Pattern Recognition Letters 202 (2026) 114–119

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Why retrieve when you can edit: A fast conditional StyleGAN latent editing method

Andrei Radu [a,b,*], Yue Song [b], Ana Neacsu [a], Nicu Sebe [b]

[a] Signal Modelling and Analysis Laboratory, National University of Science and Technology "Politehnica" of Bucharest, Iuliu Maniu Blvd., 6D, Bucharest, 061103, Romania
[b] Department of Information Engineering and Computer Science, University of Trento (UniTN), via Sommarive, 9, Trento, 38123, Italy

ARTICLE INFO

Keywords:
Image generation
Image editing
Similar image retrieval
Identity preservation
StyleGAN

ABSTRACT

Text-to-image diffusion models represent the de facto tools for image editing, but come with the disadvantage of a time-consuming multi-step approach, while also having a considerably large number of parameters. Recently, various methods have been proposed to increase the speed of the editing process, most focusing on searching the latent space of Generative Adversarial Networks (GANs) for semantically meaningful directions and synthesising the desired features from there. However, this task often requires extensive training to extract meaningful editing directions. As such, we propose a new training paradigm, related to that of a teacher-student technique, which leverages the remarkable conditional generation performances of StyleGAN for image attribute insertion via text conditioning. Our method computes the required changes of the latent style space by morphing the textual embeddings with the style space inside a Transformer architecture. We studied the editing capabilities of our approach on three benchmark datasets to demonstrate the intrinsic information acquired during the training of a conditional StyleGAN (teacher) and the transfer efficiency of information to the student network, outperforming other SOTA methods while requiring fewer resources.

1. Introduction

Image attribute editing refers to the process of altering particular characteristics of an image while maintaining the overall structure, preserving the existing features, and retaining the realism of the original image. This field has experienced substantial advancements in two distinct branches. The first branch involves the use of Generative Adversarial Networks (GANs) [1–3] to harness rich features inside their latent space, which can be more easily interpreted, while the second applies an iterative denoising process to generate the final image. As a result, many researchers are focused on exploring novel approaches to enhance the understanding and disentanglement of these latent spaces [4–9]. The interest lies especially in architectures such as StyleGAN [1,2] or BigGAN [3], which are trained with this perspective in mind. Despite these advancements, the challenge remains in achieving exact and localised editing capabilities without introducing unwanted artefacts or other features, commonly referred to as the entanglement of semantics in the latent space. With the introduction of encoding4editing (e4e) networks [10], many of the newly proposed methods [11–14] have shifted their focus from the GANs generator network as a latent space encoder.

We underscore the potential of StyleGAN for attribute manipulation by employing it within a teacher-student-like framework. We leverage the conditional capabilities of this GAN to generate on-the-fly data for the student network through a novel and faster training approach. The student network, referred to as *MiddleNet*, is designed as an extension of the original StyleGAN and is situated between the teacher's Mapping and Synthesis networks. This network is specifically trained to learn different transformations inside the latent space to achieve the desired attribute given a text.

Our method addresses two significant limitations of existing approaches. Firstly, we address specific and localised editing, effectively proposing modifications within the latent style space of the aforementioned network. We select the semantic features to be altered based on a text prompt, which is entirely processed by the MiddleNet by managing latent style changes for each layer in StyleGAN's $\mathcal{W}+$. Consequently, we obtain more localised edits than the teacher network. The second challenge we address regards the improvement in identity preservation, as a by-product of changing only the layers related to the requested feature, ensuring minimal changes of unrelated attributes in the overall image's original feature and structure.

* Corresponding author.
E-mail addresses: andrei_radu.danila@upb.ro (A. Radu), yue.song@unitn.it (Y. Song), ana_antonia.neacsu@upb.ro (A. Neacsu), niculae.sebe@unitn.it (N. Sebe).

https://doi.org/10.1016/j.patrec.2026.02.009
Received 25 July 2025; Received in revised form 13 December 2025; Accepted 8 February 2026
Available online 10 February 2026

