INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 15, Issue: 2, Month: April, Year: 2020
Article Number: 3862, doi.org/10.15837/ijccc.2020.2.3862

Pose Manipulation with Identity Preservation

A. T. Ardelean, L. M. Sasu

Andrei-Timotei Ardelean*
Transilvania University of Brasov
500036 Brasov, B-dul Eroilor, 29, Romania
*Corresponding author: andrei.ardelean@student.unitbv.ro

Lucian Mircea Sasu
1. Transilvania University of Brasov
500036 Brasov, B-dul Eroilor, 29, Romania
lmsasu@unitbv.ro
2. Xperi Corporation
500152 Brasov, Turnului, 5, Romania
lucian.sasu@xperi.com

Abstract

This paper describes a new model which generates images in novel poses e.g. by altering face expression and orientation, from just a few instances of a human subject. Unlike previous approaches which require large datasets of a specific person for training, our approach may start from a scarce set of images, even from a single image. To this end, we introduce Character Adaptive Identity Normalization GAN (CainGAN) which uses spatial characteristic features extracted by an embedder and combined across source images. The identity information is propagated throughout the network by applying conditional normalization. After extensive adversarial training, CainGAN receives figures of faces from a certain individual and produces new ones while preserving the person's identity. Experimental results show that the quality of generated images scales with the size of the input set used during inference. Furthermore, quantitative measurements indicate that CainGAN performs better compared to other methods when training data is limited.

Keywords: pose manipulation, image generation, adaptive normalization, Generative Adversarial Network.

1 Introduction

Developing a way to easily manipulate face expression and head pose of an individual has been the focus of many research groups in the last decade. The baseline solution for this task involves creating a 3D model and animate it accordingly. However, building a photo-realistic head model using standard techniques can take substantial amount of time for a human artist. Many industries could benefit from optimizing this process such as cinematic, advertising and video games; furthermore, it has potential for image enhancement and editing software.

A traditional automatic method for face manipulation is based on 3DMM fitting [2]. Parameters can be estimated from a single image and then changed to obtain different expressions. This method is not sufficient by itself to work with hidden regions, e.g. teeth and closed eyes [27].

Other approaches rely on warping from one or more source images to the desired pose to generate guided head images [26]. A drawback of this solution is the limited amount of variation between the source and target pose it can manage without great loss of quality [28].

New approaches for face image generation have been brought by recent advances in generative adversarial networks (GANs) [10]. Seminal papers in this area [14, 15] show that one can generate high resolution realistic figures of human faces using GANs.

Our contribution is a new model, Character Adaptive Identity Normalization GAN (CainGAN), that receives figures of faces from a certain individual and produces new ones while preserving the person's identity. CainGAN generates images in novel poses starting from a small set of source pictures with the individual, i.e. a few-shot setting, without any fine-tuning as found in [9, 28].

We conducted experiments to compare images generated by CainGAN, with alternative systems using image-to-image translation [13, 25] and conditioning based on Adaptive Instance Normalization [12] using computed embeddings [28]. By performing quantitative measurements on the self-reenactment task we show that our model is able to achieve state-of-the-art results using less data compared to other methods and without fine-tuning.

The rest of the paper is structured as follows: In section 2 we make a literature review; the subsequent section describes CainGAN in detail; section 4 contains a comparison between different methods and an ablation study. Eventually, we summarize our contributions in section 5. Code for the implementation of CainGAN is available at https://github.com/TArdelean/CainGAN

2 Related work

A number of works on face generation with preservation of identity focus on the talking face task, i.e. the area of interest is the mouth region with motion driven by either audio sources [4, 5, 22, 30] or video to be imitated [21, 30]. These methods cannot be easily extended to synthesize full head images that require handling more variation between poses or hidden elements in the source images. While it is possible to replace the face from an existing head footage, by using a face modeling approach as in [23], the result of pose manipulation would be limited to face expression.

Providing conditional information to several layers of the generator has been widely used to prevent input constraints from vanishing. Good results were obtained especially by modulating activations using AdaINs [6, 12, 15] and SPADE [19] that employs spatial denormalization to introduce semantic map constraints.

Our model is based on a conditional GAN framework, i.e. instead of generating starting from noise as done traditionally [10], the input of the generative network can take different forms including images [13, 19] as done in this work. The use of multiple discriminators to stabilize GAN training has been recently studied in several works [1, 8, 18]. Specifically our model uses two discriminators with different objectives.

