International Journal of Computer Vision (2019) 127:824–844
https://doi.org/10.1007/s11263-019-01155-7

Disentangling Geometry and Appearance with Regularised Geometry-Aware Generative Adversarial Networks

Linh Tran[1] · **Jean Kossaifi**[1] · **Yannis Panagakis**[1,2] · **Maja Pantic**[1,3]

Received: 22 February 2018 / Accepted: 29 January 2019 / Published online: 2 March 2019

Abstract
Deep generative models have significantly advanced image generation, enabling generation of visually pleasing images with realistic texture. Apart from the texture, it is the shape geometry of objects that strongly dictates their appearance. However, currently available generative models do not incorporate geometric information into the image generation process. This often yields visual objects of degenerated quality. In this work, we propose a regularized Geometry-Aware Generative Adversarial Network (GAGAN) which disentangles appearance and shape in the latent space. This regularized GAGAN enables the generation of images with both realistic texture and shape. Specifically, we condition the generator on a statistical shape prior. The prior is enforced through mapping the generated images onto a canonical coordinate frame using a differentiable geometric transformation. In addition to incorporating geometric information, this constrains the search space and increases the model's robustness. We show that our approach is versatile, able to generalise across domains (faces, sketches, hands and cats) and sample sizes (from as little as $\sim$ 200−30,000 to more than 200,000). We demonstrate superior performance through extensive quantitative and qualitative experiments in a variety of tasks and settings. Finally, we leverage our model to automatically and accurately detect errors or drifting in facial landmarks detection and tracking in-the-wild.

Keywords Generative adversarial network · Image generation · Active shape model · Disentanglement · Representation learning · Face analysis · Deep learning · Generative models · GAN

1 Introduction

The generation of realistic images is a longstanding problem in computer vision and graphics with numerous applications, including photo-editing, computer-aided design, image stylisation (Johnson et al. 2016; Zhu et al. 2017) as well as image de-noising (Vincent et al. 2008; Jain and Seung 2009), in-painting (Pathak et al. 2016; Xie et al. 2012), and super-resolution (Tipping and Bishop 2003; Yang et al. 2010; Ledig et al. 2016), to mention but a few examples. While a surge of computational, data-driven methods that rely on variational inference (Kingma and Welling 2014; Rezende et al. 2014) and autoregressive modelling (van den Oord et al. 2016; Salimans et al. 2017) have recently proposed for image generation, it is the introduction of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) that significantly advanced image generation enabling creation of imagery with realistic visual texture.

Despite their merit, GANs and their variants (Radford et al. 2015; Odena et al. 2016; Mirza and Osindero 2014) cannot adequately model sets of images with large visual variability in a fine-grained manner. Consequently, the quality of the generated images is severely affected in terms of shape and appearance. Specific to faces, visual texture (e.g., skin texture of faces, lighting) as well as pose and deformations (e.g., facial expressions, view angle) affect the appearance of a visual object. The interactions of these texture and geometric factors emulate the entangled variability,

Communicated by Xiaoou Tang.

✉ Linh Tran
linh.tran@imperial.ac.uk

Jean Kossaifi
jean.kossaifi@imperial.ac.uk

Yannis Panagakis
i.panagakis@imperial.ac.uk

Maja Pantic
m.pantic@imperial.ac.uk

[1] Imperial College London, London , UK

[2] Middlesex University London, London, UK

[3] Samsung AI, Cambridge, UK

Fig. 1 Samples of faces generated by different GAN architectures. Random samples were extracted from the CelebA (training) dataset (row 1), the original paper of the popular GAN architectures DCGAN (Radford et al. 2015) (row 2) and our implementation from WGAN (Arjovsky et al. 2017) (row 3). The last row shows images generated by the proposed GAGAN architecture

giving rise to the rich structure of visual object appearance. The vast majority of deep generative models, including GANs, do not allow to incorporate geometric information into the image generation process without explicit labels. As a result, the shape of the generated visual object cannot be controlled explicitly and the visual quality of the produced images degenerates significantly as for instance, depicted in Fig. 1. In particular, while GAN-based models (Radford et al. 2015; Arjovsky et al. 2017; Goodfellow et al. 2014) (cf. Sect. 2.1 for a brief overview) generate realistic visual texture, e.g., facial texture in this example, geometry is not precisely followed.

In this paper, we address the challenge of incorporating geometric information about the shape of visual objects into deep generative models. In particular, we introduce Geometry-Aware GAN (GAGAN) which disentangles the latent space corresponding to shape and texture by employing a statistical shape model. The statistical shape model is built based on a the wealth of existing annotations for fiducial points, and takes advantage of robust and reliable estimation for facial points detection (Bulat and Tzimiropoulos 2017). By mapping the output of the generator to the coordinate frame of a canonical shape through a differentiable geometric transformation, we strongly enforce the geometry of the objects. A visual overview of GAGAN is shown in Fig. 2.

A preliminary version of the proposed model appeared in the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Kossaifi et al. 2017). Here, we build on that model and significantly extend it, theoretically, methodologically and empirically. Specifically, we augment our preliminary model as follows:

- We propose a novel method to address the issue of dataset shift, specifically label shift and covariate shift (Quionero-Candela et al. 2009).
- We extend the model to automatically detect poor tracking and landmarks detection results.
- We extend GAGAN to generate entire images, including the actual visual object and the background.
- We demonstrate the versatility of our model in a variety of settings and demonstrate superior performance across domains, sample sizes, image sizes, and GAN network architectures.
- We demonstrate power of our model in terms of representation and generalisation by performing cross-database experiments.

In summary, the contributions of the paper are as follows:

- By encoding prior knowledge and forcing the generated images to follow a specified statistical shape prior, GAGAN generates morphological credible images.
- By leveraging domain specific information such as symmetries and local geometric invariances, GAGAN is able to disentangle the shape from the appearance of the objects.
- By employing a flexible differentiable transformation, GAGAN can be seen as a meta-algorithm and used to augment any existing GAN architecture.
- By constraining the search space using a shape model built in a strongly supervised way, GAGAN works well on very small datasets unlike existing approaches.

We describe related work and background in Sect. 2. GAGAN is introduced in Sect. 3, along with the mechanisms used for augmentation via perturbations and (α, β) regularisation. The performance of GAGAN is assessed in Sect. 4 by conducting extensive experiments on (i) human face generation, (ii) generation of sketches of human faces, (iii) generation of hands in various poses and (iv) generation of faces of cats. 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 as well respecting a given geometrical prior. 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

In this section, we review related work and background in image generation with generative models in Sect. 2.1 and statistical models of shape and their use in Sect. 2.2.

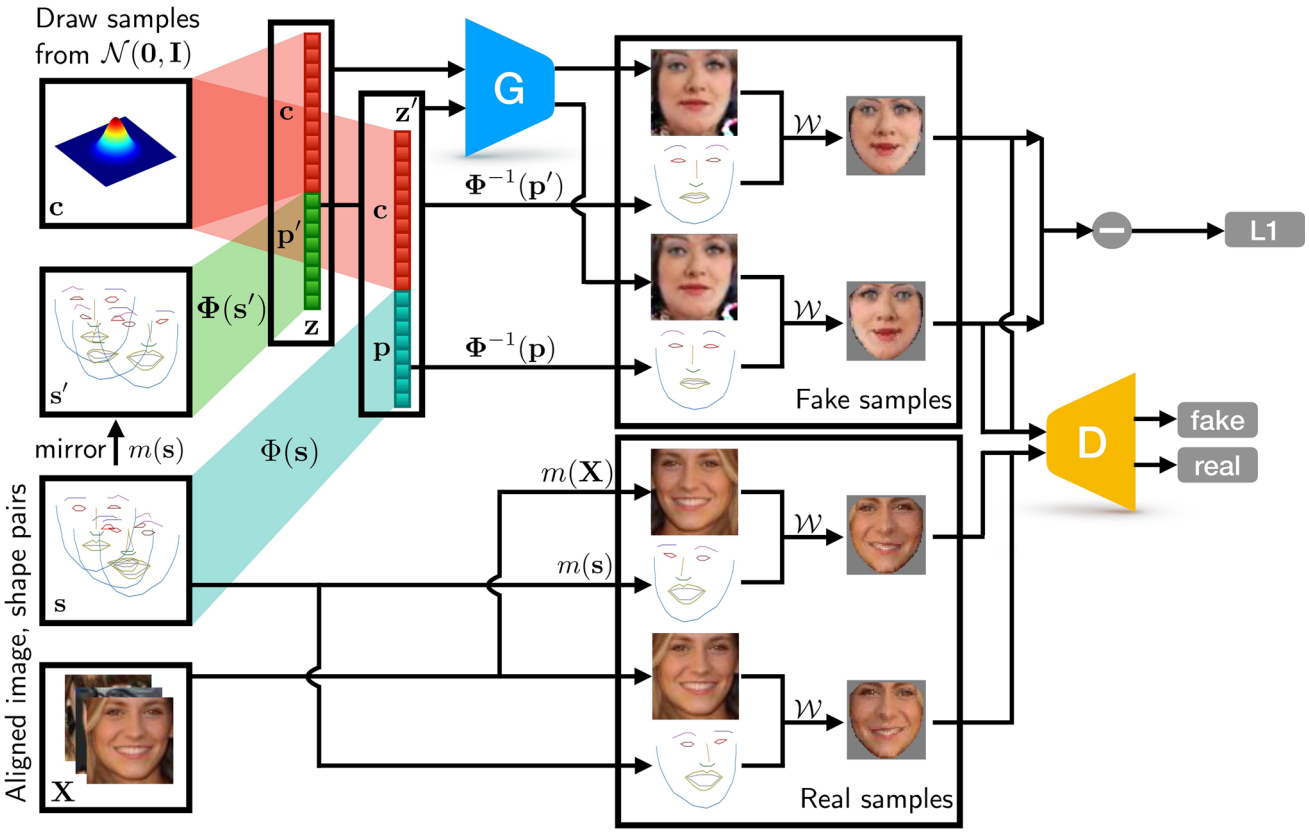


Fig. 2 Illustration of the proposed GAGAN method

2.1 Generative Models for Image Generation

Current methods for realistic image generation mainly rely on the three types of deep generative models, namely Variational Autoencoders (VAEs), autoregressive models, and Generative Adversarial Networks (GANs). Albeit different, the above mentioned deep generative models share a common setup. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ denote a set of N real images drawn from a true data distribution $P_{data}(\mathbf{x})$. Deep generative models, implicitly or explicitly, estimate a distribution $P_G(\mathbf{x}, \theta)$ by learning a non-linear mapping $G(\mathbf{z})$ parametrised with θ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The generated samples are compared to the true data distribution through a probability distance metric, e.g., Kullback–Leibler divergence or Jenson–Shannon divergence. New images are then generated by sampling from $P_G(\mathbf{z})$.

Variational Autoencoders VAEs approximate the probability distribution of the training data with a known distribution. Inference is performed by finding the parameters of the model that maximise a lower bound on the log-likelihood of the marginal distribution (Kingma and Welling 2014; Reed et al. 2016). Typically, VAEs jointly train a top-down decoder with a bottom-up encoder for inference. For images, VAE decoders model the output pixels as conditionally independent given the latent variables. This makes them straightforward to train, but results in a restrictive approximate posterior distribution (Rezende and Mohamed 2015; Kingma et al. 2016). In particular, they do not model any spatial structure in the images and fail to capture small-scale features such as texture and sharp edges, which significantly hurts both log-likelihood and quality of generated samples compared to other models (Larsen et al. 2016). Invertible density estimators were introduced by Rezende and Mohamed (2015), Kingma et al. (2016) and Dinh et al. (2017) to transform latent variables, which allows for exact log-likelihood computation and exact inference. However, the invertibility constraint is restrictive as the actual calculation of the inverse needs to be done in a computationally efficient manner.

