The Effect of Latent Space Dimension on the Quality of Synthesized Human Face Images

Ivana Marin, Sven Gotovac, Mladen Russo, and Dunja Božić-Štulić

***Abstract*—In recent years Generative Adversarial Networks (GANs) have achieved remarkable results in the task of realistic image synthesis. Despite their continued success and advances, there still lacks a thorough understanding of how precisely GANs map random latent vectors to realistic-looking images and how the priors set on the latent space affect the learned mapping. In this work, we analyze the effect of the chosen latent dimension on the final quality of synthesized images of human faces and learned data representations. We show that GANs can generate images plausibly even with latent dimensions significantly smaller than the standard dimensions like 100 or 512. Although one might expect that larger latent dimensions encourage the generation of more diverse and enhanced quality images, we show that an increase of latent dimension after some point does not lead to visible improvements in perceptual image quality nor in quantitative estimates of its generalization abilities.**

***Index Terms*—Generative Adversarial Networks, Latent space exploration, Latent dimension, Evaluation, Fréchet Inception Distance (FID), Image synthesis.**

I. Introduction

SINCE 2014., when Ian Goodfellow and his colleagues at the University of Montreal introduced them, *Generative Adversarial Networks* (*GANs*) [1] have achieved outstanding results in diverse data generation tasks. In particular, they have been successfully utilized for purposes such as photorealistic image synthesis [2]–[5], single image super-resolution [6], image-to-image translation [7, 8], image inpainting [9, 10], face swapping and reenactment [11], text-to-image translation [12], speech synthesis [13] and video generation [14, 15].

GAN's learning is set up as an adversarial game between two players, *generator* G and *discriminator* D, each represented with one neural network. Let $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_d$ denote the parameters of the generator and discriminator network, respectively. The generator network transforms latent noise vectors $\boldsymbol{z} \sim p_z$ into new data points $\boldsymbol{x}^* = G(\boldsymbol{z}; \boldsymbol{\theta}_g)$ from high-dimensional data distribution p_g of generated values. Its goal is to produce fake data indistinguishable from the real data, i.e., to achieve $p_g \approx p_{data}$ where p_{data} denotes the real data distribution. On the other hand, the discriminator network aims to distinguish fake data from real data; classify both real and generated data correctly. Discriminator as input takes $\boldsymbol{x}$ that can be either generated or real data from the training set and outputs the probability $D(\boldsymbol{x}; \boldsymbol{\theta}_d)$ that $\boldsymbol{x}$ is real data. Thus, during adversarial training, the discriminator maximizes

$$\mathbb{E}_{\boldsymbol{x} \sim p_{data}} \log D(\boldsymbol{x}; \boldsymbol{\theta}_d) + \mathbb{E}_{\boldsymbol{x}^* \sim p_g} \log \left(1 - D(\boldsymbol{x}^*; \boldsymbol{\theta}_d)\right) \quad (1)$$

while the generator simultaneously tries to minimize it. It is yet unclear what kind of data distribution the generator ends up learning. An empirical study was conducted in [16] to gain insight into a GAN's capability to learn the targeted data distribution. Experiments suggest they can fall short in learning the targeted data distribution and often suffer from *mode collapse*, i.e., a low amount of diverse samples generatable from the learned mapping.

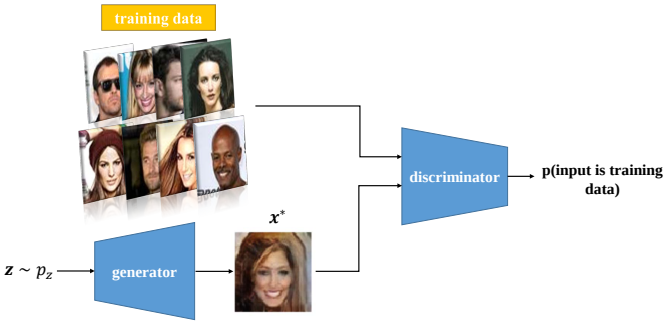

Fig. 1. Illustration of GAN for face image synthesis [17].

This work focuses on employing generative adversarial networks for face image synthesis in an unsupervised manner; more precisely, the focus is set on the *black-box* models in which the learned data distribution is not explicitly known, but we can sample from it. Through adversarial training, the generator models data distribution p_g by learning a non-linear mapping $G : \mathcal{Z} \rightarrow \mathcal{X}$ from n-dimensional latent space $\mathcal{Z} \subseteq \mathbb{R}^n$ to the real image space $\mathcal{X}$ that enables generation of new photorealistic images of human faces from random latent (noise) vectors $z \in \mathcal{Z}$.

The relation of semantic concepts shown on generated images and latent codes of different generative models has been investigated in many works. Shen et al. in [18] explore the GANs' latent space encodings of varying semantics for face image synthesis and leverage interpreted semantics together with specific manipulation techniques to control facial at-

Manuscript received February 12, 2021; revised April 2, 2021. Date of publication May 24, 2021. Date of current version May 24, 2021. The associate editor prof. Pascal Lorenz has been coordinating the review of this manuscript and approved it for publication.

I. Marin is with the Faculty of Science, University of Split, Croatia (e-mail: Ivana.Marin@pmfst.hr). S. Gotovac, M. Russo and Dunja Božić-Štulić are with the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Croatia (e-mails: {Sven.Gotovac, Mladen.Russo, Dunja.Gotovac}@fesb.hr).

Part of this work was presented at the International Conference on Software, Telecommunications and Computer Networks, SoftCOM '20, Hvar, Croatia, September 17-19, 2020.

Digital Object Identifier (DOI): 10.24138/jcomss-2021-0035

tributes on synthesized images. Semantically meaningful transformations of images generated from latent vectors acquired with arithmetic operations in latent space are observed in [19, 20]. In [21], the authors tackle the problem of disentangling *Variational Autoencoder's (VAE)* latent space by dividing it into class relevant and irrelevant dimensions. Bojanowski et al. in [22] introduce the *Generative Latent Optimization (GLO)* framework in which the latent vectors are learned freely in a non-parametric manner.

