TD-GEM: Text-Driven Garment Editing Mapper

Reza Dadfar*
dadfar@kth.se
Sanaz Sabzevari*
sanazsab@kth.se
Mårten Björkman
celle@kth.se
Danica Kragic
dani@kth.se

KTH Royal Institute of Technology
Stockholm, Sweden

Abstract

Language-based fashion image editing allows users to try out variations of desired garments through provided text prompts. Inspired by research on manipulating latent representations in StyleCLIP and HairCLIP, we focus on these latent spaces for editing fashion items of full-body human datasets. Currently, there is a gap in handling fashion image editing due to the complexity of garment shapes and textures and the diversity of human poses. In this paper, we propose an editing optimizer scheme method called Text-Driven Garment Editing Mapper (TD-GEM), aiming to edit fashion items in a disentangled way. To this end, we initially obtain a latent representation of an image through generative adversarial network inversions such as Encoder for Editing (e4e) or Pivotal Tuning Inversion (PTI) for more accurate results. An optimization-based Contrastive Language-Image Pre-training (CLIP) is then utilized to guide the latent representation of a fashion image in the direction of a target attribute expressed in terms of a text prompt. Our TD-GEM manipulates the image accurately according to the target attribute, while other parts of the image are kept untouched. In the experiments, we evaluate TD-GEM on two different attributes (*i.e.,* "color" and "sleeve length"), which effectively generates realistic images compared to the recent manipulation schemes.

1 Introduction

Text-driven garment editing frameworks provide a convenient digital tool for end-users to edit fashion items. The application of high-quality synthesized images for visualization of not yet produced garments allows for a more sustainable online fashion industry, ultimately decreasing the retailer's costs and environmental carbon footprint [12]. Recently, Generative Adversarial Networks (GANs) [7] have been used for generating photo-realistic images for various datasets. It is extensively employed in Virtual Try-ONs (VTONs) [3, 18, 25, 29]

*Equal contribution

Figure 1: **Disentangled garment manipulation** for prompt texts "lengthening sleeve", "green", and "blue" using our proposed TD-GEM architecture. The resulting manipulations visually have the same postures, shapes, and contours as the original image, preserving fine-grained details.

and outfit generators [8]. Despite tremendous development in this domain, text-conditioned human outfit editing has not yet been well explored.

Image attribute manipulation requires an accurate latent mapping between the text embedding space and latent visual space of the synthesized image often implemented by StyleGAN-based approaches [9, 10, 11, 23]. Recently, this has been done in pioneering studies such as StyleCLIP [19] and TediGAN [32] to edit images based on a target text prompt. They find the latent visual subspace aligned to the text embedding space. However, StyleCLIP attains text-based semantic image editing through Contrastive Language-Image Pre-training (CLIP) encoding [21]. To generate high-quality images, StyleGAN2 [29] has shown great promise across various applications. The majority of research studies have focused on face, car, and building datasets despite limiting exploration in the domain of human clothing [5]. Due to the diverse range of human poses and intricate textures and shapes of garments, manipulating human outfits through generative models is a challenging task [13]. Thanks to work [5], a large-scale fashion image dataset called Stylish-Humans-HQ (SHHQ) was collected and trained through the StyleGAN2 network.

This paper addresses image manipulation, including text-conditioned editing in the fashion domain using the SHHQ dataset. We learn a mapping between text prompt embeddings and latent representations of input images while generating disentangled output images based on text descriptions. The proposed Text-Driven Garment Editing Mapper (TD-GEM) edits attributes of the input image according to the input text, *e.g.* the sleeve length or color of the garment, using a single mapper and inversion space. It successfully preserves the irrelevant attributes of the input image. The primary contributions of this paper are

- We provide a text-driven image manipulation framework, TD-GEM, for full-body fashion images using CLIP and GAN inversion.
- We improve the speed of the process by training a single network for each input text rather than solving an optimization problem per image.
- TD-GEM consists of a modulation network, acting in a disentangled semantic space, that allows changes in *e.g.,* color and sleeve length based on user requests.

2 Related work

2.1 Text-Conditioned Image Editing

There is a vast array of studies on image editing in the literature [14] for various datasets, but we focus on using text prompts as input in image manipulation here, like [22]. Starting from [2] that employs natural language description to synthesize images conditioned on the given text and image embeddings, several works [17, 27] enhance the quality of synthesized images and disentangle visual attributes. To make it more concrete, [17] proposes text-adaptive GAN, creating world-level local discriminators based on the text prompt to represent a visual attribute. It leads to generating images that only modify regions associated with the given text. Another line of work [27] focuses on semantic editing using a novel GAN called ManiGAN to preserve irrelevant content in the input image. It entails two modules to initially select regions relevant to the input text and then refine missing contents of the synthetic image. It is worth noting that ManiGAN only applies to the CUB and COCO datasets. InterFaceGAN [28] interprets a GAN model for disentangled and controllable face representation and identifies facial semantics encoded in the latent space. Although a simple, effective approach for face editing, it aligns attributes with a linear subspace of the latent space resulting in failure for long-distance manipulation.

