StyleGAN-Canvas: Augmenting StyleGAN3 for Real-Time Human-AI Co-Creation

Shuoyang Zheng[1,*]

[1]Creative Computing Institute, University of the Arts London, 45-65 Peckham Rd, London, UK

Abstract

Motivated by the mixed initiative generative AI interfaces (MIGAI), we propose bridging the gap between StyleGAN3 and human-AI co-creative patterns by augmenting the latent variable model with the ability of image-conditional generation. We modify the existing generator architecture in StyleGAN3, enabling it to use high-level visual ideas to guide the human-AI co-creation. The resulting model, StyleGAN-Canvas, can solve various image-to-image translation tasks while maintaining the internal behaviour of StyleGAN3. We deploy our models to a real-time graphic interface and conduct qualitative human opinion studies. We use the MIGAI framework to frame our findings and present a preliminary evaluation of our models' usability in a generic co-creative context.

Keywords

generative adversarial networks, human-AI co-creation, creativity support tools,

Figure 1: A prototype interface encapsulating a StyleGAN-Canvas model translating paper card layout into faces. The user adjusts layouts while the model provides synchronous generation based on visual similarity. A screen recording is available at: https://youtu.be/9AsfsT8uXGY

1. Introduction

Generative adversarial networks (GANs) [1] have recently been rapidly developed and have become a powerful tool for creating high-quality digital artefacts. In the case of images, modern approaches to improving model quality have successfully brought the generated outcomes from coarse low-resolution interpretations to realistic portraits with high diversity [2] and visual fidelity [3]. Notably, introducing continuous convolution in StyleGAN3 (alias-free GAN) [4] has enabled the generative network to perform equally well regardless of pixel coordinates, paving the way for more flexible human-AI interaction. Closely following the advances in deep generative neural networks, StyleGAN models have been found to be widely used as creativity support tools, creating unconventional visual aesthetics [5, 6, 7] and novel human-AI co-creative experiences [8, 9, 10, 11]. This motivates research on human-AI co-creative applications, offering insight into interaction possibilities between human creators and AI enabled by GANs [12].

Muller et al. [13] adapt notations introduced by Spoto and Oleynik [14], presenting mixed initiative generative AI interfaces (MIGAI), an analytical framework with 11 vocabularies of actions to describe a human-AI interaction process. These actions are analysed into sequences to form generic human-AI co-creative patterns. However, Grabe et al. [15] identify a gap between the MIGAI framework and latent variable models such as GANs. This is partially due to latent variable models' deficiency of ability in interpreting visual design concepts such as sketches and semantic labels [16], leading to the left-out of the action *ideate* [15], which describes using high-level concepts to guide or shape the generation [5]. Therefore, Grabe et al. [15] suggested tailoring the MIGAI framework to fit co-creative GAN applications.

Motivated by the gap between GANs and human-AI co-creative patterns, we suggest an alternative approach to bridge the latent variable model, StyleGAN3, to the co-creative framework by modifying the model's technical functioning. Specifically, by augmenting StyleGAN3 with image-conditional generation ability, we enable it to transform visual ideas into generation. This enables a tightly-coupled human-AI interaction process that emphasises using high-level visual concepts to guide the artefact and fulfil the action *ideate* [13], aligning with the co-creative patterns mentioned in the MIGAI framework. We limit our study to StyleGAN3 because its introduction of continuous convolution facilitates more flexible

Joint Proceedings of the ACM IUI Workshops 2023, March 2023, Sydney, Australia

*Corresponding author.

j.zheng0320211@arts.ac.uk (S. Zheng)

https://alaskawinter.cc/ (S. Zheng)

0000-0002-5483-6028 (S. Zheng)

human inputs, which is a crucial feature required by our approach.

Therefore, the primary aim of this research is to augment StyleGAN3 with image-conditional generation ability for a co-creative context. To achieve this, we adapt the existing model architecture in StyleGAN3, which takes a latent vector and a class label as the model's input [4], and propose an encoder network to extract features from the conditional image. We also adapt the architecture previously applied to various image-to-image translation models [17, 18, 19] to connect the proposed encoder and StyleGAN3's generator. The modified model, StyleGAN-Canvas, takes a latent vector and an accompanying image as inputs to guide the generation. We show results from our models trained for various image-to-image translation tasks while maintaining the internal behaviours of StyleGAN3, providing more flexible and intuitive control to ideate the co-creation.

To evaluate our model in a generic co-creative context, we build a graphic interface to facilitate the real-time interaction between users and the model. We conduct qualitative human opinion studies, identify potential co-creative patterns using the MIGAI analytical framework, and present an exploratory human subject study on our model. By aligning StyleGAN-Canvas with the actions set in MIGAI, we hope to bring its capability into the discussion of co-creative design processes, and provide a preliminary insight into the unexplored interaction possibilities enabled by StyleGAN.

The rest of the paper is structured as follows. We summarise related works on StyleGAN, image-conditional generation, and the co-creative pattern in Section 2. Then, we present our modification to the model's architecture in Section 3. We conduct experiments on our models and showcase the results and applications in Section 4. Next, we evaluate the model in a co-creative context in Section 5. Section 6 highlights limitations and future studies.

2. Related Works

2.1. Alias-Free GAN

Our model architecture is extended from StyleGAN3 (Alias-Free GAN). This section summarises the background of StyleGAN and reviews the continuous convolution approach introduced by StyleGAN3 that enables the translation and rotation equivariant feature.

