State-of-the-Art in the Architecture, Methods and Applications of StyleGAN

AMIT H. BERMANO, Tel Aviv University
RINON GAL, Tel Aviv University
YUVAL ALALUF, Tel Aviv University
RON MOKADY, Tel Aviv University
YOTAM NITZAN, Tel Aviv University
OMER TOV, Tel Aviv University
OR PATASHNIK, Tel Aviv University
DANIEL COHEN-OR, Tel Aviv University

Generative Adversarial Networks (GANs) have established themselves as a prevalent approach to image synthesis. Of these, StyleGAN offers a fascinating case study, owing to its remarkable visual quality and an ability to support a large array of downstream tasks. This state-of-the-art report covers the StyleGAN architecture, and the ways it has been employed since its conception, while also analyzing its severe limitations. It aims to be of use for both newcomers, who wish to get a grasp of the field, and for more experienced readers that might benefit from seeing current research trends and existing tools laid out.

Among StyleGAN's most interesting aspects is its learned latent space. Despite being learned with no supervision, it is surprisingly well-behaved and remarkably disentangled. Combined with StyleGAN's visual quality, these properties gave rise to unparalleled editing capabilities. However, the control offered by StyleGAN is inherently limited to the generator's learned distribution, and can only be applied to images generated by StyleGAN itself. Seeking to bring StyleGAN's latent control to real-world scenarios, the study of GAN inversion and latent space embedding has quickly gained in popularity. Meanwhile, this same study has helped shed light on the inner workings and limitations of StyleGAN. We map out StyleGAN's impressive story through these investigations, and discuss the details that have made StyleGAN the go-to generator. We further elaborate on the visual priors StyleGAN constructs, and discuss their use in downstream discriminative tasks. Looking forward, we point out StyleGAN's limitations and speculate on current trends and promising directions for future research, such as task and target specific fine-tuning.

CCS Concepts: • **Computing methodologies** → **Computer graphics**; *Learning latent representations*; *Image manipulation*; Neural networks.

1 INTRODUCTION

The ability of GANs to generate images of phenomenal realism at high resolutions is revolutionizing the field of image synthesis and manipulation. More specifically, StyleGAN [Karras et al. 2019] has reached the forefront of image synthesis, gaining recognition as the state-of-the-art generator for high-quality images and becoming the de-facto golden standard for the editing of facial images. See Figure 1, top for some visual examples.

StyleGAN presents a fascinating phenomenon. It is unsupervised, and yet its latent space is surprisingly well behaved. As it turns out, it is so well behaved that it even supports linear latent arithmetic. For example, it supports adding a vector representing age to a set of latent codes, resulting in images representing the original individuals, but older. Similarly, it has been demonstrated that StyleGAN arranges its latent space not only linearly, but also in a disentangled manner, where traversal directions exist that alter only specific image properties, while not affecting others. Such properties include global, domain-agnostic aspects (e.g., viewing angles or zoom), but also domain-specific properties such as expressions or gender for human faces, car colors, dog breeds, and more (see Figure 1, and Figure 2). Exploring what these qualities entail, recent StyleGAN-based work has presented astounding realism, impressive control, and inspiring insights into how neural networks operate.

Fig. 1. Images synthesized by StyleGAN, its followups and derivative works.

As groundbreaking as it may be, these powerful editing capabilities only reside within the model's latent space, and hence only operate on images generated by StyleGAN itself. Seeking to bring real-world images to the power of StyleGAN's latent control, inversion into StyleGAN's latent space has received considerable attention. Further harnessing StyleGAN's powers, other applications have also arisen, bringing contributions to the worlds of segmentation, augmentation, explainability, and others.

In this report, we map out StyleGAN's phenomenal success story, along with analyzing its severe drawbacks. We start by discussing the architecture itself and analyze the role it plays in creating the leading generative model since its conception in 2018. We then shift the discussion to the resources and characteristics StyleGAN's training requires, and lay out the work that reduces, re-uses, and recycles it.

Fig. 2. Editing a real image of Scarlett Johansson (on the top left) with StyleGAN. We show both in-domain and out-of-domain manipulations.

In Section 3, we discuss StyleGAN's latent spaces. We show how linear editing directions can be found, encouraged, and leveraged into powerful semantic editing. We inquire into what properties StyleGAN can and cannot disentangle well and dive into a surprisingly wide variety of approaches to achieve meaningful semantic latent editing.

In Section 4, the quest for applying StyleGAN's power in real-world scenarios turns to a discussion about StyleGAN inversion. To express a given real image in StyleGAN's domain, many different approaches have been suggested, all of which thoroughly analyze and exploit the generator's architecture. Some propose latent code optimization and others apply data-driven inference. Some works seek an appropriate input seed vector, while others interface with StyleGAN at other points along the inference path, greatly increasing its expressive power. Unsurprisingly though, it turns out that the well-behaved nature of StyleGAN's latent space diminishes in regions far from its well-sampled distribution. This in practice means that given a real-life image, its accurate reconstruction quality (or *distortion*) comes at the cost of *editability*. Finding different desired points on this reconstruction-editability trade-off is a main point of discussion in the works covered in this section.

Encoding an image into StyleGAN's latent space has more merit than for image inversion per se. There are many applications where the image being encoded is not the one the desired latent code should represent. Such encoding allows for various image-to-image translation methods [Alaluf et al. 2021a; Nitzan et al. 2020; Richardson et al. 2021]. In Section 4, we present and discuss such supervised and unsupervised methods.

In Section 6, we show the competence of StyleGAN beyond its generative power and discuss the discriminative capabilities StyleGAN can be leveraged for. This includes applications in explainability, regression, segmentation, and more.

In most works and applications, the pre-trained StyleGAN generator is kept fixed. However, in Section 7, we present recent works that fine-tune the StyleGAN generator and modify its weights to bridge the gap between the training domain (in-domain) and the target domain, which could possibly be out-of-domain.

Each section addresses both the newcomer, with basic concepts and conceptual intuition, and the experienced, with a summary of the most established and promising approaches, along with some pointers regarding when to use them.

2 STYLEGAN ARCHITECTURES

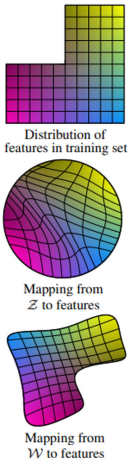
This report addresses the benefits hidden in Generative Adversarial Networks (GANs). First introduced by Goodfellow et al. [2014], GANs pose an interesting and unique approach. Two networks are interlocked in a perpetual game during training. One network, the Generator, seeks to generate images that are from the target distribution, while the other network, the Discriminator, seeks to distinguish between actual images from the training set and those created by the generator. The two networks start the training without any knowledge of the domain and spend the entire training process learning from each other. Conceptually, this could be thought of as a repetitive process where the generator finds a way to fool the discriminator, and the discriminator, in turn, finds a way to detect this "attack". This approach allows self-supervision, and hence these networks can be trained without explicit labeling.

StyleGAN, however, seems to do much more than reproduce samples from the target distribution. While following the adversarial learning process, it turns out that StyleGAN, more than other GANs, constructs a remarkably well-behaved latent space. Without any supervision, StyleGAN arranges examples it sees in a smooth, highly disentangled order, driven by powerful semantic understanding.

In this Section, we portray how StyleGAN's architecture is built, try to understand why this architecture induces such cutting-edge emerging disentanglement, and how the architecture can be improved to match specific needs, according to relevant literature.

StyleGAN1. The style-based generator architecture for generative adversarial networks, or StyleGAN for short, was first proposed by Karras et al. [2019]. At the core of StyleGAN's architecture lie the style modulation layers, from which StyleGAN draws its name. Borrowing from style-transfer literature, these layers are designed to enable control over the "style" of generated images by adjusting the statistics of the feature maps along the generative path. The generative path starts from a learned constant C, representing the epicenter of the distribution, and all the information and generative power of the network is injected through the style and an additional random noise vector n. In the first version of the architecture, [Karras et al. 2019], the style injection layers utilized the Adaptive Instance Normalization (AdaIN) mechanism [Huang and Belongie 2017]; each channel of the feature maps is first normalized to zero mean and unit variance, followed by re-scaling using new means and variances predicted from a given latent code.

However, the use of AdaIN layers was not the only major change proposed. Rather than injecting the network with a latent code z sampled directly from some Gaussian prior $\mathcal{Z}$, StyleGAN introduces a novel mapping network which converts these normal-distributed codes into vectors in an intermediate latent space $\mathcal{W}$. The authors propose an intuitive argument for adding such a network: the probability for sampling a particular combination of image attributes in the latent space should eventually match the probability for that combination to appear in the real dataset. For those cases where the data is not uniform with respect to these attributes, it follows that the mapping from $\mathcal{Z}$ to the image features must become curved in order to diminish the incidence rate of rare attribute combinations. A learned mapping network, however, could learn to "unwrap" the latent space back to a flat form, and simply account for probability densities by mapping fewer codes to regions that would otherwise portray a rare combination of attributes (see inset figure, from Karras et al. [2019]). Karras et al. postulate that this linearly-disentangled space is a more natural representation for the network, allowing it to more easily recreate a wide range of variations. As Karras et al. and follow-up works demonstrate, the learned latent spaces of StyleGAN offer considerable disentanglement. These innovations give rise to a network that, at the time, was unrivaled in quality, invertibility, and support for a wide range of generative and discriminative tasks.



StyleGAN2. With StyleGAN being quickly adopted into widespread use, it was inevitable that artifacts inherent to the model would come to light. These included characteristic water-droplet-shaped blobs which consistently appeared in all images. Additionally, a "texture-sticking" effect was observed, where certain attributes of the generated image, such as the teeth or eyes, would display strong spatial bias, remaining fixed to specific image coordinates even through latent space interpolations. In a follow-up work, Karras et al. [2020b] identify the source of these artifacts, and re-design key aspects of the network to correct them. They first identify that the water-droplets are a manifestation of a flaw in StyleGAN's normalization scheme - by normalizing each feature map separately, any information found in their relative magnitudes is destroyed. The generator would then hide information about the signal strength through localized spikes that dominate the statistics. They overcome this hurdle by shifting normalization from the feature maps, where the adaptive layer normalization forces new statistics, and onto a modulation of the convolutional kernels themselves. By doing so, they apply weaker normalization, based on the expected feature statistics rather than exact signal strength, and the network no longer needs to hide signal strength information - which in turn makes the blob-shaped artifacts disappear. This technique has also been shown to promote disentanglement between geometry and appearance in other scenarios [Yang et al. 2020b].

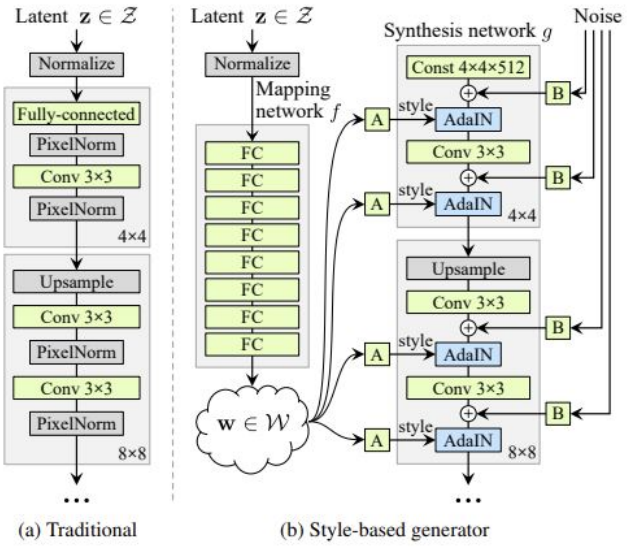


Fig. 3. The StyleGAN1 [Karras et al. 2019] architecture. The novel architecture is based on the progressive growing approach (b, right), combined with a Style injection mechanism (b, middle). In addition, a mapping network (b, left) deforms the Gaussian Z space to better match the distribution of the training data.

Fig. 4. An example for the "texture sticking" effect [Karras et al. 2020b]. As can be seen, the teeth do not follow the head when rotated, but rather remain attached to their absolute position in the image.

The "texture sticking" effect, meanwhile, was hypothesized to be an artifact of the progressive growth scheme. Karras et al. suggest that in such a setup, every resolution block serves as an output block for some stage of the training process. In such a scenario, the network attempts to create excessive high-frequency detail in these intermediate resolutions, which leads to aliasing along the generative path and in turn breaks shift-invariance [Zhang 2019]. They address this issue by revisiting progressive growing and replacing it with a skip-connection-based architecture, where each resolution block outputs a residual, which is summed up and up-scaled. These modifications, coupled with a novel path-length regularization loss and in-depth analysis of network capacity, lead to improvements both in standard quality metrics such as FID, but also in the ability to invert images into the latent space of the GAN.

