The Conditional Boundary Equilibrium Generative Adversarial Network and its Application to Facial Attributes

Ahmed Marzouk, Pablo Barros, Manfred Eppe, and Stefan Wermter,
Knowledge Technology, Department of Informatics,
University of Hamburg, Germany
Hamburg, Germany
{4elshina,barros,eppe,wermter}@informatik.uni-hamburg.de

***Abstract*—We propose an extension of the Boundary Equilibrium GAN (BEGAN) neural network, named Conditional BEGAN (CBEGAN), as a general generative and transformational approach for data processing. As a novelty, the system is able of both data generation and transformation under conditional input. We evaluate our approach for conditional image generation and editing using five controllable attributes for images of faces from the CelebA dataset: age, smiling, cheekbones, eyeglasses and gender. We perform a set of objective quantitative experiments to evaluate the model's performance and a qualitative user study to evaluate how humans assess the generated and edited images. Both evaluations yield coinciding results which show that the generated facial attributes are recognizable in more than 80% of all new testing samples.**

***Index Terms*—Conditional GAN, image generation, image translation**

I. Introduction

Automatic data synthesis from conditional attributes with artificial neural networks is among the most popular researched topics in the last few years [1], [4], [10], [14], [18], [20]. The objective of a generative model is to precisely capture the real data distribution and reproduce samples from the same distribution that mimic the real data distribution, and Generative Adversarial Networks (GAN) have been introduced as a general solution to this problem [11]. Conditional generative modeling extends this paradigm by allowing for an additional input that conditions the generated output with respect to certain attributes. As an example, consider the generation of paintings conditioned to the style of a specific painter [9].

A popular application domain for conditional generative modeling is the generation of images of faces (e.g. [1], [12]). The conditioned face generation based on visual characteristics is a complex task and has a wide range of applications such as face recognition, human-computer interaction and security [11]. Training neural networks to generate artificial images is a hard problem for conventional machine learning concepts because the generated images do not appear natural to a human eye [18].

Approaches that perform facial image generation exist (e.g. [4], [12], [14], [17]), but, as we outline in Sec II, the approaches are limited in either the resolution of the output images, their capability to perform both translation and generation, or their capability of accepting conditional input. Furthermore, most systems have only been evaluated by means of neural machine classification approaches, and the quality of the generated images has not been assessed by humans. The lack of such systems motivates our following research question:

How can we realize a GAN architecture for both image generation and translation that is controllable by being sensitive to a conditional input, and that produces high-quality images according to human judgement.

To address this question, we build on the Boundary Equilibrium Generative Adversarial Networks (BEGAN) architecture proposed by Berthelot et al. [3], which is based on the reconstruction loss as a proxy for matching the distributions of the real and the generated data. The total loss is then measured from the Wasserstein distance [2] between the reconstruction losses of real and generated data. The contribution of this work consists of the following extension to the BEGAN:

1) We extend the BEGAN architecture by adding conditional boundaries. We refer to the resulting architecture as Conditional Boundary Equilibrium GAN (CBEGAN).
2) To this end, we introduce the conditional Wasserstein distance as a novel metric for the training loss.
3) We optimize the resulting system for the domain of face image generation.
4) We evaluate the system using both machine classification and a user study.

*The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169), the Volkswagen Stiftung and the NVIDIA corporation.

978-1-7281-2009-6/$31.00 paper N-20167.pdf

II. Related Work: Adversarial Models For Face Generation

Adversarial models are a recently introduced deep learning framework [11] and are widely used to generate plausibly looking images. Conditional face generation is a complex task with many potential applications, ranging from security to e-commerce. Choi et al. [4] have created a variant of the first-generation generative models [11] to produce images of faces while allowing for passing conditional information about the facial appearance to the generator. The conditional information includes smiling, cheekbones, age, gender and other attributes, and the authors developed this architecture with a focus on conditioned face generation while preserving identity. The CelebA dataset [16] was the authors' choice for the training data. Although their research was a significant step forward in generating conditioned face images, the authors conclude that the sample quality still provides significant potential for improvements [4].