An alternative perspective on this area of research is to manipulate the image using visual-semantic alignment or image-text matching rather than word-level training feedback. In this approach, semantics are mapped from text to images. Aligned to this strategy, TediGAN [32] generates diverse and high-quality images using a control structure based on style-mixing with multi-modal inputs such as sketches or text prompts. It can manipulate images with particular attributes through the common latent space of input text and images. Another improved approach to discovering semantically latent manipulation without using an annotated collection of images is explored by StyleCLIP [19]. It develops a text-guided latent manipulation for StyleGAN image manipulation, using CLIP in an optimization scheme as a loss network. Benefiting from image text representation like StyleCLIP, human hair editing is introduced by [30] referred to as HairCLIP. It trains a mapper network to map the input references into embedded latent code and exploits the text encoder and image decoder of CLIP. Recently, the latent mapping between the StyleGAN latent space and the text embedding space of CLIP is designed in [34], introducing Free-Form CLIP (FFCLIP) to handle free-form text prompts. It leverages input text with multiple semantic meanings to edit images. Nevertheless, it is worth noting that certain disentanglement challenges remain due to the presence of human biases [6].

2.2 Image Synthesis and Editing in Fashion domain

Manipulating images in the fashion industry becomes increasingly complicated when dealing with the full human body, compared to editing images of specific body parts, like the face or hair. Typically, an image editing pipeline involves translating an input image into a latent space representation using inversion techniques, followed by decoding the modified latent representation to generate an output image [5]. In the context of the fashion domain, another essential aspect to consider is establishing a proper mapping from the garment to the human body while preserving the identity of humans and the rest of the fashion items in the original images. The pioneering work of [33] known as FashionGAN approaches end-to-end virtual garment display by training a conditional GAN [16]. FashionGAN trains an encoder using

the real fabric pattern image, which results in the latent vector containing solely the material and color information of the fabric pattern image. Afterwards, a supplementary local loss module is integrated to regulate the texture synthesis process carried out by the generator. Another concurrent work to FashionGAN, [24], provides a rich dataset, including extensive annotations called Fashion-Gen applied for the text-to-image task. However, the quality of synthesized images using the StackGAN method is blurry, especially for a face. Another aforementioned dataset (SHHQ) collected by [6] is utilized for StyleGAN-based structures. Their investigation focused on analyzing how various factors, such as the size of the data set, the distribution of the data, and the alignment of the person in the image, influence the quality of generated images. To validate editing techniques with this dataset, SOTA facial StyleGAN-based architectures like InterFaceGAN [28], StyleSpace [31], and SeFa [26] are evaluated. Text2Human [8] also uses the SHHQ dataset and generates synthesized images by applying a human posture, a textual description of the garment's texture and shape as inputs. To encode the images, they implemented a hierarchical Vector-Quantized Variational Autoencoder (VQVAE) framework [4] with a texture-aware codebook. In contrast to earlier research, which constrains the verbal proficiency of the input text owing to sparsing the text into a closed set of categories, a recent work named FICE [20] suggests a latent-code regularization approach using a text-conditioned editing model. While the FICE model performs high-quality image editing, experiments on full-body human images are not explored.

3 Fashion Image Editing using proposed TD-GEM

To achieve image manipulation, it is necessary to obtain a latent representation of the source image within the latent space, which can be carried out by GAN inversion. Then images can be edited using methodologies such as latent optimizer, StyleCLIP mapper, or our proposed mapper network. Our approach involves a two-stage process for image editing, wherein we explore the surrounding area of the latent code to identify a latent representation corresponding to the edited image through a loss function. Once this representation has been determined, we input the code into the GAN architecture to generate the desired edited image. Further information on GAN inversion and other editing platforms, including latent optimizer and StyleCLIP mapper, can be found in the Supplementary Materials.

In this paper, we introduce TD-GEM as an innovative approach to manipulating garment attributes using a single mapper network. We drew inspiration from the HairCLIP technique developed by [30], which enhanced previous works such as StyleCLIP. HairCLIP uses additional loss functions and changes the mapper architecture to enable accurate attribute manipulation while preserving irrelevant parts of the image.

TD-GEM extends HairCLIP's capabilities by adapting the loss functions to the fashion domain allowing for simultaneous editing of both the length of the sleeves and the color of the clothing using a single mapper network. The mapper's definition is expressed as $M = (M_c, M_m, M_f)$. The structure of the mapper is comprised of three distinct sub-modules, each represented as M_c, M_m, and M_f. These sub-modules correspond to the varying degrees of detail present in the images that are generated (Figure 2). We inject information about the form and shape of the clothing (t_s) into all three sub-modules and color information (t_c) into the final one. This is in contrast to HairCLIP, where hairstyle information is injected into the first two layers and color information into the last one. Our mapper receives the latent code of an image through a GAN inversion operation, such as PTI, and text embeddings obtained by feeding the textual description into the encoder of a pre-trained CLIP network.