StyleGAN3. At first, StyleGAN2 appeared to address the "texture-sticking" problem. However, more careful analysis revealed that, while the issue was resolved for large-scale objects such as the mouth or the eyes, it remained present when examining finer details such as hair or beards. To resolve this issue, Karras et al. sought out the various sources through which spatial information could leak into the convolutional operations, with the aim of fully restoring translational invariance to the network. These sources include the image borders, per-pixel noise inputs, positional encoding, and the aliasing caused by careless treatment of upsampling filters and non-linearities such as ReLUs. Through a series of small architectural changes coupled with a rigorous signal processing approach, these sources of unwanted information were removed, and translation and rotational equivariance was restored. The novel architecture of StyleGAN3 [Karras et al. 2021] brought with it remarkable improvements, leading to considerably smoother interpolations. However, the new approach brought with it new challenges. Karras et al. observe that when conducting layer mixing experiments, some properties were not cleanly inherited from just one of the codes. Preliminary investigations of the network also revealed newly introduced artifacts, from the tendency of generated faces to have a single frontal tooth, to the appearance of a faint "grid" to which background features and fine details such as hair would often get stuck. These phenomena suggest degraded disentangled properties, however, as of writing these words, the novel alias-free architecture is still in its infancy, and it remains to be seen what unique uses, improvements or challenges arise from it.

Parallel to these improvements, various works sought to identify areas in which StyleGAN could be improved. Lin et al. [2021] note that the high computational cost of full-resolution image generation makes it impractical to utilize the network for interactive editing on edge devices. They proposed an elastic generator architecture that could produce previews at lower resolutions while retaining the same latent semantics. A user could then edit these previews with a fraction of the computational budget, restoring the output to its full resolution only as a final step. [Gal et al. 2021a] followed prior observations which revealed a flaw in the generator's ability to produce high-frequency details. They demonstrated that some patterns are beyond the network's ability to recreate and linked the flaw to the inherent spectral bias of neural networks. They proposed to tackle this by shifting the generation to the frequency domain, realized by a first-level wavelet decomposition. By doing so, they reduced the network's need to learn high-frequency functions and achieved a more faithful generation of high-frequency patterns.

In an alternative approach to synthesis, Anokhin et al. [2021] forgo convolutions and instead design a style-based network which, given the coordinates of a pixel and a style code, predicts the color of that pixel. This conditionally-independent pixel synthesis approach (CIPS) was able to rival the quality of images produced by traditional convolutional methods while enabling novel synthesis applications such as the creation of cylindrical panoramas. Sendik et al. [2020] hypothesize that the single learned constant at the root of the generative path is a limiting factor when training on sets that contain multiple modalities. They hence develop a multi-constant model, where the generator could better represent the dataset modalities by assigning them different mixtures of constants. Kwon et al. [2021]

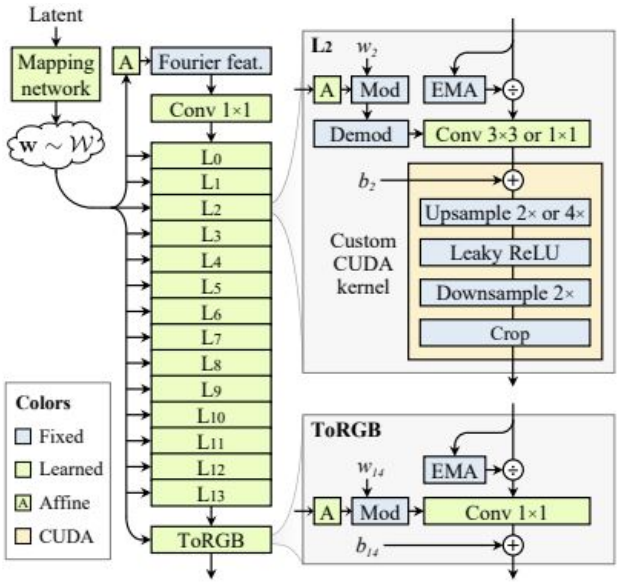


Fig. 5. StyleGAN3 [Karras et al. 2021] architecture. The main components of the architecture remain similar to previous versions. A series of small architectural changes, derived from rigorous signal processing analysis, renders the new version of StyleGAN equivariant to rotation and translation.

propose augmenting the network with Diagonal Spatial Attention (DAT) layers, which modulate the network's feature maps along the spatial directions. These modulations are in turn controlled through an additional latent code. Through this addition and an appropriate loss term, the authors disentangle "content" from "style", allowing a user to control spatial features such as pose or expression, without affecting style traits such as color or makeup. Casanova et al. [2021] suggest training a GAN conditioned on a single input. Their intuition is that unconditional GANs face difficulties in reproducing complex distributions [Liu et al. 2020b; Lučić et al. 2019] such as ImageNet [Russakovsky et al. 2015]. Typical conditional models seek to resolve this challenge [Brock et al. 2018] by conditioning the synthesis process on class labels, thereby partitioning the data into multiple clusters which are more easily modeled. However, acquiring such labels is labor intensive. Instead, they suggest partitioning the data into overlapping neighborhoods by clustering the data in some pre-trained feature space. The "label" associated with an image is then the feature vector in this space, and real images observed by the discriminator when conditioned on such a vector are sampled from the group of images with representations most similar to the given vector. In this way, the network learns to generate images sharing visual and semantic traits with a given sample.

While not strictly extensions of StyleGAN itself, a large body of work nevertheless draws inspiration from its novel architecture. These works typically repurpose the style-based modulation layers or mapping network and incorporate them into new generative frameworks. One line of work aims to merge the growing Transformer [Vaswani et al. 2017] literature with image synthesis. Hudson et al. [2021] proposed the Generative Adversarial Transformer, which utilizes a bipartite mechanism through which the latent codes and image features attend and influence each other.

Others have proposed to entirely replace the convolutional blocks with transformer-based modules such as ViT [Dosovitskiy et al. 2021; Lee et al. 2022], Linformer [Park and Kim 2021; Wang et al. 2020], or the Swin Transformer [Liu et al. 2021b; Zhang et al. 2021b]. While these have yet to achieve the same fidelity or widespread use as their progenitor, they have already shown considerable progress in layout control and convergence times. Moving towards 3D representations, a set of recent works propose to marry the style-based architecture with implicit models, such as Signed Distance Functions [Park et al. 2019] or Neural Radiance Fields [Mildenhall et al. 2020]. These models leverage weight and feature modulations [Gu et al. 2022; Or-El et al. 2021; Xu et al. 2021a; Zhou et al. 2021] or directly employ a StyleGAN network to predict a set of feature planes that serve as inputs to a small implicit network [Chan et al. 2021a]. These works achieve impressive visual quality, enable explicit control over pose, and can be used to predict detailed surface representations. However, their increased memory requirements have so far prevented them from reaching the resolution and quality of StyleGAN itself.

Training Data. "An open secret in contemporary machine learning is that many models work beautifully on standard benchmarks but fail to generalize outside the lab" [Jahanian et al. 2019]. Indeed, StyleGAN is no different. It is recognized in the literature that unsupervised training is more difficult when learning a complex domain [Casanova et al. 2021]. In the case of StyleGAN, the learned domain seems to require strict structure. The data domain should be almost convex, i.e., between every two points there should be valid samples that interpolate them on the data manifold. For this reason, for example, it is difficult to construct a full human body model. For the same reasons, StyleGAN does not handle multi-modal distributions well and behaves poorly for scenes where objects do not have specific potential locations. In recent work, Sauer et al. [2022] demonstrate that some of these challenges may be overcome through careful model scaling, though whether or not StyleGAN's unique latent-space properties persist through this modification remains an open question. In the future, we will likely witness additional works that address explicit data issues, i.e., works that try to apply StyleGAN to other types of data, perhaps by dropping or adding examples during training to make the data's landscape more smooth, by transfer learning between datasets (see Section 7), by more directly addressing multi-modalities in the data, or by incorporating more elaborate attention mechanisms into the architecture.

2.1 Latent Spaces

Unlike common GANs, StyleGAN has more than one innate latent space. Moreover, to increase the expressive power of StyleGAN, it is common to work with extensions of these spaces, illustrated in Figure 6. Here, we review the commonly used spaces and describe the differences between them.

- The first latent space is $\mathcal{Z}$ in the sense that random latent codes can be sampled from it to be inserted into the generator itself. $\mathcal{Z}$ is defined to be a normally distributed space, and it is the only space that has a closed-form definition. Therefore, images that belong to the GAN's manifold can be easily sampled from $\mathcal{Z}$.
- Latent codes from $\mathcal{Z}$ are transformed to latent codes in $\mathcal{W}$ through an MLP, commonly referred to as the *mapping network*. In a sense, the distribution of $\mathcal{W}$ is learned, and therefore better matches the distribution of the real data compared to the original $\mathcal{Z}$ space. This learned distribution provides the virtue of disentanglement. Many works employ this disentanglement property to achieve semantic image editing by traversing the latent space.
- Latent codes in $\mathcal{W}$ are not directly inserted into the synthesis network. Instead, each latent code in $\mathcal{W}$ is first transformed through a learned affine transformation. Such an affine transformation is learned during the training of each layer of the synthesis network. The space spanned by the outputs of these transformations is commonly referred to as the *StyleSpace*, or $\mathcal{S}$. Unlike $\mathcal{W}$ in which a single latent code is used for generating an image, in $\mathcal{S}$ there are several latent codes for a single image, one for each affine transformation block (e.g., 26 for a generator with a 1024×1024 output resolution). It has been shown [Wu et al. 2020] that $\mathcal{S}$ is even more disentangled than $\mathcal{W}$. More specifically, each dimension, or channel, of $\mathcal{S}$ tends to control a single semantic attribute of the generated image. Therefore, by carefully manipulating the dimensions of $\mathcal{S}$ it is possible to obtain highly disentangled edits.
- Representing real images with StyleGAN remains a challenge. The good properties of $\mathcal{W}$ have attracted most works aiming at representing real images to focus on it. Abdal et al. [2019] propose working in an extended latent space, denoted by $\mathcal{W}+$ In $\mathcal{W}+$, one inserts a different latent code for each layer of the synthesis network (e.g., 18 for a generator with a 1024×1024 resolution). StyleGAN was not trained on $\mathcal{W}+$ and thus images sampled from it do not necessarily have high quality. Moreover, it should be noted that oftentimes, when operating in $\mathcal{W}+$, it is possible to reach areas that are outside the learned distribution of $\mathcal{W}$. Such areas further push the latent code outside the distribution over which the generator was trained on. As the distribution of $\mathcal{W}$ cannot be explicitly modeled, keeping the latent code in the trained distribution is a challenging task.
- To alleviate the need of preserving the latent code inside the distribution of $\mathcal{W}$, it is possible to work with an extension of $\mathcal{Z}$ instead of $\mathcal{W}$. Similarly to the definition of $\mathcal{W}+$, in $\mathcal{Z}+$ [Song et al. 2021] a different latent code is sampled for each layer of the synthesis network (e.g., 18 for a 1024×1024-resolution generator). Note, that in $\mathcal{S}$ there is no notion of $\mathcal{S}+$ as the latent codes for each layer are different by design.

3 LATENT SPACE EDITING

Perhaps the most exciting aspect of GAN learning is the way the latent space is arranged in a well-trained GAN. Traditionally, GANs in general, and StyleGAN specifically, can be used to simply generate a wide variety of images of the same kind. These can serve as a form of data augmentation for downstream training (see Section 6). However, it has been shown that GANs tend to arrange their latent space smoothly, *i.e.* such that close regions in the latent space depict similar images.

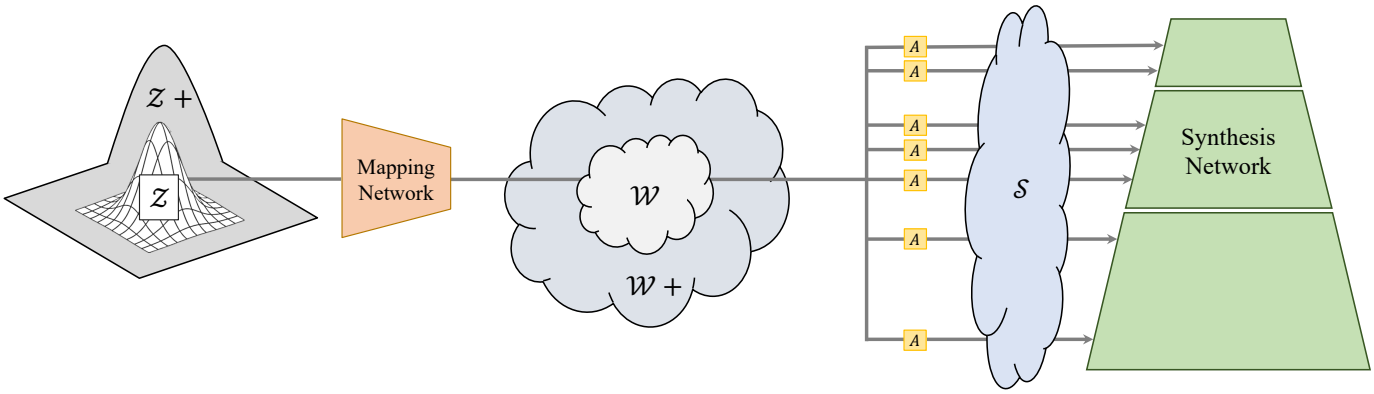


Fig. 6. The StyleGAN architecture and its latent spaces. A random latent code z is sampled from the normally distributed latent space $\mathcal{Z}$ (on the left), then transformed to the learned latent space $\mathcal{W}$ through an MLP mapping, passed through a set of different learned affine transformations (denoted by A) to reach the $\mathcal{S}$ space, and finally inserted into the synthesis network. It is common to work in the extended spaces of $\mathcal{Z}$ and $\mathcal{W}$, referred to as $\mathcal{Z}$+ and $\mathcal{W}$+, respectively.