StyleGAN [20] is a style-based generator with a regularised latent space offering high-level style control over image generation. The StyleGAN generator comprises a mapping network M that transforms the initial latent code z to intermediate latent code $w \sim W$, and a synthesis network G with a sequence of N synthesis blocks, each comprising convolutions controlled by the intermediate latent code w, non-linearities, and upsampling, and eventually produces the output image $z_N = G(z_0; w)$. Its high-level style control is achieved by adaptive instance normalisation (AdaIN) [21], an approach to amplifying specific channels of feature maps on a per-sample basis. In practice, a learned affine transform is applied to the intermediate latent space W to obtain style codes y, which are then used as scalar components to modulate the corresponding feature maps before each convolution layer. This architecture was then revised in StyleGAN2 [22] by replacing instance normalisation with feature map demodulation and inherited by StyleGAN3.

In a later work, the continuous convolution approach [23] implemented in StyleGAN3 (alias-free GAN) [4] has dramatically changed the internal representations in the synthesis network G. Signals in G are operated in the continuous domain rather than the discrete domain to make the network equivariant, which means any operation f in the network is equivariant to a spatial transformation t ($t \circ f = f \circ t$). This eliminates positional references; therefore, the model can be trained on unaligned data, and the "texture sticking" behaviour in standard GAN models is removed.

Moreover, the input of the synthesis network of StyleGAN3 uses a spatial map $z0$ defined by Fourier features [24] to precisely model the translation and rotation. The spatial map $z0$ is sampled from uniformly distributed frequencies, fixed after initialisation, and spatially infinite [4]. Its translation and rotation parameters are calculated by a learned affine layer based on intermediate latent space W. The spatial map $z0$ acts as a coordinate map that allows later layers to grab on and therefore defines the global transformations in the synthesis network [25].

2.2. Image-Conditional Generation

Methods for image-conditional generation aim to generate a corresponding image given an image from the source domain, depicting objects or scenes in different styles or conditions. Current solutions to this are usually categorised under two approaches. The first approach uses a linear encoder-decoder architecture [26], in which the input image is encoded to a vector to match the target domain, and then decoded into an output image. This method was later extended to a generic framework for various tasks, such as sketches and layout to image, facial frontalisation, inpainting, and super-resolution [27].

The second direction uses conditional GANs [17] with U-net architectures [28]. It uses a similar encoder-decoder setting, but instead of encoding the input image into a vector, it uses the residual feature maps from the encoder as spatial information and propagates them to the decoder through skip connections [29]. The propagated features are concatenated with the outputs from

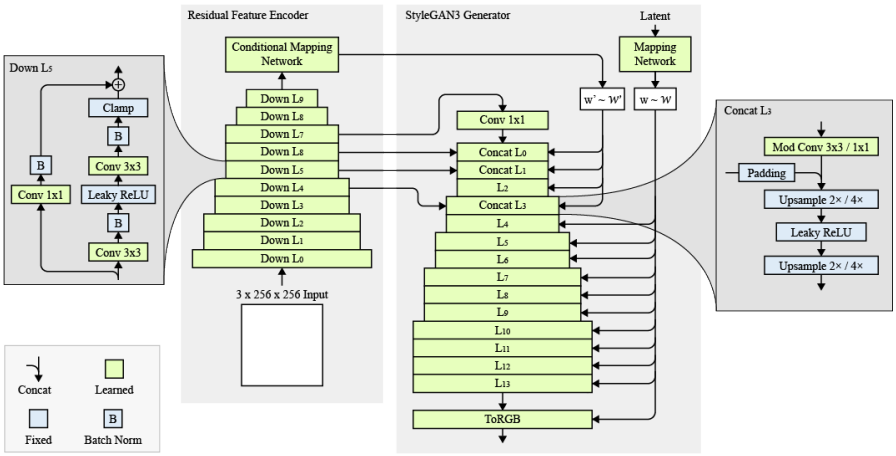

Figure 2: We build a residual feature encoder and adapt the StyleGAN3 generator. The main datapath consists of (i) 10 downsampling residual blocks (Section 3.1.1), each consisting of a mapped shortcut with a 1×1 convolutional layer and bath normalisation, and a downsampling block with two convolutional layers, an activation layer (Leaky ReLU), bath normalisations and a clamping layer, (ii) a conditional mapping network (Section 3.1.2), (iii) StyleGAN3 mapping network, (iv) adapted StyleGAN3 synthesis blocks (Section 3.2).

corresponding decoder layers. This method aims to provide the generator with a mechanism to circumvent the bottleneck layer and allow spatial information to be shuttled from the encoder to the decoder [28]. Therefore, it introduces a strong locality bias [26] into the generation, which means each pixel in the output has a positional reference to the input, and the general image structure is preserved during translation. This method has been extended for various tasks such as line-conditioned generation [19], layout-to-image synthesis [30], and semantic region-adaptive synthesis [31].

