FICE: Text-Conditioned Fashion-Image Editing With Guided GAN Inversion

Martin Pernuš[a], Clinton Fookes[b], Vitomir Štruc[a,*], Simon Dobrišek[a]

[a] *Faculty of Electrical Engineering, University of Ljubljana, Trzaska 25, Ljubljana, 1000, Slovenia, Slovenia*
[b] *School of Electrical Engineering & Robotics, Queensland University of Technology, 2 George St, Brisbane, 4000, Queensland, Australia*

Abstract

Fashion-image editing is a challenging computer-vision task where the goal is to incorporate selected apparel into a given input image. Most existing techniques, known as Virtual Try-On methods, deal with this task by first selecting an example image of the desired apparel and then transferring the clothing onto the target person. Conversely, in this paper, we consider editing fashion images with text descriptions. Such an approach has several advantages over example-based virtual try-on techniques: (i) it does not require an image of the target fashion item, and (ii) it allows the expression of a wide variety of visual concepts through the use of natural language. Existing image-editing methods that work with language inputs are heavily constrained by their requirement for training sets with rich attribute annotations or they are only able to handle simple text descriptions. We address these constraints by proposing a novel text-conditioned editing model called FICE (Fashion Image CLIP Editing) that is capable of handling a wide variety of diverse text descriptions to guide the editing procedure. Specifically, with FICE, we extend the common GAN-inversion process by including semantic, pose-related, and image-level constraints when generating images. We leverage the capabilities of the CLIP model to enforce the text-provided semantics, due to its impressive image–text association capabilities. We furthermore propose a latent-code regularization technique that provides the means to better control the fidelity of the synthesized images. We validate the FICE through rigorous experiments on a combination of VITON images and Fashion-Gen text descriptions and in comparison with several state-of-the-art, text-conditioned, image-editing approaches. Experimental results demonstrate that the FICE generates very realistic fashion images and leads to better editing than existing, competing approaches. The source code is publicly available from: https://github.com/MartinPernus/FICE.

Keywords:
Generative adversarial networks, Image editing, Deep learning, Multimodality

[*] Corresponding author
Email addresses: martin.pernus@fe.uni-lj.si (Martin Pernuš), c.fookes@qut.edu.au (Clinton Fookes), vitomir.struc@fe.uni-lj.si (Vitomir Štruc), simon.dobrisek@fe.uni-lj.si (Simon Dobrišek)

1. Introduction

Fashion-image editing refers to the task of changing the appearance of a person in a given image by incorporating a desired fashion item (e.g., different apparel) in a realistic and visually convincing manner. Successful applications of such algorithms enable users to visualize and virtually try-on items of clothing from the comfort of their homes. This functionality has the potential to enable easier sales of online apparel, reduce the costs for retailers, and reduce the environmental footprint of the fashion industry by minimizing returns [1]. As a result, research has been directed towards fashion-image manipulation (or Virtual Try-On, VTON) techniques that deliver convincing photorealistic editing results [2–5]. Furthermore, the methodologies and techniques developed in this sphere can have broader applicability; they can be adapted to tackle other image-editing challenges, such as interior design for homes [6], cosmetic simulations [7], or historical restorations [8]. Recent works in this field include various methods for 3D clothed-human reconstruction and dynamic garment animation [9–16], which further enhance the realism and functionality of VTON systems.

VTON solutions have achieved great success in synthesizing photorealistic fashion images[1] by building on advances made in convolutional neural networks and adversarial training objectives [17–19]. Most of the existing techniques in this area condition their editing models on example images of the target clothing, which is typically warped and stitched onto the given input image. Considerably less attention has been given to text-conditioned, fashion-image editing, despite the fact that such methods represent an attractive alternative to example-based editing techniques. Leveraging natural language descriptions allows for a more intuitive mechanism to drive the editing process, facilitating an easier interface for users to express their preferences. Furthermore, users can articulate nuanced design specifications through natural language inputs, without having to provide visual garment representations. While, to the best of our knowledge, only a modest amount of work has been conducted on this topic so far, existing text-conditioned methods are commonly limited to very basic descriptions, mostly due to the small size of the suitable training datasets that are publicly available [20]. To mitigate these problems, some text-conditioned fashion works were proposed to parse the input text into closed sets of categories [21] for easier text processing, simplifying the task into a more basic, categorical problem.

Meanwhile, various image–text association models have emerged. These models are trained on hundreds of millions of image–text pairs [22] and represent powerful tools for associating visual data and language descriptions [23, 24]. As a result, they have been successfully deployed for text-conditioned image editing in combination with recent state-of-the-art generative adversarial networks (GANs) [25]. Such solutions typically first embed the given input image into the latent space of a pre-trained GAN model through a process referred to as GAN inversion [26], and then perform text-conditioned manipulations in the latent

[1] While different definitions of the term *fashion image* can be found in the literature, we define it in this paper as an image of a subject that is focused on accentuating fashion garments against a clean, uncluttered background.

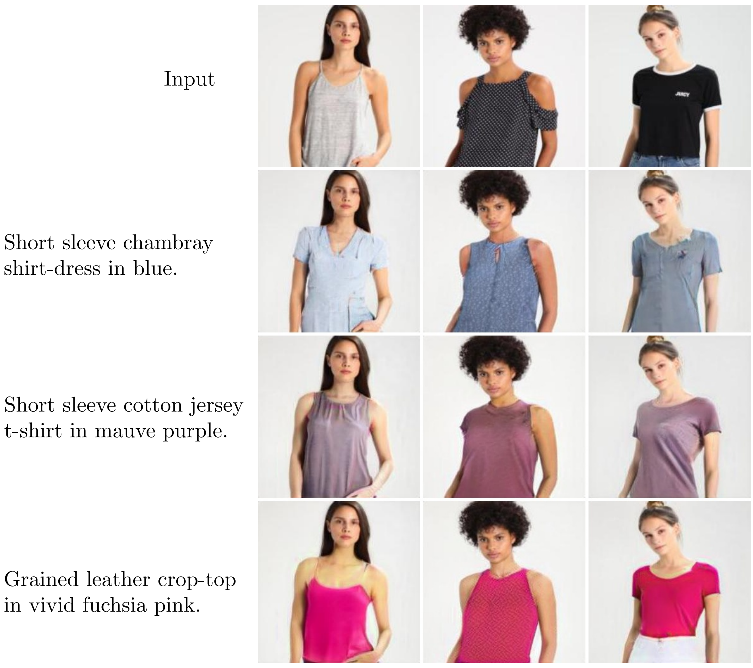

Input
Short sleeve chambray shirt-dress in blue.
Short sleeve cotton jersey t-shirt in mauve purple.
Grained leather crop-top in vivid fuchsia pink.

Figure 1: **Fashion-image editing with language-based inputs.** In this paper we propose FICE (**F**ashion **I**mage **CLIP** **E**diting), a text-conditioned image-editing model, capable of handling a wide variety of text inputs with the goal of manipulating fashion images toward the desired target appearance.

space that eventually lead to semantically meaningful changes in the corresponding output images [25, 27, 28]. While general-purpose, text-conditioned, GAN-based editing techniques have shown success in various settings, a straightforward application to the fashion domain is challenging and, more importantly, does not guarantee optimal editing results. This is due to the inherent reconstruction–editability trade-off [29] of such techniques, which typically results in a loss of identity information as well as pose changes when inverting an image into a GAN latent code. Furthermore, despite the recent advances in disentangled editing in the GAN latent space [25, 30], such methods are still problematic to use in the context of text-conditioned editing due to the high sensitivity to hyperparameter choices [31].

In this paper we address these open challenges through the introduction of FICE (Fashion Image CLIP Editing) – a novel text-conditioned, image-editing approach tailored towards fashion images. FICE introduces a novel GAN-inversion framework, allowing for the integration of capabilities that enable text-conditioned fashion-image editing. This stands in contrast to traditional VTON methods that rely heavily on categorical attributes for editing, which often necessitates extensive data curation. Our approach simplifies the editing process, enabling users to modify images using only text inputs and thereby removing the complexities associated with the management and curation of additional categorical or visual data. This streamlines the workflow, makes fashion image manipulation more accessible and reduces the barriers typically associated with digital fashion editing.

To facilitate the editing of fashion images with FICE, we propose an *iterative GAN-inversion procedure* that utilizes several constraints when optimizing for the latent code with the desired target semantics, i.e.: (*i*) a *pose-preservation constraint* that ensures the pose of the subject in the image is not altered during the editing process, (*ii*) a *composition constraint* that uses a segmentation model (i.e., a body parser) to identify regions (head and garment areas) in the input image to preserve and/or alter, and (*iii*) a *semantic-content constraint* that enforces the semantics expressed in the provided text descriptions. We use various differentiable deep-learning models to implement the constraints and leverage the CLIP model, a recent state-of-the-art image–text association approach, to enforce the desired semantics. Furthermore, we propose a latent-code regularization objective to ensure more realistic editing results. Finally, we also utilize an image-stitching step to combine the relevant image regions from the original and edited images in the final overall result. It is worth noting that with the proposed FICE model, we are the first to introduce an extended GAN-inversion approach that allows for text-conditioned image editing in the fashion domain.

To demonstrate the capabilities of FICE, we perform rigorous experiments on images from the VITON and MPV image datasets [2, 32], combined with text descriptions from the Fashion-Gen dataset [33]. We compare FICE to several general text-conditioned GAN-based editing methods, diffusion-based techniques as well as its closest competitor FashionGAN [20] and show that the proposed approach leads to superior editing results for fashion images. A few of these results can be seen in Fig. 1 for three different text descriptions. Our research leads to the following main contributions that are presented in this paper:

- We propose FICE, a model for text-conditioned fashion-image editing, which can be used with a wide variety of textual inputs and leads to realistic and visually convincing editing results. To the best of our knowledge, FICE is the first GAN-inversion approach presented in the literature that allows for the incorporation of text-provided semantics when embedding an existing fashion image into the GAN latent space.
- We introduce a novel GAN-inversion approach that incorporates constraints relevant for VTON, such as pose-preservation and image-composition constraints, as well as a special regularization technique that minimizes the generation of images outside the GAN-learned distribution.
- Through quantitative and qualitative evaluations, we show the benefits of the text-based editing of fashion images and demonstrate that FICE convincingly outperforms competing, state-of-the-art, text-based editing techniques.

2. Related Work

In this section we review relevant prior work and discuss existing research on (*i*) generative adversarial networks, (*ii*) text-conditioned image generation and editing, (*iii*) GAN-inversion techniques, and (*iv*) the use of computer vision in fashion. The goal of the section is to provide the necessary background for our work. A more comprehensive coverage of these topics can be found in some of recent surveys, e.g., [26, 34, 35].

2.1. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [19] have, in recent years, become the *de-facto* method for unconditional image synthesis, leading to convincing, high-resolution image synthesis and reasonable training times with consumer-grade hardware. DCGAN [36] introduced convolutional GANs and provided architectural pointers to achieve successful GAN convergence. ProGAN [37] was the first GAN model that achieved megapixel-sized images thanks to a progressive learning scheme. StyleGAN [38] introduced a non-linear mapping of the latent space and an alternative generator design, inspired by the style-transfer literature. StyleGAN2 [39] further adjusted the generator architecture to remove the frequent droplet artifacts and regularized the training with path-length regularization. StyleGAN2-ADA [40] proposed several augmentation techniques to enable learning a high-quality GAN with limited training data. Additional details relating to modern GAN architectures, training constraints, and loss functions can be found in a recent survey [41].

2.2. Text-Conditioned Image Generation and Editing

Text-conditioned, image-generation models are focused on generating realistic images that match the semantics of the provided text descriptions. Conversely, corresponding *editing techniques* try to realistically manipulate images in a way that preserves the image characteristics that are irrelevant to the text description. Thus, text-conditioned image editing aims to alter only the semantic content that is expressed in the text description, while preserving all other parts of the data.

Image Generation. The seminal work of Reed *et al.* [42] proposed a text-conditioned GAN model by feeding the text information to both the generator and discriminator of the GAN design. StackGAN [43] and StackGAN++ [44] proposed stacked generators where the resolution of the generated images increased progressively with each generator in the stack. AttnGAN [45] proposed an attention mechanism to attend to relevant words on which to condition the image-generation process. MirrorGAN [46] proposed a cyclic GAN architecture that regularized the generated image by enforcing correct (re)descriptions.

While these early models discussed above already provided for competitive performance, more recent text-conditioned generative models are several orders of magnitude larger in size and are trained on several orders of magnitude larger datasets. DALL-E [24], for example, trained a 12-billion parameter autoregressive transformer on 250 million image–text pairs, and considerably outperformed previous models in the considered zero-shot evaluation experiments. The model was further improved in DALL-E 2 [47] through the use of diffusion techniques. Imagen [48] proposed to increase the size of the text encoder to achieve better results in terms of image fidelity and image–text alignment.

Image Editing. An early approach to text-conditioned image-editing methods was described by Dong *et al.* in [49]. Here, the authors proposed a conditional encoder-decoder GAN model. Nam *et al.* [50]

proposed so-called word-subset local discriminators that enabled fine-grained image editing. ManiGAN [51] proposed a modified strategy for merging image and text representations while adding a detail-correction module for enhanced image quality.

