High Resolution Zero-Shot Domain Adaptation of Synthetically Rendered Face Images

Stephan J. Garbin, Marek Kowalski, Matthew Johnson, and Jamie Shotton
Microsoft

Abstract. Generating photorealistic images of human faces at scale remains a prohibitively difficult task using computer graphics approaches. This is because these require the simulation of light to be photorealistic, which in turn requires physically accurate modelling of geometry, materials, and light sources, for both the head and the surrounding scene. Non-photorealistic renders however are increasingly easy to produce. In contrast to computer graphics approaches, generative models learned from more readily available 2D image data have been shown to produce samples of human faces that are hard to distinguish from real data. The process of learning usually corresponds to a loss of control over the shape and appearance of the generated images. For instance, even simple disentangling tasks such as modifying the hair independently of the face, which is trivial to accomplish in a computer graphics approach, remains an open research question. In this work, we propose an algorithm that matches a non-photorealistic, synthetically generated image to a latent vector of a pretrained StyleGAN2 model which, in turn, maps the vector to a photorealistic image of a person of the same pose, expression, hair, and lighting. In contrast to most previous work, we require no synthetic training data. To the best of our knowledge, this is the first algorithm of its kind to work at a resolution of 1K and represents a significant leap forward in visual realism.

1 Introduction

Generating photorealistic images of human faces remains a challenge in computer graphics. While we consider the arguably easier problem of still images as opposed to animated ones, we note that both pose unsolved research questions. This is because of the complicated and varied appearance of human tissue found in the skin [39], eyes [6] and teeth of the face region. The problem is further complicated by the fact that humans are highly attuned to the appearance of faces and thus skilled at spotting any unnatural aspect of a synthetic render [30]. Machine learning has recently seen great success in generating still images of faces that are nearly indistinguishable from the domain of natural images to a non-expert observer. This gives methods like StyleGAN2 (SG2) [23] a clear advantage over computer graphics if the goal is to generate photorealistic image samples only. The limitation of models like SG2 is that we get RGB data only, and that such samples are often only useful if annotations such as head pose,
Fig. 1: Pairs of synthetic input images, and samples from our algorithm at 1K resolution. Despite the real dataset used (FFHQ) mostly containing images of people looking directly at the camera with neutral or smiling expressions, we are able to match more diverse poses and facial expressions. Best viewed zoomed in, and in colour.

UVs, or expression parameters are available for downstream tasks. The second major issue is that generative models necessarily inherit the bias of the data they were trained on. For large image collections, this may be hard to assess [42]. In computer graphics on the other hand, annotations such as UVs can be trivially obtained for an image. Since the assets that define the data input into the renderer need to be explicitly created, bias control becomes more feasible.

In this paper, we propose to play to the strengths of both fields by using machine learning to change the appearance of non-photorealistic renders to be more natural, while keeping semantics such as the shape, the expression of the face, and the lighting consistent. This means that the annotations obtained from the renders are still valid for the images after domain transfer. Because we do not require photo-realism from the synthetic renders, we can produce them at scale and with significant variety using a traditional graphics pipeline.

In contrast to much recent work in the field, we require no synthetic images for training at all, and thus no paired data. In fact, our method only requires a pre-trained StyleGAN2 model, and a small number of manual annotations from the data it was trained on as outlined below. For a given synthetic image, our methods works best if masks of the hair and background are available. These can be easily obtained from any renderer. Our method works by finding an embedding in the latent space of SG2 that produces an image which is perceptually similar to a synthetic sample, but still has the characteristic features of the data the GAN was trained with. In terms of the scale space of an image [9,8], we attempt to match the coarser levels of an image pyramid to the synthetic data,
and replace fine detail with that of a photorealistic image. Another way to interpret this is that we attempt to steer StyleGAN2 with the help of synthetic data [20].