3 The CainGAN model

In this section we present the architecture of the proposed model, followed by the training algorithm. Eventually we give some implementation details.

3.1 The Architecture

The goal is to train a generative model that is able to synthesize new images starting from K existing source pictures with the same person. Let x_i denote the i-th image of an image sequence $\mathbf{x}$; we uniformly sample $K+1$ distinct images from $\mathbf{x}$. The model receives as input $x_{i_1}, x_{i_2}, ..., x_{i_K}$ along with their corresponding landmark images [3] $L(x_{i_1}), L(x_{i_2}), ..., L(x_{i_K})$ and target landmark $L(x_{i_{K+1}})$, then generates a new image $\hat{x}_t$ that must follow target landmark and preserve identity of the person

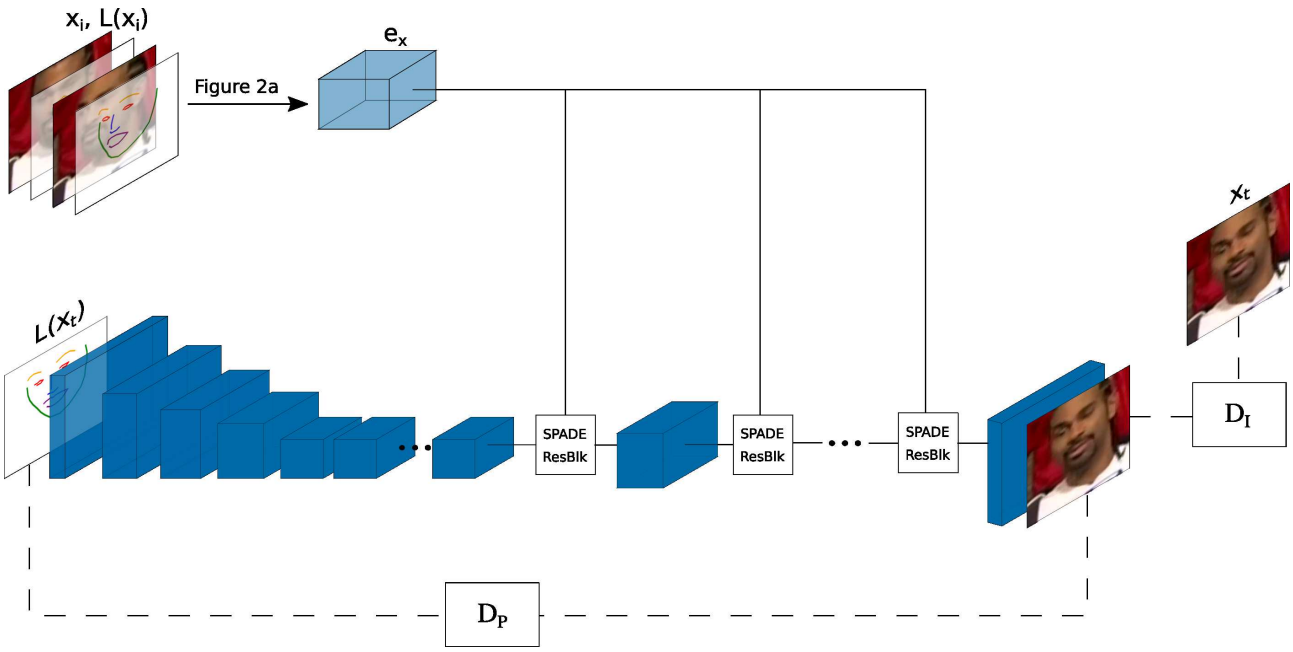


Figure 1: Full architecture and model workflow. Top side depicts the embedding which is used at each upsampling step to spatially modulate activations. The generator (shown in the bottom side of the figure) starts from the target landmarks $L(x_t)$ to synthesize a new image $\hat{x}_t$. D_I and D_P represent the identity and pose discriminators, respectively.

from the K input images. The generated image is expected to be similar to the ground truth image $x_t := x_{i_{K+1}}$.

We propose CainGAN, a model that consists of 4 networks (Figure 1) which we will describe in the following.