Another notable group of methods performs image editing by first converting a given image into the latent code of some pre-trained GAN, in a process known as GAN inversion, then performing various latent-code manipulations to achieve the desired edits. InterFaceGAN [27], for example, identified directions in the latent space of StyleGAN2 that corresponded to specific semantic changes (given by binary attribute labels) in the corresponding output image. Image2StyleGAN [52, 53] performed several face-image edits using GAN inversion, and StyleCLIP [25] proposed different methods for text-guided image editing, where the general idea is based on combining the generative capabilities of StyleGAN with the image–text matching capabilities of the CLIP model [22]. TediGAN [54] introduced a control mechanism based on style mixing in StyleGAN to achieve the desired semantics in face images driven by text descriptions. [28] introduced a lightweight adapter layer and a refinement module in the StyleGAN2 latent space to achieve text-conditioned image manipulation.

2.3. GAN Inversion

As can be seen from the literature review presented above, many of the existing editing techniques deal with the process of GAN inversion to retrieve the image's latent code for editing. How to conduct the GAN inversion is a key consideration with these techniques that has an impact on the final editing capabilities. Richardson *et al.* [55], for example, proposed an encoder, called pSp, to project images into the StyleGAN latent space, and demonstrated its feasibility through several image-to-image translation tasks. E4e [29] adjusted the pSp model so that the latent codes follow a similar distribution to the original StyleGAN latent codes and performed image edits with several latent-code manipulation techniques. ReStyle [56] presented an iterative procedure to obtain the latent code, while HyperStyle [57] adjusted the StyleGAN generator weights on a per-sample basis to achieve better image reconstructions. Additional GAN-inversion methods can be found in the recent survey [26].

Similar to the techniques discussed above, the proposed FICE model also relies on GAN inversion to perform fashion-image editing. However, in contrast to alternative techniques, our model is geared towards the characteristics of fashion images and exploits an iterative inversion process that explicitly considers pose preservation and image-stitching constraints in addition to the targeted semantics to ensure both the desired garment appearance as well as identity preservation.

2.4. Computer Vision in Fashion

A considerable cross-section of fashion-related computer-vision research focuses on VTON technology, where, given an image of a person and some target garment, the goal is to realistically fit the garment, while preserving the original person's pose and appearance. Numerous studies have been conducted on various aspects concerning 3D human and garment modelling [9–16, 58–62], as well as the image-based fitting of fashion items onto subjects [2–5, 63–65].

Several works introduced text modality when processing fashion images. In the field of image generation and editing, FashionGAN edited images conditioned on text inputs, segmentation masks and several image-specific categorical attributes using encoder–decoder GANs. Fashion-Gen [33] introduced a dataset of image–text pairs and experimented with unconditional and text-conditioned image generation. Recently, Text2Human [21] proposed the idea of generating human images based on a description of the shape and texture of the clothing. With this approach, the text-encoding method mapped the input text into a number of closed sets of categories, which limits the linguistic expressiveness of the input text. Another popular task is fashion-image editing and generation, e.g., [66, 67].

Unlike the methods presented above, FICE uses text as the only condition/input for the processing of fashion images. Furthermore, the model is not focused on image generation, but rather on image editing, i.e., on re-dressing people, similar to VTON methods, but using only text descriptions as the input instead of image examples. Finally, unlike some methods, that parse the text into a closed set of categorical attributes, FICE relies on CLIP [22], an image–text association model, trained on 400 million image–text pairs, which allows the proposed approach to process a wider set of linguistic concepts than alternative solutions from the literature.

3. Methodology

The main contribution of this paper is a novel text-conditioned model for fashion-image editing named FICE (**F**ashion **I**mage **C**LIP **E**diting), whose key component is a new GAN-inversion procedure that allows for the incorporation of natural language descriptions into the editing procedure. In this section we provide an in-depth description of the proposed model and elaborate on its characteristics.

3.1. Problem Formulation and FICE Overview

The aim of FICE is to edit the given fashion image $I \in \mathcal{R}^{3\times n\times n}$ in accordance with some (appearance-related) text description t and to synthesize a corresponding output image $I_f \in \mathcal{R}^{3\times n\times n}$ that adheres as closely as possible to the semantics expressed in t. Here, the synthesis process needs to meet the following criteria: (1) the synthesized output image I_f should preserve the pose, identity and other appearance characteristics of the subject in I, (2) the editing process should be local and only affect the desired fashion items (e.g., apparel), while leaving other parts of I unchanged, (3) clothing appearances, encoded in t, need to be realistically and seamlessly integrated into I_f, taking the initial pose and body shape into account, and (4) a wide variety of textures and clothing designs need to be supported. Thus, the goal of FICE is to implement an image-to-image mapping ψ_t conditioned on t, i.e.

$$\psi_t : I \rightarrow I_f \in \mathcal{R}^{3\times n\times n}, \tag{1}$$

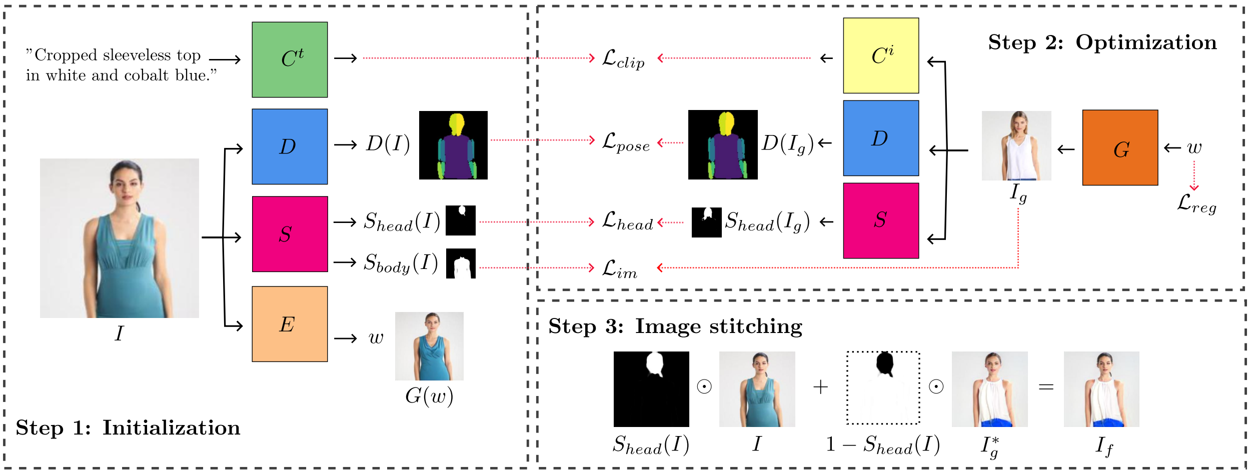


Figure 2: **High-level overview of the proposed FICE editing model**. The model uses a constrained GAN-inversion procedure over the latent space of the pre-trained StyleGANv2 model G to incorporate the semantics expressed in the provided text description into the output image I_g. Here, the optimization-based GAN inversion is performed by enforcing: (i) the presence of semantic content (determined by the image–text association model C, consisting of the image encoder C^i and the text encoder C^t), (ii) pose preservation (ensured by the pose parser D), (iii) image consistency (through the segmentation model S), and (iv) latent-code regularization. An image-stitching step is also incorporated into FICE to improve the image's fidelity and further enhance the model's identity-preservation capabilities.

under the constraints discussed above. As illustrated in Fig. 2, FICE defines the mapping through a three-stage procedure that consists of: (i) an *Initialization* stage, (ii) a *Constrained GAN-inversion* stage, and (iii) an *Image Stitching* stage. A high-level summary of the three key steps is given below:

- **Initialization.** In the first step, the model initializes a latent code w based on the image input I using a GAN-inversion encoder E, which serves as a first approximation of the targeted latent code. This initial code approximates the original appearance of the input image, when interpreted through the pre-trained GAN generator G, i.e., $I \approx G(w)$, as shown on the left-hand side of Fig. 2. Additionally, a dense-pose representation and segmentation masks corresponding to different body parts are also computed from I during this step.

- **Constrained GAN Inversion.** Next, the initial latent code w is further optimized using a constrained, optimization-based GAN-inversion technique. During each step of the optimization procedure, the latent code w is fed to the generator G to synthesize an intermediate image $I_g = G(w)$. An optimization objective (loss) $\mathcal{L}$ is then defined over I_g to drive the GAN inversion that ensures: (i) the semantics defined in t are present in I_g, (ii) the poses in I and I_g are the same, (iii) the editing procedure appears natural, and that (iv) the optimized latent code lies in a well-defined part of the

GAN latent space. The result of the fixed-step optimization procedure is a latent code w^*,

$$w^* = \arg\min_{w}\{\mathcal{L}(I, G(w), t)\}, \tag{2}$$

that corresponds to the final optimized output image $I_g^* = G(w^*)$. A number of auxiliary differentiable models are utilized to facilitate the optimization procedure – see Step 2 in Fig. 2. Details on the models are provided in the following sections.

- **Image Stitching.** In the final stage, a simple image-composition process is utilized to combine the optimized GAN synthesized/optimized image I_g^* with the original image I to ensure identity preservation and to produce the final output image I_f.

3.2. FICE Structure

FICE relies on several differentiable models to infer information about the high-level characteristics of the given input image for the constrained GAN-inversion procedure:

- **Generator** (G). The key component of FICE is the pre-trained GAN generator G, which is responsible for synthesizing the fashion images based on the provided latent code w and defines the characteristics of the generated data. To this end, we select a state-of-the-art, unconditional, image-generation model, i.e., StyleGANv2 [39], which is capable of generating high-quality images with a resolution up to 1024 × 1024 px, and train it for the generation of fashion images.

- **CLIP** (C). The Contrastive Language–Image Pre-training model (or CLIP for short) [22] is a neural network trained for pairing images and text. It consists of a separate image (C^i) and text encoder (C^t) that produce embeddings for the pairing. The model has been trained on 400 million (image, text) pairs and was demonstrated to have a strong zero-shot performance on various downstream tasks. Inspired by these zero-shot capabilities, we use CLIP to provide semantic knowledge to our model and edit the image according to the provided text description t.

- **DensePose** (D). A critical aspect for realistic results, when editing fashion images, is the pose preservation. To preserve the complete body structure and pose information, we employ a sophisticated parsing model, i.e., DensePose [68], capable of parsing individual human body parts from the given input image. The model provides for a comprehensive and dense pose description that is utilized in FICE to ensure that the subject's pose in the original I and optimized images I_g^* match.

- **Segmentation Model** (S). A segmentation model is used to identify the image regions for either alteration or preservation. We chose DeepLabv3 [69] as our segmentation model architecture due to its robust performance and efficient computation. Additionally, the goal of the model is to ensure consistent image characteristics, so the final stitched image is artifact-free, photo-realistic and visually convincing.

- **Encoder** (E). The last model needed for image editing with FICE is a GAN-inversion encoder that computes the initial latent-space embedding w from the input image I. We select the E4e (Encoder for editing [29]) encoder for its ability to predict latent codes that adhere to the StyleGAN latent-code distribution. This property is crucial to ensuring that the optimization procedure maintains the latent codes within the range of the trained latent distribution, thus also preserving the integrity of the resulting image distribution.

3.3. Constrained GAN Inversion

The main component of FICE is a novel *constrained GAN-inversion technique* that, given the input image I and text description t, optimizes for the latent code w^* that fits a number of predefined constraints – see Eq. (2). After initializing the latent code w and computing the corresponding output image $I_g = G(w)$, the goal of the optimization-based inversion process is to adjust the latent code to best match the text description while preserving various appearance characteristics of the input image. Thus, given a latent code w and the associated image $I_g = G(w)$, we define the several optimization objectives for FICE, which constrain the inversion process, as detailed below.

Semantic Content. To ensure that the appearance-related semantic content, expressed in the text description t, is present in the generated image I_g, we implement a CLIP-based optimization objective/constrain,

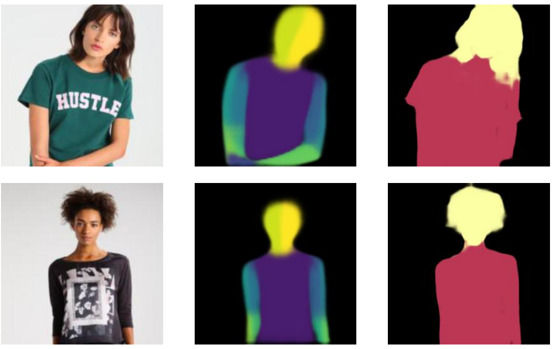
$$\mathcal{L}_{clip} = 1 - \cos(C^i(I_g), C^t(t)), \tag{3}$$

where $\cos(\cdot)$ represents the cosine similarity, and C^i and C^t represent the image and text encoder of the CLIP model, respectively. Note that each encoder returns a L2-normalized embedding, making the angular difference the most natural choice. The objective/constrain penalizes angular differences between the image and text embeddings and thus promotes the presence of the semantics from t in the generated image I_g, where $I_g = G(w)$.

