convolutional layers, using a 'latent code', which combined with noise injection leads to very good separation of high-level attributes (such as pose or gender) from small variations (skin texture, hair color etc.) [43] [44].

The last approach to content generation, the Normalizing Flows concept, seeks for a chain of invertible transformations that enable mappings between a complex distribution of training samples (e.g. face images) and a simple one (e.g. Gaussian) that can be viewed as some latent data representation. A basis for flow-generation is provided by the change of variables theorem. To ensure computational efficiency for deriving the transformations, consecutive bijections are characterized by a triangular form of the corresponding Jacobian matrix, which makes its determinant calculation trivial. As a consequence, flow transformations can be seen as a sequence of autoregressive models, where argument vector permutations at different stages of the processing chain enable discovery of arbitrary feature dependencies. Simplicity of a distribution assumed as a latent representation of target data, combined with invertibility of learned transformations, makes normalizing flows an excellent tool for manipulating high-level contents of data to be processed. Remarkable results of various normalizing flow-based concepts were reported both in image contents generation (e.g. using GLOW algorithm [45] or autoregressive Pixel-RNN [46]) or audio contents generation [8].

The approach proposed in the paper can be seen as a combination of Autoencoder-based and normalized flows-based processing. The adopted SKPCA transformation can be considered as a single-step 'encoding' of a complex distribution of Autoencoder's latent space variables onto manageable distributions of some underlying priors, that model different high-level visual attributes. In contrast to Autoencoder and GAN-based content generation, we propose a 'serial' processing scheme, in which attribute-related content is sought in a derived, intermediate representation. Moreover, the proposed approach is aimed at editing visual characteristics of a specific input object. On the other hand, unlike it occurs for normalizing flows, we propose only a single transformation that converts some complex distribution to a one that is simple and easily-manageable.

III. Proposed algorithm

Computational architecture of the presented algorithm has been shown in Fig. 1. The proposed Attribute Transformation Module extracts samples ($\mathbf{z}$) that contain disentangled information on considered face appearance attributes from latent Autoencoder's representations ($\mathbf{y}$) of input images ($\mathbf{x}$). Then, it modifies their contents, producing vectors $\tilde{\mathbf{z}}$, and reassembles latent representations ($\tilde{\mathbf{y}}$) that are to be finally decoded onto output images ($\tilde{\mathbf{x}}$).

A space of $\mathbf{z}$ vectors. referred to as *attribute-space*, should provide decorrelated, monotonous (possibly linear) representations of attribute expression-level intensities. To meet these objectives, transformation that maximizes correlations between

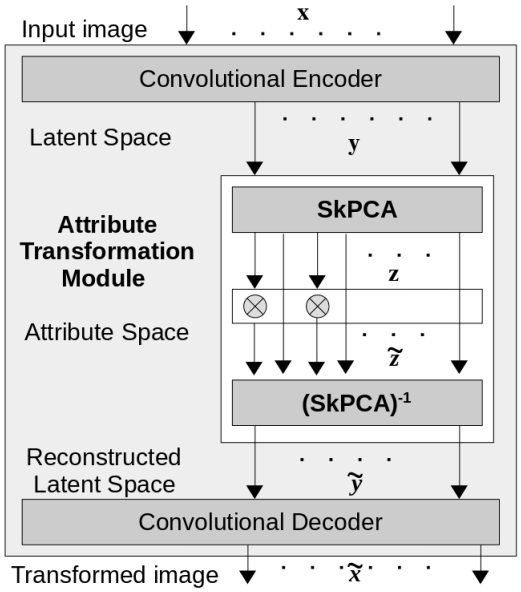

Fig. 1. Computational architecture of the algorithm.