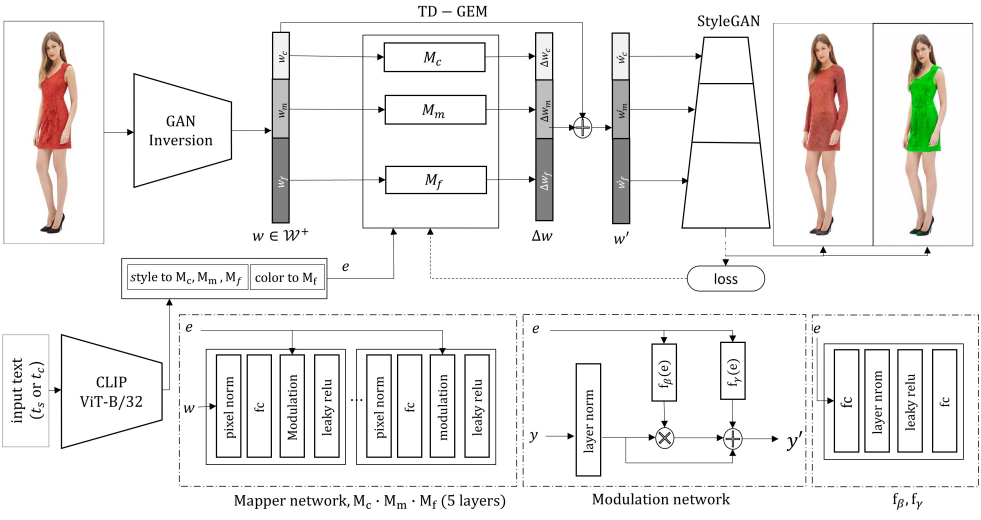

Figure 2: The TD-GEM architecture includes an input image that undergoes inversion via PTI. The output is then passed through the mapper network to obtain a residual, Δw. The mapper network is composed of three parts, all of which receive text-conditioned input related to the clothing's form and shape. The final part also receives color-conditioned information. The resulting latent code, w', is subsequently fed into a pre-trained StyleGAN generator to produce the edited image. The loss function is designed to modify the image's attributes as described in the text while preserving the irrelevant parts.

Based on this methodology, the mapper produces a residual latent code Δw. This code is then added to the original latent code of the image, resulting in $w' = w + \Delta w$, and provided to the StyleGAN architecture to generate the edited image.

Each sub-mapper consists of five layers, each consisting of a pixel-norm, fully connected layer, modulation layer, and leaky ReLU activation. The modulation layer of the network encodes the text input information and receives the information from text embeddings and the previous fully connected layer [30]. It processes the text input:

$$y' = 1 + f_\gamma(e)\frac{y - \mu_y}{\sigma_y} + f_\beta(e) \tag{1}$$

where y' in the output of the modulation network, e refers to the text embedding obtained from the input text t, and the parameters μ_y and σ_y represent the mean and standard deviation of the intermediate feature y. The neural networks f_γ and f_β consist of a fully connected layer with layer normalization and leaky ReLU activation, and another fully connected layer. The modulation layer provides semantic alignment and its architecture is illustrated in Figure 2.

To enhance the manipulation accuracy and encourage disentanglement during the editing process, the following loss function is utilized for training the TD-GEM mapper network.

$$\mathcal{L}_t = \lambda_{\text{CLIP}}\mathcal{L}_{\text{CLIP}} + \lambda_2\mathcal{L}_{norm} + \lambda_{ID}\mathcal{L}_{ID} + \lambda_{color}\mathcal{L}_{color} + \lambda_{BG}\mathcal{L}_{BG} \tag{2}$$

The CLIP and identity loss functions, $\mathcal{L}_{\text{CLIP}}, \mathcal{L}_{ID}$, are as prescribed in [19]. The locality of

the edit is ensured by the $\mathcal{L}_{norm}$ loss as

$$\mathcal{L}_{norm} = \|M(w, E_{\text{CLIP}}(t))\|_2, \tag{3}$$

The color composition of the clothing is maintained by introducing a color loss $\mathcal{L}_{color}$. A pre-trained parsing network P [1] is employed to segment the human instance into foreground and background parts. The foreground area of the image comprises the shirt, dress, coats, neck, and arms, and the rest is assumed as the background. The color loss is only applied to the foreground as

$$\mathcal{L}_{color} = \|avg(G(w') \cdot P(G(w'))) - avg(G(w) \cdot P(G(w)))\|_1, \tag{4}$$

where $\|\cdot\|_1$ is the L_1 norm, $G(w)$ is the original image, $G(w')$ is the edited image, $G() \cdot P()$ is the foreground and avg is the mean value for each channel. The image is first transformed from RGB color to XYZ and then to LAB coordinates to obtain the average.

