ual: creative computing institute

Creative Making: MSc Advanced Project

University of the Arts London
Creative Computing Institute

Expandable StyleGAN3 Canvas for Image-to-image Translation and Expressive Feature Exploration

Author
Shuoyang Zheng (Jasper)

21009460

Supervisor
Prof. Mick Grierson

Link to the video presentation
https://vimeo.com/773414893

November 21, 2022

Acknowledgements

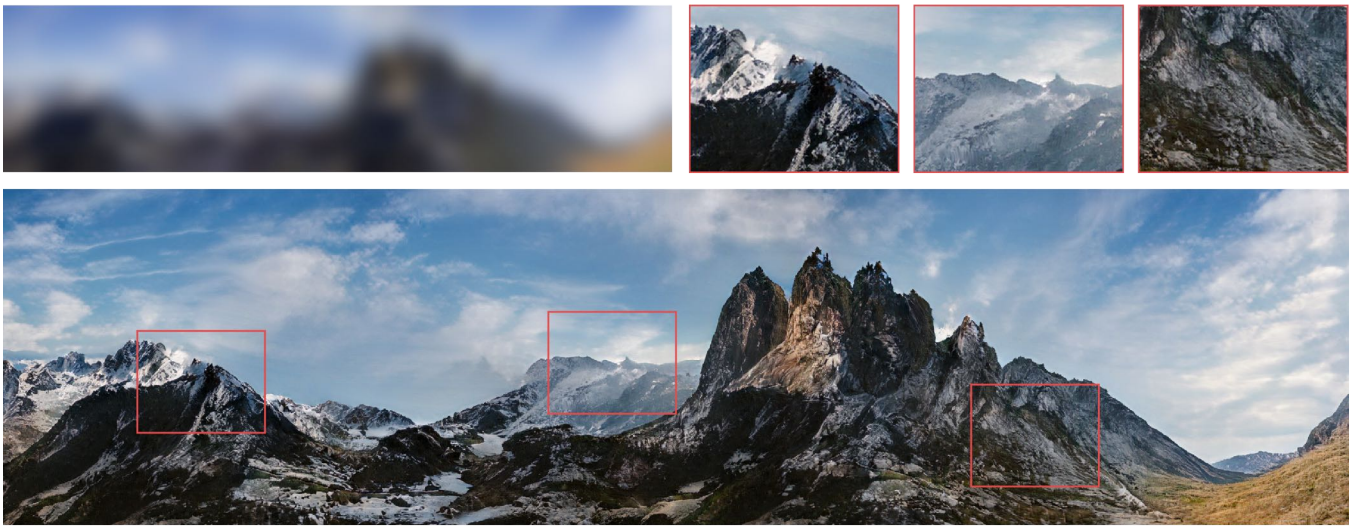
I would like to thank my supervisor Prof. Mick Grierson, for allowing me to work on his workstation. And I sincerely appreciate the treasured feedback and comments from him.

My appreciation also goes to my parents for their continuous support during my study.

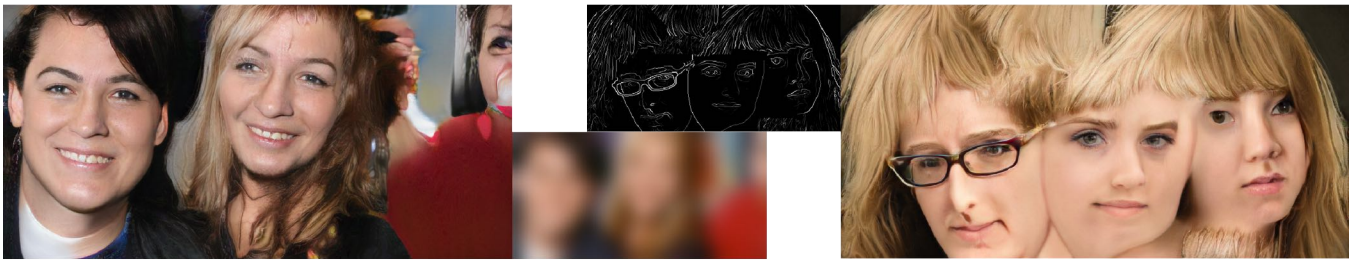
My gratitude extends to Moodie and Lucy for their encouragement.

Contents

Expandable StyleGAN3 Canvas for Image-to-image Translation and Expressive Feature Exploration



(a)



(b) (c)

Figure 1: The proposed image-to-image translation framework can be trained on a lower resolution and later expand to a larger canvas. (a) The deblurring model is trained on 512×512 images and expanded to 512×2048. (b) and (c) The extendable canvas has broadened the possibility for exploration and creation, allowing more space for expressive outcomes to emerge.

ABSTRACT

We present a new framework for real-time feature exploration and manipulation using style-based image-conditional generative adversarial networks (image-conditional StyleGANs). Our framework follows previously introduced StyleGAN3 architecture, extends it for image-to-image translation tasks by appending a feature extraction encoder that creates low-level skip connections to the generator. We first demonstrate that our framework solves a variety of image-to-image translation tasks while maintaining the image quality and the internal behaviour of StyleGAN3. Our approach also induces an extendable canvas that can be trained on a lower resolution and later expanded to a larger resolution. Next, implement our framework on network bending, and build a graphic interface for real-time interaction with the model's internal features. Finally, we conduct qualitative human opinion studies to evaluate its usability in the creative context, and demonstrate its potential to broaden the possibility for expressive outcomes to emerge. Code is available at https://github.com/jasper-zheng/pix2styleGAN3

1 INTRODUCTION

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have been rapidly developed recently and have become a powerful tool for creating high-quality digital artefacts. In the case of images, modern approaches to improve model quality have successfully brought the generated outcomes from coarse low-res interpretations to realistic portraits with high diversity (Sauer et al., 2022) and visual fidelity (Karras, Aittala, Laine, Härkönen, Hellsten, Lehtinen and Aila, 2021). Meanwhile, deep generative models have been found to be widely used as a tool for artists to introduce novel and surprising visual aesthetics (Berns et al., 2021).

Berns and Colton (2020) presented active divergence, a research theme of reusing encoded knowledge in trained models and intervening in the generative process in pursuing creativity. Later, more ways of re-organising and rewriting features in generative models were highlighted and taxonomised by Broad et al. (2021). Although the ways of exploring and evaluating features in current approaches remain partially in a blackbox (Bau, Zhu, Strobelt, Lapedriza, Zhou and Torralba, 2020), works from artists (Schultz, 2020, Som, 2020) have been attempting to bridge the gap between autonomous network decisions and computational creativity (Berns and Colton, 2020). Their works demonstrated unexplored potential for reusing features and diverging from the approximate distribution in a network. They have also outlined the need for more predictable possibilities (Broad et al., 2020) and human intervention (Hertzmann, 2020) during inference. Motivated by active divergence, we propose an alternative approach to image-to-image translation that allows the features to be enlarged and interpolated after the network is trained. This provides an extendable generation canvas and significantly broadens the possibility for creative outcomes.

Our research aims to: (i) Extend the current state-of-the-art GAN-based generative model to expandable image-to-image translation; and (ii) Investigate the potential of our models in generic creative contexts.

