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 hair, skin [45], eyes [7] 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 [36].

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) [28] 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,

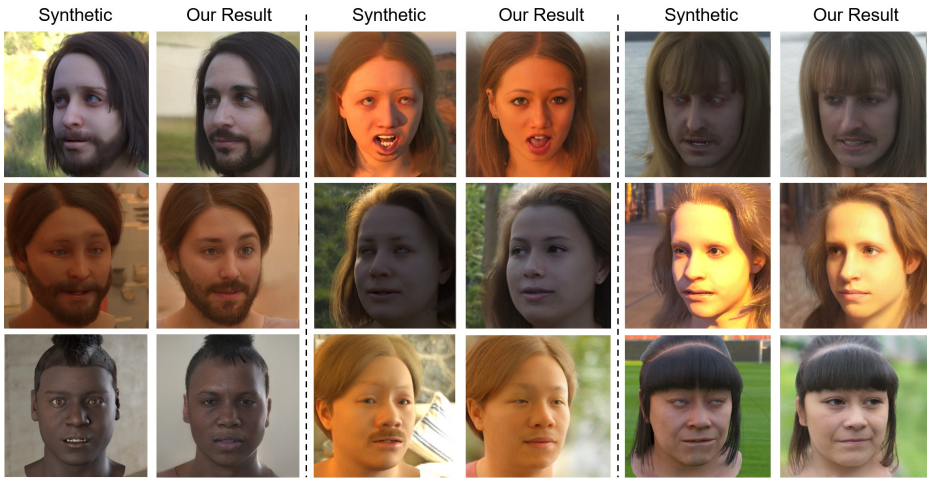

Fig. 1: Pairs of synthetic input images, and samples from our algorithm at 1K resolution. Best viewed zoomed in, and in colour.

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 [48]. 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 as consistent as possible given constraints imposed by the training data. This means that the annotations obtained from the renders are still largely 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 other work using non-photorealistic renders to train models that map from one domain to another (e.g. [17] for faces, or [8]), 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 (generated with [5]), 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 [10, 11], 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 [25].

While embedding images in the latent space of SG2 is not a new concept [1, 2], 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 [37]. 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 MMD [40], Noise Contrastive Estimation [20], Mixture Density Networks [9], Neural Autoregressive Distribution Estimators [33, 41], Diffusion Process Models [39], and flow-based models [13, 14, 32]), the most popular ones are the family of Variational Autoencoders (VAEs) [26, 30], and Generative Adversarial Networks (GANs) [19].

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) [28]. In any GAN, a neural sampler called the generator is trained to map samples from a simple distribution to the true data distribution, defined by samples (the training set). 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 algorithm. Since SG2 uses a variant of Adaptive Instance Normalisation (AdaIn)[22], 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 [2]. The authors in [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 methods in the literature that require only one source domain operating at comparable resolution. By domain adaptation we mean the ability to make images from dataset A look like images from dataset B, while preserving content. While one-shot methods like [49] or [6] have been proposed, they work at significantly lower resolution than ours and still require one sample from the target domain. ZstGAN [34], 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 signifantly 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 [24] and its successors (e.g. Pix2Pix HD [43], which uses multiple discriminators at multiple scales to produce high resolution images) can be used effectively for domain adaptation.

CycleGAN does not require paired training data [50]. 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 suggeted to improve CycleGAN. HarmonicGAN adds an additional smoothness constraint to reduce artefacts in the outputs [47], Sem-GAN exploits additional information [12], as does [3], Discriminative Region Proposal Adversarial Networks (DRPAN) [42] add steps to fix errors, Geometry-Consistent GANs (GcGAN) [15] use consistency under simple transformations as an additional constraint. Some methods also model a distribution of over possible outputs, such as MUNIT [23] or FUNIT [35].