vectors $\mathbf{z}$ and image label vectors $\mathbf{l}$ (comprising interval-type appearance attribute descriptors) needs to be found. A possible candidate for this transformation is the aforementioned SKPCA, which is built on maximizing Hilbert-Schmidt independence criterion [47]. Denoting cross-covariance between vectors $\mathbf{y}$ and image label vectors $\mathbf{l}$ by $\mathbf{C}_{y,l}$:

$$\mathbf{C}_{y,l} = E(\mathbf{y} - \boldsymbol{\mu}_y)(\mathbf{l} - \boldsymbol{\mu}_l)^T, \tag{1}$$

where $\boldsymbol{\mu}_y, \boldsymbol{\mu}_l$ denote corresponding means, supervised dimensionality reduction seeks such unit-length projection bases: $\mathbf{u}_1 ... \mathbf{u}_k$ that maximize dependence between projections of $\mathbf{y}$ samples ($\mathbf{z} = [\mathbf{u}_1 ... \mathbf{u}_k]^T \mathbf{y}$) and labels $\mathbf{l}$, which can be expressed as:

$$\mathbf{u} = \arg\max_{\mathbf{v}} \left(\mathbf{v}^T tr(\mathbf{C}_{y,l} \mathbf{C}_{y,l}^T) \mathbf{v} \right), \quad ||\mathbf{v}|| = 1, \tag{2}$$

where $tr(.)$ denotes a trace. As linear transformations involved in (2) are unlikely to provide decorrelation of probably highly nonlinear visual attribute encodings that exist in latent representation, the kernel-based SKPCA can be used to solve the problem. SKPCA captures the required relationships in some implicit high-dimensional space, where both samples $\mathbf{y}$ and labels $\mathbf{l}$ get nonlinearly transformed ($\hat{\mathbf{y}} \leftarrow \phi(\mathbf{y})$ and $\hat{\mathbf{l}} \leftarrow \psi(\mathbf{l})$). Due to a kernel trick, all necessary computations can be done in original spaces, by introducing kernels that evaluate similarity both on samples: $\mathbf{K} = [k_y(\mathbf{y}_i, \mathbf{y}_j)]$ and on labels: $\mathbf{L} = [k_l(\mathbf{l}_i, \mathbf{l}_j)]$, where $k_y(x, y) = \langle \phi(x), \phi(y) \rangle$ and $k_l(x, y) = \langle \psi(x), \psi(y) \rangle$.

Attribute space derived by SKPCA is made up of mutually uncorrelated features that are expected to correspond to individual appearance attributes. Each individual component of a vector $\mathbf{z}' = F(\mathbf{y}', \mathbf{K}, \mathbf{L})$, which is to be modified, reflects expression-level intensity of a specific appearance attribute of input image $\mathbf{x}'$. It follows that appearance modifications can be performed selectively for each attribute, by altering appropriate entries of attribute-space vectors. Moreover, expected functional relation between a specific feature value and the corresponding attribute expression intensity should be linear.

Once attribute modifications have been introduced, the resulting attribute-space vector needs to be converted back to a latent space, so that it can be correctly decoded onto the output image. As SKPCA transformation is highly nonlinear, we pose the inversion problem in terms of mean-squared error minimization: for all latent vectors from a training set we seek a matrix $\mathbf{M}$ that attempts to reconstruct original latent vectors from unaltered attribute space samples:

$$\mathbf{M} = \arg\min_{\mathbf{M}'} E \left((\mathbf{M}'\mathbf{z} - \mathbf{y})^T (\mathbf{M}'\mathbf{z} - \mathbf{y}) \right), \tag{3}$$

where $\mathbf{z} = F(\mathbf{y}, \mathbf{K}, \mathbf{L})$.

The proposed image modification procedure can be summarized as a sequence of the following operations (Fig. 1). First, input face image $\mathbf{x}$ is transformed to its latent representation $\mathbf{y}$ in CAE's encoding module. Next, SKPCA transformation is performed, resulting in an attribute-space vector $\mathbf{z}$. This vector is subject to alterations that produce its modified version $\tilde{\mathbf{z}}$, which is subsequently projected onto the reconstructed latent space using the linear transformation involving a matrix $\mathbf{M}$. The result - reconstructed latent vector $\hat{\mathbf{y}}$ is finally decoded to CAE's output $\hat{\mathbf{x}}$ in Autoencoder's decoding module.

