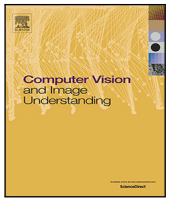
Computer Vision and Image Understanding 244 (2024) 104026

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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



Conditioning diffusion models via attributes and semantic masks for face generation

Giuseppe Lisanti *, Nico Giambi

Department of Computer Science and Engineering, University of Bologna, Mura Anteo Zamboni 7, Bologna, 40126, Italy

ARTICLE INFO

Keywords:
Diffusion model
Multi-conditioning
Face generation
Garment generation

ABSTRACT

Deep generative models have shown impressive results in generating realistic images of faces. GANs managed to generate high-quality, high-fidelity images when conditioned on semantic masks, but they still lack the ability to diversify their output. Diffusion models partially solve this problem and are able to generate diverse samples given the same condition. This paper introduces a novel strategy for enhancing diffusion models through multi-conditioning, harnessing cross-attention mechanisms to utilize multiple feature sets, ultimately enabling the generation of high-quality and controllable images. The proposed method extends previous approaches by introducing conditioning on both attributes and semantic masks, ensuring finer control over the generated face images. In order to improve the training time and the generation quality, the impact of applying perceptual-focused loss weighting into the latent space instead of the pixel space is also investigated. The proposed solution has been evaluated on the CelebA-HQ dataset, and it can generate realistic and diverse samples while allowing for fine-grained control over multiple attributes and semantic regions. Experiments on the DeepFashion dataset have also been performed in order to analyze the capability of the proposed model to generalize to different domains. In addition, an ablation study has been conducted to evaluate the impact of different conditioning strategies on the quality and diversity of the generated images.

1. Introduction

Image synthesis has recently become a hot topic, mostly thanks to the vast number of successful applications proposed in the literature. Among the different generation tasks, several works have focused attention on semantic face image synthesis (Li et al., 2023; Tan et al., 2021; Rombach et al., 2022). Most of these solutions rely on GANs' and their ability to generate high-quality and high-fidelity results (Tan et al., 2021; Li et al., 2019; Xiao et al., 2021). However, their uni-modal nature prevents them from generating diverse samples (Wang et al., 2022). Diffusion Models (DM) have proven to compete with GANs in both quality and fidelity while being multi-modal generators (Ho et al., 2020; Rombach et al., 2022; Wang et al., 2022; Choi et al., 2022; Dhariwal and Nichol, 2021). They are parametrized Markov chains that optimize the variational lower bound on the likelihood function to generate samples matching the data distribution. In order to generate an image, DMs iteratively refine a Gaussian noise via a denoising process, that is implemented with a UNet (Ronneberger et al., 2015) backbone.

In this paper, we show how to achieve and surpass the actual state-of-the-art for semantic face image synthesis, following three main evaluation criteria: quality, fidelity, and diversity. To improve the generation quality, a reweighed loss function is employed (Choi et al., 2022), which forces perceptual quality over unperceivable high-frequency details. Improved fidelity is achieved by utilizing a powerful conditioning mechanism, which in this case is cross-attention (Vaswani et al., 2017), in conjunction with semantically and spatially rich encodings. Diversity is then examined by capitalizing on Diffusion Models' innate ability to generate multi-modal images. This is accomplished by applying stricter or looser conditioning, resulting in more consistent or diverse generated images, respectively. Finally, a method is proposed to harness cross-attention for conditioning a diffusion model with multiple features simultaneously, enabling a greater degree of control over the generation process. In our solution, both facial attributes and semantic masks are considered, but the same idea could be extended to any other domain and set of features. For example, it could be possible to condition a model with both a semantic layout and a certain time of the day in order to generate landscapes with the right colors and shading or combine sketches and textual descriptions in order to generate images of suspects in the forensics field.

Our contributions can be summarized as follows:

- the analysis of perception prioritizing loss weighting (Choi et al., 2022) in the latent space of Latent Diffusion Models (Rombach et al., 2022), which enhances the quality of generated samples without increasing the model's size or training/sampling time.

* Corresponding author.
E-mail address: giuseppe.lisanti@unibo.it (G. Lisanti).

https://doi.org/10.1016/j.cviu.2024.104026
Received 28 September 2023; Received in revised form 27 March 2024; Accepted 19 April 2024
Available online 27 April 2024

- a multi-conditioning solution to impose more strict and precise control over the generated images. This mechanism lets the user combine spatial-only conditioning (*e.g.*, semantic masks), with descriptive features (*e.g.*, colors or level of detail from attributes). Additionally, we show that multi-conditioning causes a slight decrease in quality but results in high fidelity on both the provided conditioning.
- a state-of-the-art model for semantic face image synthesis, surpassing previous works in terms of generated images' quality, fidelity, and diversity.

2. Related works

In the following, we provide an analysis of some of the most recent works related to the proposed solution.

2.1. Denoising diffusion models

Recently, Diffusion Models (Sohl-Dickstein et al., 2015), have achieved state-of-the-art results in various generative tasks, such as Image Synthesis (Ho et al., 2020; Dhariwal and Nichol, 2021; Ho et al., 2022), Image Inpainting (Rombach et al., 2022) and Image-to-Image Translation (Wu and De la Torre, 2022). Ho et al. (2020) performed an empirical analysis to propose a reweighted loss function. As an extension, Choi et al. (2022) generalized this concept in order to establish a Perception Prioritized (P2) Weighting of the training objective. Recently, Rombach et al. (2022) obtained outstanding results by introducing a Latent Diffusion Model (LDM) in order to compress data and denoise them in a smaller latent space, reducing by a great margin the amount of resources used for both the training and the sampling stages. Wu and De la Torre (2022) analyzed the latent variables of different implementations of DMs (Song et al., 2020; Xiao et al., 2021) to perform Unpaired Image-to-Image Translation. Our solution builds upon (Rombach et al., 2022) and Choi et al. (2022) to obtain a model that is able to maximize the quality, fidelity, and diversity generation criteria while also introducing a multi-conditioning mechanism.