The embedder computes a spatial embedding e_{x_i} from a single input image: $e_{x_i} = E(x_i, L(x_i))$. Implementation details can be found in section 3.3. In order to use multiple source images, a method to combine the embeddings must be devised. Zakharov et al. in [28] simply averages the one-dimensional tensors computed by the embedder. We observed that weighing the features by their relevance helps to better capture the identity, therefore a responsibility based combining method was developed (eq. (1), Figure 2a). To achieve this, the embedder will also output a weighing tensor r_{x_i} representing the certainty for the computed features. The final identity embedding is calculated by a function Ψ as:

$$e_x = \Psi\left((e_{x_1}, r_{x_1}), \ldots, (e_{x_K}, r_{x_K})\right) = \frac{\sum_{i\in\{i_1,\ldots,i_K\}} e_{x_i} \cdot r_{x_i}}{\sum_{i\in\{i_1,\ldots,i_K\}} r_{x_i}} \tag{1}$$

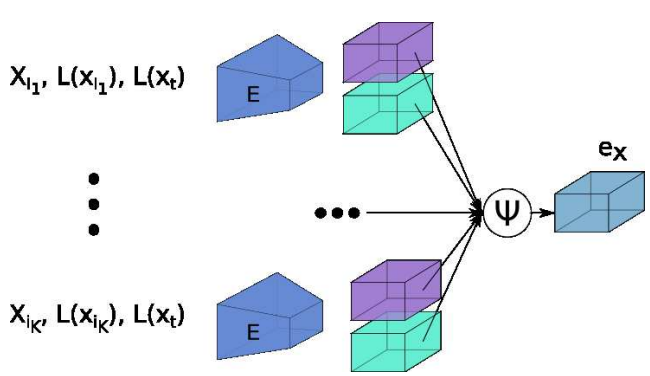
We explored the use of target landmarks $L(x_t)$ as input to the embedder along with x_i and $L(x_i)$ and found this helpful for the generation process, since the identity features can be aligned to the final pose earlier. Hence, this is the version used in our experiments, denoted Targeted Embedder, as illustrated in Figure 2b.

The generator starts from the target landmark and generates a new image $\hat{x}_t = G(L(x_t), e_x)$. The landmark image is provided through the input layer while the combined spatial embedding e_x is used to modulate the activations at several resolutions with SPADE blocks. In order to assess the quality of the generated image we employ two discriminators: identity discriminator and pose discriminator.

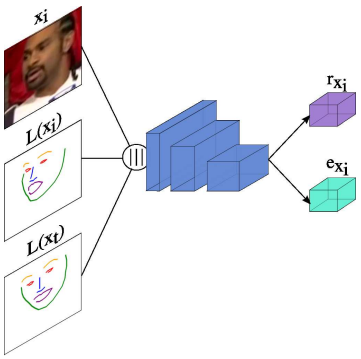
Identity discriminator $D_I(x_a, x_b)$ follows the multi-scale architecture from pix2pixHD [25] and is used to estimate identity resemblance. At this stage we only consider characteristic traits and the results are expected to be invariant to pose. Thus, the input consists of two RGB images of the same person in arbitrary positions.

Pose discriminator $D_P(x_a, L(x_a))$ has the same architecture as D_I with its own set of parameters. The network receives a frame and the appropriate target landmarks and checks the correspondence.

While both discriminators will also assess general realism of a given image, D_P is used specifically to avoid pose mismatch, whereas D_I encourages identity preservation. This disengagement allows us



(a) Combining embeddings based on responsibility, as in equation (1). The weights of the Targeted Embedder (E) are the same for each of the K inputs. The Targeted Embedder is detailed in Figure 2b.



(b) The structure of the Targeted Embedder. The real frame x_i, its landmark image $L(x_i)$ and target landmarks $L(x_t)$ are concatenated and processed to get r_{x_i} and e_{x_i}.

Figure 2: Embedder network

to assign different importance factors to each discriminator while training. Different combinations may give results that correlate better with human visual perception.