The proposed architecture includes an appended encoder network facilitating feature extraction and the StyleGAN3 mapping and synthesis network (Karras, Aittala, Laine, Härkönen, Hellsten, Lehtinen and Aila, 2021). The encoder is based on a U-net architecture with standard ResNet backbone, encoding an image into a 512-dimensional latent vector. Meanwhile, it creates skip connections that directly map features from the encoder into the StyleGAN3 synthesis network to preserve locality bias. We found that adding skip connections between only lower-level layers (i.e. feature maps with smaller resolution) is sufficient for it to shuttle precise features while preserving the internal behaviour of StyleGAN3 (i.e. encode phase information instead of signal magnitudes). In addition, as StyleGAN3 replaced the first generator layer with a frequency sampler with Fourier features to define a spatially infinite map, and therefore facilitate translation and rotation of the input, however, we propose extracting the first generator layer directly from the encoder, allowing the generator to inherit exact sub-pixel position. We show results from models with this architecture trained for a variety of image-to-image translation tasks while maintaining the image quality of StyleGAN3.

After defining the base image-conditional generative model, we follow the spirit of active divergence and extend our framework to facilitate feature exploration and manipulation. We propose implementing our framework on network bending (Broad et al., 2020), which is an approach for grouping and manipulating features via deterministic transformations to create semantically meaningful divergences. First, we build a graphic interface to facilitate the interaction between users and the model, allowing them to explore internal features in a real-time manner. Next, motivated by the network bending approach, we implement the clustering algorithm to group similar features together, and allow the user to insert and alter transformations to the computational graph interactively. Finally, we conduct qualitative-based human opinion studies on our framework. We show that it can potentially facilitate feature exploration to create expressive outcomes, providing a certain level of control over the inference while triggering unpredictable possibilities.

2 RELATED WORKS

2.1 Alias-Free GAN

Our appended network builds on StyleGAN3 (Alias-Free GAN). This section reviews the background of style-based generators and the improvements over the second (StyleGAN2) and third (StyleGAN3) upgrades. StyleGAN3 (Alias-Free GAN) is the latest style-based GAN architecture with dramatically different internal representations compared with previous versions (Karras, Aittala, Laine, Härkönen, Hellsten, Lehtinen and Aila, 2021). It interprets signals in the network as continuous representations, equivariant to translation and rotation. Therefore, the model can be trained on unaligned data, and eliminate "texture sticking" behaviours in standard GAN architectures. Practically, the generator consists of a mapping network M that transforms the initial latent code z to intermediate latent code $w \sim W$, and a synthesis network G that first applies transformation t on continuous input map $z_0 = t[g(z_0, w)]$, where t is based on a learned affine layer that output the translation and rotation parameter given w. Then the generator performs a sequence of N synthesis blocks, each controlled by the intermediate latent code w, and produces the output image $z_N = G(z_0; w)$. Our adapted architecture aims to sample transformed z_0 directly from the input image using an appended encoder.

Style Mixing. Style mixing was proposed as a key feature in StyleGAN (Karras et al., 2018), i.e. amplifying specific feature maps on a per-sample basis by adaptive instance normalisation (AdaIN). In practice, apply learned affine transform on intermediate latent space W and obtain styles y, which are then used as scalar components to modulate and bias the corresponding feature maps. This architecture was then revised in StyleGAN2 (Karras et al., 2019) by replacing instance normalisation with feature map demodulation, and inherited by StyleGAN3.

Transformed Fourier features. StyleGAN3 replaces discrete internal representation in the network with continuous signal. It introduces equivariant generator layers to model unaligned and oriented data. These layers are equivariant to transformations (i.e. translations and rotations) applied on intermediate continuous signals.

The input feature map $z0$ defines the global transformations of z_N. It is a spatially infinite map sampled from uniformly distributed frequencies, fixed after initialisation. The translation and rotation parameters of z_0 are calculated by a learned affine layer based on intermediate latent space W. It acts as a coordinate map that allows later layers to grab on and therefore defines a positional encoding pattern (Alaluf et al., 2022).

2.2 Image-to-Image Translation

Image-to-image translation methods aim to generate a corresponding image given an image from the source domain, depicting objects or scenes in different styles or conditions. Recent solutions to image-to-image translation usually fall into two directions: (i) utilising conditional GANs with a U-net architecture (Ronneberger et al., 2015), first introduced by Isola et al. (2016), using the residual

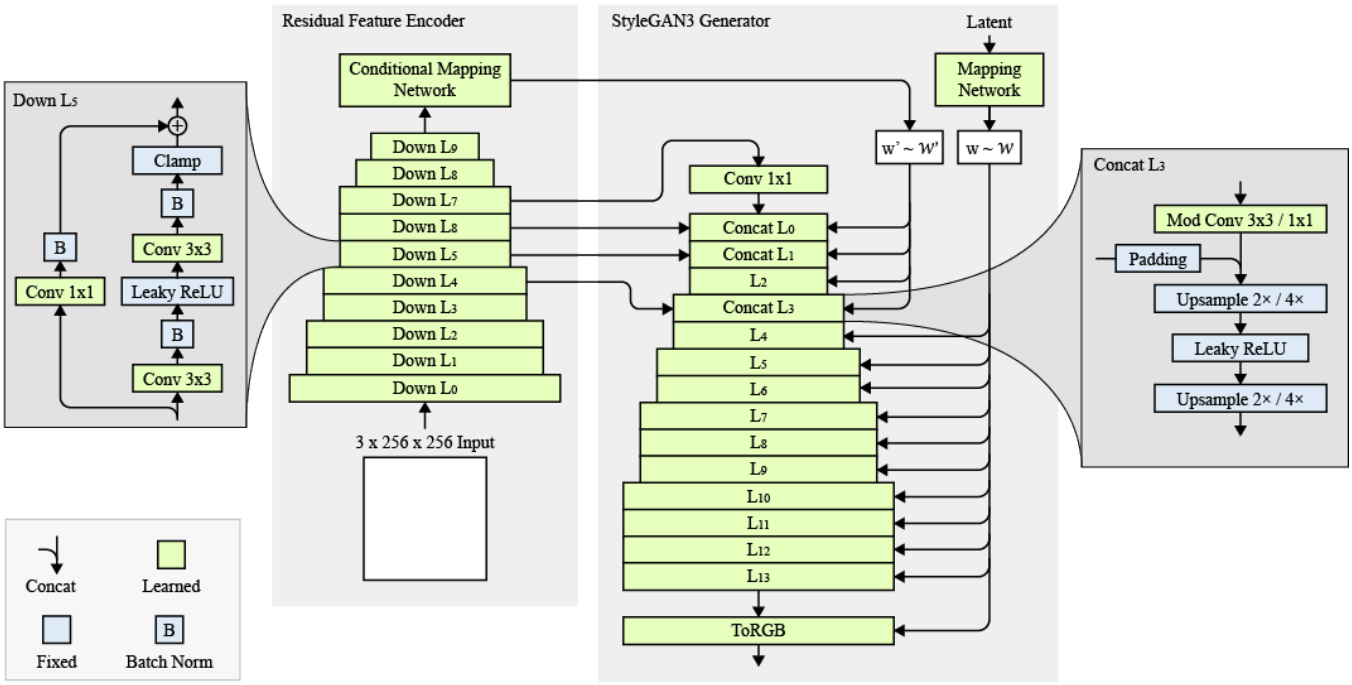


Figure 2: We build a residual feature encoder and adapt the StyleGAN3 generator. The main datapath consists of (i) downsampling residual blocks (Section 3.1), each consisting of a mapped shortcut with a 1x1 convolutional layer and bath normalisation, and a downsampling block with two convolutional layers, an activation layer (Leaky ReLU), bath normalisations and a clamping layer, (ii) a conditional mapping network (Section 3.1), (iii) StyleGAN3 mapping network, (iv) adapted StyleGAN3 synthesis blocks (Section 3.2).

feature maps from the encoder as skip connections. This method aims to provide the generator with a mechanism to circumvent the bottleneck layer and create precise translation (Isola et al., 2017). This method has been extended for various tasks such as high-resolution synthesis (Wang et al., 2017), line-conditioned generation (Li et al., 2019), layout to image synthesis (Liu et al., 2019), semantic region-adaptive synthesis (Zhu et al., 2020). Although this method usually creates strong locality bias, and was later revised in the pix2pixHD method, we found it still provides effective results when combined with StyleGAN generators. (ii) using a linear encoder-decoder architecture with no spatial dimension, which was introduced by Richardson and Weiss (2020). This method operates globally and without requiring pixel-to-pixel correspondence. It was later extended to a generic framework for a variety of tasks such as sketches and layout to image, facial frontalization, inpainting and super-resolution (Richardson et al., 2020).

