GAGAN: Geometry-Aware Generative Adversarial Networks

Jean Kossaifi*   Linh Tran*   Yannis Panagakis*,†   Maja Pantic*

*Imperial College London

† Middlesex University London

{jean.kossaifi;linh.tran;i.panagakis;m.pantic}@imperial.ac.uk

Abstract

Deep generative models learned through adversarial training have become increasingly popular for their ability to generate naturalistic image textures. However, aside from their texture, the visual appearance of objects is significantly influenced by their shape geometry; information which is not taken into account by existing generative models. This paper introduces the Geometry-Aware Generative Adversarial Networks (GAGAN) for incorporating geometric information into the image generation process. Specifically, in GAGAN the generator samples latent variables from the probability space of a statistical shape model. By mapping the output of the generator to a canonical coordinate frame through a differentiable geometric transformation, we enforce the geometry of the objects and add an implicit connection from the prior to the generated object. Experimental results on face generation indicate that the GAGAN can generate realistic images of faces with arbitrary facial attributes such as facial expression, pose, and morphology, that are of better quality than current GAN-based methods. Our method can be used to augment any existing GAN architecture and improve the quality of the images generated.

1. Introduction

Generating images that look authentic to human observers is a longstanding problem in computer vision and graphics. Benefitting from the rapid development of deep learning methods and the easy access to large amounts of data, image generation techniques have made significant advances in recent years. In particular, Generative Adversarial Networks [14] (GANs) have become increasingly popular for their ability to generate visually pleasing results, without the need to explicitly compute probability densities over the underlying distribution.

However, GAN-based models still face many unsolved difficulties. The visual appearance of objects is not only dictated by their visual texture but also depends heavily on

Figure 1: **Samples generated by GANs trained on the CelebA [24]**. The first row shows some real images used for training. The middles rows present results obtained with popular GAN architectures, namely DCGAN [32] (row 2) and WGAN [2] (row 3). Images generated by our proposed GAGAN architecture (last row) look more realistic and the represented objects follows an imposed geometry, expressed by a given shape prior.

their shape geometry. Unfortunately, GANs do not allow to incorporate such geometric information into the image generation process. As a result, the shape of the generated visual object cannot be explicitly controlled. This significantly degenerates the visual quality of the produced images. Figure 1 demonstrates the challenges for face generation with different GAN architectures (DCGAN [32] and WGAN [2]) that have been trained on the celebA dataset [24]. Whilst GANs [14, 32] and Wasserstein GANs (WGANs) [2] generate crisp realistic objects (e.g. faces), their geometry is not followed. There have been attempts to include such information in the prior, for instance the recently proposed Boundary Equilibrium GANs (BEGAN) [4], or to learn latent codes for identities and observations [11]. However, whilst these approaches in some cases improved image generation, they still fail to explicitly model the geometry of the problem. As a result, the wealth of existing annotations for fiducial points, for example from the

facial alignment field, as well as the methods to automatically and reliably detect those [5], remain largely unused in the GAN literature.

In this paper, we address the challenge of incorporating geometric information about the objects into the image generation process. To this end, the Geometry-Aware GAN (GAGAN) is proposed in Section 3. Specifically, in GAGAN the generator samples latent variables from the probability space of a statistical shape model. By mapping the output of the generator to the coordinate frame of the mean shape through a differentiable geometric transformation, we implicitly enforce the geometry of the objects and add an implicit skip connection from the prior to the generated object. The proposed method exhibits several advantages over the available GAN-based generative models, allowing the following contributions:

- GAGAN can be easily incorporated into and improve any existing GAN architecture

- GAGAN generates morphologically-credible images using prior knowledge from the data distribution (adversarial training) and allows to control the geometry of the generated images

- GAGAN leverages domain specific information such as symmetry and local invariance in the geometry of the objects as additional prior. This allows to exactly recover the lost information inherent in generation from a small latent space

- By leveraging the structure in the problem, unlike existing approaches, GAGAN works with small datasets (less than $25,000$ images).

We assessed the performance of GAGAN in Section 4 by conducting experiments on face generation. The experimental results indicate that GAGAN produces superior results with respect to the visual quality of the images produced by existing state-of-the-art GAN-based methods. In addition, by sampling from the statistical shape model we can generate faces with arbitrary facial attributes such as facial expression, pose, and morphology.

2. Background and related work

Generative Adversarial Networks [14] approach the training of deep generative models from a game theory perspective using a minimax game. That is, GANs learn a distribution $P_G(\mathbf{x})$ that matches the real data distribution $P_{data}(\mathbf{x})$, hence their ability to generate new image instances by sampling from $P_G(\mathbf{x})$. Instead of explicitly assigning a probability to each point in the data distribution, the generator G learns a (non-linear) mapping function from a prior noise distribution $P_{\mathbf{z}}(\mathbf{z})$ to the data space as

$G(\mathbf{z}; \theta)$. This is achieved during training, where the generator G "plays" a zero-sum game against an adversarial discriminator network D. The latter aims at distinguishing between fake samples from the generator's distribution $P_G(\mathbf{x})$ and real samples from the true data distribution $P_{data}(\mathbf{x})$. For a given generator, the optimal discriminator is then $D(\mathbf{x}) = \frac{P_{data}(\mathbf{x})}{P_{data}(\mathbf{x}) + P_G(\mathbf{x})}$. Formally, the minimax game is:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim P_{data}} \Big[\log D(\mathbf{x}) \Big] + \\ \mathbb{E}_{\mathbf{z} \sim noise} \Big[\log \Big(1 - D(G(\mathbf{z})) \Big) \Big]$$

The ability to train extremely flexible generating functions, without explicitly computing likelihoods or performing inference, while targeting more mode-seeking divergences, has made GANs extremely successful in image generation [32, 29, 28, 39]. The flexibility of GANs has also enabled various extensions, for instance to support structured prediction [28, 29], to train energy based models [48] and combine adversarial loss with an information loss [6]. Additionally, GAN-based generative models have found numerous applications in computer vision, including text-to-image [33, 47], image-to-image[49, 16], style transfer [17], image super-resolution [23] and image inpainting [31].