More recent architectures focused on image reconstruction by disentangling salient information while mapping those facial attributes to their respective labels [14]. The result of the author's approach is an architecture that can generate various versions of an input image by varying the facial attribute value. The authors' model [14] allows for controlling how much a particular attribute is distinguishable in the translated generated image, but lacks support for image generation from scratch.

Other recent approaches focus on building generative models using autoencoders [6]. Adversarial autoencoder models have shown that they can learn to map data deterministically (via the encoder) to a latent space and learn a mapping (via the decoder) that allows reconstructing samples from the latent space again (e.g. [17]). Such architectures are very flexible and simple [6] while producing very good reconstruction results. However as the authors conclude, their model is not capable of performing conditional face image generation. Berthelot et al. [3] propose a model that is based on adversarial autoencoder, where the autoencoder is trained with dual objectives, i.e., reconstruction error criteria. Such formulation allows for high quality and high-resolution facial image generation but not with controllable attributes.

Durugkar et al. [8] propose a generative adversarial system known as the Generative Multi Adversarial Network (GMAN). The GMAN has multiple, symmetrical discriminator models and a single generator model. The GMAN discriminators are instantiated with marginally differing parameters but share the same architecture and are trained in a similar fashion to regular GANs. Each discriminator assesses and yields its scores on the currently generated sample by the generator. The scores are evaluated through a selection metric before being utilized to train the generator. This process results in two different sets of discriminators: a sinister adversary & a friendly critic. The sinister adversary is the set of discriminators that are set to boost their own scores by giving strict feedback to the solitary generator. A sample generated by the generator must please all the discriminators to get a higher rating. On the other hand, the friendly critic limits the discriminator models to be more positive towards the generated samples. The feedback from the discriminators are collected and averaged before sending them to the generator. In addition, the generator is permitted to restrain the performance of the discriminators if they become too strong. Durugkar et al. [8] conduct their experiments with the CIFAR-10 dataset [13] and the MNIST dataset [15]. The authors found that all variations of GMAN needed fewer iterations of training to reach a state of high-quality samples compared to a regular single discriminator GAN. The authors also claim that GMAN architecture is resistant against mode collapse since the GMAN generator must satisfy multiple discriminators.

Shrivastava et al. [19] utilize generative adversarial models as part of a larger machine learning system, with the goal of improving the realism of image generation while keeping the annotated information. The proposed system is composed of two components: a simulator that is capable of generating synthetic images with annotations and a refiner network (a critic) that uses GAN to improve the quality and maintaining the annotations. The simulator generates images based on labeled data, but the problem is that these images are not realistic and contain artifacts. Shrivastava et al. replace a large portion of the input batch, almost half of the current batch with previous images, and randomly update half the buffer for each iteration of training. The authors claim [19] that their system is capable of generating images that are of better quality than regular GAN [11].

Choi, Yunjey, et al. [5] propose a generative model called StarGAN that is able to perform the image-to-image translation in more than one domain. The model-generated output is of high quality as compared to other discussed models that came before. Although the model is scalable in performing image-to-image translation among multiple domains with high-quality visual output in comparison to other approaches [5], the model does not handle multiple attribute swap.

The approach by Hinz et al. [12] is an exception within the state of the art, in the sense that it allows for conditional image generation and translation using disentangled representations. However, the authors' system is limited in the resolution of the generated images, and it has not been evaluated by human assessment.

III. Background: BEGAN

Our work is based on the BEGAN [3] architecture, which has shown impressive results in generating high fidelity images at a resolution of 128x128 pixels. Like other GANs, BEGAN uses a generator G and a discriminator D. However, in contrast