Despite their astounding performance reported across several works, Diffusion Models come with a pretty heavy burden, which is mostly represented by the computational costs both at train and test time. For this reason, several solutions that try to overcome this issue emerged in the most recent years. This newer paradigm allows for avoiding training the whole Diffusion Model from scratch when a new condition is introduced, lowering the computational requirement. It revolves around the introduction of an additional neural network, much smaller than the Diffusion Model, trained to inject the condition into pre-trained DM (Ham et al., 2023; Zhang et al., 2023; Huang et al., 2023; Yu et al., 2023; Liu et al., 2023; Mou et al., 2023). Among the solutions based on DMs, ControlNet (Zhang et al., 2023) and T2i-Adapter (Mou et al., 2023) explored the possibility of using several conditioning methods, like semantic masks, text, Canny edges, human pose, and their combination, to control the output of a Stable Diffusion model. Differently, FreeDoM (Yu et al., 2023) adopted this paradigm to introduce a network capable of conditioning various diffusion-based architectures and conducted several experiments to demonstrate the adaptability of their solution. Collaborative Diffusion (Huang et al., 2023) proposed to use a group of uni-modal pre-trained DM to achieve Multi-modal face generation. Despite the reduced computational requirements, these solutions exhibit a drop in performance compared to solutions that have been trained with a specific condition or a set of conditions as in our proposed solution.

2.2. Attributes controlled generation

Attributes Controlled Generation can both indicate synthesis and editing. In the last few years, *attributes-controlled* image editing has received a lot of attention (Choi et al., 2020; Gao et al., 2021; Hou et al., 2022), while *attributes-conditioned* image synthesis has not been of major interest. In Li et al. (2022), the authors proposed a text-to-image generation process that relies on the text transposition of the CelebA-HQ attributes and compared their results with other similar studies (Xia et al., 2021; Li et al., 2019; Ruan et al., 2021).

Since the study conducted in this paper focuses on attributes-conditioned image synthesis, we will compare the performance of our model to the methods proposed in Xia et al. (2021), Li et al. (2019) and Ruan et al. (2021) that are the closest to our solution, while the approaches in Choi et al. (2020), Gao et al. (2021) and Hou et al. (2022) are not directly comparable.

2.3. Semantic image synthesis

Over the years, semantic image synthesis has been mainly addressed by exploiting GAN-based models (Park et al., 2019; Zhu et al., 2020, 2017; Tan et al., 2021; Richardson et al., 2021). Among these, SPADE (Park et al., 2019) and SEAN (Zhu et al., 2020) focused on generating unimodal images while other works like BycicleGAN (Zhu et al., 2017) and INADE (Tan et al., 2021) explored multimodal generation, which consists in generating high-fidelity and diverse samples. Recently, diffusion models have proved to obtain generation results with higher diversity and fidelity (Rombach et al., 2022; Wang et al., 2022). Wang et al. proposed Semantic Diffusion Model (SDM) (Wang et al., 2022), for semantic image synthesis through DMs. SDM processes the semantic layout and the noisy image separately, in particular, it feeds the noisy image to the encoder stage of the U-Net model and the semantic layout to the decoder, using multi-layer spatially-adaptive normalization operators. This results in higher quality and semantic correlation of the generated images.

Differently from these approaches, cross-attention (Vaswani et al., 2017) allows more flexible and powerful control over the generation results, enabling us to execute multi-conditioning of a DM by utilizing both semantic layouts and facial attributes.

3. Multi-conditioning of latent diffusion model

This section initially offers an overview of latent diffusion models. It then delves into a comprehensive explanation of the approach devised for integrating semantic masks and attributes to influence the generation process. Lastly, it shows how the loss weighting method proposed in Choi et al. (2022) can be integrated into the training of the proposed model.

3.1. Latent diffusion model

Rombach et al. introduced the Latent Diffusion Model (LDM) (Rombach et al., 2022) to minimize DMs' computational demands while maximizing the generated samples' quality. They proposed a general purpose, perceptually focused Encoder ($\mathcal{E}$) to project the high-quality input image from pixel space to a lower dimensionality, semantically equivalent, latent space. The smaller input helps to speed up the training since it is possible to feed the model with bigger batches, but the most important advantage can be observed during the sampling. The iterative denoising process, indeed, usually requires about 500 steps. Therefore, reducing the Gaussian Noise size by a factor of 4, on each spatial dimension, results in a much faster sampling in the DM's space. Additionally, both the Encoder and the Decoder only need a single pass, meaning they bring a negligible overhead to the denoising process computational cost. This Encoding–Decoding process separates the *semantic compression* and *perceptual compression* phases. The first is completely handled by the Encoder–Decoder while the latter is managed through the U-Net backbone, which can use all its parameters to focus on the perceptual part of the denoising. Since LDM achieved outstanding results for various Unconditioned and Conditioned tasks, we decided to base our work on this framework.

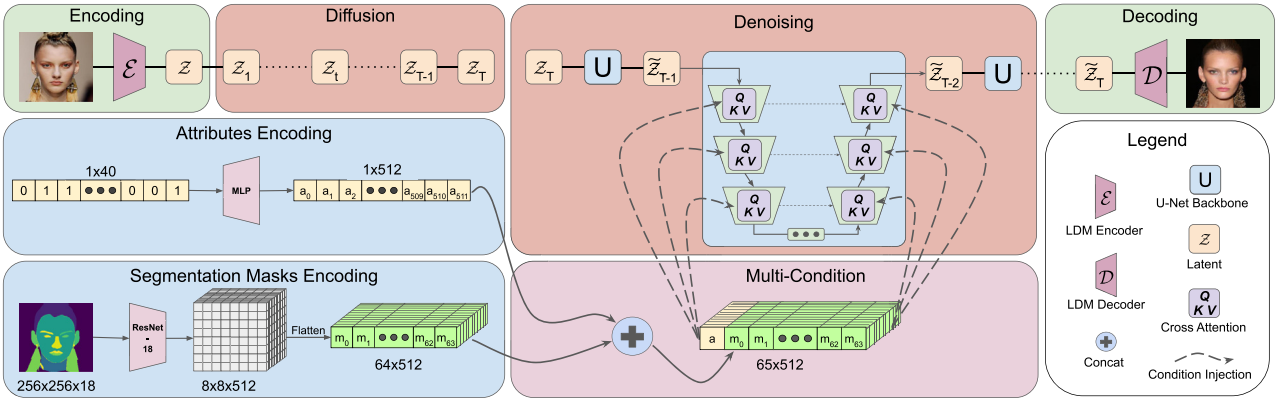


Fig. 1. Multi condition model schema. Single conditioning and Unconditioned generation are simplifications of this model.