However, most GAN formulations employ a simple input noise vector $\mathbf{z}$ without any restriction on the manner in which the generator may use this noise. As a consequence, it is impossible for the latter to disentangle the noise and $\mathbf{z}$ does not correspond to any semantic feature of the data. However, many domains naturally decompose into a set of semantically meaningful latent representations. For instance, when generating faces for the celebA dataset, it would be ideal if the model automatically chose to allocate continuous random variables to represent different factors, e.g. head pose, expression and texture. This limitation is partially addressed by recent methods [6, 26, 46, 41, 11] that are able to learn meaningful latent spaces, explaining generative factors of variation in the data. However, to the best of our knowledge, there has been no work explicitly disentangling the latent space for object geometry of GANs.

Statistical Shape Models were first introduced by Cootes et al. in [7] where the authors argue that existing methods tend to favor variability over simplicity and, in doing so, sacrifice model specificity and robustness during testing. The authors propose to remedy this by building a statistical model of the shape able to deform only to represent the object to be modeled, in a way consistent with the training samples. This model was subsequently improved upon with Active Appearance Models (AAMs) to not only model the shape of the objects but also their textures [12, 8]. AAMs operate by first building a statistical model of shape. All calculations are then done in a shape variation-free canoni-
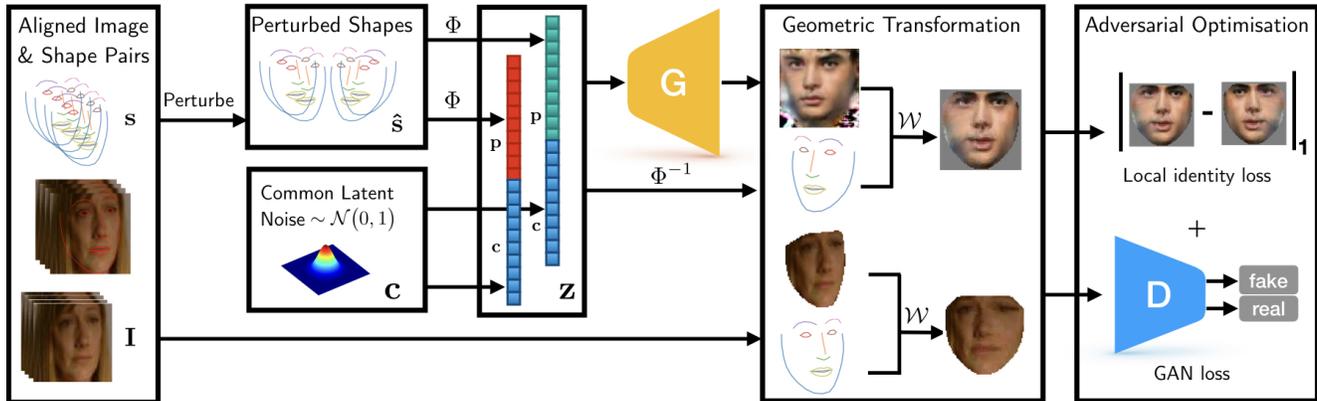
Figure 2: **Overview of our proposed GAGAN method**. (i) For each training image $\mathbf{I}$, we leverage the corresponding shape $\mathbf{s}$. Using the geometry of the object, as learned in the statistical shape model, perturbations $\hat{\mathbf{s}}_1, \cdots, \hat{\mathbf{s}}_n$ of that shape are created. (ii) These perturbed shapes are projected onto a normally distributed latent subspace using the normalised statistical shape model. That projection $\Phi(\mathbf{s})$ is concatenated with a latent component $\mathbf{c}$, shared by all perturbed versions of a same shape. (iii) The resulting vectors $\hat{\mathbf{z}}_1, \cdots, \hat{\mathbf{z}}_n$ are used as inputs to the Generator which generate fake images $\hat{\mathbf{I}}_1, \cdots, \hat{\mathbf{I}}_n$. The geometry imposed by the shape prior is enforced by a geometric transformation $\mathcal{W}$ (in this paper, a piecewise affine warping) that, given a shape $\hat{\mathbf{s}}_k$, maps the corresponding image $\hat{\mathbf{I}}_k$ onto the canonical shape. These images, thus normalised according to the shape prior, are classified by the Discriminator as fake or real. The final loss is the sum of the GAN loss and an ℓ_1 loss enforcing that images generated by perturbations of the same shape be visually similar in the canonical coordinate frame.

cal coordinate frame. The texture in that coordinate frame is expressed as a linear model of appearance. However, using row pixels as features for building the appearance model does not yield satisfactory results. Generally, the crux of successfully training such a model lies in constructing an appearance model rich and robust enough to model the variability in the data. In particular, as is the case in most applications in computer vision, changes in illumination, pose and occlusion are particularly challenging. There has been extensive efforts in the field to design features robust to these changes such as Histograms of Oriented Gradients [9] (HOG), Image Gradient Orientation kernel (IGO) [45], Local Binary Patterns [30] (LBP) or SIFT features [25]. The latter are considered the most robust for fitting AAMs [1]. Using these features, AAMs have been shown to give state-of-the-art results in facial landmarks localisation when trained on data collected in-the-wild [43, 42, 1, 21, 44]. Their generative nature make them more interpretable than discriminative approaches while they require less data than deep approaches. Lately, thanks to the democratisation of large corpora of annotated data, deep methods tend to outperform traditional approaches for areas such as facial landmarks localisation, including AAMs, and allow learning the features end-to-end rather than relying on hand-crafted ones. However, the statistical shape model employed by Active Appearance Model has several advantages. By constraining the search space, the statistical shape model allows methods that leverage it to be trained with smaller dataset.

