THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學
Pao Yue-kong Library
包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

http://www.lib.polyu.edu.hk

DEEP LEARNING-BASED INTELLIGENT FASHION IMAGE GENERATION SYSTEM

LIAO FANGJIAN

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University
School of Fashion and Textiles

Deep Learning-based Intelligent Fashion Image Generation System

Liao Fangjian

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

March 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

Liao Fangjian (Name of student)

Abstract

A Fashion image generation engine is a project that specializes in enhancing writing and elucidating concepts by providing visual representations that align with the associated text or idea. However, translating drawings into effective visual representations can often appear as a particularly challenging task, and to my knowledge, there are no similar techniques or tools available in the market that encompass the same combination of capabilities and benefits to assist designers in accelerating creation. This tool aims not only to generate novel concepts for designers but also to ignite inspiration. Simultaneously, the engine seeks to enhance the efficiency of AIGC tools and bring new trends to the academic field.

In the present scenario, deep learning networks have made significant strides in various domains, including traditional object detection, object segmentation, image pose transformation, and text generation models. However, when it comes to applying these advanced technologies in the field of fashion, it poses unique and complex challenges since the scenario is different. Currently, research efforts in the field of image-to-image translation have primarily focused on translating the details and appearances of objects. However, there is insufficient research dedicated to translating painting styles, with limited ability to preserve the essence of the style translation. Additionally, when it comes to uninformative images such as illustrations, conducting effective translation using existing methods remains a challenging task. The designing process includes several key steps such as ideation, sketching, refinement, and more. During this process, the illustrator faces the challenge of finding creative solutions to effectively simplify or visually communicate complex ideas. Additionally, illustration work often operates under tight deadlines, making efficient time management while maintaining high-quality output a challenge for illustrators.

This thesis aims to develop an intelligent fashion image generation system to overcome the aforementioned limitations. First, an automatic drawing image generation engine referencing tops and bottoms is proposed. The engine is combined with a deep neural network focused on keypoint detection and clothing segmentation, enabling effective and quick pixel-level clothing mapping. Keypoint mapping between the detected clothing and the model enables true virtual try-on. Additionally, with the assistance of the segmentation model, the garment can be better fitted on the model when the front and back pieces of the clothing are presented as separate elements.

Secondly, a novel pipeline of drawings-to-images-to-illustrations generation is proposed to spark the inspiration of creation. A state-of-the-art generation model is applied to generate images with the reference of the generated drawing images. In detail, the boundary of the drawing images is extracted as a conditional image, and the generated model combines the extracted conditional image with the drawing image to generate product images in the real domain. Face refinement will be applied to adjust the flaws in the face part of the generated image in the second stage. Additionally, a novel fashion image-to-image generation method named Uni-Duallora is proposed, which optimizes generative capabilities and reduces the number of additional parameters required, to obtain illustrative images.

Thirdly, a novel method is introduced to achieve pose-guided runway image generation. This method leverages the advantages of attention and affine transformation operations, while introducing a novel confidence map in the attention operation to enhance the performance of the image synthesis. The hierarchical structure is applied to generate the final runway image with the conditioned pose.

The contributions of this thesis lie in developing and advancing an intelligent fashion image generation engine. This comprehensive framework not only streamlines the workflow for professionals but also provides a well of inspiration. Simultaneously, its objective is to broaden the design enjoyment realm to individuals without specialized expertise. This thesis also proposes a comprehensive survey on human pose transfer, including its limitations, and provides insights for further exploration.

Publications arising from the thesis

1. Fangjian Liao, Xingxing Zou, and Waikeung Wong. Uni-dllora: Style fine-tuning for fashion image translation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6404–6413, 2024

2. Fangjian Liao, Xingxing Zou, and Waikeung Wong. Deep fabric prints generation for fashion. *Design and Semantics of Form and Movement*, 88, 2023

3. Kaicheng Pang, Xingxing Zou, Fangjian Liao, and Waikeung Wong. M-vton: Multi-layer virtual try-on system. *Design and Semantics of Form and Movement*, 266, 2023

4. Fangjian Liao, Xingxing Zou, and Waikeung Wong. Appearance and pose-guided human generation: A survey. *ACM Computing Surveys*, 56(5):1–35, 2024

5. Fangjian Liao, Xingxing Zou, and Wai Keung Wong. Attentional pixel-wise deformation for pose-based human image generation. *Expert Systems with Applications*, 246:123073, 2024

6. Fangjian Liao, Xingxing Zou, and Waikeung Wong. Minigan: Toward informative and uninformative image transferring. *AATCC Journal of Research*, page 24723444221136635, 2023

Acknowledgements

I am deeply honored to pursue my academic career within such an outstanding university and research team. First and foremost, I would like to express my sincerest gratitude to my supervisor, Prof. WaiKeung Wong, for his professional guidance, invaluable insights, and unwavering support throughout my doctoral journey. His academic advice and steadfast support have been immensely helpful to me. His encouragement has also driven me to continually strive for excellence.

Moreover, I am also deeply grateful to the faculty and staff in the school of Fashion and Textiles for providing me with the indispensable research environment and precious, abundant learning resources during my PhD studies.

I would like to thank my colleagues and fellow researchers, including Dr. Xingxing Zou, Dr. Dongmei Mo, Miss Shumin Zhu, Mr. Sibo Wang, as well as my junior and senior colleagues who have worked alongside me in the office. I would also like to express my gratitude to the friends who have supported me every step of the way.

Furthermore, I would like to extend my heartfelt thanks to my family, especially my wife, Mrs. Lifei Yan. Without her help, I would not have been able to actively face the various difficulties encountered during this arduous journey. It is the love, encouragement, and understanding of my family that have given me the strength and determination to overcome the difficulties. I am deeply grateful for the sacrifices and unwavering support my family has provided during my doctoral studies.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Background

The fashion industry is one of the largest and most influential sectors globally, generating trillions of dollars in revenue each year [6]. Beyond its economic impact, fashion is an art form allowing individuals to express their identity. For brand designers and industry professionals, creating new outfits serves as a medium to convey both personal and brand philosophies and values. However, digital illustration tools present both new opportunities and challenges for designers and brands [224]. On the one hand, it is widely held that within the process of fashion design, the integration of computational tools has the potential to alleviate the pressures and burdens encountered during creation [225]. On the other hand, the potentially complex operational procedures may significantly increase the learning curve for brands and designers.

To address these opportunities and challenges, integrating AI into design and production tools has emerged as a pressing need with significant market potential. As shown in the Figure 1.1, the current projected market input is expected to reach $770.9 billion , with an annual growth rate of 8.94%, resulting in a projected market volume of $1,183 billion by 2029. Additionally, the number of users in the fashion market is expected to reach 2.8 billion by 2029. User penetration is projected to be 33.3% in 2024 and is anticipated to increase to 37.8% by 2029.

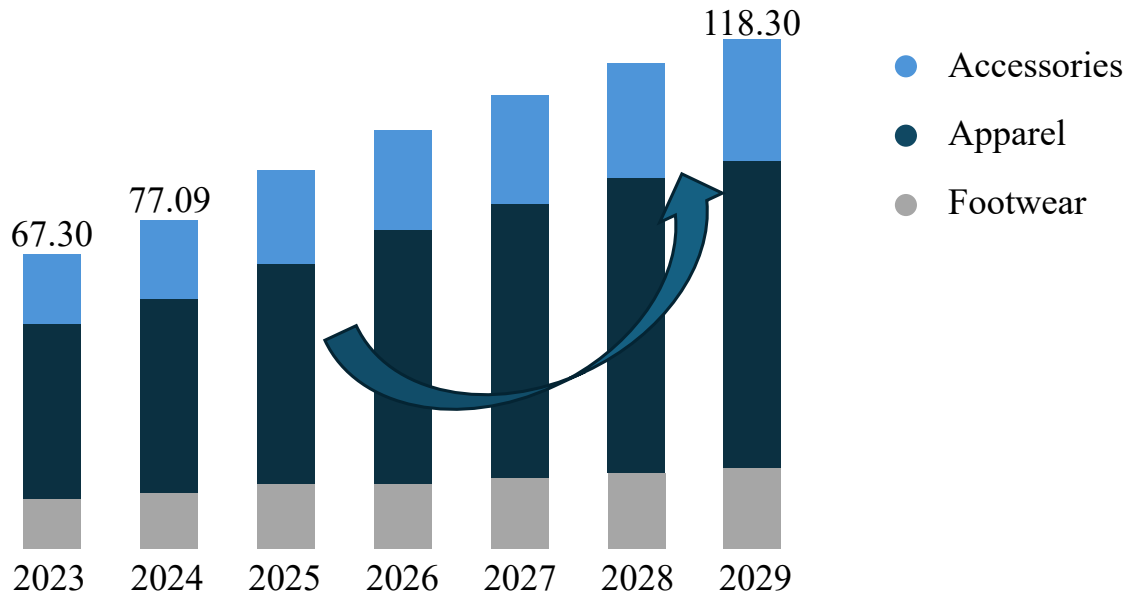
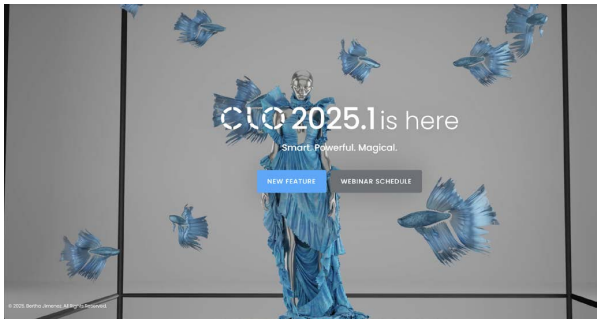


Figure 1.1: The market size of fashion grow with projected to reach $1,183 billion by 2029, with an annual growth rate of 8.94%.

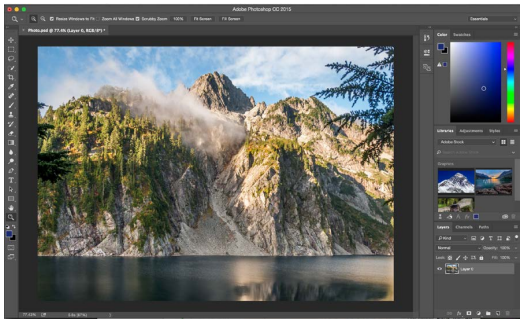
With the advent of the AI era, novel AI capabilities have the potential to inject new vitality into traditional design paradigms. AI is a multimodal field that spans areas such as computer vision and natural language processing and human-computer interaction [14]. Currently, fashion design research topics can be categorized into two main areas:

1. **Image Attribute Analysis**: This encompasses several aspects, including *Attribute Recognition* [222, 16, 191], which extracts fashion attributes from given fashion items. *Fashion Parsing Analysis* [39, 150, 182], which involves segmenting elements such as clothing, trousers, and human bodies with different labels. *Landmark Detection* [112, 113, 191] where the objective of the deep network is to locate the positions of key points defined on garments.

2. **Image Generation**: This area includes techniques such as style transfer [158, 15]. It also involves appearance and pose-guided image generation [69, 104], which generates images while preserving the target person and their desired postures. Additionally, human image animation [70] assists designers in better showcasing their expressions.



Homepage of Clo3D



Adobe Photoshop interface

Figure 1.2: Leading digital fashion design tools: Adobe Photoshop and Clo3D platforms.

Among the existing digital tools, Adobe software (such as Adobe Illustrator and Photoshop) and 3D garment simulation platforms like Clo3D shown in Figure 1.2 have become widely adopted in the fashion industry. These tools offer significant advantages: Adobe provides powerful vector and raster editing capabilities, enabling designers to create detailed sketches and manipulate images with high precision. Clo3D, on the other hand, allows for realistic 3D garment visualization, virtual fitting, and fabric simulation, which greatly enhances the efficiency of prototyping and reduces the need for physical samples.

However, both Adobe and Clo3D present notable limitations. Adobe tools require substantial manual effort and a steep learning curve, making the design process time-consuming and less accessible for beginners. Clo3D, while effective for 3D visualization, often demands high computational resources and specialized expertise. Interviews are conducted with professional fashion designers regarding their current design workflows. These discussions revealed that existing commercial software, while powerful, still demands significant time investment for basic design tasks. More importantly, neither platform offers automated solutions for integrating AI-driven content generation, such as automatic sketch creation, style transfer, or pose-guided image synthesis. This gap highlights the need for intelligent systems that can bridge the divide between traditional design workflows and advanced AI capabilities, ultimately empowering designers with more efficient and creative tools.Specifically, the primary aim is to develop an automated hand-drawing sketch generation engine, enabling designers to swiftly and accurately generate sketches

without the need for manual intervention. The second objective involves creating an image-to-image translation model capable of transforming uninformative images, like runway images or illustrations, into diverse styles. Lastly, the third model focuses on achieving pose-guided human image translation not only in the real image domain but also in illustrations and hand-drawing sketches, enhancing productivity by automatically generating images from various viewpoints. These three directions will be further expounded upon in the following paragraph.

1.1.1 Fashion Image Generation Modeling

Currently, tasks involving hand-drawing primarily rely on manual methods, often employing tools like Adobe software, to achieve the final sketch. However, this approach is both time-consuming and labor-intensive. In the current landscape, deep learning techniques have the capacity to acquire low-dimensional information like landmarks from clothing. Some methods [113, 196, 35, 135, 13, 36, 12, 86, 191, 93] focus on landmark detection to better analyze clothing in low-level visual features. Other methods [208, 24, 189, 167, 222, 4, 204, 1, 131] analyze the attributes and parsing images of the clothing to obtain high-level information. However, Up until the present moment, a methodology for the automated generation of hand-drawn sketches applying the acquired keypoints and segmentation from deep learning models remains conspicuously absent.

1.1.2 Fashion Image Style Translation Modeling

The task of fashion image-to-image translation involves the transformation of an input image from one visual domain to another. This intricate process aims to create a meaningful relationship between the source and target domains shown in Figure 1.3, ensuring that crucial perceptual features are retained while converting the image's overall appearance. As one of the most prominent directions within the realm of generative models, several methods [62, 107, 219, 218] have focused on this area and achieved remarkably impressive outcomes. The task of fashion image-to-image translation can generally be roughly categorized into two approaches: one

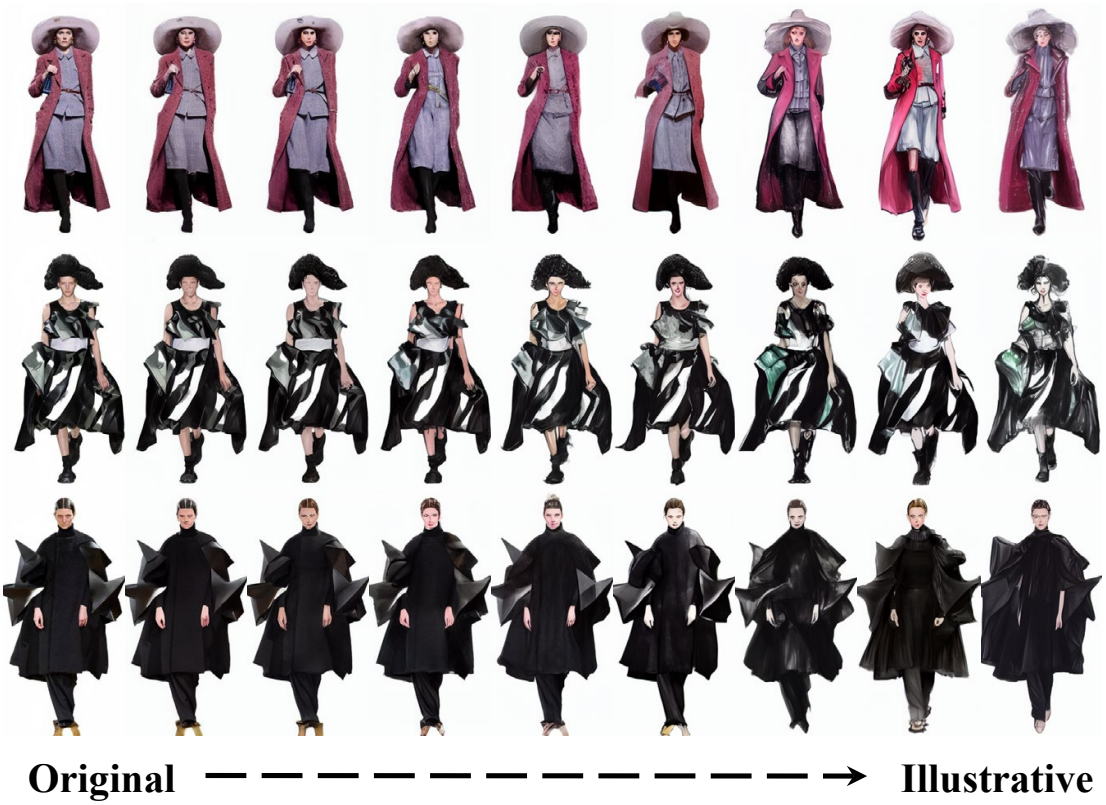


Figure 1.3: Images synthesized by combining a source image with varying style strengths. Generated images progressively carry the illustrative style.

relies on GAN-based approaches [138, 73, 74], while the other employs Diffusion-based methods [49, 127] for image generation. For the GAN-based approach, recently several methods [146, 132] are proposed to translate images from domain A to domain B. They achieved commendable image quality by preserving the original image's appearance information, while also capturing the style or semantics of target information. For Diffusion-based methods, several laboratories [176, 134, 88] have introduced innovative models to address this task. For diffusion-based approaches, there is an enhancement in the quality of generated images, coupled with an augmentation in the presence of more pronounced semantic information. However, fashion illustration translation represents a task within the realm of fashion image-to-image translation, characterized by its distinct attribute of having limited available information and demanding a high degree of precision. Certain existing methods have demonstrated limitations in achieving satisfactory outcomes in this particular context.

Figure 1.4: The Illustration of appearance and pose-guided human image generation. The generator synthesizes images with texture details of the reference image and pose style of the target image. Samples are from Ren *et al.* [144]

1.1.3 Fashion Pose-Guided Image Transferring Modeling

Fashion Pose Transfer has attracted substantial research attention due to its expansive potential applications, encompassing domains such as virtual fitting and video animation.

The primary objective of full-body pose-based generation is to synthesize realistic images or videos based on reference images and specific target requirements, which is show in Figure 1.4. Leveraging the powerful image generation capabilities of VAEs [79] and GANs [40], human image generation has achieved remarkable results. For instance, InsetGAN [31] and StyleGAN-Human [32] produce human images with unconditional appearance and texture details. Recently, diffusion models [49, 165] have emerged as state-of-the-art in synthesis results both qualitatively and quantitatively. Images synthesized using Stable Diffusion [148], Imagen [153], and DALLE·2 [140] achieve high-quality performance with intricate details, gaining significant attention in both research and commercial fields. However, the appearance and pose-guided human generation task is more challenging than unsupervised or naive text-driven image generation. Specifically, in a conditional human generation, the main focuses are: 1) achieving realistic image generation and 2) ensuring semantic correctness when applying for pose transfer. Moreover, human video animation poses further challenges, requiring maintaining semantic consistency, spatial

coherence, and temporal coherence. This novel deep learning-based function holds great potential for practical applications. Appearance and pose-guided human image and video generation find suitability in various scenarios, such as video creation, virtual try-on, and fitting with different poses. Several approaches, namely Top-down, Bottom-up, Hybrid, and Diffusion-based approaches focus on this challenging task. Generally speaking, Top-down methods proposed in recent years primarily involve directly transferring the target pose to the reference image. However, this direct transfer using a naive encoder-decoder network presents challenges in preserving identities, as delicate image details such as faces and clothing might not be accurately preserved. Therefore, two possible solutions are adding additional loss functions to constrain the final generation and designing efficient feature extractors. Bottom-up methods, proposed in the past two years, have effectively applied spatial transferring during generation. However, methods that utilize feature extraction and guidance incur higher computational costs due to the presence of extra encoders and spatial transform models within the network. The Hybrid model combines the concepts of both Top-down and Bottom-up to achieve accurate appearance transfer while preserving details, but optimizing the parameters of the entire network remains challenging [144]. Meanwhile, Diffusion-based models [17] have shown the ability to generate real-class images, but the computational cost is extremely high as it requires multiple steps for generating each sample during inference. In general, obtaining high-quality generated images and a compact transfer network size proves challenging, as it involves a trade-off between image quality and computational efficiency.

However, these methods achieve good performance in pose images with pure background, they are limited to pose transfer in sketch-based datasets and illustration-based datasets since images in these two datasets are more abstract and uninformative. Thus the details and appearance are easily lost when transferring.

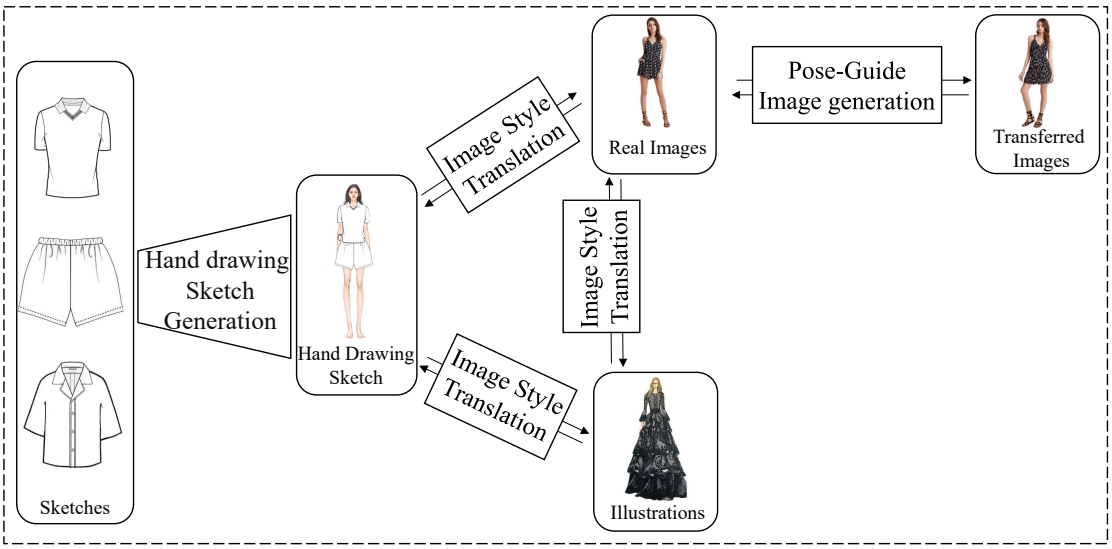


Figure 1.5: The processing framework of the intelligent fashion image generation engine.