3.2. Attributes and mask conditioning

Conditioning a generative model consists in injecting some kind of information, such that the generated samples will reflect this property. In GANs this information is usually injected exploiting a *normalization layer*, like semantic region-adaptive normalization in SEAN (Zhu et al., 2020), spatially conditioned normalization in SCGAN (Wang et al., 2021) and instance-adaptive denormalization in INADE (Tan et al., 2021). DMs use a similar process to inject information into the denoising process. For example, Dhariwal and Nichol (2021) proposed the adaptive group normalization (AdaGN) to condition the DM on both the class embedding and the time-step after each group normalization layer, while Wang et al. (2022) proposed the multi-layer spatially-adaptive normalization in order to feed the segmentation masks into the decoder stage of the denoising U-Net. Rombach et al. (2022), instead, exploited the transformer (Vaswani et al., 2017) as a flexible and powerful conditioning mechanism to be applied to a subset of layers of the U-Net. It is composed of three distinct components, the first of which is a self-attention mechanism, computed on the set of features from the relative U-Net layer. The output of the self-attention is then summed to the input features via residual connection and provided as input to a cross-attention mechanism which combines information from the previous layer and the condition. The output is again summed to the input of the cross-attention and passed through an expansion–compression feed-forward neural network (Vaswani et al., 2017) which provides the output, that represents the conditioned set of features. Our solution follows this approach in order to condition the model with: (i) an encoding of binary attributes; (ii) an encoding of segmentation masks; (iii) a sequence obtained as the concatenation of the encoding from both attributes and segmentation masks (Fig. 1).

As described above, among the different layers composing the transformer, the cross-attention (CA) is the one responsible for the injection of the condition and is defined as:

$$CA(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V, \qquad (1)$$
$$Q \in \mathbb{R}^{\phi\times(h\cdot d)}, \quad K \in \mathbb{R}^{\psi\times(h\cdot d)}, \quad V \in \mathbb{R}^{\psi\times(h\cdot d)},$$

where d is the dimension of each attention head output (i.e., $d = 64$ as in Vaswani et al. (2017) and Rombach et al. (2022)), h is the number of attention heads, $K, V \in \mathbb{R}^{\psi\times(h\cdot d)}$ are computed from the encoded conditioning while $Q \in \mathbb{R}^{\phi\times(h\cdot d)}$ is a representation obtained from the corresponding U-Net layer on which the transformer is applied. The dimension ϕ results from flattening the U-net activations of the relative layer, while the dimension ψ represents the length of the conditioning sequence.

The final output of the conditioning $CA(Q,K,V)$ will have the same dimension as the initial input $Q \in \mathbb{R}^{\phi\times(d\cdot h)}$ and will be provided as *conditioned* input to the next layer of the U-Net. It can be observed that the output shape does not depend on the conditioning sequence length ψ, and this allows providing a variable set of conditions. In our solution, three different conditioning are considered:

- binary attributes conditioning, which is obtained through an MLP that maps the 40 attributes to $\mathcal{Z}_a \in \mathbb{R}^{\psi_a\times(d\cdot h)}$;
- mask conditioning, $\mathcal{Z}_m \in \mathbb{R}^{\psi_m\times(d\cdot h)}$, which is obtained by feeding the semantic mask to a ResNet-18;
- multi-conditioning, $\mathcal{Z}_{mc} \in \mathbb{R}^{(\psi_a+\psi_m)\times(d\cdot h)}$, which is obtained by concatenating the two encodings along the ψ axis;

For the last point, in order to preserve more high-level semantic spatial information, the ResNet-18 features before the Global Average Pooling layer have been extracted. Working with 256×256 images, the ResNet-18 encoder maps the masks $m \in \mathbb{R}^{256\times256\times18}$ into $\mathcal{Z}_m \in \mathbb{R}^{(8\cdot8)\times(d\cdot h)}$. Consequently, the multi-condition encoder will generate $\mathcal{Z}_{mc} \in \mathbb{R}^{(1+64)\times(d\cdot h)}$, which is composed by one embedding for the attributes and 64 for the flattened masks features. The whole pipeline with the conditioning mechanism is illustrated in Fig. 1.

3.3. Perception prioritized loss weighting

Choi et al. (2022) analyzed the performance of the different stages of the DMs denoising process. By using perceptual measures like LPIPS (Zhang et al., 2018), they separate the diffusion process into three stages, parametrized on a Signal-to-Noise Ratio (SNR) (Kingma et al., 2021) depending on the variance schedule. These stages define when different levels of detail are lost during the diffusion, or vice-versa when they are generated in the denoising process. In the first stage of denoising, coarse details like color and shapes are generated. Then, in the content stage, more distinguishable features come up. In the final stage, the fine-grained high-frequency details are refined and most of them are not perceivable by the human eyes.

To this end, a Perception Prioritized (P2) Weighting of DM's loss function has been introduced:

$$L^t_{P2} = \frac{1}{(k + SNR(t))^\gamma}\mathbf{E}_{\mathbf{x},\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right] \qquad (2)$$

where k is a stabilizing factor that avoids exploding weights for small SNR values, usually set to 1, and γ is an arbitrary exponent that gives more or less importance to the re-weighting.