Pose Preservation. To achieve perceptually convincing editing results, it is critical to preserve the position of all body parts from the input image in the edited output. Because our goal is a detailed and accurate pose preservation, where individual body parts retain their size and shape, regardless of the overlaid clothing, we utilize the powerful DensePose model D, which predicts the position and shape of individual body parts in a *clothing-agnostic* manner. The pose-preservation constraint is, therefore, defined through the following loss,

$$\mathcal{L}_{pose} = \frac{1}{N_D} \sum_{j=1}^{N_D} ||D_j(I) - D_j(I_g)||_2^2, \tag{4}$$

where N_D is the number of parsed body parts, D is the pose-parsing model and $D_j(\cdot)$ denotes the parsed mask of the j-th body part. A couple of example outputs produced by the pose parser D are shown in Fig. 3. The aim of the pose constraint is to encourage the optimization procedure to produce inverted latent codes that correspond to images with the same pose as the original input image.



Input Parsed by D Parsed by S

Figure 3: **Sample images parsed with D and S.** The DensePose model D generates a parsed body representation consisting of 24 body parts in a clothing-agnostic manner. The segmentation model S, on the other hand, parses the input image into three classes: head (with hair), body (with clothing) and background.

Latent-Code Regularization. FICE operates in the extended latent vector space $\mathcal{W}^+$ of the pre-trained StyleGAN generator, which is commonly used with the GAN-inversion editing techniques from the literature [53]. The latent code $w \in \mathcal{W}^+$ consists of several latent codes, each impacting on an individual convolutional layer in the StyleGAN generator network G. The number of individual latent codes depends on the resolution of the generator network.

It is important to note that $\mathcal{W}^+$ is an artificial extension of the original latent space $\mathcal{W}$. As was shown in [52, 53], the extended latent space can be used to increase the expressiveness of the StyleGAN generator by having independent latent codes instead of the same code for each convolutional layer, as is the case in the original StyleGAN2. However, such an extension of the latent space does not guarantee the synthesis of an image distribution that corresponds to the learned GAN image distribution [29].

We propose a simple regularization mechanism to mitigate the distribution discrepancy that arises from the mismatch between the extended latent codes. This involves minimizing the distance between the coarsest latent code to other latent codes. We chose the coarsest latent code as the target latent code, since, as it operates on the lowest resolution, it has the most impact on the overall appearance of the image. Concretely, given an extended latent code $w = \{w^1, w^2, ..., w^{N_w}\}$, where N_w denotes the number of individual codes, we define the following loss function,

$$\mathcal{L}_{reg} = \frac{1}{N_w - 1} \sum_{j=2}^{N_w} ||w^1 - w^j||_2^2. \tag{5}$$

The presented loss term aims at minimizing the differences between the latent codes corresponding to different convolutional layers in the StyleGAN model, which directly correlates with the minimization of the image-distribution discrepancy that arises due to the mismatch between the vanilla and the extended latent spaces.

Image Composition. Finally, to ensure seamless image stitching, FICE uses an additional segmentation model S that parses the following three categories from the given input image: 'background' (S_{bg}), 'body' (S_{body}) and 'head' (S_{head}). In contrast to the pose-parsing model D, the components inferred by S contain additional features, i.e., S_{body} captures the clothing of the subject and the S_{head} captures the hair shape, as shown in Fig. 3.

We define two optimization objectives based on the segmentation model S. The first (the image loss $\mathcal{L}_{im}$) aims to preserve the background and face regions of the image. This term is necessary because although the face region is stitched with the synthesized image in the final step, this term helps to preserve the skin tone of the subject and the color characteristics of the input image. The loss term is defined as,

$$M = 1 - S_{body}(I) \tag{6}$$

$$\mathcal{L}_{im} = ||M \odot (I_g - I)||_2^2, \tag{7}$$

where $1 \in \mathbb{R}^{n\times n}$ is a matrix of all ones and $\odot$ denotes the Hadamard product. Furthermore, we use a second loss term that preserves the 'head' region,

$$\mathcal{L}_{head} = ||S_{head}(I) - S_{head}(I_g)||_2^2. \tag{8}$$

This loss is used primarily to preserve the hair of the input image and to account for interactions of the hair/head and body regions that cannot be managed solely by the image-loss term $\mathcal{L}_{im}$ defined above.

Final Optimization Objective. The final optimization objective for the constrained GAN-inversion techniques used by FICE is defined as a weighted superposition of the individual losses, as follows,

$$\begin{aligned}\mathcal{L} =& \lambda_{clip}\mathcal{L}_{clip} + \lambda_{pose}\mathcal{L}_{pose} + \lambda_{reg}\mathcal{L}_{reg} \\ &+ \lambda_{im}\mathcal{L}_{im} + \lambda_{head}\mathcal{L}_{head},\end{aligned} \tag{9}$$

where λ_{clip},λ_{pose},λ_{reg},λ_{im}, and λ_{head} are balancing weights.

3.4. Image Stitching

To preserve the identity of the input person, we perform image stitching as the final step of FICE. The final image I_f is obtained as,

$$I_f = S_{head}(I) \odot I + (1 - S_{head}(I)) \odot I_g^*, \tag{10}$$

where $I_g^* = G(w^*)$ is the image that corresponds to the optimized latent code w^* based on Eq. (2).

4. Experimental Setup

In this section we describe the experimental setup used to demonstrate the capabilities of FICE. Specifically, we discuss datasets and evaluation protocols, the baseline techniques considered as well as relevant implementation details.

Table 1: **Overview of the fashion datasets used in the paper.** Four different datasets are selected for the experiments and used to train and test various components of FICE.

Dataset	**# Images**	**Segmentations**	**Text**	**# Text Descriptions**	**Aim**
VITON [2]	$16,253$	Coarse	✗	n/a	G, E training/testing†
MPV [32]	$35,687$	Coarse	✗	n/a	Testing‡
DeepFashion Retrieval [70]	$52,713$	Fine	✗	n/a	S training
Fashion-Gen [33]	$293,018$	✗	✓	$293,018$	Source of text descriptions

† Training and testing data are disjoint.

‡ We test FICE on qualitative experiments on MPV.

Figure 4: **Sample images from the experimental datasets.** We train and test FICE with fashion images from the VITON, Fashion-Gen and DeepFashion datasets. The MPV dataset is used exclusively for testing.

4.1. Datasets

Here, we outline the data used for evaluation purposes. Four types of datasets are selected for the experiments. The selected datasets, summarized in Table 1, provide image and text data for training and testing of the main FICE components. Example images from the four datasets are presented in Fig. 4.

Image Dataset. We use the VITON dataset as our main image database. VITON is a fashion dataset with $16,253$ frontal-view images of female models with different tops. The images are split into a training set and a test set with $14,221$ and $2,032$ images, respectively. We use the training set to train the GAN model (including G) and E4e encoder (E) and the disjoint test set for performance evaluations. Similar to related studies from the literature [53, 71], we select a test set of manageable size and use the first 120 images of the official VITON test set for the experiments. Additionally, we also use a subset of fashion images from the MPV dataset [32] for the evaluation of FICE. Unlike VITON, MPV images exhibit larger appearance variability with more distinct differences in terms of zoom level and view point.

Segmentation Dataset. To train the segmentation model S needed for FICE, we utilize the DeepFashion dataset [70], specifically, the In-shop Clothes Retrieval part of the dataset. We minimize the discrepancy with the distribution of the VITON data by filtering the images to women subjects with frontal pose and restricting the fashion categories to 'Blouses & Shirts' and 'Tees & Tanks'. Furthermore, the reference segmentation masks of different classes are merged to fit the requirements of FICE, resulting in three final segmentation targets: Body, Face & Hair and Background. The images and segmentation masks are padded to a square shape and then down-scaled to a resolution of 256×256 px, as shown in Fig. 3.

Text Dataset. The last dataset used for the experiments is Fashion-Gen [33]. We use this dataset to obtain the clothing descriptions needed for testing. Fashion-Gen consists of $293,008$ images, each paired with a corresponding text description. We perform our experiments with image–text pairs that belong to the 'top' fashion category to match the VITON dataset's characteristics.

Test Dataset. To construct the dataset for evaluation purposes, we gather the VITON image data and combine it with the Fashion-Gen text data. The Fashion-Gen text descriptions were created by professional stylists and, as a result, contain many fashion-specific technical terms that are regarded as difficult for general text–image matching models, such as CLIP, due to the shift in comparison to the distribution of the training data. Thus, we only consider the sentences with a high match rate to the corresponding image. Specifically, we process each image and its corresponding description to obtain a CLIP matching score, then sort the image–text pairs by the magnitude of the match score. We keep 120 sentences with the highest matching scores, as these are the sentences that the CLIP model 'understands' the best. During testing we combine each test image with every text example to construct all possible image–text combinations, resulting in a total of $120 \cdot 120 = 14,400$ test combinations.

4.2. FICE Implementation Details

Here, we outline the specifics of the FICE implementation. The generator G of FICE is based on StyleGAN2 [39] and was trained on the training split of the VITON dataset. Prior to training, the images from the dataset were first cropped to 192×192 px by removing the bottom part of the image and then re-sized to 256×256 px using bilinear interpolation. The training was performed for a total of $450,000$ iterations, achieving a final Fréchet Inception Distance (FID) of 3.83.

For the pose-parsing we utilized the DensePose model [68] D with a ResNet50 backbone and the Panoptic FPN head [72]. We used the pre-trained model from the Detectron2 repository[2]. The selected model is capable of parsing 24 individual body parts (e.g., upper-left arm). For our implementation, we only considered body parts that are suitable (or applicable) for the VITON dataset, i.e., the upper-body indices.

The segmentation model S was trained on the DeepFashion dataset [70]. We used cross-entropy as the learning objective and weighted the predictions according to the class imbalance of the training split

[2] Available from: https://github.com/facebookresearch/detectron2

of the dataset. The learning procedure was performed for 19 epochs (until convergence) using the Adam optimizer [73] due to its competitive performance [74, 75] with a fixed learning rate of $\eta = 10^{-4}$.

The GAN inversion E4e [29] encoder E was trained on the VITON training dataset, using G as the target GAN model to invert. Finally, the CLIP RN50x4 network architecture was chosen as the image encoder for the CLIP model. For the optimization-based GAN-inversion procedure in Eq. (2), balancing weights were chosen based on preliminary experiments and a visual inspection of the results on the training data, so that $\lambda_{clip} = 1, \lambda_{im} = 30, \lambda_{pose} = 10, \lambda_{head} = 1, \lambda_{reg} = 1$. The number of GAN-inversion optimization iterations was fixed and set to 500 for all the experiments. We used the Adam optimization algorithm with a learning rate of $\eta = 5 \cdot 10^{-2}$ when optimizing the latent code to compute w^*.

4.3. Baseline Models

We use multiple baseline methods to compare against FICE. In this section we provide a brief outline of the selected methods to make the paper self-contained.

4.3.1. FashionGAN [20]

The closest competitor to FICE is **FashionGAN** [20]. FashionGAN involves a two-step generation phase. The first stage defines a plausible segmentation map corresponding to the input subject. The second stage takes the generated segmentation map, text, and categorical attributes to generate the edited image. It is worth noting that this method uses additional categorical data besides the raw text, such as the skin's mean RGB value and sleeve information, to perform the editing. We removed this conditioning to perform a direct comparison. The text encoder used in FashionGAN is not directly suitable for wide-vocabulary text due to its reliance on a predetermined word-level tokenization vocabulary that was constructed from a relatively small text dataset. Thus, we trained the method using the CLIP text encoder to allow for a fair comparison. The final step of the method involves image blending, similar to FICE.

4.3.2. Prompt to Prompt [76]

Prompt to Prompt (P2P) is an editing technique for diffusion generative models. It requires a default text prompt as well as a target text prompt. P2P works by modifying the textual prompts from default to target prompt while preserving the attention maps generated by the default prompt. These attention maps strongly influence the generated image's spatial layout, ensuring greater consistency during the editing process.

The editing process relies on a guidance scale hyperparameter during classifier-free guidance to control the semantic impact of the edit [76]. Higher guidance scales produce outputs strongly aligned with the target prompt, but this can introduce artifacts within the generated image. In our implementation, a guidance scale of 4.0 was used, and attention maps were preserved across all the steps for image generation. Processing real images is achieved through an inverse de-noising process before the P2P editing process.

Experiments indicated that this method is very sensitive to the specific wording of the target text prompt. Throughout the initial testing on the VITON image dataset, the following prompt proved most effective: *"A photo of a woman wearing* $\{t\}$*, full body, high-quality."*, where t denotes the target text fashion description. An empty prompt is used as the neutral prompt. We utilize Stable Diffusion [77], a powerful open-source, text-to-image diffusion model for the image generation. Note the substantial training-data difference: our GAN is trained on 14,221 images compared to Stable Diffusion's dataset of 5 billion images.

4.3.3. GAN-inversion-based methods

Next, we implement several GAN-inversion based methods. Specifically, we use various methods to obtain the latent code of the image, and then process the latent code using the global StyleCLIP method [25], which modifies the latent code to enforce the desired semantics.