2.3. Human-AI Co-Creative Pattern

Mixed Initiative Generative AI Interfaces (MIGAI) [13] describe modes of interaction in which both human and generative AI is engaged in a creative process. It aims to frame subprocesses in a creative flow by an action set with 11 vocabularies. A generic co-creative pattern is identified using the MIGAI framework, in which the AI system first learns a target domain, then the human ideates a design concept to guide the artefact; subsequently, the human and AI system take turns to evaluate and adjust, eventually produce the outcome [15]. Our study focuses on ideating, a process of conceptualising design solutions to high-level abstractions. Later exploratory study emphasising the potential of enhancing the creative process through human-AI collaborative ideation [16]. Methods for this have been advanced in diverse fields of application: co-creative game content [32], collaborative text editing [33], and text-guided image generation [34].

3. Extending StyleGAN to Image-conditional Generation

The objective of our approach is to allow StyleGAN3 to use images as conditions to guide the generation. This section will discuss our modification to the current StyleGAN3 architecture to address this objective.

As mentioned in Section 2.1, current StyleGAN3 models learn a mapping from a random noise vector z to an output image $Z_N = G(z)$. The modification aims to extend the input from a vector z to a conditional image x combined with z, and generate Z_N that is close to the corresponding ground truth y. To do this, we first need a feature extraction encoder E that extracts features from x, then adapts the generator G to produce outputs based on the extracted features and the input vector z, i.e. $Z_N = G(E(x), z)$. Besides, the training objective should push the generation closer to the corresponding image y.

3.1. Feature Extraction Encoder

3.1.1. Adapted Residual Network

The feature extraction encoder E employs an adapted ResNet [29] architecture as the encoder backbone, which has been previously used for feature maps extraction in image-to-image translation works [27]. As shown in Figure 2 (left), the encoder network downsamples feature maps to $(x/2^6, y/2^6)$, where x and y denote the width and height of the input image. In addition, as StyleGAN uses mixed-precision to speed up training and inferences

[24], we utilise similar techniques and reduce the precision to FP16 for the first five residual blocks in the encoder. Consequently, it requires pre-normalisation and an extra clamping layer that clamps the output of the convolutional layers to $\pm 2^9$ [35].

3.1.2. Conditional Mapping Network

Previous works in StyleGAN encoder [27] have highlighted that replacing select layers of the extended latent space $W+$ by computed latent codes 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 and editability [36]. 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

As mentioned in Section 2.1, the StyleGAN3 generator consists of a mapping network and a synthesis network, our modification aims to connect its synthesis network to extracted information from the feature extraction encoder.

We start by modifying the input layer in its synthesis network. The original network utilises a fixed-size spatial map $z0$ defined by Fourier features [24] as its input to model translation and rotation parameters. However, the entire input layer is left out, and we use feature maps from the last layer of the encoder directly as the synthesis network's input. This lets the translation and rotation parameters be inherited from the spatial feature maps.

Next, connections between the feature encoder aim to provide precise structural information about input images. A U-net [28] architecture is well-suited for propagating high-level details from the encoder to the decoder [26]. However, as mentioned in Section 2.1, experiments on StyleGAN3 have demonstrated that high-level feature maps in the synthesis network encode information in continuous domains instead of discrete domains [4], relying on skip connections to propagates discrete features from the encoder may deviate from StyleGAN3's internal generation behaviour. To tackle this, we first move the concatenation node from the end of each synthesis block to the point before the filtered non-linearities layer, shown in Figure 2 (right). We also remove the padding layers in each synthesis block to ensure the dimension matches the skip connections. Additionally, research on U-Net and its variants has demonstrated that a simplified structure with fewer feature fusions can achieve reasonable results [37, 38]. Therefore, we reduce the number of feature fusions and limit connections to only the first N 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 $i \in (0, 5]$. The experiment in Section **??** will show that $N = 5$ is the best configuration that leads to more stable training and efficient generation. 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 regularisation 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 [4]. 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:

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

The second phase starts after the training reaches 300k images. We add a perceptual loss L_{VGG} utilising a pre-trained VGG19 [39] network, which has been used in the training of previous conditional GANs and has led to more finer details in the resulting images [18], defined as follow:

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

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

$$L_{phase2}(D,G,E) = \lambda_1 L_2(G,E) + \lambda_2 L_{VGG}(G,E) + L_{GAN}(D,G,E) + L_{reg}(E,M) \quad (6)$$

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

In the following Section 4.1, we analyse the effectiveness of the U-net proposed in Section 3.2 by a set of ablation studies. Next, in Section 4.2, we demonstrate the training process of our model for several image-to-image translation tasks with different datasets and showcase their results. We also experiment with scaling the model to larger canvases in Section 4.3. Finally, we build a graphic interface that implements our models with a set of transformation filters in Section 4.4.

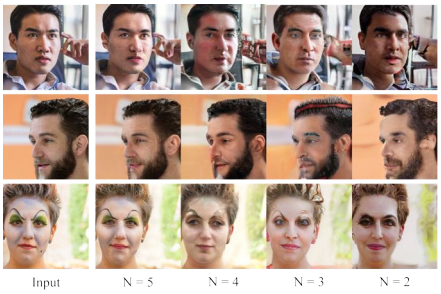

Figure 3: Ablating the skip connections

4.1. Analysis of The Skip Connections

Methodology. Section 3.2 proposed reducing the number of skip connections to only the first N layers. To verify the feasibility of this design and find the most effective configuration of N, we first trained six models on Flickr-Faces-HQ (FFHQ) [20] dataset with 512×512 resolution for an ablation test [40] on the skip connections, which is a method to investigate knowledge representations in artificial neural networks by disabling specific nodes in a network. We use inversion tasks [41] as the training goal, in which the models are trained to reconstruct a given image without translation, aiming to test the efficiency of the encoder. 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 in these six models. The resulting outputs are compared across six models to decide the final configuration.