2.3 Network Bending and the Reuse of Features

Network bending (Broad et al., 2020) is a method that manipulates the feature maps in a trained model to create active divergences (Broad et al., 2021). It alters a model's computational graph by applying image filters to the convolutional features during inference, therefore allowing the model to generate novel samples that are diverse from the training data. It has been used to create novel artefacts, e.g., StyleGAN Playground (Pinkney, 2020), GAN Bending (Pouliot, 2020), 99 Ways to use a Dataset (Schultz, 2019), as well as extend to the domain of audio synthesis (McCallum and Yee-King, 2020). Previous works on GAN Dissection (Bau et al., 2018), GANPaint (Bau et al., 2019), and model rewriting (Bau, Liu, Wang, Zhu and Torralba, 2020) have indicated that the knowledge encoded in a deep neural network is semantically related to the spatial information in its feature maps. Therefore, the approach of re-organising and reusing these features can potentially lead to semantically meaningful outcomes. Likewise, in network bending, Broad et al. (2020) present a clustering algorithm that groups spatially similar features and applies filters to selected groups. It can discover sets of features responsible for generating specific components of the generated output. Our interactive framework aims to facilitate the exploration and exploitation of features. By liberating the tool for artists without prior knowledge of generative models, we investigate how they would end up utilising this approach to create works.

3 EXTENDING STYLEGAN TO IMAGE-TO-IMAGE TRANSLATION

As mentioned in Section 2.1, current StyleGAN3 models learn a mapping from random noise vector z to image $Z_N = G(z)$. To extend the input from vector z to the ground truth image x combined with z, we first need a feature extraction encoder E that extracts features from x, then adapts the generator G to produce outputs based on the features and the input vector z, i.e. $Z_N = G(E(x), z)$.

Besides, the training objective should consider the input image and push the generator closer to the ground truth image x.

3.1 Residual Feature Encoder

ResNet. ResNet (He et al., 2015) architecture has been used in previous works on image-to-image translation for feature maps extraction (Richardson et al., 2020). Following this approach, as shown in Figure 2 (left), the residual feature encoder employs an adapted ResNet architecture as the encoder backbone. The encoder downsamples feature maps to $(x/2^6, y/2^6)$, where x and y denote the width and height of the input image.

Mixed-precision. In addition, StyleGAN employs mixed-precision FP16/FP32 training for optimisation, where the generator casts precision to FP16 for the four highest-resolution layers. And it results in about 60% speedup while maintaining virtually identical generation outputs. We utilise similar techniques and reduce the precision to FP16 for the first five residual blocks. Therefore, it also requires pre-normalisation and an extra clamping layer that clamps the output of the convolutional layers to $\pm 2^9$.

Conditional mapping network and combined latent vector. Previous works in StyleGAN encoder have highlighted that replacing select layers of the extended latent space $W+$ by computed latents can facilitate multi-modal synthesis. And the extended latent space $W+$ can be roughly divided into coarse, medium, and fine layers, corresponding to different levels of detail (Richardson et al., 2020) and editability (Mao et al., 2022). This motivates us to add a conditional mapping network as the epilogue layer of our residual feature encoder, which uses the same mapping network architecture in StyleGAN, but takes a flattened 512 vector sampled from the encoder's bottleneck and produces a replacement latent space $W+\prime$. And $W+\prime$ is then concatenated with a portion of the latent space $W+$ produced from the original mapping network. This aims to facilitate multi-modal generation.

3.2 Adapting Generator

Conditioned Synthesis Input Layer To model unaligned and randomly oriented data, StyleGAN3 utilise a fixed-size canvas cropped from an infinite spatial extent as feature maps (Karras, Aittala, Laine, Härkönen, Hellsten, Lehtinen and Aila, 2021). Meanwhile, the translation and rotation are determined by Fourier features in the first synthesis layer. Therefore, the desired input layer for the synthesis network should either learn the translation and rotation parameters and then apply the transformation on the canvas, or use the feature maps from the encoder and directly output the transformed canvas for later layers to grab on. However, experiments in Section 4.1 show that the latter method leads to more stable training.

U-net Architecture with Lower-Level Skip Connections The connection between the feature encoder aims to provide precise spatial information of the input image. A U-net (Ronneberger et al., 2015) architecture is well-suited for shuttling high-level details from the encoder to the decoder corresponding to low-level semantic feature maps. However, since the experiments on StyleGAN3 (Karras, Aittala, Laine, Härkönen, Hellsten, Lehtinen and Aila, 2021) have demonstrated that high-level feature maps in the synthesis network encode phase information instead of signal magnitudes, relying on fusing skip connections to the synthesis layer for high-level features will deviate from StyleGAN3's internal generation behaviour. To tackle this, we first move the feature fusion node from the end of each synthesis block to the point before the filtered nonlinearities layer, shown in Figure 2 (right). Additionally, research on U-Net and its variants has demonstrated that a simplified structure with fewer feature fusions can achieve reasonable results (Huang et al., 2020, Lu et al., 2022). Therefore, we reduce the number of feature fusions and limit connections to only the first five layers. In practice, skip connections in these five models each connect layer $n-i$ of the encoder to layer i, where n is the total number of layers in the encoder, and i is limited to $i \in (0, 5]$. This reduction in the network maintains unification of the network's internal behaviour, while taking advantage of the efficiency of U-shaped structural models.

3.3 Loss Functions

Standard StyleGAN loss function consists of the standard GAN loss function (i.e., logistic loss) and regularization terms (i.e., R_1). We incorporate the training objectives in StyleGAN with pixel-wise distance and perceptual loss that have been used in conditional GANs.