1.2 Research Objectives

Despite the notable achievements AI has made in the realm of generative tasks, the application of AI for customized generation within the fashion domain has remained somewhat overlooked. This discrepancy can be attributed to the existing gap in communication and collaboration between experts in computer science and those in the field of fashion. The intricate nature of fashion, with its fusion of aesthetics, style, and functionality, demands a nuanced understanding that may not always be readily comprehensible to those without a background in the domain. This gap in communication results in a mismatch between the capabilities of AI systems and the actual requirements of fashion-related generative tasks. While computer scientists and AI researchers may excel at developing sophisticated algorithms, they might not fully grasp the intricacies of fashion design, leading to a lack of contextual accuracy and authenticity in the generated outputs. Therefore, the research focus centers on cultivating a deeper understanding of the intricate symbiosis between AI technology and the fashion domain, emphasizing the need for interdisciplinary collaboration to unlock the full potential of AI in fashion-related generative tasks. The aims of the task could be summarized in above the analysis:

1) Develop an automated hand-drawing sketch generation engine suitable for

the design scenario, adapting to the proportion of clothing and models. This engine not only frees designers' hands but also accelerates their efficiency.

2) Build a model that can transform real images and hand-drawing sketches (whether it's from runway shows or posed poses) into illustrations that reflect the designer's style, aiming to enhance designers' efficiency and inspire them with more creativity.

3) Obtain a powerful generation model that generates real images or illustrations with different viewpoints while preserving the human body and clothing attributes and generating images with new poses based on specific configurations. With this model, designers can evaluate the effect of current clothing from different angles, providing them with more information and improving their workflow efficiency.

The structure of the intelligent fashion image generation system is depicted in Figure 1.5. The overall framework of fashion image generation system is illustrated in 3. For hand-drawing sketch image generation 4, two main research objectives are: implementing a landmark keypoint detection model and developing a fashion segmentation network. The empirical practice of uninformative image style transferring is defined in the second part of 4. The methods used for style transfer (illustrated in Chapter 4 and 5) and pose-guided human image transfer (chapter 6) will be covered in the subsequent Chapters.

In conclusion, the objective of this task is to build an engine that aids designers in improving efficiency while igniting their creativity.

1.3 Research Methodologies

1.3.1 Modeling Fashion Hand-drawing Sketch Generation Engine

The primary objective of developing a fashion hand-drawing sketch engine is to extract features from clothing items and then map them into body models. By uti-

lizing deep learning models, different types of clothing can be quickly and accurately mapped onto the body models. In details, this study mainly focuses on three main points: 1) Creating a dataset that contains keypoint-labeled sketches and using deep learning networks for training, enabling the model to accurately identify locations of keypoints from various types of clothings. 2) Developing an algorithm for mapping that places the selected sketch clothing onto body models based on the predicted keypoints from the detection model and the pre-set keypoints on the body models defined by designers. 3) Implementing deep learning network to segment clothing items into front and back views while considering the type of clothing, allowing the algorithm to achieve a layering effect.

1.3.2 Modeling Fashion Image Style Translation

The ability to interchange real images and illustrative images can help designers and industry professionals gain a more intuitive understanding of their designs and make subsequent adjustments. Unlike image-to-image translation in computer vision, images with less information are more commonly used in fashion. Due to the unique character of fashion images, datasets with high image quality and specialized captions have been developed to facilitate the task of fashion image style translation. Two deep learning models are employed to realize this task: a GAN-based generator with strong generative capabilities to control output, and guiding image generation with a more precise loss function is proposed to solve this task. The second model involves fine-tuning the best pre-trained diffusion-based model to achieve identity-preserving style transfer.

(1). Task Introduction, Dataset Construction. Previous methods in computer vision primarily focus on image-to-image translation and image style transfer, achieving consistent results. They perform well with informative images, which are characterized by (1) rich colors, (2) diverse spatial distribution, and (3) multi-level depth. However, transferring illustrative images poses challenges due to their characteristics: (1) relatively simple colors, (2) line drawings, and (3) a reduced sense of spatial depth. Existing methods struggle with style transfer involving pure backgrounds. To address this issue, an uninformative dataset composed of 10,000

high-resolution sketches has been created to support this task and facilitate further exploration. Furthermore, baesd on fashion paired dataset stylishU [225], SwinIR is initially utilized [94] in conjunction with LDSR [148] to perform a super-resolution version StylishU-SR, thereby obtaining images with a resolution of 512×512 for further research.

(2). Hierarchical Generative Network. To address the problem of uninformative image style transfer, a new method with a streamlined and effective approach is proposed for transferring informative and uninformative images. This method is based on an Encoder-Decoder structure and is inspired by StyleGAN [74, 73], incorporating a style coder C for style mixing to generate images and achieve style transfer. Besides, style transform block inside the network are applied to trasnfer the high-level features and support the generation of the target-style image. Low- and high-level features are both considered when the discriminator distinguishes whether the input image is real or fake. Adversarial loss and style loss are considered to constrain the generative network.

(3). Improving Quality by Fine-tuning Method. With the rise of diffusion models, their powerful generative capabilities and scalability provide favorable conditions for addressing this task. For the proposed Methods, To effectively capture both texture and spatial information, hidden details are extracted from the source image using image and sketch extraction modules. Specifically, a novel multi-layer module called Uni-Adapter is employed to gather spatial and texture information separately. Additionally, two distinct style adaptation modules, named Dual-LoRA, are integrated into the UNet denoiser to capture styles from different domains. The inclusion of parameters in both the image feature extractor and the fixed sketch feature extractor enhances the model's ability to extract spatial and textural information from the input. Specialized style adaptation modules with learnable parameters are incorporated into the UNet denoiser to refine the style of the synthesized images, facilitating content and style disentanglement.

1.3.3 Modeling Pose-Guided Human generation

The primary goal of pose-guided human generation is to help designers clearly visualize the appearance of a human figure and their garments from different perspectives, enabling them to make more precise adjustments to the detailed aspects of the garments. This work focuses on two key components: 1): *Pose-Guided Attention Estimator*: It extracts the high-level features of the clothing and the target person in detail. 2): *Image Synthesis with Confidence Map*: It balances prior knowledge and predicted information to generate human images in the target pose with rich details. These efforts make the generated images with different viewpoints more realistic, providing designers and professionals with valuable feedback and improving their efficiency.

(1). Pose-guided Attention Estimator. To reassemble the referred image according to the provided modifications, the correspondence between the referred image $\mathbf{I}_r$ and target pose $\mathbf{I}_{st}$ mapped in the same domain S is estimated. This correspondence is built using an attention estimator, which consists of three parts: (a) Attention Correspondence Module, (b) Affine Transformation Module, and (c) Feature Fusion Module. Final warp images is generated from the feature fusion module and applied for further process with the attention correspondence matrix obtained by attention correspondence module. Attention correspondence matrix is obained by correspondence between features extracted from referred images and target pose. Affine transformation is a geometric transformation including translation, rotation, scaling and sharing. With this transformation, the pixel-level information is preserved and not lost. The attention warp exemplar and affine warp guidance have complementary advantages: the former preserves the global information of the body figure, while the latter maintains the details of the human body. Therefore, a fusion model is proposed to obtain pose-guided warp exemplar based on the attention warp exemplar and affine warp.

(2). Image Synthesis with Confidence Map. The image synthesis approach uses a generator that combines the target skeletons with the attention warp exemplar. Firstly, the affine warp guidance and the warped image are weighing using the fusion map. Then, the SPADE resblock is utilized with the attention-based warp

image to render the target-pose skeleton. The final image is synthesized by the network using target skeletons, Confidence map, and correlation matrix generated from the Pose-guided Attention Estimator. The pose-guided attention warp examplar is used as input to the final generator, which is based on the progressive image generator [71, 133]. The generator synthesizes the final image by incorporating both the reliability of the warp examplar from the confidence map and the guidance of the target skeleton.

1.4 Research Significance

With the development of technology, the design process for designers is becoming increasingly digitized. However, due to the complexity of design and the tediousness of the workflow, establishing an efficient engine that can assist designers has become a meaningful endeavor.

(1). Deployment of Multi-layer Hand-Drawing Sketch Generation Engine. Compared to mainstream deep generative models*Generative Adversarial Network* [40], the Hand-Drawing Sketch Generation Engine offers significant advantages. This method not only ensures high reliability but also preserves the original textures of the clothing in the generated images. The system enables designers and professionals to intuitively visualize how the clothes would appear by providing a realistic representation of garments. Moreover, this method introduces a new data source for clothing representation from in terms of data representation.

(2). Improvement of Performance for fashion Image Style Transfer. this study firstly propose the a the differences between informative and uninformative images and introduce a novel task that *uninformative fashion image style transfer*. A proposed neat and effective MiniGAN, that achieves good result in unsupervised style transfer while preserving the color information of the source image to deal with this task. A StyleTransform Block which contains 9 independent residual blocks to carry out target-style image transferring. Besides, Style Coder is appliedto support the pixel-wise image generation. Moreover, a uninformative dataset which is composed of 10000 sketches is proposed to support this task and further exploration.

Due to the limitations of existing methods in uninformative fashion image style transfer, a novel approach named *Uni-DlLoRA* is proposed. This approach focuses on fine-tuning diffusion models for fashion image synthesis while improving style disentanglement. The method addresses current challenges by incorporating image-conditioned information through the proposed Uni-adapter and adapting the UNet denoiser with the Dual-LoRA module. This enables better capture of spatial and textural details from both real and illustrative domains.

(3). Introduction of Multimodal Fashion Dataset StylishU-SR. SwinIR [94] and LDSR [148] are utilized to enhance the resolution and clarity of images in the StylishU dataset by performing super-resolution. The caption of each image is extracted by BLIP [89] and refined by fashion experts for text-conditioning. By sharing this dataset, researchers can more easily conduct follow-up studies to generate higher-quality fashion image style transfer tasks.

(4). Improvement of Pose-Guided Human Transfer and Comprehensive Insights. This study contributes by proposing a novel pose-guided human transfer method, which utilizes a *Pose-Guided Attention Estimator* and *Image Synthesis with a Confidence Map*. These techniques ensure that the generated images retain the original appearance while matching the target pose. The effective results can aid designers in improving efficiency while sparking their creativity. Moreover, a Comprehensive research about appearance and pose-guided human transfer which contains introduction, problem definition, preliminaries, Methodologies, applicaitons, challenges. This research provides a comprehensive understanding of each of these properties and how they contribute to the generation process. Furthermore, this research explores the diverse applications of appearance and pose-guided human generation, such as video creation, virtual try-on, and fitting with different poses. The potential practical uses of this technology in various fields are discussed in detail. However, despite the remarkable progress in appearance and pose-guided human generation, there are still challenges to address. High-quality image synthesis with minimal artefacts, semantic-level editing, and accurate feature disentanglement remain difficult. These challenges will be disscused and potential future research directions will be outlined in this domain.

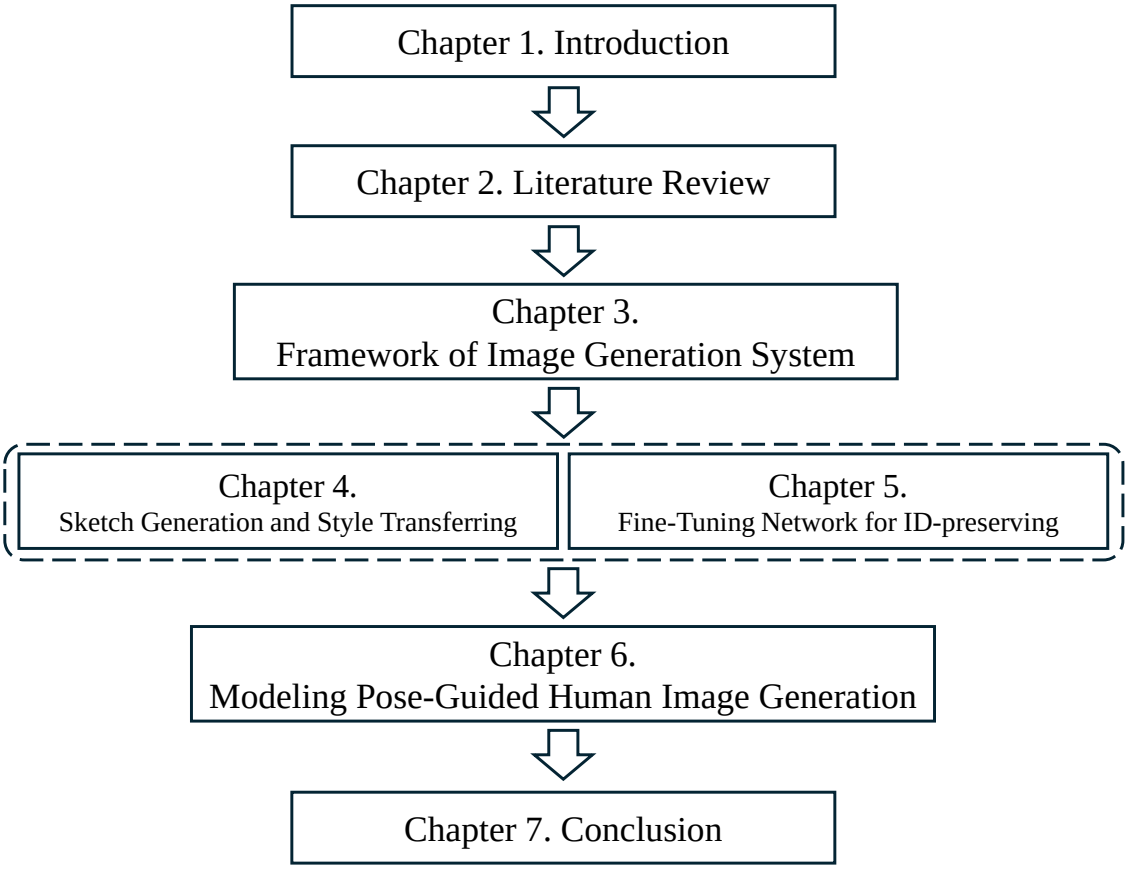


Figure 1.6: The overview organization of this thesis.

1.5 Organization of the Thesis

The outline of this thesis is illustrated in the Figure 1.6.

Chapter 1 presents the research background, objectives, and significance of this thesis, offering an introduction to the overall topic and establishing the context for the research.

Chapter 2 provides a detailed literature review of intelligent fashion image generation systems. It covers the state-of-the-art methods for each specific task within the field of computer vision. Additionally, this chapter discusses the available datasets and popular evaluation metrics. A comprehensive overview of related research is presented to inspire further investigation.

Chapters 3, 4, 5, and 6 detail the developed models for each task within the intelligent fashion image generation system. Specifically, Chapter 3 introduces an automatic fashion hand-drawing sketch generation engine capable of producing multi-

layer tops and bottoms on a body model. Chapter 4 proposes the challenges of informative and uninformative image style transferring, and presents a GAN-based image style transfer method to enhance performance in this area. Chapter 5 describes a novel approach to fashion image style transfer, introducing new methods and a multi-modal dataset for future research. Chapter 6 proposes an innovative approach for pose-guided human generation. Moreover, a comprehensive research providing critical insights is disscused in this chapter.

Chapter 7 concludes the thesis by summarizing the key findings and limitations. It also offers an overview of potential future work.

Chapter 2

Literature Review

The primary objective of this research endeavor was to develop of a fashion hand-drawing sketch generator. Two core research objectives remained are: the establishment of a fashion pose transfer model, and the implementation of a fashion image style transfer model. In recent years, numerous researchers focusing on the field of computer vision and methods, datasets, and evaluation metrics have been developed to tackle these three tasks. This chapter provides a comprehensive review of previous research efforts related to each specific task. Specifically, to create a machine capable of automatically generating hand-drawn sketches, understanding essential low-level information such as key points and parsing images is crucial, as described in Chapter 2.1 which provides an overview of low-level image analysis. Regarding landmark detection in real images and fashion images segmentation, Chapter 2.2 reviews relevant work about fashion image style trasnfer. Additionally, the engine possesses the ability to transform real images, illustrations, or hand-drawn sketches based on provided poses, making it particularly important to grasp pose-guided image generation. And the related work of this ablitiy will be illustrated in Chapter 2.3.

2.1 Fashion Low-level Clothing Features Analysis

Low-level clothing feature analysis is a research area that focuses on extracting and analyzing various low-level visual features from clothing images. These

features include color, texture, shape, and pattern, which are crucial for various applications such as fashion recognition, clothing recommendation, and virtual try-on systems. For Color-based Feature Analysis, Color is one of the most important low-level features used in clothing analysis. Various studies have focused on color-based feature extraction and analysis. For instance, Deole *et al.* [22] proposed a color histogram-based approach to extract color features from images, which were then used for classification. They achieved promising results in terms of accuracy and efficiency. Similarly, Di *et al.* [23] proposed a color-based clothing retrieval system that utilized color coherence vectors to represent clothing images and achieved significant improvements in retrieval performance. Moreover, some researchers are focusing on Texture-based Feature Analysis. Texture analysis plays a crucial role in low-level clothing feature analysis. Researchers have developed various methods to extract and analyze texture features from clothing images. For example, Rivaldy [147] proposed a texture descriptor called Local Ternary Patterns (LTP) for clothing image classification. Their experiments demonstrated that LTP-based features outperformed other texture descriptors in terms of accuracy and robustness. Another study by Zhang *et al.*[208] utilized a combination of texture features, including Local Binary Patterns (LBP) and Gabor filters, to improve clothing recognition performance. Shape analysis is another important aspect of low-level clothing feature analysis. Researchers have explored different techniques to extract and analyze shape features from clothing images. Kita *et al.* [80] introduced a shape-based clothing recognition system that utilized shape context features and achieved high accuracy in classifying different clothing categories. Patterns are distinctive visual features found in clothing items and are crucial for clothing analysis. Researchers have developed various methods to extract and analyze pattern features from clothing images. For example, Surakarin *et al.* [170] proposed a pattern-based clothing retrieval system that utilized Local Binary Patterns (LBP) to capture the pattern information of clothing items. Their experiments demonstrated the effectiveness of pattern features in clothing retrieval tasks. Similarly, Pan *et al.* [128] proposed a pattern-based clothing recognition system that employed a combination of texture and pattern features, achieving significant improvements in recognition accuracy.

2.1.1 Fashion Detection and Fashion Landmark Detection

Fashion detection serves the purpose of precisely localizing fashion items within a provided image through regression techniques. This technological advancement finds diverse applications within the realm of fashion, contributing to a range of innovative and practical functionalities.

At its core, the primary objective of fashion detection is to undertake object localization within images, often realized through the delineation of bounding boxes. This enables the accurate identification of the spatial coordinates of fashion items, facilitating subsequent analyses and manipulations. Furthermore, fashion detection extends its utility to encompass the intricate task of identifying clothing landmarks. By identifying and locating specific points of interest in garments, this facet of fashion detection enhances the understanding of clothing's spatial relationships and forms the foundation for various downstream applications.

Fashion detection utilizing bounding boxes constitutes a pivotal approach for discerning distinct fashion articles within an individual's image. This technique effectively captures various elements of attire, including dresses, coats, pants, shoes, bags, and more. Several notable methodologies have been established to achieve this endeavor, such as Regions with Convolutional Neural Networks (R-CNN) [38], fast R-CNN [37], faster R-CNN [143], Single Shot multi-box Detector (SSD) [109], Region-based Fully Convolutional Networks (R-FCN) [19], and "you only look once" (YOLO) [142]. These prominent frameworks can be harnessed to execute the task of clothing detection through bounding boxes.

The underpinning of clothing detection is frequently adapted from a generalized object detection framework [83, 162]. For instance, Hara et al. [46] integrated the R-CNN object detector with a spectrum of domain-specific priors, encompassing factors like pose, object morphology, and dimensions. This fusion of priors was combined with an appearance-derived posterior, computed using Support Vector Machines (SVM), culminating in a comprehensive posterior distribution. This final distribution was then employed to predict the class associated with each bounding box, thereby culminating in an effective clothing detection solution.

The goal of clothing detection based on landmarks is to precisely estimate the coordinates of landmark locations. Clothing landmarks play a vital role in creating more robust feature representations, particularly when dealing with clothing deformation or occlusion. This reinforcement contributes to achieving greater accuracy in predictions for various clothing-related tasks. The domain of clothing landmark detection bears a resemblance to person pose estimation.

In current research, the pursuit of clothing landmark detection generally adheres to a foundational structure akin to single-person pose estimation. This structure can be categorized into two primary methods, determined by how landmarks are predicted: coordinate-based methods [112, 113, 38] and heatmap-based methods [53, 191, 208, 93, 181, 196]. Coordinate-based approaches utilize learned feature maps to directly regress landmark coordinates. In contrast, heatmap-based methods first generate heatmaps and subsequently predict landmark coordinates based on these generated heatmaps. Within these heatmaps, pixel values denote the probabilities of landmark existence.

Coordinate-based landmark detection aims at predicting the positions of M functional key points of a fashion item. Among the coordinate-based landmark detection methods, Liu et al. introduced the Deep Fashion Alignment (DFA) framework [113]. This framework adopts a multi-stage structure that involves the sequential integration of three CNN-based regression models. In the initial stage of DFA, a raw image is utilized as input to predict preliminary landmark positions along with their respective labels. Subsequent stages build upon the outputs generated in earlier stages. It's worth noting, however, that DFA operates under the assumption that clothing bounding boxes are provided as prior information during the training phase. Regrettably, this assumption often proves impractical in real-world scenarios. To address this inherent limitation, Yan et al. devised the Deep Landmark Network (DLAN) [55]. DLAN incorporates a selective dilated convolution mechanism and a hierarchical recurrent spatial transformer. These components synergistically tackle issues related to scale discrepancies and cluttered backgrounds. Unlike DFA, DLAN circumvents the need for predefined clothing bounding boxes, making it a more adaptable and versatile solution for clothing landmark detection.

Heatmap-based landmark detection methods learn to predict a heatmap, which constitutes a positional distribution for each landmark in an image. In the realm of landmark detection, Wang et al. introduced a Bidirectional Convolutional Recurrent Neural Network (BCRNN) [53]. This innovative framework facilitates the exchange of messages across fashion grammar, a concept used to convey kinematic and symmetric relationships between clothing landmarks. In practical application, visual attention mechanisms play a pivotal role in focusing feature learning on crucial landmark positions. This focused attention enhances the capacity for discriminating representations based on location. Furthermore, the task of fashion landmark detection demands both location information and an extensive receptive field. This combined information is essential for determining whether a particular location qualifies as a landmark and for ascertaining the specific class to which the landmark belongs. In a related context, Li et al. devised a Spatial-Aware Non-Local (SANL) block [181]. This block augments the original non-local block [183] by integrating spatial information as an attention mechanism. This enhancement bolsters the model's capability to understand spatial relationships within the data. On a similar note, Huang et al. proposed a comprehensive end-to-end architecture [93] that leverages Part Affinity Fields (PAFs). This approach capitalizes on the associations between landmark locations to elevate the precision of landmark coordinate predictions. By integrating these advancements, researchers continue to advance the state-of-the-art in fashion landmark detection, pushing the boundaries of accuracy and performance.