Autoregressive Models Unlike VAEs, autoregressive models, most notably PixelCNN (van den Oord et al. 2016) and PixelCNN++ (Salimans et al. 2017), directly model the conditional probability distribution over pixels. These models are capable of capturing fine details in images and thus generate outstanding samples, but at the cost of slow sampling speed. As opposed to conventional convolutional architectures, autoregressive models do not apply down-sampling between layers, and in order capture dependencies between distant pixels, the depth of a PixelCNN grows linearly with

the size of the images. PixelCNNs also do not explicitly learn a latent representation of the data, and therefore do not allow control over the image generation.

Generative Adversarial Networks GANs (Goodfellow et al. 2014) are deep generative models that learn a distribution $P_G(\mathbf{x})$ that approximates the real data distribution $P_{data}(\mathbf{x})$ by solving a minimax optimisation problem. GANs involve two networks, namely a generator G, and a discriminator D. Instead of explicitly assigning a probability to each data point in the distribution, the generator G learns a non-linear function $G(\mathbf{z};\theta)$ from a prior noise distribution $p_{\mathbf{z}}(\mathbf{z})$ to the data space. This is achieved during training, where the generator is "playing" a zero-sum game against the adversarial discriminator network D that aims to distinguish between fake samples from the generator's distribution $P_G(\mathbf{x})$ and real samples from the true data distribution $P_{data}(\mathbf{x})$. Therefore, for a given generator, the optimal discriminator is $D(\mathbf{x}) = \frac{P_{data}(\mathbf{x})}{P_{data}(\mathbf{x})+P_G(\mathbf{x})}$. More formally, the following minimax optimisation problem is solved:

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

GANs have been extremely successful in image generation (Radford et al. 2015; Odena et al. 2016; Mirza and Osindero 2014; Salimans et al. 2016) due to their ability to learn extremely flexible generator functions, without explicitly computing (often intractable) likelihoods or performing inference. The flexibility of GANs has also enabled various extensions, for instance to support structured prediction (Odena et al. 2016; Mirza and Osindero 2014), to train energy based models (Zhao et al. 2016) and combine adversarial loss with an information loss (Chen et al. 2016). One common limitation of the available GAN-based architectures is the usage of a simple input noise vector $\mathbf{z}$ without any restrictions 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 features of the data. However, many domains naturally decompose into a set of semantically meaningful latent representation. For instance, when generating faces 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 (Chen et al. 2016; Mathieu et al. 2016; Wang et al. 2017; Tran et al. 2017) 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.

2.2 Statistical Shape Models

Statistical shape models were first introduced by Cootes et al. (1995). By exploiting a statistical model of the shape statistical shape models are able to accurately represent the object's deformations based on training data. Improved statistical shape models include Active Appearance Models (AAMs), where both the shape and texture is modelled (Edwards et al. 1998; Cootes et al. 2001). In AAMs, a statistical model of shape is built first and then the texture is described by employing a linear model of appearance in a shape variation-free canonical coordinate frame. Fitting the AAM to a new instance is then done by deforming the target image (*forward* framework) or the template (*inverse* framework) (Matthews and Baker 2004), or both simultaneously (*bidirectional* framework) (Kossaifi et al. 2015). The resulting problem can be solved analytically and effectively using Gauss–Newton optimisation (Tzimiropoulos and Pantic 2016) or second-order methods based on Newton optimisation (Kossaifi et al. 2014). However, using pixel intensities for building the appearance model does not yield satisfactory results due to their variability in the presence of illumination, pose and occlusion variations. To remedy this issue, several robust image descriptors (or features) have been proposed, including Histograms of Oriented Gradients (HOG) (Dalal and Triggs 2005), Image Gradient Orientation kernel (IGO) (Tzimiropoulos et al. 2012), Local Binary Patterns (LBP) (Ojala et al. 2002) or SIFT features (Lowe 2004). The latter are considered the most robust for fitting AAMs (Antonakos et al. 2015). Using these features, AAMs have been shown to give state-of-the-art results in facial landmarks localisation, even for in-the-wild data (Tzimiropoulos and Pantic 2016, 2014a; Antonakos et al. 2015; Kossaifi et al. 2017; Tzimiropoulos and Pantic 2017).

AAMs naturally belong to the class of generative models. As such they are more interpretable and typically require less data than their discriminative counterparts, such as deep learning-based approaches (Kossaifi et al. 2017; Tzimiropoulos and Pantic 2017; Sagonas et al. 2013a). 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 features end-to-end rather than relying on hand-crafted ones. However, the statistical shape model employed by AAMs has several advantages. In particular, by constraining the search space, it allows methods that can be trained on smaller datasets. Thanks to their generative nature, AAMs can also be used to sample new instances, unseen during training, that respect the morphology of the training shapes.

In this work, we depart from the existing approaches and propose a new method, detailed in the next section, that retains the advantages of a GAN while constraining its out-

put on statistical shape models, built in a strongly supervised way, akin to that of Active Shape Models (Cootes et al. 1995) and AAMs. To this end, we impose a shape prior on the output of the generator, hence explicitly controlling the shape of the generated object.

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. We do this by separating the shape $\mathbf{p} \in \mathbb{R}^{N\times n}$ of objects from their appearance $\mathbf{c} \in \mathbb{R}^{N\times k}$. Their concatenation $\mathbf{z} = [\mathbf{p}, \mathbf{c}]$ is used as input to the model.

We first model the geometry of N images $\mathbf{X} = \{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)}\} \in \mathbb{R}^{N\times h\times w}$ using a collection of fiducial points $\mathbf{s} = \{\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(N)}\} \in \mathbb{N}^{N\times m\times 2}$, where h and w represent height and width of a given image and m denotes the number of fiducial points. The set of all fiducial points of a sample composes its shape. Using a statistical shape model, we can compactly represent any shape $\mathbf{s}$ as a set of normal distributed variables $\mathbf{p}$ (cf. Sect. 3.1). We enforce the geometry by conditioning the output of the generator. 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*, explained in Sect. 3.2). By assuming a factorised distribution for the latent variables, we propose GAGAN, a conditional GAN to disentangle the latent space (cf. Sect. 3.4). We further extend GAGAN by perturbation-motivated data augmentation (cf. Sect. 3.5) and α, β regularisation (cf. Sect. 3.6).

3.1 Building the Shape Model

Each shape, composed of m fiducial points, is represented by a vector of size $2m$ of their 2D coordinates $\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2, \ldots, \mathbf{x}_m, \mathbf{y}_m$. First, similarities—translation, rotation and scaling—are removed using Generalised Procrustes Analysis (Cootes et al. 1995). 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) and their corresponding eigenvalues. The first $n-4$ eigenvectors associated with the largest eigenvalues $\lambda_1, \ldots, \lambda_n$ are kept and 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 these to the model before re-orthonormalising the whole set of vectors (Matthews and Baker 2004). 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.

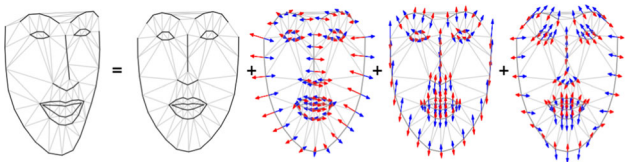


Fig. 3 Illustration of the statistical model of shape. An arbitrary shape can be expressed as a canonical shape plus a linear combination of shape eigenvectors. These components can be further interpreted as modelling pose (components 1 and 2) and smile/expression (component 3)

Given a shape $\mathbf{s}$, we can express it as:

$$\mathbf{s} = \mathbf{s}_0 + \mathbf{S}\mathbf{p}. \tag{2}$$

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

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

This transformation is invertible, and its inverse Φ^{-1} is given by $\Phi^{-1}: \mathbf{p} \mapsto \mathbf{s}_0 + \mathbf{S}\mathbf{S}^\top(\mathbf{s} - \mathbf{s}_0)$. We visualise this transformation in Fig. 3.

We can interpret our model from a probabilistic standpoint, where the shape parameters $\mathbf{p}_1, \ldots, \mathbf{p}_n$ are independent Gaussian variables with zero mean and variance $\lambda_1, \ldots, \lambda_n$ (Davies et al. 2008). By using the normalised shape parameters $\frac{\mathbf{p}_1}{\sqrt{\lambda_1}}, \ldots, \frac{\mathbf{p}_n}{\sqrt{\lambda_n}}$, we enforce them to be independent and normal distributed, suitable as input to our generator. This also gives us a criterion to assess how realistic a shape is by 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 (Davies et al. 2008).

3.2 Enforcing the Geometric Prior

To constrain the output of the generator to correctly respect the geometric prior, we propose to use a differentiable geometric function. 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. If the motion model is constrained to be differentiable, we can backpropagate from the discriminator through that transformation to the generator.

In this work, we use a piecewise affine warping as the motion model. The piecewise affine warping maps the pixels of a source shape onto a target shape. In this work, we employ the canonical shape. This is done by first triangulating both shapes, typically done as a Delaunay triangulation. An affine transformation is then used to map the points inside each simplex of the source shape to the corresponding triangle in

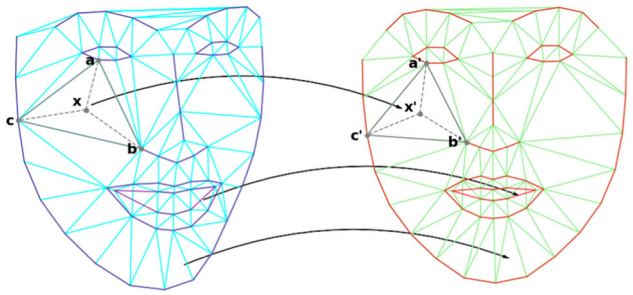


Fig. 4 Illustration of the piecewise affine warping from an arbitrary shape (left) onto the canonical shape (right). After the shapes have been triangulated, the points inside each of the simplices of the source shape are mapped to the corresponding simplex in the target shape. Specifically, a point 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

the target shape, using their barycentric coordinates in terms of the vertices of that simplex. The corresponding value is decided using the nearest neighbour or interpolation. This process is illustrated in Fig. 4.

3.3 Local Appearance Preservation

The statistical shape model provides 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., a baby's face has a different shape and appearance from that of a woman or a man. However, we also know that certain transformations should preserve appearance and identity. For instance, differences in head pose should ideally not affect appearance.

Rather than feeding directly the training shapes, we create several appearance-preserving variations of each shape, feed them to the generator, and ensure that the resulting samples have similar appearance. Consequently, for each sample we generate several variants by mirroring it, projecting it into the normalised shape space, adding random normal distributed noise sampled, and then use these perturbed shape as input. As the outputs will have different shapes and thus should look different, 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 local appearance preservations.