The algorithm involves two training procedures. Firstly, CAE needs to be derived, using appropriately large set of unlabeled face images, to obtain an appropriate latent representation of input visual information. Then, attribute transformation module: SKPCA transformation and its approximate inversion need to be derived, using a set of labeled facial images.

IV. Experimental evaluation

The proposed method has been evaluated using annotated face images from CelebA database, which comprises over 200 000 images from over 10 000 classes. Since original face image annotations are binary variables (presence/absence of some visual attribute), they were inappropriate from the point of view of the presented research. Therefore, a subset of 50 thousand images was additionally labeled with ordinal features (such as age, facial hair intensity or smile intensity) using Microsoft Cognitive Services Vision API [48].

The Convolutional Autoencoder, trained on all CelebA images, of size 208x176x3, was built based on a basic architecture available at [49]. It comprised eleven convolutional, six dropout, and four max-pooling layers at the encoder, and eleven transposed-convolution layers at the decoder module. CAE's latent representation was composed of 4096 elements (see Table I). The Autoencoder was trained using Adam optimizer ($\beta 1$ = 0.9, $\beta 2$ = 0.999) with the learning rate of 1×10^{-4} and a batch size of 16. The Attribute Transformation Module was trained on sixteen thousand annotated samples, selected to provide balanced representation among all attribute expression levels for the four considered appearance attributes: gender, age, facial hair intensity and smile intensity. To derive

the attribute-space, Radial Basis Function (RBF) kernel was used for assessing latent sample similarities ($\mathbf{K}$), with its parameter γ set to 0.25, whereas image labels were processed using a linear kernel ($\mathbf{L}$).

Although a typical machine learning approach requires that training and test datasets are disjoint, for the considered problem, processing of samples that were present in a training dataset is well-justified. Therefore, throughout experiments we focused on a scenario, where a sample to be transformed was known during training of both CAE and ATM.

TABLE I
DETAILS OF CAES USED IN FACE APPEARANCE MODIFICATIONS.

Encoder	**Decoder**
Conv2D (256, 6, 1), ReLU	Dense (9152)
GaussianDropout (0.3)	Reshape (target = 13, 11, 64)
Conv2D (256, 6, 1), ReLU	DeConv2D (128, 2, 1), ReLU
GaussianDropout (0.3)	DeConv2D (128, 2, 2), ReLU
MaxPooling2D	DeConv2D (64, 3, 1), ReLU
Conv2D (128, 5, 1), ReLU	DeConv2D (64, 3, 2), ReLU
GaussianDropout (0.3)	DeConv2D (64, 4, 1), ReLU
Conv2D (128, 5, 1), ReLU	DeConv2D (64, 4, 2), ReLU
GaussianDropout (0.3)	DeConv2D (64, 3, 1), ReLU
MaxPooling2D	DeConv2D (64, 3, 2), ReLU
Conv2D (128, 4, 1), ReLU	DeConv2D (64, 4, 1), ReLU
GaussianDropout (0.3)	DeConv2D (64, 2, 1), ReLU
Conv2D (128, 4, 1), ReLU	DeConv2D (3, 3, 1), ReLU
GaussianDropout (0.3)	
MaxPooling2D	
Conv2D (128, 3, 1), ReLU	
GaussianDropout (0.3)	
Conv2D (128, 3, 1), ReLU	
GaussianDropout (0.3)	
MaxPooling2D	
Conv2D (128, 2, 1), ReLU	
GaussianDropout (0.3)	
Conv2D (128, 2, 1), ReLU	
GaussianDropout (0.3)	
Flatten (9152)	
Dense (4096)	

* Conv2D(d,k,s) and DeConv2D(d,k,s) denote the 2D convolutional layer and 2D transposed convolutional layer, d is a dimension, k - a kernel size and s is a stride

A. Attribute mappings in the SKPCA feature space

Each feature produced by SKPCA corresponds to some sub-manifold in CAE's 4096-dimensional space that orders data samples in a way that provides maximum correlation with image labels. To assess, whether such attribute intensity-ordering maps have actually been discovered, relationships that exist between image labels and the corresponding attribute-space features have been examined. As it can be seen in Fig.2, the resulting relations are almost perfectly linear for all considered interval attributes. This proves good performance of SKPCA in extraction of attribute-related information from a latent space and gives a promising basis for subsequent appearance manipulations.