GAN inversion. Given our trained GAN model, we train several GAN-inversion encoder models to use as baselines in the experiments presented in the next section. Details about the considered encoding techniques are given below:

- **pSp** [55]. The pSp model was proposed for the tasks of conditional image synthesis, face frontalization, inpainting, and super-resolution, and due to this a broad application range was also selected for our experiments. The architecture of pSp is based on a ResNet-style feature pyramid [78] and multiple encoder networks. The encoders predict a particular StyleGAN latent code from each convolutional layer of the feature pyramid and, in this way, embed images into the StyleGAN latent space.

- **E4e** [29]. E4e follows the architecture of the pSp model, but uses a distinct training process to ensure that the predicted, extended latent code approximates the code defined in the original StyleGAN latent space. This allows the model to perform more convincing image manipulations when using techniques that try to alter the given latent code to achieve semantically meaningful edits.

- **ReStyle** [56]. While pSp and E4e embed images in the latent space in a single, forward pass, ReStyle uses an iterative procedure that gradually improves the embedding so it corresponds better to the given input image. We consider two ReStyle versions for the experiments, one with the pSp and one with the E4e encoder.

- **HyperStyle** [57]. In contrast to the inversion methods above, HyperStyle predicts the latent code of an input image, while also modifying the weights of the StyleGAN generator on a per-sample basis. The model first predicts an approximate latent code using the vanilla StyleGAN generator. This initial prediction, along with the original image, then serve as an input to a hyper-network that predicts the offsets for the StyleGAN weights. The weights are finally modulated with the offsets and the resulting StyleGAN is used to synthesize the final image.

StyleCLIP for Editing. To provide editing capabilities for the GAN-inversion techniques, we use StyleCLIP's global editing method, which displaces the latent code of the given input image along a certain editing direction in accordance with the semantics encoded in some text description. StyleCLIP relies on two hyperparameters (α, β), where α defines the magnitude of the displacement in the latent space and β represents a threshold that controls the entanglement of the edited attributes. The choice of α determines the strength of the semantic content from the text present in the edited image. However, large values of α are known to degrade the image quality and alter both the pose and identity of the input subject. These adverse effects can be reduced by using a disentanglement mechanism with a certain value of β. However, large values of β reduce the presence of the desired semantic presence. A suitable trade-off is therefore required. We consider this trade-off in our comparative assessment and perform StyleCLIP experiments with several different combinations of (α, β) values. Specifically, we use α values of $\{3.0, 4.0, 5.0, 7.5, 10.0\}$ and β values of $\{0.00, 0.025, 0.050, 0.075, 0.100, 0.125\}$, which represent a reasonable cross-section of values for the evaluation.

4.4. Evaluation Metrics

To evaluate FICE against comparative methods, we define four distinct criteria that address different aspects of the editing process, i.e.:

- **Semantic Relevance.** The synthesized images should contain semantics that are relevant with respect to the input text t. We evaluate this aspect in the experiments with the CLIP model, defining a CLIP-similarity score between a given image I and the target text t as,

  $$S(I, t) = \cos(C^i(I), C^t(t)), \tag{11}$$

  where C^i and C^t are again the CLIP image and text encoders, respectively.

- **Pose Preservation.** The edited image should preserve the pose of the input person. We evaluate the pose-preservation capabilities using Intersection over Union (IoU), a commonly used metric in the field of semantic segmentation [79–81]. We utilize DensePose for pose prediction because the model is insensitive with respect to overlaid clothing and compute an IoU score between the pose predictions from the input and edited images.

- **Identity Similarity.** The editing model should preserve the identity and facial appearance of the input image. We therefore use the RetinaFace model [82] to detect the face region, and then process it with ArcFace [83], a face-recognition model, to extract a face-embedding vector. To quantify the similarity between the input and edited image in terms of identity, we compute the cosine similarity between the corresponding face-embedding vectors.

- **Image Fidelity.** The editing methods should synthesize high-fidelity images that are comparable to the inputs. To measure the quality of the generated images, we use the Fréchet Inception Distance (FID) [84]. The metric is based on the difference between the statistics of image embeddings as extracted by the InceptionV3 neural network [85] and has been used as one of the main metrics to evaluate the image quality of different image-synthesis models [37–40].

5. Results and Discussion

In this section we report the results that: (*i*) demonstrate the capabilities of FICE for fashion-image editing through several qualitative examples, (*ii*) compare the proposed model to several competing techniques, (*iii*) highlight the importance of various components through rigorous ablation studies, (*iv*) investigate the sensitivity of FICE to text rephrasing, (*v*) analyze the generalization capabilities of FICE, and (*vi*) explore the model's limitations. Some additional results are also available in the Supplementary Material.

5.1. Qualitative Results

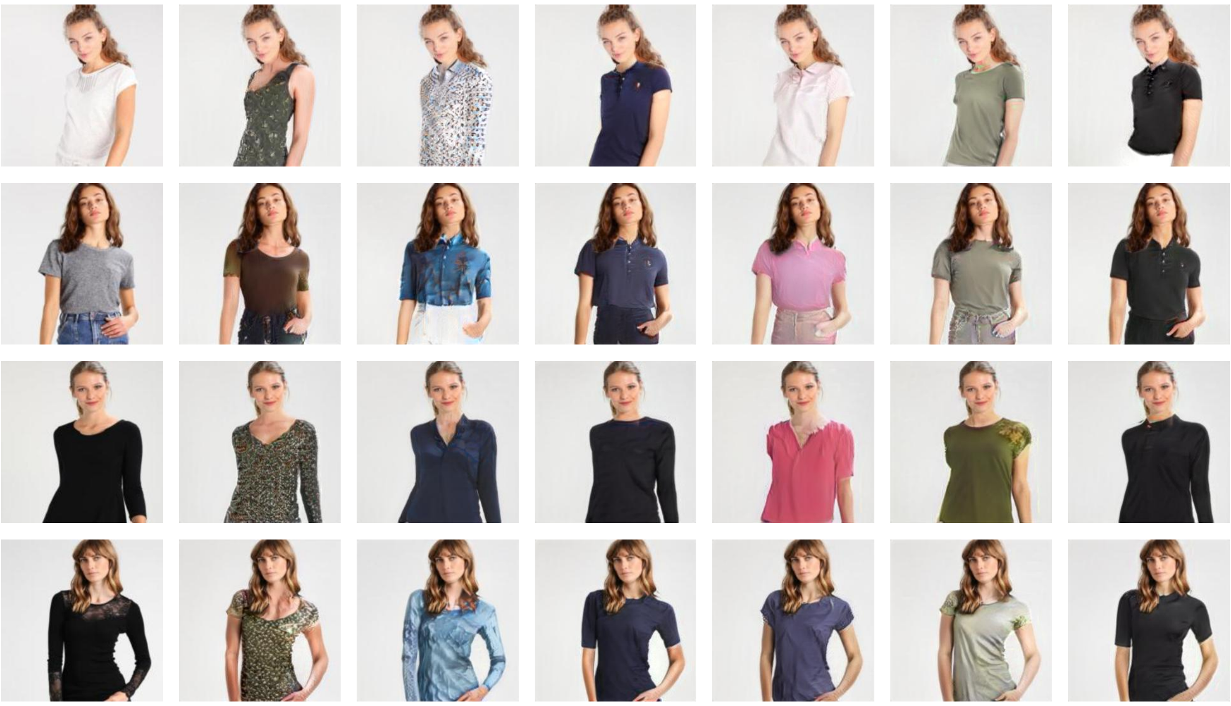
FICE Evaluation. We first evaluate the editing capabilities of FICE in Fig. 5 over a number of test images with different pose characteristics and initial clothing styles (e.g., long and short sleeves, different material, designs, etc.) and different text descriptions. We observe that FICE is able to: (*i*) synthesize complex clothing styles (see, for example, the camouflage pattern in the second column), (*ii*) convincingly incorporate the semantics expressed in the text descriptions into the edited images, (*iii*) preserve the pose and identity of the subjects, (*iv*) add or remove sleeves from the initial clothing (and hallucinate initially obscured objects, e.g., arms), and (*v*) ensure a seamless fit with realistic clothing characteristics (e.g., with creases) without explicit 3D modelling. The presented examples highlight the impressive editing capabilities of FICE and illustrate the flexibility of text-based editing.

Comparative Evaluation. Next, we compare FICE to the competing models introduced in the previous section. For the evaluation, we consider two distinct cases:

- *Same image–different text (SI-DT).* In this configuration we test all the models on the same input image and pair it with different target text descriptions to demonstrate how the models handle different types of semantics.

- *Same text–different image (ST-DI).* Here, we use different input images and pair them with the same target text description to explore the consistency of the edits made by the models with diverse inputs.

From the results in Fig. 6 (SI-DT) and Fig. 7 (ST-DI), we see that both FashionGAN and FICE preserve the subject identity very well. This is due to the final stitching operation, which is only possible due to the pose-preservation techniques of both FICE and FashionGAN. However, we also observe that the FashionGAN-generated results are much less convincing than the remaining methods, often generating



Input

“Sleeveless hand-embroidered sequin camisole featuring a camo pattern throughout in tones of green, brown, and black”

“Long-sleeve silk crepe de chine shirt featuring graphic pattern printed in tones of blue”

“Short-sleeve fine mercerized cotton piqué polo in navy”

“Short-sleeve ruffled cotton poplin and jersey blouse in pink”

“Short-sleeve cotton jersey t-shirt featuring floral pattern in tones of army green and off-white.”

“Short-sleeve cotton piqué polo in soft black”

Figure 5: **Example results generated by FICE for various text descriptions.** As can be seen, FICE is capable of synthesizing complex fashion styles (e.g., see the camouflage pattern in the second column), while preserving the pose and identity of the subjects as well as other image characteristics. Note the realism and seamless integration of different clothing items (in various designs, materials and shapes) without any 3D modelling. Best viewed electronically and zoomed-in.

fashion garments with visual artifacts. In contrast, FICE manages to generate results where the fashion garment is convincing and photo-realistic.

The analysis of the P2P results reveals that while the generated images semantically align with the target prompt, they exhibit poor correspondence to the input images, indicating significant entanglement issues. Specifically, significant pose and identity changes occur, with some results also changing the background (2. row of Fig. 7, zooming in/out (1. row of Fig. 7), and even lacking the upper body entirely (2. row of Fig. 6).

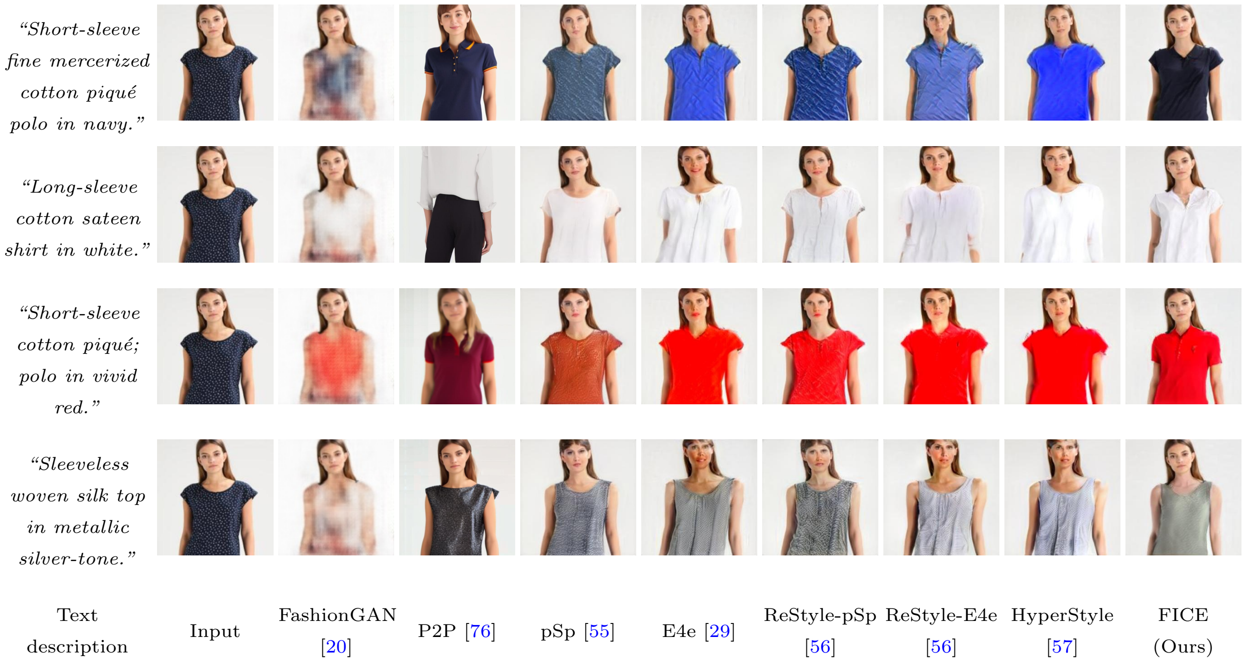

"Short-sleeve fine mercerized cotton piqué polo in navy."
"Long-sleeve cotton sateen shirt in white."
"Short-sleeve cotton piqué; polo in vivid red."
"Sleeveless woven silk top in metallic silver-tone."
Text description
Input
FashionGAN [20]
P2P [76]
pSp [55]
E4e [29]
ReStyle-pSp [56]
ReStyle-E4e [56]
HyperStyle [57]
FICE (Ours)

Figure 6: **Example results with the *Same image–different text* (SI-DT) configuration.** While most techniques are able to incorporate the desired semantics, FICE leads to the most realistic results and most convincingly preserves the identity.