While embedding images in the latent space of SG2 is not a new concept [21], the issue with using existing approaches is that they either do not, or struggle to, enforce constraints to keep the results belonging to the distribution of real images. In fact, the authors in [2] explicitly note that almost any image can be embedded in a StyleGAN latent space. If closeness to the domain of real images is not enforced, we simply get an image back from the generator that looks exactly like the synthetic input whose appearance we wish to change.

We make the observation that samples from the prior distribution usually approximately form a convex set, i.e. that convex combinations of any set of such points mapped through the generator are statistically similar to samples from the data distribution the GAN was trained on. We also note that showing interpolations between pairs of latent vectors is a common strategy of evaluating the quality of the latent embeddings of generative models [31]. As part of our method, we propose an algorithm, Convex Set Approximate Nearest Neighbour Search (CS-ANNS), which can be used to traverse the latent space of a generative model while ensuring that the reconstructed images closely adhere to the prior. This algorithm optimises for the combination of a set of samples from the SG2 prior by gradient descent, and is detailed in the method section below.

In summary, our contributions are:

1. The first zero-shot domain transfer method to work at 1K and with only limited annotations of the real data, and
2. a novel algorithm for approximate nearest neighbour search in the latent spaces of generative models.

2 Related Work

2.1 Generative Models

Generative models are of paramount importance to deep learning research. In this work, we care about those that map samples from a latent space to images. While many models have been proposed (such as Optimized Maximum Mean Discrepancy [34], Noise Contrastive Estimation [15], Mixture Density Networks [7], Neural Autoregressive Distribution Estimators [27], Real-Valued Neural Autoregressive Distribution Estimators [35], Diffusion Process Models [33]), the most popular ones are the family of Variational Autoencoders (VAEs) [21,25], and Generative Adversarial Networks (GANs) [14].

Because GANs are (at the time of writing) capable of achieving the highest quality image samples, we focus on them in this work, and specifically on the current state of the art for face images, StyleGAN2 (SG2) [23]. In any GAN, a neural sampler called the generator is trained to transform samples from a simple distribution to the true distribution over the data space. A second network, called
the discriminator, is trained to differentiate samples produced by the generator
and those from the data space.

Our method takes as input only the pretrained generator of SG2, and while
we backpropagate through it, we do not modify its weights as part of our algo-

rithm. Since SG2 uses a variant of Adaptive Instance Normalisation (AdaIn)\cite{17},
its latent space is mapped directly to the AdaIn parameters at 18 different layers.

We do not use the additional noise inputs at each layer. This way of controlling
the generator output via the AdaIn inputs is the same methodology as used in
the Image2StyleGAN work\cite{2}. The authors in\cite{2} also consider style transfer by
blending two latent codes together, but choose very different image modalities
such as cartoons and photographs. We build on their work by defining a process
that finds a close nearest neighbour to blend with, thereby creating believable
appearance transfer that preserves semantics.

2.2 Zero-Shot Domain Transfer

To the best of our knowledge, there are no zero-shot image domain transfer meth-

ods in the literature that require only one source domain. By domain adaptation
we mean the ability to make images from dataset\textit{A} look like images from dataset
\textit{B}, while preserving content. While one-shot methods like\cite{43} or\cite{5} have been
proposed, they work at significantly lower resolution than ours and still require
one sample from the target domain.

ZstGAN\cite{28}, the closest neighbour, requires many source domains (that
could for example be extracted from image labels of one dataset). The highest
resolution handled in that work is $128^2$, which is significantly lower than our
method. The categories are used to bootstrap the appearance transfer problem,
as if multiple datasets were available. Without labels for dividing the data into
categories, we were unable to use it as a baseline.

2.3 Domain Adaptation

If paired training data from two domains is available, Pix2Pix\cite{19} and its suc-

cessors (e.g. Pix2Pix HD\cite{37}, which changes to architecture to use multiple
discriminators at multiple scales to produce high resolution images) can be used
effectively for domain adaptation.