3.2 Training

Videos are used as image sequences during training. For each epoch we sample a sequence of $K+1$ distinct frames for each training video and compute landmark images: $L(x_{i_1}), L(x_{i_2}), ..., L(x_{i_{K+1}})$ using a pretrained face alignment network [3]. The embedding tensor e_x is then computed according to (1). Using $L(x_t)$ and e_x, CainGAN generates the new image $\hat{x}_t$ which is further fed to the identity discriminator $D_I(\hat{x}_t, x_{i_1})$ along with a source frame x_{i_1}. The choice of the identity frame index i_1 is arbitrary as the generation process is not influenced by the order of input images. $\hat{x}_t$ and $L(x_t)$ are also fed to the pose discriminator $D_P(\hat{x}_t, L(x_t))$. Importance factors λ_I and λ_P control the weight of each discriminator in the objective function given by the hinge loss [16, 17]:

$$\begin{aligned}\mathcal{L}_{Adv}(G,E,D) = &-\lambda_I \cdot D_I(G(L(x_t), e_x), x_{i_1}) \\ &-\lambda_P \cdot D_P(G(L(x_t), e_x), L(x_t))\end{aligned} \tag{2}$$

$$\mathcal{L}_{Id}(D_I) = \max(0, 1 - D_I(x_t, x_{i_1})) + \max(0, 1 + D_I(\hat{x}_t, x_{i_1})) \tag{3}$$

$$\mathcal{L}_{Pose}(D_P) = \max(0, 1 - D_P(x_t, L(x_t)) + \max(0, 1 + D_P(\hat{x}_t, L(x_t)) \tag{4}$$

The identity and pose discriminators are updated using $\mathcal{L}_{Id}$ and $\mathcal{L}_{Pose}$, respectively. The weights of the generator and the embedder are updated together using the full objective:

$$\begin{aligned}\mathcal{L}(G,E,D) = \mathcal{L}_{Adv}(G,E,D) + \lambda_{FM}(\lambda_I \cdot \mathcal{L}_{FM}(D_I) + \\ \lambda_P \cdot \mathcal{L}_{FM}(D_P)) + \lambda_{VGG} \cdot \mathcal{L}_{VGG}\end{aligned} \tag{5}$$

λ_{FM} and λ_{VGG} represent hyperparameters that control the importance of the loss in the full objective. $\mathcal{L}_{VGG}$ is a perceptual loss that compares features extracted at several layers by a pretrained VGGNet [20] from the original and the generated image. Feature matching loss $\mathcal{L}_{FM}$ is also a perceptual loss, comparing activations in the layers of the discriminator according to (6). Importance factors λ_I and λ_P also affect the weight of feature matching losses. The FM loss is similar to the one used in [25] as we employ multiscale discriminators:

$$\mathcal{L}_{FM}(D) = \sum_{i=1}^{S} \sum_{j=1}^{T} \frac{L_1(D_i^{(j)}(x_t, y), D_i^{(j)}(\hat{x}_t, y))}{N_j} \tag{6}$$

where S represents the number of scales, T is the number of layers in D, N_j is the number of elements in layer j and L_1 denotes the standard Manhattan distance. y is either the value of $L(x_t)$ when

using the perceptual loss induced by the pose discriminator, or x_{i_1} for the identity discriminator, respectively.

To ease the training process we start with a low importance factor for the identity discriminator and linearly increase it to its maximum value over the first 10 epochs. This allows the model to learn the easy task first, generating realistic face images in given pose, after which we gradually impose identity preservation.

We alternate between (G, E) and D updates, with twice more steps for the discriminator and using two time-scale update rule [11] to stabilize training.

3.3 Implementation details

The generator resembles an encoder-decoder architecture for image translation. There are 4 downsamplings residual blocks with learned skip connections, 3 same resolution residual blocks and 4 upsampling layers. Instance normalization [24] is used after every downsampling and upsampling layer. A SPADE residual block with spectral normalization [17, 29] is used after each upsampling. Nearest interpolation is used to bring the spatial embeddings to the appropriate resolution for each SPADE block. The discriminators are based on the architecture proposed by Wang et al. in [25], and use 2 scales as the images are relatively small. The embedder consists of 2 downsampling and 4 same resolution residual blocks which are shared while computing r_{x_i} and e_{x_i}. Two independent same resolution residual blocks are then used to get r_{x_i} and e_{x_i}.

4 Experiments and results

To evaluate our approach, we conduct extensive experiments on the VoxCeleb2 [7] dataset. To emphasize the ability of our model to learn from less data, we only use a small subset of the actual dataset. Originally, the train set contains almost 6000 different speakers featuring more than a million videos. For our experiments we randomly selected 150 speakers and their corresponding videos (around 30,000), less than 3% of the grand total. A video dataset was used since it is an accessible way to obtain multiple images with the same identity.