3.4 GAGAN

We assume our input can be described as pairs $(\mathbf{X}^{(i)}, \mathbf{s}^{(i)})$ of N images $\mathbf{X}$ with their associated shapes $\mathbf{s}$. The corresponding shape parameters are given by $\mathbf{p}^{(i)} = \Phi(\mathbf{s}^{(i)}) \in \mathbb{R}^n$, $i = 1, \ldots, N$. We model $\mathbf{p}_j^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as a set of $\mathbf{p}$ structured independent latent variables which represents the geometric shape of the output objects. For simplicity, we may assume a factored distribution, given by

$$P(\mathbf{p}_1^{(i)}, \ldots, \mathbf{p}_n^{(i)}) = \prod_j P(\mathbf{p}_j^{(i)}), \quad i = 1, \ldots, N, \quad j = 1, \ldots, n. \tag{3}$$

We propose a method for discovering these latent variables in a supervised way: the generator network G uses both noise $\mathbf{c}^{(i)}$ and the latent code $\mathbf{p}^{(i)}$, thus the form of the generator becomes $G(\mathbf{c}^{(i)}, \mathbf{p}^{(i)})$. However, in standard GANs when given a large latent space, the generator is able to ignore the additional latent code $\mathbf{p}^{(i)}$ by finding a solution satisfying $P_G(\mathbf{X}^{(i)}|\mathbf{p}^{(i)}) = P_G(\mathbf{X}^{(i)})$. To cope with the problem of trivial latent representation, we employ a differentiable geometric transformation that maps the appearance from a generated image to a canonical reference frame. We denote this function $\mathcal{W}$. This constraints $P_G(\mathbf{X}^{(i)}|\mathbf{p}^{(i)})$ to

$$P_G(\mathbf{X}^{(i)}|\mathbf{p}^{(i)}) = P_G(\mathcal{W}(\mathbf{X}^{(i)}, \Phi^{-1}(\mathbf{p}^{(i)}))|\mathbf{X}^{(i)}, \mathbf{p}^{(i)}). \tag{4}$$

In this work, we employ a piecewise affine warping which maps $\mathbf{s}$ 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 usage of a piecewise affine warping has an intuitive interpretation: The better the generator follows the given geometric shape, the better the presentation when warping to the mean shape. For ease of notation, we will use latent variable $\mathbf{z}^{(i)}$ to concatenate variables $\mathbf{p}^{(i)}$ and $\mathbf{c}^{(i)}$, i.e., $\mathbf{z}^{(i)} = (\mathbf{p}^{(i)}, \mathbf{c}^{(i)})$.

Therefore, we propose to solve the following affine-warping-regularised value function:

$$\begin{aligned} V_{\text{GAGAN}}(D, G) = \mathbb{E}_{\mathbf{X},\mathbf{s}}\Big[\log D\Big(\mathcal{W}(\mathbf{X}, \mathbf{s})\Big)\Big] \\ + \mathbb{E}_{\mathbf{z}}\Big[\log\Big(1 - D(\mathcal{W}(G(\mathbf{z}), \mathbf{s}))\Big)\Big] \end{aligned} \tag{5}$$

Additionally to Eq. 5, we add a regularisation term to preserve appearance locally. While comparing the appearance of the faces from two different images is a hard problem, our geometric transformation allows us to do so easily, by warping them onto a common canonical shape where they can directly compared, e.g. , using an ℓ_1 or ℓ_2 norm. In practice, we define $\mathbf{X}_M^{(i)}$, $i = 1, \ldots, N$ as the mirrored image of $\mathbf{X}^{(i)}$. The corresponding mirrored shape and shape parameter are denoted by $\mathbf{s}_M^{(i)}$ and $\mathbf{p}_M^{(i)}$. The mirrored shapes $\mathbf{s}_M$ and the corresponding $\mathbf{p}_M$ are used to build the entire latent space $\mathbf{z}_M \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Finally, we define $m(\cdot)$ as mirroring function, meaning it flips every image and shape horizontally. The local appearance preservation (LAP) regularisation term is defined as follows:

$$\begin{aligned}\text{LAP} = \ell 1\Big[\mathcal{W}\Big(G(\mathbf{z}), \mathbf{s}\Big), \mathcal{W}\Big(m(G(\mathbf{z}')), m(\mathbf{s}')\Big)\Big] \\ + \ell 1\Big[\mathcal{W}\Big(G(m(\mathbf{z}), \mathbf{s}'\Big), \mathcal{W}\Big(G(\mathbf{z}'), \mathbf{s}'\Big)\Big].\end{aligned} \tag{6}$$

We can similarly maintain appearance for local (minimal) shape variations. Adding the local appearance preservation to the minimax optimisation value function, we end up with the following objective for the GAGAN min–max value function:

$$\begin{aligned}V_{\text{GAGAN}}(D, G) = \mathbb{E}_{\mathbf{X},\mathbf{s}}\Big[\log D\Big(\mathcal{W}(\mathbf{X}, \mathbf{s})\Big)\Big] \\ + \mathbb{E}_{\mathbf{z}}\Big[\log\Big(1 - D(\mathcal{W}(G(\mathbf{z}), \mathbf{s}))\Big)\Big] \\ + \lambda \cdot \text{LAP}\end{aligned} \tag{7}$$

A visual overview of the method is shown in Fig. 2.

3.5 Data Augmentation with Perturbations

In order to provide more variety in shapes and avoid the generator learning to produce only faces for shape priors it has seen, we augment the set of training shapes by adding large amount of random small perturbations to these. These are sampled from a Gaussian distribution in the normalised shape space, and projected back onto the original space, therefore enforcing their correctness according to the statistical shape model.

Specifically, we generate L perturbed versions of $\mathbf{s}^{(i)}$, namely $\tilde{\mathbf{s}}^{(j)}$, $j = 1, \ldots, L$. Each perturbed shape $\tilde{\mathbf{s}}^{(j)}$ is obtained by first projecting $\mathbf{s}^{(i)}$ onto the normalised shape space, before adding random noise sampled from a Normal distribution to them and finally projecting back:

$$\tilde{\mathbf{s}}^{(j)} = \Phi^{-1}\left(\Phi(\mathbf{s}^{(i)}) + \epsilon\right), \ \epsilon \sim \mathcal{N}(0, \gamma\mathbf{I}) \tag{8}$$

$\gamma = 0.01$ was determined experimentally. We denote $\tilde{\mathbf{p}}^{(j)}$, $j = 1, \ldots, L$ as their projection onto the normalised shape space, obtained by $\tilde{\mathbf{p}}^{(j)} = \Phi(\tilde{\mathbf{s}}^{(j)})$, $j = 1, \ldots, L$. We proceed with the perturbations as input shapes instead of $\mathbf{s}$ for Eq. 7.

3.6 (α, β) Regularisation

We aim at using the representational power of the discriminator to precisely evaluate the quality of facial landmark estimation. This requires training on certain datasets and their underlying probability distribution, and testing/evaluating on on different distributions. Due to the *in-the-wild* nature of the images, this can lead to covariate shift.

In addition, the annotations for the various datasets were obtained differently for the various datasets with sometimes large variations. For instance, most of the data used for our small set of human faces was annotated in a semi-automatic way, while for CelebA, we used a state-of-the-art facial landmarks detector. This difference in labelling leads to label shift which needs to be tackled during training.

In other words, the discriminator needs to accept a certain amount of variability coming from the difference in labelling, while retaining the ability to generalise well on new datasets. For this purpose, we introduce $\mathbf{s}_+$ as slightly perturbed shapes which should still be accepted and $\mathbf{s}_-$ as largely perturbed shape that should be recognised as fakes:

$$\mathbf{s}_+ = \Phi^{-1}(\mathbf{p} + \epsilon_+), \ \epsilon_+ \sim \mathcal{N}(\mathbf{0}, \alpha\mathbf{I}) \tag{9}$$

$$\mathbf{s}_- = \Phi^{-1}(\mathbf{p} + \epsilon_-), \ \epsilon_- \sim \mathcal{N}(\mathbf{0}, \beta\mathbf{I}) \tag{10}$$

Given that $\mathbf{X}$ and $\mathbf{s}_-$ are badly aligned for a large enough value of β, and $\mathbf{X}$ and $\mathbf{s}_+$ are still well aligned a small enough value of α (we use $\alpha \ll 1$ and $\beta \gg 1$), we can let GAGAN explicitly learn how to rank image and shape pairs by how well they match, as measured by the point-to-point error. To do so, we augment the objective function with two regularisation terms, with coefficients α and β, respectively. The final objective function for GAGAN is then:

$$\begin{aligned}V_{(\alpha,\beta)}(D, G) = V_{\text{GAGAN}}(D, G) \\ + \mathbb{E}_{\mathbf{X},\mathbf{s}}\Big[\log\big(1 - D(W(\mathbf{X}, \mathbf{s}_-))\big)\Big] \\ + \mathbb{E}_{\mathbf{X},\mathbf{s}}\Big[\log\big(D(W(\mathbf{X}, \mathbf{s}_+))\big)\Big].\end{aligned} \tag{11}$$

We evaluate the effect of the parameters α and β on the discriminator and compare GAGAN with existing models in Sect. 4.

4 Experimental Evaluation

In this section, we investigate the performance of GAGAN. In particular, we have four primary goals for our evaluation. The first goal is to show the generality of our model in terms of image domains, image sizes and GAN architecture. In Sect. 4.2, we will show that regularised GAGAN can be applied to different domain, not just faces, different image sizes and different GAN architectures. Further, we will also discuss limitations of GAGAN. Following this Subsection, we will compare regularised GAGAN against GAGAN in Sect. 4.5, specifically addressing image quality and the ability to detect badly aligned image-shape pairs. The qualitative and quantitative assessment of regularised GAGAN against state-of-the-art conditional GAN (CGAN) models are presented in Sects. 4.3 and 4.4. This includes investigating the ability to disentangle the latent space and thus generate images with control over shape and appearance. Furthermore, we quanti-

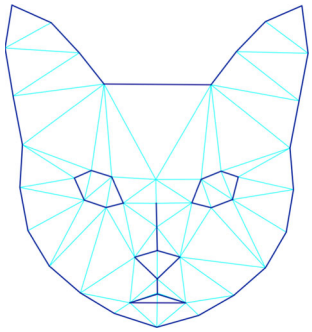


Fig. 5 Canonical shape obtained after building the shape model on the cats dataset

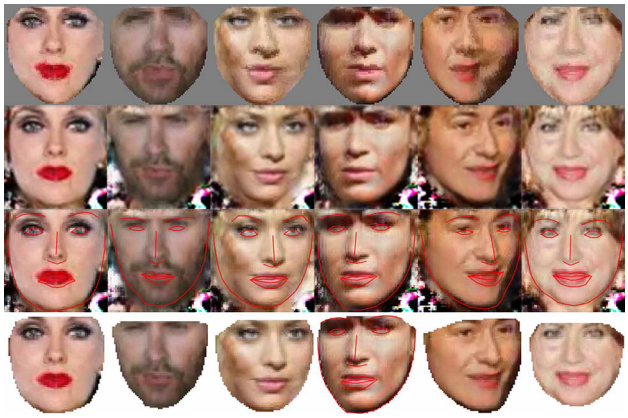


Fig. 6 Visualisation of images generated by GAGAN trained on the CelebA dataset. The discriminator never sees the background pixels as the images are mapped onto a canonical coordinate frame (first row). The second row shows the corresponding generated image folllowed by that same image superimposed by the shape prior used for generation. When displaying results obtained with GAGAN we only show pixels inside the shape used for generation (last row)

tatively assess how precise the generator can generate images with given shapes and how accurately the discriminator can discriminate when given shape and image are not aligned. In particular, we verify that, given an image and a corresponding shape, the discriminator accurately assesses how well the two corresponds. In all experiments, we compare our model with existing conditional GAN variations and GAGAN without regularisation. Finally, we show in an extensive ablation study the influence of the regularisation in Sect. 4.6.