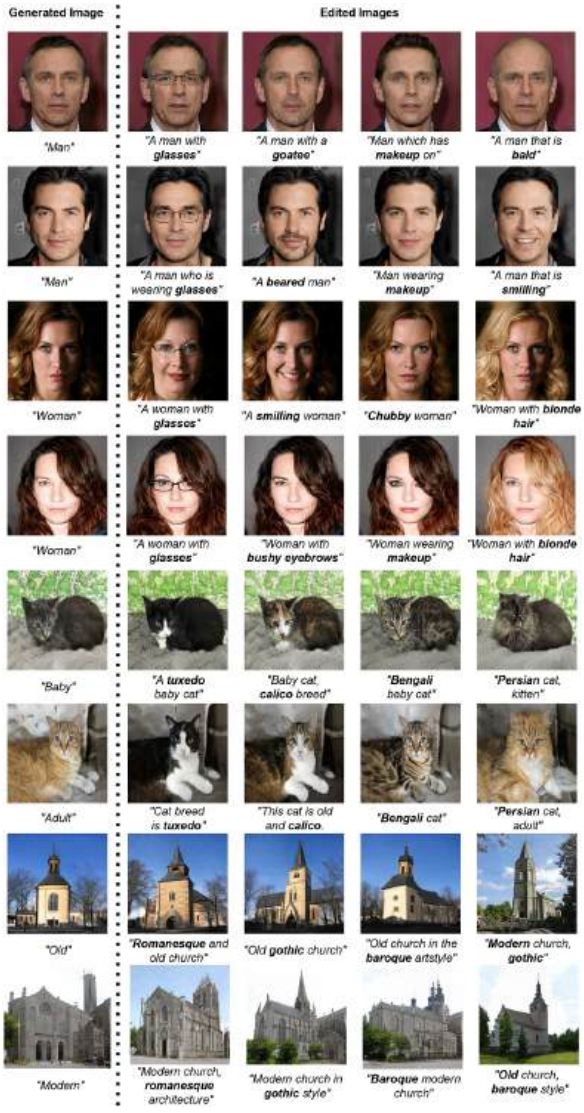

Fig. 1. Examples of generated and edited images. On the right side, the generated images use the class conditioning generation of StyleGAN in order to get a base image. The remaining images represent the edited ones, obtained using the given text prompt mentioned below each image. We used **bold** to highlight the requested feature.

This method serves as a base for an image retrieval system by storing only the encoded images. We argue that editing solutions are also viable when only small details need adjustment [15,16] leading to a larger similarity between the query and retrieved images. Attribute editing shares conceptual similarities with image retrieval. In particular, feature representation is a key step in both manipulating and retrieving content[17], where latent space proximity allows for an efficient search in a retrieval system and localised edits in attribute manipulation. Works such as [18–20] assert similarities in methodology between the two domains using latent space navigation, similarity search and user-guided interactions. To summarise, our main contributions are as follows:

- We propose a novel, faster and more precise image editing method through latent space modifications enabled by text prompting, which better preserves the original appearance of the image, as illustrated in Fig. 1. Our method presents minimal additions in terms of model size and inference speed, making it ideal for systems where the response time is crucial or resources are limited. On consumer-level hardware, the overhead is approximately **4.5 milliseconds** and **16M-parameters**.
- We address the problem of similar image retrieval by localised image attribute modification with a new text token manipulation method over a trained from scratch small encoder rather than heavily relying on larger, pre-trained ones like CLIP. Our text manipulation aims to uncover a more disentangled direction within each layer of the latent style space by addressing a specific token for each style layer.
- To our knowledge, this approach is the first method to take advantage of both **conditional generation** and **teacher-student training**. We combine those by leveraging the benefits of conditional training to train a distilled student which operates directly in the latent space. As opposed to traditional teacher-student distillation, we train the distilled version to enhance the generative capabilities of the main network, while also adding more control over it.
- Our extensive experiments show that the proposed method achieves state-of-the-art results across three distinct datasets, in terms of image subject and context.

2. Related work

GANs for Image Manipulation. GANs have revolutionised computer vision generative domains, starting with image generation, super-resolution and image (or image style) translation, which has been laying the groundwork for Image Attribute Editing. The research interest in editing image attributes has gained momentum once GAN architectures were able to control the features in the images more easily. Such examples are given by BigGAN [3], ProjGAN [21] or StyleGAN-2 [2]. The latter has been the subject of numerous works on both image [22,23] and video [24,25] tasks thanks to its extremely disentangled latent space. Earlier approaches [5,26–28] have focused on finding disentangled directions in this space, allowing users to change certain aspects of the final image w.r.t. mathematical operations such as PCA [27] or SVD [5].

Image Manipulation using text. Since textual prompts are easier to understand for humans, approaches such as ManiGAN [29] have relied on their image manipulation method on them. TAGAN [30] proposes a latent space editing guided by the provided text by using a dual-attention mechanism, stressing the modification only in the necessary areas. ManiGAN introduces a dynamic attention mechanism with aligns the text descriptions and the matching regions in the images, allowing for even more localised edits. With the introduction of CLIP and LLMs, a new paradigm has been developed, which proposes image editing using text prompts. StyleCLIP [11] was one of the pioneers when it comes to combining the power of StyleGAN and CLIP, by optimising the latent to generate images that become more similar to the provided text. StyleMC [12] continued this idea by including multi-channel control to obtain faster and more accurate control over the generated attributes. HairCLIP [13] introduces the idea of a Mapping network, trained to modulate the provided latent codes to the desired ones by injecting the CLIP text embeddings into this network. A more recent work [14] extends this to more general face attribute manipulation by modifying the latent codes in three steps. We argue that an encoder trained specifically to learn text representations for image editing on StyleGAN's style space, denoted by $\mathcal{W}+$, can achieve similar control as a large, multi-modal solution such as CLIP. We based our preposition on the fact that all the aforementioned methods also include a morphing of the CLIP text embeddings to match this space. As such, we choose to train a Transformer encoder to process the text instead of relying on pre-trained ones.