2.1.2 Fashion Clothing Segmentation

Fashion Clothing Segmentation serves the purpose of segmenting fashion elements by allocating category labels to individual pixels on a person within an image. This intricate task holds significance across various real-world applications, including personalized recommendations and virtual try-ons. In scenarios where fashion items like shoes or bags constitute only a small fraction of an image, while the bulk of the image remains unrelated to a shopper's intent, the need to precisely localize or segment these small fashion items arises. This localization process proves essential

to efficiently retrieve highly pertinent items from a visual search system, rather than forwarding the entire image. A distinction emerges between clothing parsing and human parsing, hinging on the nature of labels. Clothing parsing involves granular clothing categories as labels, whereas human parsing employs labels about body parts and general clothing categories. Clothing parsing necessitates a higher-order understanding rooted in the semantics of intricate clothing varieties. Furthermore, it must grapple with the challenge of accounting for the distorted configuration of clothing on individuals within an image, thereby introducing an additional layer of complexity. Conventional solutions build upon traditional methods by incorporating predefined rules for label inference [95, 209, 190, 63, 163, 28]. In contrast, deep learning-based approaches primarily center on the Fully Convolutional Network (FCN) image segmentation pipeline, augmented with various auxiliary modules. These modules are designed to encapsulate diverse fashion-related aspects, such as edges, outfit consistency, and texture [114, 96, 64, 76, 59, 205]. This integration of inherent fashion item knowledge augments the model's capacity to accurately recognize and categorize fine fashion classes during the segmentation process.

Deep learning methods excel in extracting contextual information from the human body through receptive fields within deep architectures. This obviates the need for explicitly treating the human body as an additional preprocessing step for a given human image. The effectiveness of this approach contrasts with earlier methods that required separate human body identification as a preliminary procedure. A notable breakthrough in this field has been achieved with Fully Convolutional Networks (FCNs), significantly elevating performance in semantic image segmentation [114]. Two prominent research trajectories emerge for conducting semantic clothing image segmentation. The first involves refining FCN architectures through the integration of supplementary discriminative classifiers [124, 96, 64, 76]. The second branch entails the incorporation of Conditional Random Fields (CRFs) [174, 130] into parsing neural networks, thereby facilitating end-to-end trainable models. However, there is no task focusing on the back-front clothing segmentation.

2.2 Fashion Image Style Translation Modeling

The fundamental objective of image-to-image translation revolves around acquiring the ability to discern and implement a conversion mechanism from a given source domain to a designated target domain while preserving the essential characteristics of the images.

2.2.1 State-of-the-art Methods

Recent strides in this field have led to promising breakthroughs. A notable pioneer in this domain is Pix2pix [62], which marks a pivotal advancement by leveraging Generative Adversarial Networks (GANs) to effectuate the transformation of images between disparate domains. However, it's noteworthy that Pix2pix [62] necessitates a meticulously matched dataset, known as paired data, to yield images of commendable quality. This prerequisite for paired data poses a challenge, prompting the exploration of innovative strategies.

In response to this challenge, novel architectures have emerged. Two noteworthy instances include the concept of cycle consistency [218] and the notion of a shared latent space [107]. These architectural paradigms strive to overcome the constraint of paired datasets. More recently, cutting-edge algorithms [132, 57, 84] have been conceived, building upon the principles of these architectures to elevate the quality of translated images. These advancements have contributed to refining the translation process, culminating in enhanced image quality.

While the current landscape of image-to-image translation has demonstrated its prowess in generating images of commendable quality and embracing a diverse array of possibilities, it's imperative to recognize that the scope of application is bounded by the prerequisite of having informative datasets for both source and target domains. This limitation implies that the existing methods hinge on datasets rich in meaningful correspondences.

Surprisingly, despite the significant progress in image-to-image translation, there remains a conspicuous gap in the research landscape. The focus has yet to extend

to the domain of uninformative style transfer, wherein the intention would be to perform transformations without the presence of profound correspondences between the datasets. This uncharted territory poses intriguing possibilities and challenges that warrant exploration to pave the way for even more comprehensive and versatile image translation methodologies.

2.2.2 Benchmark Datasets

The table provides an overview of benchmark datasets applied to the task of image style transfer. Since this task falls under the domain of computer vision, specifically the task of image-to-image translation, some datasets from image-to-image translation can also be applied to image style transfer. For example, datasets proposed in pix2pix [62], such as monet2photo, summer2winter, ukiyoe2photo, and vangogh2photo, clearly and accurately define images of different styles. While these datasets feature diverse styles, their resolutions are no longer mainstream, and the data is unpaired. StylishGAN [225] introduced a fashion dataset called StylishU, which contains paired images including real photos and their corresponding fashion illustration; however, it remains relatively rough and requires further post-processing to achieve higher-quality paired data. Additionally, some researchers have created new datasets by combining images from WikiArt [1] and Places365 [212]. However, these high-quality datasets are also unpaired. Despite these limitations, these benchmark datasets serve as valuable prior resources for evaluating and advancing techniques in fashion image style transfer.

2.3 Fashion Pose-Guided Image Transfer Modeling

This chapter illustrate the pose-guided human image transferring based on the input human images and target pose.

[1] `https://www.wikiart.org/`

2.3.1 State-of-the-art Methods

Pose-guided human image synthesis has gained significant attention in recent years. The task focuses on generating a target image conditioned on a given skeleton pose and a reference image, while preserving the appearance and fine details of the reference [186, 72, 62, 218, 57, 213]. While earlier works, such as the two-stage network proposed by [116], demonstrated improved performance, these methods come with high computational costs.

To address the efficiency challenges, [30] introduced a hybrid approach that combines Variational Auto-Encoders (VAEs) [25] with U-Net [149]. However, this method relies on 1-D embedding features, which are insufficient for capturing detailed appearance information, leading to a drop in the quality of the generated images. Additionally, the skip connections in U-Net often cause feature misalignment, negatively impacting synthesis performance. To mitigate these issues, [223] proposed a network with transfer blocks that establish connections between regions of interest in the reference and target poses, resulting in better alignment and quality.

Other approaches, such as CoCosNet [201], CoCosNetV2 [216], and NETD [144], introduced novel frameworks that generate images by leveraging semantic warp images. These methods rely on mapping semantically corresponding patches from the reference image to the target pose. While effective in improving semantic consistency, these approaches struggle to transfer fine patterns from the reference image. Moreover, their reliance on additional parsing images and attribute annotations limits their applicability in practical scenarios.

Flow-based methods, including those proposed by [215] and [44], offer greater flexibility compared to affine transformation techniques, as they are not constrained by predefined transformation components. However, these methods perform warping at the pixel level rather than the feature level, which limits their ability to generate rich and coherent content. To address this, [145] proposed incorporating local attention to enhance feature representation during synthesis. Similarly, [173] introduced a method using local flow fields to capture semantic correlations between features. Despite these advancements, these approaches face challenges when there

is a significant discrepancy between the source and target poses, as the local flow fields often produce blurred features, leading to a decline in image quality.

Dataset	Type	Scale	Resolution
Market1501 [210]	Images	32,668 images, from 1501 people	128×64
DeepFashion [112]	Images	52,712 images with over 200,000 pairs	256×256
DeepFashionHD [112]	Images	52,712 images with over 200,000 paris	1101×750
MPV [26]	Images	62,780 three-tuples	256×192
MVC [106]	Images	161,638 images in 37,499 items	1920×2240
Human3.6M [61]	Images	3,600,000 images from 11 people	1000×1000
LookBook [194]	Images	75,016 images with 9732 items	256×256
FashionOn [51]	Images	10,895 paired images	288×192
FashionTryOn [211]	Images	28,714 triplets images	256×192
Deepfashion2 [36]	Images	491,895 images with clothes pairs	$*$
Deepfashion-MM [66]	Images	44,096 images mannually annotated	1101×750
SHHQ [32]	Images	230,000 images	1024×512
Penn action [206]	videos	2,326 videos in 15 actions	640×480
Tai Chi [175]	videos	4,500 video clips	256×256
Fashion Dataset [197]	videos	600 videos with roughly 350 frames each	940×720
iPER [110]	videos	206 video sequences with 241,564 frames	256×256
iPER-HD [110]	videos	206 video sequences with 241,564 frames	1024×1024
VVT [27]	videos	791 videos with totally 190101 frames	256×192

Table 2.1: List of the main datasets for appearance and pose-guided human generation including name, types of datasets and the size of datasets, and the resolution of the datasets. Types of datasets include images and videos.

2.3.2 Benchmark Datasets

Table 2.1 illustrate the overview of benchmark datasets applied for the task of pose-guided human tranferring. List of the main datasets for appearance and pose-guided human generation including name, types of datasets and the size of datasets, and the resolution of the datasets. Types of datasets include images and videos. In human image generation, Market1501 [210] is the first and most common dataset, which contains 32,668 images composed of 1501 identities with six different view-

points, and the resolution of the images is 128×64. To enhance the quality of the dataset and expand its size with a clean background, Deepfashion [112] is proposed, and it stands as the most popular dataset adopted for human pose transfer. It consists of 52,712 in-shop clothes images with a clean background, and the resolution of Deepfashion is 256×176. The new Deepfashion2 dataset is introduced with a higher resolution (1101×750) to enhance the capabilities of the models. MPV [26] is also adopted to evaluate appearance and pose-guided image generation. Deepfashion2 [36] proposed a dataset that contains over 491,895 images with 873,000 Commercial-Consumer clothes pairs. StyleGAN-Human (SHHQ) [32] provide 40k images with the resolution of 1024×512 in which the parsing images are manually annotated. Very recently, DeepFashion-MultiModal [66] proposed high-resolution 1101×750 datasets with extra parsing images, keypoints maps and Densepose and textual descriptions. With these elements, models are the potential to carry out overall human generation by being pose-driven and text-driven.

For appearance and pose-guided video generation, Tai Chi [175] is the most popular dataset applied to video reconstruction. Table 2.1 illustrates the main datasets adopted in most experiments.

hese benchmark datasets serve as valuable resources for evaluating and advancing pose-guided human generation techniques.

2.4 Evaluation Matrix

This chapter primarily discusses the evaluation metrics previously used for the task of image style transfer and pose-guided human image generation.

Since both tasks focus on image generation, the evaluation metrics are quite similar with only minor differences. In terms of assessing the quality of the generated images, IS[154], FID[48], and LPIPS [203] are the most widely used metrics. Additionally, in recent years, scores derived from pretrained networks, such as DS, have been utilized to evaluate the confidence of person detection. High-level distances, such as the Frchet Segmentation Distance (FSD) introduced by Bau [3], and the Sliced Wasserstein Distance (SWD) [136], are applied to evaluate the quality of

the generated images at the feature map level. In details,

Detection Score (DS)[161] measures the confidence of detecting a human in generated images. It uses the highest person-class score from a pretrained SSD[109] model, reflecting how realistic the generated human appears.

Inception Score (IS)[154] evaluates the quality and diversity of images generated by GAN models. It calculates the latent distributions of features extracted by an Inception-v3[172] model pretrained on ImageNet [20]. Higher scores indicate higher diversity and clearer image generation, particularly for diverse and complex scenes.

Sliced Wasserstein Distance (SWD)[136] evaluates the Wasserstein distance[152] between distributions of projected feature maps from real and generated images. It is particularly useful for assessing fine-grained differences in feature representations.

Learned Perceptual Image Patch Similarity (LPIPS)[203] measures perceptual differences between two images using deep neural networks such as VGG[164] and SqueezeNet [58], both pretrained on ImageNet [82]. A lower LPIPS score indicates higher perceptual similarity between the images.

Faithfulness metrics evaluate the similarity between real and generated images. Commonly used metrics include:

Peak Signal-to-Noise Ratio (PSNR) is a widely used metric for assessing generated image quality. It calculates the ratio between the maximum possible pixel value and the mean squared error (MSE) between two images. Higher PSNR values indicate better image quality.

Structural Similarity Index (SSIM) [185] measures the similarity between real and generated images by comparing luminance, contrast, and structural details. Higher SSIM scores indicate greater similarity.

2.5 Chapter Summary

This chapter organizes the literature review for the task of building a Deep Learning-based Intelligent Fashion Image Generation System across three key aspects.

Specifically, it begins by discussing foundational models related to the auto-

matic multi-layer hand-drawing sketch generation engine. Models for fashion landmark detection and fashion clothing image segmentation are also reviewed, as they play an essential role in designing a comprehensive generation engine.

For the task of fashion image style transfer, models originating from computer vision tasks such as image-to-image translation and image style transfer are explored in detail. Methods in image-to-image translation focus on learning the weights of specific styles to generate corresponding outputs, while image style transfer techniques rely on feature fusion to enable few-shot style adaptation, eliminating the need for extensive training. Furthermore, the limitations of both types of models in the context of fashion image style transfer are analyzed from various perspectives.

For the task of appearance and pose-guided human image transfer, several methodologies such as feature-based, flow-based, and attention-based approaches are discussed. Additionally, the limitations and challenges associated with these methods are thoroughly examined.

This chapter also reviews the benchmark datasets relevant to these tasks. These datasets provide valuable resources for training and evaluating algorithms and models in the field of intelligent fashion image generation.

Finally, Chapter 2.4 offers an overview of the most commonly used evaluation metrics in previous studies. These metrics primarily focus on assessing either the recommendation accuracy or the retrieval performance within the context of fashion recommendation tasks.

Chapter 3

The Framework of Intelligent Fashion Image Generation System

3.1 Introduction

The intelligent fashion image generation system framework can be divided into three main components:

1): Automated generation of multi-layer sketch images from hand drawings (corresponding to Part A in the figure). In details, the system employs two specialized models: a keypoint detection model and a segmentation model. These models work in conjunction to process the input sketches.

2): The hand-drawing sketch images generated from the system are processed through an image style transfer model to create corresponding realistic and illustrative images (shown in Part B). For the task of sketch images to illustrative images translation, a novel method is proposed to tackle the problem of both informative and uninformative image transferring.

3): The final component of the framework involves using generative models to transform the pose of real images into specified poses, allowing for better garment presentation (Part C).

These three components form a comprehensive Intelligent Fashion Image Gen-

eration System, which not only assists users and designers in rapidly creating designs but also provides them with enhanced inspiration and creativity. And Figure 3.1 illustrate the overall framework.

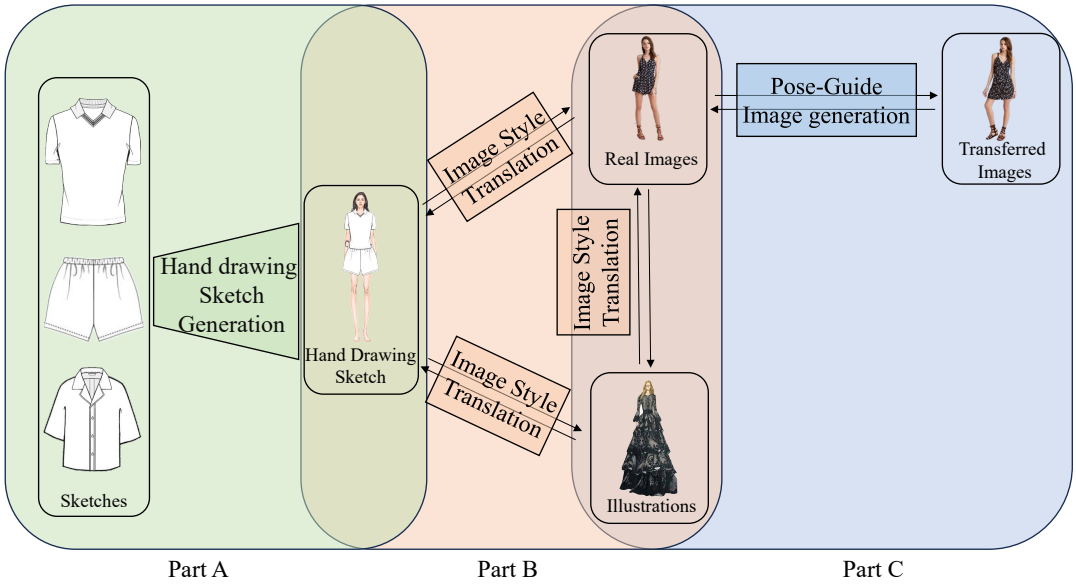


Figure 3.1: The figure illustrate the overall framework of intelligent fashion image generation system. In details, Part A illustrate the process of hand-drawing sketch image generation. Part B demonstrates the interaction between hand-drawing sketch images, illustrative images and real images. Part C elaborates the pose guided image generation based on proposed poses in the real domain.

In this chapter, the main components that constitute the framework of the Intelligent Fashion Image Generation System will be introduced. In terms of hand-drawn image generation, the keypoint detection model within this task extracts clothing landmarks, while the segmentation model generates segmentation maps of the sketches. To determine the precise placement of garments, the system calculates their scale and position by aligning the garment-specific landmarks with corresponding human body keypoints. The final output image is systematically generated through a layered synthesis process, where individual garment items are composited onto the baseboard following a predetermined order of placement. Specifically, the Chapter4.2 provides detailed insights into the implementation methods.

Once the hand-drawn sketches are automatically obtained, realistic images and illustrative images are generated to provide users or designers with more inspiration

during the image style translation process. For the task of translating sketch images into illustrative images, a novel method is proposed to address the challenges of both informative and uninformative image transfer. Specifically, a StyleTransform Block, consisting of 9 independent residual blocks, is introduced to facilitate target-style image translation. Additionally, a Style Coder, inspired by StyleGANv2 [74], is employed to support pixel-wise image generation by capturing multi-level features of the target image for final synthesis. For the task of translating real images into illustrative images, a novel model named Uni-DlLoRA is introduced to overcome existing limitations. This model integrates the original images with a pretrained diffusion-based model using the proposed Uni-adapter extractors, while leveraging the Dual-LoRA module to provide distinct style guidance. The detailed composition and implementation of these two innovative models are thoroughly discussed in Chapter 4.5 and Chapter 5.4, respectively.

For the task of pose transformation based on human figures, a novel method is proposed to accomplish this objective. Specifically, attention-based spatial transformation modules and affine transformation modules are utilized to generate accurate appearances and extract pixel-wise details in local regions, producing intermediate results. Additionally, a confidence map is introduced to refine spatial information during the final image synthesis process. The detailed composition and training methodology of this approach are comprehensively described in the Chapter**??**.

3.2 Hand-Drawing Sketch Generation Engine

For the task of hand-drawn sketch image generation, accurately matching the sketch to the model body is the most critical factor. The relevant knowledge and components of this process will be introduced in the following section.

3.2.1 Methodology

The engine of fashion hand-drawing sketch generation is realized through three steps, namely fashion landmark detection, fashion segmentation, and fashion item

composition. The detailed process will be discussed in this section.

3.2.1.1 Fashion Landmark Keypoint Detection

For the purpose of identifying the key landmarks of fashion items, the knowledge-guided fashion network, as introduced by Hrnet[168], offers a suitable backbone. Instead of directly predicting the positions of landmarks, a confidence map depicting positional probabilities for each landmark is employed to facilitate the generation of these fashion-specific landmarks. A novel head named ViPNAS [188] is introduced to access this confidence map.

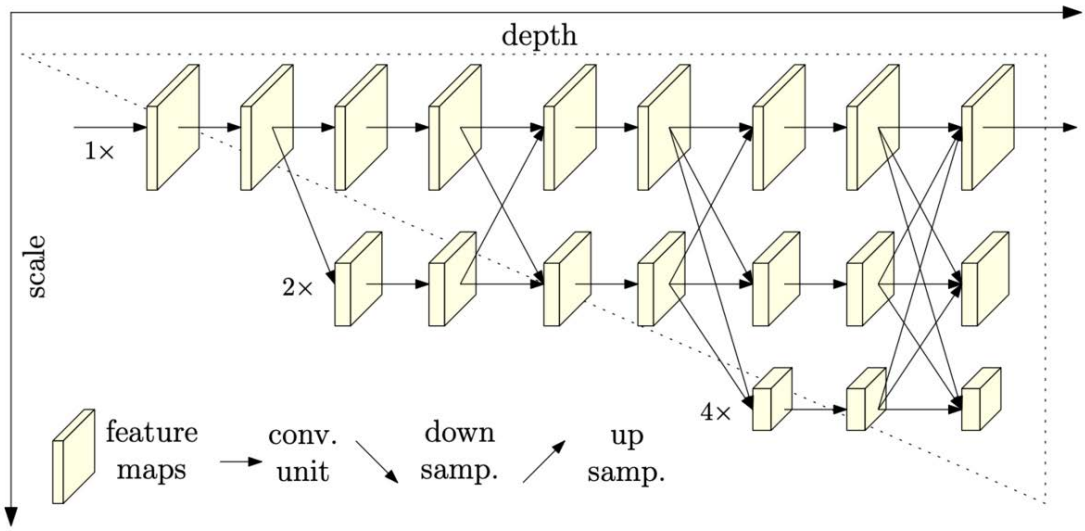


Figure 3.2: The structure of the Hrnet [168].

Figure 3.2 illustrates the structure of the Hrnet [168] that the network starts from a high-resolution convolution stream as the first stage, gradually adds high-to-low resolution streams one by one, forming new stages, and connects the multiresolution streams in parallel. As a result, the resolutions for the parallel streams of a later stage consist of the resolutions from the previous stage, and an extra lower one.

To achieve a better trade-off between accuracy and efficiency, the novel neural architecture search (NAS) method, termed ViPNAS [188], is utilized to search networks in both spatial and temporal levels for fast landmark and keypoint estimation.

Figure 3.3: Examples of existing fashion segmentation datasets.

3.2.1.2 Fashion Image Segmentation

The importance of fashion segmentation for hand-drawing sketch generation is that it makes the generated images more realistic. Especially in the case of item overlapping, taking an outwear and a shirt for instance, the front part of the outwear only covers part of the skirt instead of rough attaching the shirt directly to the outwear. To obtain an available fashion segmentation model, the creation of a fashion segmentation dataset is critical for this task. The current fashion dataset released focuses on fashion category segmentation, as shown in Figure 3.3. However, this task requires classifying the front part, and back part of an item image. In this case, a new fashion segmentation dataset is needed to classify.

3.2.1.3 Automatic Hand-Drawing Sketch Generation

After the clothes are segmented, conventional computer vision processing techniques are used to realize the item composition process. First, the scale and position of the garments are calculated by matching the landmarks of the garment with the pose key points of the human body. The output image is generated by synthesizing each item into the baseboard in the given order.

3.2.2 Role in the Framework

The Hand-Drawing Sketch Generation Engine is the first component of the entire framework, enabling automatic and rapid sketch generation (Part A in the Figure 3.1). It serves as an indispensable part of the overall framework.