The problem of modeling adequate latent space remains an open subject. In [23], the authors analyze different parametric distribution priors for the GANs' latent space, such as Gaussian, uniform, gamma and Cauchy, and propose a way to obtain appropriate non-parametric priors. However, in practice, latent space elements are usually sampled using simple distribution priors as multivariate Gaussian $\mathcal{N}(\mathbf{0}, \mathrm{I})$ [2, 3] or uniform [19, 24] prior. The latent space dimension's impact on generative models' final performance is rarely discussed and somewhat arbitrarily chosen. In [25], the authors analyze the chosen autoencoder latent space dimension's effect on its final performance. The latent dimension of GAN is typically set to 100 [19, 24], but in literature, other dimensions are also used, e.g., NVIDIA researchers use 512-dimensional vectors in [3]–[5], [26] uses 64-dimensional latent vectors, while the default latent dimension in [2] is set to 128.

To examine the latent space dimension's effect on the data distribution learned by the generator, we need an efficient and objective way to evaluate multiple GAN models. Evaluation of GANs often relies on visual inspection of the perceptual quality of generated images and related learned representations that requires human intervention and is thus very time-consuming. Despite the significant progress in generative image modeling, the field still lacks quantitative evaluation measure that 1) is easy to calculate 2) encodes desired properties such as quality and diversity of generated images 3) detects overfitting, i.e., "punishes" pure memorization of the appropriate subset of the training data.

This paper extends the conference paper [17], which gives an empirical comparison of evaluation techniques commonly used to evaluate GANs. While work from [17] focuses on evaluating GAN models trained exclusively with conventional 100-dimensional latent vectors, the current paper quantitatively and qualitatively evaluates data distributions learned using various latent dimensions. We show that the GAN model can learn semantically meaningful mappings from latent to image space even when the latent dimension is reduced from the usual 100 to a value such as 16. However, an increase of latent size from the standard value 100 did not lead to visible improvements in the quality of generated samples nor the model's generalization ability.

The rest of the paper is structured as follows. Section II describes methods used to quantitatively and qualitatively evaluate models trained using different latent dimensions. Section III describes the dataset used to train GANs (III-A) and gives implementation details (III-B). Obtained results are discussed in Section IV. In Section V, concluding remarks and directions for future work are given.

II. Evaluation Methods

Many different methods for GAN evaluation have been proposed so far. Visual inspection of produced images is the most intuitive way to evaluate generative models used for image synthesis. However, visual inspection of generated samples by human validators is expensive, biased towards overfitted, low diversity models and limited by the number of samples reviewable in a reasonable time. Thus, easy to calculate and objective quantitative evaluation metrics are desired for benchmarking generative models and final model selection.

A. Quantitative Evaluation

The two most commonly used quantitative evaluation metrics, *Inception Score* [27] and *Fréchet Inception Distance* [28], are both calculated using the *Inception v3 Network* [29] pre-trained on the *ImageNet* [30] dataset.

1) Inception Score: Salimans *et al.* [27] proposed the quantitative *Inception Score (IS)* metric as an automatic alternative to human validators. The idea is to capture the *quality* and *diversity* of generated images in a single number using the pre-trained Inception classifier. Inception Network's output for generated image $\boldsymbol{x}^*$ is a vector $p(y|\boldsymbol{x}^*)$ of probabilities assigned to each of 1000 ImageNet classes. The IS is calculated as

$$IS = e^{\mathbb{E}_{\boldsymbol{x}^* \sim p_g} D_{KL}\left(p(y|\boldsymbol{x}^*) \,||\, p(y)\right)}, \tag{2}$$

where $D_{KL}\big(p(y|\boldsymbol{x}^*) \,||\, p(y)\big)$ denotes the KL-divergence between conditional label distribution $p(y|\boldsymbol{x}^*)$ and marginal label distribution $p(y)$.

When $\boldsymbol{x}^*$ is high-quality, i.e., contains a clear object, then $p(y|\boldsymbol{x}^*)$ should be low entropy. High entropy of marginal distribution $p(y)$ is expected when generated images are diverse. If both desired properties are satisfied, then KL-divergence, and hence IS as well, should be large. Thus, the least favorable value of the IS is one; IS $\in [1, 1000]$ [31]. In practice, an estimate $\hat{p}(y) = \frac{1}{N}\sum_{i=1}^{N} p(y|\boldsymbol{x}_i^*)$ calculated on N generated images $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_N^*$ is used instead of real marginal distribution $p(y)$. Together with the approximation of the expected value of KL-divergence that gives $IS \approx e^{\frac{1}{N}\sum\limits_{i=1}^{N} D_{KL}\left(p(y|\boldsymbol{x}_i^*) \,||\, \hat{p}(y)\right)}$. Calculation of ten IS estimates with $N = 5000$ and reporting mean value is recommended [27].

The IS has several drawbacks: it is sensitive to small changes of Inception network's weights [31], it is unable to detect overfitting, and does not correlate well with human judgment on problems that require synthesis of images that are not similar to the ones from ImageNet [31, 32]. Despite its shortcomings, the IS is still one of the most frequently used quantitative metrics for GAN evaluation [31, 33].

2) Fréchet Inception Distance: Unlike the IS, which considers only generated data, the *Fréchet Inception Distance (FID)* [28] takes into account the similarity of generated images and real images from the training set. The idea is to embed generated and real images into a 2048-dimensional feature space using the Inception Network's coding layer to capture vision-specific features. Then, calculate Fréchet

Distance, i.e., Wasserstein-2 distance between obtained image embeddings distributions of real and generated data, which are assumed to follow multidimensional Gaussian distribution.

Let $X_g \sim \mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ and $X_r \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ be embedding distributions of generated and real data, respectively. The FID is calculated as follows:

$$FID = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \mathrm{tr}\Big(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2\left(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g\right)^{1/2}\Big), \quad (3)$$

where $\boldsymbol{\mu}_r$, $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_r$, $\boldsymbol{\Sigma}_g$ are sample estimates of given means and covariance matrices. For estimate calculation, the minimum sample size of 10000 is recommended to avoid underestimating the real FID value. Lower FID values correspond to more similar real and generated samples, and hence better GAN models. [28] showed that FID is more robust to noise and image distortions than IS. However, FID also comes with its limitations. The assumption of Gaussian image embeddings does not always hold in practice. Furthermore, "memory" GANs that only memorize a subset of the training data also achieve low FID values.

B. Qualitative Evaluation

While the quantitative approach to GAN evaluation is less subjective, it may not correspond to human perception of generated images' quality. To complement the results of quantitative evaluation and check whether overfitting happened, qualitative evaluation methods, which give better insight into the model's learned data representations and generalization ability can be used.