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

3 Method

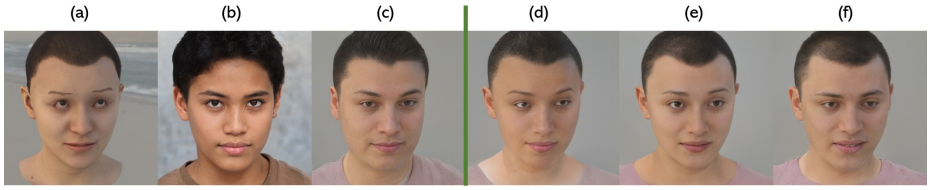

Fig. 2: Illustration of the different steps of our method. (a) is the input synthetic render, (b) the output of the sampling in step 1, (c) the result of Convex Set Approximate Nearest Neighbour Search in step 2, and (d-f) results from step 3.

In the following, any variable containing w refers to the 18×512 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 ws 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. Please note that additional details and results can be found in the supplementary material.

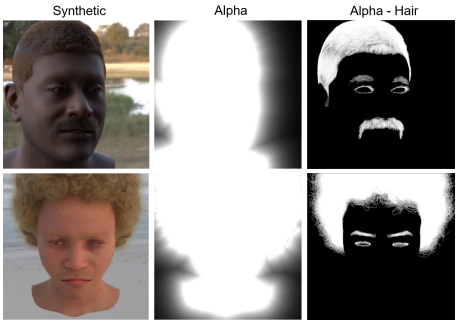

Fig. 3: Example tuples of renders and alpha masks, $\{I^s, I^a, I^{a_{hair}}\}$, derived from synthetic images. Note that we apply a falloff at sharp boundaries to preserve them as described in the text.

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 [27], 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 $\{I^s, I^a, I^{a_{hair}}\}$ 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, for a finite number of samples, 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 conclusively that this leads to greater overall fairness [16], and acknowledge that this sensitive issue needs closer examination in future work.

Fig. 4: Example of adding our control vectors to the 'mean' face of SG2: (a) face angle; (b) hair length (including headgear); (c) beard length; (d) hair curlyness. Note that these can be found by only rough annotations of a small number of samples.

The loss used in the sampling step is a combination of the LPIPS distance [46], an L1 loss with different weights for colour and luminance, and landmark loss based on 68 points computed with DLIB [29].

This loss is computed at a quarter resolution of 256^2 pixels after low pass filtering, and multiplication with the mask I^a. We do not compute the loss at full resolution because our synthetics do not exhibit fine-scale details, and we use a low-pass filter to not penalise their presence in the result. The entire loss function for the sampling step is thus:

$$\begin{aligned} L_{sampling} = L_{LPIPS}(r(I^s * I^a), r(G(w^s) * I^a)) \\ +\lambda_{lum} * \|y(r(I^s * I^a)) - y(r(G(w^s) * I^a))\|^1 \\ +\lambda_{col} * \|u(r(I^s * I^a)) - u(r(G(w^s) * I^a))\|^1 \\ +\lambda_{landm} * \|l(r(I^s * I^a)) - l(r(G(w^s) * I^a))\|^2, \end{aligned} \tag{1}$$

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. λ_{lum} is set to 0.1, λ_{col} to 0.01, and λ_{landm} to $1e-5$.

For each sample at this stage of our method, we pick one of the centroids $v_{centroid}$ 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:

$$\begin{aligned} w_i^s = s(v_{centroid}) + \mathcal{N}(0.0, \sigma^2) \\ +v_{control} * N_{uniform} * 2.0, \end{aligned} \tag{2}$$

where s is the random centroid selection function, $\sigma^2 = 0.25$, and $N_{uniform}$ 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.

Fig. 5: Our manually curated set of centroids for sampling.

3.2 Step 2: Latent Code Refinement

In step 2, we refine the previously obtained w^s while keeping the results constrained to the set of photorealistic images the SG2 generator can produce. 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. Each of these sample proposals w^p is obtained as:

$$w_i^p = s(v_{centroid}) + \mathcal{N}(0.0, \sigma^2). \tag{3}$$

To ensure samples are sufficiently close to the current w^n, we interpolate each w_i^p with w^n using a random weight drawn at uniform from the range $0.25 - 0.75$.