3.3 Sketch to Illustrative Fashion Image Generation

Sketch to Illustrative Fashion Image Generation is a specialized sub-task within the realm of image-to-image translation, aimed at infusing a specific fashion style while preserving the structural integrity of the original image. To tackle both informative and uninformative fashion image style transfer challenges, a novel model named MiniGAN has been proposed. The design principles, related work, and methodology of this model will be elaborated in the subsequent sections.

3.3.1 Methodology

The primary objective is to address the transfer of both informative and uninformative images. For GAN-based algorithms, the task involves generating vivid images with rich details, such as realistic textures and easily recognizable shapes.

As show in Figure 3.4 To enhance the details of the generated images, Gram matrices, which capture high-level target-specific style statistics, are utilized to facilitate image style transfer. Additionally, several loss functions are employed to assist in constructing the final output. In terms of the model's architecture, a Style Transform Block consisting of 9 independent residual blocks is proposed to achieve target-style image transfer. Furthermore, a Style Coder is incorporated to support pixel-wise image generation. Inspired by StyleGANv2 [74], the Style Coder captures multi-level features of the target image to enable high-quality final synthesis. The details will be illustrated in the Chapter 4.4.

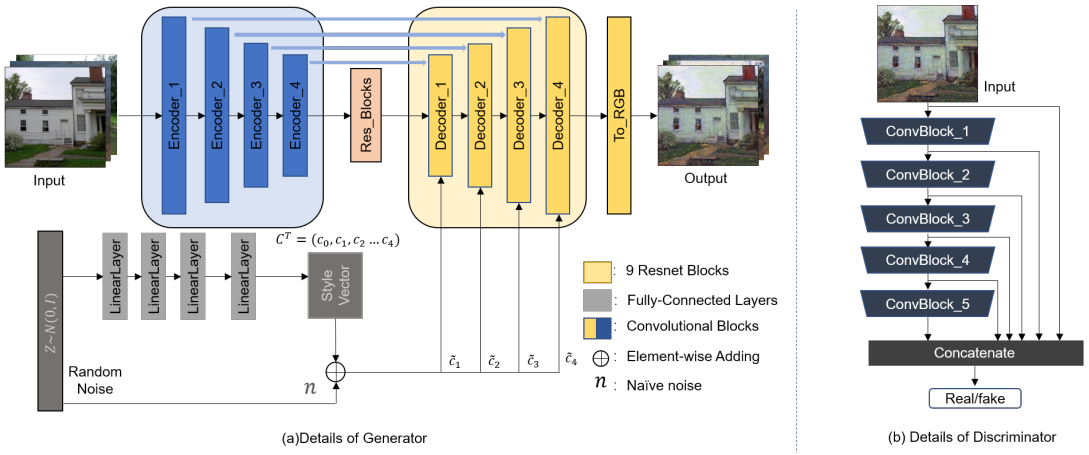


Figure 3.4: Overview of the main architecture of an encoder-decoder based network. Figure (a) shows the details of generator. In addition to the style latent vector applied in StyleGAN and StyleGANv2, deep-level feature maps extracted by encoder are also considered during skip-connection to help image generation. The StyleTransform Block which contains 9 residual convolutional layers transferring the style from content image to target image. Figure (b) illustrates the process of multi scale discriminator. From Multi scale values the discriminator supports generator captures both low level information and high level features.

3.3.2 Role in the Framework

The implementation of sketch-to-illustrative fashion image generation achieves to inspire users and designers with more creativity. As part of Part B in the framework, this functionality bridges the gap between automatically generated images and the ability to transform perspectives based on model images.

3.4 ID-Preserved Real to Illustraive Image Translation

Image-to-image (i2i) translation has achieved significant success, yet it remains challenging in certain scenarios, such as real-to-illustrative style transfer for fashion. Existing methods primarily focus on enhancing the generative model's diversity but often fall short in achieving ID-preserved domain translation. To address this lim-

itation, a novel model named Uni-DlLoRA is proposed. The related work will be discussed in this chapter, and the methodology section will introduce the fundamental principles of the model's structure.

3.4.1 Methodology

Stable Diffusion (SD) is a text-to-image model known for its robust performance in generating images from both textual and visual inputs. The architecture of this diffusion model comprises two primary modules: Autoencoders [178] and a modified UNet [149] denoiser. During the training process, the autoencoder embedded within the network is employed to encode images into a latent space, where the latent features are systematically noised in a stepwise fashion. Subsequently, the modified UNet denoiser is trained to progressively denoise these latent features, ultimately reconstructing the original image with high fidelity.

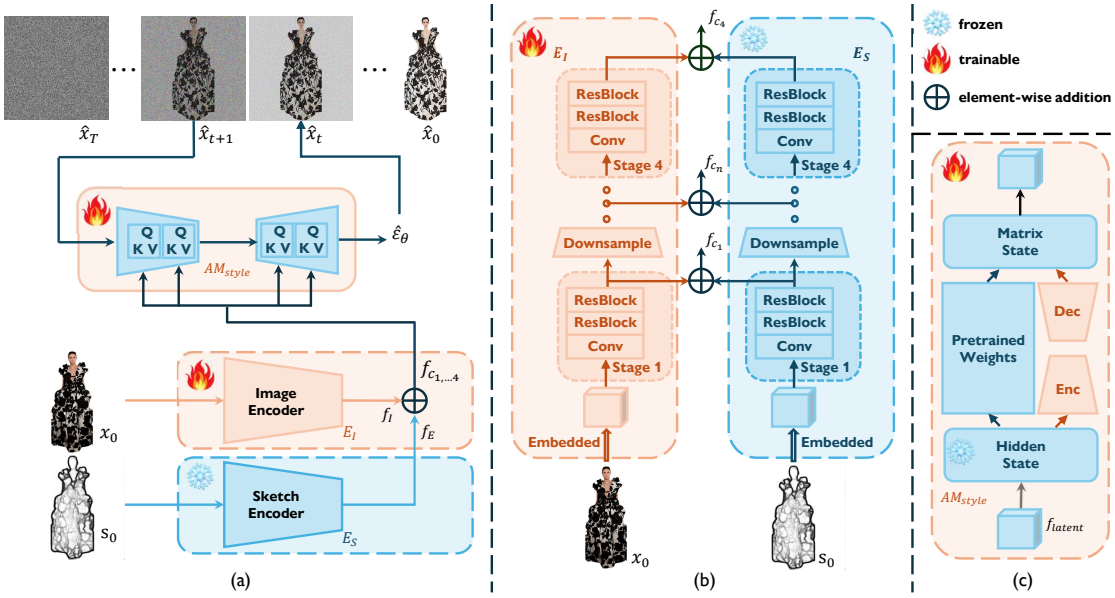


Figure 3.5: The proposed Uni-DlLoRA network includes (a) an overview, (b) a detailed process for obtaining mixed conditional embedding from multi-layer features of the image and its sketch, and (c) the specifics of the style adaptation module.

To effectively capture both texture and spatial information, hidden details are extracted from the source image using a novel multi-layer module named Uni-adapter, as illustrated in Figure 3.5. The extraction of style from target images or domains, followed by its seamless integration into source images, plays a pivotal

role in the style transfer task. Drawing inspiration from [81], two distinct style adaptation modules, termed Dual-LoRA, were incorporated into the UNet denoiser to effectively capture and adapt styles across different domains.The details will be thoroughly elaborated in Chapter 5.3.

3.4.2 Role in the Framework

The implementation of integrating specified fashion styles into real images while preserving texture details can significantly assist users and designers in completing their creative tasks more efficiently. As part of Part B, this functionality will not only accelerate the creative process for users and designers but also expand the expressive range of their work.

3.5 Capable Fashion Pose-Guided Image Transfer Modeling

Human pose transfer aims to synthesize human images with target poses based on reference inputs, a task that holds substantial economic potential for applications in e-commerce and virtual reality. In this section, a novel approach is proposed termed the Attentional Pixel-wise Deformation Network (APD-Net), designed to generate realistic human images by leveraging guided pose conditions and reference appearance details. This method focuses on precise spatial deformation and pixel-level fidelity to achieve high-quality pose-aligned synthesis.

3.5.1 Methodology

The goal of this study is to learn how to transform poses from the skeleton domain $\mathcal{A}$ to the real image domain $\mathcal{B}$, with the help of an input image $y_f \in B$. APD-Net, which learns cross-domain correspondences to provide better guidance for image translation and employs a flexible affine transformation to capture textures and local region deformations is proposed to achieve the goal. The generator

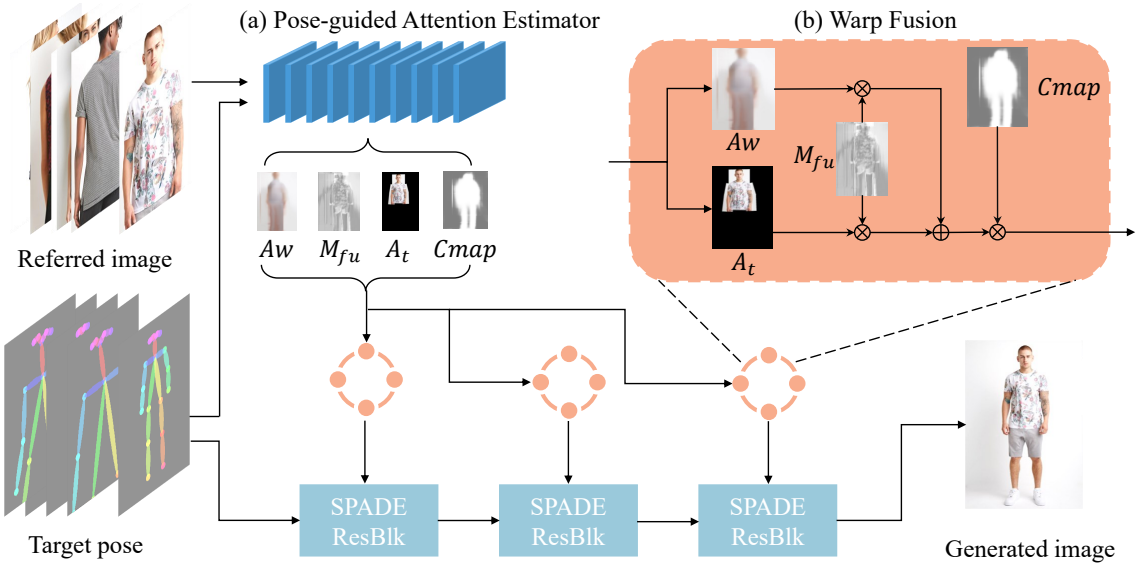


Figure 3.6: The pipeline of APD-Net for generating pose-transferred image.

combines the outputs from this estimator to obtain the final image synthesis using a SPADE resblock [133]. Specifically, the entire process is illustrated in Figure 3.6. It involves two main steps for generating pose-transferred images: 1) a Pose-guided Attention Estimator for pose estimation and 2) an Image Synthesis Module for image generation. Specifically, four elements—namely attention warp exemplar, fusion map, affine warp guidance, and confidence map in part a)—are synthesized, and attention pose-guided warp exemplars are multiplied with the confidence map in part b) to enhance image synthesis during the first stage in Chapter 6.2.1 Subsequently, cross-domain image synthesis module based on SPADE resblock [133] in Chapter 6.2.2 is proposed, aiming to synthesize final results by leveraging high-level features of reference information and target pose. Finally, the training details are explained in Chapter 6.2.3.

The objective of this study is to learn how to transform poses from the skeleton domain $\mathcal{A}$ to the real image domain $\mathcal{B}$, utilizing an input image $y_f \in B$. To achieve this, APD-Net, which learns cross-domain correspondences to provide enhanced guidance for image translation and employs a flexible affine transformation to capture textures and local region deformations, is proposed. The generator integrates the outputs from this estimator to produce the final synthesized image using a SPADE resblock [133]. The entire process, illustrated in Figure 3.6, consists of two main steps for generating pose-transferred images:

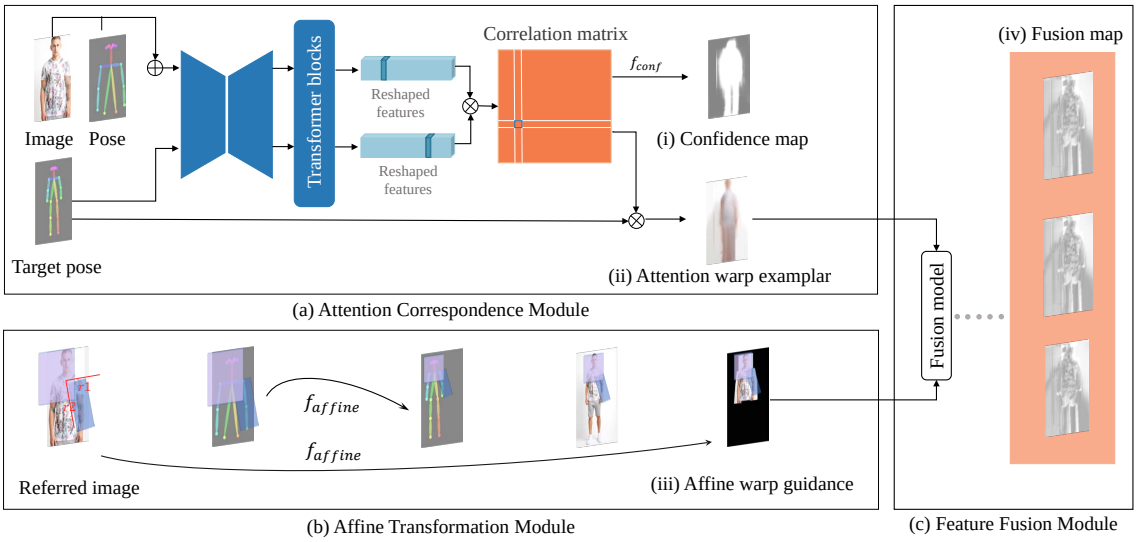


Figure 3.7: The details of the Pose-guided Attention Estimator.

- A Pose-guided Attention Estimator for pose estimation, and
- An Image Synthesis Module for image generation.

Specifically, four key elements—attention warp exemplar, fusion map, affine warp guidance, and confidence map (part a)—are synthesized. The attention pose-guided warp exemplars are then multiplied with the confidence map (part b) to refine image synthesis during the first stage, as detailed in Chapter 6.2.1. Subsequently, a cross-domain image synthesis module based on the SPADE resblock [133] is introduced in Chapter 6.2.2, aiming to synthesize the final results by leveraging high-level features from the reference information and target pose. Finally, the training details are comprehensively explained in Chapter 6.2.3.

3.5.2 Role in the Framework

The proposed pose-guided human transfer network completes the final piece of the framework 3.1 in Part C, enabling the generation of human images in the real domain according to arbitrary target poses. These target poses are defined by keypoints following the rules of OpenPose [10]. This allows the system to generate real images in any pose, providing designers with more visible information while enhancing workflow efficiency.

3.6 The Overall Fashion Image Generation Framework

The complete Fashion Image Generation Framework has been meticulously constructed as described above. For the initial sketch generation, an Automatic Hand-Drawing Sketch Generation method is employed to rapidly produce sketches. Once the hand-drawing sketch images, derived from clothing and model inputs, are obtained, various methods in Part B can transform these sketches into images with different styles. Finally, leveraging real images and desired poses, the Capable Fashion Pose-Guided Image Transfer Model enables the transformation of images into arbitrary poses. This not only enhances the efficiency of users and designers but also serves as a source of inspiration for their creative endeavors.

3.7 Chapter Summary

This chapter outlines the comprehensive framework of an intelligent fashion image generation system. Figure 3.1 provides a detailed illustration of the framework's pipeline.

Within the overall framework, the automatic hand-drawn sketch generator in Part A is discussed in Chapter 3.2. The engine incorporates a fashion landmark detection model and a fashion segmentation model, which are utilized to predict landmark keypoints and distinguish between front and back sides, respectively. The predicted results are then aligned and matched to generate the final automatic hand-drawn sketch image.

In Part B of the framework, the process of fashion image-to-image translation is addressed. MiniGAN 3.3 and Uni-DlLoRA 3.4 are introduced to facilitate fashion style transfer based on reference images from different domains. Specifically, MiniGAN employs a Style Transform Block and a Style Coder to achieve target-style image transfer, utilizing the StyleGANv2 architecture for final image synthesis. On the other hand, Uni-DlLoRA incorporates two independent LoRA structures within an SD framework to generate illustrative images while preserving the details of the

content image.

For multi-pose model image generation in Part C, the Attentional Pixel-wise Deformation Network (APD-Net) is proposed. This network leverages cross-domain correspondences and a SPADE generation block to achieve consistent results, effectively bridging the gap between Part B and Part C in the overall framework 3.1.

Chapter 4

Toward Informative and Uninformative Image Transferring

This chapter introduces the preprocess work of uninformative hand-drawing sketch image generation and an empirical approach using a generative network named MiniGAN to address both informative and uninformative image transfer tasks. These two methods achieves the developement of the automated hand-drawing sketch generation engine and transform sketch images to illustrative images, respectively.

In details, the fashion hand-drawing sketch generation engine is accomplished through three steps: fashion landmark detection, fashion segmentation, and fashion item composition. This engine utilizes a landmark detection model and a fashion segmentation model, integrating them to create an automated fashion sketch image generation system.

For MiniGAN, the generator is inspired by StyleGANv2, incorporating an Encoder and a StyleTransform Block. These components are crucial for extracting high-level feature maps from the source image and capturing the latent representation of the target image, respectively. This information guides the generator in producing the final image. MiniGAN demonstrates superior performance in style transfer, maintaining color fidelity in informative images.

The preprocessing generation work which proposed in Chatper 3.2 implement

the processes of part A and MiniGAN 3.3 implement the fashion image to image translation work in part B in the overrall framework of intelligent fashion image generation system 3.1 illustrated in Chapter 3.

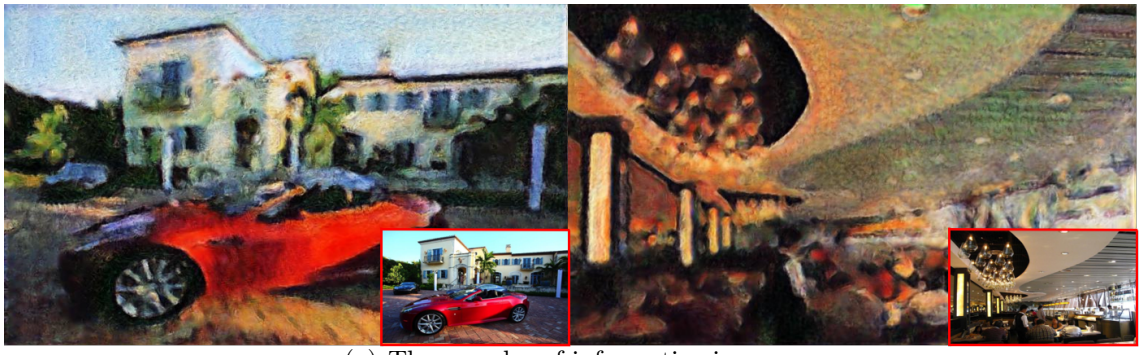
4.1 Introduction

The task of Image to Image Translation [218, 62] is to learn an appropriate mapping function from source image to target image. Recently generative adversarial networks (GAN) based methods perform well to colorize the gray-scale real images [126, 9], combine the content image with the style of the target image [68, 62, 155] and two or multi objects translation [218, 132]. Although these methods achieve promising results in some scenarios that dealing the informative images such as day scene to night scene and real photo to Monet-style photo, to our best knowledge, tackling the problem of uninformative image transforms is absent.

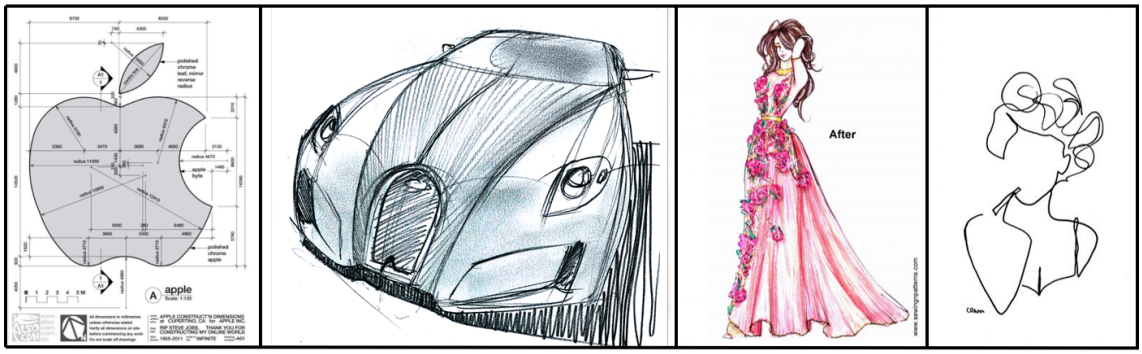
As shown in Figure 4.1(a), most images are informative enough. On the other hand, there still have large amount of images belonging to Minimalism (as shown in Figure 4.1(b)). This kind of images mainly appears in design creation process, such as illustration, technical drawing or hand sketches. Uninformative image transferring thus has highly practical value as the same as informative image transferring.

In this chapter, a novel method is proposed to tackle the problem of both informative and uninformative image transferring while the informative images have the characteristics including: (1) richful color (2) diverse space distribution (3) multi level depth. On the other side, the characteristics of uninformative images are: (1) relatively simple color (2) line drawings (3) less feeling of spaciousness. Specifically, StyleTransform Block which contains 9 independent residual blocks is proposed to carry out target-style image transferring. Besides, Style Coder is applied to support the pixel-wise image generation. The Style Coder inspired by StyleGANv2 [74] captures the multi-level features of the target imagefor final synthesis.

Qualitative and quantitative comparison is evaluated to verify the effectiveness of proposed model. Additionally, a new dataset with images which are composed of fashion line drawings to test the performance in uninformative dataset. Extensive



(a) The samples of informative images.



(b) The samples of uninformative images.

Figure 4.1: Samples of the informative images and uninformative images. The upper image shows the samples of informative images. While the input image is full of details and there is no pure color exist in big region. Besides, the image is layered. The under images illustrate the samples of uninformative images. The background of the input is clean and the color of the input is simple and pure. Besides, there is no multi-depth in the sketch.

experiments in both informative and uninformative images show that the proposed model outperforms the aforementioned models in image details preservation. The analysis of challenges of uninformative image transferring can be a good reference for future exploration in the related tasks.

The remainder of this chapter is organized as follows: Chapter 4.3 offers an overview of related work concerning current techniques in fashion image style transfer. Following this, Chapter 4.4 details the methodology employed by the proposed network. In Chapter 4.5, The experimental results of the proposed method is presented to demonstrate its effectiveness. Finally, Chapter 4.6 concludes the chapter by summarizing the key findings.