Diffusion-based methods for Image Editing. The majority of image editing techniques are now based on diffusion models [31–34]. Pioneering works in this domain include GLIDE [32] and Stable Diffusion [35]. GLIDE editing is based on image inpainting, combined with classifier-free guidance. Stable Diffusion optimises the process by applying the diffusion process inside the latent space. The nature of the iterative steps in diffusion makes these methods extremely time-consuming.

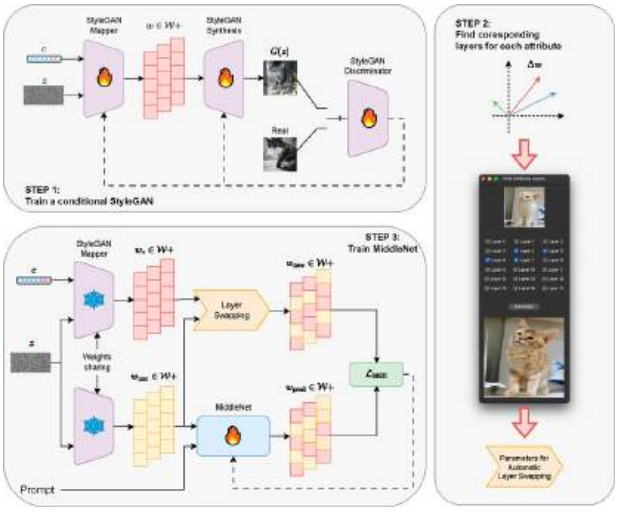

Fig. 2. Our framework is composed of three main steps. Firstly, we trained a conditional StyleGAN to be used as a teacher network. Then, we find the correspondence between the latent layers and the class conditioning. Finally, we train the MiddleNet (student) to reconstruct the selected conditional latent code layers based on text prompts, which are processed by an encoder trained from scratch.

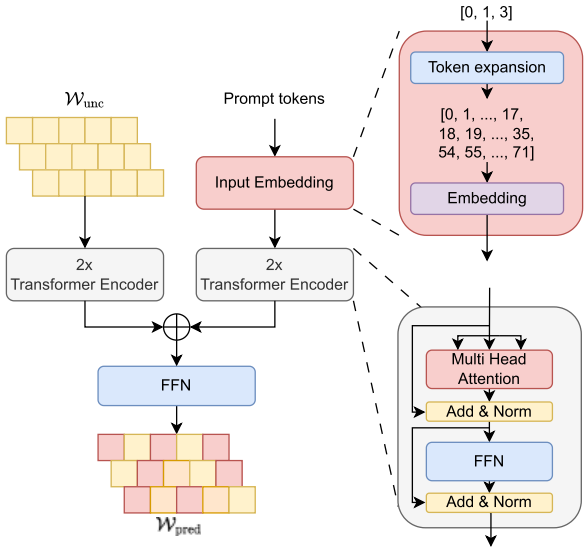

Fig. 3. MiddleNet (student) architecture. We convert both the text tokens and w style code into an intermediate representation, which we concatenate and feed into an FFN to obtain the final representation from which the images are synthesised.

As such, many works [36–39] have been dedicated to speeding up the process either by requiring fewer diffusion steps [39] or by optimising the noise level [37].

3. Methodology

Our approach utilises the conditional generative capabilities of StyleGAN in a teacher-student training framework. We introduce text prompts to further control the generation process in a more intuitive manner than class conditioning. An overview of our method is shown in Fig. 2. The process consists of several steps:

1. Training a class-conditional StyleGAN (*teacher*) to learn the representation of the features specific to each class inside the style space $\mathcal{W}+$. This model functions as both the instructor, generating soft style codes that will be treated as labels, and as a means of utilising the synthesis network to produce the final images. Moreover, the vast pool of pre-trained unconditional StyleGANs speeds this process by only requiring a class-conditional fine-tuning process.
2. For each class, we automatically find the layers which contain the most relevant change in direction for the features from the predicted latent spaces of the teacher.
3. Train a new network (*student*) that learns to map any style code s.t. it respects the text prompt, and aligns it with the initial teacher's condition. Compared to other methods, our approach achieves a significant improvement, i.e., **between 5% and 30%**, in identity preservation without sacrificing image quality.