to other GANs, the generator and the discriminator in this architecture are both based on autoencoders with the Wasserstein distance defined as the training loss function. The BEGAN training method allows incorporating a convergence measure, which reflects the quality of the generated images. Berthelot et al. [3] were first to introduce a diversity hyperparameter which, as their experiments have shown, has been used to automatically set a balanced trade-off between image diversity and quality of generation. The core principle of BEGAN is utilizing the autoencoder reconstruction abilities to optimize the training loss function for the entire model. Since the architecture incorporated autoencoders, the loss function is defined as the function of the quality of reconstruction achieved by the discriminator D on real and generated images. The reconstruction loss in this context is the error associated with reconstructing images (whether real or generated) through the discriminator. BEGAN uses the matching of the reconstruction loss distributions as a proxy for matching data distributions, and is optimized with respect to the total loss which is defined as the Wasserstein distance between the reconstruction losses of real and generated data/images. Hence, BEGAN model networks are trained by optimizing the total loss in conjunction with the equilibrium term. Formally, the objective function for the discriminator loss L_D and the generator loss L_G are defined as in the following Eq. (1):

$$\begin{aligned} L_D &= L(x) - k_t.L(G(z)) \\ L_G &= L(G(z)) \\ K_{t+1} &= k_t + \lambda * (\gamma.L(x) - L(G(z)) \end{aligned} \tag{1}$$

where L_D and L_G are the reconstruction errors for D and G, k_t is an adaptive term that allows balancing the losses at each step t, and λ is the learning rate for k_t. γ is the diversity ratio, defined as the ratio between the 2 losses: $\gamma = \frac{E[L(G(z))]}{E[L(x)]}$ Note that γ is within a range of [0,1].

IV. Conditional Boundary Equilibrium GAN (CBEGAN)

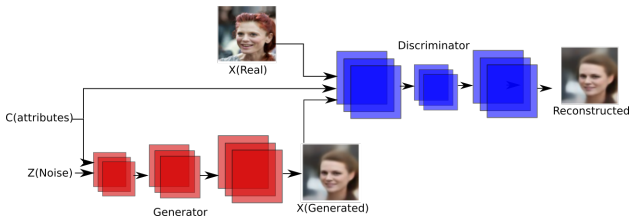

Fig. 1. Conditional Adversarial Autoencoder(CBEGAN)

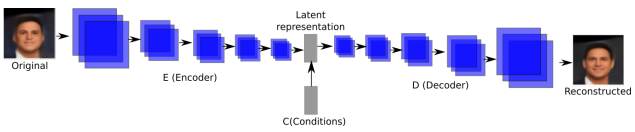

Fig. 2. Discriminator Architecture

The CBEGAN is an encoder-decoder architecture that is trained with an adversarial objective. It extends the BEGAN with conditional boundaries as described in the following:

A. Encoder-decoder Architecture

Our model described in Figure 2 is based on an encoder-decoder architecture. The discriminator D is an autoencoder (encoder-decoder) and c is the conditional information passed on to the encoder of D and the generator G. G has a similar architecture as the decoder of the autoencoder. During training, the generator will use the noise z and the condition c to generate a sample $x_{generated}$. This sample is then passed on to the encoder of the discriminator D along with the condition c. The encoder will encode the generated sample producing a latent space encoding representing the generated sample. The encoder will then apply the condition c that was passed along on the encoded representation of the image, thus influencing its reconstruction as the decoder will use this information to reconstruct the image based on the conditional information c that was passed on with the encoded image. Simultaneously, the encoder is also presented with a real sample x_{real}(an image from the training dataset) along with its respective label/condition, which the encoder will encode as well and perform the same operation as described above. Similarly, the decoder will receive the encoded information of the real image along with its respective conditional information and perform reconstruction based on the conditional information passed as described above. These conditionally encoded representations of the generated and the real images are passed on to the decoder component of the discriminator D that reconstructs the generated images $r_{generated}$ and real images r_{real} based on the condition applied.

B. Conditional Wasserstein distance and adversarial objective

With the CBEGAN, we introduce the conditional Wasserstein distance which considers the condition vector passed to the decoder when computing the reconstruction loss, as stated in Eq. 2.