4.2 Preprocess Work

As outlined in Chapter 1.2, the primary objective of hand-drawing sketch image generation was to develop of a fashion hand-drawing sketch generator. Two core research objectives inside this chapter are: the implementation of a landmark keypoint detection model, and the implementation of a fashion segmentation network.

4.2.1 Fashion Hand-drawing Sketch Generation

To achieve auto fashion hand-drawing sketch generation, the previously mentioned keypoint detection model and segmentation model, as discussed in the methodology, will be employed. These components are poised to play a crucial role as intermediate outcomes within the process of hand-drawing sketch generation.

4.2.1.1 Implementation

Since there is no existing dataset that focuses on the front and back classification, a new sketch dataset that carries this information is collected for model training. After that, Hrnet is applied to obtain the front and back region of the sketch as Figure 4.2 illustrates.

Figure 4.2: Examples of sketch segmentation results. The red region illustrates the back part and the purple region denotes the front part.

Given that keypoint positioning constitutes vital information for hand-drawing sketch generation, the fusion of [168] and [188] methodologies ensures accurate keypoint determination within the sketch. As illustrated in Figure 4.3, the provided samples showcase the successfully identified clothing keypoints through the employed model.

Figure 4.3: Samples of recognized clothing keypoints.

The final hand-drawing sketch generation is based on the keypoint and the seg-

mentation map. The scale and position of the garments are calculated by matching the landmarks of the garment with the keypoints of the human body. The output image is generated by synthesizing each item into the baseboard in the given order.

Figure 4.4: Samples of hand-drawing sketch generation results.

Figure 4.4 illustrates the samples of hand-drawing sketch generation results. From the results, it can be shown that the engine achieves good performance.

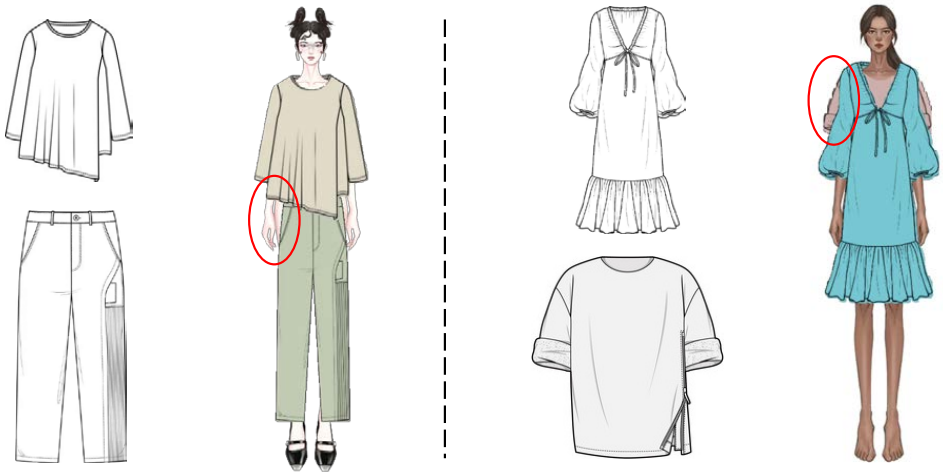


Figure 4.5: Samples of failed hand-drawing sketch generation results.

Figure 4.5 demonstrates several failure cases in the hand-drawing sketch generation process. Since the system cannot perform image deformation transformations, misalignment issues arise when the garment features do not properly match the model's body positions. For instance, when the sleeve cuffs do not align correctly with the model's arm positions, defective results are produced where the clothing items improperly positioned relative to the human figure.

4.3 Related Work

Generative adversarial network [40] is an algorithm carrying out to image synthesis. There are two key issues that need to address while first is to improve the quality of generated image while the other is to avoid mode collapse when doing synthesis. In recent, a set of progressive-like generators [71, 7, 73, 74] have been proposed to generate image with textures and details. However, these algorithms need expensive computation cost and detailed datasets. Earth-Mover distance based GAN [2] is proposed to address the problem of mode collapse. Later, methods like [42] and [123] have been proposed to stablize the quality of image generation. For these methods which achieved promising results in image generation, they are lack of ability of controlling the mode of the generated image when doing image synthesis.

Style Transfer is a task to generate a new target-like image using linear mapping way based on the content information and style information extracted from the content input and the target input. Gatys et al.[34] firstly proposed a novel algorithm that the generator using iterative optimisation ways learns the matrices-wise correlation in deep feature space extracted by pretrained deep neural networks. While the generated image fuses the content from the content input and style from the target input to generate positive result, the computation cost is relatively high. To achieve faster style mixing, single forward neural networks [90, 56, 91, 158, 195] are introduced to sharply decrease the computation time. Yao et al.[193] adopted the advantage of single forward network with multi-stroke consideration and proposed an attention-aware method to improve the quality of generated image. However, these methods require more or less style images as necessary input. Furthermore, this type of methods alters the not only texture and details but also color distribution when doing style mixing. In other words, these methods are limited to transfer the source images to the target-like images while preserving the color information of the source image.

Image to Image Translationis to learn a mapping from source domain to target domain. Recently some researches have achieved promising outcomes. Pix2pix [62] is the first GAN-based method to transfer the image from two different

domains. However, it still needs paired dataset to generate images with high quality. To overcome this, several architectures like cycle-consistency [218] and shared latent space [107] were introduced. Very recently, algorithms [57, 85, 132] based on these two architectures have been introduced to improve the quality of image. Though image to image translation can achieve good quality as well as multi-modal results, the scenario is limited as the two unpaired dataset are all informative datasets. Furthermore, there has no research focused on uninformative style transferring currently.

4.4 Methodology

The main goal is to deal with informative image and uninformative image transferring. For GAN-related algorithms, the task is to generate vivid image with rich details like reasonable texture and easily-recognised shape. Several image-to-image based models [132, 85, 146] work in two information-rich domains like real photo to paintings, etc. However, as indicated in Figure 4.6, those methods perform not good enough when the two domains are lack of information.

In order to fix this problem, Gram matrices, which captures the high-level target-specific style statistics, is adapted to carry out image style transfer. Moreover, several loss functions are applied to help construct the final output. Figure 3.4 illustrates the main structure of generator. In (a), given source image as an input, the model applies a encoder with residual blocks[47] to extract low-level details as well as high-level features of the source image. After being transferred by Style-Transform Block, high-level features maps will be fed into the generator to support the generation of the target-style like image. Inspired by StyleGAN [73, 74], a Style Coder C was involved to do style-mixing to assist generator for producing target-like image. The skip-connection supports the generator preserve the details of source image. In Figure 3.4 (b), it demonstrates the architecture of multi scale discriminator. Low level features and high level features are both considered when the discriminator distinguish whether the input image is true or fake.

Encoder In order to get latent feature maps, network based on four residual-

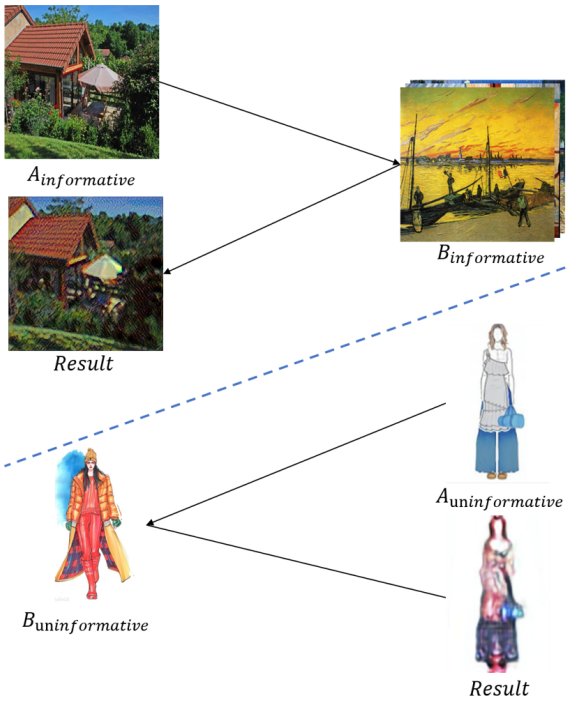


Figure 4.6: different scenarios when doing image to image style transfer. The top of the figure illustrates the sample of two traditional informative image domains. The left hand side is from the real image dataset while the right hand side shows the artistic image. In the bottom of the figure A_{un} samples from the uninformative simple generated images while the other side shows the target image named Fashion illustration.

based blocks [47] are adopted to extract features from the source image. $E_i, i \in \{0, \cdots, N\}$ while E_i is the ith block in residual-based network. Then the extracted feature maps f_i can be written as:

$$f_i = \begin{cases} E_0(\mathbf{S}), & \text{if } i = 0 \\ E_{i-1}(f_{i-1}), & \text{otherwise} \end{cases} \tag{4.1}$$

where $\mathbf{S}$ means the source image and f_{i-1} means the output of $i-1$th residual block.

On the other side, an intermediate latent vector was applied [74, 73] for style refinement. Like the implementation in StyleGANv2, the input $\mathbf{z}$ is sampled from the original latent space $\mathcal{Z}$, then a network composed of 8 fully-connected layers $f : \mathcal{Z} \to \mathcal{C}$ maps input to intermediate latent space $\mathbf{c} \in \mathcal{C}$. The dimension of both

z and **c** are 512. Style Coder C captures the style in details while feature maps f_i have the latent representation of source images. Both of them will then be utilized to fed into the generator (depicted as Decoder in Figure 3.4).

Style Transform Block From the previous style transfer methods most models do style transfer effectively in global representation. Take conditional style transfer [171] as an example, generated images are high-quality with fine details and rational textures. However the shortcoming is the color distribution of output image which is so similar as that of the target image, making it unreasonable when comparing with the original input. To tackle this problem, inspired by Style-aware [155], Style Transform Block composed of nine residual convolutional blocks is applied to transfer the image to the target-like image in latent representation.

Generator Given the multi scale features f_i and the Style Coder C, a StyleGAN-based generator is set as our main generator to carry out mixing style image generation. In order to transfer the input image to the target style while remaining the color distribution better, refined convolutional block and skip-connection [149] are applied in each style block. Furthermore, to obtain stochastic details in final output, noise inputs are applied in refined style block. The format is shown as:

$$g_i = \begin{cases} G_i\left(\mathbf{c}_0, f_N, noise\right), & \text{if } i = 0 \\ G_i\left(\mathbf{c}_i, g_{i-1}, f_{N-i}, noise\right), & \text{otherwise} \end{cases} \tag{4.2}$$

where G_i is the refined style block in generator, g_i is the output of the style block and c_i is the ith part of style vector generated by Style Coder.

At the end of the generator there is an additional convolutional block named RGB block, which represents the output to the final image.

Loss Functions *Adversarial Loss:* GANs [40] is an effective tool to help match the distribution of source image to that of target image by playing an min-max game. In other words, generator tries to deceive the discriminators through generating as same distributions of target domain as it can while discriminator learns to distinguish the differences between real target domain and fake output. Instead of using prevalent methods [2, 42]which is difficult to achieve balance between this adversarial loss and other loss in scale, least square adversarial loss [119] is applied to supervise the generator:

$$\begin{aligned}\min_D V_{\mathrm{GAN}}(D) &= \frac{1}{2}\mathbb{E}_{p_{\mathrm{data}}(\boldsymbol{x})}\left[(D(\boldsymbol{x})-b)^2\right] + \frac{1}{2}\mathbb{E}_{p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}))-a)^2\right] \\ \min_G V_{\mathrm{GAN}}(G) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}))-c)^2\right]\end{aligned} \tag{4.3}$$

where $b = c = 1$ and $a = 0$ in this work.

Inspired by [184, 115], two more feature maps are extracted as guided feature maps,these two empirical prior are applied to support discriminator to distinguish. To obtain margin features from image, traditional Sobel kernel is utilized to extract the margin of the image. Besides the structure difference in various image, texture difference is another key objective. Although it is challenging to obtain texture features in traditional RGB channel images, transferring images into luminance and color information like YUV or Lab domain release the difficulty as the first channel represents the texture information and the other two channels show the color information which influences texture little. In order to obtain more information from shape to details, the multi-scale discriminator is adopted as shown in Figure 3.4 (b) where the discriminator is composed of several Convolution blocks, to distinguish input in both low level feature maps and high level feature maps.

Style Loss: Style loss is introduced to capture the high level feature structure as well as the texture information. Gram-matrices based style loss [34] is adopted in our work. Given Gram metrices G:

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l \tag{4.4}$$

where G_{ij} is the metrices calculated by vectorised feature maps F_{ik} and F_{jk} in layer l. Then, mean-squared loss is adopted to measure the style distance between generated image and target image. Given g as generated image and i as input image, total style loss is

$$\mathcal{L}_{\mathrm{style}}(g,i) = \sum_{l=0}^{L} w_l \frac{1}{4N_l^2 M_l^2} \sum_{i,j}\left(G_{ij}^l - I_{ij}^l\right)^2 \tag{4.5}$$

where N_l and M_l represents the number of channels in layer l and number of pixels in feature maps in layer l, respectively. G_{ij}^l and I_{ij}^l are the Gram metrices from multi scale feature maps extracted by encoder from generated image and target image, respectively. While it is an trade-off that the generated image should be shown as

similar to both target image and content image, only style loss is calculated in the deepest feature maps.

Content Loss: Content loss is utilized to preserve the global structure of the input image. Mean square loss is adopted to calculate the distance of deep feature maps extracted by content extractor. Content extractor is defined as $\mathcal{CB}$. Loss is written as:

$$\mathcal{L}_{\text{content}} = \left(\mathcal{CB}\left(G\left(\boldsymbol{I}\right)\right) - \mathcal{CB}\left(\boldsymbol{I}\right)\right)^2 \tag{4.6}$$

where $G\left(\boldsymbol{I}\right)$ and $\boldsymbol{I}$ is generated image and input image, respectively.

Total Variance Loss: Due to the specific characteristics, the frequency information of painting images is different from that of real photo, which makes it difficult to generalize. To maintain the continuity of the image, total variance loss was adopted to decrease the probability of unwanted noise. The loss function is illustrated as:

$$\mathcal{L}_{tv} = \frac{1}{H * W * C} \sum_{i} \left\| \nabla_x \left(G\left(\boldsymbol{i}\right)\right) + \nabla_y \left(G\left(\boldsymbol{i}\right)\right) \right\| \tag{4.7}$$

where H, W, C means the height, width and channel, respectively. i means each pixel in the image and ∇ means the direction of the axis.

Full Structure: Overall our model can be illustrated as a network composed of encoder, style transformer, generator and an multi-scale discriminator. The full loss function is used to optimize generator in high level features as well as textures and details representation. The formulate is shown as:

$$\mathcal{L}_{\text{final}} = \lambda_1 * \mathcal{L}_{\text{multi-adv}} + \lambda_2 * \mathcal{L}_{\text{style}} + \lambda_3 * \mathcal{L}_{\text{content}} + \lambda_4 * \mathcal{L}_{\text{tv-loss}} \tag{4.8}$$

where λ_1, λ_2, λ_3, λ_4 are hyper-parameters which could be modified to generate various style images.

4.5 Experiments

Due to the limited size of the dataset, an early stopping strategy was employed during training to mitigate the risk of overfitting. Extensive experiments on

(a) informative images (b) uninformative images

Figure 4.7: Sample of images from Place365 (real photo), artworks, BeautyU (illustrations) and sketches, respectively. Figure (a) shows the samples of informative images and Figure (b) shows the samples are uninformative images.

state-of-the-art cycle-consistency [218] based models *i.e.*GDWCT [15], MUNIT [57], DRIT [85], CycleGAN [218] and style-transfer based models like Style-aware [155] and our model were conducted to evaluate the performance in both informative dataset and uninformative dataset. Training performance and general results analysis in both informative and uninformative domains are presented in the following part.

4.5.1 Implementations

For optimization problem, Adam [78] algorithm was adopted in both generator and discriminator with $\beta_1 = 0.5, \beta_2 = 0.999$. The initial learning rate for generator and discriminator are $lr_g = 10^{-4}$ and $lr_d = 10^{-5}$, respectively. The batch size was set as 2 and the model was trained with about 100000 epoch or until was reached convergence.

Hyper-parameter The default hyper-parameters were set as: $\lambda_1 : 1, \lambda_2 : 100, \lambda_3 : 100, \lambda_4 : 1e^{-2}$ while the multi-scale weights in adversarial loss were all set as 1. The default parameter is based on the training dataset which has high bias in real domain and target domain. Refined parameters were used in other domain-based scenarios to ensure satisfactory performance.

Dataset Images is sampled from Places365[212] training dataset as our source domain. For target domain, the collection of Cezanne and Van Gogh images is adopted from WikiArt [1]. For source domain there are over 300000 images while there are 999 images in target domain. To get more images in target domain, data augmentation like rotation and flipping were used to help create "new artistic image". Images shown in this paper are sampled from testing image dataset of Place365. For validation dataset, several images sampled from Place365 testing dataset is applied to measure and compare. In order to obtain high-resolution images, all images in two domains were resized to $512 * 512$ resolution. For the purpose of uninformative image style transfer, 5817 samples were used in both BeautyU and in single sketch dataset. For those uninformative datasets, the size of images is set as $384 * 256$ which is the same as the size of the images in BeautyU. Figure 4.7 shows the sample of each dataset.

4.5.2 Qualitative Comparisons

Figure 4.8 illustrates the comparison between the four benchmarked methods and our method in informative domain image style transfer. Due to the effectiveness of skip-connection, the image generated from our method has clear contours and details. For images generated by style image guided algorithms, from the aspect of style, generated images contain rich information of texture. However, they inevitably have the color information from guided images, which is opposite to our expectation. For unsupervised algorithms such as MUNIT and DRIT, while they capture both content and style latent representation of target images, the outputs lose the color information in style transferring process. For image-guided algorithm GDWCT, it is difficult for outputs to obtain the style of target images. Furthermore, to obtain color-invariant generated images is another challenge.

[1] `https://www.wikiart.org/`

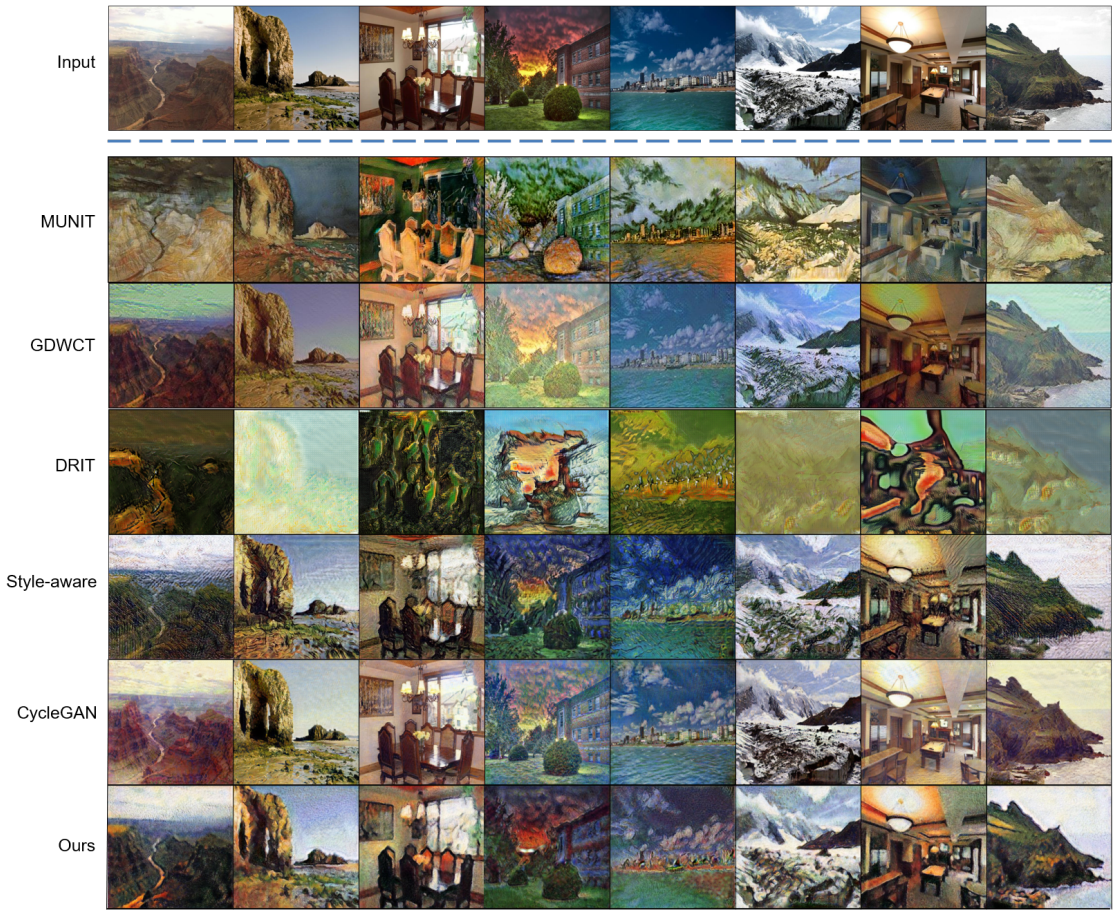


Figure 4.8: Qualitative comparisons on informative style transfer. To get the stroke of the target domain as well as remain the global structure of the content image, Our models outperforms other models in details preserving as well as style representation. For Style-aware, the content of the generated image lose much that the contour of the image looks messy.

4.5.3 Quantitative Comparisons

Frechet Inception Distance (FID) [48] which is an algorithm to calculate the Frechet distance between two Gaussian-Mixed based probabilities is adopted in this work as it is an ideal distance to evaluate how close two probabilities are and evaluate the quality of generated images. As is illustrated below:

$$\mathrm{FID} = \|\mu_g - \mu_t\|^2 + \mathrm{Tr}\left(\Sigma_g + \Sigma_t - 2\left(\Sigma_g \Sigma_t\right)^{\frac{1}{2}}\right) \tag{4.9}$$

where in the format $V_g \sim \mathcal{N}(\mu_g, \Sigma_g)$ and $V_t \sim \mathcal{N}(\mu_t, \Sigma_t)$ demonstrate the mean and variance of two vectors of generated and target images extracted from

pre-trained Inception-v3 [172] model, respectively. The lower the FID scores, the better quality the image synthesize as well as better performance the model has.

Model	Resolution	FID_{photo}	FID_{paint}	LPIPS
Style-aware [155]	512*512	161.99	96.78	**0.583**
DRIT [85]	512*512	139.55	106.20	0.749
MUNIT [57]	512*512	78.82	**84.73**	0.663
GDWCT [15]	512*512	29.76	112.08	0.836
CycleGAN [218]	512*512	**20.00**	130.12	0.78
Ours	512*512	133.83	105.99	**0.632**
Photo	512*512	\	157.36	\

Table 4.1: Table illustrates the FID distance as well as LPIPS scores between the real photo dataset and painting dataset. The lower score indicates better stylization results. The aim of the task is to get lower score from photo dataset as well as from painting dataset while it is an trade-off. For LPIPS, lower score means the generated image is more similar to the original image. In other words, the lower score means the better the generated image preserve details. While our model is not achieved the best result in all measurements, it achieves the balenced FID scores and the second best result in LPIPS.