The compared GAN-inversion models (pSp, E4e, ReStyle-pSp, ReStyle-E4e, and HyperStyle) and P2P do not preserve identity nearly as well as FICE and FashionGAN. The reason for that is twofold. First, when an image is embedded in the latent space, there is some loss of identity information due to imperfect reconstruction through the generative model. The second reason lies in the entanglement of the latent-code space that causes several undesired changes to the edited image. StyleCLIP, which is used with GAN-inversion models, integrates a disentanglement mechanism; however, a shift of the latent code still partially affects the person's identity. FICE, on the other hand, preserves the identity due to the image-stitching formulation of the model. Moreover, we observe that the editing results of the competing models exhibit certain entanglement issues, as can be seen in the third row of Fig. 6, where a change to red clothes also adds red lipstick. Finally, FICE performs better when editing certain clothing styles such as 'polo shirt' as seen in a more convincing collar in the first and third rows of Fig. 6. While the semantics from the target text can be seen in most edited images, the results for the proposed model are visually the most convincing.

When looking at Fig. 7, we see that the compared GAN-inversion methods often have problems with exact positioning of the person in the edited image. The person is often shifted with respect to the original image, as best seen in the result of the E4e model in Fig. 7. Such pose shifts make it impossible to integrate image-stitching techniques into the competing models. The main reason for the observed behavior can be

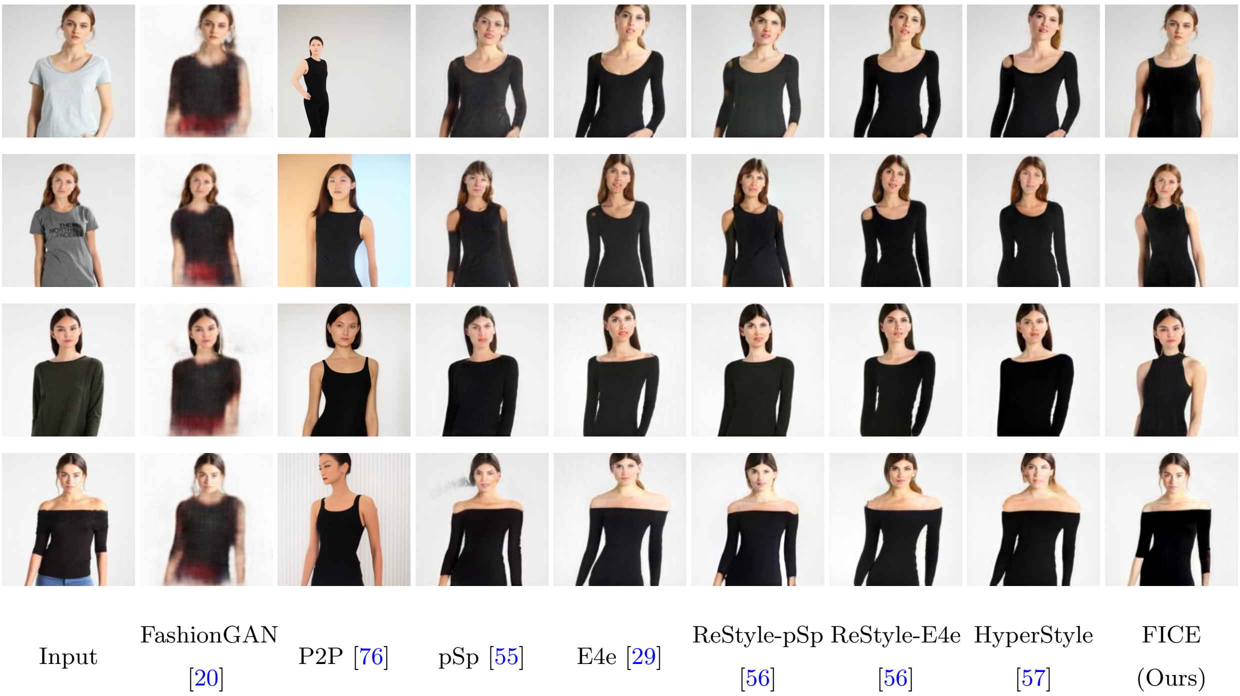


Figure 7: **Visual examples generated with the *Same text–different image* (ST-DI) experimental configuration.** All models were tested with the same target description *"Sleeveless rib knit wool bodysuit in black"*. Note how all models are able to infuse the targeted semantics, but except for FICE, often also produce pose/position changes as well as visual artifacts.

attributed to the use of perceptual losses when training the encoder models. The perceptual losses generally increase the image fidelity at the expense of a precise object localization. While such a trade-off is reasonable for other image-processing tasks, it is not optimal for the task of VTON, where the pose and position of the input person should be preserved exactly. FICE integrates a pose-preserving loss term with a pixel-level loss term to avoid such positioning problems. We note that all the tested models generate consistent edits in terms of the desired semantics, but except for FICE, also often introduce considerable visual artifacts.

5.2. Quantitative Results

In this section, we perform a quantitative evaluation, comparing FICE to the competing methods. The evaluation metrics were computed using the test set, consisting of every pair of 120 VITON images and 120 Fashion-Gen text descriptions, resulting in a total of $14,400$ image–text pairs. When calculating the final scores for performance reporting, the scores for images and text descriptions are averaged for the semantic relevance score, IoU, and identity-similarity metrics. The FID score is averaged only over the text descriptions, since the metric is already calculated over a set of images.

In general, all the GAN-inversion compared methods exhibit some trade-off, where a higher semantic rel-

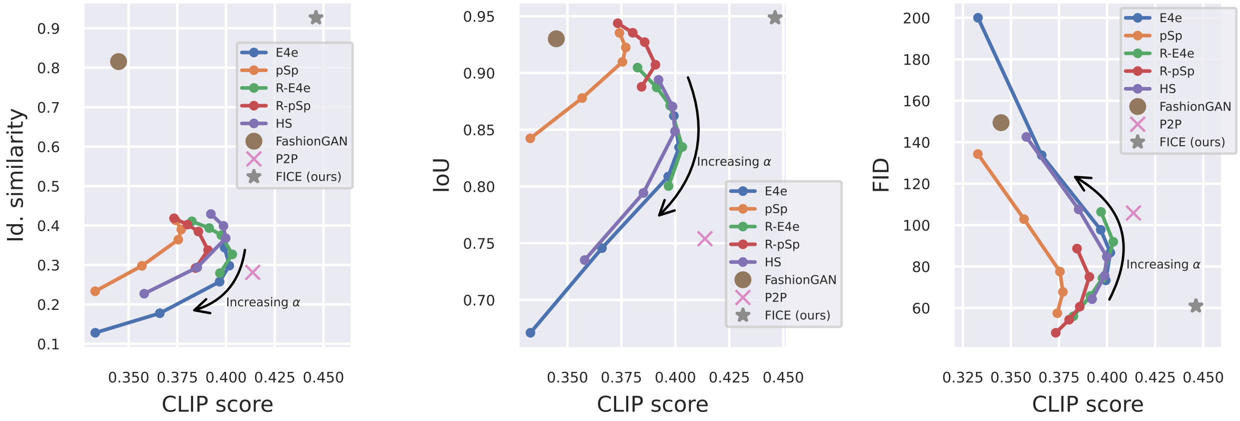


Figure 8: **Fine-grained comparison across multiple hyperparameter values.** For the StyleCLIP-based models, we consider the β value of the best-performing (α, β) combination and $\alpha \in \{3.0, 4.0, 5.0, 7.5, 10.0\}$. Increasing α generally increases the semantic-relevance score, but above a certain point degrades the results.

evance of the image is associated with worse performance for all the other performance indicators considered. Therefore, we decide to evaluate the results for the competing methods for several sets of hyperparameter values. We directly compare FICE with the competitors for the hyperparameters values (α, β), where the maximum semantic score was obtained to ensure a reasonable and fair evaluation.

Table 2 presents the average experimental results. FICE outperforms the competing methods across all the evaluated metrics. The image-stitching technique contributes to FICE's superior identity similarity compared to other GAN-inversion models. FashionGAN, which also employs image stitching, achieves a high identity similarity, but falls slightly short of the FICE. This is likely due to FashionGAN's tendency to generate image artifacts within the face bounding box, degrading the identity-similarity scores. FICE achieves the highest identity similarities, with only minor artifacts of the image stitching preventing it from achieving a perfect score of 1.

The results of the outlined experiments are shown in Table 2 as average, overall results. As can be seen, FICE outperforms the competing methods in terms of all the evaluated metrics. The image-stitching technique helps FICE achieve a superior result in terms of identity preservation when compared to other GAN-inversion models. FashionGAN, which uses an identical image-stitching technique, also achieves a high identity similarity, but one that is slightly worse than FICE. We attribute this to a higher likelihood of FashionGAN generating image artifacts within the face bounding box, which degrade the identity similarity scores. FICE achieves the highest identity similarities, with only the minor artifacts of the image stitching preventing it from achieving a perfect score of 1.

FICE also excels in pose preservation, achieving higher IoU scores than the competing methods. The

Table 2: **Quantitative comparison.** FICE is compared to several competing techniques and across four different performance indicators. The arrows denote whether higher or lower scores imply better performance.

Model	**Semantics** (↑)	**Identity sim.** (↑)	**IoU** (↑)	**FID** (↓)
FashionGAN [20]	0.345	0.816	0.930	149.42
P2P [76]	0.414	0.281	0.754	105.82
pSp [55]	0.377	0.390	0.922	67.74
e4e [29]	0.402	0.298	0.834	86.74
ReStyle-pSp [56]	0.403	0.337	0.907	75.01
ReStyle-e4e [56]	0.391	0.327	0.835	91.92
HyperStyle [57]	0.399	0.368	0.849	84.70
FICE (Ours)	**0.446**	**0.926**	**0.949**	**60.96**

competing GAN-inversion methods and P2P tend to alter the subject's pose as they do not have an explicit subject-pose-conditioning mechanism. FashionGAN achieves the next-best pose-preservation score. Its worse performance can be attributed to the fact that, only the second step in the generation process is conditioned on the subject-segmentation map, while the final pose is not explicitly enforced in the training procedure. Conversely, FICE enforces pose consistency in every iteration of the GAN-inversion process.

Finally, FICE also achieves the best FID scores (lower is better), pointing to the highest visual quality among all the evaluated models. FashionGAN, on the other hand, has degraded FID results, again likely due to the presence of image artefacts in the generated images. P2P achieves fairly weak FID results, which can be attributed to the discrepancy between the test-set image distribution and the generated one due to the P2P's entanglement issues.

In Fig. 8 we show the results for several α values, where the β value is fixed to the value from the best (α, β) combination. Increasing the α value leads to a higher CLIP score (higher semantic relevance), but only up to a certain point, after which the results become worse. All the other metrics worsen when the α value is increased.

5.3. Ablation Study

In order to facilitate understanding of the importance of individual FICE components, we perform an ablation study of the impact of: (i) the latent-space choice made, (ii) the initialization scheme used, and (iii) the contribution of different loss terms. The motivation for the ablation is to further justify the design choices made with the proposed editing approach.

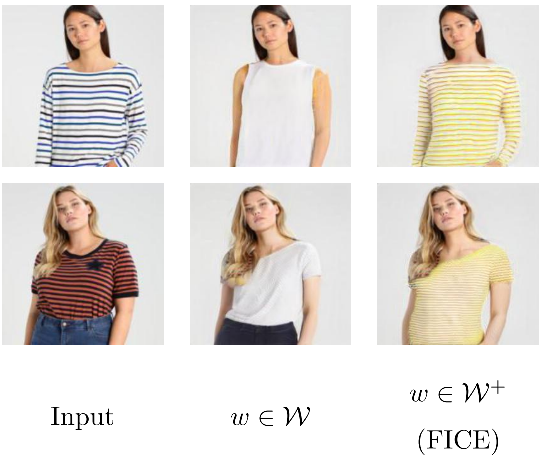


Figure 9: **Visual ablation study results – latent space.** Examples of FICE–edited images when the latent space is restricted to $w \in \mathcal{W}$. The target description is set to *"Structured knit bandeau top in yellow and off-white stripes"*. Due to lower expressiveness of the vanilla StyleGAN latent space $\mathcal{W}$, the generated results exhibit limited pose preservation and semantic correctness.

Table 3: **Ablation study w.r.t. latent spaces.** We analyze the choice of the latent space (vanilla $\mathcal{W}$ and extended $\mathcal{W}^+$) as well as the procedure for initializing the latent code w.