The model is trained using different combinations of objectives at two different training phases. The first phase starts from zero to the first 300k images, and this is also the phase where the training images are blurred with a Gaussian filter to prevent early collapses (Karras, Aittala, Laine, Härkönen, Hellsten, Lehtinen and Aila, 2021). During this phase, the pixel-wise loss L_2 distance between input images x and target images y, logistic loss L_{GAN} and regularization terms L_{reg} as follows:

$$L_2(G,E) = \mathbb{E}_{x,y,z}[\|y - G(E(x),z)\|_2] \tag{1}$$

$$\begin{aligned} L_{GAN}(D,G,E) &= \mathbb{E}_y[logD(y)] \\ &+\mathbb{E}_{x,z}[log(1 - D(G(E(x),z)))] \end{aligned} \tag{2}$$

$$L_{reg}(E,M) = \mathbb{E}_{x,z}[\|E(x) - \bar{w}\|_2 + \|M(z) - \bar{w}\|_2] \tag{3}$$

where G and D denote the generator and the discriminator, E denotes the feature encoder, and M denotes the mapping network in the generator. Then, the training loss L_{phase1} is defined as:

$$L_{phase1}(D,G,E) = \lambda_1 L_2(G,E) + L_{GAN}(D,G,E) + L_{reg}(E,M) \tag{4}$$

The second phase starts after the training reaches 300k images. We add a perceptual loss L_{VGG} utilising a pre-trained VGG19 (Simonyan and Zisserman, 2014) network, which has been used in the training of previous conditional GANs and has led to more finer details in the resulting images (Wang et al., 2017), defined as follow:

$$L_{VGG}(G,E) = \mathbb{E}_{x,y,z}[\|F(y) - F(G(E(x),z))\|_2] \tag{5}$$

Then, the phase 2 loss is then calculated as follows:

$$\begin{aligned} L_{phase2}(D,G,E) &= \lambda_1 L_2(G,E) + \lambda_2 L_{VGG}(G,E) \\ &+L_{GAN}(D,G,E) + L_{reg}(E,M) \end{aligned} \tag{6}$$

where F denotes the pre-trained VGG19 feature extractor. λ_1 and λ_2 are constant numbers used to weigh the loss parameters, which vary across different training data and configurations.

4 EXPERIMENTS AND APPLICATIONS

Section 4.1 compares two approaches for the conditioned synthesis input layer proposed in Sec Section 3.2. We then analyse the effectiveness of skip connections and feature fusion proposed in Section 3.2 by a set of ablation studies in Section 4.2. Then, we showcase the results of our approach on several image-to-image translation tasks with different datasets in Section 4.3, as well as demonstrate the model's ability to enlarging canvas in Section 4.4. Finally, we build a graphic interface that implements the deblurring models with transformation filters in Section 4.5.

4.1 Analysis of The Synthesis Input Layer

Methodology. In Section 3.2, we proposed two approaches for implementing the conditioned synthesis input layer. The first method reuses the synthesis input layer in StyleGAN3, a spatially infinite map with learned orientation. In contrast, the second method directly uses the feature maps from the last layer of the encoder. First, we train two models implemented with these two different input methods identically on Flickr-Faces-HQ (FFHQ) (Karras et al., 2018) dataset (512x512 resolution) for inversion, which is to train the model aims to reconstruct a given image. Then, we analyse the input layer in both approaches by conducting an ablation study (Meyes et al., 2019), a method to investigate knowledge representations in artificial neural networks. After the two models are trained for 1680k samples, we first run the inference on unchanged models, then manually set the weights of their input layer to zeros, and then run the inference on the same input again and compare the damages to assess the layer's contribution to the output generations.

Figure 3: Inversion results from the model with two approaches and the ablation tests

Results. Figure 3 shows the results of the two approaches and the ablation tests. The second approach (direct mapping from the encoder's feature maps) is able to produce more accurate details. Besides, the ablation tends to degrade the model trained on the second approach, indicating the second approach leads to a more effective input layer.

Another benefit of using directly mapped input layers is that it allows later layers to work on the generation according to the extracted features independently regardless of the pixel coordinates, taking full advantage of translation equivariant CNNs. One of the disadvantages of using canvas cropped from an infinite space is that the generation is limited to the target region. As shown in Figure 4, arbitrarily expanding the canvas of a StyleGAN3 model leads to convergence failure in areas outside the canvas. However, replacing the cropped input map with feature maps from the encoder will free the generation from a fixed boundary. Figure 5 shows inverting a 1024x1024 image by a network initially trained on images with 512x512 resolution, and compares it with projecting the same image with 2048x2048 resolution to the StyleGAN3 model trained on 1024x1024 images.

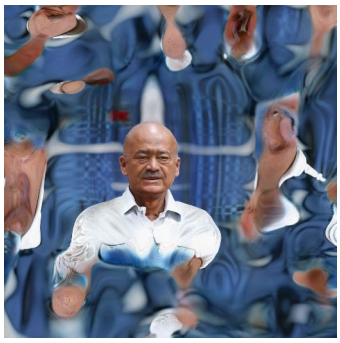


Figure 4: Arbitrarily expanding the canvas of a StyleGAN3 model

Figure 5: Inversion task on a extended canvas

4.2 Analysis of The Skip Connections

Methodology. Section 3.2 proposed reducing the number of feature fusions and limiting skip connections to only the first five layers. To verify the feasibility and effectiveness of this design, we conduct ablation studies on feature fusion. We train six inversion models on 512 x 512 FFHQ dataset. Skip connections are installed between the encoder's last N layers and the synthesis network's first N layers, where N progressively reduces from 5 to 0 (without any additional connections) in these six models. Besides, we analyse the feature maps in the final model with five skip connections to verify the internal representations.

Results. Figure 6 shows the results of the fusion ablation test for the $N = 5$ to $N = 2$ models trained on 1680k samples. The rest of the two models ($N = 1$ and $N = 0$) are unable to converge to sensible results after 800k samples and were therefore aborted. The results show notable improvements when increasing the number of skip connections, especially in preserving details such as background,

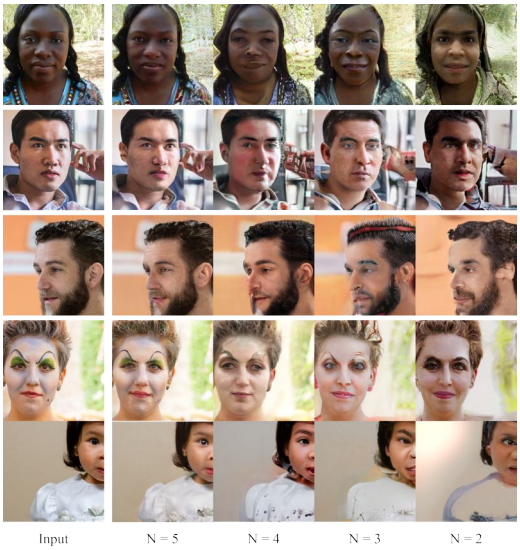


Figure 6: Ablating the skip connections

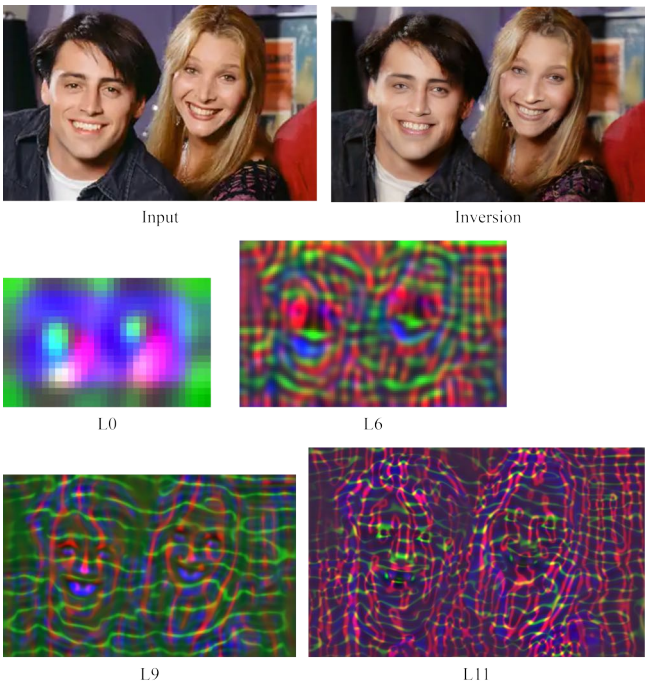


Figure 7: Examples of internal representations in selected layers

hands, unique make-up and even the details in hairs. Therefore, $N = 5$ is decided as the final model's setting.