To ensure that irrelevant parts of the image are kept untouched, a background loss is applied. It is defined as

$$\mathcal{L}_{BG} = \|(G(w') - G(w)) \cdot (\neg P(G(w')) \cap \neg P(G(w)))\|_2. \tag{5}$$

where $\|\cdot\|_2$ denotes the L_2 norm. During training, we obtain foreground masks for both the original and edited images at each iteration. We combine these masks to obtain a union, which represents the editable area, while the rest of the image constitutes the background. Mathematically, we express the background as $\neg(P(G(w')) \cup P(G(w)))$, which is equivalent to $\neg P(G(w')) \cap \neg P(G(w)))$. By applying the L_2 norm in the background loss, we ensure that the original and edited images are similar for the areas included in the background.

4 Experiments

4.1 Dataset

We employ the SHHQ dataset for the experiments in this paper. It consists of high-quality, full-body fashion, human-centric images. It contains $230K$ fashion images in diverse poses and textures, with a resolution of 1024×512 pixels. However, only $40K$ images are presently accessible to external researchers. We have selected 2200 samples with mostly short-sleeved or sleeveless attributes to either lengthen sleeves or change the color. The dataset is split into a 90/10 ratio for training and testing, using 2000 samples for training and 200 for testing.

4.2 Implementation Detail

For this experiment, the development is carried out within a dockerized environment utilizing a NVIDIA GeForce 3090 GPU with 24GB VRAM. The code is implemented using PyTorch 1.9.1 and Cuda 11.4. In this work, we train a mapper, M, to manipulate images, leveraging a pre-trained StyleGAN2-ADA generator G [11], a pre-trained CLIP model [21], and a pre-trained parsing network, P [1].

4.3 Comparisons and Evaluation

This section presents the results of TD-GEM for the manipulation of full-body human images in the fashion domain. To assess the preservation of the shape and patterns of the clothing,

Figure 3: The TD-GEM network is employed to alter the length of the sleeves and the color of the clothing to blue and green.

several samples are presented in Figure 3. The last two experiments are conducted where the color is changed to blue and green, respectively (last two rows). The body configurations, faces, and hairs are well preserved, while the garment color is changed to blue. The degree of color saturation varies, but the color attribute is successfully modified in all images. Color leakage is observed in one image, where the trousers are painted green (Figure 3b). The results indicate that the patterns of garments are well-preserved, even for complicated patterns with more detailed garments, as seen in Figure 3(d-h), where the garment has a more complicated shape and wrinkles.

A quantitative comparison is made between different approaches, including the latent optimizer, StyleCLIP mapper, and TD-GEM, all in the context of sleeve lengthening. The outcomes of this comparison, using four distinct measures, are presented in Table 1.

The scores for the background represent the degree of disentanglement in the image manipulation. In the TD-GEM case, the background is preserved with the same quality as the StyleCLIP mapper case, with the Fréchet Inception Distance (FID) score of $O(10^{-2})$ for both cases. Although the FID score in TD-GEM is worse than the StyleCLIP mapper case, the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR) scores show an improvement. The color composition in both cases is almost the same, where the better results belong to the TD-GEM network. The evaluation of the image manipulation in the foreground is performed based on the qualitative results, as the metrics do not accurately

reflect the editing quality occurring in the foreground.

We also investigate the advantages of incorporating semantic injection into the fine module, as presented in Table 2. This approach enhances the preservation of background details in the edited image, as indicated by the improvement across all metrics. For a more detailed analysis, please refer to the supplementary materials.

A comparison between the TD-GEM and supervised and unsupervised editing methods by [5] is shown in Figure 4. Two samples from the test dataset are edited using four approaches, and the results are illustrated in the top and bottom rows. InterFaceGAN radically changes the shape and form of the garment, while StyleSpace provides a better solution but still has some issues with the form of the clothing. SeFa presents promising results, but the disentanglement property is not successfully enforced. The shape of the shoes is not preserved, as can be seen in Figure 4b. On the other hand, the TD-GEM performs the best by successfully lengthening sleeves up to the wrist and keeping the other unrelated attributes untouched. The quantitative scores for TD-GEM are compared with the three aforementioned methods in Table 3. FID scores show an order of magnitude better performance for TD-GEM in the background region. The SSIM and PSNR scores are also substantially higher for our proposed approach. Furthermore, ACD scores in TD-GEM are superior to InterFaceGAN and SeFa, and similar to StyleSpace. It's important to highlight that our measurement scores may not provide a precise assessment of the quality in the foreground; hence the comparison was conducted using the outcomes of qualitative analysis. As per our analysis, the approach we proposed yields superior outcomes when compared to the baseline methodologies.