Fig.3 indicates that attribute-space features are indeed uncorrelated. The presented 2D plots show projections of images labeled with pairs of attributes (age plus facial hair and age plus smile intensity) onto the corresponding two attribute-space features. Brightness of each dot is proportional to a value of 'Age' image label (for the left column) and to a value of the second image label (right column). To identify, which attribute-space feature corresponds to which object's attribute, we evaluate correlation coefficients between attribute descriptors and the corresponding feature values (for the considered task, simple Pearson's correlation was sufficient to get robust results).

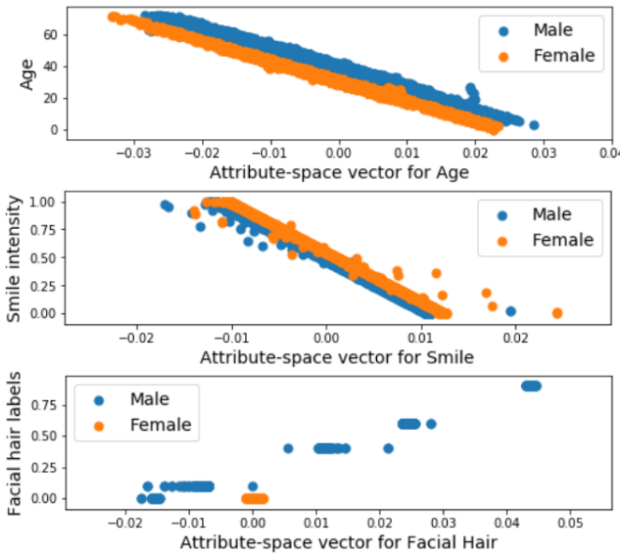

Fig. 2. Image projections onto attribute-space features for (from top to bottom): age, smile intensity and facial hair, shown separately for female and male faces.

B. Qualitative assessment of appearance modifications

To qualitatively assess effects of appearance modifications performed in the attribute-space, the following procedure has been adopted. Given a trained CAE together with the trained ATM, a test sample is fed to CAE's input and its attribute expression levels are transformed by modifying either a single feature or selected feature combinations.

Results of face appearance modifications are summarized in Fig.4 through Fig.6. Modifications of expression-level for a single attribute are presented in Fig.4, whereas results of alterations introduced simultaneously to a couple of attributes have been shown in Fig. 5. In the latter case, both components of attribute-space vectors were updated by the same amount. As it can be seen from Fig.4, realistically looking appearance alterations can be obtained. Also, inducing a combined change in two attributes results in plausible facial appearance.

An interesting consequence of the adopted appearance modification strategy is an opportunity to examine impact of interpolating binary attributes. Fig.6 shows face appearance transitions induced by changes in a value of the 'Gender' attribute, between the two extreme values.

The presented experiments proved that SKPCA is a promising method for extracting attribute-related information from Autoencoder's latent space. Also, attribute-space features