In addition, Figure 7 shows the inversion result on a larger canvas and its feature maps in selected layers, further verifying that the internal representation complies with the original StyGAN3 model, encoding phases instead of signal magnitudes.

4.3 Conditional Image Synthesis

We test our architecture on four conditional image synthesis tasks: generating face images from blurry images and canny edges, generating realistic landscape images from blurry images. We train the models on Flickr-Faces-HQ (FFHQ) (Karras et al., 2018) dataset and the Landscapes High-Quality (LHQ) (Skorokhodov et al., 2021) dataset, both on 512x512 resolution.

We use StyleGAN3-R, the translational and rotational equivariant configuration of StyleGAN3, as the generator backbone for the FFHQ dataset; and StyleGAN3-T, the translational equivariant configuration of StyleGAN3, as the generator backbone for the LHQ dataset. The training configuration is identical to StyleGAN3.

For the deblurring task, the conditional input was pre-processed by first resizing the image to 256x256, and then applying a Gaussian Filter with a blur sigma of 21. Figure 8 shows examples of results from the model on FFHQ, Figure 9 shows examples of re sults from the model on LHQ. Latent vectors are randomly selected for multi-modal generation.

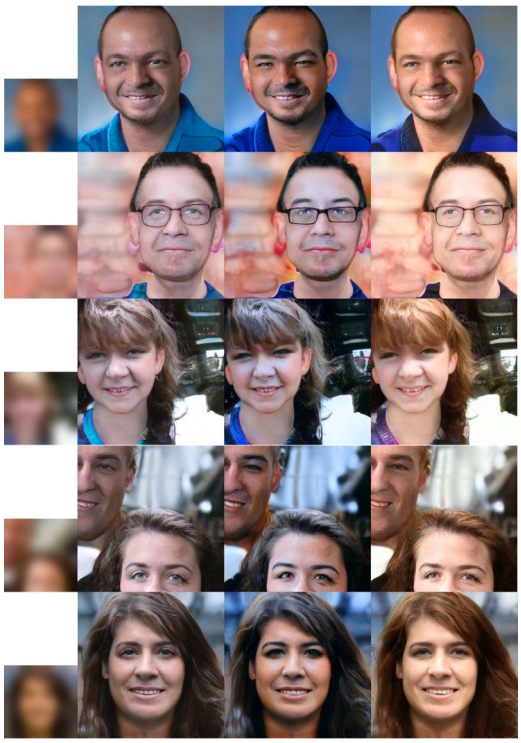


Figure 8: Results of our model for deblurring on FFHQ

For the canny edges model, the conditional input was pre-processed by first resizing the image to 256x256, and then applying a Canny edge detector. Figure 10 shows examples of results with different latent vectors.

4.4 Expandable Canvas

Next, we enlarge the size of the input to test the network's ability on a larger canvas. Both models originally take inputs with 256x256 and output 512x512 generations. Figure 11 shows examples of results when enlarging the input to 256x768 (LHQ) and 256x512 (FFHQ), and the model's outputs expand to 512x1536 (LHQ) and 512x1024

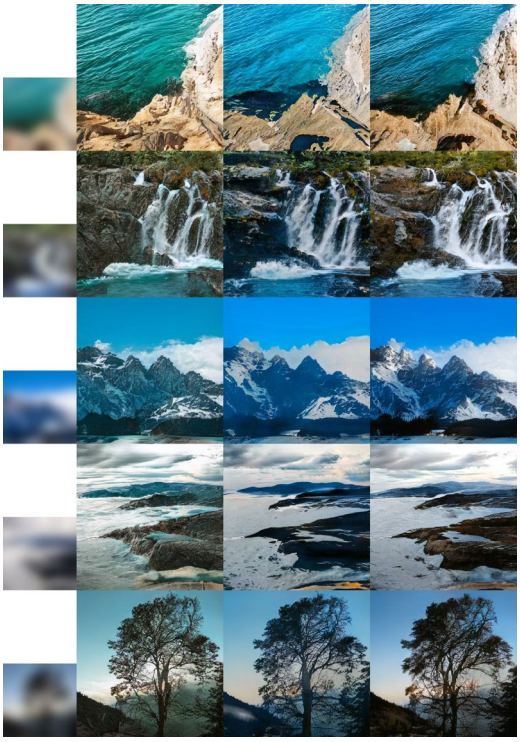


Figure 9: Results of our model for deblurring on LHQ

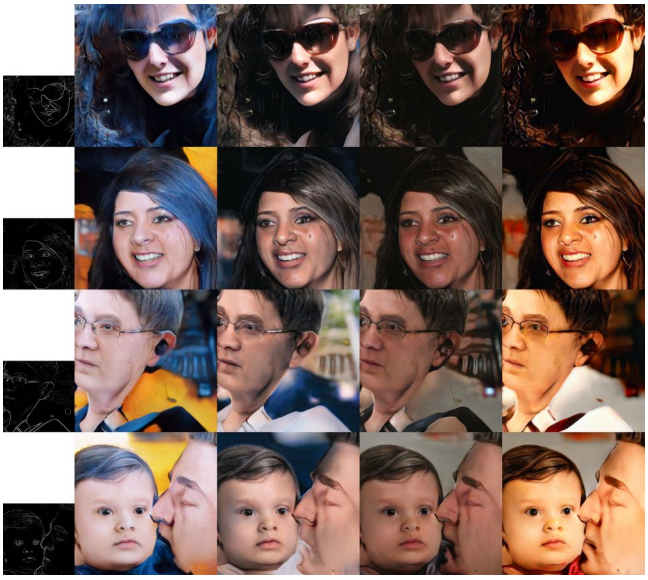


Figure 10: Results of our model for edge to faces on FFHQ

Figure 11: Examples of results with 512x512 model originally trained on 256x256 dataset taking input with 256x768 / 256x512

(FFHQ). Figure 12 shows the result when we further expand the input to 256x1024, and the model's outputs expand to 512x2048.

4.5 Combine the Network Bending Approach

We implement the feature clustering method proposed in network bending, which is to train a softmax feature extraction CNN for each layer, and cluster the feature vector from the bottleneck by the k-means algorithm. However, we increase the length of the narrow bottleneck from 10 to 32 ($\vec{v} \in R^{32}$) to confront larger numbers of channels in StyleGAN3-R. Next, inspired by Learning to See (Akten et al., 2019), we build a graphic interface that implements the deblurring models and takes the webcam as input. A screenshot of the interface is shown in Figure 13. The real-time framework using