Generative by nature, it is also interpretable and as such can be used to sample new sets of points, unseen during training, that respect the morphology of the training shapes.

In this work, we depart from existing approaches and propose a new method that leverages a statistical model of shape, built in a strongly supervised way, akin to that of ASM and AAM, while retaining the advantages of GANs. We do so by imposing a shape prior on the output of the generator. We enforce the corresponding geometry on the object outputted by the generator using a differentiable geometric function that depends on the shape prior. Our method does not require complex architectures and can be used to augment any existing GAN architecture.

3. Geometry-Aware GAN

In GAGAN, we disentangle the input random noise vector $\mathbf{z}$ to enforce a geometric prior and learn a meaningful latent representation. To do so, we model the geometry of objects using a collection of fiducial points. The set of all fiducial points of a sample composes its shape. Using the set of these shapes on the training set, we first build a statistical shape model capable of compactly representing them as a set of normal distributed variables. We enforce that geometry by conditioning the output of the generator on shape parameter representation of the object. The discriminator, instead of being fed the output of the generator, sees the images mapped onto the canonical coordinate frame by a differentiable geometric transformation (*motion model*).
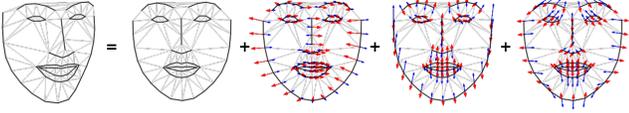
Figure 3: **Illustration of the statistical model of shape**. An arbitrary shape can be expressed as a canonical shape plus a linear combination of shape eigenvectors. These components can be further interpreted as modeling pose (components 1 and 2), smile/expression (component 3), etc.

Building the shape model Each shape, composed of m fiducial points is represented by a vector of size $2m$ of their 2D coordinates $\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2, \cdots, \mathbf{x}_m, \mathbf{y}_m$. First, similarities – that is, translation, rotation and scaling– are removed from these using Generalised Procrustes Analysis [7]. Principal Component Analysis is then applied to the similarity free shapes to obtain the mean shape $\mathbf{s}_0$ and a set of eigenvectors (the principal components) associated with the eigenvalues . The first $n-4$ eigenvectors associated with the largest eigenvalues $\lambda_1, \cdots, \lambda_n$ are kept and these compose the shape space. However, since this model was obtained on similarity free-shapes it is unable to model translation, rotation and scaling. We therefore mathematically build 4 additional components to model these similarities and append them to the model before re-orthonormalising the whole set of vectors [27]. By stacking the set of all n components as the columns of a matrix $\mathbf{S}$ of size $(2m, n)$, we obtain the shape model. A shape $\mathbf{s}$ can then be expressed as:

$$\mathbf{s} = \mathbf{s}_0 + \mathbf{S}\mathbf{p}, \tag{1}$$

We define ϕ the mapping from the shape space to the parameter space:

$$\phi\colon \mathbb{R}^{2m} \to \mathbb{R}^n$$
$$\mathbf{s} \mapsto \mathbf{S}^\top(\mathbf{s} - \mathbf{s}_0) = \mathbf{p}$$

This transformation is invertible, and its inverse, ϕ^{-1} is given by $\phi^{-1}\colon \mathbf{p} \mapsto \mathbf{s}_0 + \mathbf{S}\mathbf{S}^\top(\mathbf{s} - \mathbf{s}_0)$.

We can interpret our model from a probabilistic standpoint [10], where the shape parameters $\mathbf{p}_1, \cdots, \mathbf{p}_n$ are independent Gaussian variable with variance $\lambda_1, \cdots, \lambda_n$ and zero mean. By using the normalised shape parameters $\frac{\mathbf{p}_1}{\sqrt{\lambda_1}}, \cdots, \frac{\mathbf{p}_n}{\sqrt{\lambda_n}}$, we enforce that they be independent and normal distributed, suitable as input to our generator. This also gives us a criteria to assess how realistic a shape is using the sum of its normalised parameters $\sum_{k=1}^n \frac{\mathbf{p}_k}{\sqrt{\lambda_k}} \sim \chi^2$, which follows a Chi squared distribution [10].

Enforcing the geometric prior To constrain the output of the generator to correctly respect the geometric prior, we propose the use of the differentiable geometric function.
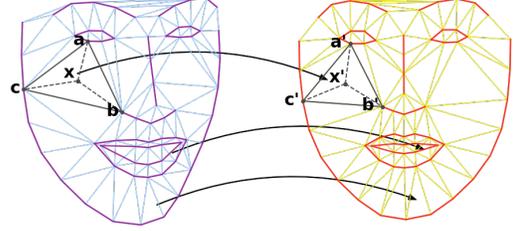

Figure 4: **Illustration of the piecewise affine warping from an arbitrary shape (left) onto the canonical shape (right)**. After the shapes have been triangulated, the points inside each of the simplices of the source shape are mapped to the corresponding simple in the target shape. Specifically, a point x is expressed in barycentric coordinates as a function of the vertices of the simplex it lays in. Using these barycentric coordinates, it is mapped onto x' in the target simplex.