We then optimise for a set of weights, α, which determine how the w_i^ps and current w^n are combined. It is an important detail that we use sets of α for each of the 18 StyleGAN2 latent space inputs for this optimisation, *i.e.* α is a matrix of shape $[512 + 1, 18]$ (note how the current w^n is included).

We constrain the optimisation to make sure each row of α sums to 1 using the softmax function, ensuring a convex combination of the samples. In addition to α, we include the control vectors in the optimisation, which are scaled by a learnable parameter β. Because this last step could potentially lead to solutions far outside the space of plausible images, we clamp β to 2.0. The loss is the same as Equation 1, just without the non-differentiable landmark term, *i.e.* with λ_{landm} set to 0.

We use 96 outer iterations for which the sample proposals w^p are redrawn, and α and β reset so that the current w^n is the starting point (*i.e.* β is set to zero, and *alpha* to one only for the current w^n). For each of these outer loops, we optimise α and β using Adam [31] 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 [28] 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 apprximate 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:

$$w^{final} = w_s * \sqrt{\alpha} + w_n * \sqrt{1.0 - \alpha}, \tag{4}$$

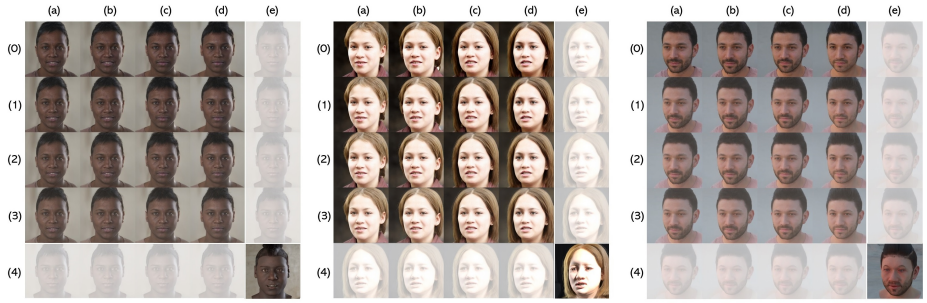

Fig. 6: Interpolating beteen the output of step 2 ($a0$) and the output of the exact fit to the synthetics ($e4$). ($a - e$) represent the number of latent codes used for blending, and ($0 - 4$) the floating point weights for them.

where w^{final} is a candidate for the final output of our method. We generate candidates by letting α 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: Result 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) [44] metric at a resoluton 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 [18], 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 [28], wich 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 [50], 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 [11]. 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 quantitive 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 8 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 examing 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.