3.1. Preliminaries

StyleGAN represents a class of generative adversarial network (GAN) architectures which are known for their high control over the styles at different levels, *e.g.* coarse, medium and fine. The generator of this network operates in two steps: mapping and synthesis. This dual approach allows for the generation of high-quality images with nuanced stylistic variations, facilitating extensive customisation and manipulation of the generated outputs. First, the *Mapping Network*, composed of 8 fully connected layers, takes a noisy input and converts it to a style code. Second, the *Synthesis Network* takes a set of style codes to generate the final image. This methodology allows for a more disentangled style space [40], in which we can perform more precise attribute manipulation. We use similar notations to [2], where: $z \in \mathcal{N}(\mu, \sigma^2)$ denotes the noisy input, $w \in \mathcal{W} \subseteq \mathbb{R}^{512}$ represents a single style code, while $\mathcal{W}+ \subseteq \mathbb{R}^{(18\times512)}$ is the extended style space.

3.2. Network architecture

Our model, named *MiddleNet*, lies inside the StyleGAN generators modules, between the *Mappiing* and *Synthesis* networks. The architecture is depicted in Fig. 3. Our objective is to develop a compact, efficient, and precise feature editing network that proposes modifications in the latent space corresponding to each query in an image-based retrieval system. The requested changes are intended to be integrated with the pre-existing latent representation extracted from the query image, facilitating the preservation of the identity and inherent features in the original image.

To achieve this, we use a transformer-like architecture to optimise both the style codes generated by the StyleGAN Mapper, or the e4e network and the text tokens. For each text token, we are training an intermediate representation $\Delta w \in \mathcal{W}+$ responsible for guiding the editing aspects of the final image. To match the layer dimension of $\mathcal{W}+$, we expand each text token $t^{(i)}$ into an 18-size vector. Specifically, $t_j^{(i)}$ represents a deterministic offset, producing one toker per network layer:

$$\xi(t^{(i)}) = \{t_j^{(i)} \mid t_j^{(i)} = L \cdot t^{(i)} + j,\ j \in \{1, \dots, L\}\} \tag{1}$$

where $\xi(\cdot)$ is the expansion function and L denotes the number of the style space layer, which is 18 in our case. We feed each of these new tokens to an embedding module to further expand their dimension s.t. it matches $\mathcal{W}+$. By concatenating the resulting representation for each expanded token, we obtain $\Delta w^{(i)} = [w_1^{(i)}, \dots, w_L^{(i)}]$, which represents the change in direction for each layer individually w.r.t. the text token $t^{(i)}$. During our experiments, we observed that the best results are obtained by employing a two-layers encoder-only transformer with an 8-head attention mechanism for each branch. For each layer, we used a 512-dimension embedding, to match the dimension of the style codes. Each FNN has two layers with an intermediate representation that is 1024-dimensional. To integrate prompts of arbitrary length, we aggregate all per-token direction changes into a single global style shift $\Delta w \in \mathcal{W}+$ derived from the text prompt. The original style codes, denoted by $w \in \mathcal{W}+$ are also processed by *MiddleNet*, yielding an intermediate representation that is particularly susceptible to modifications in the middle

layers of $\mathcal{W}+$ (please refer to the Appendix for details). We concatenate these with the embedded text and employ a feed-forward network (FFN) to efficiently integrate these newly processed style codes. This approach facilitates the creation of an abstract non-linear combination of the two different representations while smoothing the overall output.

3.3. Training procedure

StyleGAN training represents the first step in our framework, enabling us to harness the robust generative capabilities of a guided GAN to enhance the performance of the student network. In this phase, we employ both the Generator and the Discriminator. We selected the second generation of StyleGAN (StyleGAN2) and focused our attention on the F training scheme, which is used to preserve the initial disentanglement of features inside the style space when working with pre-trained models. Class conditioning is achieved by modifying both the generator's input as well as the discriminator architecture. Firstly, the class-conditioned one-hot vector c is concatenated with the noise z before feeding it into the Mapping network, influencing how the generator transforms the w latent codes. For the discriminator, the required modification regards the addition of new classes.