Specifically, the discriminator never directly sees the output of the generator. Instead, we leverage a motion model that, given an image and a corresponding set of landmarks, maps the image onto the canonical coordinate frame. The only constraint on that motion model is that it has to be differentiable. We then backpropagate from the discriminator to the generator, through that transformation.

In this work, we use a piecewise affine warping as the motion model. The piecewise affine warping works by mapping the pixels from a source shape onto the target shape. In this work, we employ the canonical shape. This is done by first triangulating both shapes, typically as a Delaunay triangulation. The points inside each simplex of the source shape are then mapped to the corresponding triangle in the target shape, using its barycentric coordinates in terms of the vertices of that simplex, and the corresponding value is decided using the nearest neighbor or interpolation. This process is illustrated in Figure 4.

GAGAN We consider our input as a pair of N images $\mathbf{I} \in \mathbb{R}^{N \times h \times w}$ and their associated shapes (or set of fiducial points) $\mathbf{s} \in \mathbb{N}^{N \times k \times 2}$, where h and w represent height and width of a given image, and k denotes the number of fiducial points. From each shape $\mathbf{s}^{(i)}$, $i = 1, \ldots, N$, we generate K perturbed version: $\hat{\mathbf{s}}^{(i)} = (\hat{\mathbf{s}}_1^{(i)}, \ldots, \hat{\mathbf{s}}_K^{(i)})$. We denote $\hat{\mathbf{p}}^{(i)} = (\hat{\mathbf{p}}_1^{(i)}, \cdots, \hat{\mathbf{p}}_K^{(i)})$ their projection onto the normalised shape space, obtained by $\hat{\mathbf{p}}_j^{(i)} = \Phi(\mathbf{s}_j^{(i)})$, $j = 1, \cdots, K$. We model $\hat{\mathbf{p}}_j^{(i)} \sim \mathcal{N}(0, 1)$ as a set of structured latent variables which represents the geometric shape of the output objects. For simplicity, we may assume a factored distribution, given by $P(\hat{\mathbf{p}}_1^{(i)}, \ldots, \hat{\mathbf{p}}_n^{(i)}) = \prod_j P(\hat{\mathbf{p}}_j^{(i)})$, $i = 1, \cdots, N$, $j = 1, \cdots, n$.
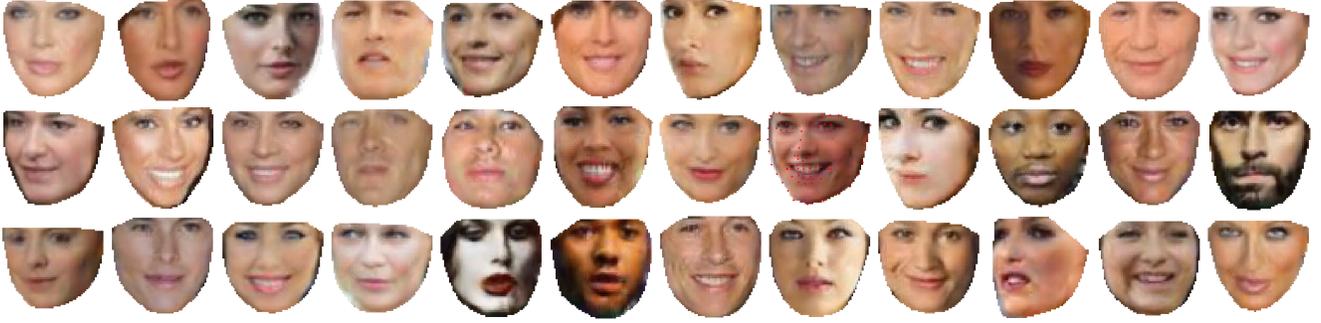
Figure 5: Random 64x64 samples from GAGAN (ours).

We now propose a method for discovering these latent variables in a supervised way: we provide the generator network G with both a the latent code vector $\hat{\mathbf{p}}_j^{(i)}$ and an associated noise vector $\mathbf{c}_j^{(i)}$, so the form of the generator becomes $G(\mathbf{c}_j^{(i)}, \hat{\mathbf{p}}_j^{(i)})$. However, in standard GAN and given a large latent space, the generator is able to ignore the additional latent code $\mathbf{p}_j^{(i)}$ by finding a solution satisfying $P_G(\mathbf{x}^{(i)}|\mathbf{p}_j^{(i)}) = P_G(\mathbf{x}^{(i)})$. To cope with the problem of trivial latent representation, we propose to employ a differentiable geometric transformation $\mathcal{W}$, also called motion model, that maps the appearance from a generated image to a canonical reference frame. In this work, we employ a piecewise affine warping and map onto the mean shape $\mathbf{s}_0$). The discriminator only sees fake and real samples after they have been mapped onto the mean shape. Discriminating between real and fake is then equivalent to jointly assessing the quality of the appearance produced as well as the accuracy of the shape parameters on the generated geometric object. The use of a piecewise affine warping has an intuitive interpretation: the better the generator follows the given geometric shape, the better the result when warping to the mean shape. For ease of notation, we denote $\hat{\mathbf{z}}^{(i)}$ the latent variable concatenating $\hat{\mathbf{p}}^{(i)}$ and $\mathbf{c}^{(i)}$, $\hat{\mathbf{z}}^{(i)} = (\hat{\mathbf{p}}^{(i)}, \mathbf{c}^{(i)})$.