This, combined with the notion that GANs produce images that are within the distribution of the target domain gives rise to *latent-based editing*. In other words, the two concepts suggest that traversing the latent space yields a path of smoothly changing images, each of them on their own belonging to the target domain (*e.g.* realistic human faces). This could be thought of as geodesic traversal on the manifold of all valid images. Even the first works in generative modeling already demonstrated how latent code interpolation between two examples yields a natural morphing between them [Goodfellow et al. 2014]. As it turns out, careful traversal in the latent space can also produce desirable semantic changes in the resulting image that would otherwise be very difficult to perform. These include changes in viewpoint, lighting conditions, and domain-specific attributes such as expressions for faces, colors for cars, or widths of buildings. Of course, the most desirable edits are the disentangled ones — those that change one attribute without affecting any other. Applications of such powerful editing tools are endless, from automatically adding smiles to facial images, through interior design explorations, to rapid car design.

In this aspect, StyleGAN shines. As previously discussed (Section 2), StyleGAN operates best on well-structured data. When trained on such data, StyleGAN constructs a highly disentangled latent space in an unsupervised manner, simply by virtue of inductive bias. Many techniques have been proposed to traverse this latent space and facilitate semantically disentangled latent-based editing. Of all sections in this report, the editing art is the most diverse, presenting creative approaches borrowed from different fields.

Early approaches to this task pointed out that StyleGAN's latent space is so well behaved and disentangled, that it even supports linear latent space arithmetics. These linear editing works demonstrate, for example, that to make a face older, one can traverse in a specific, pre-computed, direction. These works come in two main flavors — supervised and unsupervised. The first works in the field have presented a thorough analysis of GAN behavior [Jahanian et al. 2019; Liu et al. 2020a] (including StyleGAN), and showed how one can identify linear traversal directions that present high disentangled qualities. Using edits that are easily attainable in image space (e.g. 2D rotation or zoom, and pan), they look for directions in the $\mathcal{W}$ space (Section 2.1) that produce the same effect. Changing the magnitude of traversal along these directions induces a disentangled edit that is weaker or stronger according to the step size. This early work also drew conclusions regarding the extent of the space's linearization. That is, they show that going too far along a direction will eventually break the disentanglement, and affect other crucial factors of the image. They also offer an analysis on, and a way to improve, the extent of the linearization.

Supervised Linear Approaches. The most natural approach to finding editing directions is to do so explicitly, through full supervision. Perhaps one of the most noteworthy works in linear editing is InterFaceGAN [Shen et al. 2020a,b]. This work leverages per-image binary annotations to identify hyper-planes in the latent space that separate the two binary attribute values. These planes can be found using Support Vector Machines (SVMs). Then, to edit one attribute without affecting others, one finds a direction that is orthogonal to one plane and parallel to the others. Figure 7a depicts some of the typical editing directions extracted by this method. Yang et al. [2020a] further propose a way to evaluate how well the activations of specific layers are correlated with semantic attributes, based on 105 pretrained attribute classifiers. Recently, Wu et al. [2020] employ a pretrained classifier, or use a few images for direction identification. Their key idea is to identify correspondence between the most active channels and the semantics corresponding generated images depict (termed *semantic consistency*). The authors show that this correspondence indicates specific channels in the generator's activations that control very disentangled image characteristics. This offers a fine-grained approach to latent editing that is different from the popular latent-editing approaches which modify all activations of a layer (or more). Editing in $\mathcal{S}$ space (see Section 2.1) is shown to provide highly disentangled, spatially adaptive directions for editing.

Unsupervised Linear Approaches. In many cases, collecting the data required for supervised editing can be difficult or prohibitively expensive. To expand the range of available editing directions, despite these limitations, unsupervised editing methods have been proposed. Perhaps the first was proposed by Voynov et al. [2020]. The core idea is to predict a set of traversal directions and concurrently try to infer their meaning from the images corresponding to the code before and after the edit. They propose to jointly learn a set of directions and a model to identify the corresponding image transformations. Under this paradigm, the assumption is that directions that are easy to identify with high accuracy are likely candidates for disentangled editing directions. GANSpace [Härkönen et al. 2020], take a more natural approach, and simply search for the dominant directions in the latent codes of a dataset, using Principal Component Analysis (PCA). Alharbi et al. [2020] propose editing through adding random noise to the input learned constant, rather than augmenting the style input. They show that by enforcing a spatial structure to the noise, spatial disentanglement can be encouraged, and can be paired with the semantic disentanglement StyleGAN already offers. In all three cases, manual inspection is used to identify whether these directions indeed produce valuable edits, and infer their semantic meaning. SeFa [Shen and Zhou 2021] takes a different approach to the unsupervised editing problem. They propose analyzing the weights of the pretrained generator and identifying principle directions that are most affected by these weights. To do so, they perform an eigenanalysis of the matrix representing the latent-to-image space projection. This analysis is closed-form, meaning it is fast and does not require even sampling the network. This approach is still valuable and has been used for other domains and GAN architectures as well [Spingarn et al. 2020].

Non-linear Approaches. As may be expected, non-linear approaches can present higher quality editing at the cost of simplicity. Hou et al. [2022], operate similarly to Yang et al. [2020a] by using classifiers. However, they propose to move beyond global, linear directions and towards a non-linear traversal paradigm. In their case, a different direction is generated per example for the same editing operation. The editing is then performed by changing the latent code of only one layer at a time, in a style-mixing manner (see Section 2), thereby improving disentanglement. StyleFlow [Abdal et al. 2020b] is a seminal work in the realm of facial editing, presenting one of the most versatile and stable editing approaches, disentangled enough to produce a realistic result even when performing several editing operations serially, as can be seen in Figure 7b. The core idea for this work is the clever employment of normalizing flows — a method through which a bi-directional mapping can be obtained between the latent space and an input code, conditioned on specific attributes. This mapping is trained in a supervised manner through an elaborate multi-attribute classifier. This promising normalizing flow-based approach has also seen follow-up work in an unsupervised setting [Liang et al. 2021]. Alaluf et al. [2021a] use an age regression network to provide control over age in human faces. Looking towards more recent works along this line [Wang et al. 2021c], perhaps the state-of-the-art lies with DyStyle [Li et al. 2021a]. The main contribution of this supervised approach is a dynamic network, trained to handle multiple edits at the same time. Here, a different network is trained for each attribute, producing its own latent editing direction. For every training example, consisting of a different composition of desired edits, only the relevant networks are applied, with their outputted codes fused into one using a self-attention mechanism. This approach enables high-quality editing in flexible domains, especially when composing several edits together. The combined dynamic approach seems not only to improve sequential editing, but also provide enough regularization to improve the state-of-the-art for a single edit as well (see Figure 7c). Aiming for video editing, Yao et al. [2021] train a dedicated latent-code transformer to achieve more disentangled edits.

Different Supervision Modalities. Other approaches have been proposed that leverage supervision, but differ in nature from explicit attribute supervision or classification-based techniques. StyleRig [Tewari et al. 2020] suggests employing synthetic data to guide the editing process. They acquire a roughly 200-parameter 3D Morphable Face Model (3DMM) using traditional PCA over 200 input faces. This model can be used in a self-supervised manner to train a network to perform the editing over $\mathcal{W}$. Through a plethora of synthetically generated paired examples, the method finds high-quality edits. This is because perfect labeling can be assigned to images that are rendered by specific parameter changes in the 3DMM model. This approach, however, was only able to find high-quality editing directions for a subset of the face model parameters. Perhaps unsurprisingly, the successfully found directions do not enable more diverse edits compared to less supervised methods. A similar approach has also been proposed [Zhang et al. 2021a], using general meshes instead of 3DMM, for more diverse objects. Through differentiable rendering, parameters like camera position and object shape can be self-supervised easily. Taking this line of work a step further, Ghosh et al. [2020] propose generating the parameters of a 3D facial model learned from 4D scans. In this paradigm, the geometry is constructed through a learned 3D model (FLAME [Li et al. 2017]), and StyleGAN generates appearance and texture. Combining the two models offers more expressive facial variations in shape and expression, and an inherent disentanglement between geometry and appearance. FreeStyleGAN [Leimkühler and Drettakis 2021] use standard calibration tools to construct pairs of facial images and associated camera parameters. These pairs are used to learn explicit control over image views within the GAN's aligned image manifold. Taking this approach a step further, the authors use a flow-based model to learn an image mapping module that can transform the generated images beyond StyleGAN's aligned domain. HistoGAN [Afifi et al. 2021] employs color histograms to recolor images and paintings.

Several works employ the power of language. They guide edits by using textual descriptions, which are more global and abstract in nature. Patashnik et al. [2021], one of the first works to propose this approach, employs CLIP [Radford et al. 2021], a powerful pre-trained model that embeds text and imagery to a joint latent space. By finding traversal directions that bring the produced image and the desired text description closer together, this method demonstrated new and exciting semantic editing operations, such as makeup removal and specific hairstyles for human faces (see Figure 7e). TediGAN [Xia et al. 2021a] employ a novel architecture

and training process for the language model to be trained along with the generator. While potentially powerful, the resulting networks are not as expressive as language models pre-trained on web-scale data. Hence, they fail to achieve the same quality. Chefer et al. [2021] utilize CLIP to blend two facial images, demonstrating better preservation of the original identity while successfully transferring meaningful semantic features from the desired target images. Abdal et al. [2021a] find meaningful directions in CLIP-space in an unsupervised manner, map them to latent-space directions, and use CLIP to automatically generate natural language descriptions for these directions.

Finer Control. Several of the latest works propose operating in a more disentangled latent space — the $\mathcal{S}$ space [Liu et al. 2020a; Wu et al. 2020; Xu et al. 2021b] (see Section 2.1). However, it is significantly larger, posing a computational challenge. Furthermore, augmenting the generator activations themselves after the AdaIN (StyleGAN [Karras et al. 2019]) or Modulation (StyleGAN2 [Karras et al. 2020b]) layers, provides even finer control. This allows applying local changes in the image maps, rather than a global change. For example, Bau et al. [2021] offers users the ability to paint a mask in a given image and to describe in free text what this region of the image should depict. They do this by feeding the same modulation layer different style codes, according to the spatial location in the resulting map: one code for regions inside the painted mask, and one for the rest.

Albahar et al. [2021] suggest spatial control through the initial input constant, while leveraging the inherent semantic understanding StyleGAN naturally develops, reinforced by human pose labeling. Unlike most editing works, which manipulate the behavior of a pretrained StyleGAN, this work proposes architectural changes to the generator, to adapt it to human pose inputs. Through full supervision [Cao et al. 2019], they train StyleGAN to change the pose of human clothing models. Through pose labeling and paired UV coordinates, the clothes are warped in UV space to better match the new pose (see Figure 7c). Similarly, Abdal et al. [2020a] change the spatial activations to allow scribble-level control for the user (see Section 4 for more details). StyleFusion [Kafri et al. 2021] propose a new mapping architecture for StyleGAN to better disentangle a target attribute. This results in a learned blending between style codes, resulting in fine-grained local control of the edited images. They also introduce an additional latent code for controlling global aspects of the images (*e.g.* pose, lighting, background). Finally, StyleMapGAN [Kim et al. 2021] suggest an architectural change where the global $\mathcal{W}$ latent code is replaced with a spatial map, and the global style infusion layer (*i.e.* AdaIN or weight modulation) is replaced with a spatially adaptive one. This allows blending two images very naturally, with a high level of detail and finer local control (See Figure 7f).