Results. Figure 3 (left) shows the results of the 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, hands, unique make-up and even the details in hairs. Therefore, $N = 5$ is decided as the final setting.

4.2. Conditional Image Synthesis and Editing

Methodology. Conditional image synthesis uses image-conditional models to generate image Z corresponding to the ground truth image y, given input condition image x. We tested our architecture on three conditional image synthesis tasks: (i) generating face images from blurry images, (ii) generating face images from canny edges, and (iii) generating realistic landscape images from blurry images.

First, the ground truth y was pre-processed into processed condition x. In deblurring models, the pre-process pipeline is a resizing layer that scales the resolution of y to 256×256, and a Gaussian filter with sigma $\sigma = 28$ that process resized y to blurred images as condition x. In the edge-to-face model, the resolution of y is first resized to 256×256, and then applying a Canny edge detector [42] to process resized y to edges as condition x. y and x are provided to the training as paired data. After the models were trained, the same pipelines were applied to pre-process input $x\prime$ for inference. The training process is illustrated in Figure 4 (top).

The dataset used for face generation models was Flickr-Faces-HQ (FFHQ) [20], and the dataset used for landscape photos generation was Landscapes High-Quality (LHQ) [43], both with 512×512 resolution. We used 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 was identical to StyleGAN3.

Results. The deblurring model on the FFHQ dataset was trained with 3700k samples, and the deblurring model on the LHQ dataset was trained with 6700k samples. We compared ground truth samples y, conditions x, and generation outcomes Z_N in Figure 5 and Figure 6. The edge-to-face model on the FFHQ dataset was trained with 3700k samples. In Figure 7, we compared ground truth samples y, conditions x, and generation outcomes Z_N with three randomly selected latent vectors for multi-modal generations.

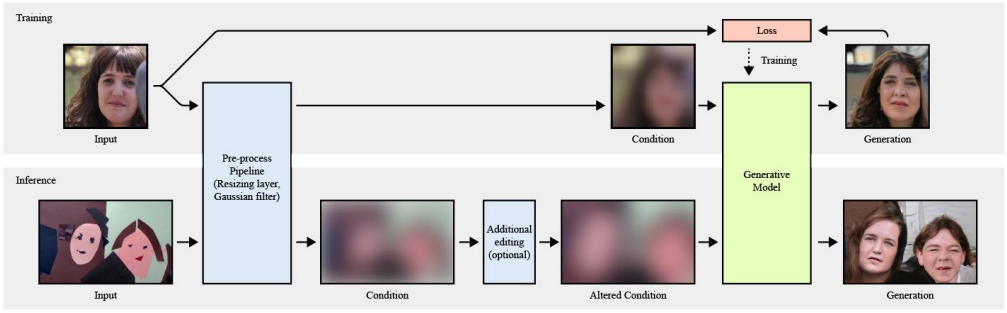

Figure 4: During training, target images are processed into condition, the model generates fake images, and the fake images and the target images are used to calculate the loss.

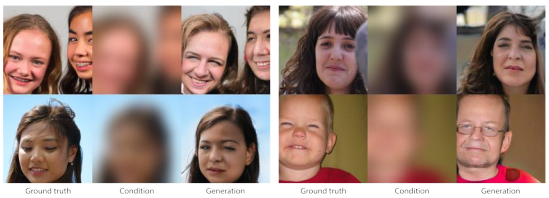

Figure 5: Results of our model for deblurring on FFHQ 512×512 (left: ground truth samples, middle: processed conditions, right: generations

Figure 6: Results of our model for deblurring on LHQ 512×512 (left: ground truth samples, middle: processed conditions, right: generations

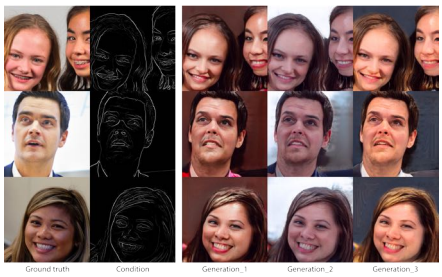

Figure 7: Results of our model for edge-to-face model on FFHQ 512×512 (ground truth samples, processed conditions, and three generations with different latent vectors

Figure 8: Results of our model for edge-to-faces on FFHQ, and local editing

In addition, the canny edges model provides an alternative approach to local editing. Modifying edges in the condition image allows the model to alter semantical elements in the generation. Illustrated in Figure 4 (bottom), the modified conditions can be obtained by combining and adapting existing conditions from other images. For example, in Figure 8, we superimposed edges processed from other images to the original edges to add hair fringe, glasses and smile; we painted on the original edges to modify eyes and add sunglasses.

4.3. Large Canvas

As mentioned in Section 3.2, the padding layers in the synthesis network are removed, and the entire generator does not have unintentional positional references with absolute positional references [44], we hypothesised that our modified architecture induced an extendable generation canvas that can be enlarged after trained on a fixed resolution, without additional training required. Therefore, we enlarge the input resolution to test the model's ability on a larger canvas.