We propose to solve the following affine-warping-regularized minimax game:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{I},\mathbf{s}\sim P_{data}}\Big[\log D\big(\mathcal{W}(\mathbf{I},\mathbf{s})\big)\Big]$$
$$+ \mathbb{E}_{\tilde{\mathbf{z}}\sim\mathcal{N}(0,1)}\Big[\log\Big(1 - D\big(\mathcal{W}(G(\tilde{\mathbf{z}}),\hat{\mathbf{s}})\big)\Big)\Big]$$
(2)

Local appearance preservation The generative model of shape provides us rich information about the images being generated. In particular, it is desirable for the appearance of a face to be dependent on the set of fiducial points that compose it (i.e. an infant's face has a different shape and appearance from that of an adult male or female face). However,

we also know that certain transformations should preserve appearance and identity. For instance, differences in head pose should ideally not affect appearance.

To enforce this, rather than directly feeding the training shapes to the generator, we feed it several appearance-preserving variations of each shape, and ensure that the resulting samples have similar appearance. Specifically, for each sample, we generate several variants by mirroring it, projecting it into the normalised shape space, adding random noise sampled from a Gaussian distribution there, and using these perturbed shapes as input. Since the outputs should look different (as they have different poses for instance), we cannot directly compare them. However, the geometric transformation projects these onto a canonical coordinate frame where they can be compared, allowing us to add a loss to account for these local appearance preservations.

Formally, we mirror the images and denote the corresponding mirrored shape and shape parameter are denoted by $\hat{\mathbf{s}}_M^{(i)}$ and $\mathbf{p}_{jM}^{(i)}$. The mirrored, normalised shape parameters $\hat{\mathbf{p}}_M$ are then used to build the latent space $\hat{\mathbf{z}}_M \sim \mathcal{N}(0,1)$. For simplicity, we define perturbed shapes $\tilde{\mathbf{s}} = (\hat{\mathbf{s}}, \hat{\mathbf{s}}_M)$, normalised parameters $\tilde{\mathbf{p}} = (\mathbf{p}, \mathbf{p}_M)$ and latent vectors $\tilde{\mathbf{z}} = (\hat{\mathbf{z}}, \hat{\mathbf{z}}_M)$ that share a common noise vector $\mathbf{c}$. Finally, we define the mirroring function $m(\cdot)$, that flips every image or shape horizontally. The local appearance preservation loss (LAP) is then defined as:

$$LAP = \left| \mathcal{W}\big(G(\hat{\mathbf{z}}),\hat{\mathbf{s}}\big) - \mathcal{W}\big(m\big(G(\hat{\mathbf{z}}_M)\big), m\left(\hat{\mathbf{s}}_M\right)\big) \right|_1 +$$
$$\left| \mathcal{W}\big(m\big(G(\hat{\mathbf{z}})\big), m(\hat{\mathbf{s}})\big) - \mathcal{W}\big(G(\hat{\mathbf{z}}_M),\hat{\mathbf{s}}_M\big) \right|_1$$
(3)

Adding the local appearance preservation to the minimax

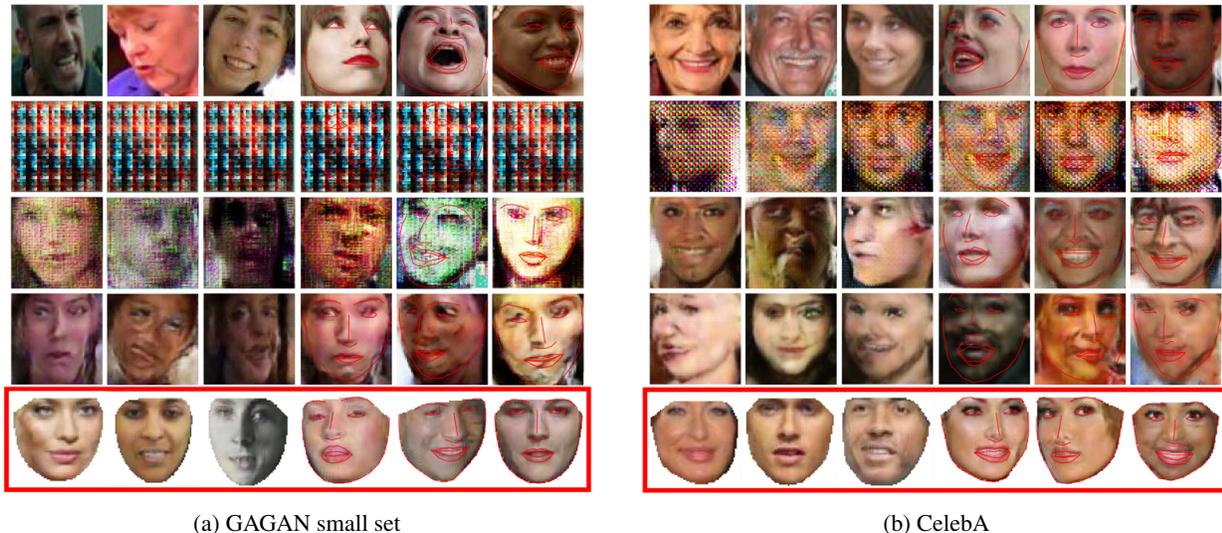(a) GAGAN small set   (b) CelebA

Figure 6: **Comparison between samples of faces generated by the baseline models and our model GAGAN for the GAGAN-small set (left column) and celebA (right column)**. The first row shows some real images. The following rows presents results obtained with our baseline models: row (2): Shape-CGAN, row (3): P-CGAN and row (4): Heatmap-CGAN. The last row present some images generated by our proposed GAGAN architecture. The first three columns show generated samples solely, while we visualize the shape prior, overlaid on the generated images, in the last three columns.