Studying the wide and versatile editing works, it is clear that latent-based editing holds great potential and sparks the curiosity of many. Some of the most recent works present unprecedented quality, showcasing the expressive powers of GANs in general and of StyleGAN in particular. However, all of these works still operate in lab conditions. They present a handful of novel editing operations. These, however, are still restricted (*e.g.* only specific expressions

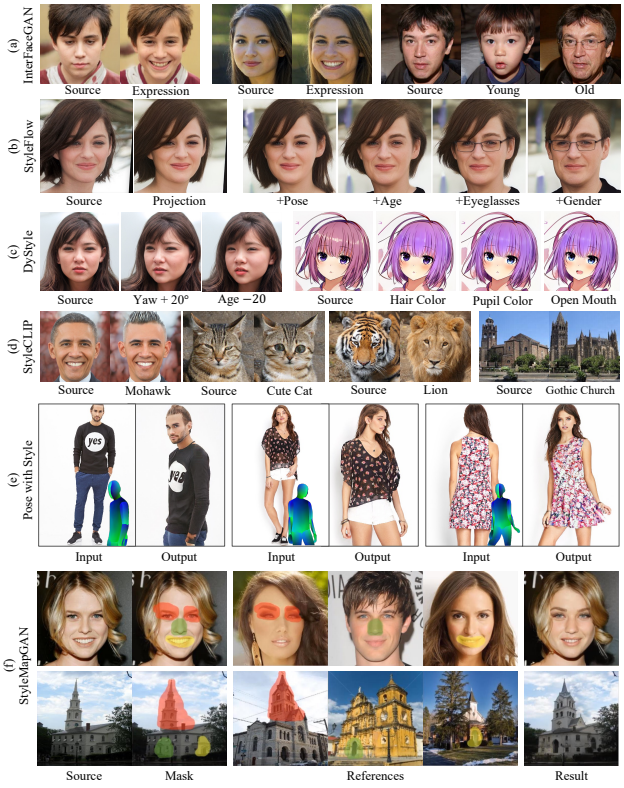


Fig. 7. Examples of prominent editing works. (a) InterfaceGAN [Shen et al. 2020b] extracts linear editing directions through attribute level supervision. (b) StyleFlow [Abdal et al. 2020b] is the first to present editing that is stable enough to be composed, through employing normalizing flows and attribute-level supervision. (c) DyStyle [Li et al. 2021a] addresses compositional editing directly, producing more accurate, elborate, and diverse editing. (d) StyleCLIP [Patashnik et al. 2021] employs free textual editing, through a visual-linguistic pretrained model [Radford et al. 2021]. (e) Pose with Style [AlBahar et al. 2021] employs human pose supervision to edit body poses and clothing. (f) StyleMapGAN [Kim et al. 2021] provides localized editing by augmenting StyleGAN's architecture with spatially adaptive modulation. Zoomed-in viewing recommended.

can be altered and the degree of possible changes in pose is limited). These restrictions pose practical challenges when employing StyleGAN for industrial or in-the-wild use. Furthermore, they bear the burden of the generator's limited capabilities regarding the versatility and structure of the training data (see Section 2). In the future, we will probably witness more works that adapt to new data on the fly, possibly using techniques such as fine-tuning (Section 7), or layer mixing, where several different models are trained, and their layers are mixed according to specific applications [Park et al. 2020; Pinkney and Adler 2020]. In any case, it seems that a core challenge editing works face is the evaluation of their quality, as discussed in Section 5.

4 ENCODING AND INVERSION

The success of the aforementioned latent space editing techniques results in a natural question of how to apply such techniques to edit real images (i.e., images not necessarily residing within the GAN's domain). To do so, we need to find the latent representation of a given image, a task commonly referred to as *GAN Inversion*. First introduced by Zhu et al. [2016], the inversion task aims to find a latent vector from which a pre-trained GAN can most accurately reconstruct the given image. Formally, given an input image x, we want to minimize the distortion of the reconstructed image obtained from the inverted latent code w using a well-trained generator G:

$$w^* = \arg\min_w \mathcal{L}\left(x, G(w)\right), \tag{1}$$

where $\mathcal{L}$ is some reconstruction loss (e.g., the LPIPS perceptual loss [Zhang et al. 2018] and/or the pixel-wise L2 loss). In the following, we explore the various core approaches for performing this inversion process, outlined in Figure 8.

4.1 GAN Inversion

Existing optimization-based GAN inversion methods search for the desired latent vector via a per-image latent vector optimization by solving Equation 1 [Abdal et al. 2019, 2020a; Bau et al. 2019; Creswell and Bharath 2018; Gu et al. 2020; Lipton and Tripathi 2017; Wulff and Torralba 2020; Yeh et al. 2017; Zhu et al. 2016, 2020b]. Early works performing optimization attempted to invert into StyleGAN's learned latent space $\mathcal{W}$. However, it has been shown that inverting a real image into a 512-dimensional vector $w \in \mathcal{W}$ is not expressive enough to accurately encode and reconstruct real images. As such, it has become common practice to invert images into an extended latent space $\mathcal{W}+$ [Abdal et al. 2020a] defined by a concatenation of multiple w vectors, one for each input of StyleGAN. While optimization techniques often result in near-perfect reconstructions of the input, they typically require several minutes to do so for a single image.

To accelerate this optimization process, some works trained an encoder over a large collection of images to learn a direct mapping from an image to its latent representation [Luo et al. 2017; Perarnau et al. 2016]. Here, the training objective can be defined by,

$$\theta_E^* = \arg\min_{\theta_E} \sum_i \mathcal{L}(\mathbf{x}_i, G(E_{\theta_E}(\mathbf{x}_i))), \tag{2}$$

where the weights θ_E^* of the encoder are sought. Pidhorskyi et al. [2020] propose a StyleGAN-based autoencoder, where the encoder network E is trained alongside the generator.

Many works have explored various avenues for improving the performance of encoder-based inversion methods in an attempt to close the gap in performance with optimization techniques. Some have explored various encoder architectures for improving the inversion quality.

Richardson et al. [2021] and Xu et al. [2021b] explore a hierarchical encoder based on a feature pyramid network (FPN) to better match the coarse, medium, and fine-level details of StyleGAN's hierarchical structure. For extracting the learned styles from the encoder's feature maps, Richardson et al. [2021] introduce 18 separate map2style modules, one for each input layer of StyleGAN. Wei et al. [2021] and

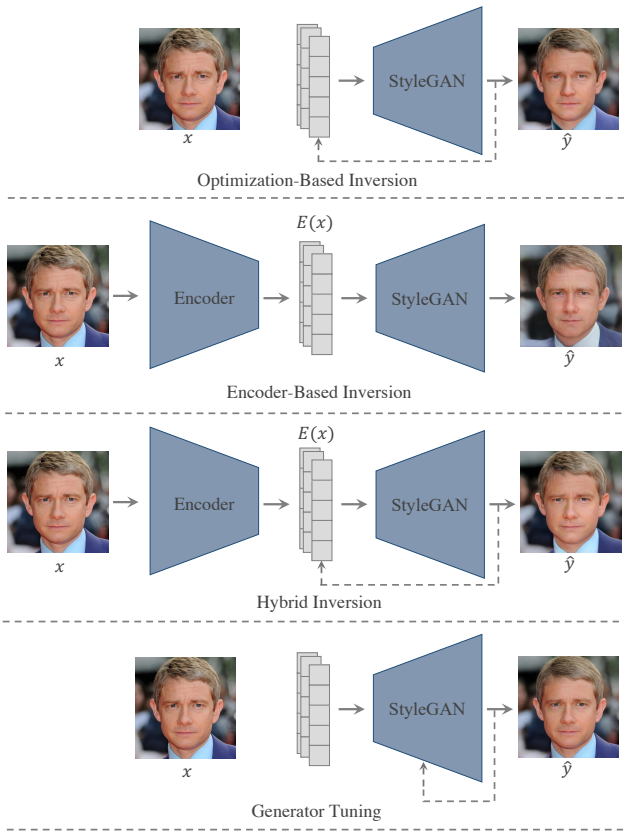


Fig. 8. Various approaches for GAN Inversion. **Optimization-based** techniques perform a per-image optimization procedure on the latent vector to minimize the reconstruction loss between x and $\hat{y}$. **Encoder-based** schemes aim to learn a direct mapping between the image x to its latent representation $E(x)$. **Hybrid** techniques attempt to combine "the best of both worlds" by initializing the optimization procedure with the inversion prediction of a trained encoder. Finally, recent **generating tuning** methods fix a latent code and learn to modify the generator itself to obtain the reconstruction of the given image. Figure layout adopted from Xia et al. [2021b].

Alaluf et al. [2021b] find that a complex hierarchical encoder is unnecessary, especially in unstructured domains (e.g., cars, churches, horses) and instead propose simpler backbones. Wei et al. [2021] further replace the 18 map2style blocks with a simple block comprised of a single average pooling layer and fully connected layer. Rather than encoding an image into a set of *style vectors*, Kim et al. [2021] instead invert images into an intermediate latent space with a spatial dimension, resulting in more accurate reconstructions compared to other encoder networks. They also demonstrate that this extended latent space enables reference-guided local edits of real images. More recently, Wang et al. [2021d] explored inverting into multiple latent spaces to achieve higher-fidelity inversions. They first invert an image into $\mathcal{W}$, to capture low-frequency details. A second encoder is then trained to map the distortion map — the difference between the given image and its initial inversion — into a set of spatial feature modulation maps that capture the remaining high-frequency image information.

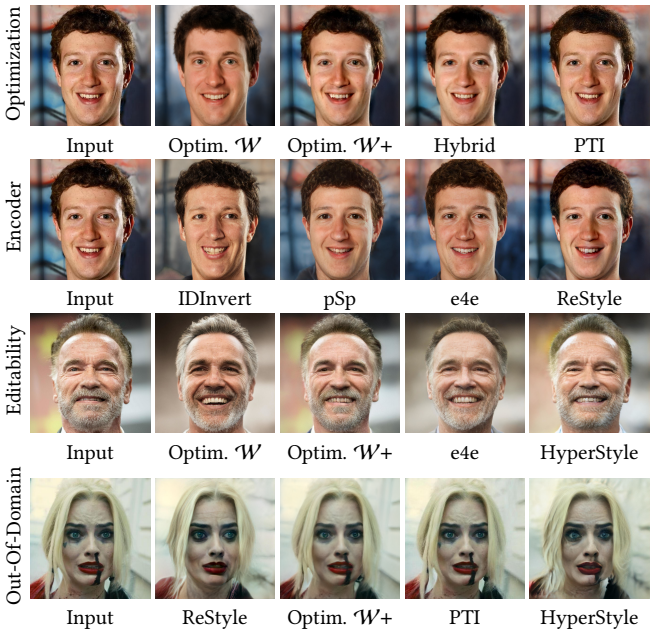


Fig. 9. StyleGAN inversion. Upper row presents inversion results of optimization methods: optimization to $\mathcal{W}$+ as proposed by Karras et al. [2020b], Hybrid approach where pSp [Richardson et al. 2021] and optimization to $\mathcal{W}$+ are employed, and PTI [Roich et al. 2021]. Second raw demonstrates the inversion using encoders, IDInvert [Zhu et al. 2020c], pSp [Richardson et al. 2021], e4e [Tov et al. 2021] and ReStyle [Alaluf et al. 2021b], over the same input image. The third row illustrates the editability of different regions in the latent space. The same smile editing was applied over inversion to $\mathcal{W}$+ space, $\mathcal{W}$ space and well-behaved regions of $\mathcal{W}$+ using the e4e [Tov et al. 2021] encoder. As can be seen, optimization to $\mathcal{W}$+ achieves high-quality reconstruction but poor editability. PTI mitigates this tradeoff by using $\mathcal{W}$ space and tuning the generator weights, but suffer from extensive time consumption. Like PTI, HyperStyle [Alaluf et al. 2021c] uses the $\mathcal{W}$ space for editing, but efficiently learn to modify the generator weights rather than perform time-intensive optimization. Lastly, the ability of PTI and HyperStyle to handle out-of-domain attributes, such as face painting, is presented at the bottom row. Zoom-in is recommended.

Another direction for improving the inversion of encoders is the improvement of the loss objectives used to learn the direct mapping from an image to its latent representation. Zhu et al. [2020c] employ a discriminator for an adversarial-based training of the encoder network and use the discriminator as an additional loss to the encoder. To improve the inversion on the human facial domain, Richardson et al. [2021] introduce a dedicated identity loss using a pre-trained facial recognition network. Tov et al. [2021] extend this to additional domains by employing a similarity loss based on a MoCo [Chen et al. 2020] feature extractor pre-trained on ImageNet. Wei et al. [2021] utilize a pre-trained face parsing network to achieve more localized supervision during the encoder training.

While encoder-based techniques result in an efficient inference scheme, taking a fraction of a second per image, the reconstructions are typically less accurate than optimization-based approaches. In an attempt to close the gap between the two methodologies, Alaluf et al. [2021b] introduce an iterative refinement scheme over standard encoder-based inversion techniques. Instead of directly outputting the inferred latent code using a single forward pass through the network, the encoder outputs a sequence of residuals used to iteratively improve the inverted latent code and corresponding reconstruction. Others [Zhu et al. 2020c, 2016] exploit the advantages of both of the above approaches and employ a hybrid technique combining the two. First, an initial approximate latent code $w_{initial}$ is inferred via a trained encoder. This latent code is then used to initialize the optimization procedure. In [Guan et al. 2020], the encoder network is used to initialize an optimization process, which in turn supervises the training of the encoder network via a set of reconstruction losses. We refer the reader to Figure 9 for a comparison of various optimization-based and encoder-based inversion techniques. Xia et al. [2021b] provide a comprehensive survey and analysis of recent inversion methods, exploring the three aforementioned methodologies and their use in various editing applications.