Latent-space variant	**Semantics** (↑)	**Id. sim.** (↑)	**IoU** (↑)	**FID** (↓)
$\bar{w}, \mathcal{W}$	0.422	0.923	0.914	72.11
$\bar{w}, \mathcal{W}^+$	0.443	0.918	0.919	77.13
$\hat{E}(I), \mathcal{W}$	0.378	0.911	0.886	63.16
$E(I), \mathcal{W}^+$ (FICE)	**0.446**	**0.926**	**0.949**	**60.96**

Latent Space and Initialization. To demonstrate the importance of the latent-space choice made for FICE as well as the importance of the initialization procedure, we consider the vanilla $\mathcal{W}$ and the extended $\mathcal{W}^+$ latent-code spaces, and initializations with either the latent-code mean $w \leftarrow \bar{w}$ or the GAN-inversion encoder predicted latent code $w \leftarrow E(I)$. Because the utilized GAN-inversion E encoder (E4e) is based on the extended $\mathcal{W}^+$ space, we train a separate E4e encoder (denoted as $\hat{E}$) for the vanilla latent-code space $\mathcal{W}$ using the proposed method from the E4e repository[3].

From the results in Table 3, we see that the vanilla latent space $\mathcal{W}$ has limited expressive power, which adversely affects the semantic-relevance scores. The extended latent space $\mathcal{W}^+$ contributes towards higher semantic scores, but requires a careful initialization scheme. The trivial mean-initialization scheme leads to a degradation in image fidelity, whereas a GAN-inversion based initialization leads to fidelity gains. We

[3] Available at https://github.com/omertov/encoder4editing

Table 4: **Ablation study results w.r.t. loss terms.** We analyze the impact of individual FICE optimization objectives across four performance measures.

Objective	**Semantics** (↑)	**Id. sim.** (↑)	**IoU** (↑)	**FID** (↓)
Semantic term ($\mathcal{L}_{clip}$)	0.453	0.819	0.781	82.52
+ composition ($\mathcal{L}_{im/head}$)	**0.466**	0.924	0.888	68.94
+ latent reg. ($\mathcal{L}_{reg}$)	0.450	0.924	0.884	64.14
+ pose pres. ($\mathcal{L}_{pose}$)†	0.446	**0.926**	**0.949**	**60.96**

† Complete FICE

show visual examples that illustrate the impact of the latent-space choice in Fig. 9. Note how due to the limited expressiveness of the $\mathcal{W}$ space (compared to $\mathcal{W}^{+}$) the editing procedure is not able to infuse proper semantics into the images shown in the middle column.

Impact of Loss Terms. Next, we evaluate the contribution of the individual loss terms from Eq. (9) on the generated results. Specifically, we ablate three distinct terms by setting the corresponding loss weights to 0: (*i*) the pose-preservation term ($\lambda_{pose} = 0$), (*ii*) the image-composition term ($\lambda_{im} = 0$ and $\lambda_{head} = 0$), and (*iii*) the latent-code regularization term ($\lambda_{reg} = 0$). We keep the semantic-related CLIP term, as text-based editing is not possible without it.

The results of loss-related ablations are shown quantitatively in Table 4 and qualitatively in Fig. 10. We observe that the image-composition term improves all the metrics by ensuring that composition does not cause unnatural visual artifacts. Interestingly, the absence of the composition term strongly degrades the identity-similarity score, even though the face is preserved by the image-stitching step. Regularization of the latent code slightly degrades the semantic-relevance score, but contributes to image fidelity. Finally, the pose-preservation term also slightly lowers the semantic-relevance score, but contributes to the IoU index and further improves the FID score.

Sensitivity of Balancing Weights. Here, we inspect the sensitivity of FICE with respect to the values of the balancing weights. We modify the image-composition loss term $\mathcal{L}_{im}$ by adjusting its corresponding balancing weight λ_{im}. The default FICE balancing weight is $\lambda_{im} = 30$, and we perform experiments by changing its value to $\{1, 15, 45, 60\}$. The results are shown in Figure 11. We observe that the results are robust to large deviations of this weighting term, but low values tend to fail in preserving the skin tone, while high values might cause the model to focus less on the correct semantic content.

5.4. Sensitivity to Textual Rephrasing

In the next series of experiments, we explore FICE's sensitivity to rephrasing of the input prompts. We conduct an experiment by using all 120 textual descriptions in the testing dataset, rephrasing each

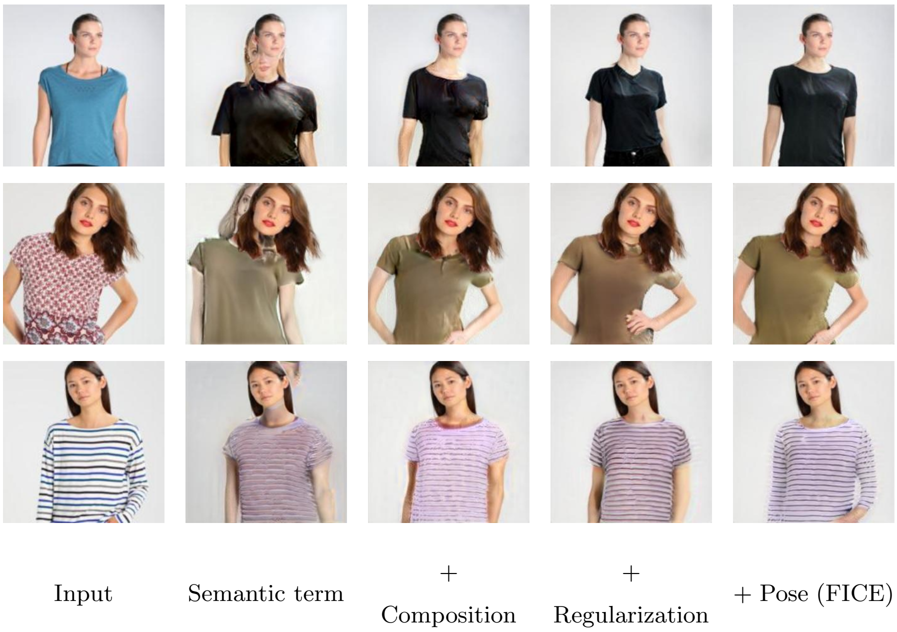


Figure 10: **Visual ablation study results – loss terms.** An optimization objective that relies exclusively on the semantic term (2nd column) produces unsatisfactory image-composition results. Adding the image-composition term (3rd column) ensures pleasing image-stitching results, but the output image still often contains visual artifacts. The addition of the latent-code regularization term (3rd column) helps mitigate these artifacts. Finally, the pose-preservation accurate ensures more realistic and true results compared to the input.

description into a new one. For this purpose, we resort to the GPT4 large language model [86] with the following prompt: "Rephrase the following fashion-related sentence while preserving the original meaning: *#text*", where *#text* is the textual description that originates from the initial dataset. We collect the returned sentences and process them with FICE on the testing-image dataset. Example results are shown in Fig. 12. We observe that the edited images exhibit slightly different garment colors and clothing styles. However, the overall changes are minute and semantically correct and do not introduce any variations in pose or facial appearance.

5.5. Generalization Analysis With Other Data

In this section, we examine how FICE generalizes across different types of data and datasets.

Image Data. We explore the generalization capabilities of FICE by investigating text-conditioned image editing on the MPV image dataset [32], which was not used at any point of the training procedure. We process the images with the same pre-processing operations as used for the VITON dataset, first cropping the bottom part of the image to 192 × 192 px, then resizing the cropped image to 256 × 256 px. For the experiments, we again create various image–text combinations, where the text descriptions stem from the

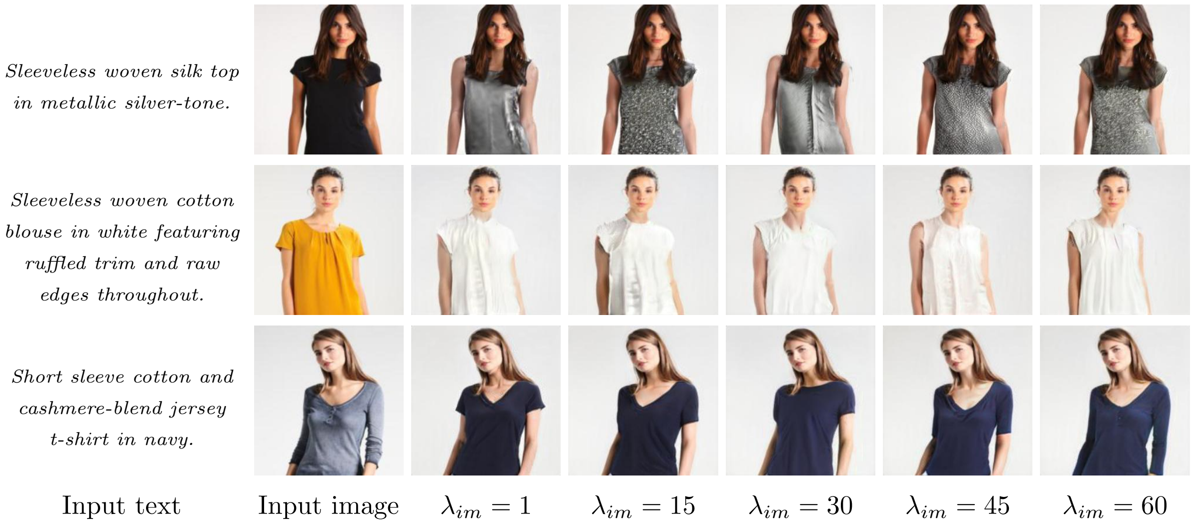

Sleeveless woven silk top in metallic silver-tone.
Sleeveless woven cotton blouse in white featuring ruffled trim and raw edges throughout.
Short sleeve cotton and cashmere-blend jersey t-shirt in navy.
Input text
Input image
$\lambda_{im} = 1$
$\lambda_{im} = 15$
$\lambda_{im} = 30$
$\lambda_{im} = 45$
$\lambda_{im} = 60$

Figure 11: **Visual ablation study results – balancing weights.** The results obtained when varying the image-composition loss term weight λ_{im} are fairly robust. However, some undesired deviations can occur. When the image-composition weight is low, the optimization focuses less on the original skin tone and it might not be well preserved, as seen in the results for $\lambda_{im} = 1$ in the first and second rows. On the other hand, if the weight is too high, it might overpower the other loss terms, such as the semantic content term, which can cause the model to not generate all the target semantics. An example can be seen for the higher values of λ_{im} in the last row, where the depicted subject wears longer sleeves even though the target text specifies short sleeves. Note that $\lambda_{im} = 30$ corresponds to the default FICE settings.

Fashion-Gen dataset. Qualitative example results are shown in Fig. 13. Note how FICE is again able to convincingly infuse the semantics from the text descriptions into the MPV images, while preserving the pose and identity of the subjects in the input images.

Text Data. Here, we explore the FICE generalization capabilities with respect to the textual inputs. Our aim is to synthesize novel textual descriptions with the least amount of dataset-specific bias. We collect new text descriptions using the GPT4 large language model [86]. The input prompt was specified to match the fashion categories present in our testing-image dataset (VITON). The GPT4 input prompt was set to *”Generate random textual descriptions, that would describe a top or a blouse, while focusing on visual aspects.”*. Example results of this experiment are shown in Fig. 14. We observe that FICE is robust with respect to a wide range of input textual descriptions, generating convincing image edits across a wide range of fashion concepts.

5.6. Limitations

While the proposed FICE model achieves high-quality, competitive image-editing results, it still has certain limitations. Specifically, the model inherits the constraints of CLIP on the length of the text

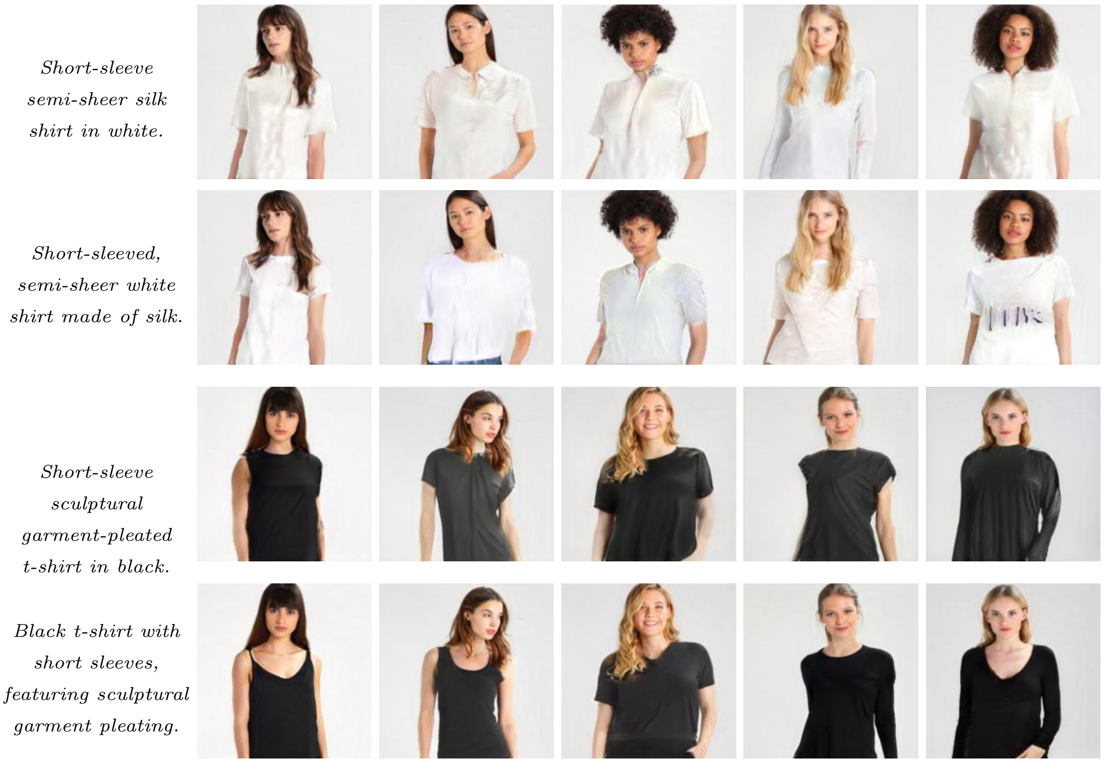

Short-sleeve semi-sheer silk shirt in white.
Short-sleeved, semi-sheer white shirt made of silk.
Short-sleeve sculptural garment-pleated t-shirt in black.
Black t-shirt with short sleeves, featuring sculptural garment pleating.