1) Nearest neighbors: One way to check whether overfitting happened is by displaying samples of generated images together with their nearest neighbors from the training set - training images most similar to them. The nearest neighbors can be calculated pixel-wise using some similarity measure, typically Euclidean distance. A drawback of Euclidean distance in this context is sensitivity to minor perturbations such as shifting an image for a few pixels. Consequently, GANs that generate transformed training data can pass such tests of overfitting. The previous problem can be reduced by using perceptual metrics and showing more than one nearest neighbor [34]. The alternative approach uses features, i.e., embeddings obtained by a classifier trained on a large-scale dataset, and calculates feature-wise similarity to find the nearest neighbors.

2) Latent space interpolation: The generator's ability to translate interpolated points in the latent space in semantic interpolations in the resulting images is one sign of good generalization. The most common form of latent interpolation is *linear interpolation*. Let $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{Z}$ be any two points in the latent space. Linearly-interpolated latent points are given by

$$\boldsymbol{z}_t = (1-t)\boldsymbol{z}_1 + t\boldsymbol{z}_2, \quad t \in \langle 0, 1\rangle. \quad (4)$$

We are interested in how generated images $G(\boldsymbol{z}_t)$ change as we move from $\boldsymbol{z}_1$ to $\boldsymbol{z}_2$ via a *linear path* in the latent space. Smooth, semantically meaningful transitions in produced images infer that the model has learned relevant data representations, while sharp transitions between images indicate that data memorization has occurred [19].

3) Latent space arithmetic: Vector arithmetic can be used to access points in the latent space from which new images with desired properties will be generated, e.g., we can use a linear combination $\alpha_1\boldsymbol{z}_1 + \alpha_2\boldsymbol{z}_2 + \cdots + \alpha_k\boldsymbol{z}_k$, $\alpha_i \in \mathbb{R}$ as the generator's input to produce a new image with combined semantic properties encoded in used latent vectors $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k$. If necessary, the linear combination should be scaled appropriately to prevent possible diverging from a model's prior distribution. Linear interpolation, addition and subtraction are special cases of linear combinations.

4) Disentangled representations: The quality of learned data representation can be assessed by checking the disentanglement property. Disentangled representation allocates a separate set of directions in the latent space to different semantic concepts in targeted data [35]. The presence of such semantically meaningful directions can be verified by varying appropriate components of latent vectors and observing changes in resulting images.

III. Experimental setup

The experimental study's goal is to examine the effect of the used latent space dimension on the final quality of samples generated by the baseline GAN model. Quantitative and qualitative evaluation methods mentioned in Section II were used to assess the quality of generated images. For implementation, we used *Python 3.7.6* programming language together with *TensorFlow 2.1.0.* machine learning framework and *Tensorflow.Keras* API.

A. Dataset

As training data, we use large-scale *CelebFaces Attributes (CelebA) Dataset* [36], which provides more than 200k images of celebrity faces. The dataset contains images with large variations in pose and background clutter, diverse people and rich annotations. In all experiments, we use the resized 64×64 pixel version of celebrity images. Besides its use in an unsupervised manner, such as image synthesis, the CelebA dataset can also be employed for face attribute recognition, face detection, facial part localization, and face editing tasks.

B. Model Architecture and Training Settings

For GAN models, we use *Deep Convolutional Generative Adversarial Network* (*DCGAN*) architecture from [19], with some incorporated recommendations from [37] for stable training throughout all experiments as in [17]. Detailed descriptions of generator and discriminator architectures are given in Table I and Table II, respectively. We train multiple GAN models, which differ only in the dimension of latent space.

Latent vectors $\boldsymbol{z}$ are sampled from a Gaussian rather than from a uniform distribution [37]. LeakyReLU activation function with parameter $\alpha = 0.2$ is used on all non-output layers [19, 37]. *Batch Normalization* [38] is applied before the activation function on all model layers except the generator's input and discriminator's output layer [19]. *Adam* [39] optimization algorithm with learning rate 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$ [19], and mini-batches of size 128 are used for training. All

TABLE I
GENERATOR ARCHITECTURE
$n \in \{4, 10, 16, 50, 64, 100, 128, 256, 512\}$

Generator		
Input	n-dim latent vector $\boldsymbol{z}$	
Dense	$4 \cdot 4 \cdot 1024$	LeakyReLU
RESHAPE	$(4, 4, 1024)$	
Conv2D Transpose	512 5×5 filters, stride 2	LeakyReLU
Conv2D Transpose	256 5×5 filters, stride 2	LeakyReLU
Conv2D Transpose	128 5×5 filters, stride 2	LeakyReLU
Conv2D Transpose	3 5×5 filters, stride 2	tanh
Output	$64 \times 64 \times 3$ array	
Num of params:	n-dependent	

TABLE II
DISCRIMINATOR ARCHITECTURE

Discriminator		
Input	$64 \times 64 \times 3$ array	
Conv2D	128 5×5 filters, stride 2	LeakyReLU
Conv2D	256 5×5 filters, stride 2	LeakyReLU
Conv2D	512 5×5 filters, stride 2	LeakyReLU
Conv2D	1024 5×5 filters, stride 2	LeakyReLU
Conv2D	1024 5×5 filters, stride 2	LeakyReLU
FLATTEN		
Dense	1	sigmoid
Output	probability that input is training data	
Num of params:	17 238 273	

networks' weights are initialized using Gaussian distribution $\mathcal{N}(0, 0.2^2)$ [19]. Training images are scaled between -1 and 1 to match the range of the generator's output values.

During the training, checkpoints of the current model state are saved and afterward used for final model selection. The IS is calculated as the original paper proposes, with ten estimates calculated using 5000 samples. For FID calculation, at least 10000 samples are required to avoid underestimation; we used 25000 in all experiments.