Our solution explores the possibility of employing this loss weighting in the latent space of LDM,[1] instead of the pixel space as performed in Choi et al. (2022). This is achieved by modifying the original loss formulation of (Rombach et al., 2022):

$$L^t_{LDM} = \mathbf{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(y))\|^2\right] \qquad (3)$$

[1] For the detailed mathematical derivation, please refer to the supplementary material.

by introducing the weighting factor from Eq. (2):

$$L^t_{LDM} = \mathbf{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\frac{\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(y))\|^2}{(k + SNR(t))^\gamma}\right]. \quad (4)$$

In both Eqs. (3) and (4), z_t is the latent representation of the input image obtained by the Encoder $\mathcal{E}$ at diffusion timestep t, τ_θ is the condition encoder model and y is its input which, for the proposed method, can be a segmentation mask or an attribute array.

4. Experimental results

This section reports on a set of experiments conducted to validate the proposed approach. First, some details about the datasets and settings used in all experiments are provided. Then, quantitative and qualitative generation results with unconditioned models and with models conditioned using attributes, semantic masks, or both are discussed.

Dataset. All experiments were conducted on the CelebAMask-HQ (Lee et al., 2020) and DeepFashion (Liu et al., 2016) datasets. The CelebAMask-HQ samples have been divided in a train/validation split of 25.000/5.000, as in LDM (Rombach et al., 2022), while for DeepFashion 11.500 samples have been used for training and 1.200 for validation. All images have been resized to 256 × 256 pixels and 256 × 384 pixels for CelebAMask-HQ and DeepFashion, respectively.

Metrics. Visual quality has been assessed by computing the Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018). Since, for conditioned tasks, it is also important to validate the correspondence between the generated samples and the condition, an accuracy score has been computed for the samples conditioned by masks, attributes or both.

As an additional correspondence metric, the mean Intersection over Union (mIoU) of segmentation masks obtained from mask-conditioned and multi-conditioned generation have also been measured. This has been computed using off-the-shelf segmentation models[2] to extract semantic masks from the generated images and compare them to their relative ground truth masks on which they were originally conditioned.

Finally, to evaluate diversity among samples conditioned on the same set of features, the LPIPS (Zhang et al., 2018) metric has been employed. In particular, for each conditioning method, the LPIPS has been computed among 10 samples.

In all tables, when a number follows a metric's name, it means that all results shown in that table are computed on that specific amount of samples. For unconditioned generation, the metrics have been computed on 50K generated samples, while for conditioned generations we sample as many images as in the validation set (e.g., 5K samples), using the set of attributes or masks provided with the validation samples. Each table includes metrics denoted by ↑ if higher is better, ↓ if lower is better.

In all tables, "*d.*" and "*s.*" denote that the result has been obtained using a deterministic ($\eta = 0.0$) or a stochastic ($\eta = 1.0$) sampling, respectively.

Train and test settings. The LDM's pre-trained encoder ($\mathcal{E}$) has been used to map images from the pixel space to a VQ-regularized latent space with a reduction factor of 4, hence performing diffusion and denoising on a 64 × 64 space. The latent space denoising U-Net (**U**), the image decoder ($\mathcal{D}$), the attributes encoder ($\mathcal{E}_a$), the ResNet-18 mask encoders ($\mathcal{E}_m, \mathcal{E}_{mnp}$) and the multi-condition encoder $\mathcal{E}_{mc}$ have all been trained from scratch for all our models. In particular, each model has been trained for 500 epochs considering a variance schedule of 1000 denoising steps, as in Choi et al. (2022). At test time, all samples are generated with 500 DDIM (Song et al., 2020) denoising steps, following Song et al. (2020) and Choi et al. (2022). It is worth highlighting that the solution proposed in Choi et al. (2022), at test time, uses a different number of sampling steps depending on the dataset under analysis. However, for the experiments conducted on CelebAMask-HQ, 500 denoising steps have been employed and for this reason, we employed the same number of sampling steps at test time.

[2] Available at: https://github.com/zllrunning/face-parsing.PyTorch.

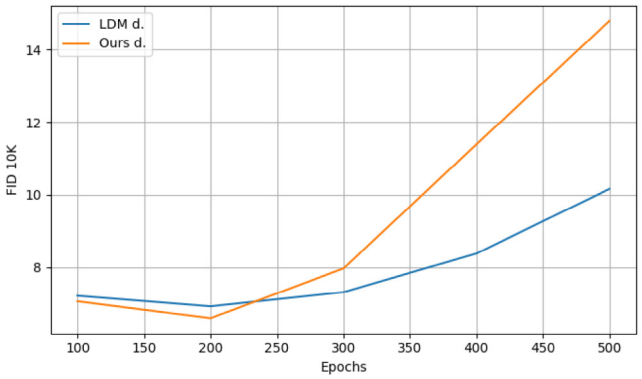


Fig. 2. FID 10K at various training epochs for LDM baseline and our P2 weighted LDM (Ours).

Table 1
Qualitative metrics computed on 50K samples.

Method	FID 50K↓	KID 50K↓
PGGAN[b] (Karras et al., 2017)	8.00	–
DDGAN[a] (Xiao et al., 2021)	7.64	–
UDM[b] (Kim et al., 2021)	7.16	–
WaveDiff[a] (Phung et al., 2022)	5.94	–
LDM[b] (Rombach et al., 2022)	5.11	–
StyleSwin[c] (Zhang et al., 2022)	**3.25**	–
P2 (Choi et al., 2022)	13.20	–
LDM d.	5.88	0.0034
LDM s.	6.60	0.0036
Ours d.	5.42	**0.0032**
Ours s.	6.15	0.0033

[a] Means the corresponding result is taken from Rombach et al. (2022).
[b] Means the corresponding result is taken from Phung et al. (2022).
[c] Means the corresponding result is taken from Zhang et al. (2022).