The layer selection step is automatically implemented to ensure that the student network achieves the necessary output image with minimal alterations while preserving as much of the original image as possible. As demonstrated in previous works [5–7,41,42], each layer of the StyleGAN governs one or more semantic attributes in the generated image, allowing for a transition between two images by continuously interpolating their latent spaces. Inspired by the [5] and [7], we implement a latent semantics discovery module w.r.t the condition class. First, we trained an attribute classifier from scratch for each dataset. We then use this classifier to identify the most influential layers. At each iteration, we generate two samples, each with its corresponding latent: one conditioned and one unconditional. Next, we apply stochastic layer swapping, replacing selected components of the unconditional latent with the corresponding components from the conditioned latent, and generate a new image. For each layer, we quantify its effect by measuring the change in the classifier's predicted probability for the target attribute. After multiple iterations, we select the layers that produce the largest shifts towards the desired class for each attribute. The full procedure is detailed in Appendix Algorithm 1. In addition, we implemented a Graphical User Interface (GUI), depicted in Fig. 2 only to validate the correctness of the layer selection module, as an optional step. More details about our findings regarding this can be found in Section A.1.

Training MiddleNet as a student represents the final step in our framework. Inspired by Latent Diffusion Models (LDMs) [35], we train this network to be effective in the extended style space $\mathcal{W}+$. To achieve this, we use the *Mapping* network of the teacher to generate two sets of latent code at each training iteration. Firstly, we select the random noise z and the class-conditioning one-hot vector c, subsequently feeding them into the *Mapper* in two distinct ways: in the first configuration, both z and c are input simultaneously, while in the second configuration, only the noise z is used. As such, we obtain two separate latent codes, which we denote by w_{c} for the conditional input and w_{unc} for the unconditional one. We apply the same layer swapping algorithm to obtain a new set of style codes, denoted w_{new}, which will serve as ground truths. Our MiddleNet is trained to predict style codes that resemble w_{new} by only receiving the unconditional latent codes w_{unc} alongside a text prompt. We assert this by utilising the Mean Squared Error (MSE) as a loss function between the predicted w_{pred} and w_{new}.

Our training framework ensures that the generalisation capabilities of the teacher are transferred to the student by sharing a common *Mapping* and *Synthesis* network. We took inspiration from the LDMs' approach of sharing the VAE encoder and decoder network, while all the transformations are done inside the latent space. As illustrated in Fig. 2, the student network is capable of applying more localised modifications compared to the teacher counterpart. This characteristic also results in the features predicted inside the pre-trained $\mathcal{W}+$ being more effectively disentangled once they are passed through MiddleNet.

4. Experiments

4.1. Datasets

To prove the efficacy of our method, we selected three benchmark datasets that are commonly used in image generation tasks: CelebA and two subsets of the LSUN , namely, Cats and Churches. The datasets were selected based on their dissimilarities with one another, resulting in a vast and complex testing framework. More details about the datasets and their limitations are presented in Section A.2.

4.2. Metrics

We evaluated our method against several state-of-the-art image editing methods, based either on StyleGAN2 or diffusion methods. We compared the generation from three distinct perspectives: realism, identity preservation and the ability to edit specific attributes. We also indicated **the highest-performing** and second-best methods.

FID. Firstly, we are interested in the quality of the generated samples in terms of realism, which is also important to ensure that the StyleGAN has been trained enough. As such, we relied on the Fréchet Inception Distance (FID).

ID. Preservation of identity–facial features for human faces, breed characteristics for cats, and architectural styles for churches–is a critical aspect of an efficient attribute editing method using image generation. We used the Identity similarity (ID) to measure this, which is computed using ArcFace , for human faces, and CLIP image embeddings for the other datasets. This results in a new but similar metric to ID. We termed this metric the CLIP Image Similarity Score (CISS), which calculates similarity using CLIP embeddings instead of relying on face-specific features.

AMA and CMP. To evaluate the accuracy of the attribute manipulation, we took inspiration from [14], in which two new metrics are proposed: Attribute Manipulation Accuracy (AMA) and CLIP Manipulative Precision (CMP). For the AMA(s), we selected a *single* attribute to manipulate for each image. Although the model was not trained specifically on *multiple* attribute manipulation, we also evaluated it using AMA(m). In this setting, we sampled two or three attributes per image, as proposed by the authors of the metric. The CLIP Manipulative Precision (CMP) score represents a variation of the Manipulative Precision, first presented in [29].

4.3. Quantitative results

Table 1 presents the obtained results on CelebA, while Table 2 is dedicated to results on the LSUN Cats and Churches datasets. We have evaluated our method using the same evaluation protocol as [14].

On CelebA, our methods obtained a significantly lower FID score while also improving the ID and AMA(s). The images generated by our proposed method tend to be more realistic in terms of added features than other methods, especially from diffusion ones, while also having more diversity in terms of generation, as denoted by the obtained FID. In terms of the manipulation of the required attribute, our approach achieves the best results, as illustrated by the CMP and AMA(s) scores, with the main drawback in multiple-attribute control. We can assert these to the training scheme, which did not include multiple arguments. However, amongst the StyleGAN-based methods, our approach achieves the second-best results even in this category. Since we concentrate on localised edits and their representation in the latent space, we effectively maintain identity, as indicated by the ID score.