IV. RESULTS

A. Quantitative Evaluation

In [17], an empirical evaluation of the DCGAN model trained with commonly used 100-dimensional latent space was presented. Fig. 2 shows obtained results for two quantitative evaluation metrics, IS and FID. The ranking of saved checkpoints by IS and FID considerably differs. Some models ranked as the best by the IS are among the worst ones according to the FID. For example, FID values of the latest checkpoints are among greater ones, while the IS among these checkpoints finds two out of the five best models (iterations 24500 and 28500). Fig. 3 shows a random sample of 36 images generated by each of the top-5 models according to quantitative metrics' evaluation results. The same seed is used for all models, i.e., the same 36 latent vectors $\boldsymbol{z}_i \in \mathbb{R}^{100}$, $i = 1, \ldots, 36$. In this way, we can observe how synthesized

Fig. 2. Quantitative evaluation scores for $dim(z) = 100$.

images change as the learning of the DCGAN progresses and gain insight into conflicting rankings by visually examining generated samples. Visual inspection of generated images shows that FID models' ranking agrees better with human understanding of generated images' perceptual quality. Third and fourth-best ranked models by the IS suffer from a mode collapse and produce distorted images. The third-best model outputs only a small number of perceptually different images, e.g., the first two images in row one (Fig. 3) are synthesized multiple times from several different latent vectors. On the other hand, FID properly punishes these models.

The IS is bounded with $1 \leq \text{IS} \leq 1000$ (since there are 1000 classes in the ImageNet data) [31], where the highest possible value denotes the best possible score. All IS values computed on the synthesized face images are pretty low. They range from 1.660 to the highest value of 2.785 from the iteration 9000. A low IS is obtained even on the real data. Latter is the consequence of the discrepancy between face images and images from the ImageNet dataset. Although widely used for evaluating generative models when different datasets are used, including datasets with face, flower, and bedroom images, as the previous discussion suggests, the IS can be misleading in the case of face image synthesis. Achieved results empirically substantiate the allegation from [31] that IS should not be used for GAN evaluation when generated images are not similar to ImageNet data. Hence, in the following experiments, we exclusively use the FID metric for the quantitative evaluation of models.

In additional experiments, the DCGAN model is trained using the latent dimensions ranging from 4 to 512. In Fig. 4, we can see how FID values change during the training for four different latent dimensions $dim(z) \in \{4, 16, 100, 512\}$. The FID values corresponding to the smallest latent dimension $dim(z) = 4$ are almost always higher than ones from the other three dimensions and hence worse. However, some regular patterns can not be observed for the other dimensions from Fig. 4. Although there is a large difference between latent dimensions 16 and 512, their FID values intertwine during the training and $dim(z) = 16$ outperforms $dim(z) = 512$ in some iterations.

From boxplots in Fig. 5, we can see a more significant decrease in medial and first quartile values of FIDs calculated during the training for the first three latent dimensions 4,

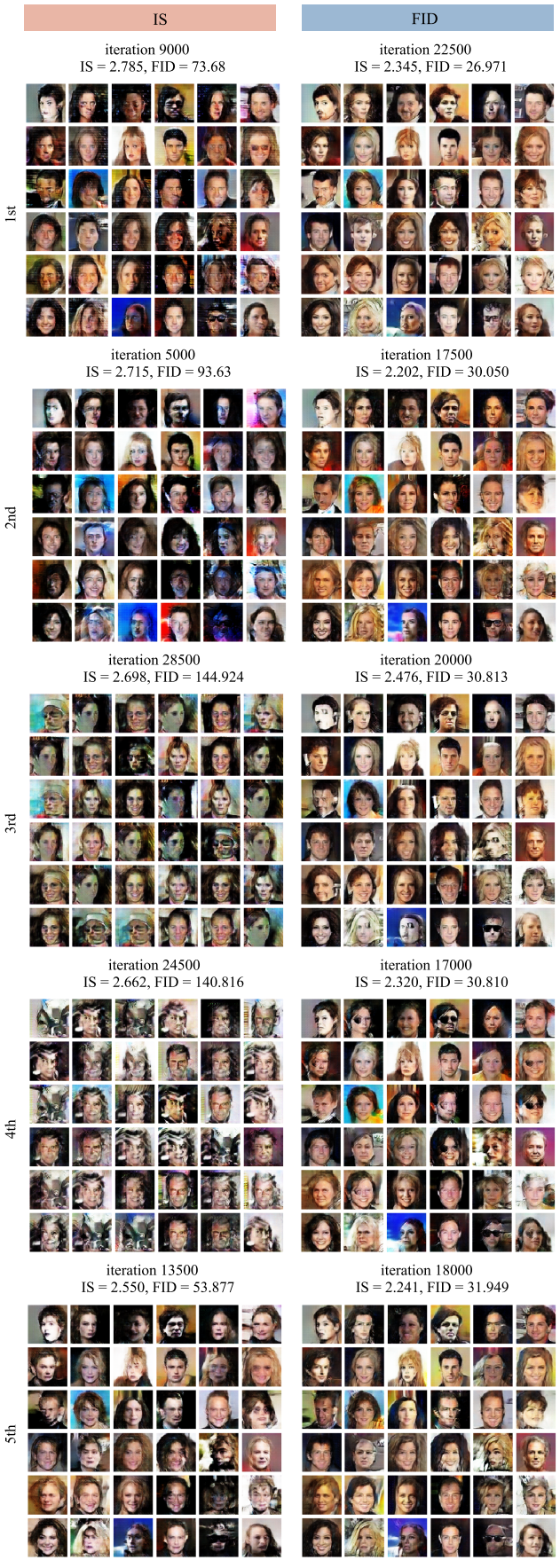

Fig. 3. Top-5 ranked models for $dim(z) = 100$.

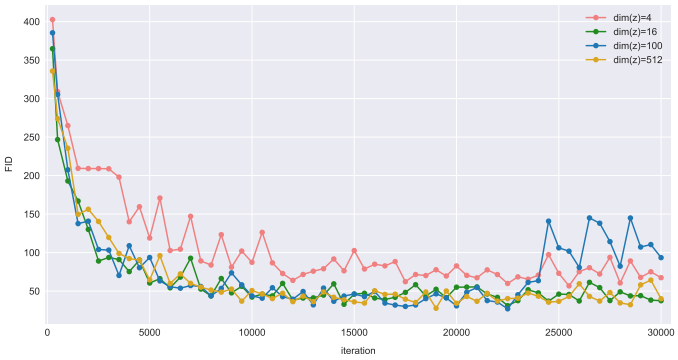

Fig. 4. FID values during the training.