While the inversion process is a well-studied problem, it remains an open challenge. Numerous works [Abdal et al. 2020a; Tov et al. 2021; Wulff and Torralba 2020; Zhu et al. 2020c,b] demonstrate the existence of a reconstruction-editability trade-off. Whereas $\mathcal{W}$+ has been shown to be more expressive than $\mathcal{W}$ [Abdal et al. 2020a], supporting more accurate reconstructions, its use leads to latent codes which lie in regions of the latent space that were unobserved during the generator training. In these regions, the semantic structure of the latent space deteriorates, resulting in degraded performance of latent space traversal editing methods, as demonstrated in Figure. 9. Some works searched for a good point on this trade-off curve. Tov et al. [2021] design an encoder to embed images into $\mathcal{W}$+ that are close to $\mathcal{W}$, resulting in a good balance between reconstruction quality and editability. Zhu et al. [2020b] analyze various latent spaces to achieve more control over the reconstruction-editability trade-off. In an attempt to side-step this trade-off, Roich et al. [2021] propose a pivotal tuning method to inject new identities into well-behaved, editable regions of StyleGAN's latent space. They first use a standard optimization procedure to find a latent code $w \in \mathcal{W}$ approximating the input image. This is followed by a per-image fine-tuning session where the generator weights are modified to improve the reconstruction quality. Other generator tuning approaches have also been proposed for achieving high fidelity reconstructions (see Section 7).

While most generator tuning approaches improve the image inversion via a per-image optimization of the generator weights, such an approach is costly in terms of inference time. To reduce this inference overhead, Alaluf et al. [2021c] and Dinh et al. [2021] propose a hypernetwork-based encoder that *learns* how to modify the pre-trained generator weights to best reconstruct a given image. Such a learned approach results in high-fidelity reconstructions and edits, at a fraction of the time compared to optimization-based tuning approaches.

Finally, while most works studying inversion focus on encoding and editing still images, when it comes to video editing new challenges arise. Specifically, video inversion should be temporally consistent. Tzaban et al. [2022] demonstrate that by combining encoders [Tov et al. 2021] with generator tuning techniques [Roich et al. 2021], the consistency of the original video can be maintained. Another challenge can be found in the texture-sticking phenomenon observed in StyleGAN1 and StyleGAN2 [Karras et al. 2021],

which hinders the realism of generated and manipulated videos. To overcome this, Alaluf et al. [2022] combine the PTI [Roich et al. 2021] and ReStyle [Alaluf et al. 2021b] encoding techniques for encoding and editing videos with the StyleGAN3 [Karras et al. 2021] generator. Further leveraging the equivariance of StyleGAN3, they demonstrate the ability to expand the field of view when working on a video with a cropped subject resulting in more uniform video editing.

4.2 Latent Space Embedding

Image inversion provides a latent code that reconstructs a given image. As the image itself is given, the produced latent code is usually not of interest on its own. Rather, one applies inversion to then manipulate the latent code to produce a new latent code that corresponds to a novel image.

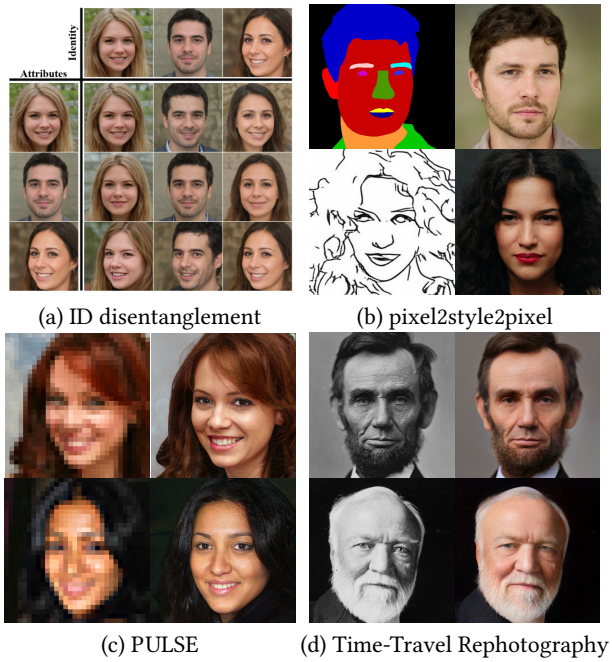
In this light, the limitations of inversion are clear. First, inversion methods require that the input image be invertible. That is, it must reside within one of the latent spaces of StyleGAN. Second, it is assumed that there is a known global transformation in the latent space to produce the desired manipulated code. However, for some applications, at least one of these assumptions is not true. For example, consider commonly studied image-to-image tasks such as semantic map-to-image and sketch-to-image [Isola et al. 2017]. As StyleGAN is trained on one domain, usually the natural image domain, the sketch image would not be invertible to StyleGAN's latent space.

Hence, several works have analyzed this limitation and have proposed a broader task of *Latent Space Embedding*. In this setting, for some image x, one seeks a function f such that $G(f(x)) \sim h(x)$, where G is the pretrained StyleGAN generator and h is some conceptually known function in image space (e.g., sketch-to-image). Under this perspective, inversion is a special case in which h is the identity function. However, many methods have proposed training such function f for specific transformations h.

First, Nitzan et al. [2020] propose using StyleGAN to disentangle identity from other facial attributes and recompose novel images. They do so by extracting identity and attribute representations from different images, combining them, and then training a mapping network to directly produce the latent code that fuses the two representations, resulting in a novel face image, see Figure 10(a).

Next, Richardson et al. [2021] proposed a generic framework, pixel2style2pixel (pSp), to perform a wide variety of image-to-image tasks, such as the aforementioned sketch-to-face and semantic map-to-face. pSp employs an encoder architecture based on a feature pyramid network (FPN), to naturally match the StyleGAN hierarchical generative path. Through the right inductive bias, this work demonstrates state-of-the-art inversion quality, along with various other successful encoding tasks for human faces, including in-painting, super-resolution, unsupervised frontalization, colorization, and more, see Figure 10(b).

Several works used the above concept of latent space embedding for a variety of tasks, often obtaining state-of-the-art performance. Most notably for the task of restoring corrupted images. PULSE [Menon et al. 2020] solves super resolution of facial images. Specifically, PULSE performs a latent-space optimization to recover a code, from which StyleGAN synthesis followed by downsampling reconstructs the original low-resolution input image (see Figure 10(c)). Time-Travel Rephotography [Luo et al. 2021] (Figure 10(d)) restores old photographs, transforming them to modern imagery. They do so by following a similar approach to PULSE, with a dedicated degradation module replacing the down-sampling step. GFPGAN [Wang et al. 2021b] solves blind face restoration by constructing dedicated losses and architecture. GLEAN [Chan et al. 2021b] use an encoder-latent bank-decoder architecture to solve super-resolution tasks. Once more, the decoder is a well-trained StyleGAN generator.



(a) ID disentanglement (b) pixel2style2pixel

(c) PULSE (d) Time-Travel Rephotography

Fig. 10. Examples of prominent works leveraging latent space embedding. (a) Nitzan et al. [2020] disentangle identity from other face attributes and recompose them to generate novel face images. (b) pSp [Richardson et al. 2021] proposed a generic pix2pix-like architecture for embedding into StyleGAN's latent space. (c) PULSE [Menon et al. 2020] perform super resolution by recovering the StyleGAN latent code that after downsampling reconstructs the original image. (d) Time-Travel Rephotography [Luo et al. 2021] restore old photos with a similar approach to PULSE, using a "old photo" degredation module instead of downsampling.

Of the aforementioned tasks, a task receiving considerable attention is that of sketch-to-image due to its immediate application to a variety of real-world settings. Building on the multi-modal sketch-to-image approach from pSp [Richardson et al. 2021], Wei et al. [2021] utilize a specialized face parsing loss to improve the alignment between an input sketch or semantic map and the output realistic facial image. Finally, Wang et al. [2021a] modify a pretrained StyleGAN for transforming a given sketch into a realistic image. As the generator tuning is subtle, the inherent characteristics of the original generator (e.g., color, texture) are well-preserved while supporting multi-modal synthesis.

A myriad of other applications has also been explored via the task of latent space embedding. Chai et al. [2021a] train an encoder for performing image composition and image completion by leveraging the strong image prior of a pre-trained StyleGAN generator. Alaluf et al. [2021a] pair a pSp encoder [Richardson et al. 2021] and pre-trained age regressor [Rothe et al. 2015] for performing age transformation on real images via StyleGAN's latent domain. Jang et al. [2021] transform real facial images to caricatures by altering the specific layers of a pre-trained StyleGAN. Specifically, they leverage the hierarchical nature of StyleGAN and modify the coarse input layers controlling head shape while keeping the fine layers controlling style and color unchanged.

Xu et al. [2021] and Ling et al. [2021] edit a given image in the domain of its part-segmentation. The fundamental observation is that one can train a simple function, f inferring a semantic segmentation, corresponding to the image generated by a latent code, from intermediate activations of StyleGAN on that latent code. Specifically, they use the up-sampled and concatenated per-layer activations as input to f. This observation and construction was concurrently proposed by other works [Tritrong et al. 2021; Zhang et al. 2021c] for different applications and are discussed in Section 6. Given an image, Xu et al. [2021] propose to compute its semantic segmentation using off-the-shelf methods. Then, the segmentation map is edited with some desired effect and finally, it is embedded into the latent space of StyleGAN through StyleGAN's own layers as well as those of the function f. The resulting latent may then be forwarded through StyleGAN to generate the edited image.

Other works [Wei et al. 2021; Yang et al. 2021a; Zhu et al. 2021b] examined the task of face swapping by blending the latent representations of two input images embedded via a learned encoder. Zhu et al. [2021a] study the task of hairstyle transfer. In their work, they decompose an input latent code into a pair of latent codes representing structure and appearance. To transfer a given hairstyle, they blend between several images by taking specific regions of the structure latent codes and combining them with a target appearance. Finally, Chandran et al. [2021] combine traditional and neural synthesis approaches by projecting high-quality skin maps into the latent space of StyleGAN, which is tasked with filling in regions that traditional methods struggle with — such as the eyes, inner mouth, or hair. While all the aforementioned works showed incredible results and promise in real-world scenarios, they are limited in the domains they operate over. Some works have explored going beyond the facial domain and have explored applying StyleGAN for full-body synthesis in various applications such as virtual try-on and portrait reposing [AlBahar et al. 2021; Lewis et al. 2021].

5 EVALUATION METRICS

While many aspects of GAN quality can be evaluated qualitatively, it is often desirable to assess the model quality more objectively. Evaluation metrics can be used to produce reliable, standardized benchmarks and to better gauge the advancement of the field. As we discuss below, this problem is not restricted to StyleGAN editing alone, but to the evaluation of most GANs and editing operations.

GAN Evaluation. The evaluation of generative models is straightforward when ground truth is at hand. For example, GAN inversion can be measured by various metrics assessing the distortion, such as pixel-wise distance using mean-squared error, perceptual similarity using LPIPS [Zhang et al. 2018], structural similarity using MS-SSIM [Wang et al. 2003], or identity similarity [Marriott et al. 2020], employed for facial images using a face recognition network [Deng et al. 2019]. In the absence of such ground truth for the task of unconditional image synthesis, the evaluation of GAN quality remains an open challenge. Undoubtedly, the most popular metric is the Frechet Inception Distance (FID) [Heusel et al. 2017]. FID measures the similarity between two distributions using the Frechet Distance, where each distribution consists of visual features extracted by utilizing a pretrained recognition network [Szegedy et al. 2016]. Namely, given two sets of images, low FID indicates these sets share similar visual statistics. For the case of GANs, the target dataset is compared to the same number of random synthesized images, showing the similarity between these distributions.

Former to FID, Inception Score (IS) [Salimans et al. 2016] was introduced for the same purpose, measuring KL divergence over the same feature statistics. An additional approach has been suggested to measure the distance between real and generated images using the Sliced Wasserstein Distance (SWD) [Rabin et al. 2011], which computes the statistical similarity between local image patches extracted from the Laplacian pyramid of the images. However, as FID is shown to be better correlated with the human perception of high-quality images, it has become the most widely used metric.

Despite its vast popularity, the FID metric does have drawbacks. As the extracted visual features are local, FID struggles to grasp a global structure. For facial images, which bear a simple structure, FID is still effective. Yet, images containing extremely unrealistic structures but high-quality textures, such as a cat with eight legs, can still achieve a good FID score undesirably. Another major concern is the employment of the popular truncation trick [Karras et al. 2019; Marchesi 2017]. Many works generate images using a latent truncation but measure FID without it, as it alters the distribution substantially and leads to a deterioration of FID values [Katzir et al. 2022].

Sajjadi [Sajjadi et al. 2018] proposed a solution to this exact problem by breaking down the GAN evaluation into recall and precision. High precision indicates high quality and realistic image generation, while high recall refers to generating a large amount of variation which is similar in diversity to the original data.

Editing Evaluation. For most practical cases, acquiring ground truth data and labeling to directly evaluate editing is infeasible or altogether impossible. As such, creative solutions have been proposed to tackle the problem of editing quality evaluation. Contrary to disentanglement or GAN quality, the evaluation of StyleGAN's editing ability has not been widely studied. A few key aspects need to be analyzed for the evaluation of these editing procedures. Consider the example of adding a smile to a facial image. The most important aspect is the semantic meaning, namely, whether the editing successfully implants a smile. For binary editing, this could be easily performed using a classifier [Lample et al. 2017; Mokady et al. 2019], but in most cases, continuous editing is required. A regression model can be adopted for this case. However, for many attributes, these models are unavailable or require a vast amount of

annotations to be trained. For example, recent works [Roich et al. 2021; Zhu et al. 2020b] used the Microsoft Face API [Microsoft 2020] to measure face rotation but fail to measure the smile extent continuously. Furthermore, Zhu et al. [2020b] demonstrate that the semantic editing magnitude when employing fixed editing is larger for the more native and editable regions of StyleGAN, and hypothesize that the magnitude could be utilized as an editability metric.