The model for landscape photo generation was trained on the dataset with 512×512 resolution, taking inputs with 256×256 resolution. To enlarge the generation canvas, we doubled and tripled the width of inputs, expanding their resolution to 1024×256 and 768×256. Then, the expanded inputs were taken directly into the generator and convolved by each convolutional layer. Therefore, the expected output resolutions are 2048×512 and

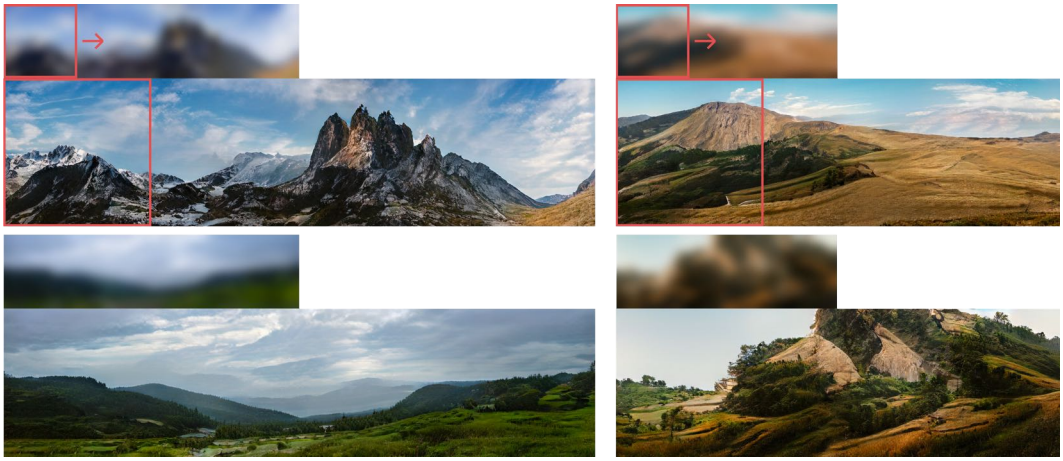

Figure 9: Examples of results with 512×2048 / 512×1536 image generated from model originally trained on 512×512 dataset. More results can be accessed from https://github.com/jasper-zheng/StyleGAN-Canvas

1536×512. Additional training is not required during the experiment.

Figure 9 shows the resulting generations. While the outputs are expanded up to four times larger than the original canvas, the generation quality remains unchanged. This ensures that the input canvas can be flexibly expanded when the models are implemented into user interfaces.

4.4. User Interface

Implementation. The deblurring models were deployed to a graphic user interface. The models were running on a cloud server. We used Flask and SocketIO for bi-directional communications between the web client and the server. The generation runs in real-time at roughly ten frames per second on an NVIDIA RTX A4000 GPU. The code for our implementation is available at https://github.com/jasper-zheng/realtime-flask-model.

Combining network bending. In addition to the baseline model, the generation system is implemented with network bending [45], an approach to alter a trained model's computational graph by inserting transformation filters between different convolutional layers, allowing the model to generate novel samples that are diverse from the training data [46]. We used the clustering algorithm presented by Broad et al. [45] to group spatially similar feature maps in selected layers and allow the transformation filters to be inserted into specific groups. We trained softmax feature extraction CNNs for each layer and clustered each flatten layer by the k-means algorithm. However, different from the implementation in Broad et al.'s work, we increase the length of the flatten vector from 10 to 32 ($\vec{v} \in R^{32}$) to confront larger numbers of channels in StyleGAN3-R.

Interface design. Figure 10 shows a screenshot of the deployed system. The interface allows inputs from a webcam or locally selected image files (top left). The interface allows users to pause, resume generation, and switch the input between the webcam and local files. Users can insert transformation filters into certain groups in certain layers. The list on the bottom left side presents layers available for network bending operations. Once a specific layer is selected, the system shows current clusters of feature maps in the layer. Users can regenerate clusters according to feature maps from the current frame. Then, users activate the 'route' button to apply transformation filters to the cluster. The system provides basic transformation filters, including erosion and dilation, multiply, translations, rotation, and scale.

Figure 10: A screenshot of our interface.

5. Human Subjects Study and Evaluation

This section presents a human subjects study to evaluate our models in a human-AI co-creation context. The study uses a thematic analysis approach to identify potential co-creative patterns defined by the MIGAI framework [13] that underlies the interaction.

5.1. Methodology

In the user study, we asked six participants to use the generation interface described in Section 4.4. For the webcam inputs, coloured paper cards and scissors were available to the participants. Participants were asked to arrange and layout paper cards in front of a webcam pointing to a paper canvas, aiming to control the generation until it reached something they liked. Besides, the transformation filters were also available to fine-tune the creation further. Models used in Section 4.2 are available to the participants, including the deblurring model and the edge-to-face model trained on Flickr-Faces-HQ (FFHQ) [20], the deblurring model trained on Landscapes High-Quality (LHQ) [43].

The experiment followed the qualitative research procedure described by Adams et al. Adams et al. [47]. Six participants were divided into two groups of three. We conducted the study on the first group and ran an analysis to emphasise issues raised by the participants based on their frequency and fundamentality, leading to tentative findings. Then, the interview questions are revised for the second group to probe and grow these findings. The study was divided into two parts, each lasting 15 minutes. The first part asked participants to familiarise the interface and explore the system components. The second part asked the participant to create a work they liked. Observation is conducted while participants interact with the framework, and a 5-minute semi-structured interview with open-ended questions follows each part of the experiment.

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. We loosely followed the interview template during the interviews while ensuring these four topics were covered.