4.1. Unconditioned image synthesis

The aim of this experiment is to analyze the improvement obtained by introducing P2 Weighting (Choi et al., 2022) into LDMs. The baseline LDM and the P2 weighted model have been trained from scratch. For the latter $\gamma = 0.5$ was used, as suggested in Choi et al. (2022) for CelebA-HQ. The two models have the exact same architecture and are both trained for 500 epochs, the only difference is in the objective function.

Fig. 2 shows the FID performance for different training checkpoints on 10K images generated with deterministic sampling ($\eta = 0.0$). We used linearly spaced checkpoints (epochs 100, 200, 300, 400, 500). P2 improves the baseline FID at each checkpoint by 0.5 points until divergence (around epoch 300), without increasing the model's number of parameters or its sampling time.

Table 1, instead, reports the FID and KID results, also compared to previous works. Here, for each model, 50K samples have been generated using 500 DDIM (Song et al., 2020) steps, both deterministically ($\eta = 0.0$) and stochastically ($\eta = 1.0$). These results show that the proposed LDM, both with and without P2 weighting, obtains lower FIDs compared to most of the existing solutions. The sole approach that manages to attain a lower FID score in unconditioned generation is StyleSwin Zhang et al. (2022). Nevertheless, it is essential to note that our solution has primarily been tailored for conditioned generation, where it achieves state-of-the-art results.

Some qualitative samples generated from the same latent, considering the best training checkpoints for both the P2-weighted model and the baseline LDM are shown in Fig. 3. It is possible to appreciate that, after 100 epochs, the proposed model has already reached satisfying generation stability while the baseline is still trying to converge.

From now on, our solution with the P2 weighting scheme will be employed in all subsequent experiments.

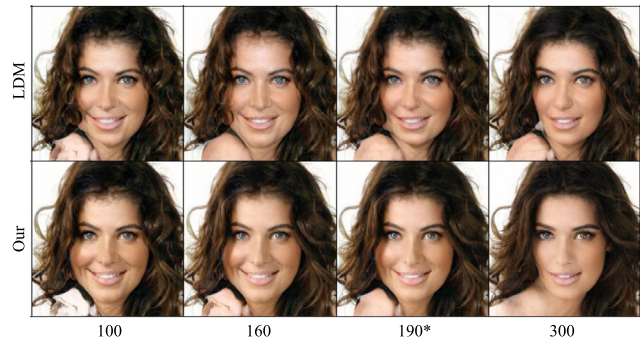


Fig. 3. Unconditioned samples from LDM and the P2 weighted model (Our) at various epochs (*columns*). All the samples are generated from the same latent. * denotes the checkpoint used in the evaluation.

Table 2
FID, KID, accuracy (Acc.), and LPIPS for attributes conditioned synthesis (bottom) and text conditioned synthesis (top).
Source: The text-conditioned results (top) are taken from StyleT2I (Li et al., 2022).

Method	FID ↓	KID ↓	Acc. (%)↑	LPIPS ↑
ControlGAN (Li et al., 2019)	31.38	–	–	–
DAE-GAN (Ruan et al., 2021)	30.74	–	–	–
TediGAN-B (Xia et al., 2021)	15.46	–	–	–
StyleT2I (Li et al., 2022)	17.46	–	–	–
Ours d. (conditioned with $\mathcal{Z}_a$)	**8.83**	**0.0028**	90.53	–
Ours s. (conditioned with $\mathcal{Z}_a$)	9.18	**0.0028**	**91.14**	**0.549**

4.2. Attributes conditioned synthesis

This section shows how a simple attributes encoding can successfully condition DMs via cross-attention, both quantitatively and qualitatively. The attributes encoder is implemented as a simple MLP which maps the set of 40 binary attributes into an embedding of dimension $d = 512$ which is fed to the diffusion model via cross-attention, as detailed in Section 3.2. We did not find any significant previous work on this specific task, the closest family of solutions performs text-to-face generation by conditioning the models on prompts usually formed by the keywords corresponding to the name of the binary attributes, as in StyleT2I (Li et al., 2022).

Table 2 shows that the proposed conditioned model outperforms these solutions in terms of FID. It is important to highlight that the methods introduced by Li et al. (2019), Ruan et al. (2021), Xia et al. (2021) and Li et al. (2022) do not work directly on the raw set of attributes, as in our solution, but perform some pre-processing which may limit the capability of the model to generate samples that reflect the original attributes.

In addition, to assess the conditioning fidelity of the proposed model a ResNet-18 network has been fine-tuned, on the CelebA-HQ training set, to perform a multi-label attribute classification. The classifier obtains a **90.85%** accuracy on the ground truth validation images, while the samples generated by the proposed model (i.e., obtained by conditioning with the set of attributes from the validation set) obtain a classification accuracy of **90.53%**, which confirms the capability of our model to generate samples which reflect the provided attributes.

Fig. 4 shows some samples generated from the same noise. It could be observed that the generated faces share a similar physiognomy, which differs just for the presence or absence of different attributes. This behavior was also observed in Wu and De la Torre (2022).

4.3. Semantic image synthesis

As for the attributes conditioned synthesis discussed in Section 4.2, it is possible to employ cross-attention for injecting semantic information into our model. This time, the encoder backbone is a pruned ResNet-18, with 18 input channels representing binary masks, one for each available part of the face, background excluded. Two different conditions have been tested depending on the layers of the ResNet-18 encoder from which the features are extracted. The first version, $\mathcal{E}_m$, is the full ResNet-18 backbone except for the classification layer, while the second, $\mathcal{E}_{mnp}$, also discards the Global Average Pooling layer, in order to preserve spatially relevant semantic information. The corresponding latent encodings are $\mathcal{Z}_m \in \mathbb{R}^{1\times1\times512}$ and $\mathcal{Z}_{mnp} \in \mathbb{R}^{8\times8\times512}$, which differ just for the spatial size.

Table 3
FID, accuracy (Acc.), mean Intersection over Union (mIoU) and LPIPS for masks conditioned synthesis. Ground-Truth refers to masks parsed from the original validation set.