optimization value function, we get the final objective:

$$\min_{G} \max_{D} V(D,G) = \mathbb{E}_{\mathbf{I,s} \sim P_{data}}\Big[\log D\big(\mathcal{W}(\mathbf{I}, \mathbf{s})\big) \Big]$$
$$+ \mathbb{E}_{\tilde{\mathbf{z}} \sim \mathcal{N}(0,1)}\Big[\log \Big(1 - D\big(\mathcal{W}(G(\tilde{\mathbf{z}}), \tilde{\mathbf{s}})\big)\Big) \Big]$$
$$+ \lambda \cdot LAP$$

$$(4)$$

A visual overview of the method can be found in Figure 2 and Figure 5 presents samples generated with GAGAN.

4. Experimental results

In this section, we introduce the experimental setting and demonstrate the performance of the GAGAN quantitatively and qualitatively on what is arguably the most popular application for GANs, namely face generation. Experimental results indicate that the proposed method outperforms existing architectures while respecting the geometry of the faces.

4.1. Experimental setting

Datasets   To train our method, we used widely established databases for facial landmarks estimation, namely Helen [13], LFPW [3], AFW [50] and iBUG [38]. In all cases we used 68 landmarks, in the Multi-Pie configuration [15] as annotated for the 300-W challenge [37, 38]. We also used the test set of the $300-W$ challenge [34] and sampled frames from the video of the 300-VW challenge [40], as well as the videos of the AFEW-VA dataset [22]. We coin

the set of all these images and shapes the *GAGAN-small set*. To allow for comparison with other traditional GAN methods, we also used the CelebA dataset [24], which contains $202,599$ images of celebrities. Finally, to demonstrate the versatility of the method, we apply it to the cat faces dataset introduced in [35, 36]

Pre-processing   All images where processed in the following way: first the shape in the image was rescaled to a size of 60×60. The corresponding image was resized using the same factors and then cropped into a size of 64×64 so that the shape is in the center with a margin of 2 pixels on all sides. Since the celebA dataset is only annotated for 5 fiducial points, we use the recent deep learning based face alignment method introduced in [5] to detect these. This method has been shown to provide remarkable accuracy, often superior to that of humans annotators [5].

Implementation and training details   We used a standard DC-GAN architectures as defined in [32], with input images of size 64×64. The latent vector $\mathbf{z}$ of the generator has size 100 and is composed of the 50 normalised shape parameters concatenated with i.i.d. random noise sampled from a normal distribution. We trained our model using Adam [18], with a learning rate of 0.0002 for the discriminator and a learning rate of 0.001 for the generator. Model collapse has been observed with high learning rates. Reducing the learning rate was sufficient to avoid this issue.
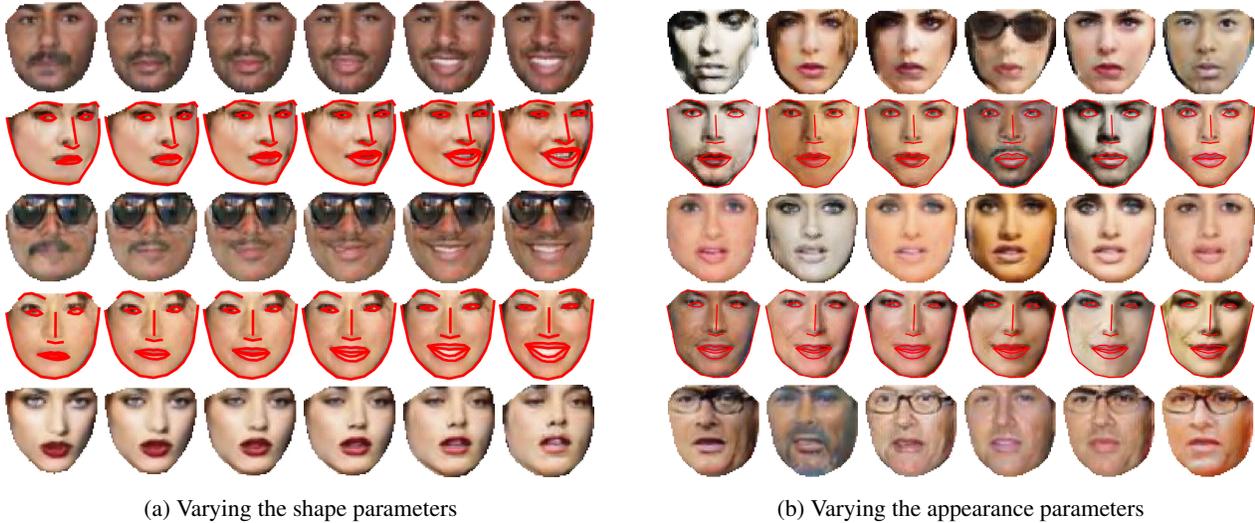
(a) Varying the shape parameters

(b) Varying the appearance parameters

Figure 7: **Images obtained by varying only the parameters from the shape prior (left) or non geometric prior (right)**. In (7a), we vary only a few shape parameters, while keeping all others fixed. The components of the statistical shape space can easily be interpreted in term of pose, morphology, smile, etc. Conversely, by varying only the non geometric latent vector **c** and keeping the shape parameters fixed (7b), we can control appearance and generate more diverse images.