The compared StyleGAN-based methods use dedicated encoding networks [10,46] to obtain the latent codes from images, which has been

Table 1
Quantitative evaluation metrics on the CelebA dataset. The resulting FIDs are computed on the images with modified attributes. *Italicized* method are diffusion-based.

Method	FID ↓	ID ↑	CMP ↑	AMA(s) ↑	AMA(m) ↑
StyleGAN-based methods					
TediGAN-B [43]	55.424	37.97	**0.285**	11.286	1.142
StyleCLIP-LO [11]	80.833	29.69	0.210	15.857	3.429
StyleCLIP-GD [11]	82.393	57.37	0.191	33.143	11.429
StyleMC [12]	84.088	30.05	0.187	12.143	2.857
HairCLIP [13]	93.523	57.50	0.218	41.571	15.143
CLIPInverter [14]	97.210	52.14	0.221	61.429	41.714
Diffusion model-based methods					
TurboEdit [44]	90.462	54.94	0.215	59.868	59.372
Imagic [45]	57.695	41.64	0.179	68.593	**60.037**
Ours	**54.552**	**64.68**	0.222	**70.635**	28.173

Table 2
Quantitative results for LSUN Cats and Churches. The FID is computed after the attribute is manipulated. *Italicized* method are diffusion-based.

Dataset	Method	FID ↓	CISS ↑	CMP ↑	AMA(s) ↑
LSUN Cats	*TurboEdit* [44]	8.973	78.71	0.159	66.94
	Imagic [45]	**8.411**	87.42	0.141	**76.07**
	Ours	9.370	**91.94**	**0.208**	75.73
LSUN Churches	*TurboEdit* [44]	24.060	81.25	0.170	**62.22**
	Imagic [45]	45.286	**86.19**	0.145	61.76
	Ours	**15.869**	85.99	**0.188**	57.71

proven to be a fast and efficient way of inverting images to their corresponding style codes [40]. Since there are no publicly available inverters for images containing cats, we decided to compare our method to more advanced ones, based on diffusion. We turned our attention to the fastest editing methods available and selected two of them: TurboEdit [44] and Imagic [45], which are based on StableDiffusion-1.5 and StableDiffusion-XL Turbo [47], respectively. To ensure a fair comparison, both methods were evaluated using the default settings recommended by their respective authors.

While diffusion-based models demonstrate superior generative capabilities, this advantage is offset by substantial resource demands, necessitating significant VRAM and incurring prolonged inference times. In contrast, our proposed method achieves comparable FID and AMA(s) scores with significantly improved computational efficiency.

4.4. Qualitative comparison

A qualitative assessment of our results is presented in Fig. 4, which displays generated images from our method and four comparative approaches, including a diffusion-based one. The first column shows the reference images; subsequent columns present results from TediGAN (selected for its second-best FID score among StyleGAN-based methods), StyleMC (representing an optimisation-based approach), CLIPInverter (demonstrating strong AMA performance, especially in a multiple attribute scenario among StyleGAN methods), and TurboEdit (a recently proposed, faster diffusion-based method). This selection allows for a comprehensive comparison across diverse methodologies.

Unlike other methods, our approach produces highly localised edits, preserving identity while modifying only the specified regions of the image. For example, TediGAN and StyleMC tend to add incomplete features, such as eyeglasses on the top row or the incomplete beard on the second row, while CLIPInverter sometimes exaggerates, as shown in the fourth and final rows. TurboEdit, employing a fundamentally different image editing paradigm, serves as a valuable comparative point with newer methods. First of all, the images generated by this method tend to have hyper-saturated colours, resulting in an unnatural aesthetic, which represents a common limitation for diffusion models. Furthermore, modifications to extensive attributes, such as glasses or hair, often lead to unintended alterations in adjacent features; for instance, nose modifications may accompany changes to glasses, and hair position changes may occur alongside hair edits. All implementations struggle to generate realistic representations of the scarce images from the dataset, such as *bald women*. Moreover, in those cases, we can observe biased examples in both GANs and diffusion-based approaches. For more qualitative results on LSUN Churches and Cats, and a discussion of the two aforementioned problems, please refer to the supplementary.

Fig. 4. Qualitative comparison between different methods on the CelebA dataset. Our method is presented in the last column.