Figure 12: **FICE Sensitivity to Textual Rephrasing.** Rephrasing the input prompts can lead to different image-editing results. The edited images can have a different colour of the fashion garment or a slightly different clothing style. However, overall, the changes are slight, semantically correct, and do not induce any variations in pose or facial appearance.

description, which is limited to 76 tokens, as extracted by the byte-pair encoding technique [87]. FICE is also a slower method than the encoder-based methods, requirinig approximately 40 seconds for image edits, compared to less than a second for GAN-inversion-based methods and approximately 10 seconds for the P2P model. Finally, FICE sometimes struggles with accurate depictions of sleeve length (see 3rd row of Fig. 5), text prompts containing detailed descriptions of objects/logos, and the occasional unconvincing synthesis of a subject's hands. Failure cases due to complex text prompts and unconvincing synthesized hands are exemplified in Fig. 15. We speculate that a higher-capacity generative model would contribute a great deal to improvements in the FICE results.

6. Conclusion

In this paper we presented FICE, a novel model for realistic text-conditioned fashion-image editing. The core of the approach is based on a novel, extended GAN-inversion procedure guided by CLIP semantic knowledge as well as pose, regularization, and composition constraints. We showed through rigorous exper-

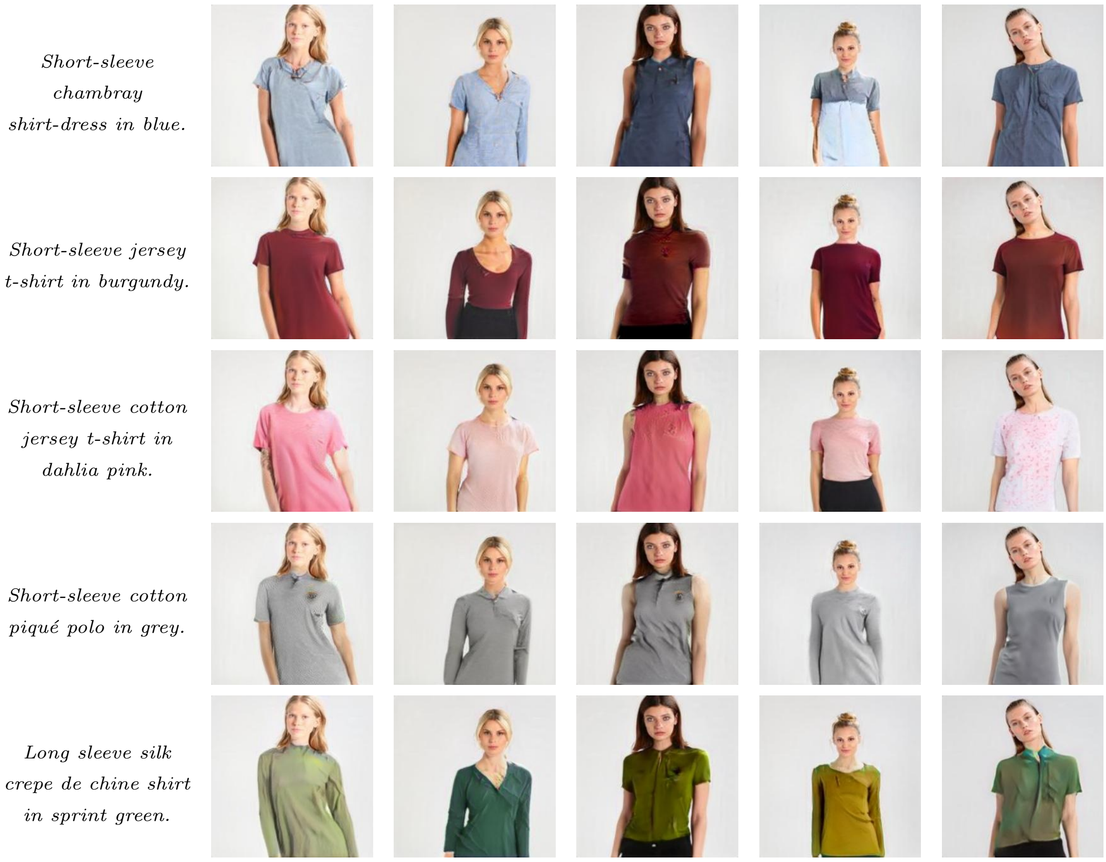

Short-sleeve chambray shirt-dress in blue.
Short-sleeve jersey t-shirt in burgundy.
Short-sleeve cotton jersey t-shirt in dahlia pink.
Short-sleeve cotton piqué polo in grey.
Long sleeve silk crepe de chine shirt in sprint green.

Figure 13: **Example results generated by FICE for various text descriptions for the MPV dataset.** We again observe that the results preserve the pose and identity of the subjects, as well as other image characteristics, such as scene illumination.

iments that FICE not only convincingly edits fashion images, but it also enhances the realism and fidelity of the synthesized fashion images.

One key insight from our research is the effectiveness of integrating text-based input with GANs for fashion-image editing. Unlike traditional methods that rely on direct image manipulation, using descriptive text allows for a more intuitive and flexible editing process. Our findings also underscore the importance of maintaining the pose and identity of the subjects in fashion images. This is achieved through pose- and image-composition-preserving constraints that help ensure that the inherent characteristics of the original images are retained and achieve a more convincing result than the compared methods. The main novelty introduced in this work, i.e., the extended GAN-inversion approach, has implications for other areas and generative tasks that require the embedding of existing images into the latent space, while also considering additional sources of information during the process.

We plan to further extend the procedure to additional sources of information (e.g., reference pose key-

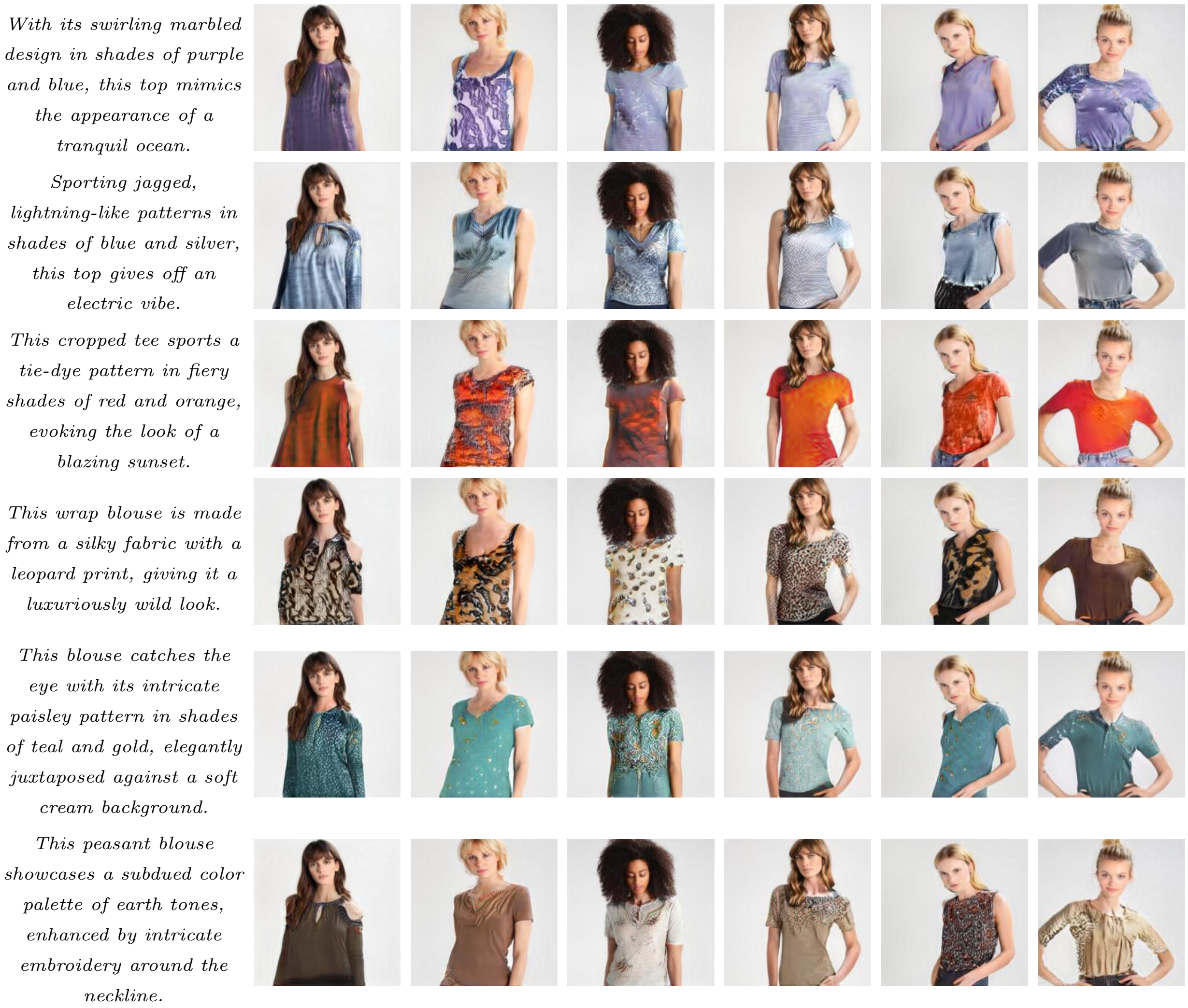

With its swirling marbled design in shades of purple and blue, this top mimics the appearance of a tranquil ocean.
Sporting jagged, lightning-like patterns in shades of blue and silver, this top gives off an electric vibe.
This cropped tee sports a tie-dye pattern in fiery shades of red and orange, evoking the look of a blazing sunset.
This wrap blouse is made from a silky fabric with a leopard print, giving it a luxuriously wild look.
This blouse catches the eye with its intricate paisley pattern in shades of teal and gold, elegantly juxtaposed against a soft cream background.
This peasant blouse showcases a subdued color palette of earth tones, enhanced by intricate embroidery around the neckline.

Figure 14: **Example results generated by FICE for GPT4-generated textual descriptions.** FICE is capable of synthesizing convincing image edits from a wide range of input textual descriptions.

points, additional visual prompts, clothing-texture information, etc.) that will allow us to alter additional characteristics of the input images, look into mechanisms for speeding up the iterative inversion process and explore possibilities to edit other clothing items beyond tops, such as lower garments or shoes.

Acknowledgements

The work was supported in parts by the ARRS Research Programme P2-0250(B) Metrology and Biometric System, the ARRS Project J2-2501(A) DeepBeauty and the ARRS junior researcher program.

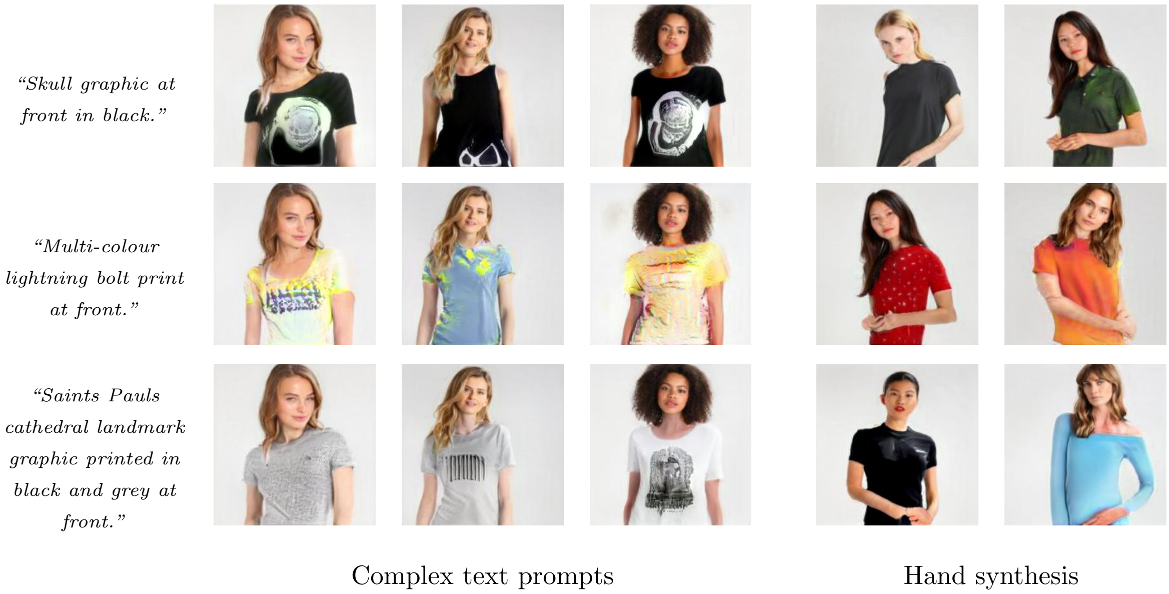

"Skull graphic at front in black."
"Multi-colour lightning bolt print at front."
"Saints Pauls cathedral landmark graphic printed in black and grey at front."
Complex text prompts
Hand synthesis

Figure 15: **Illustration of FICE limitations. Left:** when presented with text descriptions involving objects and specific logos to be shown on the garment, the model is only capable of generating approximate results. In the presented examples, the model fails to produce a convincing skull graphic (first row), a lightning bolt (middle row), or a cathedral (last row). **Right:** Our model sometimes fails to generate convincing hands of the subjects.