Furthermore, a evaluation algorithm named Learned Perceptual Image Patch Similarity [203] is also adopted to measure the quality of generated images in our work. In Table 4.1 Three cycle-consistency based models obtain the best three FID scores in the photo domain, and their scores in the painting domain are relatively high. Meanwhile, since their scores in the paint domain are relatively low, this indicates that these methods cannot effectively carry the style of the paint images. In other words, FID scores are relatively more important in the painting domain compared to those in the photo domain, because the generated images need to carry the style. Our model obtains relatively low FID scores when compared with other methods, which means that our model can capture the latent style representation from the target images. MUNIT obtains the lowest FID scores in both photo domain

and painting domain. However, the change of color distribution is undesirable. Furthermore, extremely low FID scores in real datasets mean the model changes little in the source image. With relatively lower scores compared with other methods, our method preserves the details of the content images. For LPIPS, our method and Style-aware obtain the best two scores, which means that the generated images from Style-aware and ours outperform the other methods in semantic structure representation.

4.5.4 Ablation study

In order to carry style transferring while preserving the color information. Regular encoder-decoder structure model Style-aware [155] with multi-scale discriminator is utilized as our baseline. Several components were added into our model for the sake of higher-quality image generation. Outputs generated with or without these blocks are compared to evaluate the effectiveness of them.

In Figure 4.9 (b) the images are generated with normal GAN loss. There is irrational color distribution on the whole image. Besides, the tree in the first image is in mess, which means the model is limited to transfer some objects. As the structure representation of the (b) is similar with the input, this means that the skip-connection can catch the low frequency information and high frequency information. However, it preserves the details of the content image so good that it cannot catch the style of the target images. In Figure 4.9 (c), output images are generated by the model when it trained without adding the noise. The color distribution of generated image is shifting when comparing with the input. Besides, mode collapse appeared in several places like the branch of the trees and the top of the car. Images generated from the full model alleviate the color difference while learning the latent representation of style image.

In addition to image demonstration. Table 4.2 illustrates the FID distance and LPIPS scores among three models. Model (b) obtained the highest FID distance in real dataset and LPIPS scores, which means that preserving the content of the source image and the style of target image is limited. For model (c), while it achieved the best result of LPIPS scores, the FID distance between real dataset and generated

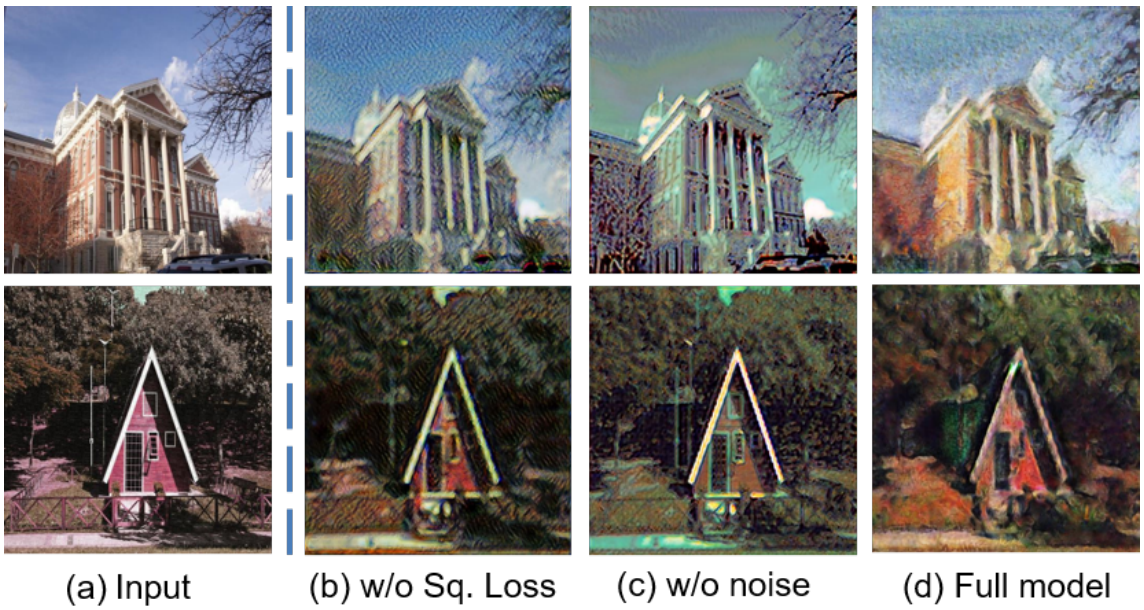


Figure 4.9: Ablation studies with or without some components. From the figure (b) while it performs well in the under image, there are some irrational black lines on the trees. For (c) two images are not realistic, and for the upper image there is black blob in some places.Also, the color distribution is not as natural as original image. For our model the generated image has more realistic color distribution as well as the painting-like stroke.

Model	FID_{photo}	FID_{paint}	LPIPS
(b) w/o Sq.loss	149.23	170.26	0.633
(c) w/o Noise	71.54	141.66	**0.468**
(d) Full Model	133.83	**105.99**	0.632
Photo	\	157.36	\

Table 4.2: Table illustrates the FID distance and LPIPS scores between the real photo dataset and painting dataset in ablation study. According to the table4.2, (c) obtained the lowest distance in FID-photo and LPIPS scores. This means (c) changed little of the input image. The full model obtained the best result in FID-paint, which means the model captured the style of the target images.

images is relatively low and that between paint dataset and generated images is high. This means the ability of style transferring is ineffective. The full model gained lowest FID distance between paint dataset and relatively high distance between real dataset. It indicates the model captured the style of the target image and preserved the information of source image well.

4.5.5 Analysis on Uninformative Dataset

From the human vision perspective, some images generated from existing algorithms can deceive the expert in a certain extent from layout to texture and color distribution. Those well performed image transferring models stand with two informative image domains. In other words, there is little research focused on uninformative image transfer. For our images which belongs to uninformative domain, the aforementioned five algorithms and ours are compared and the result is illustrated in Figure 4.10.

In Figure 4.10 the above methods perform not well in this task. For the four cycle-consistency based image translation methods (MUNIT, GDWCT, DRIT and CycleGAN), CycleGAN preserved the content of the source image in this scenario, but the style of the generated image changed little. For the other three aforementioned methods, mode collapse more or less appeared when they carry out image synthesis. In Figure 4.8, DRIT can carry the style of target domain. However, it is limited to preserve the content of the input image. For MUNIT and GDWCT, while the generated images somehow capture the style representation of target image, the color and full shape of the body in images are out of control. For Style-aware, the strokes in the images are different with both that of input images and that of target images. It has thick, straight lines, which is same as its performance in real-painting image transfer, instead of thin and curve lines. However, this method preserves body shape well. For our model, it preserves well in details in content domain but still lack of style representation from the target domain. Moreover, the output has blurry background, which decreases the quality. Due to the noise-adding structure inside the generator, the model tends to add some unexpected blurring in the background. In order to remove those unexpected blurring, two more methods are proposed. One

Figure 4.10: Uninformative style transfer. For the first three algorithms there is more or less mode collapse. Images generated from the latter three algorithms changes little from the original domain but stroke and color distribution.

method is to use mask to extract the main part and removing the other region to white clean. The other method is to combine the input with its mask to generate new 4-dimension input. And the new input will be fed into the network. In the latter way the added mask can be seen as a white-box attention. Figure 4.11 shows its results. It is noted that in the third and the fourth line the generated image is background-clean. However the model cannot learn the layout and stroke of the target image. For the former method, using mask to get foreground is easy to understand that the interest region is the same as that of mask. In other words, only the region of interests remains. For the latter method, while the channel of mask shares the region of interests, it supports too much about the ability of shape generation and limits the style of image synthesis.

Figure 4.11: Methods to clean the background from the image generated by our method. For the third and fourth line the background is cleaned while the stylish of generated images are limited.

4.6 Chapter Summary

In this chapter, the details of hand-drawing fashion sketch image generation and a novel method MiniGAN is proposed. The main structure of MiniGAN is a encoder-decoder network with style transferring module. StyleGAN-like modulated convolution layers is applied to facilitate the representation of content. Multi-scale gan loss and variational loss are applied to strengthen the quality of generated images. Qualitative and quantitative results show that this method can generate images with target style when the dataset belongs to informative dataset. For uninformative data, the proposed method performs well in details preservation but still not satisfying enough in preserving style which will be the future work.

Chapter 5

ID-preserved Fashion Domain Image Translation

To manage the task of real-to-illustrative style transfer of fashion illustrated in Chapter 3.4, Image-to-image (i2i) translation has achieved notable success but remains challenging in this specific scenarios. Existing methods focus on enhancing the generative model with diversity while lacking ID-preserved domain translation. This Chapter introduces a novel model named Uni-DlLoRA to release this constraint. The proposed model combines the original images within a pretrained diffusion-based model using the proposed Uni-adapter extractors, while adopting the proposed Dual-LoRA module to provide distinct style guidance. This approach optimizes generative capabilities and reduces the number of additional parameters required. In addition, a new multimodal dataset featuring higher-quality images with captions built upon an existing real-to-illustration dataset is proposed. Experimentation validates the effectiveness of our proposed method.

At the same time, this proposed method successfully enables image transfer from the real domain to the illustrative domain, as demonstrated in Part B of the framework 3.1 in Chatper 3. This advancement not only enhances the overall framework's capabilities but also significantly improves designers' workflow efficiency.

5.1 Introduction

The advancement of generative models has revolutionized the field of computer vision, particularly in fashion, where the creation and manipulation of images play a pivotal role[8, 43, 160, 67]. Fashion synthesis [221, 77, 41, 214, 70] has emerged as a dynamic area of research, encompassing a spectrum of applications from virtual try-on to appearance and pose transfer. Despite these advancements, the translation of fashion images between distinct domains, such as illustration and realism, remains a challenging topic. This translation is critical for fashion creation and understanding the nuanced interplay between style and content in fashion imagery.

Recent works have begun to explore the synthesis of fashion images, with StylishGAN [225] introducing a dataset that bridges the gap between real and illustrated fashion domains. However, existing methods of fashion image synthesis, while making significant strides, exhibit several limitations: (1) Lack of Dataset Quality: Current fashion illustration datasets often suffer from low resolutions and the presence of backgrounds in real domain images, which hinder the training of models to focus on fashion items exclusively. (2) Inadequate Style Capture: Existing generative models struggle to accurately capture and replicate the specific stylistic elements of fashion items, particularly when translating between domains with distinct visual characteristics. (3) Limited Style Control: Text-based style transfer methods lack precise stylistic control, resulting in inconsistent and less realistic outputs due to the insufficiency of textual descriptions to convey complex visual styles. (4) Style Adaptation Challenges: Although several methods[52, 157] mitigate catastrophic forgetting with low-rank matrices, they struggle to learn specific styles due to poor alignment between condition information and the model's internal knowledge.

To this end, Uni-DlLoRA, a novel approach that focuses on the fine-tuning of diffusion models is presented for fashion image synthesis and improving style disentanglement. SwinIR [94] and LDSR [148] are utilized to enhance the resolution and clarity of images in the StylishU dataset by performing super-resolution. The caption of each image is extracted by BLIP [89] and refined by fashion experts for text-conditioning. Our method, Uni-DlLoRA, is designed to address the limitations of current techniques by incorporating image-conditioned information using the pro-

posed Uni-adapter and adapting the UNet denoiser with the Dual-LoRA module to better capture spatial and textural details from both real and illustrative domains. Uni-DlLoRA enables the seamless translation of fashion images while preserving their essential visual features and stylistic elements. Style features are disentangled from the target images or domains and integrated into the source images to achieve stylistic consistency. Qualitative and quantitative comparisons with state-of-the-art methods demonstrate the effectiveness of Uni-DlLoRA. All in all, the contribution of this chapter can be summarized as:

- This article highlights a novel method that fully applies a Uni-adapter to extract latent features from input images and enhances learning in fashion image translation through the novel Dual-LoRA module.
- The article presents a new dataset in response to the existing challenges in the fashion field, which features graphics with better resolution and accurate textual information.
- Additionally, an innovative training method successfully generates images full of detail while effectively disentangling the content and style of the images. Detailed experiments describe the effectiveness and practicality of the method.

5.2 Related Work

5.2.1 Fashion Image Synthesis

Fashion synthesis is a burgeoning research domain within the expansive realm of computer vision. In particular, numerous approaches [45, 18, 87, 220] focus extensively on virtual try-on, a process that involves transferring desired clothing onto a specific person. Other studies [116, 144, 5] concentrate on appearance and pose-guided transfer. Recently, image editing has gained popularity, with several methods [66, 207, 221, 65] focusing on the editing of specific elements onto clothing. Some of these methods like SGDiff [169] have achieved significant results through the use of diffusion models, enabling text editing to become a reality. Nevertheless,

the translation of fashion images between illustration and real domains remains relatively unexplored compared to other areas within the fashion industry, despite being an important process in fashion creation. StylishGAN [225] first introduced this task into the field of computer vision and developed a dataset containing fashion images from both real and illustrated domains. However, there is still room to improve the quality of the dataset and the generative model.

5.2.2 Fashion Image-to-Image Translation

Image-to-image (i2i) translation is a widely studied and popular research topic introduced by Isola [62]. The main goal of this task is to accurately and effectively translate an input image into an output image while preserving important visual features and details. This can be used for various applications such as style transfer [34] and image synthesis [218]. Several methods [108, 21, 11] apply a content image and a style reference image to create an image that captures the style of the reference while retaining the content of the original during the generation process. However, the texture and color of the style images are hard to disentangle. Though AAST [54] proposed a model that transfers the images to the target domain while considering the texture and aesthetic, blurred background exists during generation. Other methods [132, 102] tried to transfer the style images to the certain style with pre-trained networks, but failed to transfer uninformative images [225] to another domain.

Afterwards, text-driven image-to-image translation has gained traction, with several methods [176, 134] achieving significant results by leveraging powerful generation models such as the diffusion model. However, the utilization of text-driven information is constrained in effectively conveying styles or emotions, as objects are easily described, while styles are challenging to articulate in words.

5.2.3 Fine-tuning based on Diffusion Models

The diffusion models [148, 153] have recently gained significant popularity and fine-tuning models based on them are widely used for downstream tasks. How-

ever, the over-fitting and mode collapse exists while training the neural network with additional training data. Extensive research paid attention to avoiding such issues. For instance, Dreambooth [151] and Textual Inversion [33] customize the content in the generated image by fine-tuning the image diffusion model with a small set of user-provided example images. However, fine-tuning the entire model has a high computational cost. Lora-Rank Adaptation (LoRA) [52] noted that over-parameterized models exist within a low intrinsic dimension subspace, and thus this method prevents catastrophic forgetting by obtaining information on the parameter offset using low-rank matrices. However, learning specific styles applying LoRA can be challenging. Based on substantial results obtained by adapter methods adopted in pretrained model [141, 166] in several downstream tasks, T2I-Adapter [125] and Controlnet [200] adapt Stable Diffusion to different external conditions and learn the alignment between condition information and internal knowledge, achieving solid results. However, T2I-Adapter finds it challenging to learn the style, while ControlNet struggles to strike a balance between model capability and computational cost.

5.3 Methodology

5.3.1 Preliminaries

The Stable Diffusion (SD) is a text-to-image model known for its strong performance in generating images from text and images. It comes with pretrained checkpoints, making it the chosen backbone model. The diffusion model consists of two major modules: Autoencoders [178] and a modified UNet [149] denoiser. In the training process, the autoencoder within the whole network will be utilized to encode the images into a latent space, and the latent features will be deliberately noised in a step-by-step manner. After this stage, the modified UNet denoiser is trained to denoise the latent features step by step. The optimization of denoising could be written as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0,\mathbf{c},\boldsymbol{\epsilon},t}\left(\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta\left(a_t\mathbf{x}_0 + \sigma_t\boldsymbol{\epsilon}, \mathbf{c}\right)\|_2^2\right), \tag{5.1}$$

where $\mathbf{x}_0$ denotes the input latent features and $\mathbf{c}$ illustrates the optional conditional informaition. $\boldsymbol{\epsilon} \in \mathcal{N}(0, \mathbf{I})$ represents the added noise and $\mathbf{x}_t = a_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$ denotes noised input latent features in step t. $\hat{\boldsymbol{\epsilon}}_\theta$ represents the predicted noise from UNet denoiser with conditional informaiton $\mathbf{c}$ according to the Classifier-Free Guidance[50] in the training stage:

$$\hat{\boldsymbol{\epsilon}}_\theta \left(\mathbf{x}_t, \mathbf{c} \right) = \omega \boldsymbol{\epsilon}_\theta \left(\mathbf{x}_t, \mathbf{c} \right) + (1 - \omega) \boldsymbol{\epsilon}_\theta \left(\mathbf{x}_t \right), \tag{5.2}$$

where ω is a guidance weight. After the denoising stage, the final image is generated from the cleaned latent features $\hat{\mathbf{x}}_0$ during the decoder part of the Autoencoders. For inference, the latent features $\mathbf{x}_T$, whether originating from random noise or noised input latent features, become progressively clearer as the predicted noise $\hat{\boldsymbol{\epsilon}}_\theta$ is applied at each step t to denoise the latent features, transforming $\mathbf{x}_T$ into $\hat{\mathbf{x}}_0$ with equation:

$$\hat{\mathbf{x}}_{T-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_T - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \hat{\boldsymbol{\epsilon}}_\theta \left(\mathbf{x}_T, c \right) \right) + \sigma_t \mathbf{z} \tag{5.3}$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ denotes the gaussian noise. To capture the textual information during the denoising stage, the pretrained CLIP [137] is applied to embed text prompts into a sequence of vectors $\mathbf{c}_v$ in the latent space. These vectors are then utilized by the cross-attention module inside the UNet denoiser to aid in the denoising process. The equation can be written as:

$$\textit{CrossAttention} \left(\mathbf{q}, \mathbf{k}, \mathbf{v} \right) = \text{softmax} \left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_k}} \right) \cdot \mathbf{v} \tag{5.4}$$

where $\mathbf{q} = \mathbf{w}_q \phi \left(\hat{\mathbf{x}}_t \right), \mathbf{k} = \mathbf{w}_k \tau(\mathbf{c}_v), \mathbf{v} = \mathbf{w}_v \tau(\mathbf{c}_v)$. $\phi(\cdot)$ and $\tau(\cdot)$ denotes the embedding matrices inside the module and $\mathbf{w}_q, \mathbf{w}_k, \mathbf{w}_v$ represents the weight of projection matrices.

5.3.2 Diffusion Model with Image Conditioned

For the basic diffusion model in the T2I task, the textual information will be embedded firstly into the latent space by pretrained CLIP[137] and then fed into the cross-attention module inside the UNet denoiser. The generated results are unstable when the input consists solely of text, as text struggles to convey spatial information effectively. The lack of alignment in the results arises from the inherent difficulty

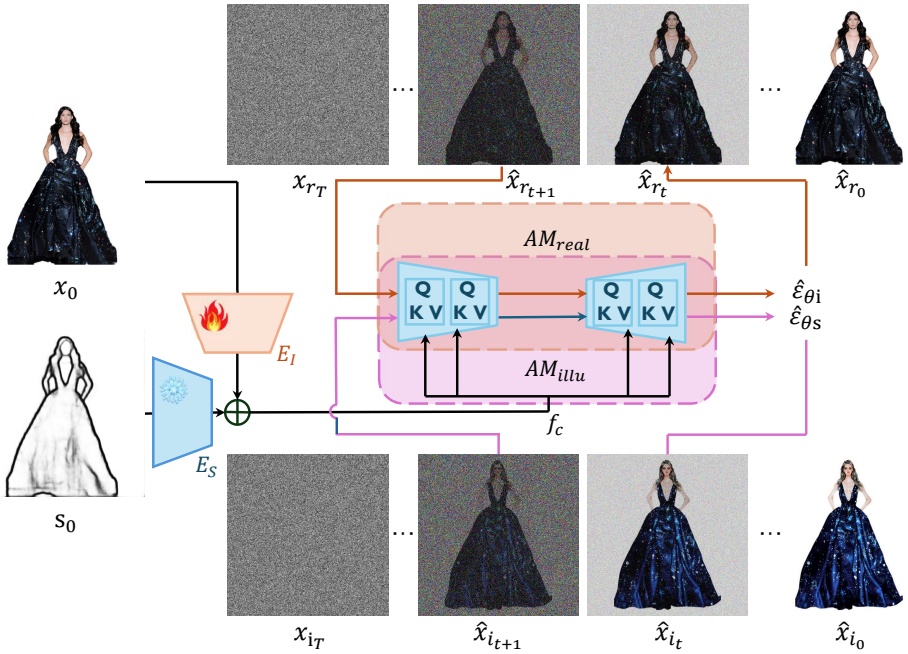


Figure 5.1: Detailed training process: The mixed conditional embedding is sent to the modified U-Net denoiser for various tasks. An illustration adaptation module is inserted for the synthesis of illustrative images, while a real adaptation module is employed to synthesize real images.

of text in offering precise external control. To effectively capture both texture and spatial information, hidden details are extracted from the source image using a novel multi-layer module named Uni-adapter, as depicted in Figure 3.5. Inspired by T2I-Adapter [125], the pixel unshuffle [159] operation inside the extraction module is firstly applied to downsample the input. The multi-convolutional layers including two residual blocks are then applied to extract the unshuffled features and multi-scale features will be obtained as: $f_c = \{f_{c_1}, f_{c_2}, f_{c_3}, f_{c_4}\}$. Due to the alignment of latent features from two same-structure extract modules, the equation for the mixed conditional embedding is:

$$f_{c_i} = \phi_{E_I^i}\left(\mathbf{x_0}, \theta\right) + \phi^*_{E_S^i}\left(\mathbf{s_0}, \theta\right), i \in \{1, 2, 3, 4\} \tag{5.5}$$

5.3.3 Style and Content Disentanglement

The extraction of style from target images or domains, followed by its integration into source images, is significant within the context of the style transfer task.