Another key aspect is refraining from distorting the unedited parts of the image, usually referred to as preserving the original identity. For example, smile editing should not result in the appearance of glasses or a change in haircut. Some works [Alaluf et al. 2021a; Nitzan et al. 2020; Richardson et al. 2021; Roich et al. 2021; Tov et al. 2021] focus on facial images, where identity preservation could be evaluated using facial recognition networks [Deng et al. 2019]. Since these networks are trained to be invariant to most attributes, adding a smile should not affect the output substantially. Therefore, an identity similarity can be measured by the cosine similarity of the facial identity representations. Nevertheless, as have been shown by Zhu et al. [2020b], the less editable latent spaces produce lower magnitude edits, leading to a bias in favor of these barely editable regions. Intuitively, the identity is better preserved better when the editing effect is reduced. Consequently, Roich et al. [2021] suggest measuring the identity similarity while performing edits of the same magnitude, e.g. rotation to a predetermined angle. Such metrics have been shown to be more robust, with the identity similarity for the less editable $\mathcal{W}+$ space inversion being inferior compared to the native $\mathcal{W}$ space. Recent works [Alaluf et al. 2021c; Yao et al. 2021] have taken the above procedure a step further, plotting the measured identity similarity along a *range* of editing magnitudes. This results in a continuous curve measuring identity preservation as a function of editing strength. In the context of videos, Tzaban et al. [2022] measured temporal coherence of edited videos, separating the evaluation of local (TL-ID) and global (TG-ID) identity consistency. Locally, they evaluate the identity similarity between pairs of adjacent frames. Globally, they measure similarity between all possible pairs, *i.e.* not necessarily adjacent.

Still, these metrics are mostly limited to facial data, as it is challenging to procure identity recognition networks for other domains such as churches, cars, or cats. To this end, Abdal et al. [2020b] focused on the setting of sequential editing and proposed to measure similarity between results obtained when applying the same semantic directions in a different order. Tov et al. [2021] suggested the latent editing consistency (LEC) metric to evaluate the editing quality realized by a given encoder E. Their method consists of performing latent editing followed by synthesis, encoding, and applying the reverse editing. Optimal editing is expected to result in minimal distortion as the editing procedure should only affect the desired attribute.

One more concern is image quality. One of StyleGAN's key benefits is high visual quality, and editing methods should aim to preserve it. To this end, the common FID metric can be used over the edited images. However, editing might cause a significant bias between the edited and the real data, leading to inaccurate evaluation. If available, a classifier or regression model can be used to balance both image collections with respect to some attribute. A further approach, presented by Zhu et al. [2020b], is to evaluate the interpolation quality. They suggest that good editability should retain the high quality of StyleGAN even for the interpolated images, and utilize the FID metric for this purpose. Lastly, a number of works utilized a user study to evaluate editing quality [Tov et al. 2021; Zhu et al. 2020b] through human judgment. Although this approach carries a profound understanding of the editing procedure, it consumes significant resources and is susceptible to unwanted manipulations. To this day, there is no widely acceptable assessment metric for latent manipulation quality.

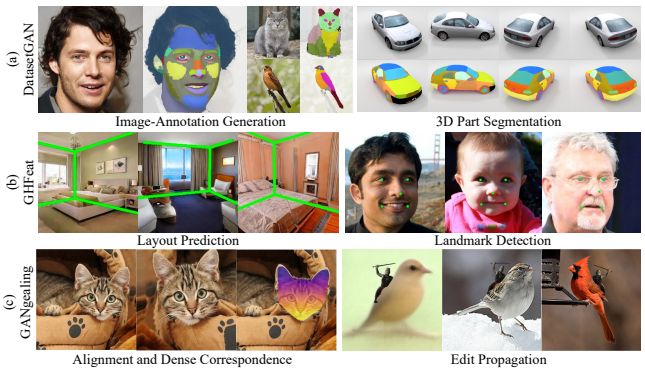


Fig. 11. Examples of discriminative applications built around the StyleGAN generator. These include the ability to synthesize highly detailed segmentation masks (a) [Zhang et al. 2021c], regress facial landmarks or room layouts (b) [Xu et al. 2021b] and even learning alignments and dense correspondences which can be used to propagate edits between different images or video frames (c) [Peebles et al. 2021].

6 DISCRIMINATIVE APPLICATIONS

While the generative capabilities of GANs, and StyleGAN in particular, are indeed groundbreaking, one may ask what *non-generative* tasks can potentially be tackled using GANs. In its most basic form, the GAN's capability to generate a large number of images, all essentially re-sampled from the same target distribution, can be used for data enrichment and augmentations for downstream training tasks. Indeed, many early works proposed using a GAN as an augmentation tool to generate more training data [Antoniou et al. 2017; Tanaka and Aranha 2019; Zhu et al. 2018], possibly also through the use of latent-space editing [Hochberg et al. 2021].

Leveraging the GAN's editing capabilities, Chai et al. [2021b] propose an ensembling method for image classification, by augmenting the input image at test-time. The input is projected into the pre-trained generator's latent space, and editing operations such as style mixing are applied to it, generating different views. The generated images are then fed into a classification network, and the final prediction of the model is based on an ensemble of the network predictions on all of the images. Unlike conventional ensembles in deep learning, where predictions of several models are combined to yield the final result, this method proposes using different views of the same image (while preserving its identity) and ensemble the classifier predictions on the images at test-time.

Aiming to leverage the semantic understanding of StyleGAN in new ways, Peebles et al. [2021] present a novel framework to approach the task of dense visual alignment. Using sampled latent codes and their corresponding images, the authors jointly learn a latent representation used for latent-space editing, and a Spatial Transformer Network to align the generated images according to the edit, as illustrated in Figure 11c. Once both manipulations converge to a single viewpoint, the authors can employ the STN to align real images. Abdal et al. [2021b] present an unsupervised segmentation method based on a pre-trained GAN. The authors recognize that nullifying some of the activations actually causes the GAN to erase the foreground object, producing an image with only a background. Hence, they form two networks, one that generates only the background for an image, and another which generates only the foreground, naturally leveraging the GAN's internal semantic understanding. They then use the two networks to train a segmentation mask generation network in an unsupervised manner. This notion, of extracting a segmentation map with the help of StyleGAN's structure, has been employed similarly by others [Lewis et al. 2021; Li et al. 2021b; Ling et al. 2021; Zhang et al. 2021c].

However, StyleGAN's well-behaved latent space offers more opportunities. In a fine example of taking full leverage of the latent space, Xu et al. [2021b] show how to exploit pre-trained generative models for a wide variety of analysis and generation tasks. In essence, they propose an adversarial feature learning technique [Donahue et al. 2017]; they extract meaningful feature maps through StyleGAN inversion and use them for a wide variety of tasks. The authors show that the channel-level modulations performed by a style code can be used as descriptive features for downstream tasks. Hence, by simply training an encoder, high-quality feature vectors can be produced in an unsupervised manner. The authors evaluate the quality of the features on different generative and discriminate downstream tasks, including image editing, image recognition, landmark detection, and more (see Figure 11b). Nitzan et al. [2021] further observe that the linear nature of semantic directions in StyleGAN's latent space, *i.e.*the same linearity exploited by the editing literature (see Section 3), can be leveraged as a tool for few-shot regression. Their basic premise is that if the space is indeed linear, then given two labeled points along a disentangled axis, any interpolated point between them should produce an interpolation of their labels as well. In other words, they show that linear editing directions are not only global, in the sense that they cause a similar effect for all inputs, but also that the magnitude of these effects is linear in the size of the traversal step. Through this realization, they achieve state-of-the-art few-shot regression performance on various properties, such as yaw angle and age estimation for human faces.

Continuing this line of thought, several papers leverage StyleGAN's intermediate representation to perform semantic segmentation. As previously discussed (see Section 4.2), a simple function, f, may be learned between the up-sampled concatenated per-layer features and a semantic segmentation of the image. As illustrated in Figure 11a, Zhang et al. [2021c] propose to use StyleGAN together with f to generate a virtually infinite paired synthetic training set for semantic segmentation. Alternatively, Tritong et al. [2021] directly use StyleGAN and f for segmentation by first inverting a real image into latent space. In the context of local editing, Collins et al. [2020] and Kafri et al. [2021] perform a simple clustering procedure over StyleGAN's internal representations to obtain a semantic segmentation of an input image. This semantic map can then be used to perform local editing over an image, guided by a target reference image. Lang et al. [2021] propose to not only exploit the emerging disentanglement properties of a pretrained StyleGAN, but to train a StyleGAN model for a specific disentangled axis. Through a clever training scheme, combining training StyleGAN along with a classifier for binary or multi-class recognition (e.g., a cat vs. dog classifier), they drive the latent space to capture classifier-specific attributes. As they demonstrate, through this joint training process, the linear editing directions that emerge from this model correspond to specific properties that the classifier searches for. For example, in the cat vs. dog case, the emerging editing directions include the shape of the eye, and the pointedness of the ears. This means that the image can be augmented to be more or less suitable for a specific label (e.g., a cat could be turned to be more dog-like), thus providing explainable examples of how the classifier makes its decision.

As can be seen, StyleGAN's unsupervised arrangement of its latent space in disentangled directions is an exciting property that could potentially be leveraged for various applications. In the near future, it is likely that more works along these directions would be introduced, maybe establishing GANs as a method useful for many downstream tasks in machine learning in general, reaching beyond data augmentation or entertainment. Furthermore, it stands to reason that future generations of GANs may be designed with more consideration to discriminative tasks.

7 FINE-TUNING THE GENERATOR

7.1 Data Reduction

Training a StyleGAN model requires substantial amounts of data, confined to quite a small domain. This means that the amount of available data serves as a strong bottleneck for adapting StyleGAN to new domains. An established way to address the lack of data is through augmentation. To stabilize training in limited data scenarios, several methods [Karras et al. 2020a; Zhao et al. 2020a,b] used differential augmentations during the training process. In contrast to classification tasks, generative augmentations pose a challenge — if the discriminator observes sufficient augmented samples, the generator might produce such augmented results by itself. In many cases, the augmentations leakage to the generated images is highly undesirable, *e.g.*when the augmentations contain rotations or unrealistic colorization changes. By monitoring overfitting indications during training and adaptively increasing augmentation strength, Karras et al. [2020a] are able to introduce additional supervision to the network through augmentations, without allowing them to leak into the generated results. They achieved state-of-the-art results in low data domains, significantly reducing the number of training samples required for training. Sinha et al. [2020] also suggest explicitly providing the discriminator with negative out-of-distribution samples to bias the generator away from unwanted samples. Another direction, proposed by Yang et al. [2021b] is to empower the discriminator to better extract knowledge from the training set by providing it with an auxiliary task in the form of instance-discrimination via a contrastive learning objective. Kumari et al. [2021] propose to

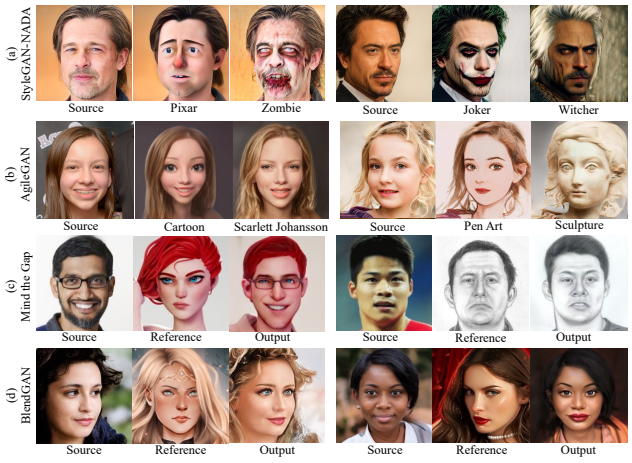


Fig. 12. Many works have approached the task of transforming images from one domain (e.g., real faces) to other, semantically similar domains (e.g., cartoons). Typically, this has been done either through guidance via a textual description of the desired target domain (a) [Gal et al. 2021b], a short fine-tuning approach trained on a handful of images (b) [Song et al. 2021], or single-shot adaptation approach with style introduced via a desired reference image (c,d) [Liu et al. 2021a; Zhu et al. 2020a].

leverage the feature space of pre-trained vision models, trained for different vision tasks. By progressively selecting and employing the models as additional discriminators, they manage to improve synthesis quality in both limited-data and large-scale settings.