We used λ in range $[0.0, 1.0, 2.5, 5.0]$. We found 2.5 to be the best regularization factor in terms of quality of generated images. All experiments were ran on a single GPU on Amazon Web Services, with an NVIDIA VOLTA GPU.

Baseline models For our baselines, we used 3 models, which the same architecture as DCGAN [32]:

Shape-CGAN is a Conditional GAN (CGAN) [28], modified to generate images conditioned on shapes **s** by channel-wise concatenation.

P-CGAN is a CGAN conditioned on the normalised shape parameters **p** (as used by GAGAN and introduced previously) by channel-wise concatenation.

Heatmap-CGAN is a novel model, based on a CGAN conditioned on shapes by heatmap concatenation. First a heatmap with value 1 at the expected position of landmarks, and 0 everywhere else is created. This is then used as an additional channel and concatenated to the image passed on to the discriminator. For the generator, the shapes are flattened and concatenated to the latent vector **z** obtained from our statistical shape model.

4.2. Qualitative results

Figure 5 shows some representative samples drawn from **z** at a resolutions of 64 x 64. We observe realistic images that closely follow the imposed shape prior, for a wide range of poses, expression, gender and illumination. Though we observed fewer older people, the proportion between men

and women sampled appears to be balanced. Interestingly, the model was able to generate accessories, such as glasses, during sampling.

We also compared the quality of images generated by GAGAN and our baseline models (Fig. 6), on the GAGAN small set and CelebA datasets. When trained on GAGAN small set, (Fig. 6a), Shape-CGAN fails to generate any meaningful image. P-CGAN, on the other hand, generates images of faces that respect the shape parameters, validating the use of such a representation. However, the generated images are highly pixelated and textures are rudimentary. Heatmap-GAN correctly generates faces according to the shapes and the textures are more realistic than P-CGAN, but the geometry is distorted. Our model, GAGAN, generates the most realistic images among all models and accurately follows the shape prior. On CelebA, generation is better for all models, including ours (Fig. 6b). As observed on the small set, the baseline models can generate meaningful images that approximately follow the shape prior, but inferior to that of GAGAN, either low quality (Shape-CGAN), or highly distorted (P-CGAN, Heatmap-CGAN). The difference in performance between the two datasets can be explained by their size, CelebA being about ten times as large as GAGAN small set. As is known, deep learning methods, including GANs, typically work best with large datasets.

4.3. Quantitative results

The facial landmark detector introduced in [5] is reported to detect fiducial points with an accuracy in most cases higher than that of human annotators. Our model takes as

input a shape prior and generates an image that respects that prior. Therefore, we propose to assess the quality of the results by running the landmark detector on the produced images and measuring the distance between the shape prior and the actual detected shape. We directly run the detector on 10,000 images created by the generator of our GAGAN, heatmap-CGAN and P-CGAN, all trained on CelebA.

Performance is evaluated in terms of the well-established normalised point-to-point error (*pt-pt-error*), as introduced in [50] and defined as the RMS error, normalised by the face-size. Following [19, 43, 50, 20, 44, 21], we produced the cumulative error distribution (CED) curve, Fig. 8. It depicts, for each value on the x-axis, the percentage of images for which the point-to-point error is lower than this value. As a baseline (*ground-truth*), we run the facial landmark detector on our challenging GAGAN-small-set, and compute the errors with the annotations provided with the data.
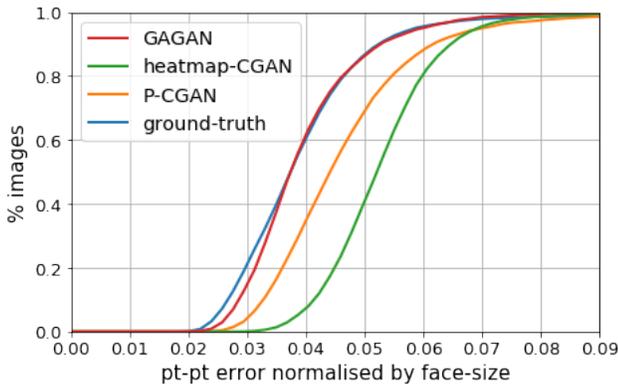

Figure 8: **Cumulative Error Distribution**. We plot the error between the landmarks estimated by the detector and those used as prior to generate the images, for GAGAN (red), heatmap-CGAN (green) and P-CGAN (orange), all trained on CelebA. As a baseline, we evaluate the performance of the landmark detector on our GAGAN-small set (*ground-truth*, blue).

As can be observed, the images generated by GAGAN accurately follow the given geometric prior used for generation, with an accuracy similar to that of the landmark detector. While heatmap-CGAN and P-CGAN also generate images that follow the prior, they do so with a significantly lower accuracy, which might also be due to the lesser quality of the images generated.

4.4. Generality of the model

To demonstrate the versatility of the model, we apply it to the generation of cats faces, using the dataset introduced in [35, 36]. Specifically, we used 348 images of cats, for which 48 facial landmarks were manually annotated [35], including the ears and boundaries of the face. We build a

statistical shape space as previously done for human faces and condition GAGAN on the resulting shape parameters.