Inspired by [81], two separate style adaption modules named Dual-LoRA were inserted in the UNet denoiser to capture the styles in different domains. As shown in Figure 5.1 (c), full-rank dense layers within the module that perform matrix multiplication are integrated into the pretrained UNet denoiser to adjust the style of the synthesized image. Specifically, the inclusion of parameters in both the image feature extractor and the fixed sketch feature extractor enhances the model's ability to extract spatial and textural information from the input. Specialized style adaptation modules with learnable parameters are inserted into the UNet denoiser to aid in refining the style of the synthesized images, as well as in content and style disentanglement. Unlike simple LoRA [52], two separate style adaption modules within Dual-LoRA are applied to assist specific noise prediction with an equation at each step t:

$$\begin{aligned}\hat{\mathbf{x}}_{r_{t-1}} &= \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{x}}_{r_t} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\epsilon}}_{\theta r}\left(\hat{\mathbf{x}}_{r_t}, f_c, \theta_r\right)\right) + \sigma_t \mathbf{z} \\ \hat{\mathbf{x}}_{i_{t-1}} &= \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{x}}_{i_t} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\epsilon}}_{\theta i}\left(\hat{\mathbf{x}}_{i_t}, f_c, \theta_i\right)\right) + \sigma_t \mathbf{z}\end{aligned} \tag{5.6}$$

Specifically, the predicted noise in the process can be written as:

$$\begin{aligned}\hat{\boldsymbol{\epsilon}}_{\theta r}\left(\hat{\mathbf{x}}_{r_t}, f_c, \theta_r\right) &= \omega \boldsymbol{\epsilon}_\theta\left(\hat{\mathbf{x}}_{r_t}, f_c, \theta_r\right) + (1-\omega)\boldsymbol{\epsilon}_\theta\left(\hat{\mathbf{x}}_{r_t}, \theta_r\right) \\ \hat{\boldsymbol{\epsilon}}_{\theta i}\left(\hat{\mathbf{x}}_{i_t}, f_c, \theta_i\right) &= \omega \boldsymbol{\epsilon}_\theta\left(\hat{\mathbf{x}}_{i_t}, f_c, \theta_i\right) + (1-\omega)\boldsymbol{\epsilon}_\theta\left(\hat{\mathbf{x}}_{i_t}, \theta_i\right)\end{aligned} \tag{5.7}$$

where $\hat{\boldsymbol{\epsilon}}_{\theta r}$ and $\hat{\boldsymbol{\epsilon}}_{\theta i}$ denotes the predicted noise for real style and illustration style images reconstruction, respectively. $\boldsymbol{\epsilon}_\theta\left(\cdot, \theta_r\right)$ and $\boldsymbol{\epsilon}_\theta\left(\cdot, \theta_i\right)$ represent the basic UNet denoiser adding real-style adaption module and illustration-style adaption module, respectively.

5.3.4 Training Objectives

As discussed in Section 5.3.1, the diffusion algorithm progressively adds the Gaussian noise into the original image $\mathbf{x}_0$ with t times and obtains noisy image $\mathbf{x}_t$. The diffusion models will implicitly learn to reconstruct an image from the noisy image by predicting the added noise depending on the timestep t and task-specific conditions c_t. During the training process of our proposed method, images in the

real domain are utilized as conditions to provide spatial and texture information, as depicted in the figure 5.1. Given that two separate style adaptation modules are implemented within the UNet denoiser to aid individual noise prediction, a dual loss can be formulated throughout the entire training process as follows:

$$\begin{aligned}\mathcal{L}^{dual} = &\mathbb{E}_{\mathbf{x}_{i0},f_c,\boldsymbol{\epsilon_i},t}\left(\|\boldsymbol{\epsilon_i} - \hat{\boldsymbol{\epsilon}}_{\theta i}\left(\mathbf{x}_{it}, f_c, \theta_i\right)\|_2^2\right) \\ &+ \mathbb{E}_{\mathbf{x}_{r0},f_c,\boldsymbol{\epsilon_r},t}\left(\|\boldsymbol{\epsilon_r} - \hat{\boldsymbol{\epsilon}}_{\theta r}\left(\mathbf{x}_{rt}, f_c, \theta_r\right)\|_2^2\right)\end{aligned} \tag{5.8}$$

where $\mathcal{L}^{dual}$ is the overall training objective of the entire diffusion model. This objective is directly applied in finetuning diffusion models with an image extractor and Dual-LoRA modules. $\boldsymbol{\epsilon_i}$ and $\boldsymbol{\epsilon_r}$ represent the added noise for images in the illustration domain and real domain, respectively. The parameters within the pretrained UNet denoiser are fixed during the training process.

5.4 Experiments

5.4.1 Implementations

Network Architecture. Diffusion models denoise the image by applying the conditions from the prompt and the given image. However, the generated image often lacks a strong correlation with the conditional source image owing to the prompt typically not conveying precise semantic information and struggles to perfectly match the spatial and textural details from the image (as shown in Figure 5.2). Two adapters, namely the image feature extractor and sketch feature extractor, are applied to carry the multi-scale spatial and texture information from size 64×64 to 8×8 that match the spatial size of the feature maps inside the UNet denoiser to address this issue. In pursuit of style disentanglement, two distinct style adaptation modules are employed to refine the style of image generation. DDIM [165] is applied to accelerate the inference process.

Dataset. In this study, there are rarely fashion illustration paired datasets. [225] gathered a dataset StylishU that comprises 3567 paired images consisting of real photos and hand-sketch illustrations. However, the resolution of the images is relatively low, and they contain backgrounds within the real domain images. SwinIR [94] is

Figure 5.2: The StylishU-SR dataset includes runway images, paired illustrative images, and captions.

initially utilized in conjunction with LDSR [148] to perform a super-resolution version StylishU-SR, thereby obtaining images with a resolution of 512×512. During the training process, 3467 high-resolution paired images are used as the training dataset, while the remaining 100 paired images are designated as the test dataset. The textual caption of each image is extracted by BLIP [89] and refined by fashion experts for further research.

Training Details. The stable-diffusion v1-5 was utilized as the backbone diffusion model. Considering the potential semantic disparity between textual and image information, as shown in Figure 5.2, **None Prompt** is provided to the UNet denoiser, while the extracted mixed conditional embedding f_c in Equation 5.5 serves as the sole condition during the training process. The proposed model was fine-tuned on the paired dataset using the AdamW optimizer with a learning rate of $5e^{-6}$. The batch size was set to 8, and the A100 was utilized to train the proposed model for 100,000 iterations. The pretrained PIDNet [187] was employed to extract the sketch from the input images, with the threshold set to 0.5. The parameters of the sketch feature extractor were kept fixed with pretrained weights obtained from training data of COCO17 [103]. Regarding the style adaption modules, the linear encoder-decoder layers with rank=16 are set within the UNet denoiser. To ensure clean background generation, the initial noise will be combined with latent features [122] extracted from images by pretrained Autoencoders.

Baselines. The original image is utilized as the conditional information for performing fashion image style transfer. The proposed method is compared with several state-of-the-art methods, including some GAN-based [218] and diffusion-based fine-tuning [52, 125, 200] methods, both qualitatively and quantitatively. The performance of fine-tuned original Stable diffusion (SD)[148] is also evaluated. The test set of the StylishU-SR is applied to the performance of the generated results from each method.

Metrics. Following the general practice, four metrics including FID [48], LPIPS [203], CLIP-image [137], and CLIP-aesthetic [156] are applied to evaluate the quality of the generated images for comparison our method with the SOTAs. While the FID score and LPIPS score focus on the latent feature distance between ground truth and generated images, the FID score emphasizes the overall distribution, while LPIPS calculates the distance between each pair of generated images and corresponding ground truth. It is worth noting that, due to the limited number of test datasets, the FID score reported in this article is derived from latent features extracted by the first block of the pretrained CNN, which is denoted as FID_{64}. Since this score is based on low-level features, it is more concerned with the similarity between the generated image and the ground truth's underlying features. For these two criteria, the lower the FID and LPIPS scores, the higher the synthesized image quality. Conversely, the CLIP image assesses the cosine similarity between the ground truth and synthesized images, where higher scores denote better alignment. Similar to the CLIP image, the CLIP-aesthetic predictor applies CLIP embeddings with an MLP layer to predict the average preference for an image. Higher scores indicate better results.

5.4.2 Quantitative Comparisons

Quantitative Comparison: Table 5.1 illustrates the quality of synthesized images between our method and other state-of-the-art methods. For diffusion-based models, our proposed method outperforms the others in terms of the LPIPS scores. The FID score of the images from our method also achieves the best results in diffusion models, which means the generated images are of higher quality than those

Table 5.1: Qunatitative evaluation and comparison with several SOTA methods.

Methods	Metrics			
	$FID_{64} \downarrow$	$LPIPS \downarrow$	CLIP-image ↑	CLIP-aes ↑
CycleGAN	**0.454**	**0.206**	86.776*	5.322
SD(add text)	2.677	0.298	74.748	5.598*
LoRA(add text)	0.605	0.233	81.530	**5.638**
SD-finetuned	0.586	0.586	83.122	5.448
ControlNet	2.078	0.216	85.863	5.415
T2I-Adapter	0.762	0.216	85.221	5.305
Ground Truth	—	—	—	5.398
Ours	0.557*	0.209*	**87.677**	5.407

The **bold** text denotes the best result. And the second-best results are denoted with *.

Table 5.2: Time and memory consumption of image synthesis

	SD	SD w. LoRA	Adapter	ControlNet	Ours
Speed(UNet)	8.13it/s	7.70it/s	8.31it/s	5.51it/s	7.93it/s
Flops(UNet)	1.36TF	1.37TF	1.36TF	1.83TF	1.37TF
Parameters	4067MB	4080MB	4362MB	5445MB	4668MB

from other methods. CycleGAN achieves favorable results on these two criteria by introducing only minor changes, though it does not fully capture the style of the illustrative image. This will be discussed in more detail in the User Study section. CLIP-image is a criterion that evaluates the quality of the synthesized images; our method performs better than the others, indicating that it carries more of the illustrative style. For CLIP-aesthetic, the score from our method is higher than that of Adapter and CycleGAN but lower than those of ControlNet, LoRA, and SD with text. The reason for this is that this criterion is derived from feature maps based on the pretrained CLIP model on the LAION-5B dataset, which contains a larger proportion of real images. The scores are assigned based on these real images, which can lead to a shift in scoring. On the other hand, the scores obtained by our method are closer to that of the ground truth, indicating that the synthesized images from our method more closely resemble the ground truth compared to those from other methods. Table 5.2 denotes the details of consumption. Image synthesis tests were

conducted on a single RTX 3090 GPU with the resolution of the synthesized images set to 512×512 pixels. The model parameter sizes are calculated based on a float32 precision format. As shown in the table, the inference time and computational cost of our method is significantly shorter than that of ControlNet's, and slightly longer than those of the Adapter and the baseline model. However, the usage of the time is comparable. In terms of memory usage for parameters, our method requires slightly more memory than the basic Stable Diffusion and T2I-Adapter, but much less than ControlNet. This is because the style adaptation module in our method has far fewer parameters than the extra UNet Denoiser. In summary, our proposed method requires only a small amount of extra memory compared to Stable Diffusion and can generate high-quality images with an illustrative style on a home-use GPU.

5.4.3 Qualitative Comparisons

The generated results include CycleGAN [218] for GAN-based models, and for diffusion-based models, including pretrained Stable Diffusion (SD), Fine-tuned SD [148], SD with LoRA [52], ControlNet [200], and T2I-Adapter [125], along with results from our method for comparison. Pretrained Stable Diffusion (SD) has zero-shot capabilities but cannot perform style transfer independently; text prompts are adopted for its synthesis. Similarly, prompts are also adopted for Fine-tuned SD and SD with LoRA. Figure 5.3 illustrates the comprehensive qualitative comparison. Generally speaking, images generated by T2I-Adapter, ControlNet, and our method are able to capture the illustrative style, while CycleGAN and SD with LoRA struggle to alter the style of the source image. Since the pretrained SD learns the illustrative style from a universal dataset, it cannot accurately capture the specific illustrative style of a real designer. Specifically, all methods can preserve the appearance of the input image in each row. However, results from CycleGAN struggle to modify the style of the images, whereas the generated images capture the style of the real images and appear more realistic when compared with illustrative images. The images generated from fine-tuned Stable Diffusion, Stable Diffusion with text, and Stable Diffusion with LoRA are able to capture the semantic information from the input images. However, they lack the detailed nuances of the illustrative

Figure 5.3: Qualitative comparison between Uni-DlLoRA and other state-of-the-art approaches. From left to right, the displayed results correspond to CycleGAN, Stable Diffusion (SD), fine-tuned SD, SD with LoRA, T2I-Adapter, ControlNet, and our method, respectively. The text caption is utilized for content synthesis in SD, fine-tuned SD, and SD with LoRA, while the prompt 'illustrative style' is used for style guidance in both SD and fine-tuned SD. The figure is best viewed when zoomed in.

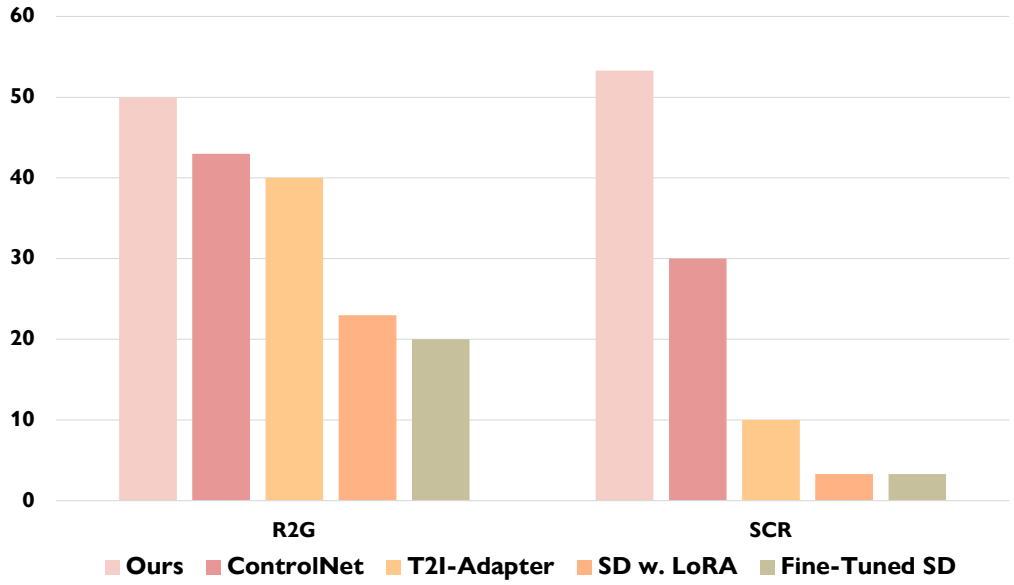


Figure 5.4: User study. The table on the left represents the G2R score, while the table on the right illustrates the SCR score.

style. Additionally, the generated images do not integrate harmoniously with the overall composition. For instance, see rows (b) and (c): the resulting images appear rigid and exhibit a discernible style conflict when compared with the ground truth. In terms of the generated results from the T2I-Adapter, ControlNet, and our own model, all are capable of conveying the illustrative style while maintaining the appearance of the runway model. However, the T2I-Adapter and ControlNet may fall short in replicating the intricacies of the clothing. For example, there is a slight color shift in the results from the T2I-Adapter evident in rows (b) and (d). Additionally, the clothing details exhibit variations in row (c). As for the images generated by ControlNet, while they effectively capture the style and general appearance, there is potential for improvement in clothing details, such as the red attire in row (c) and the gray clothing in row (a).

5.4.4 User Study

Since the evaluation of illustrations is often abstract and subject to many human perceptions, the opinions of 100 human participants will be used as the standard for assessing effectiveness. A user study was conducted to assess the abstract quality of the results from our method compared to those obtained by other methods.

Table 5.3: Qunatitative comparison between each component.

Methods	Metrics			
	$FID_{64} \downarrow$	$LPIPS \downarrow$	CLIP-image ↑	CLIP-aes ↑
Baseline (SD)	2.677	0.298	74.748	**5.598**
Uniadapter	0.814	0.214	83.789	5.291
Uni-SgLoRA	0.749*	0.213*	84.814*	5.356
Full Model	**0.557**	**0.209**	**87.677**	5.407*

The **bold** text denotes the best result and the second-best results are denoted with *.

Two approaches are adopted for this evaluation. The first employs G2R metrics, as mentioned in the research by Zhu et al. (2019) [223], which measures the percentage of generated images classified as ground truth (illustrative images). The second criterion involves the scores assigned to the highest-quality results by the participants. They are instructed to base their evaluations on the ability of each competing approach to produce accurate clothing and an illustrative style. This is quantified using another metric named *SCR*, defined as the percentage of images considered the best among all the models. Higher values in these three metrics indicate better performance. For the R2G metric, 35 real images and 35 generated images are randomly selected and shuffled. The first 10 of these are used for participant practice, while the remaining 60 images constitute the evaluation set. For the SCR metric, 30 generated images derived from 30 different real images are used for each method. Participants are then asked to choose one of the images as the best in quality. The comparative results of the study are illustrated in Figure 5.4, which clearly demonstrates that our methods surpass the others in terms of human perception: 50% of the results from our method are perceived as ground truth. Regarding the SCR metric, our *SCR* score is 53%, indicating that participants favored our approach more frequently than the competing methods.

5.4.5 Ablation study

An ablation study was conducted to evaluate the impact of each component within the proposed model in the StylishU-SR dataset. Table 5.3 illustrates the

Figure 5.5: Ablation results on the StylishU-SR. The images in this figure correspond to the ablation studies in Table 5.3.

impact of each component on the dataset. Baseline (SD) neither employs the uni-adapter module only a UNet-based noise prediction module with extra prompt "illustrative style". Although it can generate images with a precise appearance in Figure 5.5, its ability to retain the illustrative style and preserve the texture of the garments is limited. To effectively model the complex textures within the clothing, a learnable adaption module extracts image information and then sent to the UNet denoiser. When incorporating latent appearance features extracted by a pretrained adaptation module, this process is referred to as Uniadapter. Compared to the baseline, the Uniadapter reduces the FID_{64} score from 2.677 to 0.814, indicating a performance improvement. As shown in Figure 5.5, the results from Uniadapter capture more appearance and image information than the baseline model. To enhance the style translation, a style adaptation module is adopted during both training and sampling to capture the style features. From the table, it is clear to see that the style adaption module improves the results in all four criteria. The SgLoRA module not only improves the generation quality from the statistics but also in human perception shown in Figure 5.5. To further disentangle the style and content information of the source image, the Dual-LoRA module is adapted to align the output image content and style with the source image content and style. The last column in Figure 5.5 illustrates that our model can successfully catch the content and reconstruct the source image with good quality. In comparison with Uni-SgLoRA, our full model improves the FID_{64}, $LPIPS$, $CLIP\text{-}image$ and $CLIP{-}aes$ by a margin of 0.192, 0.004, 2.863 and 0.051, respectively.

5.4.6 Illustrative Style Interpolation

The proposed model is capable of modifying the final generated graphic's illustrative style by adjusting the sampling positions. The DDIM [165] sampling approach is utilized for the generation task. Specifically, the image generation task involves sampling a total of 50 times. To adjust the strength of the effect, linear interpolation is performed with values between 0 and 1. Based on this value, Gaussian noise of corresponding strength is added to the latent features extracted from the input image. Additionally, the introduction of Gaussian noise at various timesteps

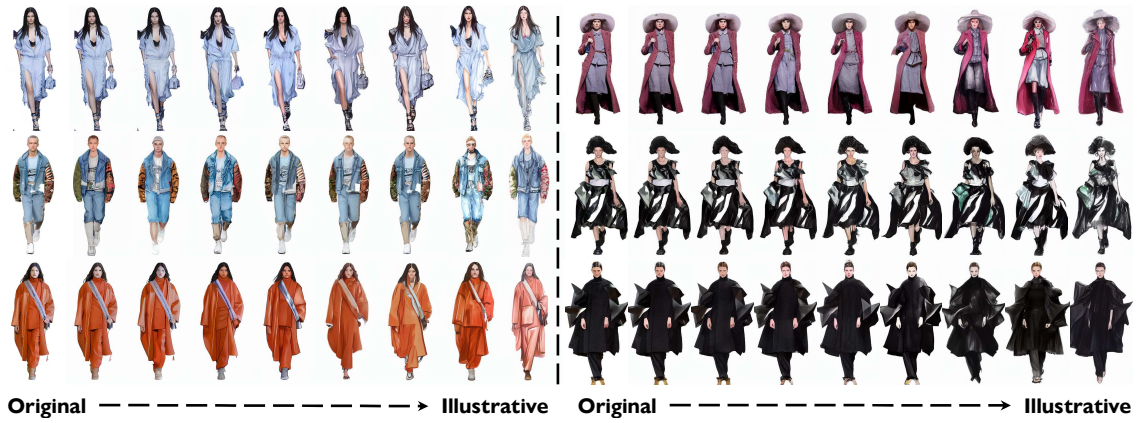


Figure 5.6: Style interpolation. Images are synthesized by combining a source image with varying style strengths ranging from 0 to 1. Generated images progressively carry the illustrative style. Both images inside the dataset and in the wild are evaluated.

is also based on the interpolated value. This allows the generated image to obtain more style information. Three samples from the test dataset and three real-world samples are selected to demonstrate the effectiveness of the illustrative style interpolation. As depicted in Figure 5.6, it is evident that the style of the images undergoes a gradual transformation from the left source image to the right source image. This gradual shift showcases the model's capability to provide a smooth transition in two different styles.

5.4.7 Generate Image in the Wild

Our model, which is fine-tuned based on a pretrained stable diffusion model, exhibits strong robustness and is also capable of performing illustrative style transfer on runway images outside of the dataset. The images in the figure showcase some successful instances of illustrative style transformation. As illustrated in the right half of the Figure 5.6. For the image that is out-of-dataset, the method not only generates images that carry illustrative style but also captures varying degrees of style based on the number of steps. This successful style evolution and the consistency observed in images from the test dataset and images in the wild underscore the robustness and strong adaptability of the proposed method.

Figure 5.7: Failure cases using the proposed method.

5.4.8 Limitations

Although our proposed method achieves solid results in most cases, it still fails in certain scenarios as shown in Figure 5.7. For instance, due to the images being formed by the overlay of noise, precise alignment remains an area in the complicated domain like fashion for improvement. As demonstrated in the figure, the generated illustrative images still exhibit noticeable differences from the original in aspects such as the texture of the clothing (the first six examples), and the shape of the garments (the rest six examples). The aforementioned examples also prove that the image transformation through this method entails a certain level of randomness and does not align as closely with the source image as might be desired. The sketch images may be insufficient to carry all the detail necessary, thus failing to constrain the final image synthesis adequately.

5.4.9 Future Work

Although the proposed method demonstrates promising results, its performance is still constrained by the limited size and diversity of the available dataset. The scarcity of high-quality, annotated data restricts the model's ability to generalize and capture the full range of fashion styles and details. In future work, expanding

the dataset with more diverse and representative samples will be crucial for further improving the quality and robustness of the generated images. Additionally, future research could explore the integration of advanced generative techniques such as diffusion models or flow-matching methods that do not require paired training samples. These approaches have shown great potential in other domains for learning complex data distributions without the need for explicit input-output pairs. Incorporating such methods may help overcome current data limitations and enable more flexible and effective style transfer in fashion image synthesis.