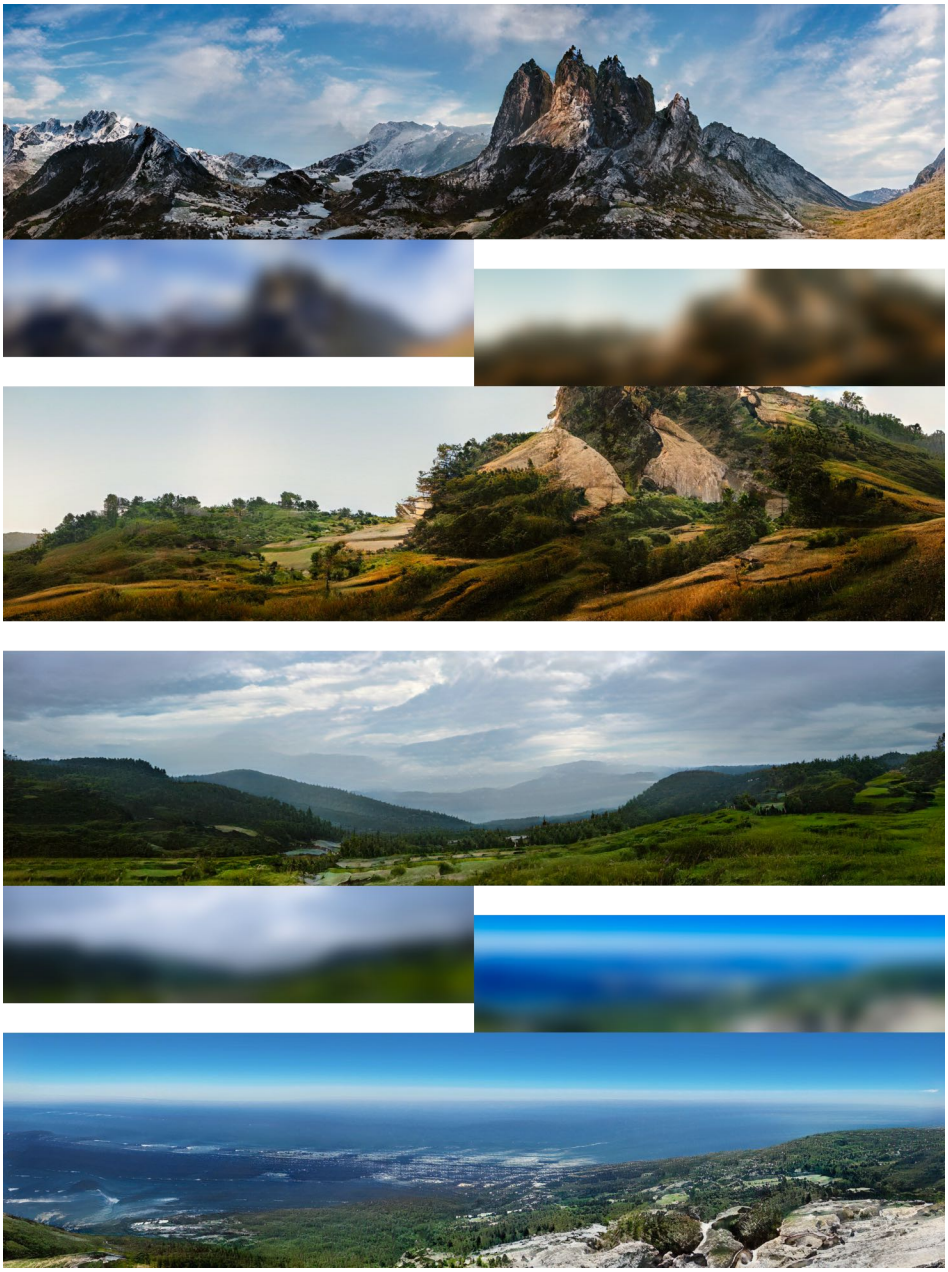


Figure 12: Examples of results with 512x2048 image generated from model originally trained on 512x512 dataset

Flask and SocketIO for bi-directional communications between the web client and the model. Our code for the interface is available at https://github.com/jasper-zheng/realtime-flask-model.

5 HUMAN OPINION STUDIES AND EVALUATION

Previous studies on network bending demonstrated that utilising stochastic layers' transformation as a tool for creating artworks for EP (extended play record) (Broad et al., 2020), produced novel and visually aesthetic results. Our study brings the framework to generic creative contexts using the thematic analysis approach. It investigates how users may potentially use this framework to create novel and expressive outcomes, and aims to identify critical factors influencing the creation process that are not previously conceptualised.

5.1 Methodology

The experiment implements two models trained on different datasets. The first is an inversion model trained on Flickr-Faces-HQ (FFHQ) (Karras et al., 2018), and the other is a deblurring model trained

Figure 13: A screenshot of the real-time interface

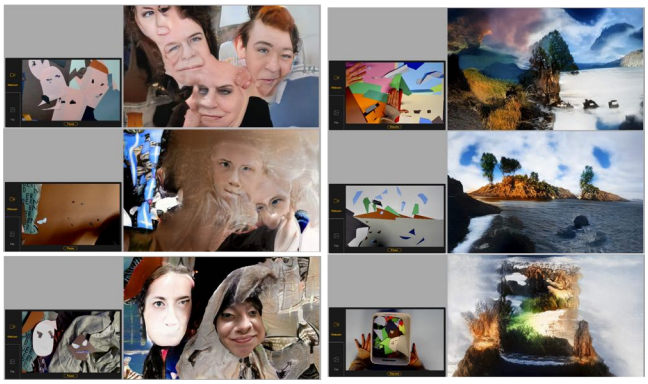


Figure 14: Examples of works created by participants

on Landscapes High-Quality (LHQ) (Skorokhodov et al., 2021). Both models were trained on 512×512 resolution and extended to 640×1024, with the fourth to the seventh layer inserted with clustered feature map transformations. Each model has a dual way of inputs, participants can choose either to use a static image frame, or the real-time webcam as input.

The experiment is divided into two parts. The first part aims to let participants get familiar with the interface and explore the system components, and the second part asks participants to create a work utilising the framework. Each part lasts 15 minutes, and the participants are asked to spend 5 minutes on each model. Observation is conducted while participants interact with the framework, and each part of the experiment is followed by a 5-minute semi-structured interview with open-ended questions.

We incorporate a thematic analysis (Braun and Clarke, 2012) to investigate human perception towards the implemented framework. Utilising a perceptual experiment with human observers and semi-structured interviews, it aims to recognise critical factors influencing exploration and creation.

The experiment follows the qualitative research procedure described by Adams et al. (2008). Six participants are divided into two groups, each with three participants. We conduct the experiment on the first group and run a thematic analysis to emphasise issues raised by the participants based on their frequency and fundamentality, leading to tentative findings. The interview questions are then revised for the second group to probe and grow these findings.

Participants were questioned on their attitudes regarding this form of interaction, the creation process, their generation outcomes, and the differences in their perceptions of different models. During the interviews, we loosely follow the interview template while ensuring these four topics were covered.

5.2 Analysis

Figure 14 shows some examples of works created by the participants. We use the thematic analysis method to aggregate comments collected from the interview.

5.2.1 Multi-level Controls. Several comments aggregated around the interpretable control (Berns and Colton, 2020) over their generative process. Some positive comments indicate that the process of creating shapes, using lights and shadows to outline the overall formation of generations turned out to be a fun aspect, whereas some negative comments suggest that the lack of controls in details (e.g., textures in the landscapes, details of the facial elements) leads to confusion and complaint. Therefore, we suggest defining multi-levels of controls when analysing a deep generation system, where low-level controls coarsely determine the shape, structure and colour, and high-level controls depict finer details of the generation.

The influence of aims. Different definitions of controls also affect by the aims of using deep generative systems. Most participants pointed out that they would use this form of generation as inspiration or a playful experience, or simply in pursuit of an abstract visual effect. However, if the aim is to create a piece of work, they would eventually switch to more stable methods with higher levels of control such as Photoshop or CAD software to refine the generation.