CycleGAN does not require paired training data\cite{44}. This brings it closer
to the application we consider. However, it still requires a complete dataset of
both image modalities. Many improvements have since been suggested to im-
prove CycleGAN. HarmonicGAN adds an additional smoothness constraint to
reduce artefacts in the outputs\cite{41}, Sem-GAN exploits additional information
in the source images\cite{10}, as does\cite{3}. Discriminative Region Proposal Adversar-

ial Networks (DRPAN)\cite{36} add steps to fix errors, Geometry-Consistent GANs
(GcGAN)\cite{11} use consistency under simple transformations as an additional
constraint. Some methods also model a distribution of over possible outputs,
such as MUNIT\cite{18} or FUNIT\cite{29}.

However, none of these methods are capable of zero-shot domain adaptation.
3 Method

Fig. 2: Illustration of the different steps of our method. (a) is the input synthetic render, (b) the output of the sampling in step 1, (c) the result of Convex Set Approximate Nearest Neighbour Search in step 2, and (d-f) results from step 3.

In the following, any variable containing $w$ refers to the $18 \times 512$ dimensional inputs of the pretrained SG2 generator, $G$. Any variable prefixed with $I$ refers to an image, either given as input, or obtaining by passing a $w$ through the generator $G$. The proposed method takes as input a synthetically rendered image $I^s$, and returns a series of $w$s that represent domain adapted versions of that input.

Our algorithm has four stages, each producing results more closely matching the input. In the first, we find the latent code, $w^s$, of an approximate nearest neighbour to a given synthetic input image, $I^s$, by sampling. This is the starting point of our method. For the second step, we propose Convex Set Approximate Nearest Neighbour Search (CS-ANNS), an algorithm to refine the initial sample by traversing the latent space while being strongly constrained to adhere to the prior. This gives us a refined latent code, $w^n$.

In the third step, we fit SG2 to the synthetic image without any constraint to obtain another latent code $w^f$ that matches $I^s$ as closely as possible. We can then combine $w^f$ and $w^n$ with varying interpolation weights to obtain a set of final images that strongly resemble $I^s$, but which have the appearance of real photographs.

Because $w^s$, $w^n$ and the results from step 3 are all valid proposals for the final result, we select the latent code that gives an image as semantically similar
to $I^s$ as possible from among them in the fourth and final step. An example of the different steps of our method can be seen in Figure 2.

We note that the SG2 model used in this section was trained on the FFHQ dataset [22], a dataset of photographs at high resolution. We use the same pre-processing and face normalisation as the authors of that work.

Since we care about closely matching the face in this work, we construct floating-point alpha masks from the synthetic renders that de-emphasize the background and allow us to separate the hair. This gives us tuples of renders and alpha masks $\{I^s, I^a, I^{ahair}\}$ for each input. We observe that in order to get accurate matching of face boundaries, the sharp opacity edges of $I^a$ that come from the renderer need to be extended outwards from the face. We compute the distance transfer for the face boundary and produce a quickly decaying falloff by mapping the resulting values, remapped to be in the range $0 - 1$, by $x^{10}$, where $x$ is the output of the distance transform at a pixel. This is illustrated in Figure 3.

3.1 Step 1: Sampling

To find a good initialisation, we could sample from the prior of SG2 and take the best match as input to the other steps. However, we found that this could fail to produce convincing results for faces at an angle, under non-frontal illumination etc. because our synthetic data is more varied in pose, lighting and ethnicity than FFHQ. To overcome this problem, we annotate a small subset of 2000 samples from SG2 with a series of simple attributes to obtain a set of 33 control vectors, $v_{control}$. These are detailed in the supplementary material. The effect of adding some of these to the mean face of SG2 is shown in Figure 4. We also select a set of centroids, $v_{centroid}$, to sample around. As can be seen in Figure 5 these are selected to be somewhat balanced in terms of sex, skin tone and age, and are chosen empirically. We are unable to prove that this leads to greater overall fairness [12], and acknowledge that this sensitive issue needs closer examination in future work.