5 Discussion and Conclusion

This paper has presented a novel approach for full-body human fashion image editing through textual input descriptions. The image manipulation process involves two stages: firstly, obtaining a latent representation of the image in the latent space of a pre-trained network, and secondly, editing the image by navigating semantically along with the relevant directions using a pre-trained language model CLIP. To this end, we employed PTI as a GAN inversion technique, given its accurate result. To proceed with attribute editing, our proposed TD-GEM can manipulate the sleeve length and color of a garment, integrating new

Table 1: Comparison of different networks in the context of lengthening sleeves

Methods	Sec.	FID ↓	SSIM ↑	PSNR ↑	ACD ↓
Latent Optimizer [19]	Back.	0.126	0.853	17.059	0.278
StyleCLIP Mapper [19]	Back.	0.021	0.919	24.293	0.165
TD-GEM	Back.	0.030	0.935	27.543	0.146

Table 2: The quantitative scores for the case with and without injection in the fine module for the lengthening sleeves

Method	Sec.	FID ↓	SSIM ↑	PSNR ↑	ACD ↓
TD-GEM	Back.	**0.030**	**0.935**	**27.543**	**0.146**
w/o fine injection	Back.	0.089	0.925	26.883	0.209

Figure 4: The figure compares various methodologies for manipulating human clothing, including InterFaceGAN, StyleSpace, SeFa, and the TD-GEM network. The first three methods are based on the work by [5]. Two samples from the testing dataset are shown in the two rows (a) and (b).

Table 3: The quantitative comparison through lengthening sleeves. TD-GEM exhibits superior performance in both foreground and background areas.

Method	Sec.	FID ↓	SSIM ↑	PSNR ↑	ACD ↓
TD-GEM	Back.	0.030	0.935	27.543	0.146
InterFaceGAN [28]	Back.	0.199	0.864	16.794	0.712
StyleSpace [31]	Back.	0.102	0.898	22.725	0.137
SeFa [26]	Back.	0.176	0.882	20.540	0.244

loss functions that accomplish full-body human image editing. We discovered that incorporating semantic injection into the fine-mapper enhances image editing outcomes, while the impact on identity loss remains relatively insignificant. Extensive experiments demonstrate that our solution can achieve high-quality fashion image editing results. It outperforms competing methods, including the latent optimizer and the StyleCLIP mapper network, in terms of accuracy and reduction of computational complexity. This is achieved by training only one network for color or sleeve-length textual descriptions. Nonetheless, text-conditioned fashion image editing still requires further exploration, particularly with complex pattern clothing. Furthermore, the findings from our study will serve as a benchmark for future processes involving full-body human fashion image editing. This foundation can be expanded upon using large text-image fashion datasets, incorporating diverse textual prompts such as "shortening sleeve length," "adding stripes," and "incorporating patterns," among others.

References

[1] Fuwei Zhao Bowen Wu. 2d-human-parsing, 2021.

[2] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE international conference on computer vision*, pages 5706–5714, 2017. doi: 10.1109/ICCV37128.2017.

[3] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9026–9035, 2019. doi: 10.1109/ICCV.2019.00912.

[4] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34:3518–3532, 2021.

[5] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 1–19. Springer, 2022. doi: 10.1007/978-3-031-19787-1_1.

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[8] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

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

[10] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

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

[12] Anika Kozlowski, Michal Bardecki, and Cory Searcy. Environmental impacts in the fashion industry: A life-cycle and stakeholder framework. *Journal of Corporate Citizenship*, (45):17–36, 2012. doi: 10.9774/gleaf.4700.2012.sp.00004.

[13] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. doi: 10.1109/CVPR42600.2020.00559.

[14] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. *Advances in Neural Information Processing Systems*, 34:16331–16345, 2021.

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. doi: 10.1109/CVPR52688.2022.01167.

[16] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[17] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *Advances in neural information processing systems*, 31, 2018.

[18] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. Image based virtual try-on network from unpaired data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5184–5193, 2020. doi: 10.1109/CVPR42600.2020.00523.

[19] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. doi: 10.1109/ICCV48922.2021.00209.

[20] Martin Pernuš, Clinton Fookes, Vitomir Štruc, and Simon Dobrišek. Fice: Text-conditioned fashion image editing with guided gan inversion. *arXiv preprint arXiv:2301.02110*, 2023. doi: 10.48550/arXiv.2301.02110.

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

[22] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[23] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1): 1–13, 2022. doi: 10.1145/3544777.

[24] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. doi: 10.48550/arXiv.1806.08317.