$$\begin{aligned} L_D &= L(x|c) - k_t.L(G(z|c)) \\ L_G &= L(G(z|c)) \\ K_{t+1} &= k_t + \lambda * (\gamma.L(x|c) - L(G(z|c)) \end{aligned} \tag{2}$$

where

- L_D & L_G are the model respective losses for D & G that both model components try to minimize
- c is conditional information (label) that is passed on to the CBEGAN model
- $L(x|c)$ is the reconstruction loss of the real images conditioned on conditional information
- $L(G(z|c))$ is the reconstruction loss of the generated images conditioned on conditional information
- γ is the diversity ratio with a range between [0,1]; γ is defined as the ratio between the 2 losses, however, in our

formulation the γ is defined over the conditioned losses $\gamma = \frac{E[L(G(z|c))]}{E[L(x|c)]}$

In comparison to the BEGAN loss defined in Eq. (1), the objective of the encoder and decoder is now to compute a latent representation that encodes the conditional information represented by c, and the objective of the decoder is to reconstruct x given c. The above formulation also allows utilizing the discriminator/autoencoder to perform attributes swap (since the autoencoder is capable of learning attribute associations), a feature which is not present in the BEGAN model. Finally, CBEGAN and BEGAN both share the same generator filters 128x128 which produce images with a 128x128 resolution.

V. Implementation Details

We adapt the architecture of our network from Berthelot et al. [3] and add an extra layer in the autoencoder architecture D to realize the concatenation of the encoded representation of the passed images (both real and generated) with the conditional information c, right before it is being passed to the decoder. The generator of our model differs from that of Berthelot et al. in terms of input-layer size, as it requires the concatenation of c with z (see Figure 2).

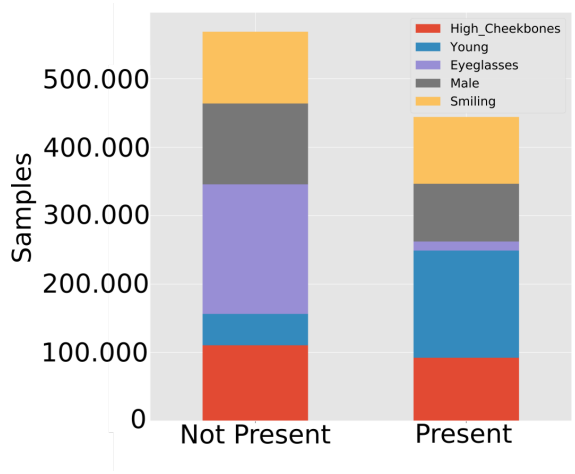

Fig. 3. The selected labels from CelebA dataset: high Cheekbones, the presence of eye glasses, gender, apparent age, the presence of a smile

The CBEGAN's discriminating autoencoder is composed of an input layer followed by the encoder which is made of eight convolutional layers, each with a 2D kernel, followed by a flattened layer, then a dense output layer. The information from the dense output layer is then concatenated with the corresponding conditional information (facial attribute) in the concatenation layer. The concatenation layer is the final output that is then passed on the decoder model for decoding. The CBEGAN hyperparameter details are presented in Table I.

The generative decoder of the CBEGAN is composed of an input layer, a single dense layer, a single reshape layer and nine convolutional layers with one upsampling layer between every two convolutional layers. The final layer produces the

Fig. 4. The CBEGAN model is capable of associating up to five features for a face image. This implies that there are 32 different attributes combinations that the CBEGAN model learns after completion of training. In the above picture, we generate images for five different condition configurations and five different noise inputs, resulting in 25 different faces. CB/NCB: cheekbones/no cheekbones, M/F: male/female, E/NE: eyeglass/no eyeglass, S/NS smiling/not smiling, Y: young.

reconstructed/generated conditioned image. The hyperparameter details and the generator's structure is summarized in Table II.

The CBEGAN decoder also serves as the model's generator component. It aims to minimize the reconstruction loss of the generated images by working adversarially to the autoencoder as explained above. The reconstruction loss is the error associated to reconstructing image samples through the discriminator, which is defined as the mean value of the element-wise absolute value. The CBEGAN decoder/generator is a conditional decoder that can process the condition (labels or facial features) information passed on to it, trained in an adversarial manner. The CBEGAN decoder playing the role of the generator $G(z)$ generates images from a noise input z. The decoder guided by the conditional information passed on to it maps z into the data space to try to fool the discriminator, i.e., produce realistic looking images that are indistinguishable from the real data distribution. The CBEGAN decoder decodes real and generated images from the latent space (the latent space that was created by the encoder) based on the conditional information (labels or facial attributes) that is passed along with the encoded representation. The CBEGAN generator eventually learns (through the reconstruction loss) to associate facial attributes (labels) with face images, this allows the