The loss used in the sampling step is a combination of the LPIPS distance [40], an L1 loss with different weights for colour and luminance, and landmark loss based on 68 points compute with DLIB [24]. The loss is computed at a quarter resolution, $256^2$ pixels, after low pass filtering with a Gaussian kernel of size 7 and standard deviation of 0.15, and multiplication with the mask $I^a$.
We found this crucial to ensure fine detail is retained in the fits. The entire loss function for the sampling step is thus:

\[
L_{\text{sampling}} = L_{\text{LPIPS}}(r(I^s * I^a), r(G(w^s) * I^a)) \\
+ \lambda_{\text{lum}} \| y(r(I^s * I^a)) - y(r(G(w^s) * I^a)) \|_1 \\
+ \lambda_{\text{col}} \| u(r(I^s * I^a)) - u(r(G(w^s) * I^a)) \|_1 \\
+ \lambda_{\text{landm}} \| l(r(I^s * I^a)) - l(r(G(w^s) * I^a)) \|_2^2, \tag{1}
\]

where \( r \) is the resampling function that changes image size after Gaussian filtering, \( u \) separates out the colour channels in the YUV colour space, \( y \) the luminance channel, \( G \) is the pretrained SG2 generator, \( I^s \) a synthetic image, \( w^s \) a latent code sample, and \( l \) the landmark detector. \( \lambda_{\text{lum}} \) is set to 0.1, \( \lambda_{\text{col}} \) to 0.01, and \( \lambda_{\text{landm}} \) to \( 1e-5 \).

For each sample at this stage of our method, we pick one of the centroids \( v_{\text{centroid}} \) with uniform probability, and add Gaussian noise to it. We then combine this with a random sample of our control vectors to vary pose, light, expression etc. The i’th sample is thus obtained as:

\[
w^s_i = s(v_{\text{centroid}}) + \mathcal{N}(0, \sigma^2) \\
+ v_{\text{control}} * \mathcal{N}_{\text{uniform}} * 2.0, \tag{2}
\]

where \( s \) is the random centroid selection function, \( \sigma^2 = 0.25 \), and \( \mathcal{N}_{\text{uniform}} \) is uniform noise to scale the control vectors.

The output of this stage is simply the best \( w^s \) under the loss in Equation (1) for any of the 512 samples taken.

3.2 Step 2: Convex Set Approximate Nearest Neighbour Search

In step 2, we refine \( w^s \) while keeping the results constrained to the domain of valid images in SG2. The intuition is that any convex combination of samples from the prior in the latent space also leads to realistic images when decoded through \( G \). We highlight that \( w^n \) is the current point in the latent space, and updated at every iteration. It is initialised with the result from step 1.

At each step, we draw 512 samples using the same procedure as before, just without adding noise, which now becomes a learnable parameter, \( \beta \). At each iteration, each of the 512 proposal samples \( w^p \) is obtained as:

\[
w^p_i = s(v_{\text{centroid}}) + \mathcal{N}(0, \sigma^2). \tag{3}
\]
We blend the centroid candidates $w_i^p$ using linear interpolation with the current $w_n$ using a uniform random $\alpha$ in the range of $0.25 - 0.75$, as we found that this stabilised the optimisation. We then optimise for a set of weights, $\alpha$, which determine how the $w_i^p$'s and current $w_n$ are combined. We use sets of $\alpha$ for each of 18 StyleGAN2 latent space inputs, i.e. $\alpha$ is a matrix of shape $[512 + 1, 18]$ (note how the current $w_n$ is included). We constrain the optimisation to make each row of $\alpha$ sum to 1 using the softmax function, ensuring a convex combination of the samples. In addition to $\alpha$, we allow learned variation for the control vectors, which are scaled by a learnable parameter $\beta$. Because this last step could potentially lead to solutions far outside the space of plausible images, we clamp $\beta$ to 2.0. The loss is the same as Equation 1, just without the landmark term, i.e. with $\lambda_{landm}$ set to 0.