Quantitative comparison is performed against two baselines: pix2pixHD [25] and previously state-of-the-art method for talking head generation [28] denoted FSHM (Few-shot Head Models). We trained the pix2pixHD model from scratch as described in the original paper and official implementation. In order to use the model without fine-tuning, the network input consists of all source frames and their landmarks as well as the target landmark; these are also given to the discriminator. We also implemented a version of FSHM (feed-forward only) in order to assess the results in a limited training data setting.

Three different metrics are used to compare the described methods: structural similarity metric (SSIM) between the ground truth and the generated image is used to measure low-level structural similarity, cosine similarity (CSIM) between embedding vectors as computed by a pretrained face recognition network and Fréchet Inception Distance (FID) [11] measuring perceptual realism which usually better captures the similarity of real and fake images. We follow the same training setup presented by Zakharov et al. [28], using 50 video sequences with 32 test frames for each.

The comparison given in Table 1 shows that CainGAN is able to get better quantitative results using only a fraction of the dataset. Additionally, the method is able to generate realistic images in the desired pose with a good preservation of identity. From qualitative comparison in Figure 3 we can see that while FSHM can synthesize the face with the right alignment there is a high identity mismatch. Clearly, small amounts of training images severely affect the ability of the FSHM model to generalize to unseen faces. We also obtain the uncanny artifacts present in images generated by pix2pixHD, as reported in [28].

4.1 Ablation study

We performed an ablation study to analyze the influence of different components of our method. Quantitative results of the experiments are visible in Table 2. The variants are: CainGAN without

Method (K)	SSIM ↑	CSIM ↑	FID ↓
pix2pixHD(1)	0.66	0.80	72.26
FSHM (1)	0.64	0.72	93.17
FSHM-FF-full (1)	0.61	N/A	46.61
FSHM-FT-full (1)	0.64	N/A	48.5
CainGAN (1)	**0.69**	**0.85**	**35**
pix2pixHD(8)	0.66	0.81	71.89
FSHM (8)	0.65	0.73	83.13
FSHM-FF-full (8)	0.64	N/A	42.2
FSHM-FT-full (8)	0.68	N/A	42.2
CainGAN (8)	**0.77**	**0.91**	**24.92**

Table 1: K is the number of source frames used for testing. For SSIM and CSIM higher is better, for FID lower is better. CainGAN (8) was stopped after 20 epochs to avoid overfitting, all other models were trained for 30 epochs. The "full" suffix refers to the models being trained on the entire dataset. These results are taken from [28]. CSIM is not reported here, as a different face recognition network was used for the original results.