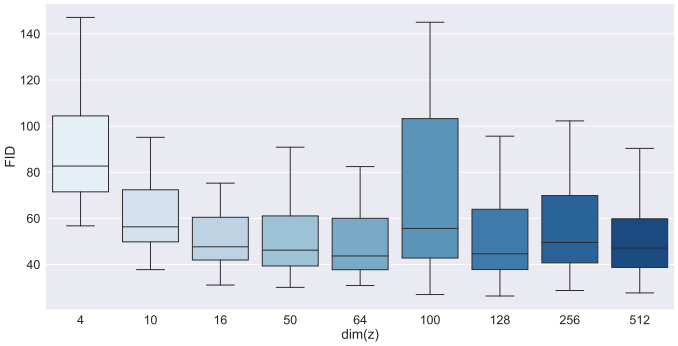

Fig. 5. Boxplots of FID values per latent dimension.

10, and 16. Further increase of latent dimension results in a slighter decrease of medial and first quartile values or even their increase. An extensive range of FID values for $dim(z) = 100$ is a consequence of the mode collapse present in the last iterations' checkpoints, consequently having greater FID values [17]. Table III shows the top-5 FID results for all latent dimensions. The best overall FID result is obtained with 128-dimensional latent space, followed by 100-dimensional space. However, the last three checkpoints of latent dimension 100 in a top-5 ranking surpass the 128-dimensional ones. From Table III and Fig. 6, we can notice how the increase of small latent dimensions at the beginning notably improves FID score, and afterward, around dimension 100, stagnation happens.

The training of GANs comes with many challenges; they are prone to different failures such as mode collapse, non-convergence, and instability during the training [40]. Such failures cause the generator to produce distorted or low-diversity images. In our experiments, the most severe failures are observed in the model with 100-dimensional latent space. Around iteration 23000, the model starts to diverge, and mode collapse happens. The latter results in high FID values in iterations towards the end of the training, even higher than the values of the 4-dimensional latent space, as can be seen in Fig 4. Mentioned contributions to the high deviation of the 100-dimensional model's FID values are noticeable on its boxplot shown in Fig. 5. However, when the best-achieved FID results from Table III are considered, the 100-dimensional model's overall performance is rated second-best. In order to fairly compare different latent dimensions and avoid false

TABLE III
TOP-5 DOCUMENTED FID RESULTS PER LATENT DIMENSION

	$dim(z)$								
	4	10	16	50	64	100	128	256	512
1.	56.749	37.772	31.084	30.070	30.873	26.971	26.335	28.717	27.671
2.	59.901	40.457	32.836	31.619	31.013	30.050	27.915	31.128	32.275
3.	60.573	41.116	36.919	31.863	31.135	30.813	32.353	32.572	34.153
4.	62.168	41.244	37.124	31.863	31.236	31.810	32.599	32.986	34.420
5.	63.903	45.313	37.498	33.572	31.935	31.949	33.825	33.593	34.655

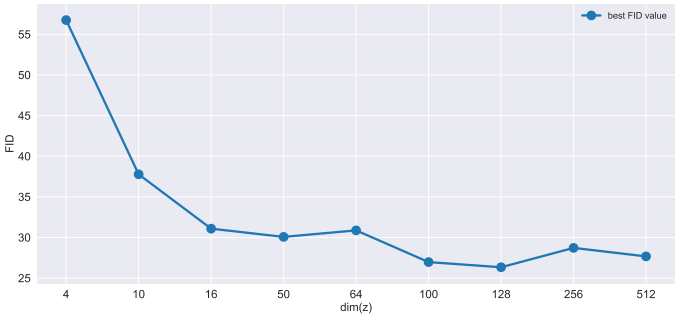

Fig. 6. Best achieved FID scores.

conclusions that can be made by considering only the final model that can suffer from one of the possible failures, for further qualitative analysis of the latent dimension effect on the quality of generated images, the checkpoint with the best, i.e., lowest, FID value is selected as the representative for a given dimension.

B. Qualitative Evaluation

To further examine the generalization ability and learned representations of the selected model, we use qualitative evaluation methods.

1) Visual Inspection: By visually examining samples of generated images, we notice that all models succeeded in generating a fair number of convincible face images. Some signs of redundancy in generated samples can be observed in all models. However, more expressive redundancy is noticed in lower-dimensional latent space models, especially in the 4-dimensional model. The 4-dimensional model generates images of just a few faces with significantly different facial attributes. For example, as shown in Fig. 7, most women have similar facial features with slight diversities such as hair color or head position. Additionally, almost all women in generated images are smiling while men are serious. Redundancies in GANs' learned data representations can uncover imbalances in the training data and can be further used for debiasing specific computer vision models.

2) Nearest Neighbors: To check whether some good-looking "fake" images are only memorized training data, i.e., whether overfitting happened, we compare them with the training images. Since manual comparison is not feasible, we use a pre-trained network to find images in the training data that are most similar to selected generated images. Fig. 8 shows five nearest neighbors (from the training data) of three chosen images per each latent dimension $dim(z) \in \{4, 16, 100, 128, 512\}$. We can notice similarities between generated images and their nearest neighbors, especially the first one. However, neither model generated images entirely identical to the training images, which means that generators succeeded in learning meaningful and generalizable data representation without severe overfitting.

Fig. 7. Sixty-four random images generated with 4-dimensional latent space. Images with similar facial attributes are similarly highlighted.

3) Linear Interpolation: Next, we explore how interpolations in latent spaces of different dimensions affect generated images. To get discrete linear interpolation with N midpoints, we calculated points $\boldsymbol{z}$ on a linear latent path as in Eq. (4) with $t \in \left\{\frac{1}{N+1}, \frac{2}{N+1}, \ldots, \frac{N}{N+1}\right\}$. On interpolated images shown in Fig. 9, we can notice smooth transitions from one synthesized face image to another with present gradual semantic changes for $dim(z) \in \{16, 100, 128, 512\}$. Here are a few examples: (i) In the first row corresponding to $dim(z) = 16$, a serious woman faced right gradually turns left and obtains a slight smile. Moreover, as we move from $\boldsymbol{z}_1$ to $\boldsymbol{z}_2$, the hair becomes more voluminous, and sunglasses start to appear on the image. (ii) If we look at the first row of $dim(z) = 100$, a smiling woman's face gradually becomes serious and then slowly turns to the right. The lack of diversity in images generated with 4-dimensional latent space is also notable in three women's