We use 96 outer iterations for which the $w_i^p$ are redrawn, and $\alpha$ and $\beta$ reset so that the current $w_n$ is the starting point (i.e. $\beta$ is set to zero, and $\alpha$ to one only for the current $w_n$). For each of these outer loops, we optimise $\alpha$ and $\beta$ using Adam [26] with a learning rate of 0.01 in an inner loop. We divide the initial learning rate by 10.0 for every 4 iterations in that inner loop, and return the best result at any point, which gives us the refined $w_n$. We name this algorithm Convex Set Approximate Nearest Neighbour Search (CS-ANN). More details can be found in the supplementary material.

3.3 Step 3: Synthetic Fit and Latent Code Interpolation

To fit SG2 to the synthetic image $I_s$, we use the method of [23] with minor modifications based on empirical observation. We set the number of total steps to 1000, the initial learning rate to 0.01, and the initial additive noise to 0.01. These changes are justified as we start from $w_n$ and so have a much-improved initialisation compared to the original algorithm. We also mask the loss using the same $I_a$ as above.

Having obtained a latent code $w_s$ that closely resembles the synthetic input image $I_s$, and a latent code that describes that approximate nearest neighbour $I_n$, we can combine them in such a way that preserves the overall facial geometry of $I_s$ but has the fine detail of $I_n$. We use simple interpolation to do this, i.e. the final latent code is obtained as:

$$w_{final} = w_s \ast \sqrt{\alpha} + w_n \ast \sqrt{1.0 - \alpha},$$

Fig. 6: Interpolating between the output of step 2 ($a0$) and the output of the exact fit to the synthetics ($e4$). ($a - e$) represent the number of latent codes used for blending, and ($0 - 4$) the floating point weights for them.
where $w_{\text{final}}$ is a candidate for the final output of our method. We generate candidates by letting $\alpha$ retain the first $\{1, 3, 5, 7\}$ of the 18 latent codes with a floating point weight of $\{1.0, 0.9, 0.8, 0.7\}$ each. An example of the effect of this interpolation can be seen in Figure 6.

### 3.4 Step 4: Sample Selection

Having obtained a sequence of proposals, from step 1-4, we simply select the one that matches input most using the Structural Similarity (SSIM) [38] metric at a resolution of $368^2$ pixels, which we empirically found to give better qualitative results than the LPIPS distance. We hypothesise that this is due to the fact that perceptual losses prioritise texture over shape [13], and alignment of facial features is important for effective domain adaptation. We note that step 1-3 are run ten times with different random seeds to ensure that even difficult samples are matched with good solutions.

### 4 Experiments

We want to establish how realistic the images generated by our method are, and how well they preserve the semantics of the synthetic images, specifically head pose and facial features. To do so, we obtain a diverse set of 1000 synthetic images, and process with them with our method, as well as two baselines.

We evaluate our algorithm quantitatively against the fitting method proposed in [23], which was designed to provide a latent embedding constrained to the domain of valid images in a pretrained SG2 model, and also operates at 1K resolution. This method is referred to as the StyleGAN2 Baseline.

To assess how well we can match face pose and expression, we additionally compare facial landmark similarity as computed by OpenFace [4]. A qualitative comparison is made to the StyleGAN2 Baseline, and CycleGAN [44], with the latter trained on the entirety of our synthetic dataset. We conducted a user study to assess the perceived realism of our results compared to the StyleGAN2 Baseline as well as the input images, and to provide an initial assessment of loss of semantics.

We make use of two variants of our results throughout this section. For Ours (only face), we replace the background and hair using the masks from the synthetic data by compositing them using a Laplacian Pyramid [9]. Because of the close alignment of our results with the input, this produces almost no visible artefacts. This allows us to ensure that the background does not impact the quantitative metrics., and to isolate just the appearance change of the face itself.

### 4.1 Qualitative Experiments

We train CycleGAN on FFHQ as well as a dataset of 12000 synthetic images, using the default training parameters suggested by the authors. Despite having access to the synthetic training data, and using a much tighter crop, we found the
results after 50 epochs unconvincing. Even at $128^2$, the images show artefacts, and lack texture detail. We show some results in Figure 7 and more in the supplementary material. Because of the overall quality of the results and because this method has access to the entire synthetic dataset during training, we do not include it in our user study. Instead, we focus on the StyleGAN2 Baseline for extensive evaluation.