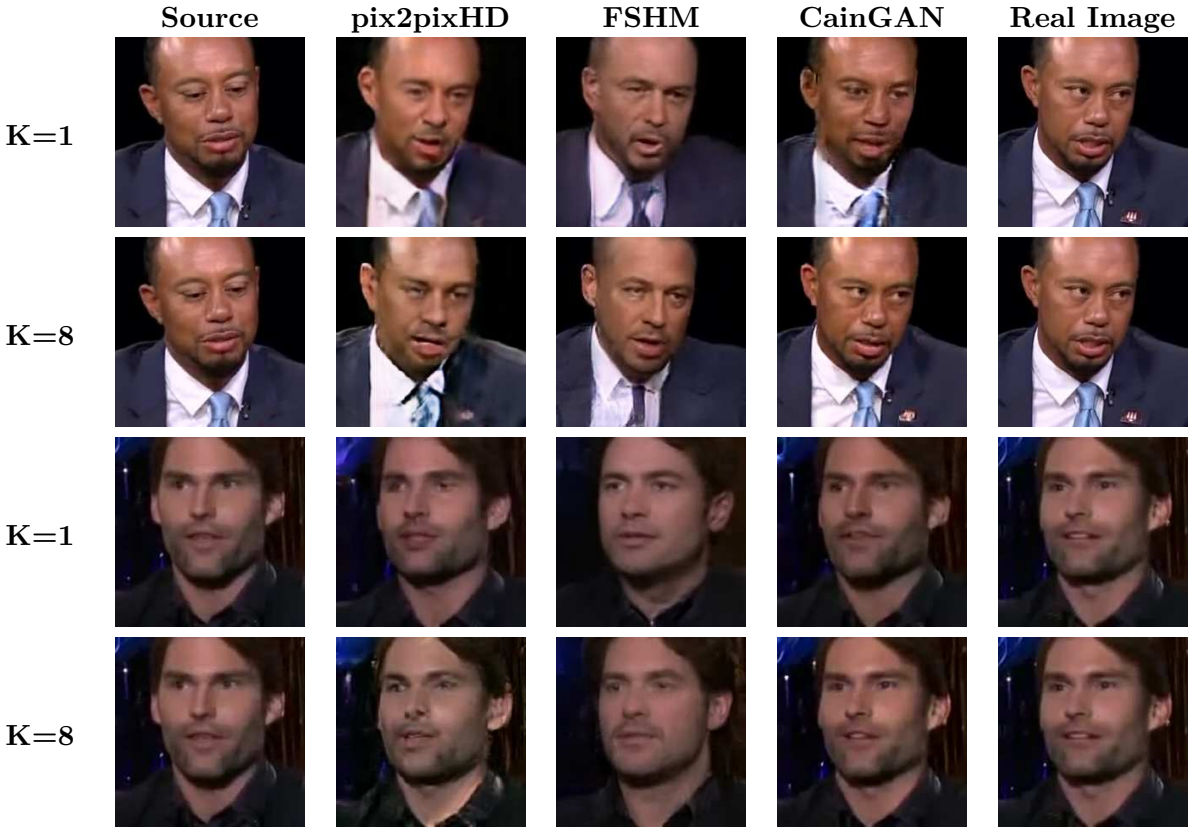


Figure 3: Visual assessment on the VoxCeleb2 dataset. First column represents the number of source frames, the next column illustrates one of the K source images and the last column contains the ground truth (x_t) images. In between are the generated frames by different methods. The figure is best viewed in color.

targeting (CainGAN w/o T) where only the source frame and its landmarks are given to the embedder, CainGAN without discriminator importance weighing (CainGAN w/o I) where $\lambda_D = \lambda_I$ are fixed and CainGAN without responsibility based embedding mixing (CainGAN w/o R) where the weighted version in equation (1) is replaced by:

$$e_x = \frac{1}{K} \sum_{i \in \{i_1, \ldots, i_K\}} e_{x_i} \tag{7}$$

Method (K)	SSIM ↑	CSIM ↑	FID ↓
CainGAN w/o T (1)	**0.69**	**0.85**	36.26
CainGAN w/o I (1)	0.68	0.84	46.44
CainGAN (1)	**0.69**	**0.85**	**35**
CainGAN w/o T (8)	0.72	0.87	38.08
CainGAN w/o I (8)	0.76	**0.91**	28.72
CainGAN w/o R (8)	0.75	0.90	30.05
CainGAN (8)	**0.77**	**0.91**	**24.92**

Table 2: Ablation study on selection of VoxCeleb2 dataset. All models were trained for 30 epochs, the best result between epochs 20 and 30 was reported. K is the number of source frames.

This variant is not applicable for $K = 1$ as in this case the two expressions yield the same result.

We can observe that all components are essential to obtain the best results. Using targeted embedder has a greater influence in the $K = 8$ setting, which is expected since more images can benefit from early alignment.

5 Conclusions and future work

We introduced a new method for synthesizing images in novel poses while preserving the identity of a given subject. CainGAN uses spatially adaptive normalization with a proper combining function of spatial feature maps in the embedding space. Experimental results show that CainGAN behaves better on scarce training sets compared to other methods. Furthermore, realistic images can be generated without the need for fine-tuning. The ablation study demonstrates CainGAN's non-redundant structure whereas the difference between scores in the $K = 1$ and $K = 8$ settings illustrate the ability to capitalize on more source images when available. Further development directions include designing a method to extend the applicability of CainGAN beyond the task of self-reenactment and closing the gap between one-shot and multi-shot results by improving on single source image generation.

References

[1] Albuquerque, I.; Monteiro, J.; Doan T.; Considine B.; Falk T.; Mitliagkas I. (2019). Multi-objective training of Generative Adversarial Networks with multiple discriminators, *arXiv preprint arXiv:1901.08680*, 2019.

[2] Blanz, V.; Vetter, T. (1999). A Morphable Model for the Synthesis of 3D Faces, *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, 187–194, 1999.

[3] Bulat, A.; Tzimiropoulos, G. (2017). How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks), *International Conference on Computer Vision*, 2017.

[4] Chen, L.; Li, Z.; Maddox, R.K.; Duan, Z.; Xu, C. (2018). Lip movements generation at a glance, *Proceedings of the European Conference on Computer Vision (ECCV)*, 520–535, 2018.