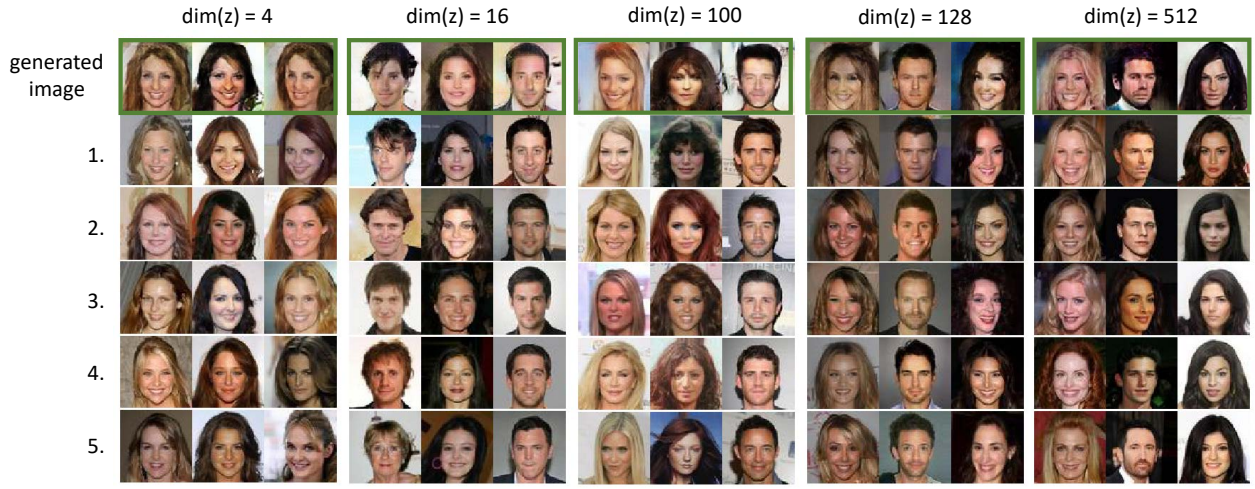

Fig. 8. Five nearest neighbors of three chosen generated images per each latent dimension $dim(z) \in \{4, 16, 100, 128, 512\}$.

images, which are interpolated in Fig. 9. In the second row, we can see the sharp transition of a dark-haired man's serious face into the blond woman's face with a wide smile. Such sharp transitions along linear paths suggest that learned data representation does not generalize well.

4) Latent Space Arithmetic: Examples in Fig. 10 show that adding and subtracting vectors in considered latent spaces result in new synthesized images that in a meaningful way combine properties of images generated from used latent vectors, i.e., we get *serious woman + serious man = smiling man* in all cases except when using latent dimension four. As already discussed, when 4-dimensional latent space is used, the generator mainly produces smiling women's images. Therefore, it was harder to find a generated image of a serious woman with satisfying perceptual quality in this case. However, the arithmetic operations result still contains combined features of the used images (longer hair, the sign of eyeglasses, darker skin tone) but not in the way we expect.

5) Disentangled Representation: In Fig. 11, we show disentangled property examination results for five latent dimensions. By moving in different directions in the latent space, we notice different semantic changes in generated images such as: adding glasses and smile to a given face $\left(d_3^{(100)}\right)$, slight head-turning to the left with the lessened smile and changes in the face attributes $\left(d_2^{(128)}\right)$, slight head-turning to the right with darker hair and more serious face expression $\left(d_2^{(512)}\right)$, darker hair and more serious face expression $\left(d_2^{(512)}\right)$, adding a wide smile to a serious face $\left(d_3^{(512)}\right)$. Desired disentanglement property of mentioned models additionally confirms that the generator has learned a meaningful mapping from latent to image space. Walk in the 4-dimensional latent space in the given directions results in images with entirely different facial attributes, which usually do not correlate visually with the originals. We also obtain almost identical new images even when moving from different starting points as in $d_2^{(4)}$ and

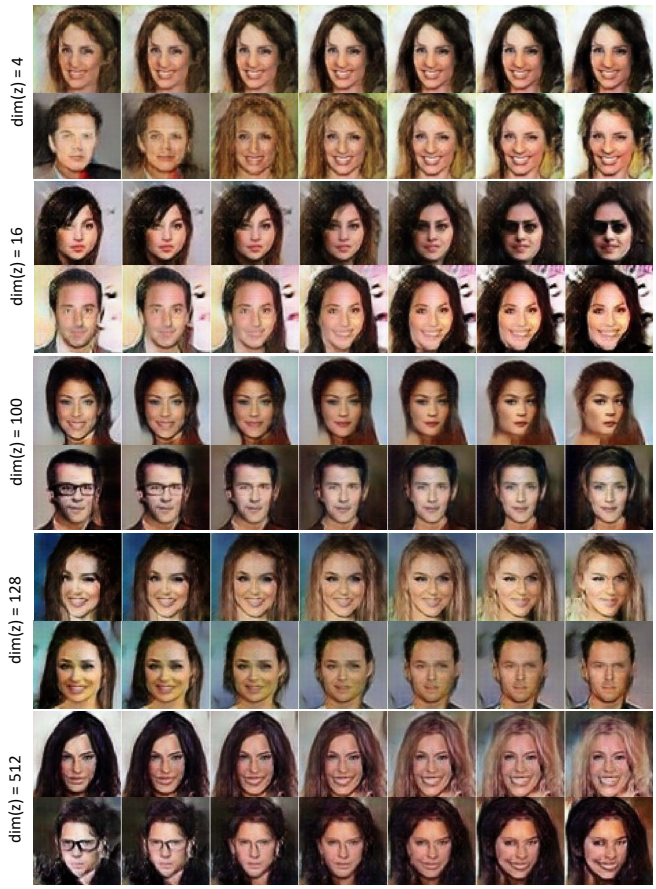

Fig. 9. Linear interpolations for $dim(z) \in \{4, 16, 100, 128, 512\}$. Every row corresponds to discrete linear latent walk from $\boldsymbol{z}_1$ to $\boldsymbol{z}_2$, with $N = 5$ equidistant midpoints. Images in the first and the last column correspond to $G(\boldsymbol{z}_1; \boldsymbol{\theta}_g)$ and $G(\boldsymbol{z}_2; \boldsymbol{\theta}_g)$.