Method	FID ↓	Acc. (%) ↑	mIoU (%) ↑	LPIPS ↑
Pix2PixHD[a] (Wang et al., 2018)	23.69	95.76	76.12	–
SPADE[a] (Park et al., 2019)	22.43	**95.93**	77.01	–
SEAN[a] (Zhu et al., 2020)	17.66	95.69	75.69	–
GroupDNet[b] (Liu et al., 2019)	25.90	–	76.10	0.365
INADE[b] (Tan et al., 2021)	21.50	–	74.10	0.415
SDM[b] (Wang et al., 2022)	18.80	–	77.00	0.422
pSp (Richardson et al., 2021)	82.93	–	–	–
Collaborative (Huang et al., 2023)	11.86	–	–	–
Ours d. (conditioned with $\mathcal{Z}_m$)	8.49	91.36	75.14	–
Ours d. (conditioned with $\mathcal{Z}_{mnp}$)	8.43	93.77	78.67	–
Ours s. (conditioned with $\mathcal{Z}_m$)	8.41	91.52	75.80	**0.469**
Ours s. (conditioned with $\mathcal{Z}_{mnp}$)	**8.31**	93.91	**79.06**	0.446
Ground truth	0.0	95.51	81.79	–

[a] Denotes results taken from SEAN (Zhu et al., 2020).
[b] Denotes results taken from SDM (Wang et al., 2022).

In Table 3 it is possible to observe that both FID and conditioning fidelity are higher when $\mathcal{E}_{mnp}$ is employed, which demonstrates the capability of the cross-attention mechanism to leverage the information provided by the larger number of embeddings. Moreover, both conditioning methods outperform previous works by Wang et al. (2018), Park et al. (2019), Zhu et al. (2020), Liu et al. (2019), Tan et al. (2021), Wang et al. (2022) in terms of FID. This may be related to the fact that the conditioning is very powerful, and both LDM and P2 optimize human perception. So, the generated images are both faithful (due to conditioning) and realistic (LDM + P2). Furthermore, by using a strong conditioning method, the distribution of generated images, at least in terms of facial shape, faithfully follows the original examples, further improving the FID.

To analyze the ability of the proposed model to adapt to noisy masks, a second experiment has been conducted in which: *(i)* a face parsing model is employed to extract the segmentation masks from the validation set (instead of extracting the mask from the generated image as in the previous experiment); *(ii)* these masks are used to condition the model (instead of using the ground-truth validation masks); *(iii)* 5K samples are generated using the new imperfect masks. The FID obtained for this experiment, **8.20**, is lower than the one obtained with the default masks, indicating a good ability of the proposed model to adapt to imperfect masks. The last row of Table 3 shows the accuracy and mIoU for the masks generated at point *(i)*.

A diversity study has also been conducted using LPIPS (Zhang et al., 2018) as the metric. Ten samples were generated for each segmentation mask in the validation set and were used to compute an intra-class diversity score for each class. The average LPIPS results compared to previous works are reported in Table 3. Note that, LPIPS have been computed only on samples generated using stochastic samplers because of the greater differences that could show up in the samples due to the variance and hence more complex latent. These results show that the proposed solution surpasses the previous methods, both GANs and DMs based, on the quality, and diversity of the generated images. As regards fidelity the proposed solution shows a slightly lower performance in terms of accuracy and a higher result in terms of mIoU.

It can be observed that fidelity and diversity show an inverse behavior depending on the degree of conditioning applied to the model.

Fig. 4. Generation examples obtained from the same latent (i.e., the same initial noise) using a deterministic DDIM. Each sample is conditioned on a random set of attributes chosen from the validation set.

Fig. 5. Samples generated using $\mathcal{E}_{mnp}$ with their relative semantic masks.

On one hand, the model conditioned with $\mathcal{E}_{m}$, uses only $1/64^{th}$ of the embedding compared to $\mathcal{E}_{mnp}$, which results in a less accurate encoding for semantic masks. This is reflected in a higher LPIPS and lower fidelity, expressed by both accuracy and mIoU. On the other hand, using more spatially relevant conditioning allows for improving the results in terms of fidelity while observing a reduction in the capability of the model to diversify the generated images. Fig. 5 shows some samples conditioned with $\mathcal{E}_{mnp}$. Non-centered faces, glasses and hats do not pose any problems.

Lastly, we conducted an additional experiment exploiting the source code and weights available for Collaborative Diffusion (Huang et al., 2023) and pixel2style2pixel (pSp) (Richardson et al., 2021). For this experiment, we generated 5K samples using the semantic masks from the validation set as the condition. The FID computed for both models is highlighted in Table 3. The solution from Richardson et al. (2021) (i.e., pSp) obtains the worst FID among all solutions. This result confirms the performance reported in Ham et al. (2023) and is mainly caused by the limitation of the training-free GAN-based solution.

In addition, in Fig. 6 we show some qualitative samples generated using these two solutions in comparison with samples generated by our model. It is possible to observe that the samples generated by Huang et al. (2023) exhibit some artifacts when the pose of the face is not frontal. This phenomenon is less noticeable in the samples generated by our solution. It is also possible to notice that the images generated by Richardson et al. (2021) in some cases are not faithful to the masks used for conditioning.

Fig. 6. Qualitative comparison between images generated using our model (i.e., Ours), Collaborative Diffusion (Huang et al., 2023) and pSp (Richardson et al., 2021). The three models were conditioned using the relative semantic mask (*top row*).

4.4. Multi condition image synthesis

As explained in Section 3.2, it is possible to exploit a property of cross-attention to inject two or more different sets of feature embeddings into any model, before providing them as a condition into the transformer.

Table 4
Comparison across various metrics for different Condition Encoders. FID and KID metrics are for sample quality, mask accuracy, attributes accuracy and mIoU are for correspondence and LPIPS for diversity. All the metrics are evaluated on 5K samples against their respective 5K images from the validation set, except for LPIPS which is computed on sets of 10 images for each of the 5K validation images and features. *(top)* attributes conditioning. *(middle)* masks conditioning. *(bottom)* multi-conditioning.