The emerging creation routine. An emerging pattern of behaviour was revealed during the experiment, we observed that most participants tend to follow a "arrange first, adjust later" routine. They primarily focus on arranging the paper card under the webcam the majority of the time and, until achieving a satisfying result, then turn to the interface and play with the transformation filters. When questioned on their attitudes toward the transformation filters, some participants commented that they tend to rely more on the webcam input since it provides more instant and intuitive changes according to the actions. The network bending approach usually acts as a bonus procedure to slightly adjust the visual effects. Figure 15 shows examples of results visualise an example of this routine.

Combining techniques. Some participants tried to tackle the lack of control by utilising other pre-processing or post-processing processes. This also inspires us to combine the model with other image-processing techniques. Figure 16 illustrate an example of an action that attempts to fix the oddity by slightly warping the intermediate image. Figure 17 illustrate a sequence of editing performed on the intermediate image to intentionally create unrealistic and novel outcomes. Combining techniques might be an interesting direction in future works.

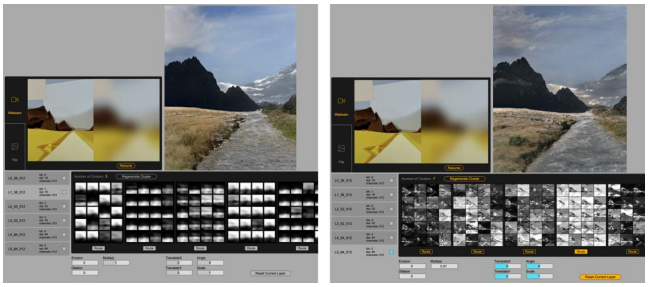


Figure 15: The participant first worked on the paper cards exclusively and achieved the results on the left. Then they experimented with the grouped filters, and eventually figured out a way to boost some detailed textures in the shadow of mountains, and achieved the results on the right.

Figure 16: Slightly warping the intermediate image fixes the peculiar effect in the waves.

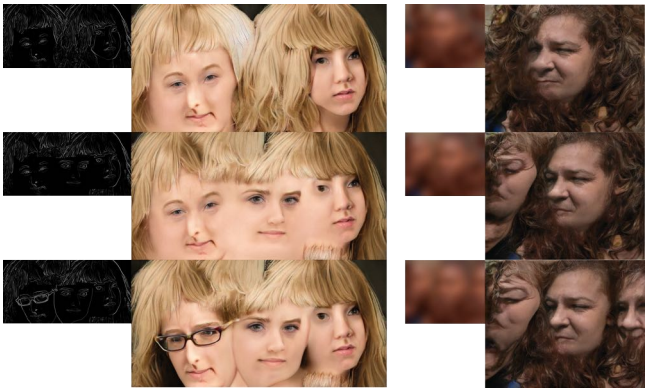


Figure 17: Editing the processed edges to intentionally create unrealistic and novel outcomes.

5.2.2 Influence of the Unpredictables. Another underlying factor is the unpredictable possibilities in the creations, which is also a natural characteristic (Bau et al., 2017) of deep generative systems. These possibilities not only influence participants' perception towards the generated outcomes, but also the creation process.

The acceptance of oddities. In general, most participants chose to use the model trained on the landscape dataset in the end. When questioned on the reason for this decision, they suggested the surprising, unexpected results produced by the landscape model are more likely to be accepted by their visual aesthetic. Although both models are able to create unrealistic imagery, distortion and oddities on human faces may easily lead to uncanny feelings and, more importantly, negative ethical issues such as bias and offences.

Besides, the choice of training data is a critical aspect of deep generative models. Models trained on objects with more complete appearances (e.g. faces, animals, machinery) might create more playful experiences at the sacrifice of utility, since suboptimal results on these models usually only lead to abuses. While models trained on more elementary objects (e.g. terrains, textures, materials) might be more useful for the creative works.

Reusing patterns. Some participants describe the system as a "cooperated partner" instead of a tool, because it always has a certain level of unpredictability. Some participants tend to memorise useful patterns of shapes and colours that may trigger satisfactory results, and then utilise these patterns to create a "blueprint" for the model to work on. And they describe this process as a way to learn the preferences of the system and reuse these patterns to create new pieces.

5.3 Discussion

The equivariant generator is an important component in re-organising and reusing features. The pixel-wise translation allows more intuitive changes according to actions without worrying much about the semantical meanings in the features, while the grouped transformation filter facilitates the adjustment in details. The extendable canvas has broadened the possibility for exploration and creation, allowing more space for the expressive outcome to emerge.

An overall criticism was the lack of interpretable control in the system. While the input image acts as a blueprint for the generation, the user also needs precise control over the details when using the model as a creative tool. This leads us to rethink the design of the intermediate representation. Our framework currently only implements deblurring models for the experiment, however, it might be more useful to use intermediate representations that encode more detailed information (e.g., boundary maps or edges) like the interactive demos in pix2pixHD.

6 CONCLUSION

In this work, we propose a feature extraction encoder with low-level skip connections that extend StyleGAN3 to image-to-image translation. Our approach allows the bounded feature maps to be enlarged even if the model is trained on a fixed resolution, leading to a more flexible canvas for expressive generations. We demonstrate our approach on several image-to-image translation tasks with different datasets. In addition, we implement network bending with a real-time interactive interface to facilitate advanced manipulation of the features. Finally, we conduct a human opinion study

to investigate the potential of our model in a generic creative context. Critical factors influencing the creation are aggregated and analysed to map out future studies' themes.

6.1 Limitation and Future Works

While Section 5.3 highlighted a few extending possibilities, the proposed architecture can also be improved in technical aspects.

Our approach takes an initial step in extending StyleGAN models to image-conditional generation. Although it has demonstrated its potential in solving several image-to-image translation tasks, the detailed architecture still needs further investigation and refinement in future works.

Besides, our model architecture utilises the equivariant generator in StyleGAN3, however, our feature extractor is not yet rotation equivariant. Therefore, the generation may suffer when the rotation is not encoder. Figure 18 show an example of failure where the rotation is not preserved by the encoder. It would be beneficial to make the feature encoder equivariant as well in future work.

Figure 18: Encoder failed to extract the rotation

6.2 Ethical Considerations and Energy Consumption

Potential negative societal impacts of images produced by GAN (Prabhu et al., 2019) were taken into consideration throughout the project. The models trained on the FFHQ dataset are for purely academic purposes and its interactive prototype will not be publicly distributed by any means. Model trainings used approximately a cumulative 300 hours of computation performed on A100 SXM4 80GB (TDP of 400W). Total emissions are estimated to be 25.92kg CO_2, as calculated by MachineLearning Impact calculator (Lacoste et al., 2019). Paper cards in the experiments were limited allocated to participants, reused during and after the experiments.

REFERENCES

Adams, A., Lunt, P. and Cairns, P. (2008), A qualititative approach to hci research, *in* P. Cairns and A. Cox, eds, 'Research Methods for Human-Computer Interaction', Cambridge University Press, Cambridge, UK, pp. 138–157.
URL: *http://oro.open.ac.uk/11911/*

Akten, M., Fiebrink, R. and Grierson, M. (2019), Learning to see: You are what you see, *in* 'ACM SIGGRAPH 2019 Art Gallery', SIGGRAPH '19, Association for Computing Machinery, New York, NY, USA.
URL: *https://doi.org/10.1145/3306211.3320143*

Alaluf, Y., Patashnik, O., Wu, Z., Zamir, A., Shechtman, E., Lischinski, D. and Cohen-Or, D. (2022), 'Third time's the charm? image and video editing with stylegan3'.
URL: *https://arxiv.org/abs/2201.13433*

Bau, D., Liu, S., Wang, T., Zhu, J.-Y. and Torralba, A. (2020), Rewriting a deep generative model, *in* 'Proceedings of the European Conference on Computer Vision (ECCV)'.