4.1 Experimental Setting

Human Faces Datasets For human face generation, we used widely established databases for facial landmarks estimation, namely Helen (Zhou and Lin 2013), LFPW (Belhumeur et al. 2011), AFW (Zhu and Ramanan 2012) and iBUG (Sagonas et al. 2013a). In all cases we used 68 landmarks, in the Multi-Pie configuration (Gross et al. 2010) as annotated for the 300-W challenge (Sagonas et al. 2013b, a). We also used the test set of the 300-W challenge (Sagonas et al. 2016) and sampled frames from the videos of the 300-VW challenge (Shen et al. 2015), as well as from the videos of the AFEW-VA dataset (Kossaifi et al. 2017). We name 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 (Liu et al. 2015), which contains 202,599 images of celebrities. Since the CelebA dataset is only annotated for 5 fiducial points, we use the recent deep learning based face alignment method introduced in Bulat and Tzimiropoulos (2017) to detect the entire set of 68 facial landmarks. This method has been shown to provide remarkable accuracy, often superior to that of humans annotators (Bulat and Tzimiropoulos 2017). For higher resolution (128×128 and 256×256) CelebA-HQ (Liu et al. 2015) was used to be aligned and landmark tracked. The resulting dataset has 29,623 images.

Cats Dataset For the generation of faces of cats, we used the dataset introduced in Sagonas et al. (2015) and Sagonas et al. (2016). In particularly, we used 348 images of cats, for which 48 facial landmarks where manually annotated (Sagonas et al. 2015), including the ears and boundaries of the face. We first build the statistical shape space as we did previously for human faces. The resulting canonical shape is represented in Fig. 5.

Hand Gestures Dataset We used the Hand Gesture Recognition (HGR) (Grzejszczak et al. 2016; Nalepa and Kawulok 2014; Kawulok et al. 2014) which contains the gestures from Polish Sign Language ('P' in the gesture's ID) and American Sign Language ('A'). We only used the subsample of HGR which has all 25 hand feature point locations, as some annotations do only include the feature points which are visually visible. This results in a training set of 276 samples.

Sketch Dataset Finally, to demonstrate the versatility of the method, we apply GAGAN to the Face Sketch in the Wild dataset (FSW) (Yang et al. 2014) which contains 450 greyscale sketches of faces. Similarly to the face databases described above, the sketches are annotated with 68 facial landmark.

Pre-processing All images were processed in the following way. First, each shape is resized to be of size 60×60. The corresponding image is then resized by the same factor. Finally, we take a central crop of size 64×64 around the centre of the shape, leaving a margin of 2 pixels on all sides as the input image and translate the shape accordingly.

Removing Background Pixels Rather than imposing an explicit condition on the shape prior, the geometry is enforced implicitly using a differentiable geometric transformation, here a piecewise affine warping. During this process of warp-

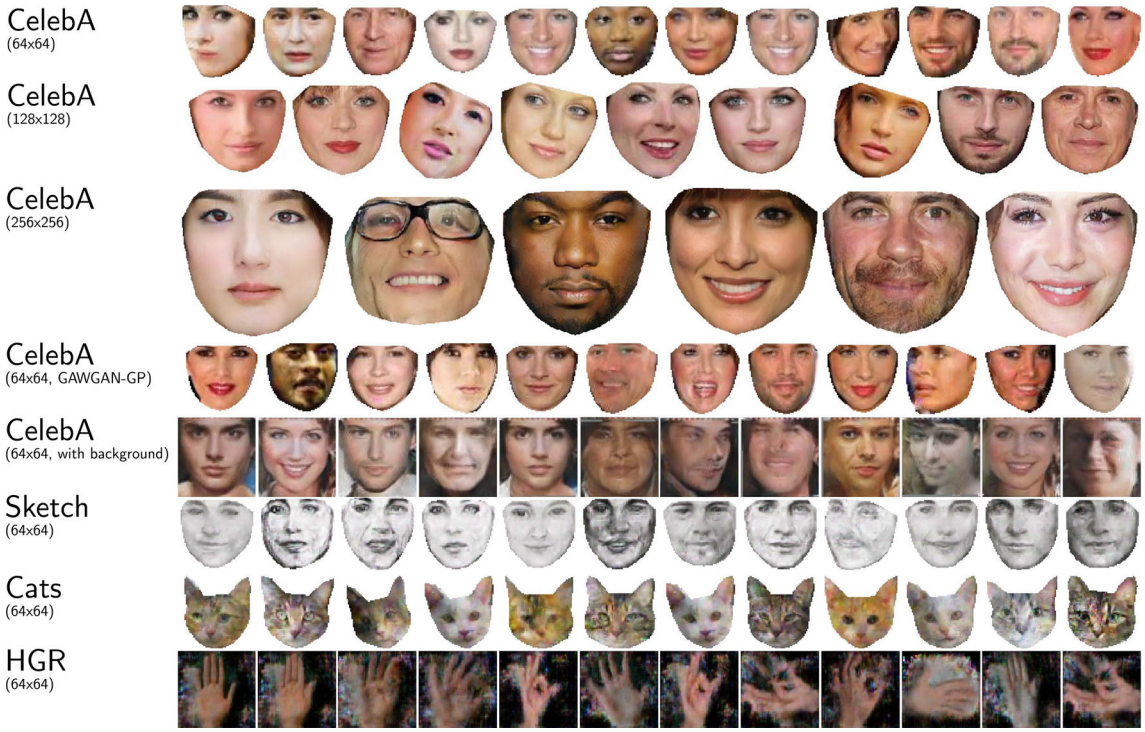


Fig. 7 Random samples from trained GAGAN models. Random CelebA samples in sizes 64 × 64 (first row), 128 × 128 (second row), 256 × 256 (third row), generated with GAWGAN-GP (fourth row) and with background (fifth row). Further samples from GAGAN (ours), obtained on face sketches (sixth row), cat faces (sixth row) and hand gesture (last row)

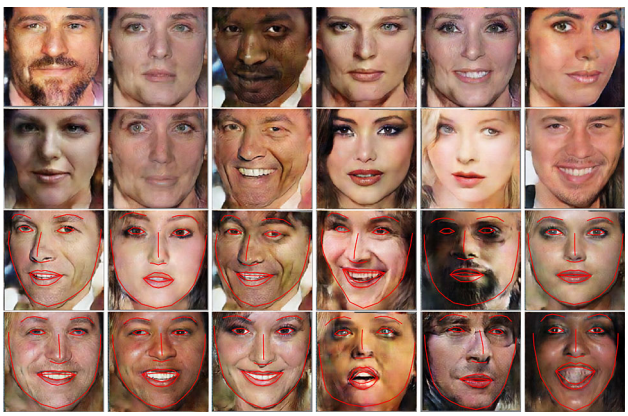
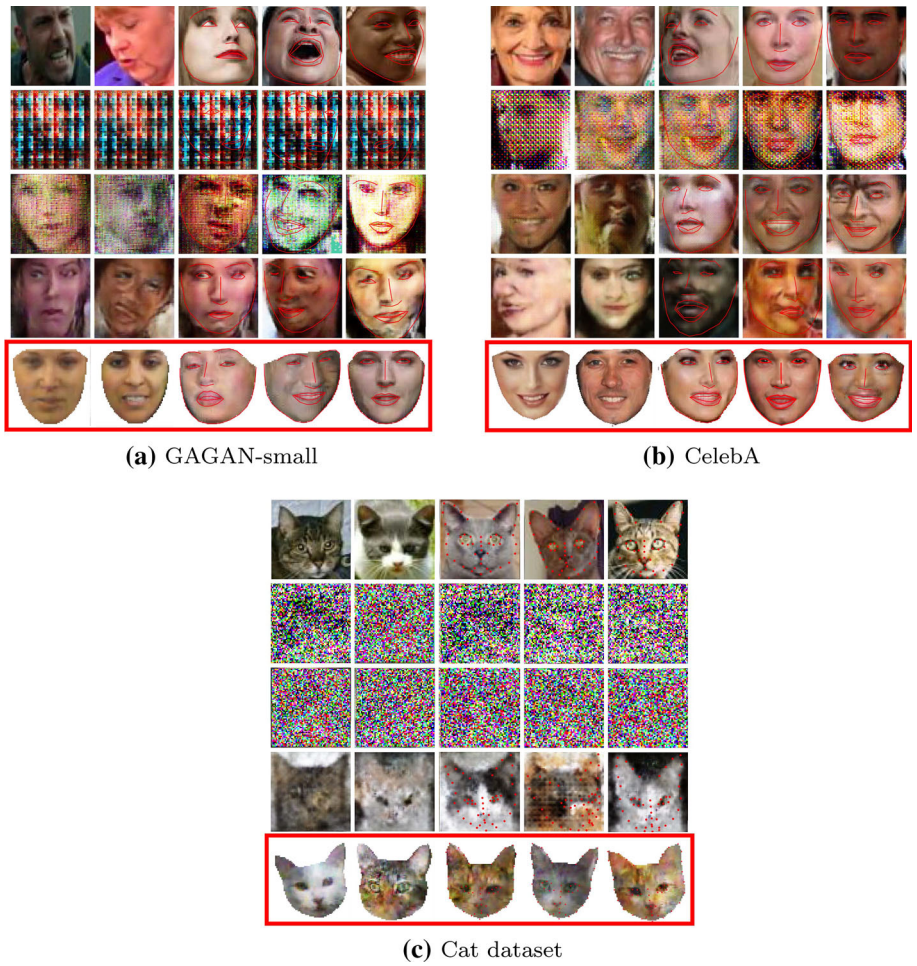


Fig. 8 Random 256 × 256 samples with background. GAGAN was trained with additional points to generate background. The last two rows have the conditional shapes overlayed

ing the input image onto a canonical coordinate frame, pixels outside the shape are discarded and the discriminator of GAGAN never sees them. Consequently no background is propagated through the networks and thus the background of the generated images is random and should be discarded. This is why we only display the points inside the shape. Specifically, we take the convex hull of the shape points and discard the pixels outside it, as illustrated in Fig. 6.

Implementation Details We use a traditional DCGAN architecture as introduced by Radford et al. (2015). The input is a dataset with either 64 × 64 pixel (rescaled) images and corresponding shapes. The latent vector $\mathbf{c}$ of the generator has size 100, is normal distributed and is concatenated with the normalised shape parameters $\mathbf{p}$. We trained our model using Adam 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 higher learning rates. Reducing the learning rate was sufficient to avoid this issue. We used values of $[0, 1.0, 2.5, 5.0]$ for λ. We found 1.0 to be the best regularisation factor in terms of quality of generated images. For α we chose values of $[0, 0.01, 0.05, 0.1]$, for β values of $[0, 0.1, 0.5, 0.8]$. By visual inspection, we excluded $\alpha > 0.1$ as then image and corresponding shape were not aligned and thus decreased quality of generation. We also ran experiments for $\beta > 0.8$, but observed no significant improvement.