Condition Enc.	FID ↓	KID ↓	Attr. Acc. (%) ↑	Masks Acc. (%) ↑	mIoU (%) ↑	LPIPS ↑
$\mathcal{E}_a$ d.	8.33	0.0028	90.53	–	–	–
$\mathcal{E}_a$ s.	9.18	0.0028	**91.14**	–	–	**0.549**
$\mathcal{E}_m$ d.	8.49	0.0024	–	91.36	75.14	–
$\mathcal{E}_m$ s.	8.41	0.0023	–	91.52	75.80	0.469
$\mathcal{E}_{mnp}$ d.	8.43	0.0025	–	93.77	78.67	–
$\mathcal{E}_{mnp}$ s.	**8.31**	**0.0021**	–	93.91	79.06	0.446
$\mathcal{E}_{mc}$ d.	8.39	0.0024	90.27	93.90	78.68	–
$\mathcal{E}_{mc}$ s.	8.39	0.0022	90.19	**94.06**	**79.20**	0.432
Ground truth	0.0	0.0	90.85	95.51	81.79	–

Fig. 7. Qualitative samples generated from our model conditioned using both attributes and masks $\mathcal{E}_{mc}$ with, in the bottom-right, the validation set images from which the segmentation mask and attributes have been taken.

Fig. 8. Qualitative samples showing the ability of our model to diversify its generated samples. *(left)* the reference image from the validation set. *(top row)* the images generated when conditioning our model on the reference image's attributes. *(central row)* the images generated when conditioning our model on the reference image's semantic mask. *(bottom row)* the results obtained with our multi-condition encoder, using both attributes and semantic masks.

In our experiments the attributes embedding, $\mathcal{Z}_a \in \mathbb{R}^{1\times512}$, and the flattened version of the mask embedding, $\mathcal{Z}_{mnp} \in \mathbb{R}^{(8\cdot8)\times512}$ are combined. This results in a multi-condition embedding $\mathcal{Z}_{mc} \in \mathbb{R}^{65\times512}$ obtained via concatenation.

Table 4 reports the results obtained using the multi-conditioned model against the attributes-conditioned and the mask-conditioned models. It is worth highlighting that, the high fidelity observed on both attributes and masks results in lower FID and LPIPS, compared to single-conditioned models.

Fig. 7 shows some multi-conditioned examples generated by exploiting the segmentation masks and attributes of a face from the validation set.

In order to compare the results obtained by the proposed model while using different conditioning, Fig. 8 shows the images generated

Fig. 9. Unconditioned samples on DeepFashion.

Fig. 10. Multi-conditioned samples on DeepFashion. The top row shows validation images, the bottom row contains the samples generated on the corresponding set of masks and attributes.

with the different modalities starting from the same attributes and mask.

5. Swapping components between masks

To further explore the ability of the proposed model to adapt to incoherent masks, we conducted an experiment swapping some components (i.e., mask channels) between pairs of segmentation masks and used the resulting mixed mask as conditioning. From Fig. 11 it is possible to observe that the proposed solution can generate samples that highly correspond to the relative mask but, at the same time, it tries to correct those face components that are no longer coherent with the rest of the mask. It is worth noting that this experiment was conducted considering two differently oriented faces and no pre-processing was performed on the mixed masks.

6. Additional experimental results on DeepFashion

This section reports on some additional experiments performed on DeepFashion (Liu et al., 2016), a dataset that, differently from CelebA-HQ does not focus on human faces. This dataset has been selected because it provides semantic masks and attributes, allowing us to directly apply the multi-conditioning. These experiments have been performed to analyze the capability of the model to adapt to other domains. In addition, DeepFashion's images are rectangular, hence it is also possible to experiment with different ratios from the default square images.

Two experiments have been conducted, one for unconditioned generation and one for multi-conditioned generation using both attributes and semantic masks. The unconditioned and multi-conditioned models have been trained for 150 epochs on the subset of the dataset that

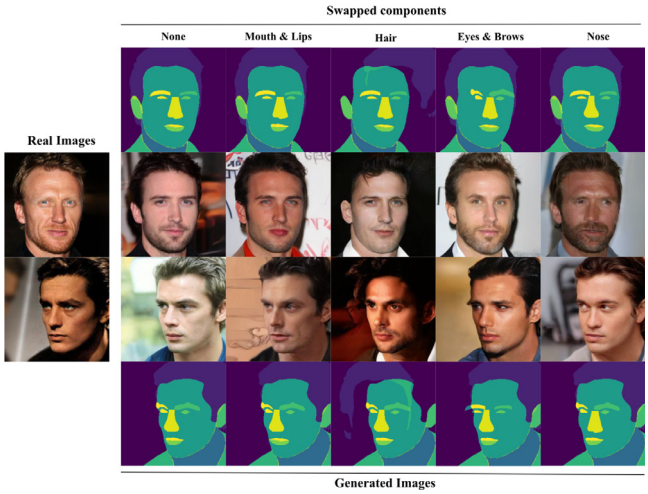


Fig. 11. Samples generated by conditioning on incoherent segmentation masks, resulting from mixing differently oriented masks' components. On the left, are the real images from the validation set from which the mask was taken. On the right, the swapped masks are shown (top and bottom), and their relative generated samples just above or below.

provides both semantic masks and attributes. The unconditioned model has been used to generate 50K samples while the multi-conditioned model was used to generate one sample for each *attribute-mask* pair. Fig. 9 shows some unconditioned samples, while Fig. 10 shows multi-conditioned samples against their corresponding original images from which the conditioning was taken. The FID score was determined using the same procedure as CelebA-HQ. Specifically, for unconditioned

generation, the FID score was calculated by comparing it to the entire dataset, resulting in a score of **4.15**. In the case of multi-conditioned generation, the FID score was computed by comparing it to the validation set, which constitutes 10% of the entire dataset, yielding a score of **17.32**. This disparity between the unconditioned and conditioned FID scores is attributed to the limited representation of the validation set, which consists of only 1200 images. Typically, FID scores are calculated based on sets containing at least 10K samples, often 50K.