[25] Sanaz Sabzevari., Ali Ghadirzadeh., Mårten Björkman., and Danica Kragic. Pg-3dvton: Pose-guided 3d virtual try-on network. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,*, pages 819–829. INSTICC, SciTePress, 2023. doi: 10.5220/0011658100003417.

[26] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1532–1540, 2021. doi: 10.1109/CVPR46437.2021.00158.

[27] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. doi: 10.1109/CVPR42600.2020.00926.

[28] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020.

[29] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018. doi: 10.1007/978-3-030-01261-8_36.

[30] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. Hairclip: Design your hair by text and reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18072–18081, 2022. doi: 10.1109/CVPR52688.2022.01754.

[31] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. doi: 10.1109/CVPR46437.2021.01267.

[32] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. doi: 10.1109/CVPR46437.2021.00229.

[33] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE international conference on computer vision*, pages 1680–1688, 2017. doi: 10.1109/ICCV.2017.186.

[34] Yiming Zhu, Hongyu Liu, Yibing Song, Xintong Han, Chun Yuan, Qifeng Chen, Jue Wang, et al. One model to edit them all: Free-form text-driven image manipulation with semantic modulations. *arXiv preprint arXiv:2210.07883*, 2022.

Supplementary Material

A Explanation editing

A.1 GAN Inversion

A crucial aspect of an effective inversion approach is its ability to balance the trade-off between distortion and editability. Specifically, the method should be capable of preserving the original appearance of an image (*i.e.,* low distortion) while enabling convincing attribute modifications (*i.e.,* high editability). One such method that claims to achieve this goal is PTI, as introduced by [23]. This technique employs an off-the-shelf encoder, such as e4e, to derive a latent code for the StyleGAN architecture. However, the encoder's output can result in distortion in the reconstructed image compared to the original, which is known as the identity gap. To address this issue, they fine-tuned the generator to preserve the image's identity. The process of adjustment can be analogized to the act of aiming a dart towards a target and subsequently realigning the board to account for a near-miss.

In this paper, we utilized a pre-trained e4e encoder for the inversion, as provided by [5], with further fine-tuning of the generator based on a specific loss term. The loss function $\mathcal{L}(\theta)$ was defined as the sum of the learned perceptual image patch similarity loss function ($\mathcal{L}_{LPIPS}$) and the pixel-wise mean square error ($\mathcal{L}_2$),

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{LPIPS}(x_i, G(w_i:\theta)) + \lambda_2 \mathcal{L}_2(x_i, G(w_i:\theta)), \tag{A1}$$

with a hyperparameter λ_2 set to 1. The aim of the optimization is to determine the optimal parameters θ^* for the generator G, based on the output of the e4e encoder w_i for each image x_i in the dataset of size N. We used the AlexNet network to calculate the perceptual loss, with the learning rate set to 5×10^{-4}, a maximum number of iterations, 3500, and a convergence tolerance of 10^{-4}.

A.2 Latent optimizer

The latent optimizer framework is an image manipulation approach that relies solely on solving a direct optimization problem [19]. This framework uses the GAN inversion to first invert an image into a latent code. Then, an optimization problem is solved using a loss function to find the latent code residual. The residual is added to the original latent representation and fed into the StyleGAN to obtain the edited image (Figure A1).

The loss function is as follows:

$$\mathcal{L}_t = \lambda_{\text{CLIP}}\mathcal{L}_{\text{CLIP}} + \lambda_2\|\Delta w\|_2 + \lambda_{ID}\mathcal{L}_{ID} \tag{A2}$$

The clip loss $\mathcal{L}_{\text{CLIP}}$ is designed to guide the optimization process toward achieving the attribute described in the input text. To accomplish this, the embeddings of both the input text (t) and the generated image ($G(w')$) are obtained in a shared space using the pre-trained CLIP encoder, and their cosine similarity is considered in the loss function:

$$\mathcal{L}_{\text{CLIP}} = 1 - cos(E_{\text{CLIP}}(G(w')), E_{\text{CLIP}}(t)), \tag{A3}$$

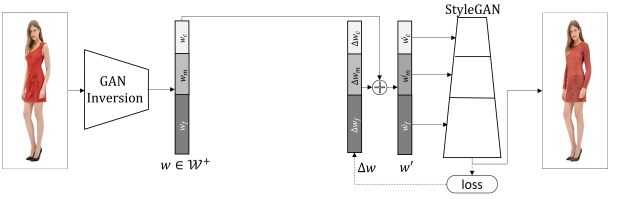

Figure A1: The manipulation of garments through lengthening sleeves via the use of the latent optimizer approach.

The term $\|\Delta w\|_2$ ensures exploration within the vicinity of the original latent representation. This term guarantees a localized manipulation of the initial image.