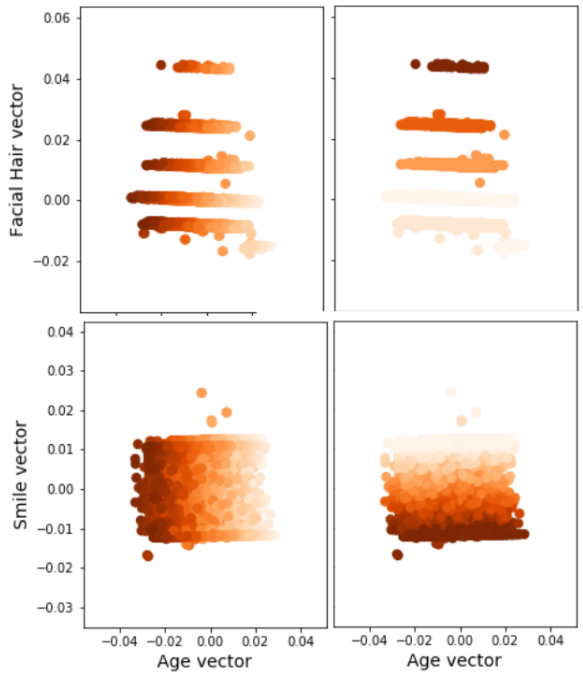

Fig. 3. Image projections onto 2D attribute space features: age vs. facial hair (top), age vs. smile intensity (bottom row). Five distinct values of facial hair intensity were present in a dataset.

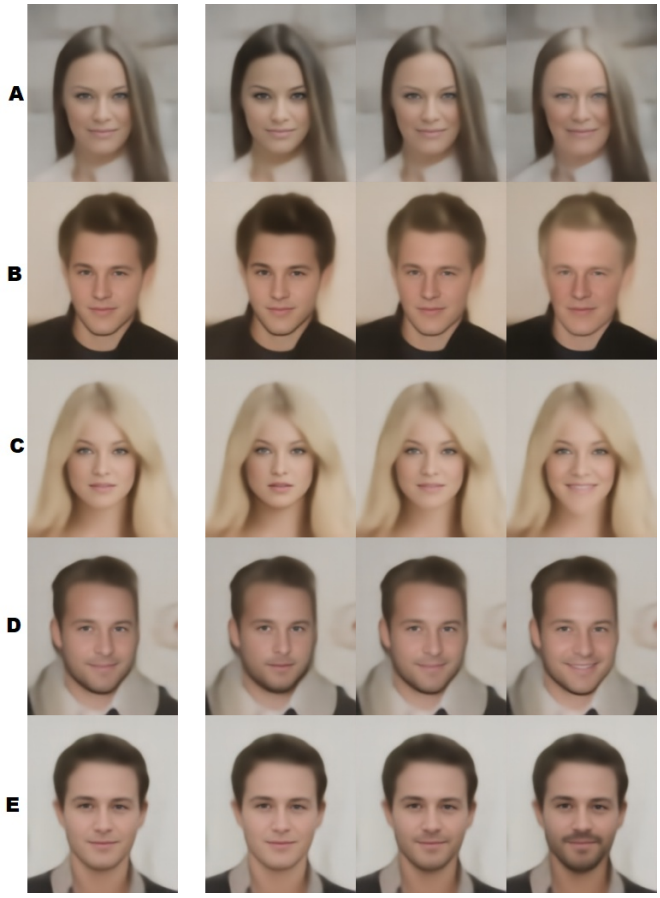

Fig. 4. Results of attribute expression-intensity modifications. Original samples are shown in the leftmost column, subsequent columns show face appearance for the minimum, medium and the maximum values of the corresponding attribute-space feature. The considered attributes are age (rows (A) and (B)), smile (rows (C) and (D)) and facial hair (E).

decorrelate information from different attributes, enabling selective appearance modifications. However, we found several issues that need to be addressed to enable photo-realistic image editing. The first one is image quality deterioration introduced by CAE. There are a few possible ways to improve image re-synthesis fidelity. A possible approach is to significantly reduce a number of classes involved in CAE training, to enable better learning of individual appearance, however this would require large amounts of training images per each of considered classes. As this is not the case for the selected dataset, we propose a two-step CAE training procedure: the first one involves all examples, whereas the second one fine-tunes the pretrained CAE to better learn only a subset of classes. The strategy proved to improve face reconstruction fidelity: sample results obtained for CAE tuned on 25 selected classes, are shown in Fig. 7.

Another limitation of photo-realistic image editing capabilities of the proposed approach is difficulty with generalization of concepts learned by CAE and ATM for previously unseen categories. Finally, the adopted method for SKPCA inversion is valid only for samples that do not deviate much from samples used for training.