7. Conclusion

In this paper, a solution for face generation using diffusion models conditioned by both attributes and masks is introduced. The proposed solution has been trained by re-weighting the loss terms of an LDM in a perception-prioritized fashion showing that this allows achieving a higher quality of the generated samples. In the conditioned generation, first attributes and segmentation masks are considered and studied separately. Then, a novel approach to multi-condition a generative model that exploits cross-attention to join the two conditions (i.e., attributes and semantic masks) is introduced. Both the single-conditioned and multi-conditioned models have been evaluated on a various range of metrics to assess quality, fidelity and diversity on CelebA-HQ. An evaluation on a different dataset, namely DeepFashion, has also been conducted in order to show the capability of our solution to generalize to different domains. We plan to investigate and analyze more efficient techniques for encoding the different conditions as well as extending the proposed solution to consider and combine more conditions.

CRediT authorship contribution statement

Giuseppe Lisanti: Methodology, Resources, Supervision, Writing – original draft, Writing – review & editing. **Nico Giambi:** Investigation, Methodology, Software, Writing – original draft, Writing – review & editing.

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.

Data availability

Experiments were conducted on publicly available datasets.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2024.104026.

References

Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., 2018. Demystifying mmd gans. arXiv preprint arXiv:1801.01401.

Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S., 2022. Perception prioritized training of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11472–11481.

Choi, Y., Uh, Y., Yoo, J., Ha, J.-W., 2020. Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8188–8197.

Dhariwal, P., Nichol, A., 2021. Diffusion models beat gans on image synthesis. Adv. Neural Inf. Process. Syst. 34, 8780–8794.

Gao, Y., Wei, F., Bao, J., Gu, S., Chen, D., Wen, F., Lian, Z., 2021. High-fidelity and arbitrary face editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16115–16124.

Ham, C., Hays, J., Lu, J., Singh, K.K., Zhang, Z., Hinz, T., 2023. Modulating pretrained diffusion models for multimodal image synthesis. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inf. Process. Syst. 30.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. 33, 6840–6851.

Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T., 2022. Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. 23 (47), 1–33.

Hou, X., Zhang, X., Liang, H., Shen, L., Lai, Z., Wan, J., 2022. Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing. Neural Netw. 145, 209–220.

Huang, Z., Chan, K.C., Jiang, Y., Liu, Z., 2023. Collaborative diffusion for multi-modal face generation and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6080–6090.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

Kim, D., Shin, S., Song, K., Kang, W., Moon, I.-C., 2021. Score matching model for unbounded data score. arXiv preprint arXiv:2106.05527, 7.

Kingma, D., Salimans, T., Poole, B., Ho, J., 2021. Variational diffusion models. Adv. Neural Inf. Process. Syst. 34, 21696–21707.

Lee, C.-H., Liu, Z., Wu, L., Luo, P., 2020. Maskgan: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5549–5558.

Li, Z., Min, M.R., Li, K., Xu, C., 2022. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18197–18207.

Li, B., Qi, X., Lukasiewicz, T., Torr, P., 2019. Controllable text-to-image generation. Adv. Neural Inf. Process. Syst. 32.

Li, Z., Zhang, S., Zhang, Z., Meng, Q., Liu, Q., Zhou, H., 2023. Attention guided domain alignment for conditional face image generation. Comput. Vis. Image Underst. 234, 103740. http://dx.doi.org/10.1016/j.cviu.2023.103740.

Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X., 2016. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. CVPR.

Liu, X., Park, D.H., Azadi, S., Zhang, G., Chopikyan, A., Hu, Y., Shi, H., Rohrbach, A., Darrell, T., 2023. More control for free! image synthesis with semantic diffusion guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 289–299.

Liu, X., Yin, G., Shao, J., Wang, X., et al., 2019. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. Adv. Neural Inf. Process. Syst. 32.

Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X., 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453.

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2337–2346.

Phung, H., Dao, Q., Tran, A., 2022. Wavelet diffusion models are fast and scalable image generators. arXiv preprint arXiv:2211.16152.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D., 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. pp. 2287–2296.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Ruan, S., Zhang, Y., Zhang, K., Fan, Y., Tang, F., Liu, Q., Chen, E., 2021. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13960–13969.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. PMLR, pp. 2256–2265.

Song, J., Meng, C., Ermon, S., 2020. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.

Tan, Z., Chai, M., Chen, D., Liao, J., Chu, Q., Liu, B., Hua, G., Yu, N., 2021. Diverse semantic image synthesis via probability distribution modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7962–7971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., Li, H., 2022. Semantic image synthesis via diffusion models. arXiv preprint arXiv:2207.00050.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8798–8807.

Wang, Y., Qi, L., Chen, Y.-C., Zhang, X., Jia, J., 2021. Image synthesis via semantic composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13749–13758.

Wu, C.H., De la Torre, F., 2022. Unifying diffusion models' latent space, with applications to CycleDiffusion and guidance. arXiv preprint arXiv:2210.05559.

Xia, W., Yang, Y., Xue, J.-H., Wu, B., 2021. Tedigan: Text-guided diverse face image generation and manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2256–2265.

Xiao, Z., Kreis, K., Vahdat, A., 2021. Tackling the generative learning trilemma with denoising diffusion GANs. arXiv preprint arXiv:2112.07804.

Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J., 2023. Freedom: Training-free energy-guided conditional diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23174–23184.

Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., Guo, B., 2022. Styleswin: Transformer-based gan for high-resolution image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11304–11314.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847.

Zhu, P., Abdal, R., Qin, Y., Wonka, P., 2020. Sean: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5104–5113.

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E., 2017. Toward multimodal image-to-image translation. Adv. Neural Inf. Process. Syst. 30.