We show each annotator three images: The synthetic input, the baseline result and our result, in random order. We ask if our result or the baseline is more photorealistic, and which image is the overall most realistic looking, i.e. comparable to a real photograph. Finally, we ask if the synthetic image and our result could be the same image of the same person. In this case, we let each annotator answer \{Definitely No, Slightly No, Slightly Yes, Definitely Yes\}.

From the annotation of 326 images, our results are considered more photoreal than the StyleGAN2 Baseline in 94.48% of cases. In 95.1% of responses, our result was considered more realistic looking than the input or the baseline.

In terms of whether the annotators thought the input and our result could be a photograph of the same person, the responses to the options \{Definitely No, Slightly No, Slightly Yes, Definitely Yes\} were selected \{18.71, 19.1, 30.67, 31.6\} percent of the time. Despite the large gap in appearance, and the fact that our results are designed to alter aspects of the face like freckles which could be considered part of identity, roughly 60% still believed our results sufficiently similar to pass as photograph of the same person at the same moment in time.

Figure 9 shows some of our results compared to the input synthetic images and the baseline.

### 4.2 Quantitative Experiments

The preservation of important facial features is also assessed quantitatively by examining the alignment of 68 landmarks [4]. On our $1024^2$ images, the median absolute error in pixels is just 20.2 horizontally, and 14.2 vertically. We illustrate alignment errors per landmark in Figure 8. The results indicate that the biggest errors occur on the boundary of the face near the ears, and that in the face region the eyebrows and lips have the highest degree of misalignment. Since FFHQ contains mostly smiling subjects, or images of people with their mouth closed, it is unsurprising that the diverse facial expressions from the synthetic data would show the greatest discrepancy in these features. We emphasize however, that these errors are less than two percent of the effective image resolution on average.

We also compute both the FID [16] and IS [32] metrics. The results are shown in Table 10 for the large and small crops used throughout this paper. Our method improves the FID significantly compared to the baseline, and slightly in
Fig. 7: Representative comparison of results of our zero-shot method vs CycleGAN trained on the whole synthetic dataset. Note how CycleGAN is unable to change the input images enough to make them look realistic. We suggest viewing this figure zoomed in.
Fig. 9: Representative comparison of results of our method vs the StyleGAN2 Baseline. Both variants of our method are able to produce substantially more realistic samples with much greater detail. We suggest viewing this figure zoomed in.
case of the IS. This backs up the user study in terms of the perceptual plausibility of our results, but a larger number of samples would be beneficial for a conclusive result.

The FID difference between Ours and Ours (only face) shows that the background can significantly impact this metric, which is not reflected in human assessment.

|                | IS   | FID (to FFHQ) |                | IS   | FID (to FFHQ) |
|----------------|------|---------------|----------------|------|---------------|
| SG2 Baseline   | 3.4965 | 90.342       | SG2 Baseline   | 3.398 | 78.185       |
| Ours (only face) | 3.3187 | 78.728       | Ours (only face) | 3.464 | 70.731       |
| Ours           | **3.653** | 81.06       | Ours           | 3.435 | 76.947       |

(a) IS and FID (to FFHQ) - large crop. (b) IS and FID (to FFHQ) - tight crop.

Fig. 10: We note that Ours (only face) uses the same background and hair as the synthetic renders, while Ours replaces the entire image with our fit. We hypothesise that the reason our results, which retain the background and hair from the synthetic renders, score better than those where the full image is optimised in terms of the FID, is because the background HDRs are visible in the renders. As these are natural images, we expect this to impact the metrics to some degree. Note that we resize the large crop to match CycleGAN resolution when calculating the IS.

5 Conclusions

We have presented a novel zero-shot algorithm for improving the realism of non-photorealistic synthetic renders of human faces. The user study indicates that it produces images which look more photorealistic than the synthetic images themselves. It also shows that previous work on embedding images in the StyleGAN2 latent space produces results of inferior visual quality.