Figure 11 shows some examples of works created by the participants.

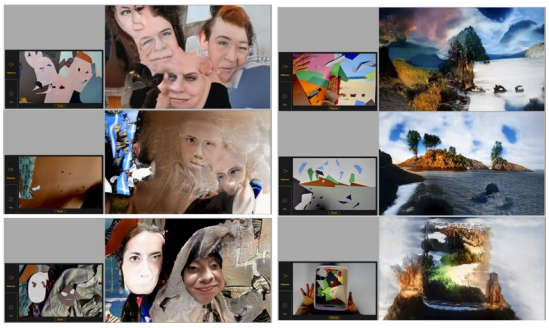

Figure 11: Examples of works created by participants

5.2. Analysis

In this section, we use the thematic analysis [48] method to aggregate comments collected from the interview, aiming to identify critical factors influencing the interaction. We used the MIGAI analytic framework [13] to frame the human-AI co-creative patterns.

5.2.1. Learn

The action *learn* in the MIGAI framework describes the AI model using training data to construct its knowledge, usually involving the choice of datasets, model architectures, and training configurations. Most participants used 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 can 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. Therefore, the choice of training data is critical for co-creation.

5.2.2. Ideate

The action *ideate* describes the process of using high-level concepts to guide the generation. We collected comments aggregated around this process. Some positive comments indicate that creating shapes using paper cards is an intuitive generation process. 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. Most participants pointed out that they would use this form of generation as inspiration, giving them a high-level sketch developed from their ideated concept. Alternatively, they may treat it as a playful experience or simply in pursuit of an abstract visual effect. However, if the generation aims to create a serious piece of work, they would eventually switch to more stable methods with lower-level control to refine the work, such as Photoshop or CAD software. This is because, in this case, they want to be manually in charge of finer details such as lighting and textures.

5.2.3. Adapt

The action *adapt* describes adjusting existing artefacts. During the study, we observed that participants primarily focused on exploring the outcome from the computational agent by arbitrarily arranging the paper cards. When they reach a layout that triggers interesting results, they iteratively adjust the arrangement until achieving a satisfying result. Some participants tried to tackle the lack of control by utilising other pre-processing or post-processing processes. Figure 12 illustrates an example of an action that attempts to fix the oddity by slightly warping the intermediate image in other image editing software. Figure 13 illustrates a sequence of editing performed on intermediate condition images to intentionally create unrealistic and novel outcomes. The edges were edited and assembled before the generation using Photoshop, and then uploaded to the interface as inputs. The user evaluated the current outcome, and then iteratively fine-tuned the edges according to their preferences.

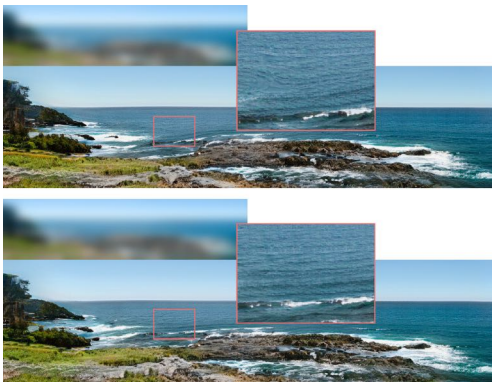

Figure 12: Slightly warping the intermediate image fixes the peculiar effect in the waves.

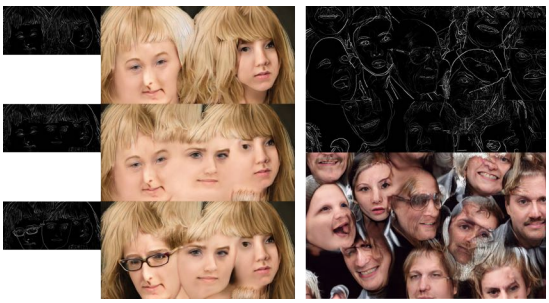

Figure 13: Editing, assembling the condition images to intentionally create unrealistic and novel outcomes

5.2.4. IterativeLoop

The MIGAI framework uses *IterativeLoop* to describe humans reflectively learning from the co-creative process. We observed that some participants might memorise useful patterns of shapes and colours that may trigger satisfactory results, and later use the paper card to reproduce these patterns. Some participants describe this process as a way to learn the preferences of the AI model to steer the co-creation through reflection.

5.3. Discussion

The motivation of this paper was to augment style-based GAN models into a co-creative context. By extending the StyleGAN model to image-conditional generation, these models allow human agents to ideate a high-level concept of the artefacts, like blurred arrangement or edges. The flexibility of input and the semantically meaningful control are critical features for effective co-creative experiences. Meanwhile, to create a sense of a "cooperated partner", the computational agent needs to maintain a certain level of unpredictability, and AI's contribution needs to partly influence the human agents' decision [49]. Therefore, the co-creative process balances machine autonomy and human creativity. Our current results are good indications that human agents act as a director who steers the model by organising and reusing learned knowledge in the computational agent. While further studies on the model architecture may improve the generation quality, the current results in this research show that bridging StyleGAN3 to the co-creative context is possible. Furthermore, it could be employed for novel and unique co-creative experiences. For example, Figure 14 shows an interactive installation with webcam inputs that produce 704×1280 images in real-time.

Figure 14: An interactive installation implementing our framework, the model runs in real-time with webcam input at 10fps.