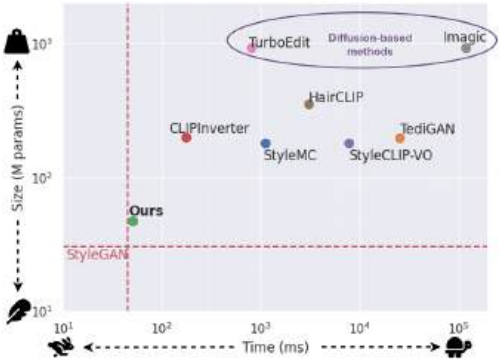

Fig. 5. Number of parameters and inference time for each model. Our proposed method is both lighter and faster than other methods. Note that the scale is logarithmic.

4.5. Performance analysis

Besides obtaining state-of-the-art attribute manipulation capabilities, our solution also demonstrates a significant improvement in computational efficiency compared to existing models. Fig. 5 displays a comprehensive analysis of both the model size and inference time. To mimic

an end user, all the tests were run on a commercially available GPU, namely an NVIDIA RTX 3060 paired with a Ryzen 5 3600 CPU. We ran 100 iterations for each method and averaged the inference time.

5. Conclusions

This work introduces a novel, efficient teacher-student-like training paradigm for controlling StyleGAN generation and editing via text prompts. Our key contributions are threefold: (i) We present a method for deriving a set of directional representations from the text embeddings, which enables precise and localised attribute manipulation, through morphing with pre-existing w ones. (ii) We propose a novel text embedding method that fosters greater independence and disentanglement of directional edits within StyleGAN's latent space. (iii) Our method achieves state-of-the-art results across multiple benchmark datasets with only a negligible number of parameters and inference time overhead to the original network, offering a lightweight and computationally efficient alternative to existing approaches.

CRediT authorship contribution statement

Andrei Radu: Writing - original draft, Methodology; **Yue Song:** Writing - review & editing, Conceptualization; **Ana Neacsu:** Writing - review & editing, Supervision; **Nicu Sebe:** Writing - review & editing, Supervision.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has been supported by the EU Horizon project ELIAS (No. 101120237) and by the FIS project GUIDANCE (No. FIS2023-03251).