To preserve the identity of the original image, an identity loss term $\mathcal{L}_{ID}$ is used. It calculates the mean square error between the features of the original and edited images obtained from the last layer of a pre-trained ConvNeXt network.

$$\mathcal{L}_{ID} = MSE(R(G(w)), R(G(w'))). \tag{A4}$$

where MSE is the mean square error, R is the pre-trained ConvNeXt network [15], and $R(G(w))$ and $R(G(w'))$ are the features of the ground truth and generated images, respectively.

A.3 StyleCLIP Mapper

The latent optimizer framework is not efficient in terms of image editing due to the need to solve an optimization problem for each image. To address this issue, [19] introduced a mapper network that can infer the manipulated image based on a given input text, making the process more efficient. The mapper network is first trained on the training dataset. It is then used to edit new images from the testing dataset. The architecture of the mapper is designed with three distinct sub-modules, each responsible for different aspects of the generated image (Figure A2). These sub-modules are divided into coarse, medium, and fine clusters, which control the corresponding structures in the image. A more detailed

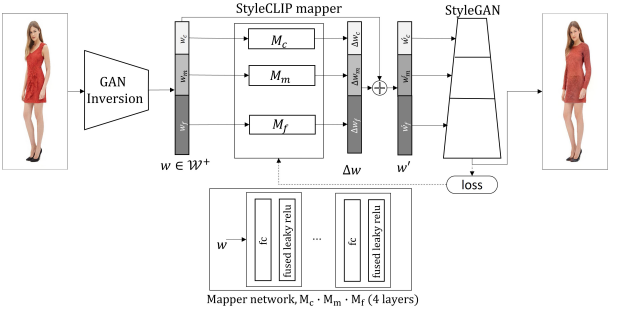

Figure A2: The StyleCLIP mapper network is utilized to increase the length of sleeves.

description of the architecture can be found in [19]. The input image is first inverted during editing, and the resulting latent code is fed into the mapper to obtain a residual latent code. This residual latent code is then added to the original latent code and passed through the pre-trained StyleGAN generator to produce the edited image. The loss function is defined as:

$$\mathcal{L}_t = \lambda_{\text{CLIP}}\mathcal{L}_{\text{CLIP}} + \lambda_2\|M_t(w)\|_2 + \lambda_{ID}\mathcal{L}_{ID} \tag{A5}$$

where $M_t(w)$ is the output of the mapper designed for the input text, t. The clip and identity losses are the same as described in section A.2.

B Detailed evaluation

B.1 Pivotal Tuning Inversion (PTI)

This section presents the results of PTI, one of the GAN inversion approaches. Figure B3 displays a selection of sample images obtained using PTI. The method successfully preserves fine-grained details, such as facial features and hair, while maintaining the correct configuration of the human body, including the posture of hands, legs, and shoes. Additionally, the clothing attributes, including the shape, color, and patterns, are accurately preserved, with no missing fashion items in the inverted results. These findings demonstrate the effectiveness of PTI in achieving high-quality GAN inversion with a faithful representation of the input images' attributes. The quantitative analysis further supports the superiority of the PTI approach over e4e, as shown in Table B1. The FID scores decrease significantly from 0.245 to 0.005 compared to e4e, while SSIM scores improve by 11.3% from 0.836 to 0.943. The PSNR score also increases from 19.136 to 32.013, demonstrating a 40.2% improvement. Finally, color composition scores exhibit significantly enhanced performance, improving from 0.108 to 0.007.

B.2 Latent optimizer

In this section, we present the outcomes of image manipulation using the latent optimizer approach. Figure B4 displays a collection of sample images obtained by lengthening the sleeves of the garments using this approach. The text prompt for this operation is "A long sleeve." The generated images demonstrate successful attribute modification, with the sleeves appear-

Figure B3: This is a collection of qualitative outcomes for PTI single-mode inversion, where the top row displays the original image, and the bottom row exhibits the corresponding inverted ones.

Table B1: The quantitative results for PTI inversion are compared with e4e using different metrics.

Method	FID ↓	SSIM ↑	PSNR ↑	ACD ↓
e4e	0.245	0.836	19.136	0.108
PTI	0.005	0.943	32.013	0.007

Figure B4: The result for manipulating garments by lengthening sleeves with latent optimizer.

ing longer in all cases. However, extensive distortion is observed in the results. The shape of the heads and faces are changed in all images of Figure B4(a-h), and the body configuration, including the posture of hands, feet, and main body, is not conserved. For example, in Figure B4(a), the angle of the hands and the posture of the legs are different from the original image. In Figure B4(g), the shape of the shoes deviates from the original ones, and the pattern of the clothing is not similar. Nonetheless, the color composition in all images is successfully maintained.

It is possible to improve the quality of the results of the latent optimizer by introducing more losses. However, the main disadvantage of the method is that an optimization problem must be solved for each individual image, which is practically inconvenient. Therefore, we examine other methodologies that do not have this limitation in the next sections.