6. Conclusion

In this work, we augmented StyleGAN3 with the ability of image-conditional generation, enabling it to turn high-level visual ideas into generation. This augmentation aligned latent variable models with the co-creative patterns mentioned in the MIGAI framework and brought StyleGAN3 into a co-creative context. We adapted the existing model architecture and proposed an encoder network to extract information from the conditional image. We demonstrated the modified architecture, StyleGAN-Canvas, on various image-to-image translation tasks with different datasets. In addition, we deployed our models to a graphic interface to facilitate the real-time interaction between users and the model. To evaluate our models in a co-creative context, we conducted qualitative human opinion studies and identified potential co-creative patterns using the MIGAI analytic framework.

6.1. Limitation and Future Works

An overall criticism was the need for more 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.

Besides, the proposed architecture can also be improved in technical aspects. Our approach proposes an alternative for 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 to improve the generation quality. Our model architecture utilises the equivariant generator in StyleGAN3. However, our feature extractor still needs to be rotation equivariant. Therefore, the generation may suffer when the rotation is not encoder. Figure 15 show an example of failure where the encoder does not preserve the rotation. It would be beneficial to make the feature encoder equivariant in future work.

Figure 15: encoder failed to extract the rotation

6.2. Ethical Considerations and Energy Consumption

Potential negative societal impacts of images produced by GAN [50] were considered 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 300 hours of computation on A100 SXM4 80GB (TDP of 400W). Total emissions are estimated to be 25.92kg CO_2, as calculated by MachineLearning Impact calculator [51]. Paper cards in the experiments were limitedly allocated to participants, reused during and after the experiments.

Acknowledgments

This research was carried out under the supervision of Prof. Mick Grierson. I sincerely appreciate and treasure feedbacks and comments from him.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014. arXiv:arXiv:1406.2661.

[2] A. Sauer, K. Schwarz, A. Geiger, Stylegan-xl: Scaling stylegan to large diverse datasets, 2022. arXiv:arXiv:2202.00273.

[3] B. Liu, Y. Zhu, K. Song, A. Elgammal, Towards faster and stabilized GAN training for high-fidelity few-shot image synthesis, CoRR abs/2101.04775 (2021). URL: https://arxiv.org/abs/2101.04775. arXiv:2101.04775.

[4] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free generative adversarial networks, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 852–863. URL: https://proceedings.neurips.cc/paper/2021/file/076ccd93ad68be51f23707988e934906-Paper.pdf.

[5] S. Berns, T. Broad, C. Guckelsberger, S. Colton, Automating generative deep learning for artistic purposes: Challenges and opportunities, 2021. arXiv:arXiv:2107.01858.

[6] M. Som, Mal som @errthangisalive, 2020. URL: http://www.aiartonline.com/highlights-2020/mal-som-errthangisalive/.

[7] D. Schultz, Artificial images, 2020. URL: https://artificial-images.com/project/you-are-here-machine-learning-film/.

[8] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2337–2346. doi:10.48550/ARXIV.1903.07291.

[9] J. Rafner, S. Langsford, A. Hjorth, M. Gajdacz, L. Philipsen, S. Risi, J. Simon, J. Sherson, Utopian or dystopian?: Using a ml-assisted image generation game to empower the general public to envision the future, in: Creativity and Cognition, Association for Computing Machinery, New York, NY, USA, 2021, p. 5. URL: https://doi.org/10.1145/3450741.3466815. doi:10.1145/3450741.3466815.

[10] Y. Wang, W. Zhou, J. Bao, W. Wang, L. Li, H. Li, Clip2gan: Towards bridging text with the latent space of gans, 2022. URL: https://arxiv.org/abs/2211.15045. doi:10.48550/ARXIV.2211.15045.

[11] J. Simon, Artbreeder, 2018. URL: https://www.artbreeder.com/about.

[12] S. Shahriar, Gan computers generate arts? a survey on visual arts, music, and literary text generation using generative adversarial network, 2021. URL: https://arxiv.org/abs/2108.03857. doi:10.48550/ARXIV.2108.03857.

[13] M. Muller, J. D. Weisz, W. Geyer, Mixed initiative generative ai interfaces: An analytic framework for generative ai applications, in: Proceedings of the Workshop The Future of Co-Creative Systems-A Workshop on Human-Computer Co-Creativity of the 11th International Conference on Computational Creativity (ICCC 2020), 2020.

[14] A. Spoto, N. Oleynik, Library of mixed-initiative creative interfaces, 2017. URL: http://mici.codingconduct.cc/.

[15] I. Grabe, M. G. Duque, S. Risi, J. Zhu, Towards a framework for human-ai interaction patterns in co-creative gan applications, in: Joint Proceedings of the ACM IUI Workshops 2022, March 2022, Helsinki, Finland, 2022. URL: https://ceur-ws.org/Vol-3124/paper9.pdf.

[16] J. Kim., M. L. Maher., S. Siddiqui., Collaborative ideation partner: Design ideation in human-ai co-creativity, in: Proceedings of the 5th International Conference on Computer-Human Interaction Research and Applications (CHIRA 2021), INSTICC, SciTePress, 2021, pp. 123–130. doi:10.5220/0010640800003060.

[17] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, CVPR (2017).

[18] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, 2017. URL: https://arxiv.org/abs/1711.11585. doi:10.48550/ARXIV.1711.11585.