References

[1] A. Kozlowski, M. Bardecki, C. Searcy, Environmental impacts in the fashion industry: A life-cycle and stakeholder framework, Journal of Corporate Citizenship (2012) 17–36.

[2] X. Han, Z. Wu, Z. Wu, R. Yu, L. S. Davis, Viton: An image-based virtual try-on network, in: Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7543–7552.

[3] B. Fele, A. Lampe, P. Peer, V. Struc, C-vton: Context-driven image-based virtual try-on network, in: IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 3144–3153.

[4] C. Ge, Y. Song, Y. Ge, H. Yang, W. Liu, P. Luo, Disentangled cycle consistency for highly-realistic virtual try-on, in: Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16928–16937.

[5] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, P. Luo, Parser-free virtual try-on via distilling appearance flows, in: Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8485–8493.

[6] Z. Lyu, X. Xu, C. Yang, D. Lin, B. Dai, Accelerating diffusion models via early stop of the diffusion process, arXiv preprint arXiv:2205.12524 (2022).

[7] R. Kips, R. Jiang, S. Ba, E. Phung, P. Aarabi, P. Gori, M. Perrot, I. Bloch, Deep graphics encoder for real-time video makeup synthesis from example, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3889–3893.

[8] A. Basu, S. Paul, S. Ghosh, S. Das, B. Chanda, C. Bhagvati, V. Snasel, Digital restoration of cultural heritage with data-driven computing: A survey, IEEE Access (2023).

[9] H. Zhang, S. Lin, R. Shao, Y. Zhang, Z. Zheng, H. Huang, Y. Guo, Y. Liu, Closet: Modeling clothed humans on continuous

surface with explicit template decomposition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 501–511.

[10] Y. Xiu, J. Yang, X. Cao, D. Tzionas, M. J. Black, Econ: Explicit clothed humans optimized via normal integration, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 512–523.

[11] Y. Cao, K. Han, K.-Y. K. Wong, Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4647–4657.

[12] D.-Y. Song, H. Lee, J. Seo, D. Cho, Difu: Depth-guided implicit function for clothed human reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8738–8747.

[13] F. Zhao, Z. Li, S. Huang, J. Weng, T. Zhou, G.-S. Xie, J. Wang, Y. Shan, Learning anchor transformations for 3d garment animation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 491–500.

[14] Y. Jafarian, T. Y. Wang, D. Ceylan, J. Yang, N. Carr, Y. Zhou, H. S. Park, Normal-guided garment uv prediction for human re-texturing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4627–4636.

[15] R. Jain, K. K. Singh, M. Hemani, J. Lu, M. Sarkar, D. Ceylan, B. Krishnamurthy, Vgflow: Visibility guided flow network for human reposing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21088–21097.

[16] L. Zhu, D. Yang, T. Zhu, F. Reda, W. Chan, C. Saharia, M. Norouzi, I. Kemelmacher-Shlizerman, Tryondiffusion: A tale of two unets, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4606–4615.

[17] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., Recent advances in convolutional neural networks, Pattern Recognition 77 (2018) 354–377.

[18] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, B. Kim, Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments, Pattern Recognition 120 (2021) 108102.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems (NIPS), 2014.

[20] S. Zhu, R. Urtasun, S. Fidler, D. Lin, C. Change Loy, Be your own prada: Fashion synthesis with structural coherence, in: International Conference on Computer Vision (ICCV), 2017, pp. 1680–1688.

[21] Y. Jiang, S. Yang, H. Qju, W. Wu, C. C. Loy, Z. Liu, Text2human: Text-driven controllable human image generation, ACM Transactions on Graphics (TOG) 41 (2022) 1–11.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[23] N. Mu, A. Kirillov, D. Wagner, S. Xie, Slip: Self-supervision meets language-image pre-training, in: European Conference on Computer Vision (ECCV), Springer, 2022, pp. 529–544.

[24] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: International Conference on Machine Learning (ICML), PMLR, 2021, pp. 8821–8831.

[25] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, D. Lischinski, Styleclip: Text-driven manipulation of stylegan imagery, in: International Conference on Computer Vision (ICCV), 2021, pp. 2085–2094.

[26] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, M.-H. Yang, GAN inversion: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[27] Y. Shen, C. Yang, X. Tang, B. Zhou, Interfacegan: Interpreting the disentangled face representation learned by gans, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[28] A. C. Baykal, A. B. Anees, D. Ceylan, E. Erdem, A. Erdem, D. Yuret, Clip-guided stylegan inversion for text-driven real

image editing, ACM Transactions on Graphics 42 (2023) 1–18.

[29] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, D. Cohen-Or, Designing an encoder for stylegan image manipulation, ACM Transactions on Graphics (TOG) 40 (2021) 1–14.

[30] Z. Wu, D. Lischinski, E. Shechtman, Stylespace analysis: Disentangled controls for stylegan image generation, in: Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12863–12872.

[31] M. Pernuš, V. Štruc, S. Dobrišek, Maskfacegan: High resolution face editing with masked gan latent code optimization, IEEE Transactions on Image Processing (2023).

[32] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, J. Yin, Towards Multi-Pose Guided Virtual Try-on Network, in: International Conference on Computer Vision (ICCV), 2019, pp. 9026–9035.

[33] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, C. Pal, Fashion-gen: The generative fashion dataset and challenge, arXiv preprint arXiv:1806.08317 (2018).

[34] X. Wu, K. Xu, P. Hall, A survey of image synthesis and editing with generative adversarial networks, Tsinghua Science and Technology 22 (2017) 660–674.

[35] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, J. Liu, Fashion meets computer vision: A survey, ACM Computing Surveys (CSUR) 54 (2021) 1–41.

[36] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: International Conference on Learning Representations (ICLR), 2016.

[37] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, in: International Conference on Learning Representations (ICLR), 2018.

[38] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401–4410.

[39] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8110–8119.

[40] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, T. Aila, Training generative adversarial networks with limited data, Advances in Neural Information Processing Systems (NIPS) (2020) 12104–12114.

[41] P. Shamsolmoali, M. Zareapoor, E. Granger, H. Zhou, R. Wang, M. E. Celebi, J. Yang, Image synthesis with adversarial networks: A comprehensive survey and case studies, Information Fusion 72 (2021) 126–146.

[42] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: International conference on machine learning (ICML), PMLR, 2016, pp. 1060–1069.

[43] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. N. Metaxas, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: International Conference on Computer Vision (ICCV), 2017.

[44] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. N. Metaxas, Stackgan++: Realistic image synthesis with stacked generative adversarial networks, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 1947–1962.

[45] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, Attngan: Fine-grained text to image generation with attentional generative adversarial networks, in: Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1316–1324.

[46] T. Qiao, J. Zhang, D. Xu, D. Tao, Mirrorgan: Learning text-to-image generation by redescription, in: Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1505–1514.

[47] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125 (2022).

[48] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic text-to-image diffusion models with deep language understanding, Advances in neural information processing systems 35 (2022) 36479–36494.

[49] H. Dong, S. Yu, C. Wu, Y. Guo, Semantic image synthesis via adversarial learning, in: International Conference on Computer Vision (ICCV), 2017, pp. 5706–5714.

[50] S. Nam, Y. Kim, S. J. Kim, Text-adaptive generative adversarial networks: manipulating images with natural language, Advances in neural information processing systems (NIPS) 31 (2018).

[51] B. Li, X. Qi, T. Lukasiewicz, P. H. Torr, Manigan: Text-guided image manipulation, in: Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7880–7889.

[52] R. Abdal, Y. Qin, P. Wonka, Image2stylegan: How to embed images into the stylegan latent space?, in: International Conference on Computer Vision (ICCV), 2019, pp. 4431–4440.

[53] R. Abdal, Y. Qin, P. Wonka, Image2stylegan++: How to edit the embedded images?, in: Computer Vision and Pattern Recognition (CVPR), 2020.

[54] W. Xia, Y. Yang, J.-H. Xue, B. Wu, Tedigan: Text-guided diverse face image generation and manipulation, in: Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2256–2265.

[55] 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: Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2287–2296.

[56] Y. Alaluf, O. Patashnik, D. Cohen-Or, Restyle: A residual-based stylegan encoder via iterative refinement, in: International Conference on Computer Vision (ICCV), 2021, pp. 6711–6720.

[57] Y. Alaluf, O. Tov, R. Mokady, R. Gal, A. H. Bermano, Hyperstyle: Stylegan inversion with hypernetworks for real image editing, in: Computer Vision and Pattern Recognition (CVPR), 2022.

[58] L. Qiu, G. Chen, J. Zhou, M. Xu, J. Wang, X. Han, Rec-mv: Reconstructing 3d dynamic cloth from monocular videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4637–4646.

[59] X. Zou, X. Han, W. Wong, Cloth4d: A dataset for clothed human reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12847–12857.

[60] A. Grigorev, M. J. Black, O. Hilliges, Hood: Hierarchical graphs for generalized modelling of clothing dynamics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16965–16974.

[61] K. Wang, G. Zhang, S. Cong, J. Yang, Clothed human performance capture with a double-layer neural radiance fields, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21098–21107.

[62] L. De Luigi, R. Li, B. Guillard, M. Salzmann, P. Fua, Drapenet: Garment generation and self-supervised draping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1451–1460.

[63] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, P. Luo, Towards photo-realistic virtual try-on by adaptively generating-preserving image content, in: Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7850–7859.

[64] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: International Conference on Computer Vision (ICCV), 2017, pp. 2223–2232.

[65] S. He, Y.-Z. Song, T. Xiang, Style-based global appearance flow for virtual try-on, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3470–3479.

[66] N. Fang, L. Qiu, S. Zhang, Z. Wang, K. Hu, K. Wang, A novel dagan for synthesizing garment images based on design attribute disentangled representation, Pattern Recognition 136 (2023) 109248.

[67] D. Zhang, C. Zuo, Q. Wu, L. Fu, X. Xiang, Unabridged adjacent modulation for clothing parsing, Pattern Recognition 127 (2022) 108594.

[68] R. A. Güler, N. Neverova, I. Kokkinos, Densepose: Dense human pose estimation in the wild, in: Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7297–7306.

[69] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, in: Computer Vision and Pattern Recognition (CVPR), 2017.

[70] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, in: Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1096–1104.

[71] S. Menon, A. Damian, S. Hu, N. Ravi, C. Rudin, Pulse: Self-supervised photo upsampling via latent space exploration of generative models, in: Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2437–2445.

[72] A. Kirillov, R. Girshick, K. He, P. Dollár, Panoptic feature pyramid networks, in: Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6399–6408.

[73] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR), 2014.

[74] A. Sezer, A. Altan, Detection of solder paste defects with an optimization-based deep learning model using image processing techniques, Soldering & Surface Mount Technology 33 (2021) 291–298.

[75] İ. Yağ, A. Altan, Artificial intelligence-based robust hybrid algorithm design and implementation for real-time detection of plant diseases in agricultural environments, Biology 11 (2022) 1732.

[76] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, D. Cohen-Or, Prompt-to-prompt image editing with cross attention control, arXiv preprint arXiv:2208.01626 (2022).

[77] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[78] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.

[79] S. Luo, Y. Li, P. Gao, Y. Wang, S. Serikawa, Meta-seg: A survey of meta-learning for image segmentation, Pattern Recognition (2022) 108586.

[80] X. Li, C. Li, M. M. Rahaman, H. Sun, X. Li, J. Wu, Y. Yao, M. Grzegorzek, A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches, Artificial Intelligence Review 55 (2022) 4809–4878.

[81] J. Zhang, C. Li, S. Kosov, M. Grzegorzek, K. Shirahama, T. Jiang, C. Sun, Z. Li, H. Li, Lcu-net: A novel low-cost u-net for environmental microorganism image segmentation, Pattern Recognition 115 (2021) 107885.

[82] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: Computer Vision and Pattern Recognition (CVPR), 2020.

[83] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4690–4699.

[84] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in neural information processing systems (NIPS), 2017, pp. 6626–6637.

[85] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.

[86] OpenAI, GPT-4 Technical Report, arXiv e-prints (2023) arXiv:2303.08774. arXiv:2303.08774.

[87] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, arXiv preprint arXiv:1508.07909 (2015).