Another established and popular approach is **domain adaptation**, where different works seek to convert pre-trained StyleGAN models into other semantically similar domains using few data exemplars. Aside from reducing the amount of data needed to train a model from scratch, Karras et al. [2020a] fine-tuned a StyleGAN generator trained over the FFHQ dataset into the domain of MetFaces, a collection of images from the Metropolitan Museum of Art, using only 1,336 training samples. Following the above, Pinkney et al. [2020] leverage the disentangled control over the coarse, medium, and fine semantic attributes StyleGAN offers and perform *domain-mixing* for depicting the geometry of one domain with the textural appearance of another. The authors fine-tune a well-trained StyleGAN for human faces using only 300 cartoon examples. They then propose blending the models, replacing high-resolution layers of the fine-tuned model with the pre-trained layers of the source domain. This yields a toonification effect, producing an output domain of cartoons, with the complex variety and structure of human faces. Several works make use of a similar form of fine-tuning. Song et al. [2021] introduce several synthesis paths within the generator for different attributes, achieving high-quality portrait stylization. Following the fine-tuning process, Jang et al. [2021] leverage the semantic correspondence between the two models. By feeding the same latent codes to both generators, an extensive paired training data set can be generated. This data is then used to train a translation network between the domains (in this case, a shape-exaggeration network operating over the caricatures domain).

Within the **few-shot** settings, several works perform domain adaptation based on as few as 10 training exemplars. When the translation is done between semantically similar domains, the methods manage to preserve the semantic properties and diversity of the source domain. Li et al. [2020] maintain diversity by applying an elastic weight consolidation loss, regularizing weights change based on Fisher information, computed from a discriminator. To facilitate few-shots adaptation, Ojha et al. [2021] propose explicitly maintaining the source domain structure, through a distance consistency loss between pairs of resulting images. In addition, a shared image and patch discriminator is applied to create patch-level adversarial similarity throughout the latent space together with image-level similarity around chosen anchors.

Taking domains with data shortage to the extreme, Gal et al. [2021b] perform zero-shot domain adaptation, fine-tuning StyleGAN without providing any exemplar images. They propose describing the desired target domain in text and using a pre-trained linguistic-visual model [Radford et al. 2021] to guide the adaptation. The GAN is trained such that the CLIP-space direction between the resulting images before and after the tuning process aligns with the CLIP-space direction between a pair of source and target description texts. Zhu et al. [2020a] perform single-shot domain adaptation by matching the reference image in the target domain with a corresponding synthesized image from the source domain, obtained through latent space optimization. Using the reference pair, in addition to the CLIP-space loss defined in StyleGAN-NADA, they introduce a new objective, which is to maintain CLIP-space direction similarity between each reference image and the current training iteration's synthesized image. Using this method, the authors manage to achieve better adaptation of pose, lighting, and expression through the domain transfer process. Taking a different approach, Yang et al. [2021c] freeze the generator weights and learn a linear transformation in $\mathcal{Z}$ space, using as little as a single reference image. We refer the reader to Figure 12 for sample domain adaptation results.

7.2 Data-Aware Generator Tuning

Arguably, the most promising direction for StyleGAN development is through data-aware model manipulation. A pre-trained StyleGAN model is phenomenal in local structure and disentanglement (see Section 2), but is relatively confined in generality. As such, it stands to reason that there could be significant benefit in adapting the model to include new, specific data points. Then, if the GAN's structure is maintained, these new points could be better processed for editing or discriminative applications. This concept has been suggested in the past already [Bojanowski et al. 2018]. For example, in the context of GANs, Bau et al. [2019] propose adapting the image prior learned by Progressive GAN [Karras et al. 2018] to image statistics of an individual image. Through minimal fine-tuning, the authors were able to faithfully reconstruct a given image, and present editing capabilities of quality unseen at the time, including synthesizing new objects seamlessly, removing unwanted objects, and changing object appearance. Similarly, Pan et al. [2020] use BiGGAN [Brock et al. 2018] to capture high-level semantic image priors such as color, texture, spatial coherence, etc., for tasks such as colorization, inpainting, morphing, and category transfer.

This is contrary to the traditional approach [Ulyanov et al. 2018], where only low-level priors are captured. Their method is based on GAN-inversion, but they overcome the difficulty of GAN-inversion methods to generate out-of-domain images by allowing the generator to be fine-tuned on the fly when searching for the latent source. They regularize the generator fine-tuning with feature matching loss from the discriminator, and use progressive fine-tuning (from shallow layers to deep).

In the context of StyleGAN, Roich et al. [2021] take a similar approach. Given a real-world image that is similar, but not included in the domain of a pretrained StyleGAN (e.g., an image of a human face, with very untypical facial features, makeup, or hairstyle), they propose finding the closest latent code to the desired appearance (termed the 'pivot'), and fine-tuning the generator so the exact appearance would be reconstructed with this code. In addition, they ensure the process does not impair the disentangled latent space through regularization. This simple approach produces significantly better reconstructions (see Fig. 9), and allows employing off-the-shelf editing techniques with high quality, essentially bypassing the notorious distortion-editability trade-off (see Section 4). Such fine-tuning sessions are typically brief, lasting an order of a single minute. As described in Section 4.1, this generator tuning can alternatively be performed as a forward pass procedure using hypernetworks [Alaluf et al. 2021c; Dinh et al. 2021]. Tzaban et al. [2022] further improve the tuning scheme to semantically edit a video while preserving temporal coherence. First, they observe that using an encoder for the initial inversion allows for a temporally-smooth edit after tuning the generator. Second, they propose to further tune the generator to better stitch the edited cropped face back to the original frame.

Bau et al. [2020] perform a similar tuning operation, but where the awareness is to the task instead of the data. The authors propose changing semantic and physical properties (or *rules*) of deep generative networks, relying on the concept of linear associative memory. While current methods for image editing allow users to manipulate single images, this method allows changing semantic rules and properties of the network, so that all images generated by the network have the desired property. This includes removing undesired patterns such as watermarks and adding objects such as human crowds or trees. Kwong et al. [2021] outline a method for cross-domain editing by inverting images into a source domain and re-synthesizing them in a fine-tuned model using the same code. Cherepkov et al. [2021] expand the range achievable by existing state-of-the-art generative models used for image editing and manipulation, such as StyleGAN2. While existing methods find interpretable directions in the model's latent space and operate on latent codes, they find interpretable directions in the space of generator parameters and use them to manipulate images and expand the range of possible visual effects. They show that their discovered manipulations, such as changing car wheel size, cannot be achieved by manipulating the latent code. Finally, Liu et al. [2022] demonstrate that brief fine-tuning sessions can be used to condition a model on labels derived from the latent-space itself, thereby "baking" editing directions into the GAN and improving treatment of rare data modalities, such as extreme poses or underrepresented ethnicities.

8 CONCLUSIONS

StyleGAN has revolutionized the field of image synthesis, bringing with it consistent, high-quality results with exceptional photo-realism across multiple domains. More interestingly, through a combination of layer-wise style modulations and a novel mapping network, StyleGAN is capable of mapping out a smooth, semantic, and highly-disentangled latent space in an entirely unsupervised manner. This enables latent-based editing, yielding effects such as photo-realistic and plausible alterations to age, hairstyles, or body poses, and even transformations into celebrities or magical beings.

However, StyleGAN struggles with domains that do not exhibit strong structure. An in-depth look over the diverse set of works covered by this report will reveal that they all demonstrate their abilities on a rather limited collection of domains, and with no regard to the temporal axis. To address these limitations, there is yet much development that the generative field must undergo.

While many works focused on re-using a pre-trained generator for downstream tasks, a recent trend has shown that some of these domain-related limitations can be overcome if one adapts the generator to their specific needs. In essence, such approaches build upon the extensive knowledge that StyleGAN can glean from a rich source domain, and transfer it to new realms such as 3D rendering, paintings, or wildlife.

Another noteworthy direction resides in extracting knowledge from StyleGAN for non-generative needs. StyleGAN has already been leveraged for regression, segmentation, and explainability, but there is doubtless more that could be learned from exploring its structured latent space. On the quest to self-supervision and learning representations that naturally disentangle and understand the elements comprising data distributions, StyleGAN is an important milestone.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In *Proceedings of the IEEE international conference on computer vision*. 4432–4441.

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020a. Image2StyleGAN++: How to Edit the Embedded Images?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8296–8305.

Rameen Abdal, Peihao Zhu, John Femiani, Niloy J Mitra, and Peter Wonka. 2021a. CLIP2StyleGAN: Unsupervised Extraction of StyleGAN Edit Directions. *arXiv preprint arXiv:2112.05219* (2021).

Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. 2020b. StyleFlow: Attribute-conditioned Exploration of StyleGAN-Generated Images using Conditional Continuous Normalizing Flows. arXiv:2008.02401 [cs.CV]

Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. 2021b. Labels4free: Unsupervised segmentation using stylegan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13970–13979.

Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. 2021. HistoGAN: Controlling Colors of GAN-Generated and Real Images via Color Histograms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. 2021a. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–12.

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. 2021b. ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. 2022. Third Time's the Charm? Image and Video Editing with StyleGAN3. arXiv:2201.13433 [cs.CV]

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H Bermano. 2021c. HyperStyle: StyleGAN Inversion with HyperNetworks for Real Image Editing. *arXiv preprint arXiv:2111.15666* (2021).

Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. 2021. Pose with Style: Detail-Preserving Pose-Guided Image Synthesis with Conditional StyleGAN. *ACM Transactions on Graphics* (2021).

Yazeed Alharbi and Peter Wonka. 2020. Disentangled Image Generation Through Structured Noise Injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov. 2021. Image generators with conditionally-independent pixel synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14278–14287.

Antreas Antoniou, Amos Storkey, and Harrison Edwards. 2017. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340* (2017).

David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. 2021. Paint by Word. arXiv:2103.10951 [cs.CV]

David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. 2020. Rewriting a deep generative model. In *European Conference on Computer Vision*. Springer, 351–369.

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. 2019. Semantic Photo Manipulation with a Generative Image Prior. *ACM Trans. Graph.* 38, 4, Article 59 (jul 2019), 11 pages. https://doi.org/10.1145/3306346.3323023

Piotr Bojanowski, Armand Joulin, David Lopez-Pas, and Arthur Szlam. 2018. Optimizing the Latent Space of Generative Networks. In *International Conference on Machine Learning*. PMLR, 600–609.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.

Arantxa Casanova, Marlène Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. 2021. Instance-conditioned gan. *Advances in Neural Information Processing Systems* 34 (2021).

Lucy Chai, Jonas Wulff, and Phillip Isola. 2021a. Using latent space regression to analyze and leverage compositionality in GANs.. In *International Conference on Learning Representations*.

Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, and Richard Zhang. 2021b. Ensembling with Deep Generative Views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14997–15007.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2021a. Efficient Geometry-aware 3D Generative Adversarial Networks. In *arXiv*.

Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. 2021b. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14245–14254.

Prashanth Chandran, Sebastian Winberg, Gaspard Zoss, Jérémy Riviere, Markus Gross, Paulo Gotardo, and Derek Bradley. 2021. Rendering with style: combining traditional and neural approaches for high-quality face rendering. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–14.

Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. 2021. Image-based clip-guided essence transfer. *arXiv preprint arXiv:2110.12427* (2021).

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved Baselines with Momentum Contrastive Learning. arXiv:2003.04297 [cs.CV]

Anton Cherepkov, Andrey Voynov, and Artem Babenko. 2021. Navigating the gan parameter space for semantic image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3671–3680.

Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. 2020. Editing in Style: Uncovering the Local Semantics of GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5771–5780.

Antonia Creswell and Anil Anthony Bharath. 2018. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems* 30, 7 (2018), 1967–1974.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. 2021. HyperInverter: Improving StyleGAN Inversion via Hypernetwork. *arXiv preprint arXiv:2112.00719* (2021).

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. 2017. Adversarial feature learning. In *International Conference on Learning Representations*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations* (2021).

Rinon Gal, Dana Cohen Hochberg, Amit Bermano, and Daniel Cohen-Or. 2021a. SWAGAN: A Style-Based Wavelet-Driven Generative Model. *ACM Trans. Graph.* 40, 4, Article 134 (July 2021), 11 pages. https://doi.org/10.1145/3450626.3459836

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021b. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. arXiv:2108.00946 [cs.CV]

Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. 2020. Gif: Generative interpretable faces. In *2020 International Conference on 3D Vision (3DV)*. IEEE, 868–878.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14)*. MIT Press, Cambridge, MA, USA, 2672–2680.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2022. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *International Conference on Learning Representations* (2022).

Jinjin Gu, Yujun Shen, and Bolei Zhou. 2020. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3012–3021.

Shanyan Guan, Ying Tai, Bingbing Ni, Feida Zhu, Feiyue Huang, and Xiaokang Yang. 2020. Collaborative Learning for Faster StyleGAN Embedding. *arXiv preprint arXiv:2007.01758* (2020).

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*. 6626–6637. arXiv:1706.08500 [cs.LG]

Dana Cohen Hochberg, Raja Giryes, and Hayit Greenspan. 2021. Style encoding for class-specific image generation. In *Medical Imaging 2021: Image Processing*, Vol. 11596. International Society for Optics and Photonics, 1159631.

Xianxu Hou, Xiaokang Zhang, Hanbang Liang, Linlin Shen, Zhihui Lai, and Jun Wan. 2022. Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing. *Neural Networks* 145 (2022), 209–220.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.