We present some examples of generated images, along with the geometrical prior used for generation in Figure 9.
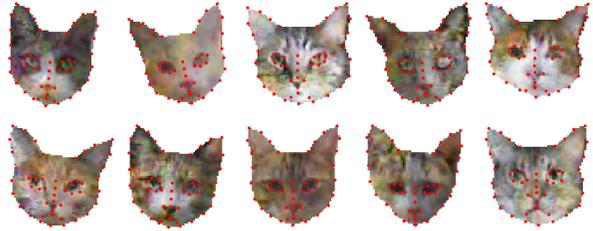

Figure 9: Samples generated by our model trained on the cats dataset, overlaid with the geometric prior used for generation (red points).

5. Conclusion and future work

We introduced GAGAN, a novel method that can be used to augment any existing GAN architecture to incorporate geometric information. Our generator samples from the probability distribution of a statistical shape model and generates faces that respect the induced geometry. This is enforced by an implicit connection from the shape parameters fed to the generator to a differentiable geometric transform applied to its output. The discriminator, being trained only on images normalised to a canonical image coordinates is able to not only discriminate on whether the produced fakes are realistic but also on whether they respect the geometry. As a result, our model is the first one, to wit, able to produce realistic images conditioned on an input shape. Going forward, we are currently working on extending our method in several ways by, i) applying it to the generation of larger images, ii) exploring more complex geometric transformations that have the potential to alleviate the deformations induced by the piecewise-affine warping and iii) augmenting traditional CNN architectures with our method for facial landmark detection.

6. Acknowledgements

References

[1] E. Antonakos, J. Alabort-i-Medina, G. Tzimiropoulos, and S. Zafeiriou. Feature-based lucas-kanade and active appearance models. *IEEE Transactions on Image Processing (TIP)*, 24(9):2617–2632, 2015. 3

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 1

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 545–552, June 2011. 6

[4] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 1

[5] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 6, 7

[6] 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 (NIPS)*, pages 2172–2180, 2016. 2

[7] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38 – 59, 1995. 2, 4

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 23, pages 681–685, 2001. 2

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, June 2005. 3

[10] R. Davies, C. Twining, and C. Taylor. *Statistical Models of Shape: Optimisation and Evaluation*. Springer Publishing Company, Incorporated, 1 edition, 2008. 4

[11] C. Donahue, Z. C. Lipton, A. Balsubramani, and J. McAuley. Semantically decomposing the latent spaces of generative adversarial networks. *arXiv preprint arXiv:1705.07904*, 2017. 1, 2

[12] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 300–305, 1998. 2

[13] J. B. F. Zhou and Z. Lin. Exemplar-based graph matching for robust facial landmark localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1025–1032, 2013. 6

[14] 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 (NIPS)*, pages 2672–2680, 2014. 1, 2

[15] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing (IVC)*, 28(5):807–813, 2010. 6

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 2

[17] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 2

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[19] J. Kossaifi, G. Tzimiropoulos, and M. Pantic. Fast newton active appearance models. In *IEEE International Conference on Image Processing (ICIP)*, pages 1420–1424, Paris, France, October 2014. 8

[20] J. Kossaifi, G. Tzimiropoulos, and M. Pantic. Fast and exact bi-directional fitting of active appearance models. In *IEEE International Conference on Image Processing (ICIP)*, pages 1135–1139, Quebec City, QC, Canada, September 2015. 8

[21] J. Kossaifi, G. Tzimiropoulos, and M. Pantic. Fast and exact newton and bidirectional fitting of active appearance models. *IEEE Transactions on Image Processing (TIP)*, 26(2):1040–1053, Feb 2017. 3, 8

[22] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65(Supplement C):23 – 36, 2017. Image and Vision Computing (IVC), Multimodal Sentiment Analysis and Mining in the Wild. 6

[23] 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. *arXiv preprint arXiv:1609.04802*, 2016. 2

[24] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 1, 6

[25] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004. 3

[26] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5040–5048, 2016. 2

[27] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision (IJCV)*, 60(2):135–164, Nov 2004. 4

[28] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2, 7

[29] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016. 2

[30] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(7):971–987, 2002. 3

[31] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 2

[32] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1, 2, 6, 7

[33] 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. 2

[34] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing (IVC), Special Issue on Facial Landmark Localisation "In-The-Wild"*, 47:3–18, 2016. 6

[35] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Robust statistical face frontalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3871–3879, Santiago, Chile, December 2015. 6, 8

[36] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Robust statistical frontalization of human and animal faces. *International Journal of Computer Vision (IJCV), Special Issue on "Machine Vision Applications"*, June 2016. 6, 8

[37] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 397–403, December 2013. 6

[38] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 896–903, 2013. 6

[39] 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 (NIPS)*, pages 2234–2242, 2016. 2

[40] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *IEEE International Conference on Computer Vision (ICCV), 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop (ICCVW'15)*, pages 50–58, December 2015. 6

[41] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 4, page 7, 2017. 2

[42] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2014. 3

[43] G. Tzimiropoulos and M. Pantic. Fast algorithms for fitting active appearance models to unconstrained images. *International Journal of Computer Vision (IJCV)*, pages 1–17, 2016. 3, 8

[44] G. Tzimiropoulos and M. Pantic. Fast algorithms for fitting active appearance models to unconstrained images. *International Journal of Computer Vision (IJCV)*, 122(1):17–33, Mar 2017. 3, 8

[45] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Subspace learning from image gradient orientations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(12):2454–2466, 2012. 3

[46] C. Wang, C. Wang, C. Xu, and D. Tao. Tag disentangled generative adversarial networks for object image re-rendering. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2901–2907, 2017. 2

[47] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*, 2016. 2

[48] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016. 2

[49] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 2

[50] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886, 2012. 6, 8