C. Quantitative assessment of method's performance

To evaluate whether face appearance modifications, introduced by manipulations performed in attribute-space, are plausible a set of face recognition experiments was performed. The classifier, based on VGGFace architecture implemented in Keras [50] for both VGG16 and ResNET backends, was initially trained on VGGFace2 dataset [51] and used for the

Fig. 5. Face appearance changes for simultaneous modifications of two attributes: age and smile (top row), age and facial hair (middle row) and smile and facial hair (bottom row). Original images are shown in the leftmost column, modifications from the minimum, through medium to the maximum attribute expression-level are given in subsequent columns.

Fig. 6. Appearance modifications induced by gradual changes in the gender-related attribute-space feature for original female face (top) and male faces (bottom row).

Fig. 7. Original faces (top) and their reconstructions generated by CAE trained using the two-step procedure (with the tuning phase - middle) and using one-step procedure (no tuning - bottom).

task realization under a framework of transfer learning. Its last three fully-connected layers were trained to recognize a subset of twenty-six CelebA face categories, that were passed through the Autoencoder without any modification. An objective of the experiments was to assess recognition accuracy for faces with appearance altered using the presented method, as a function of attribute expression intensity levels. The classification procedure was repeated ten times for random training-test set splits. Results are summarized using mean recognition accuracy (fraction of correctly classified samples) and standard deviation, separately for six scenarios, involving alterations of different attribute-space features and their combinations. In each case, five levels of attribute expression intensity were considered (i.e. each feature of an attribute space sample was assigned with either of five evenly-spaced values, ranging from the minimum one - '0', to the maximum '1'). Experiment results, provided in Table II prove plausibility of face appearance modifications. As majority of original training images featured moderate attribute expression-levels, the best results can be found for mid-values of attribute-space features. However, the differences are minor, with an exception for age, which can be attributed to relatively few samples labeled with extreme values.

TABLE II
FACE CLASSIFICATION RESULTS CONDITIONED ON APPEARANCE ATTRIBUTE-EXPRESSION INTENSITY.

Attribute	Intensity	VGG16 μ	VGG16 σ	ResNet μ	ResNet σ
Not modified	-	0.952	0.012	0.920	0.014
age	0	0.896	0.030	0.795	0.029
	0.25	0.956	0.013	0.926	0.022
	0.5	0.964	0.011	0.961	0.010
	0.75	0.967	0.010	0.942	0.014
	1	0.915	0.019	0.902	0.018
smile	0	0.971	0.015	0.962	0.013
	0.25	0.970	0.009	0.967	0.014
	0.5	0.970	0.007	0.973	0.011
	0.75	0.958	0.012	0.962	0.010
	1	0.963	0.013	0.950	0.012
facial hair	0	0.964	0.012	0.964	0.013
	0.25	0.971	0.008	0.963	0.008
	0.5	0.962	0.008	0.967	0.008
	0.75	0.952	0.019	0.968	0.008
	1	0.939	0.016	0.939	0.017
age and smile	0	0.875	0.041	0.772	0.026
	0.25	0.960	0.013	0.905	0.018
	0.5	0.965	0.013	0.940	0.016
	0.75	0.947	0.012	0.924	0.019
	1	0.872	0.032	0.840	0.028
age and facial hair	0	0.872	0.032	0.748	0.028
	0.25	0.951	0.017	0.896	0.026
	0.5	0.963	0.012	0.962	0.014
	0.75	0.908	0.017	0.906	0.019
	1	0.788	0.031	0.786	0.044
smile and facial hair	0	0.959	0.013	0.953	0.012
	0.25	0.967	0.010	0.967	0.012
	0.5	0.968	0.012	0.967	0.015
	0.75	0.946	0.011	0.957	0.017
	1	0.909	0.031	0.910	0.034

μ - average classification accuracy, σ - standard deviation

V. CONCLUSION

The presented concept for image object appearance modifications proved promising when tested on facial images. However, as the core of the method is a simple concept of extraction and alteration of specific content-related information, we believe that it can be considered as a more general framework, applicable to several other data processing contexts. Firstly, the method should be examined in altering appearance of visual objects other than faces, using other high-level appearance attributes. Secondly, the introduced Appearance Transformation Module could be considered as a component of other content-generating architectures, such as e.g. Generative Adversarial Networks. Finally, the concept could be used for introducing modifications to data that does not have to be structured as images, for example, it could be adapted to signal modification.