TABLE I
DISCRIMINATOR / AUTOENCODER STRUCTURE

Layer (type)	Output Shape	Parameter	Connected to
input_1 (InputLayer)	(None, 3, 128, 128)	0	
encoder L1 Conv1 (Conv2D)	(None, 128, 256, 256)	3584	input_1[0][0]
encoder L1 Conv2 (Conv2D)	(None, 128, 128, 128)	147584	encoder L1 Conv1[0][0]
encoder L2 Conv1 (Conv2D)	(None, 256, 128, 128)	295168	encoder/L1/Conv2[0][0]
encoder L2 Conv2 (Conv2D)	(None, 256, 64, 64)	590080	encoder L2 Conv1[0][0]
encoder L3 Conv1 (Conv2D)	(None, 384, 64, 64)	885120	encoder L2 Conv2[0][0]
encoder L3 Conv2 (Conv2D)	(None, 384, 32, 32)	1327488	encoder L3 Conv1[0][0]
encoder L4 Conv1 (Conv2D)	(None, 512, 32, 32)	1769984	encoder L3 Conv2[0][0]
encoder L4 Conv2 (Conv2D)	(None, 512, 32, 32)	2359808	encoder L4 Conv1[0][0]
flatten_1 (Flatten)	(None, 524288)	0	encoder L4 Conv2[0][0]
encoder Dense (Dense)	(None, 64)	33554496	flatten_1[0][0]
pose_aslabels (InputLayer)	(None, 3)	0	
concatenate_1 (Concatenate)	(None, 67)	0	encoder Dense[0][0] pose_aslabels[0][0]
decoder (Model)	(None, 3, 128, 128)	2303075	concatenate_1[0][0]

CBEGAN model to generate high fidelity images and also perform an image-to-image translation.

TABLE II
GENERATOR / DECODER STRUCTURE

Layer (type)	Output Shape	Parameter
input_3 (InputLayer)	(None, 67)	0
decoder Dense (Dense)	(None, 32768)	2228224
reshape_2 (Reshape)	(None, 32, 32, 32)	0
decoder L1 Conv1 (Conv2D)	(None, 32, 32, 32)	9248
decoderL1 Conv2 (Conv2D)	(None, 32, 32, 32)	9248
up_sampling2d_4 (UpSampling2)	(None, 32, 64, 64)	0
decoder L2 Conv1 (Conv2D)	(None, 32, 64, 64)	9248
decoder L2 Conv2 (Conv2D)	(None, 32, 64, 64)	9248
up_sampling2d_5 (UpSampling2)	(None, 32, 128, 128)	0
decoder L3 Conv1 (Conv2D)	(None, 32, 128, 128)	9248
decoder L3 Conv2 (Conv2D)	(None, 32, 128, 128)	9248
up_sampling2d_6 (UpSampling2)	(None, 32, 128, 128)	0
decoder L4 Conv1 (Conv2D)	(None, 32, 128, 128)	9248
decoder L4 Conv2 (Conv2D)	(None, 32, 128, 128)	9248
decoder FinalConv (Conv2D)	(None, 3, 128, 128)	867

VI. EXPERIMENTS SETUP

We use the CelebA dataset [16] as a basis to perform image generation and image translation experiments (Sec. VI-A). We performed preprocessing steps (Sec. VI-B) and used the resulting data to perform image generation (Sec. VI-C) and image translation (Sec. VI-D).

A. Dataset

The CelebA dataset [16] was selected as our training dataset due to its diverse facial attributes representation. CelebA is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. Ideally, CelebA [16] comes with only certain facial attributes pre-labeled such as age, hair color, gender, etc. We selected the following labels to be used as conditions: (high cheekbones, the presence of sun-glasses, gender, apparent age, and the presence of a smile). These characteristics were chosen due to their balanced presence within the dataset and their distinctive representation on the image level. The label distribution is illustrated in Figure 3.