Baselines and State-of-the Art Comparisons For comparison, we used a the Conditional GAN (CGAN) (Mirza and Osindero 2014), modified to generate images conditioned on the shape or shape parameters. Shape-CGAN is a CGAN conditioned on shapes by channel-wise concatenation and P-CGAN is a CGAN conditioned on the shape parameters



(a) GAGAN-small **(b)** CelebA **(c)** Cat dataset

Fig. 9 Comparison between samples of faces generated by the baseline models and our model GAGAN for **a** the GAGAN-small set, **b** CelebA and **c** cat dataset. The first row shows some real images. The following rows presents results obtained with our baseline models: shape-CGAN (row 2), P-CGAN (row 3) and Heatmap-CGAN (row 4). The last row presents images generated by our proposed GAGAN architecture. The first three columns show generated samples solely, while we visualise the shape prior on the generated images in the last three columns

by channel-wise concatenation. To be able to compare with our model, we also ran experiments on Heatmap-CGAN, a CGAN conditioned on shapes by heatmap concatenation. First a heatmap taking 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 $\mathbf{z}$. All models use the architecture of DCGAN (Radford et al. 2015).

4.2 Generality of GAGAN

This subsection presents results showing the versatility and generality of GAGAN. As such, we show the different domains GAGAN can applied to as well as GAGAN used for different image sizes (128×128 and 256×256) and different architectures [improved Wasserstein GAN (Gulrajani et al. 2017)]. Further, we extend GAGAN to generate the entire image and discuss the limitations of GAGAN.

Different Domains Figure 7 shows different image domains that GAGAN was applied to. GAGAN is a general model able to generate any structured objects, including human faces, but also cat faces, sketches and hands. Generally, it is only restricted to objects that have an underlying geometry that can be leveraged by the model. The first row of Fig. 7 shows representative samples generated faces from CelebA (Liu et al. 2015). We also trained GAGAN to successfully generates face sketches from 450 sketches annotated with 68 landmark points (sixth row). Further, GAGAN was trained on cat images annotated with 48 facial landmarks (seventh row) and a subset of HGR which include 25 hand feature point annotations (last row). More samples generated for *GAGAN-small* and CelebA can be found at the end of the paper (cf. Figs. 21, 22 and 23).

Different Sizes and Architectures GAGAN leverages a statistical shape model, as well as a differentiable piecewise affine transformation to learn a geometry-aware generative adversarial model. These concepts are not limited to a specific image size and GAN architecture. We extended the DCGAN architecture to generate images of size 128×128 and 256×256, and samples from our best model are shown in Fig. 7 (second and third row). Further, we transferred the concept of GAGAN to improved Wasserstein GAN (Gulrajani et al. 2017). We call this model Geometry-Aware WGAN-

Fig. 10 Manipulating shape latent codes on CelebA: we show the effect of the learned continuous latent factors $\mathbf{p}$ on the output for their values varying from -2.5 and 2.5. In every other row we also plot the shapes to show how images and shapes are aligned. In rows 1 and 2 we show that one of the continuous latent codes captures facial expressions, from sad over neutral to happy. In rows 3 and 4 we show that different latent codes control horizontal and vertical rotation. In rows 5 and 6 a continuous latent space learns the morphology of the face

GP (GAW-GAN-GP) and samples from the model are also depicted in Fig. 7 (third row).

Extending Regularised GAGAN to Generate Background One of the apparent limitations of GAGAN is that the discriminator only discriminates piecewise affine transformations and thus only considers the face. As a consequence background information as well as features like hair, ears and neck were not generated. In order for GAGAN to consider background, we have added four additional points to each fiducial shape set:

$$\begin{aligned}\mathbf{s}^{(i)}_{\text{corner}} = [&x_{i1}, y_{i1}, x_{i2}, y_{i2}, \ldots, x_{im}, y_{im}] \cup \\ &[0, 0, w, 0, h, 0, w, h],\end{aligned} \tag{12}$$

where w represents the width and h the height of the image. Visually, these four points are the corner points of the image and thus the warped images will include background and hair. Training GAGAN with the corner points is exactly the same as without corner points except that we have to retrain our statistical shape model. Figures 7 (fourth row) and 8 also presents samples generated from a CelebA model including the four additional corner points.

Limitations Regularised GAGAN has three dependencies: (1) shape annotations, (2) statistical shape model and (3) piecewise affine transformation. The performance of both statistical shape model and piecewise affine transformation both depends on the shape annotations. The generated hands (cf. Fig. 7, last row) do suffer in quality as both statistical shape model and piecewise affine transformation require outer shape annotations, whereas annotations of HGR (Grzejszczak et al. 2016; Nalepa and Kawulok 2014; Kawulok et al. 2014) only provide inner shape annotations. This also explains the observed thinness of the generated fingers. The second limitation is the piecewise affine transformation, which performs best when all shape meshes are visible.

Fig. 11 Manipulating appearance latent codes on CelebA: by varying $\mathbf{c}^{(i)}$ and leaving $\mathbf{p}^{(i)}$ constant, we generate images with the same shape but different identities. Each row has the same $\mathbf{p}^{(i)}$ while $\mathbf{c}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ was sampled for each row element. In every odd row we also plot the shapes to show how images and shapes are aligned

Fig. 12 Manipulating appearance latent codes on cats and hands: by varying $\mathbf{p}^{(i)}$ and leaving $\mathbf{c}^{(i)}$ constant, we generate images with the same identity but different shape

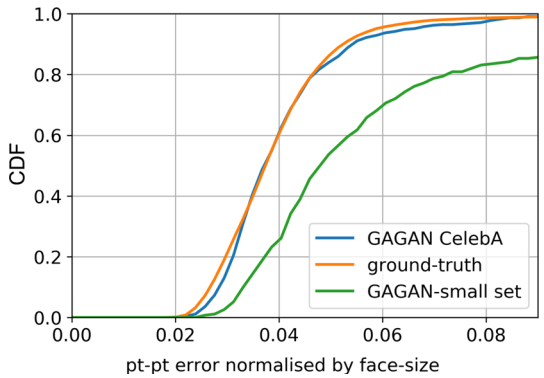


Fig. 13 Percentage of images as a function of the cumulative normalised point-to-point. The error between the landmarks detected by the detector is plotted against those used as prior to generate the images for a model trained on our GAGAN-small set (green) and CelebA (orange)

Therefore, side faces are difficult to handle. One way to address this issue to to use a part-based model (Tzimiropoulos and Pantic 2014b) based on a more flexible, patch-based transformation. However, it is worth noting that our method does not suffer as much as AAMs from this limitation since the generator creates images in their original coordinate frame. Only the discriminator sees the warped image. As such, the discriminator also learns which deformation corresponds to which shape parameters. This is why the images generated by GAGAN do not display artefacts resulting from deformation by the piecewise affine warping.

4.3 Qualitative Results

This subsection presents a qualitative evaluation of our proposed regularised GAGAN. If not further mentioned, regularised GAGAN was trained with $\alpha = 0.01$, $\beta = 0.5$, $\lambda = 1.0$.

Comparisons to Baseline Models We compared GAGAN with Heatmap-GAN, P-GAN and Shape-GAN. Figure 9 depicts samples from the dataset (first row), from all baseline models (rows 2–4) and our results (fifth row) for each (a) GAGAN-small, (b) CelebA and (c) cat dataset. Shape-GAN can only create face images if the dataset is large enough (cf. Fig. 9b), as completely fails to generate any faces when trained on the GAGAN-small set (cf. Fig. 9a). P-GAN and Heatmap-GAN generate samples that follow the shape condition, but the images are either distorted or highly pixelated. Generation is better when trained on CelebA for all models, including ours. This can be explained by the size of each dataset. CelebA is about ten times as large as GAGAN-small set and offers more variety in terms of ethnicity and pose. As known, deep learning methods, including GANs, currently require large training datasets.

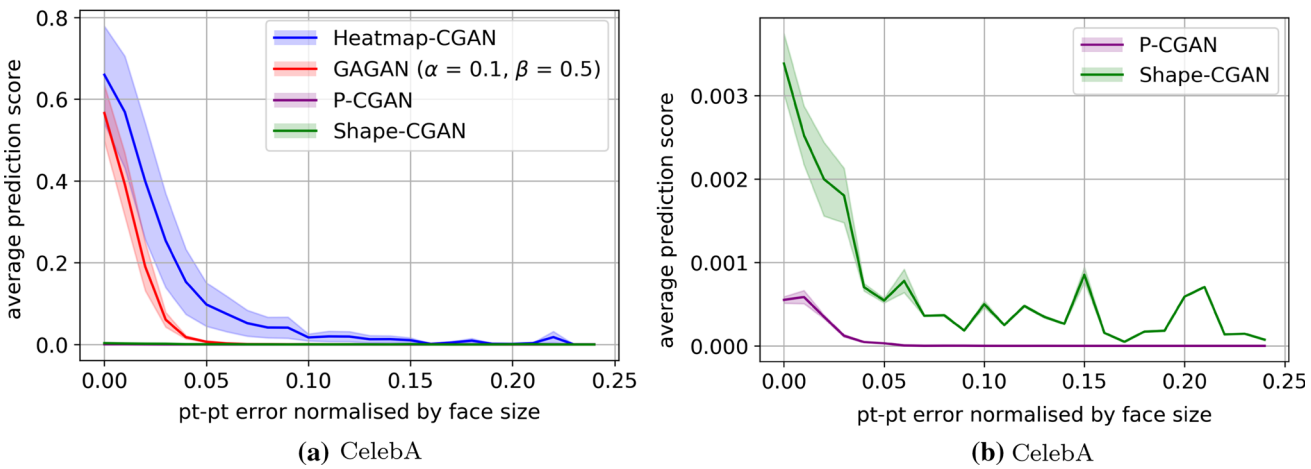
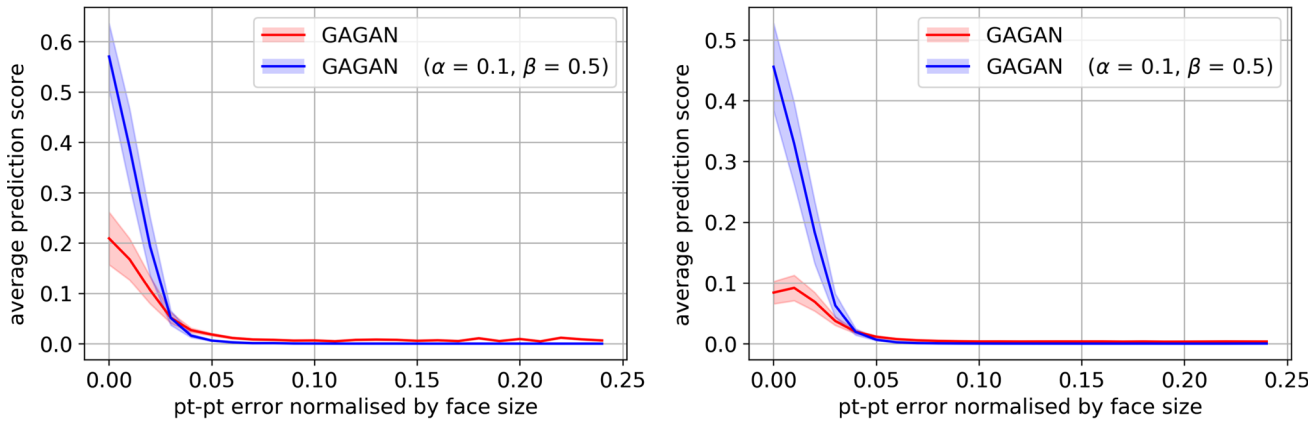