B.3 StyleCLIP

The StyleCLIP mapper network is evaluated, and the results show that the images are edited successfully, with the shape and configuration of hairs and faces maintained during editing (Figure B5). However, there are some small deviations, such as the shape of the fingers in the left hand not being the same as in the source image (Figure B5b). The color composition and pattern preservation are problematic, with the pattern of the clothing disappearing in some images and differences in color between the source and edited images being observed. The quantitative analysis reveals that the StyleCLIP mapper network has improved the background presentation (Table 1), with better FID, SSIM, PSNR scores, and color composition compared to the latent optimizer scheme. However, there are dramatic changes in the foreground in ACD scores, indicating inferior preservation of the color composition.

B.4 TD-GEM

The quantitative comparison of color and sleeve manipulations using TD-GEM is provided in Table B2. The scores for the background region are almost the same for both colors. The SSIM and PSNR values are also very close. The ACD scores in the foreground indicate a significant change in the color attribute, which shows the color attribute is successfully

Figure B5: The StyleCLIP mapper network is utilized to extend the length of the sleeves.

Table B2: TD-GEM network by editing sleeve and color

Text	Sec.	FID ↓	SSIM ↑	PSNR ↑	ACD ↓
sleeve	Back.	0.030	0.935	27.543	0.146
blue	Back.	0.017	0.956	31.607	0.191
green	Back.	0.030	0.956	31.608	0.294

modified. However, a higher value of the color score for the green color in the background could be related to the color leakage effect.

C Implementation details

Table C3 presents the coefficients for the loss function employed in this study. Notably, these coefficients differ when editing sleeve length and garment color, with a higher background coefficient used during color editing. The training process utilizes 100k steps (*e.g.*, 50 epochs) and a learning rate of 5×10^{-4} with a "Ranger" optimizer.

C.1 Ablation Study

This section investigates the efficacy of two key assumptions: identity loss and semantic injection across all layers.

We first analyze the impact of incorporating semantic injection into all three mappers while editing the sleeve length, as demonstrated in Figure C6. In scenario (a), the image generated without fine injection exhibit artificial vertical straw compared to the ground truth. However, this artifact disappears when all three mappers are employed. In scenario (b), the presence of all three mappers leads to better preservation of the wrinkle.

Next, we conduct a second ablation study focusing on the role of identity loss. Generally, identity loss has a minor influence on the outcomes. Its significance becomes more apparent during the color editing process. Figure C7 displays edited images both with and without the

Table C3: Coefficients corresponding to various terms within the loss function.

Case	λ_{CLIP}	λ_2	λ_{ID}	λ_{color}	λ_{BG}
sleeve	1.0	1.0	1.0	5×10^{-3}	0.3
color	1.0	1.0	1.0	5×10^{-3}	1

Figure C6: The effect of semantic injection into the fine mapper in lengthening the sleeves.

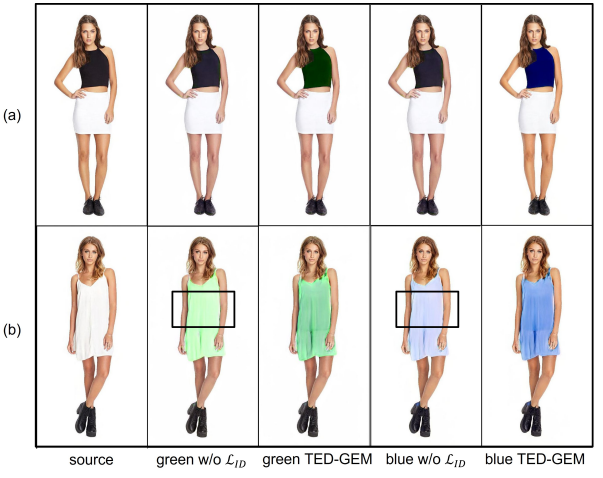

Figure C7: The effect of identity loss in editing the color.

identity term, taking into account the input description that specifies color alterations to either blue or green. When the identity loss is omitted, the color change appears muted in both scenarios (a) and (b). Conversely, the inclusion of identity loss results in more noticeable color variations.

D Additional Examples

Figure D8 shows a set of qualitative results from this method. The degree of disentanglement

Figure D8: The images are edited by TD-GEM network; the length of sleeves is increased.

in the generated images is quite acceptable, with preserved details in the person's configuration, including hands, legs, and main body, which have the same postures, shapes, and contours as the original images. In contrast to the previous StyleCLIP mapper network, the finger details are well preserved in the edited images, as seen in Figure D8b. Additionally, the faces and hair in the manipulated images are indistinguishable from the original images due to interpolation between the original and edited images, which improves the quality of those regions. The color composition in the TD-GEM network is remarkably better than the previous StyleCLIP mapper network, as seen in Figure D8h, where only slight deviations between the original yellow and edited yellow are observed.