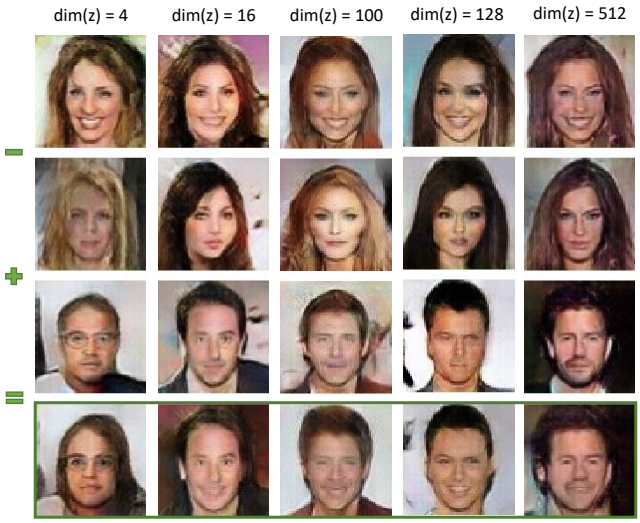

Fig. 10. Latent space arithmetic. Similar to [19], we examine what happens with generated images when we try to do the following arithmetic in the latent space *smiling woman - serious woman + serious man*.

$d_3^{(4)}$. In the 16-dimensional case, even when new images are significantly altered versions of the starting ones as in $d_1^{(16)}$, some features from the original images can still be recognized on them, such as sunglasses in the second image, hair parting (bangs), lips, and chin shape in the third image.

V. Conclusion

When it comes to setting initial priors on latent space, the choice of used latent dimension is usually reduced to selecting a commonly used value without questioning its final impact on the generative model's performance. This paper investigates the latent space dimension impact on GAN's ability to synthesize plausible and diverse face images and learn a semantically interpretable latent representation of data.

Visual inspection of the synthesized images combined with quantitative and qualitative evaluation suggests that reducing the common latent dimension 100 still enables the generative model to create new compelling face images. The increase to larger values such as 256 and 512, in our experiments, did not result in enhanced synthesis of new face images nor improved data representations. However, a significant improvement in GAN performance is captured in the initial increases of the latent dimension starting at dimension 4. After initial improvements, a further increase in the latent dimension has a milder positive effect on GAN performance until a point after which quantitative estimates show slighter degradation of learned mapping. Considering both quantitative and qualitative results, dimensions 100 and 128 seem to be the most prominent in our settings. However, experiments infer that all reasonable, not too small, latent dimensions such as standard $dim(z) \in \{100, 128, 512\}$ are a good starting point with comparable final performance and generalization ability.

The regular latent dimension 100 exhibits FID values' highest deviation. Considering its average FID ranking, it is placed as the third-worst model, while taking into account only the performance at the end of the training places it in

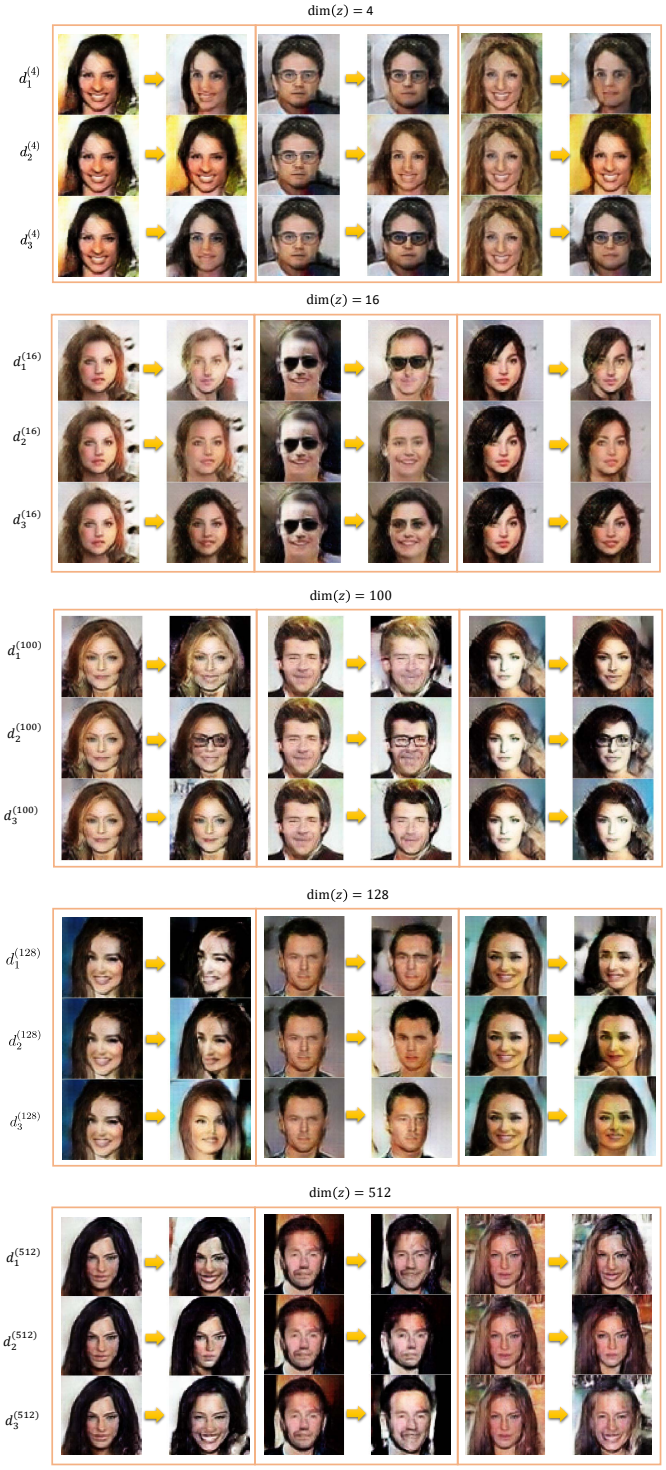

Fig. 11. Disentenglement property for $dim(z) \in \{4, 16, 100, 128, 512\}$.

the convincing last place. Still, the overall performance of the model trained with a 100-dimensional latent space place shows notable results and provides a second-best result in a single test. Because of the possible failures and instabilities in the GANs training procedure, no matter the chosen latent dimension, one should monitor the model's performance during the training and choose the best-performing model instead of the one obtained after the last training step in order to produce synthesized images of higher quality at the inference time.