[19] Y. Li, X. Chen, F. Wu, Z.-J. Zha, Linestofacephoto: Face photo generation from lines with conditional self-attention generative adversarial network, 2019. URL: https://arxiv.org/abs/1910.08914. doi:10.48550/ARXIV.1910.08914.

[20] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, 2018. URL: https://arxiv.org/abs/1812.04948. doi:10.48550/ARXIV.1812.04948.

[21] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: ICCV, 2017.

[22] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, 2019. URL: https://arxiv.org/abs/1912.04958. doi:10.48550/ARXIV.1912.04958.

[23] N. Dey, A. Chen, S. Ghafurian, Group equivariant generative adversarial networks, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=rgFNuJHHXv.

[24] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, R. Ng, Fourier features let networks learn high frequency functions in low dimensional domains, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 7537–7547. URL: https://proceedings.neurips.cc/paper/2020/file/55053683268957697aa39fba6f231c68-Paper.pdf.

[25] Y. Alaluf, O. Patashnik, Z. Wu, A. Zamir, E. Shechtman, D. Lischinski, D. Cohen-Or, Third time's the charm? image and video editing with stylegan3, 2022. URL: https://arxiv.org/abs/2201.13433. doi:10.48550/ARXIV.2201.13433.

[26] E. Richardson, Y. Weiss, The surprising effectiveness of linear unsupervised image-to-image translation, 2020. URL: https://arxiv.org/abs/2007.12568. doi:10.48550/ARXIV.2007.12568.

[27] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, D. Cohen-Or, Encoding in style: a stylegan encoder for image-to-image translation, 2020. URL: https://arxiv.org/abs/2008.00951. doi:10.48550/ARXIV.2008.00951.

[28] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, 2015. URL: https://arxiv.org/abs/1505.04597. doi:10.48550/ARXIV.1505.04597.

[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. URL: https://arxiv.org/abs/1512.03385. doi:10.48550/ARXIV.1512.03385.

[30] X. Liu, G. Yin, J. Shao, X. Wang, H. Li, Learning to

predict layout-to-image conditional convolutions for semantic image synthesis, 2019. URL: https://arxiv.org/abs/1910.06809. doi:10.48550/ARXIV.1910.06809.

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

[32] A. Liapis, G. Smith, N. Shaker, Mixed-initiative content creation, Springer International Publishing, Cham, 2016, pp. 195–214. URL: https://doi.org/10.1007/978-3-319-42716-4_11. doi:10.1007/978-3-319-42716-4_11.

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[34] R. Gal, O. Patashnik, H. Maron, G. Chechik, D. Cohen-Or, Stylegan-nada: Clip-guided domain adaptation of image generators, 2021. arXiv:2108.00946.

[35] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, T. Aila, Training generative adversarial networks with limited data, in: Proc. NeurIPS, 2020.

[36] X. Mao, L. Cao, A. T. Gnanha, Z. Yang, Q. Li, R. Ji, Cycle encoding of a stylegan encoder for improved reconstruction and editability, 2022. URL: https://arxiv.org/abs/2207.09367. doi:10.48550/ARXIV.2207.09367.

[37] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, Unet 3+: A full-scale connected unet for medical image segmentation, 2020. URL: https://arxiv.org/abs/2004.08790. doi:10.48550/ARXIV.2004.08790.

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

[39] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014. URL: https://arxiv.org/abs/1409.1556. doi:10.48550/ARXIV.1409.1556.

[40] R. Meyes, M. Lu, C. W. de Puiseau, T. Meisen, Ablation studies in artificial neural networks, 2019. URL: https://arxiv.org/abs/1901.08644. doi:10.48550/ARXIV.1901.08644.

[41] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, A. A. Efros, Generative visual manipulation on the natural image manifold, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 597–613.

[42] J. Canny, A computational approach to edge detection, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8 (1986) 679–698. doi:10.1109/TPAMI.1986.4767851.

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

[44] R. Xu, X. Wang, K. Chen, B. Zhou, C. C. Loy, Positional encoding as spatial inductive bias in gans, in: arxiv, 2020.

[45] T. Broad, F. F. Leymarie, M. Grierson, Network bending: Expressive manipulation of deep generative models, 2020. URL: https://arxiv.org/abs/2005.12420. doi:10.48550/ARXIV.2005.12420.

[46] T. Broad, S. Berns, S. Colton, M. Grierson, Active divergence with generative deep learning – a survey and taxonomy, 2021. arXiv:arXiv:2107.05599.

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

[48] V. Braun, V. Clarke, Thematic analysis., APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological. (2012) 57–71. doi:10.1037/13620-004.

[49] M. T. Llano, M. d'Inverno, M. Yee-King, J. McCormack, A. Ilsar, A. Pease, S. Colton, Explainable computational creativity (2022). URL: https://arxiv.org/abs/2205.05682. doi:10.48550/ARXIV.2205.05682.

[50] V. U. Prabhu, D. A. Yap, A. Wang, J. Whaley, Covering up bias in celeba-like datasets with markov blankets: A post-hoc cure for attribute prior avoidance, 2019. URL: https://arxiv.org/abs/1907.12917. doi:10.48550/ARXIV.1907.12917.

[51] A. Lacoste, A. Luccioni, V. Schmidt, T. Dandres, Quantifying the carbon emissions of machine learning, 2019. URL: https://arxiv.org/abs/1910.09700. doi:10.48550/ARXIV.1910.09700.