This result is reflected in quantitative terms in both the FID and IS metrics comparing our result to real images from FFHQ. CycleGAN, having access to a large dataset of synthetic images which our method never sees, and working on an inherently easier crop, is clearly not able to compare with our results qualitatively or quantitatively as well. A downside of our method is that it requires substantial processing time per image. We hypothesise that this could be amortised by training a model that predicts the StyleGAN2 embeddings directly from synthetic images once a large enough dataset has been collected. We leave temporal consistency for processing animations as future work, and show more results as well as failure cases (for which the algorithm can simply be repeated with a different random seed) in the supplementary material.
We would like to conclude by noting that our algorithm works across a wide range of synthetic styles (due to its zero-shot nature), and even with some non-photoreal images. Examples of this can be seen in Figure 11.

Fig. 11: Our method applied to synthetic characters from popular culture. Left to right, row by row, these are: Nathan Drake from Uncharted, Geralt of Rivia from the Witcher, Flynn Rider from Tangled, Aloy from Horizon Zero Dawn, Grand Moff Tarkin from Rogue One, and Ellie from The Last of Us. We note that this is the output of only step 1 and 2 of our method. This indicates that we can find visually plausible nearest neighbours even with some exaggerated facial proportions.
References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? (2019)
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? CoRR abs/1904.03189 (2019). http://arxiv.org/abs/1904.03189
3. AlBahar, B., Huang, J.B.: Guided image-to-image translation with bi-directional feature transformation (2019)
4. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018). pp. 59–66 (May 2018). https://doi.org/10.1109/FG.2018.00019
5. Benaim, S., Wolf, L.: One-shot unsupervised cross domain translation. CoRR abs/1806.06029 (2018). http://arxiv.org/abs/1806.06029
6. Béard, P., Bradley, D., Gross, M., Beeler, T.: Lightweight eye capture using a parametric model. ACM Trans. Graph. 35(4) (Jul 2016). https://doi.org/10.1145/2897824.2925962 http://doi.acm.org/10.1145/2897824.2925962
7. Bishop, C.M.: Mixture density networks. Tech. rep., Citeseer (1994)
8. Burt, P.J.: Fast filter transform for image processing. Computer Graphics and Image Processing 16(1), 20 – 51 (1981). https://doi.org/https://doi.org/10.1016/0146-664X(81)90092-7 http://www.sciencedirect.com/science/article/pii/0146664X81900927
9. BURT, P.J., ADELSON, E.H.: The laplacian pyramid as a compact image code. In: Fischler, M.A., Firschein, O. (eds.) Readings in Computer Vision, pp. 671 – 679. Morgan Kaufmann, San Francisco (CA) (1987). https://doi.org/https://doi.org/10.1016/B978-0-08-051581-6.50065-9 http://www.sciencedirect.com/science/article/pii/B9780080515816500659
10. Cherian, A., Sullivan, A.: Sem-gan: Semantically-consistent image-to-image translation. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1797–1806 (Jan 2019). https://doi.org/10.1109/WACV.2019.00196
11. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., Tao, D.: Geometry-consistent adversarial networks for one-sided unsupervised domain mapping. CoRR abs/1809.05852 (2018). http://arxiv.org/abs/1809.05852
12. Gajane, P.: On formalizing fairness in prediction with machine learning. CoRR abs/1710.03184 (2017). http://arxiv.org/abs/1710.03184
13. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. CoRR abs/1811.12231 (2018). http://arxiv.org/abs/1811.12231
14. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks. ArXiv e-prints (Jun 2014)
15. Gutmann, M., Hyvänni, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Teh, Y.W., Titterington, M. (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 9, pp. 297–304. PMLR, Chia Laguna Resort, Sardinia, Italy (13–15 May 2010). http://proceedings.mlr.press/v9/gutmann10a.html
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a nash equilibrium. CoRR abs/1706.08500 (2017). \url{http://arxiv.org/abs/1706.08500}

17. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: International Conference on Computer Vision (ICCV), Venice, Italy (2017). \url{https://vision.cornell.edu/se3/wp-content/uploads/2017/08/adain.pdf} oral

18. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: The European Conference on Computer Vision (ECCV) (September 2018)

19. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CoRR abs/1611.07004 (2016). \url{http://arxiv.org/abs/1611.07004}

20. Jahanian, A., Chai, L., Isola, P.: On the ”steerability” of generative adversarial networks. CoRR abs/1907.07171 (2019). \url{http://arxiv.org/abs/1907.07171}

21. Jimenez Rezende, D., Mohamed, S., Wierstra, D.: Stochastic Backpropagation and Approximate Inference in Deep Generative Models. ArXiv e-prints (Jan 2014)

22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CoRR abs/1812.04948 (2018). \url{http://arxiv.org/abs/1812.04948}

23. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan (2019)

24. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1867–1874 (June 2014). \url{https://doi.org/10.1109/CVPR.2014.241}

25. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. ArXiv e-prints (Dec 2013)

26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

27. Larochelle, H., Murray, I.: The neural autoregressive distribution estimator. In: The Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR: W and CP, vol. 15, pp. 29–37 (2011)

28. Lin, J., Xia, Y., Liu, S., Qin, T., Chen, Z.: Zstgan: An adversarial approach for unsupervised zero-shot image-to-image translation. CoRR abs/1906.00184 (2019). \url{http://arxiv.org/abs/1906.00184}

29. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

30. Mori, M., MacDorman, K., Kageki, N.: The uncanny valley. IEEE Robotics and Automation Magazine \textbf{19}, 98–100 (06 2012). \url{https://doi.org/10.1109/MRA.2012.2192811}

31. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR abs/1511.06434 (2015). \url{http://arxiv.org/abs/1511.06434}

32. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. CoRR abs/1606.03498 (2016). \url{http://arxiv.org/abs/1606.03498}

33. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. CoRR abs/1503.03585 (2015). \url{http://arxiv.org/abs/1503.03585}
34. Sutherland, D.J., Tung, H.Y., Strathmann, H., De, S., Ramdas, A., Smola, A., Gretton, A.: Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy. ArXiv e-prints (Nov 2016)
35. Uria, B., Murray, I., Larochelle, H.: RNADE: the real-valued neural autoregressive density-estimator. In: Advances in Neural Information Processing Systems 26, pp. 2175–2183 (2013)
36. Wang, C., Zheng, H., Yu, Z., Zheng, Z., Gu, Z., Zheng, B.: Discriminative region proposal adversarial networks for high-quality image-to-image translation. CoRR abs/1711.09554 (2017), http://arxiv.org/abs/1711.09554
37. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans (2017)
38. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thirteenth Asilomar Conference on Signals, Systems Computers, 2003. vol. 2, pp. 1398–1402 Vol.2 (Nov 2003). https://doi.org/10.1109/ACSSC.2003.1292216
39. Wrenninge, M., Villemin, R., Hery, C.: Path traced subsurface scattering using anisotropic phase functions and non-exponential free flights. Tech. rep.
40. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. CoRR abs/1801.03924 (2018), http://arxiv.org/abs/1801.03924
41. Zhang, R., Pfister, T., Li, J.: Harmonic unpaired image-to-image translation. CoRR abs/1902.09727 (2019), http://arxiv.org/abs/1902.09727
42. Zhao, S., Ren, H., Yuan, A., Song, J., Goodman, N.D., Ermon, S.: Bias and generalization in deep generative models: An empirical study. CoRR abs/1811.03259 (2018), http://arxiv.org/abs/1811.03259
43. Zheng, Z., Yu, Z., Zheng, H., Yang, Y., Shen, H.T.: One-shot image-to-image translation via part-global learning with a multi-adversarial framework. CoRR abs/1905.04729 (2019), http://arxiv.org/abs/1905.04729
44. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)