Fig. 14 Baseline comparison. GAGAN in comparison with Heatmap-CGAN, P-CGAN and Shape-CGAN in **a** and a close-up plot of P-CGAN and Shape-CGAN is provided in **b**. Note the different scale of average scores for GAGAN and Heatmap-CGAN in contrast to P-CGAN and Shape-CGAN



(a) Training on Small-set (excluding 300VW) and testing on 300VW

(b) Training on GAGAN-small set, testing on CelebA

Fig. 15 Cross-database comparison of GAGAN and regularised GAGAN. The performance of GAGAN and regularised GAGAN is compared for cross-database testing on 300 VW and CelebA

Fig. 16 GAGAN versus regularised GAGAN. Random 64×64 samples from GAGAN without (α, β) regularisation (first row) and GAGAN with (α, β) regularisation (second row). Both models were trained on GAGAN-small set

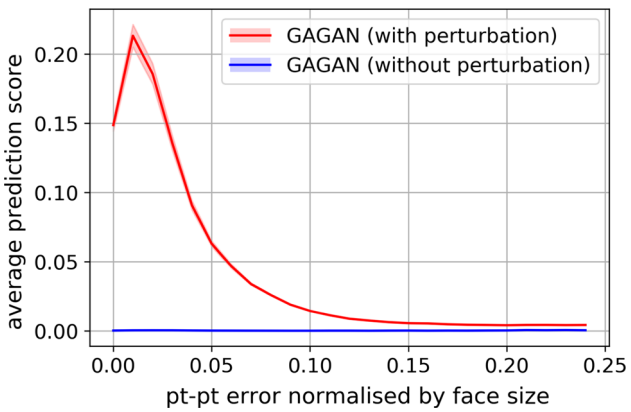
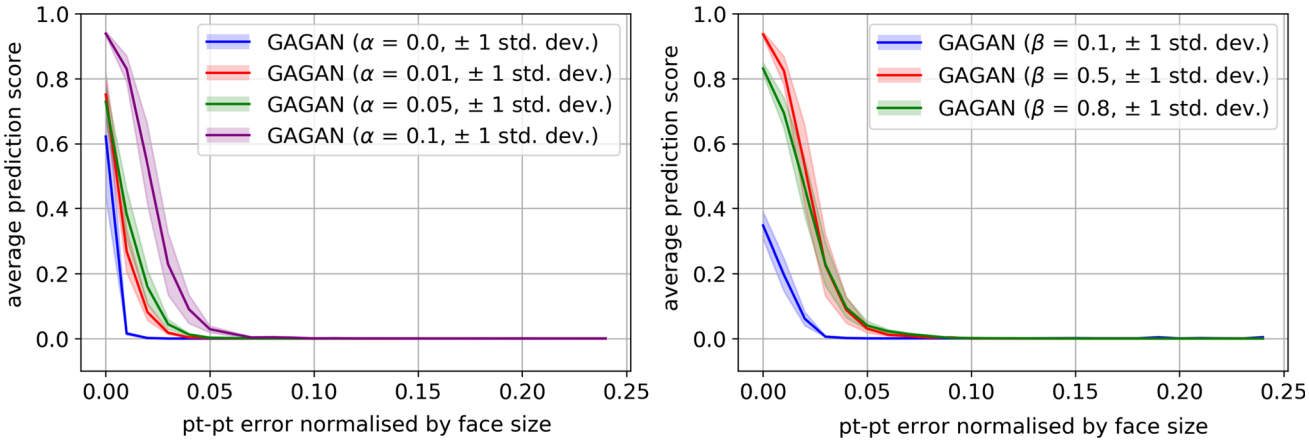


Fig. 17 Cross-database results for GAGAN with and without perturbation. Both models were trained without (α, β) regularisation on GAGAN-small and hyperparameter optimised on all other GAN hyperparameter. The test was conducted with the entire CelebA dataset

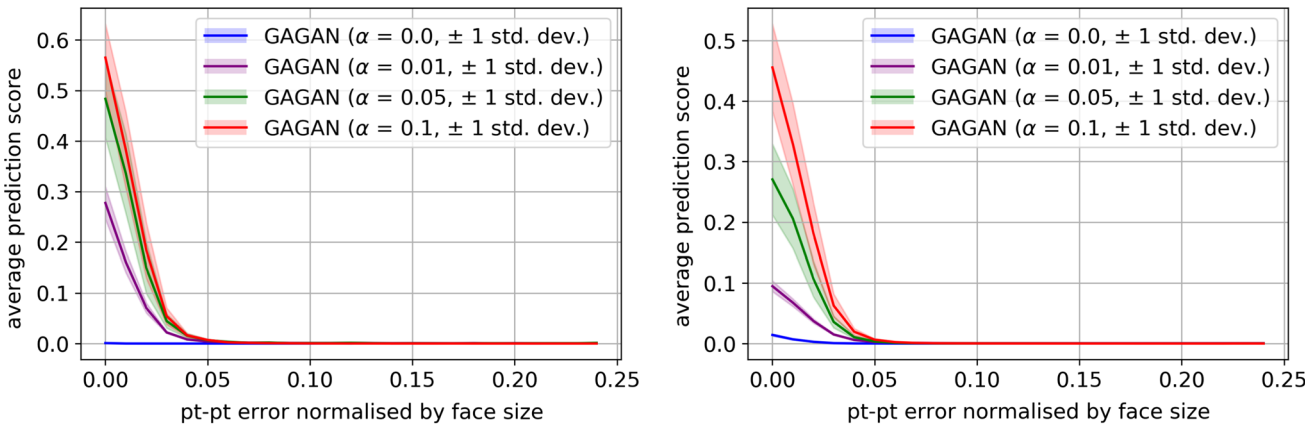
Generating Images from Small Datasets As observed with training baseline models with GAGAN-small set (cf. Fig. 9a), conditional GANs struggle to learn a joint hidden representation to generate realistic images when the sample size is not large enough. P-GAN best manages to generate faces among all baselines, although they are distorted and not realistic and require larger training datasets (see comparison between generation from GAGAN-small set and CelebA in Fig. 9a). In addition, we compared the generation of cats from Sagonas et al. (2015, 2016) for our model with the baseline models. We chose to generate cats because it was among the smallest datasets used in our experiments and the shape model differs greatly from the ones for human faces. The samples of the dataset (first row) and each model are visualised in Fig. 9c. Both Shape-GAN and P-GAN struggle with generating anything but noise whereas Heatmap-GAN generates blurry images. Our approach stands out, solely generating realistic images solely even when trained on a small dataset.

Disentangled Representation In our experiments, we visualise the disentanglement by interpolating only one continuous latent variable $\mathbf{p}_j^{(i)}$ in the range $[-2.5, 2.5]$ while keeping all other $\mathbf{p}_k^{(i)}$, $k \neq j$ and $\mathbf{c}^{(i)}$ fixed. Figure 10 shows that the continuous latent variables $\mathbf{p}$ encode visual concepts such as pose, morphology and facial expression while appearance remains constant, indicating successful disentangled. Figure 7 shows some representative samples drawn from $\mathbf{z}$ at resolutions of 64×64. Only one latent variable $\mathbf{p}_j^{(i)}$ was changed at a time for each row in Fig. 10. We observe realistic and shape-following images for a wide range of facial expressions (rows 1–2), poses (rows 3–4) and morphology (rows 5–6). We show the entire range of $[-2.5, 2.5]$ as this was the range that GAGAN was trained on. Extreme facial expressions shown in rows 1–2 (first 3 samples each) are hard to generate for GAGAN because they are less realistic and do not occur naturally, with lips too thin to generate.



(a) Varying α **(b)** Varying β

Fig. 18 Varying α and β for testing on LFPW and Helen test. The GAGAN models were trained with all samples from GAGAN-small set excluding all images from LFPW and Helen test. For **a** GAGAN was trained with keeping $\beta = 0.5$ constant, while GAGAN was trained with keeping $\alpha = 0.1$ constant



(a) Training on Small-set (excluding 300VW) and testing on 300VW **(b)** Training on GAGAN-small set, testing on CelebA

Fig. 19 Cross-database results for varying α. All models for **a** were trained on GAGAN-small set without 300 VW and tested on 300 VW, while all models for **b** were trained on the entire GAGAN-small set and tested on CelebA. For GAGAN we kept $\beta = 0.5$ constant while varying $\alpha = [0.0, 0.01, 0.05, 0.1]$

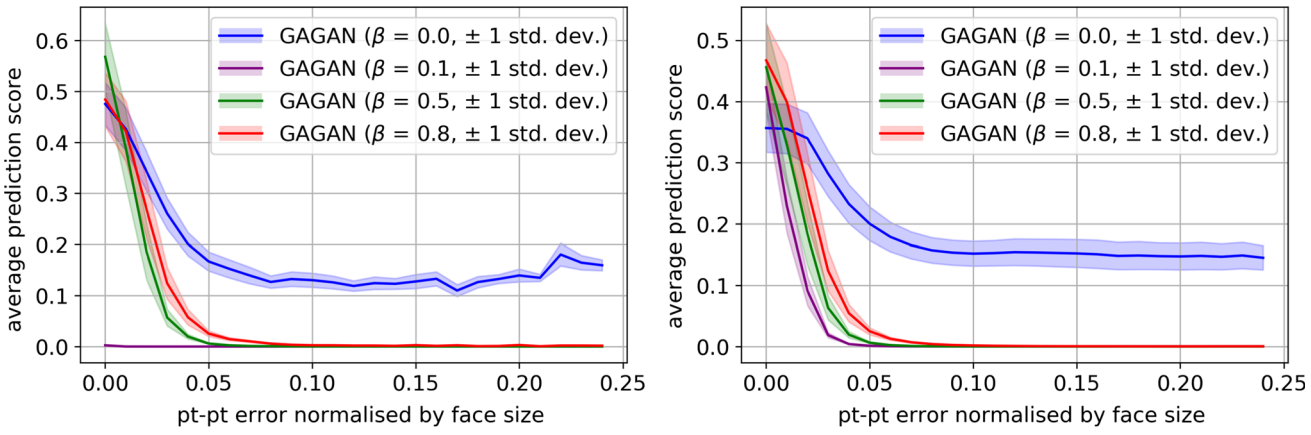
Similarly, we only randomly sampled $\mathbf{c}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and kept $\mathbf{p}^{(i)}$. The results are shown in Fig. 11. In each row we sampled different $\mathbf{c}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As depicted in Fig. 11, by only sampling $\mathbf{c}^{(i)}$ the shape is constant in every row, while the identity varies from image to image. The proportion between men and women sampled seems to be balanced, though we observed fewer older people. Interestingly, the model was able to generate accessories such as glasses during sampling.

Figure 12 shows examples of cats and hands generated by varying the shape parameter $\mathbf{p}^{(i)}$ while keeping $\mathbf{c}^{(i)}$ constant.

4.4 Quantitative Results

This section discusses quantitative results, especially we focus on the discriminative ability of GAGAN to verify landmark detections.

Generation of Aligned Faces The facial landmark detector introduced in Bulat and Tzimiropoulos (2017) detects fiducial points with an accuracy in most cases higher than that of human annotators. Since our model takes as input a shape prior and outputs an image that respects that prior, we can access how well the prior is followed by running that detector on the produced images and measuring the distance between the shape prior and the actual detected shape.