One of important features of the proposed method is a possibility of inducing real-time appearance changes in a

processing loop, where a required modification can be induced by feedback information. This can be an important property enabling improvements in operation of advanced human-computer interfaces. However, it can also be exploited in a malicious way, for example, by facilitating presentation attacks on biometric systems.

REFERENCES

[1] O. Russakovsky et al., ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115, 2015 ,pp. 1–42.

[2] A. Gupta and A. Joshi, "Speech Recognition Using Artificial Neural Network," 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2018, pp. 0068-0071. doi: 10.1109/ICCSP.2018.8524333

[3] A. Karpathy, F.-F. Li, Deep visual-semantic alignments for generating image descriptions., in: CVPR, IEEE Computer Society, 2015, pp. 3128–3137. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#KarpathyL15

[4] Y. Papanikolaou, I. Roberts, A. Pierleoni, Deep bidirectional transformers for relation extraction without supervision, 2019, pp. 67–75. doi:10.18653/v1/D19-6108.

[5] D.Silver et al. , Mastering the game of Go with deep neural networks and tree search , Nature 529(7587), 2016, pp. 484–489. doi:10.1038/nature16961.

[6] Ares Oliveira, S., Seguin, B., Kaplan, F., dhSegment: A Generic Deep-Learning Approach for Document Segmentation2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018

[7] L. Gatys, A. Ecker, M. Bethge, A neural algorithm of artistic style, Journal of Vision 16 (12), 2016 doi:10.1167/16.12.326. URL http://dx.doi.org/10.1167/16.12.326

[8] A. v. d. Oord, et al., Wavenet: A generative model for raw audio, 2016. URL http://arxiv.org/abs/1609.03499

[9] Bodnar, C., Text to Image Synthesis Using Generative Adversarial Networks., 2018, arXiv:1805.00676

[10] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H., Generative adversarial text to image synthesis, 2016, arXiv:1605.05396

[11] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks,2015, arXiv:1511.06434.

[12] D. E. Rumelhart, J. L. McClelland. Learning Internal Representations by Error Propagation. MITP, 1987

[13] Xiao-Jiao Mao, Chunhua Shen, Yu-Bin Yang., Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections, 2016

[14] D. P. Kingma, T. Salimans, M. Welling. Improving Variational Inference with Inverse Autoregressive Flow. CoRR, vol.1606.04934, 2016, http://arxiv.org/abs/1606.04934

[15] L. Dinh, D. Krueger, Y. Bengio. NICE: Non-linear Independent Components Estimation. Proc. of 3rd International Conference on Learning Representations, ICLR 2015, http://arxiv.org/abs/1410.8516

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014, arXiv:1406.266

[17] He Huang, Yu, P S. and Wang, C., An Introduction to Image Synthesis with Generative Adversarial Nets., 2018, arXiv:1803.04469

[18] Barshan, E., Ghodsi, A., Azimifar, Z. and Zolghadri Jahromi, M., Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. Pattern Recogn., 44(7), pp. 1357–1371, 2011

[19] Ślot K., Kapusta P., Kucharski J., Autoencoder-based image processing framework for object appearance modifications, unpublished.

[20] Adamiak., K, Kapusta., P., Ślot K., Article source code, URL https://github.com/LordIllidan/FacialAppearanceModificationsUsingSKPCA, doi:https://doi.org/10.5281/zenodo.3822505

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

[22] Zhang, H. et al., Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017, arXiv:1612.03242

[23] Xu, T. et al., AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks, 2017, arXiv:1711.10485

[24] Brock, A., Lim, T., Ritchie, J. M. and Weston, N., Neural photo editing with introspective adversarial networks, 2016, arXiv:1609.07093

[25] Gorijala, M. and Dukkipati, A., Image generation and editing with variational info generative adversarial networks, 2017, arXiv:1701.04568

[26] Pieters, M., and Wiering, M., Comparing Generative Adversarial Network Techniques for Image Creation and Modification, 2018, arXiv:1803.09093

[27] D. Berthelot, T. Schumm, L. Metz, Began: Boundary equilibrium generative adversarial networks, 2017, arXiv:1703.10717.