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.patrec.2026.02.009

References

[1] T. Karras, S. Laine, T. Aila, A style-Based generator architecture for generative adversarial networks, in: CVPR, 2019.
[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of styleGAN, in: CVPR, 2020.
[3] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, in: ICLR, 2019.
[4] Y. Hu, J. Xia, H. Liu, X. Wang, Unsupervised face image deblurring via disentangled representation learning, in: Pattern Recognition Letters, 2024.
[5] Y. Shen, B. Zhou, Closed-Form factorization of latent semantics in GANs, in: CVPR, 2021.
[6] J. Zhu, Y. Shen, Y. Xu, D. Zhao, Q. Chen, Region-Based semantic factorization in GANs, in: ICML, 2022.
[7] Y. Song, J. Zhang, N. Sebe, W. Wang, Householder projector for unsupervised latent semantics discovery, in: CVPR, 2023.
[8] G. Scaramuzzino, F. Becattini, A. Del Bimbo, Attribute disentanglement with gradient reversal for interactive fashion retrieval, in: Pattern Recognition Letters, 2023.
[9] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, InfoGAN: interpretable representation learning by information maximizing generative adversarial nets, in: NeurIPS, 2016.
[10] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, D. Cohen-Or, Designing an encoder for styleGAN image manipulation, in: TOG, 2021.
[11] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, D. Lischinski, StyleCLIP: text-Driven manipulation of styleGAN imagery, in: CVPR, 2021.
[12] U. Kocasari, A. Dirik, M. Tiftikci, P. Yanardag, StyleMC: multi-Channel based fast text-Guided image generation and manipulation, in: WACV, 2021.
[13] T. Wei, D. Chen, W. Zhou, J. Liao, Z. Tan, L. Yuan, W. Zhang, N. Yu, HairCLIP: design your hair by text and reference image, in: CVPR, 2022.
[14] A.C. Baykal, A.B. Anees, D. Ceylan, E. Erdem, A. Erdem, D. Yuret, CLIP-Guided styleGAN inversion for text-driven real image editing, in: TOG, 2023.
[15] A. Zaeemzadeh, S. Ghadar, B. Faieta, Z. Lin, N. Rahnavard, M. Shah, R. Kalarot, Face image retrieval with attribute manipulation, in: ICCV, 2021.
[16] H. Li, Y. Zhang, W. Wang, Enhanced blind face inpainting via structured mask prediction, in: Pattern Recognition Letters, 2024.
[17] G.R. Leticio, V.S. Kawai, L.P. Valem, D.C.G. Pedronette, R. da S. Torres, Manifold information through neighbor embedding projection for image retrieval, in: Pattern Recognition Letters, 2024.
[18] J. Tian, X. Xu, Z. Wang, F. Shen, X. Liu, Relationship-Preserving knowledge distillation for zero-Shot sketch based image retrieval, in: ACM MM, 2021.
[19] Y. Hou, E. Vig, M. Donoser, L. Bazzani, Learning attribute-Driven disentangled representations for interactive fashion retrieval, in: ICCV, 2021.
[20] A. Baldrati, M. Bertini, T. Uricchio, A. Del Bimbo, Effective conditioned and composed image retrieval combining CLIP-Based features, in: CVPR, 2022.
[21] T. Miyato, M. Koyama, CGANs with projection discriminator, in: ICLR, 2018.
[22] Y. Poirier-Ginter, J.-F. Lalonde, Robust unsupervised styleGAN image restoration, in: CVPR, 2023.
[23] Y.X. Tan, C.P. Lee, M. Neo, K.M. Lim, J.Y. Lim, Text-to-image synthesis with self-supervised bi-stage generative adversarial network, in: Pattern Recognition Letters, 2023.
[24] I. Skorokhodov, S. Tulyakov, M. Elhoseiny, StyleGAN-V: a continuous video generator with the price, image quality and perks of styleGAN2, in: CVPR, 2022.
[25] J. Choi, K. Seo, A. Ashtari, J. Noh, StylecineGAN: landscape cinemagraph generation using a pre-trained styleGAN, in: CVPR, 2024.
[26] Y. Shen, J. Gu, X. Tang, B. Zhou, Interpreting the latent space of GANs for semantic face editing, in: CVPR, 2020.
[27] E. Härkönen, A. Hertzmann, J. Lehtinen, S. Paris, GANSpace: Discovering interpretable GAN controls, in: NeurIPS, 2020.
[28] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Perez, M. Zollhofer, C. Theobalt, Stylerig: rigging styleGAN for 3D control over portrait images, in: CVPR, 2020.
[29] B. Li, X. Qi, T. Lukasiewicz, P.H.S. Torr, ManiGAN: text-Guided image manipulation, in: CVPR, 2020.
[30] S. Nam, Y. Kim, S.J. Kim, Text-Adaptive generative adversarial networks: manipulating images with natural language, in: NeurIPS, 2018.
[31] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: NeurIPS, 2020.
[32] A.Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, M. Chen, Towards photorealistic image generation and editing with text-Guided diffusion models, in: ICML, 2022.
[33] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, in: ICLR, 2020.
[34] B. Kim, K.-A. Sohn, Text-free diffusion inpainting using reference images for enhanced visual fidelity, in: Pattern Recognition Letters, 2024.
[35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: CVPR, 2022.
[36] J. Wang, W. Zhu, P. Wang, X. Yu, L. Liu, M. Omar, R. Hamid, Selective structured state-Spaces for long-Form video understanding, in: CVPR, 2023.
[37] S.X. Chen, Y. Vaxman, E. Ben Baruch, D. Asulin, A. Moreshet, K.-C. Lien, M. Sra, S. Pradeep, TiNO-Edit: timestep and noise optimization for robust diffusion-Based image editing, in: CVPR, 2024.
[38] Y. Shang, Z. Yuan, B. Xie, B. Wu, Y. Yan, Post-Training quantization on diffusion models, in: CVPR, 2023.
[39] Z. Zhou, D. Chen, C. Wang, C. Chen, Fast ODE-based sampling for diffusion models in around 5 steps, in: CVPR, 2024.
[40] Z. Wu, D. Lischinski, E. Shechtman, Stylespace analysis: disentangled controls for styleGAN image generation, in: CVPR, 2021.
[41] J. Zhu, R. Feng, Y. Shen, D. Zhao, Z.-J. Zha, J. Zhou, Q. Chen, Low-Rank subspaces in GANs, in: NeurIPS, 2021.
[42] Y. Wei, Y. Shi, X. Liu, Z. Ji, Y. Gao, Z. Wu, W. Zuo, Orthogonal jacobian regularization for unsupervised disentanglement in image generation, in: ICCV, 2021.
[43] W. Xia, Y. Yang, J.-H. Xue, B. Wu, TediGAN: text-Guided diverse face image generation and manipulation, in: CVPR, 2021.
[44] Z. Wu, N. Kolkin, J. Brandt, R. Zhang, E. Shechtman, Turboedit: instant text-based image editing, in: ECCV, 2024.
[45] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, M. Irani, Imagic: text-Based real image editing with diffusion models, in: CVPR, 2023.
[46] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, D. Cohen-Or, Encoding in style: a styleGAN encoder for image-to-Image translation, in: CVPR, 2021.
[47] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, SDXL: Improving latent diffusion models for high-Resolution image synthesis, in: ICLR, 2024.