[5] Chen, L.; Zheng, H.; Maddox, R.K.; Duan, Z.; Xu, C. (2019). Sound to Visual: Hierarchical Cross-Modal Talking Face Video Generation, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops*, 2019.

[6] Chen, T.; Lucic, M.; Houlsby, N.; Gelly, S. (2019). On Self Modulation for Generative Adversarial Networks, *International Conference on Learning Representations*, 2019.

[7] Chung, J. S.; Nagrani, A.; Zisserman, A. (2018). VoxCeleb2: Deep Speaker Recognition, *INTERSPEECH*, 2018

[8] Durugkar I. P.; Gemp, I.; Mahadevan, S. (2016). Generative Multi-Adversarial Networks, *arXiv preprint arXiv:1611.01673*, 2016.

[9] Finn, C.; Abbeel, P.; Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks, *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1126-1135, 2017.

[10] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. (2014). Generative Adversarial Nets, *Advances in Neural Information Processing Systems 27*, 2672–2680, 2014.

[11] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium, *Advances in Neural Information Processing Systems*, 6626–6637, 2017.

[12] Huang, X.; Belongie, S. (2017). Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization, *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510, 2017.

[13] Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. (2017). Image-to-Image Translation with Conditional Adversarial Networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134, 2017.

[14] Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation, *International Conference on Learning Representations*, 2018.

[15] Karras, T.; Laine, S.; Aila, T. (2018). A Style-Based Generator Architecture for Generative Adversarial Networks, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4396–4405, 2018.

[16] Lim, J.H.; Ye, J.C. (2017), Geometric gan, *arXiv preprint arXiv:1705.02894*, 2017.

[17] Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. (2018). Spectral Normalization for Generative Adversarial Networks, *arXiv preprint arXiv:1802.05957*, 2018.

[18] Nguyen, T.; Le, T.; Vu, H.; Phung, D. (2017). Dual Discriminator Generative Adversarial Nets, *Advances in Neural Information Processing Systems*, 2670–2680, 2017.

[19] Park, T.; Liu M.; Wang T.C.; Zhu, J.Y. (2019), Semantic Image Synthesis with Spatially-Adaptive Normalization, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2337–2346, 2019.

[20] Simonyan, K.; Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*, 2014.

[21] Song, Y.; Zhu, J.; Li, D.; Wang, A.; Qi, H. (2019). Talking Face Generation by Conditional Recurrent Adversarial Network, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 919–925, 2019.

[22] Suwajanakorn, S; Seitz, S.; Kemelmacher, I. (2017). Synthesizing Obama: learning lip sync from audio, *ACM Transactions on Graphics*, 36, 1–13, 2017.

[23] Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. (2018). Face2Face: Real-time face capture and reenactment of RGB videos, *Communications of the ACM*, 62, 96–104, 2018.

[24] Ulyanov, D.; Vedaldi, A.; Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization, *arXiv preprint arXiv:1607.08022*, 2016.

[25] Wang T.C.; Liu M.Y.; Zhu J.Y.; Tao A.; Kautz J.; Catanzaro B. (2018). High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8798–8807, 2018.

[26] Wiles, O.; Koepke, A.S.; Zisserman, A. (2018). X2Face: A network for controlling face generation, *European Conference on Computer Vision* 670–686, 2018.

[27] Yuan, X.; Park, I.K., (2019). Face De-occlusion using 3D Morphable Model and Generative Adversarial Network, *Proceedings of the IEEE International Conference on Computer Vision*, 10062–10071, 2019.

[28] Zakharov, E.; Shysheya, A.; Burkov, E.; Lempitsky, V. (2019). Few-Shot Adversarial Learning of Realistic Neural Talking Head Models, *arXiv preprint arXiv:1905.08233*, 2019.

[29] Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. (2018), Self-Attention Generative Adversarial Networks, *arXiv preprint arXiv:1805.08318*, 2018

[30] Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; Wang, X. (2019). Talking Face Generation by Adversarially Disentangled Audio-Visual Representation, *AAAI Conference on Artificial Intelligence*, 33, 9299–9306, 2019

Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:
Ardelean, A. T.; Sasu, L. M. (2020). Pose Manipulation with Identity Preservation, *International Journal of Computers Communications & Control*, 15(2), 3862, 2020.
https://doi.org/10.15837/ijccc.2020.2.3862