[28] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, 2018, arXiv:1802.05957.

[29] Brock, A., Donahue, J. and Simonyan, K., Large scale GAN training for high fidelity natural image synthesis, 2018, arXiv:1809.11096

[30] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, 2017, arXiv:1710.10196.

[31] Mirza, M., and Osindero, S. (2014). Conditional generative adversarial nets. CoRR, abs/1411.1784

[32] Perarnau, G., van de Weijer, J., Raducanu, B., and Álvarez, J. M., Invertible conditional gans for image editing. CoRR, 2016, abs/1611.06355

[33] Baek, K., Bang, D, and Shim, H., Editable generative adversarial networks: Generating and editing faces simultaneously. CoRR, 2018, abs/1807.07700

[34] Yan, X., Yang, J., Sohn, K., and Lee, H., Attribute2Image: Conditional Image Generation from Visual Attributes. In European Conference on Computer Vision, 2016

[35] B. Li, X. Qi, T. Lukasiewicz, P. H. S. Torr, Controllable text-to-image generation, 2019, arXiv:1909.07083.

[36] H. Zhang, et al., Stackgan++: Realistic image synthesis with stacked generative adversarial networks, 2017, arXiv:1710.10916.

[37] T.Xu, et al., Attngan: Fine-grained text to image generation with attentional generative adversarial networks, pp. 1316–1324, 2018, doi:10.1109/CVPR.2018.00143.

[38] A. El-Nouby, et al., Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction, 2018, arXiv:1811.09845.

[39] W. Yin, Y. Fu, L. Sigal, X. Xue, Semi-latent gan: Learning to generate and modify facial images from attributes, 2017, arXiv:1704.02166.

[40] Z.He,W.Zuo,M.Kan,S.Shan,X.Chen,Attgan: Facial attribute editing by only changing what you want, 2017, arXiv:1711.10678.

[41] C.-H. Lee, Z. Liu, L. Wu, P. Luo, Maskgan: Towards diverse and interactive facial image manipulation, 2019, arXiv:1907.11922.

[42] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, 2017, arXiv:1703.06868.

[43] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, 2018, arXiv:1812.04948.

[44] T.Karras, S.Laine, M.Aittala, J.Hellsten, J.Lehtinen, T.Aila, Analyzing and improving the image quality of stylegan, 2019, arXiv:1912.04958.

[45] D.P. Kingma, P. Dhariwal. Glow: Generative Flow with Invertible 1×1 Convolutions. Proc. of the 32nd International Conference on Neural Information Processing Systems, 10236–10245, 2018.

[46] A. Van Den Oord, N. Kalchbrenner, K. Kavukcuoglu. Pixel Recurrent Neural Networks. Proc. of the 33rd International Conference on International Conference on Machine Learning, vol. 48, 1747–1756, 2016.

[47] N. Aronszajn, Theory of reproducing kernels. Transactions of the American Mathematical Society, 68 (3), pp. 337-4040, 1950

[48] Microsoft, Microsoft azure cognitive services, URL https://azure.microsoft.com/en-us/services/cognitive-services/face/, accessed on 2019-12-12 (Dec 2019).

[49] M. Abadi et al., TensorFlow: Large-scale machine learning on heterogeneous systems, URL https://www.tensorflow.org/, accessed on 2019-12-12 (Dec 2019).

[50] R. C. Malli, Vggface implementation with keras framework, URL https://github.com/rcmalli/keras-vggface, accessed on 2019-12-12 (Dec 2019).

[51] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in: International Conference on Automatic Face and Gesture Recognition, 2018