Drew A Hudson and Larry Zitnick. 2021. Generative adversarial transformers. In *International Conference on Machine Learning*. PMLR, 4487–4499.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: Discovering Interpretable GAN Controls. In *Proc. NeurIPS*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

Ali Jahanian, Lucy Chai, and Phillip Isola. 2019. On the" steerability" of generative adversarial networks. In *International Conference on Learning Representations*.

Wonjong Jang, Gwangjin Ju, Yucheol Jung, Jiaolong Yang, Xin Tong, and Seungyong Lee. 2021. StyleCariGAN: caricature generation via StyleGAN feature map modulation. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–16.

Omer Kafri, Or Patashnik, Yuval Alaluf, and Daniel Cohen-Or. 2021. StyleFusion: A Generative Model for Disentangling Spatial Segments. arXiv:2107.07437 [cs.CV]

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020a. Training Generative Adversarial Networks with Limited Data. In *Proc. NeurIPS*.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* 34 (2021).

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020b. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

Oren Katzir, Vicky Perepelook, Dani Lischinski, and Daniel Cohen-Or. 2022. Multi-level Latent Space Structuring for Generative Control. arXiv:2202.05910 [cs.CV]

Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. 2021. Exploiting Spatial Dimensions of Latent in GAN for Real-time Image Editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2021. Ensembling Off-the-shelf Models for GAN Training. *arXiv preprint arXiv:2112.09130* (2021).

Gihyun Kwon and Jong Chul Ye. 2021. Diagonal attention and style-based GAN for content-style disentanglement in image generation and translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13980–13989.

Sam Kwong, Jialu Huang, and Jing Liao. 2021. Unsupervised Image-to-Image Translation via Pre-trained StyleGAN2 Network. *IEEE Transactions on Multimedia* (2021).

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DENOYER, et al. 2017. Fader Networks: Manipulating Images by Sliding Attributes. In *Advances in Neural Information Processing Systems*. 5963–5972.

Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. 2021. Explaining in Style: Training a GAN to explain a classifier in StyleSpace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 693–702.

Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. 2022. Vitgan: Training gans with vision transformers. *International Conference on Learning Representations* (2022).

Thomas Leimkühler and George Drettakis. 2021. FreeStyleGAN: Free-View Editable Portrait Rendering with the Camera Manifold. *ACM Trans. Graph.* 40, 6, Article 224 (dec 2021), 15 pages. https://doi.org/10.1145/3478513.3480538

Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. 2021. TryOnGAN: body-aware try-on via layered interpolation. *ACM Trans. Graph.* 40 (2021), 115:1–115:10.

Bingchuan Li, Shaofei Cai, Wei Liu, Peng Zhang, Miao Hua, Qian He, and Zili Yi. 2021a. DyStyle: Dynamic Neural Network for Multi-Attribute-Conditioned Style Editing. arXiv:2109.10737 [cs.CV]

Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. 2021b. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8300–8311.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.

Yijun Li, Richard Zhang, Jingwan Cynthia Lu, and Eli Shechtman. 2020. Few-shot Image Generation with Elastic Weight Consolidation. *Advances in Neural Information Processing Systems* 33 (2020), 15885–15896.

Hanbang Liang, Xianxu Hou, and Linlin Shen. 2021. SSFlow: Style-Guided Neural Spline Flows for Face Image Manipulation. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 79–87. https://doi.org/10.1145/3474085.3475454

Ji Lin, Richard Zhang, Frieder Ganz, Song Han, and Jun-Yan Zhu. 2021. Anycost gans for interactive image synthesis and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14986–14996.

Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. 2021. EditGAN: High-Precision Semantic Image Editing. *Advances in Neural Information Processing Systems* 34 (2021).

Zachary C Lipton and Subarna Tripathi. 2017. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782* (2017).

Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. 2021a. BlendGAN: Implicitly GAN Blending for Arbitrary Stylized Face Generation. *Advances in Neural Information Processing Systems* 34 (2021).

Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. 2020b. Diverse image generation via self-conditioned gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14286–14295.

Yunzhe Liu, Rinon Gal, Amit H. Bermano, Baoquan Chen, and Daniel Cohen-Or. 2022. Self-Conditioned Generative Adversarial Networks for Image Editing. arXiv:2202.04040 [cs.CV]

Yunfan Liu, Qi Li, Zhenan Sun, and Tieniu Tan. 2020a. Style Intervention: How to Achieve Spatial Disentanglement with Style-based Generators? arXiv:2011.09699 [cs.CV]

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

Mario Lučić, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, and Sylvain Gelly. 2019. High-fidelity image generation with fewer labels. In *International conference on machine learning*. PMLR, 4183–4192.

Junyu Luo, Yong Xu, Chenwei Tang, and Jiancheng Lv. 2017. Learning inverse mapping by autoencoder based generative adversarial nets. In *International Conference on Neural Information Processing*. Springer, 207–216.

Xuan Luo, Xuaner Zhang, Paul Yoo, Ricardo Martin-Brualla, Jason Lawrence, and Steven M. Seitz. 2021. Time-Travel Rephotography. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2021)* 40, 6, Article 213 (12 2021). https://doi.org/10.1145/3478513.3480485

Marco Marchesi. 2017. Megapixel size image creation using generative adversarial networks. *arXiv preprint arXiv:1706.00082* (2017).

Richard T Marriott, Safa Madiouni, Sami Romdhani, Stéphane Gentric, and Liming Chen. 2020. An assessment of GANs for identity-related applications. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 1–10.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2437–2445.

Microsoft. 2020. Azure face.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*. Springer, 405–421.

Ron Mokady, Sagie Benaim, Lior Wolf, and Amit Bermano. 2019. Masked Based Unsupervised Content Transfer. In *International Conference on Learning Representations*.

Yotam Nitzan, A Bermano, Yangyan Li, and D Cohen-Or. 2020. Face identity disentanglement via latent space mapping. *ACM Transactions on Graphics (TOG)* 39 (2020), 1–14.

Yotam Nitzan, Rinon Gal, Ofir Brenner, and Daniel Cohen-Or. 2021. LARGE: Latent-Based Regression through GAN Semantics. *arXiv preprint arXiv:2107.11186* (2021).

Utkarsh Ojha, Yijun Li, Cynthia Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. 2021. Few-shot Image Generation via Cross-domain Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2021. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. arXiv:2112.11427 [cs.CV]

Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. 2020. Exploiting deep generative prior for versatile image restoration and manipulation. In *European Conference on Computer Vision*. Springer, 262–277.

Jeeseung Park and Younggeun Kim. 2021. Styleformer: Transformer based Generative Adversarial Networks with Style Vector. *arXiv preprint arXiv:2106.07023* (2021).

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 165–174.

Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. 2020. Swapping Autoencoder for Deep Image Manipulation. In *Advances in Neural Information Processing Systems*.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.

William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei Efros, and Eli Shechtman. 2021. GAN-Supervised Dense Visual Alignment. arXiv:2112.05143 [cs.CV]

Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Álvarez. 2016. Invertible Conditional GANs for image editing. In *NIPS Workshop on Adversarial Training*.

Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. 2020. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14104–14113.

Justin N M Pinkney and Doron Adler. 2020. Resolution Dependant GAN Interpolation for Controllable Image Synthesis Between Domains. *arXiv preprint arXiv:2010.05334* (2020).

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. 2011. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, 435–446.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2287–2296.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2021. Pivotal Tuning for Latent-based Editing of Real Images. *arXiv preprint arXiv:2106.05744* (2021).

Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. DEX: Deep EXpectation of Apparent Age from a Single Image. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. 252–257. https://doi.org/10.1109/ICCVW.2015.41

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. 2018. Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems* 31.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016), 2234–2242.

Axel Sauer, Katja Schwarz, and Andreas Geiger. 2022. StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets. arXiv:2202.00273 [cs.LG]

Omry Sendik, Dani Lischinski, and Daniel Cohen-Or. 2020. Unsupervised K-modal Styled Content Generation. *ACM Transactions on Graphics (TOG)* (2020).

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020a. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9243–9252.

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020b. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence* (2020).

Yujun Shen and Bolei Zhou. 2021. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1532–1540.

Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. 2020. Negative Data Augmentation. In *International Conference on Learning Representations*.

Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. 2021. AgileGAN: stylizing portraits by inversion-consistent transfer learning. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.

Nurit Spingarn, Ron Banner, and Tomer Michaeli. 2020. GAN" Steerability" without optimization. In *International Conference on Learning Representations*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. 2019. Data augmentation using GANs. *arXiv preprint arXiv:1904.09135* (2019).

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2020. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6142–6151.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an Encoder for StyleGAN Image Manipulation. *ACM Trans. Graph.* 40, 4, Article 133 (jul 2021), 14 pages. https://doi.org/10.1145/3450626.3459838

Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. 2021. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4475–4485.

Rotem Tzaban, Ron Mokady, Rinon Gal, Amit H Bermano, and Daniel Cohen-Or. 2022. Stitch it in Time: GAN-Based Facial Editing of Real Videos. *arXiv preprint arXiv:2201.08361* (2022).

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9446–9454.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

Andrey Voynov and Artem Babenko. 2020. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*. PMLR, 9786–9796.

Hui-Po Wang, Ning Yu, and Mario Fritz. 2021c. Hijack-gan: Unintended-use of pretrained, black-box gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7872–7881.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768* (2020).

Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. 2021a. Sketch Your Own GAN. In *Proceedings of the IEEE International Conference on Computer Vision*.

Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. 2021d. High-fidelity gan inversion for image attribute editing. *arXiv preprint arXiv:2109.06590* (2021).

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021b. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. Ieee, 1398–1402.

Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Lu Yuan, Gang Hua, and Nenghai Yu. 2021. A Simple Baseline for StyleGAN Inversion. arXiv:2104.07661 [cs.CV]

Zongze Wu, Dani Lischinski, and Eli Shechtman. 2020. StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12863–12872. arXiv:2011.12799 [cs.CV]

Jonas Wulff and Antonio Torralba. 2020. Improving Inversion and Generation Diversity in StyleGAN using a Gaussianized Latent Space. arXiv:2009.06529 [cs.CV]

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021a. TediGAN: Text-Guided Diverse Face Image Generation and Manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2021b. GAN Inversion: A Survey. arXiv:2101.05278 [cs.CV]

Jianjin Xu and Changxi Zheng. 2021. Linear Semantics in Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9351–9360.

Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 2021a. 3D-aware Image Synthesis via Learning Structural and Textural Representations. (2021).

Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. 2021b. Generative hierarchical features from synthesizing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4432–4442.

Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. 2021b. Data-efficient instance generation from instance discrimination. *Advances in Neural Information Processing Systems* 34 (2021).

Ceyuan Yang, Yujun Shen, Zhiyi Zhang, Yinghao Xu, Jiapeng Zhu, Zhirong Wu, and Bolei Zhou. 2021c. One-Shot Generative Domain Adaptation. arXiv:2111.09876 [cs.CV]

Ceyuan Yang, Yujun Shen, and Bolei Zhou. 2020a. Semantic Hierarchy Emerges in Deep Generative Representations for Scene Synthesis. *International Journal of Computer Vision* (2020).

Lingchen Yang, Zefeng Shi, Yiqian Wu, Xiang Li, Kun Zhou, Hongbo Fu, and Youyi Zheng. 2020b. iOrthoPredictor: model-guided deep prediction of teeth alignment. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–15.

Shuai Yang, Kai Qiao, Ruoxi Qin, Pengfei Xie, Shuhao Shi, Ningning Liang, Linyuan Wang, Jian Chen, Guoen Hu, and Bin Yan. 2021a. ShapeEditor: A StyleGAN Encoder for Stable and High Fidelity Face Swapping.

Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. 2021. A latent transformer for disentangled face editing in images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13789–13798.

Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. 2017. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5485–5493.

Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. 2021b. StyleSwin: Transformer-based GAN for High-resolution Image Generation. arXiv:2112.10762 [cs.CV]

Richard Zhang. 2019. Making convolutional networks shift-invariant again. In *International conference on machine learning*. PMLR, 7324–7334.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. 2021a. Image GANs meet Differentiable Rendering for Inverse Graphics and Interpretable 3D Neural Rendering. In *International Conference on Learning Representations*.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. 2021c. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10145–10155.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. 2020a. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems* 33 (2020), 7559–7570.

Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. 2020b. Image Augmentations for GAN Training. arXiv:2006.02595 [cs.LG]

Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. 2021. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. *arXiv preprint arXiv:2110.09788* (2021).

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020c. In-domain GAN Inversion for Real Image Editing. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. 2016. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*. Springer, 597–613.

Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. 2020a. Mind the Gap: Domain Gap Control for Single Shot Domain Adaptation for Generative Adversarial Networks. In *International Conference on Learning Representations*.

Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. 2021a. Barbershop: GAN-Based Image Compositing Using Segmentation Masks. *ACM Trans. Graph.* 40, 6, Article 215 (dec 2021), 13 pages. https://doi.org/10.1145/3478513.3480537

Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020b. Improved StyleGAN Embedding: Where are the Good Latents? arXiv:2012.09036 [cs.CV]

Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. 2018. Emotion classification with data augmentation using generative adversarial networks. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 349–360.

Yuhao Zhu, Qi Li, Jian Wang, Chengzhong Xu, and Zhenan Sun. 2021b. One Shot Face Swapping on Megapixels. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 4834–4844.