Bau, D., Strobelt, H., Peebles, W., Wulff, J., Zhou, B., Zhu, J.-Y. and Torralba, A. (2019), 'Semantic photo manipulation with a generative image prior', *ACM Trans. Graph.* **38**(4).
URL: *https://doi.org/10.1145/3306346.3323023*

Bau, D., Zhou, B., Khosla, A., Oliva, A. and Torralba, A. (2017), 'Network dissection: Quantifying interpretability of deep visual representations'.
URL: *https://arxiv.org/abs/1704.05796*

Bau, D., Zhu, J., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T. and Torralba, A. (2018), 'GAN dissection: Visualizing and understanding generative adversarial networks', *CoRR* **abs/1811.10597**.
URL: *http://arxiv.org/abs/1811.10597*

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B. and Torralba, A. (2020), 'Understanding the role of individual units in a deep neural network', *Proceedings of the National Academy of Sciences* .
URL: *https://www.pnas.org/content/early/2020/08/31/1907375117*

Berns, S., Broad, T., Guckelsberger, C. and Colton, S. (2021), 'Automating generative deep learning for artistic purposes: Challenges and opportunities'.

Berns, S. and Colton, S. (2020), Bridging generative deep learning and computational creativity., *in* 'ICCC', pp. 406–409.

Braun, V. and Clarke, V. (2012), 'Thematic analysis.', *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* p. 57–71.

Broad, T., Berns, S., Colton, S. and Grierson, M. (2021), 'Active divergence with generative deep learning – a survey and taxonomy'.

Broad, T., Leymarie, F. F. and Grierson, M. (2020), 'Network bending: Expressive manipulation of deep generative models'.
URL: *https://arxiv.org/abs/2005.12420*

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014), 'Generative adversarial networks'.

He, K., Zhang, X., Ren, S. and Sun, J. (2015), 'Deep residual learning for image recognition'.
URL: *https://arxiv.org/abs/1512.03385*

Hertzmann, A. (2020), Visual indeterminacy in gan art, *in* 'ACM SIGGRAPH 2020 Art Gallery', SIGGRAPH '20, Association for Computing Machinery, New York, NY, USA, p. 424–428.
URL: *https://doi.org/10.1145/3386567.3388574*

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W. and Wu, J. (2020), 'Unet 3+: A full-scale connected unet for medical image segmentation'.
URL: *https://arxiv.org/abs/2004.08790*

Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A. (2016), 'Image-to-image translation with conditional adversarial networks'.
URL: *https://arxiv.org/abs/1611.07004*

Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A. (2017), 'Image-to-image translation with conditional adversarial networks', *CVPR* .

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J. and Aila, T. (2021), Alias-free generative adversarial networks, *in* M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang and J. W. Vaughan, eds, 'Advances in Neural Information Processing Systems', Vol. 34, Curran Associates, Inc., pp. 852–863.
URL: *https://proceedings.neurips.cc/paper/2021/file/076ccd93ad68be51f23707988e934906-Paper.pdf*

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J. and Aila, T. (2021), 'Alias-free generative adversarial networks'.

Karras, T., Laine, S. and Aila, T. (2018), 'A style-based generator architecture for generative adversarial networks'.
URL: *https://arxiv.org/abs/1812.04948*

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. and Aila, T. (2019), 'Analyzing and improving the image quality of stylegan'.
URL: *https://arxiv.org/abs/1912.04958*

Lacoste, A., Luccioni, A., Schmidt, V. and Dandres, T. (2019), 'Quantifying the carbon emissions of machine learning'.
URL: *https://arxiv.org/abs/1910.09700*

Li, Y., Chen, X., Wu, F. and Zha, Z.-J. (2019), 'Linestofacephoto: Face photo generation from lines with conditional self-attention generative adversarial network'.
URL: *https://arxiv.org/abs/1910.08914*

Liu, X., Yin, G., Shao, J., Wang, X. and Li, H. (2019), 'Learning to predict layout-to-image conditional convolutions for semantic image synthesis'.
URL: *https://arxiv.org/abs/1910.06809*

Lu, H., She, Y., Tie, J. and Xu, S. (2022), 'Half-unet: A simplified u-net architecture for medical image segmentation', *Frontiers in Neuroinformatics* **16**.
URL: *https://www.frontiersin.org/articles/10.3389/fninf.2022.911679*

Mao, X., Cao, L., Gnanha, A. T., Yang, Z., Li, Q. and Ji, R. (2022), 'Cycle encoding of a stylegan encoder for improved reconstruction and editability'.
URL: *https://arxiv.org/abs/2207.09367*

McCallum, L. and Yee-King, M. (2020), Network bending neural vocoders, *in* '4th Workshop on Machine Learning for Creativity and Design at NeurIPS 2020, Vancouver, Canada.'.
URL: *https://research.gold.ac.uk/id/eprint/29652/*

Meyes, R., Lu, M., de Puiseau, C. W. and Meisen, T. (2019), 'Ablation studies in artificial neural networks'.
URL: *https://arxiv.org/abs/1901.08644*

Pinkney, J. (2020), 'Matlab stylegan playground'.
URL: *https://www.justinpinkney.com/matlab-stylegan/*

Pouliot, A. (2020), 'Gan bending'.
URL: *https://darknoon.com/2020/03/03/gan-hacking/*

Prabhu, V. U., Yap, D. A., Wang, A. and Whaley, J. (2019), 'Covering up bias in celeba-like datasets with markov blankets: A post-hoc cure for attribute prior avoidance'.
URL: *https://arxiv.org/abs/1907.12917*

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S. and Cohen-Or, D. (2020), 'Encoding in style: a stylegan encoder for image-to-image translation'.
URL: *https://arxiv.org/abs/2008.00951*

Richardson, E. and Weiss, Y. (2020), 'The surprising effectiveness of linear unsupervised image-to-image translation'.
URL: *https://arxiv.org/abs/2007.12568*

Ronneberger, O., Fischer, P. and Brox, T. (2015), 'U-net: Convolutional networks for biomedical image segmentation'.
URL: *https://arxiv.org/abs/1505.04597*

Sauer, A., Schwarz, K. and Geiger, A. (2022), 'Stylegan-xl: Scaling stylegan to large diverse datasets'.

Schultz, D. (2019), '99 ways to use a dataset'.
URL: *https://artificial-images.com/project/ladiescrop-machine-learning-art-experiments/*

Schultz, D. (2020), 'Artificial images'.
URL: *https://artificial-images.com/project/you-are-here-machine-learning-film/*

Simonyan, K. and Zisserman, A. (2014), 'Very deep convolutional networks for large-scale image recognition'.
URL: *https://arxiv.org/abs/1409.1556*

Skorokhodov, I., Sotnikov, G. and Elhoseiny, M. (2021), 'Aligning latent and image spaces to connect the unconnectable', *arXiv preprint arXiv:2104.06954* .

Som, M. (2020), 'Mal som @errthangisalive'.
URL: *http://www.aiartonline.com/highlights-2020/mal-som-errthangisalive/*

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J. and Catanzaro, B. (2017), 'High-resolution image synthesis and semantic manipulation with conditional gans'.
URL: *https://arxiv.org/abs/1711.11585*

Zhu, P., Abdal, R., Qin, Y. and Wonka, P. (2020), Sean: Image synthesis with semantic region-adaptive normalization, *in* 'IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.