(a) 300VW **(b)** CelebA

Fig. 20 Cross-database results for varying β. All models for **a** were trained on GAGAN-small set without 300 VW and tested on 300 VW, while all models for **b** were trained on the entire GAGAN-small set and tested on CelebA. For GAGAN we kept $\alpha = 0.05$ constant while varying $\beta = [0.1, 0.5, 0.8]$

We directly run the method on 10,000 images generated by the generator of our GAGAN. The error is measured in terms of the established normalised point-to-point error (*pt-pt-error*), as introduced in Zhu and Ramanan (2012) and defined as the RMS error normalised by the face-size. Following Tzimiropoulos and Pantic (2016, 2017); Kossaifi et al. (2015, 2014, 2017) we produced the cumulative error distribution curve depicting for each value on the x-axis in Fig. 13, the percentage of images for which the point-to-point error was lower than this value. For comparison we run the facial landmarks detector on our GAGAN-small set and compute the error using the ground-truth provided with the data. As can be observed, most of the images are pretty well fitted for the model trained on our GAGAN-small set. When trained on CelebA, our model generates faces according to the given prior with similar accuracy as the landmark detector obtains on our training set.

Discriminative Power As GAGAN is trained to discriminate real and fake images conditioned on shapes, we measured how accurately the GAGAN discriminator can assess how well image and shape pairs are aligned. In practice, this is useful because even though state-of-the-art landmark detection performs well, verification of the landmark is usually done manually. This can now be done automatically using the regularised GAGAN to discriminate between good and bad alignment. In the following experiments, we vary the pt–pt error by adding perturbations to the shapes to assess this capability. We report for all experiments the average prediction score and one standard deviation in relation to the pt–pt-error of a given image-shape input. We compare the best GAGAN model ($\beta = 0.5, \alpha = 0.1$) to the baseline models Heatmap-CGAN, P-CGAN and Shape-CGAN. We trained each of the models with GAGAN-small set without 300 VW and tested it on 300 VW. Similarly as Fig. 20, the curves are plotted by calculating the average prediction score from image-shape pairs from CelebA and its perturbations and the corresponding pt-pt error. In Fig. 14a, Heatmap-CGAN has a similar predictive performance, with a much higher variance. We plotted P-CGAN and Shape-CGAN separately in Fig. 14b to assess the slope and predictability. Shape-CGAN performs worse than GAGAN and Heatmap-CGAN as the scale of predictions is much smaller (0.0, 0.003) and the curves show high spikes of variance, especially at the critical pt–pt error of 0.05. P-CGAN fails almost completely to detect any differences in pt–pt-error.

4.5 Improvement Through Regularisation

We also compare the original GAGAN[1] against GAGAN with ($\alpha > 0, \beta > 0$) regularisation. We compared the ability to discriminate alignment quantitatively with test samples from 300 VW and CelebA. Both GAGAN and regularised GAGAN were trained on GAGAN-small set, leaving out 300 VW samples, with $\lambda = 1.0$. In both test cases, regularised GAGAN results in a better discrimination of alignment in terms of slope and variance as observed in Fig. 15.

Further, we also investigated the generation of images of GAGAN and regularised GAGAN. Figure 16 shows samples from both models, GAGAN (first row) and GAGAN with (α, β)-regularisation (second row). While GAGAN without regularisation suffers from minor artefacts, we observe smoother textures generated from GAGAN with regularisation.

[1] Original GAGAN can be expressed as regularised GAGAN+ with ($\alpha = 0, \beta = 0$).

4.6 Ablation Study

In this subsection, we present experiments conducted to show the impact of the perturbation and (α, β) regularisation on GAGAN. Firstly, we will show qualitative and quantitative results on GAGAN trained with and without perturbation. Secondly, similarly to the quantitative experiments in Sect. 4.4, we use our trained GAGAN discriminator for automatic facial landmark verification while varying the alignment of image-shape pairs.

Cross-Database Results With and Without Perturbation We conducted cross-database experiments by training GAGAN on the GAGAN-small set and testing the ability of automatic facial landmark verification on CelebA. Figure 17 shows the pt–pt-error curves in relation to the average prediction score. GAGAN without perturbation cannot discriminate between well and badly aligned image-pairs, whereas GAGAN with perturbation establishes a smooth trend, although only in the between range 0.0 and 0.2 of prediction score. The performance of GAGAN with perturbation is also poor because we did not employ any (α, β) regularisation. This allowed for a unbiased evaluation of the perturbation.

Test on Helen and LFPW Experiments using test sets where the training set includes the corresponding training dataset (LFPW, Helen) were performed to assess how well GAGAN discriminators are able to distinguish between good and bad alignment. Since the GAGAN-small set consists of shape annotations which have been manually verified and corrected by experts, we know that annotations are well aligned to the images. We trained with all samples except the ones from the LFPW and Helen test set and varied α or β while keeping respectively $\beta = 0.5$ and $\alpha = 0.1$ constant. GAGAN was trained with $\lambda = 1.0$. Figure 18 shows that, with $\alpha = 0.1$ and $\beta = 0.5$, GAGAN predicts alignments with pt–pt-error $= 0.0$ with an average score of almost 1.0, and decreasing to an average score of 0.0 with pt–pt-error > 0.05. This average scoring is desirable as in practice any alignment with pt–pt-error > 0.05 should be manually corrected.

Cross-Database Results We also conducted cross-database experiments by training GAGAN on the GAGAN-small set, leaving out all samples from 300 VW for testing. Further, we also trained GAGAN on the entire GAGAN-small set to test it on CelebA. We ran several experiments for GAGAN varying $\alpha = [0.0, 0.01, 0.05, 0.1]$ and $\beta = [0.0, 0.1, 0.5, 0.8]$, with $\lambda = 1.0$. Figure 19 shows the average prediction score in relation to increasing pt–pt-error for testing 300 VW and CelebA for different α. In these experiments, α was trained with different values in range of $[0.0, 0.01, 0.05, 0.1]$ while keeping $\beta = 0.5$ constant. We observe an improvement in average prediction score with higher α values of 0.05 and 0.01, while $\alpha = 0.0$ fails completely to detect any difference in pt–pt-error and $\alpha = 0.0$. Further, the average prediction scores on the CelebA dataset have higher variance than the ones of 300 VW. This is likely the result it being more diverse and much larger.

We also varied $\beta = [0.1, 0.5, 0.8]$ while keeping $\alpha = 0.1$ constant in our cross-database experiments. Results are visualised in Fig. 20. By increasing β we decrease the variance in predictions and have a clearer threshold between well aligned images and shapes (pt–pt error < 0.05) and badly aligned images and shapes for both 300 VW and CelebA. With $\beta = 0.0$, there is a clear division between aligned images and shapes (pt–pt error $= 0.0$) and aligned images and shapes with pt–pt error of approx. 0.05–0.2 for 300 VW. However, the average prediction score rises with 0.2 pt–pt-error which is counter-intuitive of how the discriminator should rank the images and shapes. Further, with increasing β, the variance of the average prediction score decreases, and thus gives better precision to the predictions.

5 Conclusion

We introduced regularized GAGAN, a novel method able to produce realistic images conditioned on disentangled latent shape and appearance representations. The 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 transformation applied to its output. The discriminator, trained only on images normalised to canonical image coordinates, is able to not only differentiate realistic from fake samples, but also judge the alignment between image and shape without being explicitly conditioned on the prior. The resulting representational power allows to automatically assess the quality of facial landmarks tracking, while avoiding dataset shifts. We demonstrated superior performance compared to other methods across datasets, domains and sample sizes. Our method is general and can be used to augment any existing GAN architecture.

Appendix

See Figs. 21, 22, 23.

Fig. 21 Samples from GAGAN trained with CelebA (128 × 128): all samples are from one batch generated by GAGAN. Rows 1–2 are the samples without processing, rows 3–4 are samples with shapes superimposed, rows 5–9 are samples with the background removed during post-processing and rows 9–12 are samples with the background removed and shapes superimposed

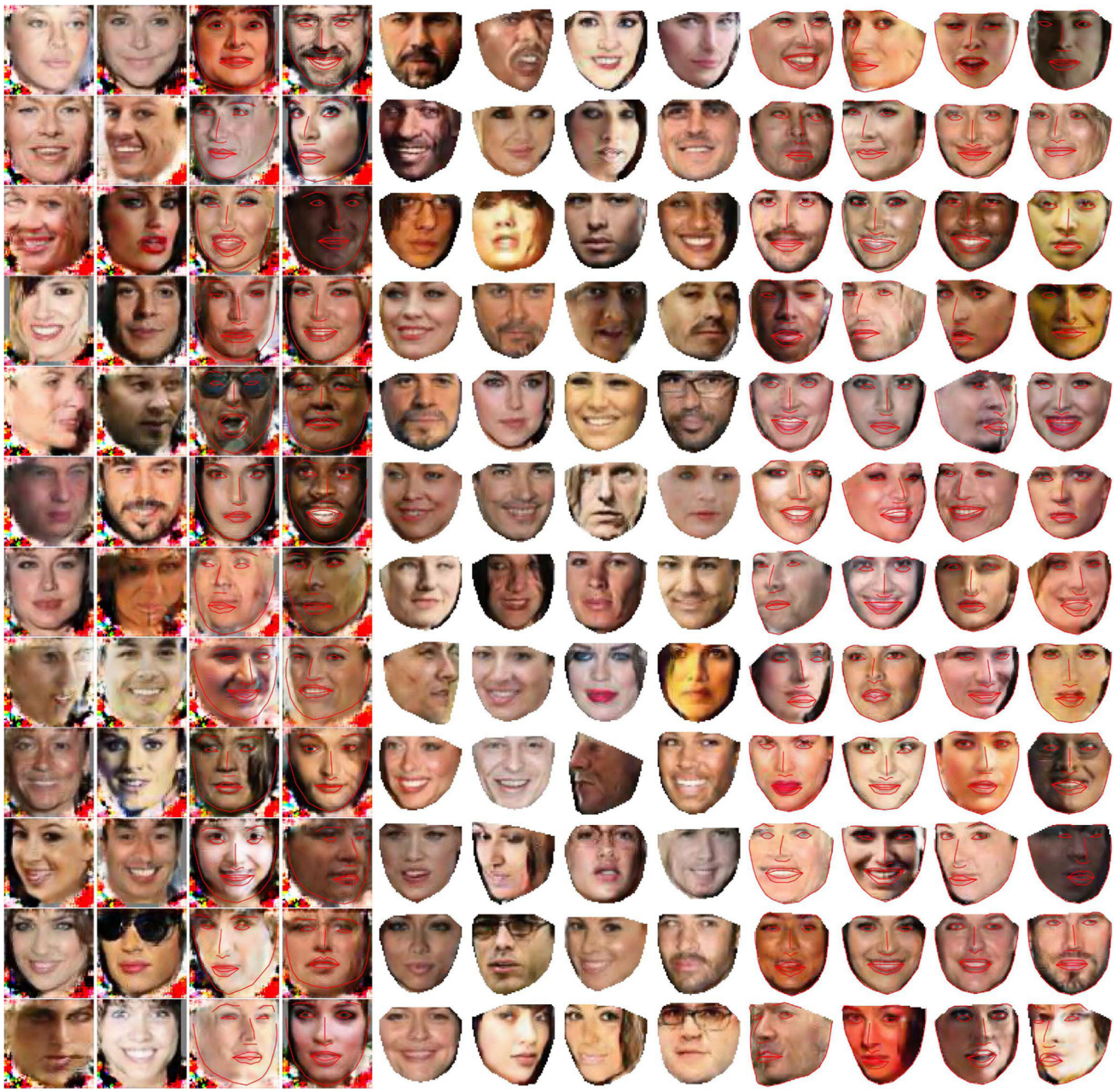


Fig. 22 Samples from GAGAN trained with CelebA (64 × 64): all samples are from one batch generated by GAGAN. Rows 1–2 are the samples without processing, rows 3–4 are samples with shapes superimposed, rows 5–9 are samples with the background removed during post-processing and rows 9–12 are samples with the background removed and shapes superimposed