In this work, we discuss the latent space dimension's effect on the quality of fixed-size generator's output ($64 \times 64 \times 3$). In the future, it would be interesting to explore the relationship of the latent space dimension, the size of synthesized images, i.e., generator's output, and final quality of generated samples; to analyze how the increases of generator's output size to values such as $128 \times 128 \times 3$ or even greater values as $512 \times 512 \times 3$ and $1024 \times 1024 \times 3$ affect the perceptual quality of synthesized images for given latent dimensions. Combined with the analysis of related computational costs, such a study could give practical guidelines for choosing appropriate combinations of latent space dimensions and image sizes.

REFERENCES

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

[2] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

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

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401–4410, doi: 10.1109/CVPR.2019.00453.

[5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119, doi: 10.1109/CVPR42600.2020.00813.

[6] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690, doi: 10.1109/CVPR.2017.19.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134, doi: 10.1109/CVPR.2017.632.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232, doi: 10.1109/ICCV.2017.244.

[9] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5485–5493, doi: 10.1109/CVPR.2017.728.

[10] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514, doi: 10.1109/CVPR.2018.00577.

[11] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 7184–7193, doi: 10.1109/ICCV.2019.00728.

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

[13] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2017, doi: 10.1109/TASLP.2017.2761547.

[14] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances in neural information processing systems*, 2016, pp. 613–621.

[15] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1526–1535, doi: 10.1109/CVPR.2018.00165.

[16] S. Arora and Y. Zhang, "Do gans actually learn the distribution? an empirical study," *arXiv preprint arXiv:1706.08224*, 2017.

[17] I. Marin, S. Gotovac, and M. Russo, "Evaluation of generative adversarial network for human face image synthesis," in *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2020, pp. 1–6, doi: 10.23919/SoftCOM50211.2020.9238203.

[18] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9240–9249, doi: 10.1109/CVPR42600.2020.00926.

[19] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[20] P. Upchurch, J. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Weinberger, "Deep feature interpolation for image content changes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, doi: 10.1109/CVPR.2017.645.

[21] Z. Zheng and L. Sun, "Disentangling latent space for vae by label relevant/irrelevant dimensions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 192–12 201, doi: 10.1109/CVPR.2019.01247.

[22] P. Bojanowski, A. Joulin, D. Lopez-Paz, and A. Szlam, "Optimizing the latent space of generative networks," *arXiv preprint arXiv:1707.05776*, 2017.

[23] R. Singh, P. Turaga, S. Jayasuriya, R. Garg, and M. W. Braun, "Nonparametric priors for generative adversarial networks," *arXiv preprint arXiv:1905.07061*, 2019.

[24] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[25] K. M. Ngoc and M. Hwang, "Finding the best k for the dimension of the latent space in autoencoders," in *International Conference on Computational Collective Intelligence*. Springer, 2020, pp. 453–464, doi: 10.1007/978-3-030-63007-2_35.

[26] D. S. Trigueros, L. Meng, and M. Hartnett, "Generating photo-realistic training data to improve face recognition accuracy," *Neural Networks*, 2020.

[27] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234–2242.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems*, 2017, pp. 6626–6637.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[30] 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, http://www.image-net.org/.

[31] S. Barratt and R. Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018.

[32] J. Yang, A. Kannan, D. Batra, and D. Parikh, "Lr-gan: Layered recursive generative adversarial networks for image generation," *arXiv preprint arXiv:1703.01560*, 2017.

[33] A. Borji, "Pros and cons of gan evaluation measures," *Computer Vision and Image Understanding*, vol. 179, pp. 41–65, 2019, doi: 10.1016/j.cviu.2018.10.009.

[34] L. Theis, A. v. d. Oord, and M. Bethge, "A note on the evaluation of generative models," *arXiv preprint arXiv:1511.01844*, 2015.

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

[36] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015, doi: 10.1109/ICCV.2015.425.
[37] S. Chintala, E. Denton, M. Arjovsky, and M. Mathieu, "How to train a gan? tips and tricks to make gans work," https://github.com/soumith/ganhacks, nips 2016.
[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[40] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

Ivana Marin received her Bachelor's degree in Mathematics from the Faculty of Science, University of Split, Croatia, in 2017, and a Master's degree in Mathematics with specialization in computing in 2019 from the same faculty. She is currently a PhD student at the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, and has been working as an assistant at the Faculty of Science, University of Split, since 2019. Her research interests include deep learning, machine learning, and data science.

Sven Gotovac was born on 07.22.1960. He graduated in 1983. at the University of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture. He received master's degree at the University of Zagreb, Faculty of Electrical Engineering in 1988 and his PhD at the TU Berlin in 1994. From 1884 he worked at the University of Split, Faculty of Electrical Engineering, Mechanical Engineering, and Naval Architecture. Currently he is full professor and lead of the Department of Computer Architecture and Operating Systems. He worked on the three national research projects, one international, and has been leader of two national and currently leader of one international project at the ALICE experiment in CERN. He was Dean of the Faculty of electrical engineering, mechanical engineering and naval architecture, University of Split from 2015. to 2020. He is co-author on about a five hundred scientific papers in indexed journals, 20 papers at international conferences and co-author of two books. He was mastering six PhDs and five master's theses (http://bib.irb.hr/lista-radova?autor=108173). He is married, father of four children. He speaks English, German and Italian.

Mladen Russo was born in Split, Croatia. He received the B.S. degree in electrical engineering from the University of Split, Croatia in 2001. He has been with the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture (FESB), University of Split, Croatia since 2001, where he received the M.S. and Ph.D. degrees in 2006 and 2010, respectively, and is currently involved as Associate Professor. His research interests include mainly signal processing and machine learning technologies, human-computer interfaces (speech recognition, emotion recognition, virtual and augmented reality technologies), wireless positioning and ambient energy harvesting. He is a member of IEEE and CCIS and associate editor of the Journal of Communications Software and Systems.

Dunja Božić-Štulić received her Bachelor's and Masters's degrees in Computer Science in 2012 and 2014, respectively, at the University of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture. She received her PhD degree in Artificial Intelligence at the same university in 2020. She is currently an assistant at the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture. Her research interests include artificial intelligence in general, but more specifically computer vision, machine learning, deep learning for natural landscape images. She is the author or co-author of 11 scientific papers on various applications of artificial intelligence.