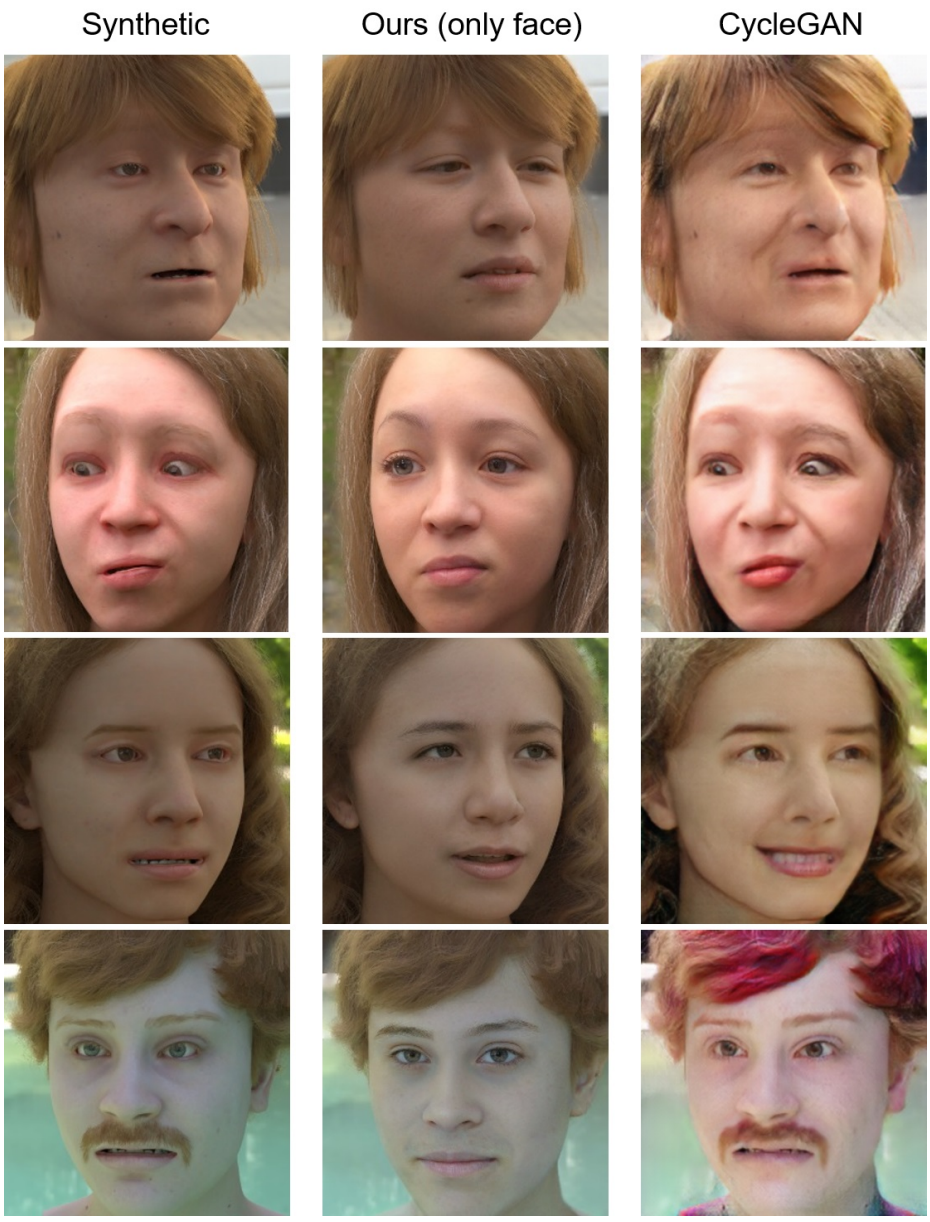

Fig. 7: Representative comparison of results of our *zero-shot* method vs CycleGAN trained on the whole synthetic dataset. Note how CycleGAN is unable to change the input images enough to make them look realistic. We suggest viewing this figure zoomed in.

Fig. 8: 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.

Fig. 9: *Ours (only face)* uses the same background and hair as the renders, while *Ours* replaces the entire image with our fit. We hypothesise that the difference in FID between *Ours (only face)* and *Ours* is because the background HDRs are visible in the renders, and those backgrounds are photographs. Note that we resize the large crop to match CycleGAN resolution when calculating the IS.

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

	IS	FID (to FFHQ)
SG2 Baseline	3.4965	90.342
Ours (only face)	3.3187	**78.728**
Ours	**3.653**	81.06

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

	IS	FID (to FFHQ)
SG2 Baseline	3.398	78.185
Ours (only face)	**3.464**	**70.731**
Ours	3.435	76.947

We illustrate alignment errors per landmark in Figure 10. 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 unsurpising 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.

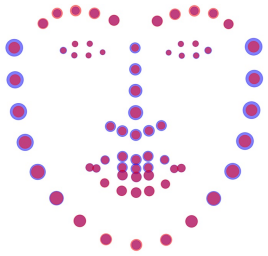

Fig. 10: Standard deviation of the L_1 landmark error in our results (scaled 20× for figure). Blue/red = horizontal/vertical error.