B. Preprocessing

The inputs of the proposed model need to have a fixed length for both the generator and the discriminator since they are both composed of convolutional neural networks. Therefore, every facial attribute value has been preprocessed and assigned a value of 1 (to imply its presence) or 0 (to imply its absence) and all images were preprocessed (re-sized while keeping their RGB channels) for instance the original CelebA image dimension is 178x218 which is re-sized into 128x128 during training. No image normalization was needed. The processed images are sent to the model along with the transformed condition vector. This results in 200K image samples of dimension (3, 128,128) and a condition vector of 5- imensions. Other than the described preprocessing steps no further preprocessing steps have been applied to the images provided from the CelebA dataset.

C. Image Generation Task

In the image generation task, images are generated from scratch, only based on the condition vector c. Our implementation supports up to five conditions (see Fig. 3), i.e. c is of size five, with each of its Boolean components indicating the presence or absence of a particular attribute. To generate an image, a condition vector is passed to the generator that uses this information for the reconstruction. The noise vector z determines the basic facial structure, and the conditions determine variations of the face. Examples of the results are illustrated in Figure 4.

D. Image-to-Image Translation Task

In the image-to-image translation task, a given image is changed according to the characteristic determined by the con-

dition vector c. This is realized by querying the discriminator's autoencoder with a condition vector c. Examples are illustrated in Figure 5.

Fig. 5. Image-to-Image Translation using real images. A real image is passed to the autoencoder and modified according to a single facial attribute.

VII. Evaluation

To assess our model, we perform a quantitative evaluation using an external neural network trained for classification, and we also perform a qualitative evaluation by performing a user study.

A. Quantitative Evaluation

To evaluate the CBEGAN capability to generate and perform image-to-image translation, we consider classification accuracy. To this end, we train a convolutional neural network to classify facial attributes based on the labeled data contained in the CelebA dataset [16]. To this end, we leverage transfer learning and use the Inception-v3 ConvNet architecture [21] that has been pre-trained on the imagenet dataset [7]. We, re-trained the last three layers of the Inception-v3 model on the CelebA subsets described above while freezing the other previously trained layers to utilize the weights they had already learned, i.e., transfer learning. We use the trained network to calculate the classification accuracy by running it over images generated by the CBEGAN model. We also use the same classifiers to classify translated images output by our proposed model. The trained classifiers accuracy can be found in Table III

TABLE III
Classifier Accuracy

Classifier(type)	Accuracy
smiling_not_smiling	75%
young_old	77%
male_female	95.4%
sunglasses_nosunglasses	87.6%
High_cheekbones_noHigh_cheekbones	74.4%

B. Qualitative Evaluation

To evaluate the subjective quality of the CBEGAN model, two qualitative studies are designed to evaluate the CBEGAN model concerning image generation and image-to-image translation respectively. To accomplish that, we generate 10 images for each of the (2^5) condition combinations, leading to a sum of 320 images in total. The images are presented to a group of 14 subjects through a simple User Interface (UI) developed to present the participant with an image and record their choice. The participants evaluate if each individual attribute is present or not. We calculate the accuracy of the subject's responses and compared it to the ground truth. The second user study was conducted analogously, aiming to evaluate the image-to-image translation quality.

VIII. Results and Discussion

A. Quantitative Results

Figure 6 demonstrates the average accuracy obtained when running the trained classifiers on the generated and translated images by the CBEGAN model.

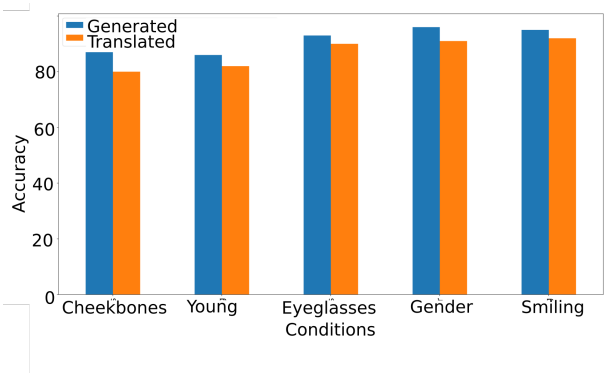

Fig. 6. Results of the trained classifiers when run on the generated and translated images by CBEGAN.