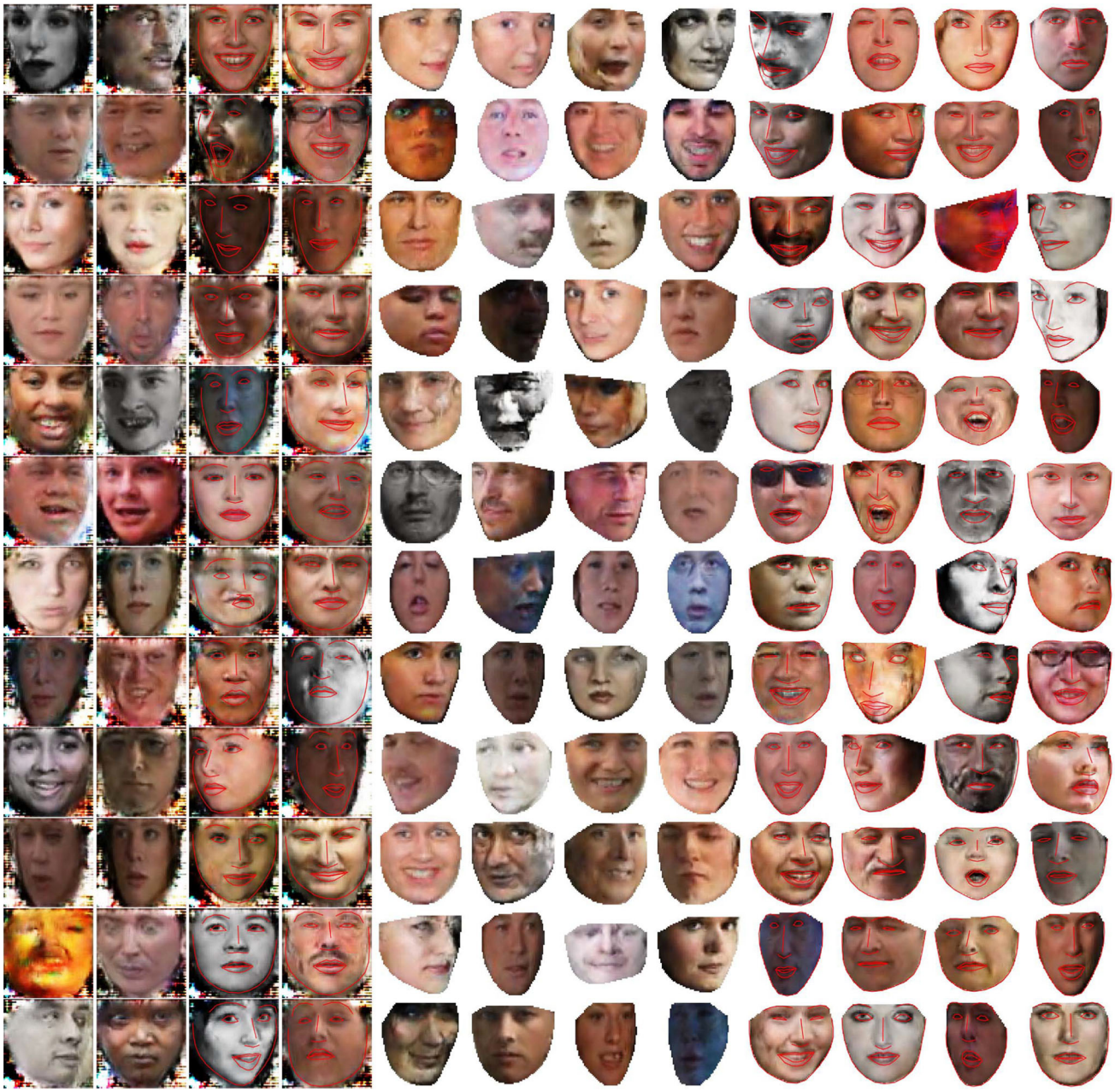


Fig. 23 Samples from GAGAN trained with GAGAN-small: all samples are from one batch generated by GAGAN. Rows 1–2 are the samples without processing, rows 3–4 are samples with shapes superimposed, rows 5–9 are samples with the background removed during post-processing and rows 9–12 are samples with the background removed and shapes superimposed

Acknowledgements The work of Linh Tran, Yannis Panagakis and Maja Pantic has been funded by the European Community Horizon 2020 under Grant Agreement Nos. 688835, 645094 (DE-ENIGMA).

References

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

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. arXiv preprint arXiv:1701.07875.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2011). Localizing parts of faces using a consensus of exemplars. In *The 24th IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 545–552).

Bulat, A., & Tzimiropoulos, G. (2017). How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International conference on computer vision*.

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

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *23*(6), 681.

Cootes, T., Taylor, C., Cooper, D., & Graham, J. (1995). Active shape models-their training and application. *Computer Vision and Image Understanding*, *61*(1), 38.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*.

Davies, R., Twining, C., & Taylor, C. (2008). *Statistical Models of Shape: Optimisation and Evaluation* (1st ed.). Berlin: Springer.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using real NVP. In *5th International conference on learning representations (ICLR)*.

Edwards, G. J., Taylor, C. J., & Cootes, T. F. (1998). Interpreting face images using active appearance models. In *IEEE international conference on automatic face and gesture recognition (FG)* (pp. 300–305).

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

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. *Image and Vision Computing (IVC)*, *28*(5), 807.

Grzejszczak, T., Kawulok, M., & Galuszka, A. (2016). Hand landmarks detection and localization in color images. *Multimedia Tools and Applications*, *75*(23), 16363. https://doi.org/10.1007/s11042-015-2934-5.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A.C. (2017). In *Advances in neural information processing systems* (pp. 5767–5777).

Jain, V., & Seung, S. (2009). Natural image denoising with convolutional networks. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (pp. 769–776). Red Hook: Curran Associates Inc.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694–711).

Kawulok, M., Kawulok, J., Nalepa, J., & Smolka, B. (2014). Self-adaptive algorithm for segmenting skin regions. *EURASIP Journal on Advances in Signal Processing*, *2014*(170), 1. https://doi.org/10.1186/1687-6180-2014-170.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *2nd international conference on learning representations (ICLR)*.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems* (pp. 4743–4751).

Kossaifi, J., Tran, L., Panagakis, Y., & Pantic, M. (2017). Gagan: Geometry-aware generative adversarial networks. In *IEEE CVPR*. arXiv:1712.00684.

Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2014). Fast newton active appearance models. In *Proceedings of the IEEE international conference on image processing (ICIP14)* (pp. 1420–1424).

Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2015). Fast and exact bi-directional fitting of active appearance models. In *Proceedings of the IEEE international conference on image processing (ICIP15)* (pp. 1135–1139).

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

Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2017). Fast and exact newton and bidirectional fitting of active appearance models. *IEEE Transactions on Image Processing*, *26*(2), 1040.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning* (pp. 1558–1566).

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., & Wang, Z. et al. (2016). Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:1609.04802.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of international conference on computer vision (ICCV)*.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, *60*(2), 91.

Mathieu, M. F., Zhao, J. J., Zhao, J., Ramesh, A., Sprechmann, P., & LeCun, Y. (2016). Disentangling factors of variation in deep representation using adversarial training. In *Advances in neural information processing systems* (pp. 5040–5048).

Matthews, I., & Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision (IJCV)*, *60*(2), 135.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Nalepa, J., & Kawulok, M. (2014). Fast and accurate hand shape classification. In *International conference: beyond databases, architectures and structures* (pp. 364–373).

Odena, A., Olah, C., & Shlens, J. (2016). Conditional image synthesis with auxiliary classifier gans. arXiv preprint arXiv:1610.09585.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2536–2544).

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset shift in machine learning*. Cambridge: The MIT Press.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396.

Rezende, D., & Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning* (pp. 1530–1538).

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic back-propagation and approximate inference in deep generative models. arXiv preprint arXiv:1401.4082.

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

Sagonas, C., Panagakis, Y., Zafeiriou, S., & Pantic, M. (2015). Robust statistical face frontalization. In *Proceedings of IEEE international conference on computer vision (ICCV 2015)*.

Sagonas, C., Panagakis, Y., Zafeiriou, S., & Pantic, M. (2016). Robust statistical frontalization of human and animal faces. International

Journal of Computer Vision. Special Issue on "Machine Vision Applications".
Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013a). A semi-automatic methodology for facial landmark annotation. In *CVPR Workshops*.
Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013b). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *The IEEE international conference on computer vision (ICCV) workshops* (pp. 397–403).
Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems* (pp. 2234–2242).
Salimans, T., Karpathy, A., Chen, X., & Kingma, D. P. (2017). Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *5th international conference on learning representations (ICLR)*.
Shen, J., Zafeiriou, S., Chrysos, G., Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of IEEE international conference on computer vision, 300 videos in the wild (300-VW): Facial landmark tracking in-the-wild challenge & workshop (ICCVW'15)* (pp. 50–58).
Tipping, M. E., & Bishop, C. M. (2003). Bayesian image super-resolution. In *Advances in neural information processing systems* (pp. 1303–1310).
Tran, L., Yin, X., & Liu, X. (2017). Disentangled representation learning gan for pose-invariant face recognition. *IEEE CVPR*, *4*(5), 7.
Tzimiropoulos, G., & Pantic, M. (2014a). Gauss-newton deformable part models for face alignment in-the-wild. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1851–1858).
Tzimiropoulos, G., & Pantic, M. (2014b). In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1851–1858).
Tzimiropoulos, G., & Pantic, M. (2016). Fast algorithms for fitting active appearance models to unconstrained images. *International Journal of Computer Vision*, *122*, 1–17.
Tzimiropoulos, G., & Pantic, M. (2017). Fast algorithms for fitting active appearance models to unconstrained images. *International Journal of Computer Vision*, *122*(1), 17.
Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2012). Subspace learning from image gradient orientations. *IEEE TPAMI*, *34*(12), 2454.
van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., & Graves, A. et al. (2016). Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems* (pp. 4790–4798).
Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096–1103).
Wang, C., Wang, C., Xu, C., & Tao, D. (2017). Tag disentangled generative adversarial networks for object image re-rendering. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI* (pp. 2901–2907).
Xie, J., Xu, L., & Chen, E. (2012). Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems* (pp. 341–349).
Yang, J., Wright, J., Huang, T. S., & Ma, Y. (2010). Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, *19*(11), 2861.
Yang, H., Zou, C., & Patras, I. (2014). Face sketch landmarks localization in the wild. *IEEE Signal Processing Letters*, *21*(11), 1321.
Zhao, J., Mathieu, M., & LeCun, Y. (2016). Energy-based generative adversarial network. arXiv preprint arXiv:1609.03126.
Zhou, J. B. F., & Lin, Z. (2013). Exemplar-based graph matching for robust facial landmark localization. In *IEEE international conference on computer vision (ICCV)* (pp. 1025–1032).
Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the 2012 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2879–2886).
Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 2242–2251).