We also compute both the FID [21] and IS [38] metrics. The results are shown in Table 9 for the large and small crops used throughout this paper. Our method improves the FID signifcantly compared to the baseline, and slightly 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.

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 quantitively 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. Baltrusaitis, T., Wood, E., Estellers, V., Hewitt, C., Dziadzio, S., Kowalski, M., Cashman, T., Johnson, M., Shotton, J.: A high fidelity synthetic face framework for computer vision. Tech. Rep. MSR-TR-2020-24, Microsoft (July 2020), https://www.microsoft.com/en-us/research/publication/high-fidelity-face-synthetics/
6. Benaim, S., Wolf, L.: One-shot unsupervised cross domain translation. CoRR **abs/1806.06029** (2018), http://arxiv.org/abs/1806.06029
7. Bérard, 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, https://doi.org/10.1145/2897824.2925962
8. Bi, S., Sunkavalli, K., Perazzi, F., Shechtman, E., Kim, V.G., Ramamoorthi, R.: Deep cg2real: Synthetic-to-real translation via image disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
9. Bishop, C.M.: Mixture density networks. Tech. rep., Citeseer (1994)
10. 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
11. 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
12. 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
13. Dinh, L., Krueger, D., Bengio, Y.: Nice: Non-linear independent components estimation (2014)
14. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp (2016)
15. 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
16. Gajane, P.: On formalizing fairness in prediction with machine learning. CoRR **abs/1710.03184** (2017), http://arxiv.org/abs/1710.03184

17. Gecer, B., Bhattarai, B., Kittler, J., Kim, T.: Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model. CoRR **abs/1804.03675** (2018), http://arxiv.org/abs/1804.03675
18. 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
19. 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)
20. Gutmann, M., Hyvärinen, 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
21. 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), http://arxiv.org/abs/1706.08500
22. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: International Conference on Computer Vision (ICCV), Venice, Italy (2017), https://vision.cornell.edu/se3/wp-content/uploads/2017/08/adain.pdf, oral
23. 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)
24. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CoRR **abs/1611.07004** (2016), http://arxiv.org/abs/1611.07004
25. Jahanian, A., Chai, L., Isola, P.: On the "steerability" of generative adversarial networks. CoRR **abs/1907.07171** (2019), http://arxiv.org/abs/1907.07171
26. Jimenez Rezende, D., Mohamed, S., Wierstra, D.: Stochastic Backpropagation and Approximate Inference in Deep Generative Models. ArXiv e-prints (Jan 2014)
27. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CoRR **abs/1812.04948** (2018), http://arxiv.org/abs/1812.04948
28. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan (2019)
29. 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). https://doi.org/10.1109/CVPR.2014.241
30. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. ArXiv e-prints (Dec 2013)
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
32. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions (2018)
33. 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)

34. 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), http://arxiv.org/abs/1906.00184
35. 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)
36. Mori, M., MacDorman, K., Kageki, N.: The uncanny valley. IEEE Robotics and Automation Magazine **19**, 98–100 (06 2012). https://doi.org/10.1109/MRA.2012.2192811
37. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR **abs/1511.06434** (2015), http://arxiv.org/abs/1511.06434
38. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. CoRR **abs/1606.03498** (2016), http://arxiv.org/abs/1606.03498
39. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. CoRR **abs/1503.03585** (2015), http://arxiv.org/abs/1503.03585
40. 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)
41. 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)
42. 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
43. 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)
44. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003. vol. 2, pp. 1398–1402 Vol.2 (Nov 2003). https://doi.org/10.1109/ACSSC.2003.1292216
45. Wrenninge, M., Villemin, R., Hery, C.: Path traced subsurface scattering using anisotropic phase functions and non-exponential free flights. Tech. rep.
46. 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
47. Zhang, R., Pfister, T., Li, J.: Harmonic unpaired image-to-image translation. CoRR **abs/1902.09727** (2019), http://arxiv.org/abs/1902.09727
48. 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
49. 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
50. 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)