B. Qualitative Results

The results of the participants of the first and second user study which show the average classification accuracy of the human participants in classifying the generated and translated images by the CBEGAN model are found in Figure 7.

C. Discussion

For both evaluations, we observe that certain attributes show higher average accuracy as compared to other used attributes, which can be attributed to their respective presence within the dataset and to their visibility on the images. However, we observe an overall accuracy of mostly above 80%. Interestingly, the generated images consistently show a higher accuracy than the translated ones, which is possibly due to the fact that they are less constrained by the input. It is also remarkable that both the user study and the classification consistently show similar results.

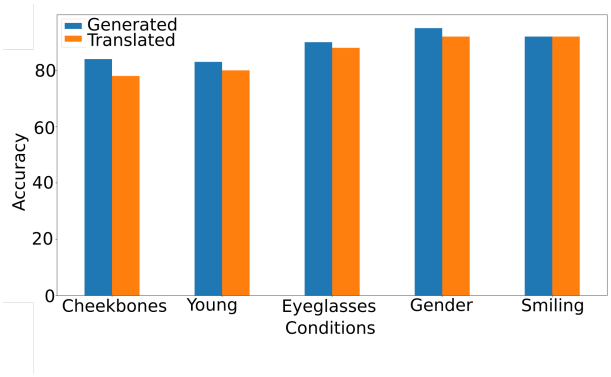

Fig. 7. Results of the user study participants classifying generated images.

IX. CONCLUSION AND FUTURE WORK

This research project aims at evaluating the hypothesis that adding conditional boundaries and utilizing the conditional Wasserstein distance as a cost function to a conditional autoencoder architecture trained in an adversarial manner would allow a real data distribution representation to achieve to high quality images with controllable high-level attributes. To address this hypothesis, we have extended the BEGAN architecture with conditional boundaries and introduced the conditional Wasserstein distance as a metric for the loss computation.

The resulting CBEGAN method has several benefits compared to other state-of-the-art approachs. It has a higher quality in terms of classification accuracy compared to at least the recent approach by Hinz et al. [12], which reports an accuracy of less than 80% in most cases, while our results consistently show an accuracy above 80%. Furthermore, our approach supports the generation and translation of images with up to five conditions, in combination with a comparably high resolution of 128x128 pixels. This has not been yet achieved by other approaches.

The results show that our model provides significant potential in application domains related to big data processing, including crime investigations, fashion and e-commerce.

In this work, we have introduced the CBEGAN model in the context of generating and translating images of faces, but the general architecture is agnostic to the kind of data to be processed. Therefore, we plan to investigate the use of CBEGAN in other domains such as speech processing, and we are also looking forward to employ it for multi-modal applications, such as audio-visual data processing. Our project has shown that conditional Wasserstein distance is applicable as a measure for mimicking real-data distribution. However, we think adding a penalty for preserving identity to the loss function might be interesting and could improve training. We plan to add extensions to increase the number of attributes as well, and we will also perform tests with larger generator filters to produce images at higher resolutions.

REFERENCES

[1] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] David Berthelot, Tom Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[4] Xuwen Cao, Subramanya Rao Dulloor, and Marcella Cindy Prasetio. Face generation with conditional generative adversarial networks.

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint*, 1711, 2017.

[6] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[8] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. *arXiv preprint arXiv:1611.01673*, 2016.

[9] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. CAN: Creative Adversarial Networks Generating Art by Learning About Styles and Deviating from Style Norms. In *International Conference on Computational Creativity (ICCC)*, 2017.

[10] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2, 2014.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] Tobias Hinz and Stefan Wermter. Image Generation and Translation with Disentangled Representations. Technical report, 2018.

[13] Alex Krizhevsky and Geoff Hinton. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7), 2010.

[14] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017.

[15] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[17] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[18] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[19] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, volume 2, page 5, 2017.

[20] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

[21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.