5.5 Chapter Summary

Leveraging the existing challenges within illustrative transformation, this chapter has created a new high-resolution real-to-illustration dataset. It also introduces a novel approach to address these challenges. The proposed model incorporates the concept of disentanglement, utilizing a shared image extractor and distinct style adaption modules to learn the content and style of images, and converts these into an illustrative style. This innovation contributes significantly to the fashion field. Nevertheless, the method has limitations, and the illustrative style transformation does not fully achieve alignment with the source images. In the future, complete content alignment is aimed to be achieved while better capturing texture information and further enhancing the style transformation.

Chapter 6

Capable Fashion Pose-Guided Image Transfer Modeling

Human pose transfer aims to synthesize referred human images with target pose and plays important roles in the intelligent fashion generation sysytem, bringing the substantial economic potential for E-commerce or virtual reality. In Chapter 3.5, a preliminary introduction to the proposed method is provided. This chapter presents a novel approach termed the Attentional Pixel-wise Deformation Network (APD-Net), which is designed for synthesizing human images based on guided poses and reference images. Specifically, attention-based spatial transformation modules and affine transformation modules are leveraged to generate accurate appearance and extract pixel-wise details in local regions to generate intermediate results. Additionally, a confidence map is introduced to refine spatial information during the final image synthesis. Domain alignment loss, cycle loss, perceptual and feature matching loss and contextual loss are applied to constrain the synthesized images while attention loss and fusion loss are benefit warp images generation. The results of the proposed method approach surpasses previously published state-of-the-art results on most evaluation metrics. At the same time, the proposed method effectively bridging the gap between Part B and Part C in the overall framework 3.1 in Chapter 3.

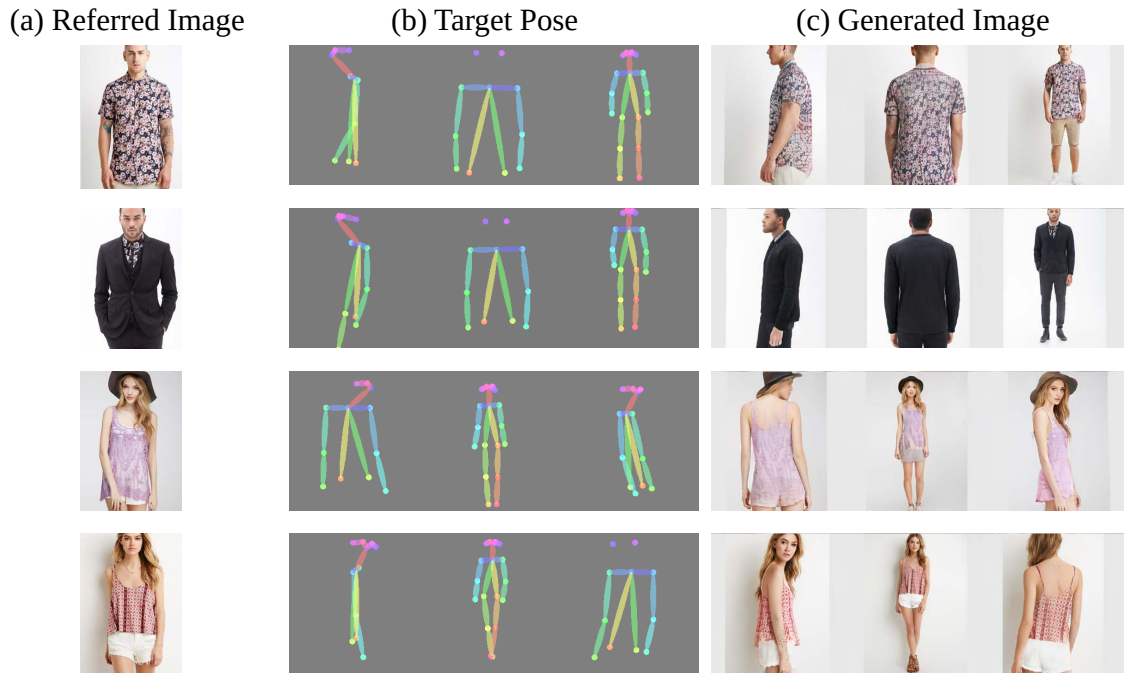


Figure 6.1: Pipeline of pose-guided human image transfer. Given a referred image (a) and target pose (b), APD-Net can generate human image with a pose transferred (c).

6.1 Introduction

Photo-realistic person image generation from pose and a referred image has become increasingly popular in recent years, with various applications in various industries such as e-commerce, virtual reality, and entertainment. However, this complex and challenging task requires advanced techniques to capture the 3-D structure of clothing and human appearance, synthesize the image with the target pose, and preserve the details of the patterns. The ability to achieve these objectives accurately and efficiently is essential for generating high-quality images that can be used for virtual try-on, pose transfer, and guided video generation. Despite the significant computational requirements and time constraints, researchers and developers continually work to improve these methods to meet the growing demand for these tools. The electronic commerce and virtual reality industries, in particular, are exploring the use of these algorithms to enhance the user experience and boost sales. As shown in Figure 6.1, the process involves a 2-D geometric transformation, which further adds to the complexity of the task. Therefore, developing efficient and

cost-effective solutions is critical for realizing the full potential of this technology.

In recent years, deep-learning-based methods leveraging the power of deep neural networks have shown great potential for image synthesis. Researchers have made significant progress in this area with notable contributions [40, 7, 73, 74]. However, synthesizing pose-guided images with the input image and target pose remains challenging, requiring the accurate transformation of human spatial information, the generation of clear human appearance, and the preservation of garment texture details. While several methods have made progress in this area, Convolutional Neural Networks (CNNs) face limitations in capturing long-term spatial information, as their convolutional operations in each layer focus on local patches [139]. Additionally, simple convolutional operations can negatively impact the details of the generated image during downsample and upsample operations.

Several methods have been proposed to improve spatial transformation using attention operations [198, 179, 183]. These methods calculate the value of feature patches using a weighted matrix to obtain the target feature patches, allowing each patch in the source features to communicate with the expected target features directly. This reduces local attention and focuses on global attention, making spatial transformation more effective and reducing local dependencies. However, attention-based methods face challenges when generating delicate details such as textures. The target patches must often focus on a very small local region, which is difficult for a simple attention matrix to guide the target patches' synthesis accurately. Furthermore, using a single attention module can make preserving prints such as logos and icons in the referred images challenging. Another effective spatial transformation network based on flow maps has been introduced to preserve image details, such as flownet [29, 60]. These methods estimate the motion of each pixel using flow-based operations, allowing the source image to be warped with predicted 2D coordinate offsets. The pixel-level features can be captured and transformed to a target position, resulting in realistic textures in the final generated image. However, obtaining an accurate flow map of the image can be challenging when there is long-term motion, as flow-based operations are only related to the local region of the features. Complex deformations can also lead to semantic failures.

In this chapter, a novel approach named Attentional Pixel-wise Deformation Network (APD-Net) is presented that combines the strengths of an attention matrix and flow-based operations. As obtaining stable flow-based operations during training is challenging, an affine transformation operation is applied as a substitute to capture the image's textures. A correspondence matrix with a confidence value are extracted to aid image synthesis. Furthermore, an affine transformation module is applied to extract details in local regions and fuse the deformation between attention and affine transformation operations to obtain the warp examplar. Finally, the final image is generated by combining the fused warp examplar with the confidence map. Our experimental results show that APD-Net achieves more photo-realistic human pose transfer with fewer parameters and faster speed than five state-of-the-art methods. All in all, the main contributions are summarized as follows:

- APD-Net, a simple yet effective framework for human pose transfer that combines attention and pixel-wise deformation operations is proposed. The model is easy to train and requires few parameters. APD-Net leverages the advantages of attention and affine transformation operations, which extract details in local regions and capture image textures.

- a novel confidence map is introduced in the attention operation to enhance the performance of the image synthesis module.

- The superiority of this method is demonstrated over state-of-the-art approaches through quantitative and qualitative evaluations.

6.2 Related Work

Recent advancements in deep neural networks have shown the capabilities of generative networks in synthesizing realistic and vivid images under user specifications, such as segmentation maps and edges. Pose-guided human image synthesis has become a popular task in this field. It aims to synthesize the target image based on the corresponding skeleton and referred image while preserving the appearance and details of the guide image [186, 72, 62, 218, 57, 213]. Previous works, such as

the two-stage network proposed by Ma *et al.* [116], use a coarse-to-fine approach for transferring a person's image to the target pose. Later, Ma *et al.* [117] achieved better performance by disentangling the pose and appearance of the referred image. However, these methods are computationally expensive.

To address this issue, Esser [30] combined Variational Auto-Encoders [25] with U-Net [149] to improve processing. However, 1-D embedding features lack the ability to capture appearance information, which results in a decline in the quality of the generated image. Additionally, skip connections in U-Net can lead to feature misalignment, hindering the performance of the generated image. To overcome these limitations, Zhu *et al.*[223] proposed a network comprising transfer blocks that connect regions of interest in the referred and target poses.

Meanwhile, CoConsnet [201], CoConsnetV2 [216] and NETD [144] proposed a novel network that generates an image based on the semantic warp image by referring to the semantically corresponding patches in the referred image. However, the patterns of the exemplar cannot be transferred. Other works, such as CtNet [192], ADGAN [121] and RSAGAN [105], disentangled semantic features at the feature level and achieved better performance. Very recently, several methods achieved good performance with delicate neural generators. For instance, CASD [217] applied Transformer-structure into the generator to obtain the final image synthesis. However, these methods used additional parsing images and attributes, limiting their practical applications.

To tackle the task of appearance and texture synthesis in specific regions, efficient spatial transformation modules such as methods [197, 110, 161, 180, 173], are proposed. Siarohin *et al.*[161] proposed a method that applies skip-connections to transform the textures spatially. At the same time, the whole deformation is decomposed to a set of sub-parts transformed by a certain affine transformation to alleviate the problem of spatial misalignment. Several methods like PATN [223] and Khatun et al. Khatun *et al.* [75] proposed a method that transform the spatial texture information with time sequence to preserve the details.However, it is difficult to synthesize the accurate shape with a single affine transformation module.

Flow-based methods are more flexible than affine transformation methods with

no limitation to transformation components. Zhou *et al.* [215] firstly estimated the target by warping the sources, and Han *et al.* [44] introduced a cascaded flow estimator to obtain flow fields without supervision in the process of synthesis. However, their warping process is under the level of pixels instead of that of features, which hinders content generation. Li *et al.* [92] proposed a method to generate flow field labels with additional 3D geometry. However, the 3D human model is necessary to acquire appearing flows where the computation cost is exceptionally high.Ren *et al.* [145] proposed introducing the features with local attention to obtain the final image, and Tang *et al.* [173] proposed the method with local flow fields to carry out the information of semantic correlations between each feature. Nevertheless, features in local fields are blurred when there is a large discrepancy between the source and target pose.

In contrast to the aforementioned approaches, a novel method is proposed that combines the strengths of attention-based and flow-based operations. Specifically, our method employs an attention-guided correspondence matrix and affine transformation module to replace the challenging flow module, enabling effective pose-guided human image synthesis. Improved performance over existing methods is demonstrated by leveraging attention-based and flow-based operations.

The goal of this study is to learn how to transform poses from the skeleton domain $\mathcal{A}$ to the real image domain $\mathcal{B}$, with the help of an input image $y_f \in B$. APD-Net, which learns cross-domain correspondences to provide better guidance for image translation and employs a flexible affine transformation to capture textures and local region deformations is proposed to achieve the goal. The generator combines the outputs from this estimator to obtain the final image synthesis using a SPADE resblock [133]. Specifically, the entire process is illustrated in Figure 3.6. It involves two main steps for generating pose-transferred images: 1) a Pose-guided Attention Estimator for pose estimation and 2) an Image Synthesis Module for image generation. Specifically, four elements—namely attention warp exemplar, fusion map, affine warp guidance, and confidence map in part a)—are synthesized, and attention pose-guided warp exemplars are multiplied with the confidence map in part b) to enhance image synthesis during the first stage in Chapter 6.2.1 Subsequently, cross-domain image synthesis module based on SPADE resblock [133] in

Chapter 6.2.2 is proposed, aiming to synthesize final results by leveraging high-level features of reference information and target pose. Finally, the training details are explained in Chapter 6.2.3.

6.2.1 Pose-guided Attention Estimator

Accurately transferring the texture and appearance of the referred image to the target image in pose-guided person image synthesis is always challenging. To reassemble the referred image according to the provided modifications, the correspondence is estimated between the referred image $\mathbf{I}_r$ and target pose $\mathbf{I}_{st}$ mapped in the same domain S. As shown in Figure 3.7, this correspondence is built using an attention estimator, which consists of three parts: (a) Attention Correspondence Module, (b) Affine Transformation Module, and (c) Feature Fusion Module. Final warp images is generated from the feature fusion module and applied for further process with the attention correspondence matrix obtained by attention correspondence module.

Attention Correspondence Module. The local and global features of both referred image and target pose are extracted by the pyramid network, which uses a pyramid-like structure to capture both fine-grained and coarse-grained information of the input information, and weighted by self-attention mechanisms [111]. They are then aligned in the same domain S, denoted as: $I_{rS} \in \mathbb{R}^{H\times W\times C}$ and $I_{tS} \in \mathbb{R}^{H\times W\times C}$ respectively. H and W are the spatial sizes of the feature maps extracted by pyramid networks and C means the channel-wise dimension. Let $\mathcal{F}_R$ and $\mathcal{F}_T$ be the extracting networks respectively, the formula can be written as:

$$I_{rS} = \mathcal{F}_R\left(I_r; \theta_{\mathcal{F},r\to S}\right) \tag{6.1}$$

$$I_{tS} = \mathcal{F}_T\left(I_{st}; \theta_{\mathcal{F},t\to S}\right) \tag{6.2}$$

where $\theta_{\mathcal{F}}$ denotes the learnable parameters. Then the correlation matrix $\mathcal{M} \in R^{HW\times HW}$ is acquired as:

$$\mathcal{M}^{i,j} = I_{rS_f}^{iT} I_{tS_f}^{j} \tag{6.3}$$

illustrating the correspondence between features extracted from referred images and target pose. Where $I_{rS_f} \in \mathbb{R}^{HW\times C}$ and $I_{tS_f} \in \mathbb{R}^{HW\times C}$ are the flatten vector of extracted feature maps obtained from the referred image and target pose, respectively.

Notations i and j denote the location of the i^{th} and the j^{th} pixel values in each feature maps, respectively. Attention warp examplar Awcould be obtained according to the matrix with coefficient α and softmax operation:

$$Aw^i = \sum_j \text{Softmax}_j(\alpha \mathcal{M}^{i,j}) \cdot I_r^j \tag{6.4}$$

where α is a parameter that is used to control the sharpness of the correlation matrix and is set to be 1 in this paper. $I_{\text{r}} \in R^{H \times W \times 3}$ denotes the referred image. In this module, the attention warp examplar can be reassembled by the referred image.

While it is difficult to correctly reassemble the attention warp image according to the referred image in every location, a confidence map is introduced based on the correlation matrix to assign the reliability, as shown in Figure 3.7. The confidence map is computed as:

$$Cmap = f_{conf}\left(\mathcal{M}\right) \tag{6.5}$$

where

$$f_{\text{conf}}\left(\mathcal{M}\right) = \text{sigmoid}\left(max_i\left(\mathcal{M}^i\right) - \overline{\mathcal{M}}_i\right) \tag{6.6}$$

and $\overline{\mathcal{M}}_i$ illustrates the pixel-wise mean of the correlation matrix. It calculates the pixel-wise confidence of the correlation matrix. Specifically, a lower weight should be given to the feature correspondence with lower reliability, which is then processed for the final image synthesis. On the other hand, a higher weight is assigned to the reliable feature correspondence, providing reliable semantic guidance.

Affine Transformation Module. OpenPose [10]is firstly used to extract 18 joints from the referred image and roughly divide the human body into 10 sub-parts rectangular regions, which include the head, torso, left/right upper/lower arms, and left/right upper/lower legs. For the left/right upper/lower arms and legs, two corresponding joints is used to construct the rectangle, with the major axis of the rectangle corresponding to the line between specific joints and the minor axis being orthogonal to r1 (as shown in Figure 3.7) with a smaller length.

For the head, 5 head joints as keypoints is applied, and the points of the left and right shoulder serve as the major axis of the rectangle. As the torso part is the largest region, the center of the left/right hips and neck is applied as the major

axis of the rectangle, with their minor axis being orthogonal to their major axis at a particular scale. Based on this binary mask maps $R_r^i \in \mathbb{R}^{H\times W\times C}$ and $R_t^i \in \mathbb{R}^{H\times W\times C}$ can be obtained where C denotes the numbers of rectangles indicating the full-body regions. Furthermore, Affine transformation $f_{affine}$is computed with Least Squares Error in keypoints to match the points in R_t^i and R_r^i. Affine transformation is a geometric transformation including translation, rotation, scaling and sharing. With this transformation, the pixel-level information is preserved and not lost. In some cases where there are occlusion and truncation in image borders, or the keypoint is missing when detected by OpenPose, the mask regions will be set in both $R_r^i \in \mathbb{R}^{H\times W\times C}$ and $R_t^i \in \mathbb{R}^{H\times W\times C}$ to zero (function affine transformation is not computed). The sub-part rectangle regions transform to achieve approximate global pose-dependent deformation:

$$A_t^i = f_{affine}\left(I_{\mathrm{r}} \odot R_r^i\right) \tag{6.7}$$

where the $\odot$ denotes the function of point-wise multiplication and A_t^i is the affine warp guidance.

Feature Fusion Module. The attention warp exemplar and affine warp guidance have complementary advantages: the former preserves the global information of the body figure, while the latter maintains the details of the human body. Therefore, a fusion map is generated$M_{fu} \in \mathbb{R}^{H\times W\times 1}$ using a fusion model:

$$M_{fu} = f_u\left(Aw, A_t\right) \tag{6.8}$$

where f_u is a generator consisting of several residual blocks and the last layer is the sigmoid function. The values in M_{fu} range from 0 to 1. By using the Feature Fusion Module, the attention pose-guided warp exemplar can be generated in both global features and details. The formula for generating the exemplar is as follows:

$$Paw = M_{fu} \odot Aw + (1 - M_{fu}) \odot A_t \tag{6.9}$$

where $\odot$ denotes the point-wise multiplication.

6.2.2 Image Synthesis

The image synthesis approach uses a generator that combines the target skeletons with the attention warp examplar. Firstly, the affine warp guidance and the warped image are calculated using the fusion map. Then, the SPADE resblock with the attention-based warp image is applied to render the target-pose skeleton. The final image is synthesized by the network using $I_t, Cmap$ and M_{fu} generated from the Pose-guided Attention Estimator. The formula for the image synthesis process can be expressed as follows:

$$\hat{I}_t = gen\left(I_{at}, Cmap, M_{fu}, Aw, A_t\right) \tag{6.10}$$

The equation illustrates the generation process that Ir,Cmap, Paw denote the target skeletons, confidence map obtained in Equation 6.5, and Paw, the examplar achieved in Equation 6.9. Mfu is the fusion map generated as Equation 6.8 while Aw and At are attention warp examplar and affine warp guidance, respectively. The pose-guided attention warp examplar is used as input to the final generator, which is based on the progressive image generator [71, 133]. The generator synthesizes the final image by incorporating both the reliability of the warp examplar from the confidence map and the guidance of the target skeleton.

$$\hat{I}_t = gen\left(Cmap \odot Paw, (1 - Cmap) \odot I_{st}\right) \tag{6.11}$$

Where $Cmap$ is the confidence map and Paw is the attention pose-guided wap examplar. Note that in our implementation, the target pose is concatenated to obtain better synthesis, denoted as I_{st}.

6.2.3 Loss Function

Our proposed model is fully differentiable end-to-end, which allows us to optimize it using back-propagation and multiple loss functions simultaneously. These loss functions include both losses in the attention warp exemplar and losses in the final output.

Losses in Warp Examplar. Although the proposed model is an end-to-end network, the intermediate output in the network is crucial as it provides guidance for

the final image synthesis. Therefore, several loss functions are set to ensure meaningful deformations. Firstly, the representation I_{rS} and its counterpart I_{tS} should lie in the same domain and have the same spatial distribution. Thus, the $L1$ loss is set for domain alignment, which can be represented as:

$$\mathcal{L}_{\text{align}} = \left\| \mathcal{F}_R \left(I_r; \theta_{\mathcal{F}, r \to S} \right) - \mathcal{F}_T \left(I_{st}; \theta_{\mathcal{F}, t \to S} \right) \right\|_1 \tag{6.12}$$

The $L1$ loss is calculated between the warp exemplar and the target image using the following formula:

$$\mathcal{L}_{attn} = \left\| Aw - I_t^{\downarrow} \right\|_1 \tag{6.13}$$

where $I_t^{\downarrow}$ indicates the downsampled target image. In order to improve the quality of the warp exemplar and minimize reconstruction errors, it is important to constrain the attention correlation matrix and promote a more meaningful correlation matrix.

Taking inspiration from the concept of cycle consistency in [218], it is expected that the learned correlation matrix will be capable of transferring the warp exemplar to the referred image, provided that the correlation matrix can establish a correspondence between the target and referred images in the domain S.

$$\mathcal{L}_{cycle} = \left\| Awr - I_r^{\downarrow} \right\|_1 \tag{6.14}$$

where Awr refers to the warp exemplar of the referred image and $I_r^{\downarrow}$ denotes the downsampled referred image. The warp exemplar can be obtained using the following formula:

$$Awr^i = \sum_i \text{softmax}_i(\alpha \mathcal{M}(i,j)) \cdot Aw^j \tag{6.15}$$

where Aw^j represents the jth pixel value of Aw and Awr^i illustrates the ith pixel value of Awr.

To obtain a set of affine transformations, the six parameters is calculated$\mathbf{k} \in \mathbb{R}^{1\times 6}$ to constrain the loss using least square error:

$$\mathcal{L}_{affine} = \min_{\mathbf{k}} \sum_{\mathbf{p}_i \in I_r, \mathbf{q}_i \in I_t} \left\| \mathbf{q}_i - f_{affine} \left(\mathbf{p}_i; \mathbf{k} \right) \right\|_2^2 \tag{6.16}$$

where $\mathbf{p}_i$ and $\mathbf{q}_i$ represent the sub-parts rectangle regions of the target and referred images, respectively. It is worth noting that in certain extreme cases, such as those previously discussed, the computation of f_{affine} may not be feasible. Additionally,

it should be noted that the affine transformation calculation can be performed prior to training the deep neural networks.

Losses in Real Image Generation. The Pose-guided Attention estimator is capable of capturing the correspondence between the input and target images. Consequently, our model generates the final image $\hat{I}_t$ with the guidance of the warp exemplar. To ensure that the generated image $\hat{I}_t$ closely resembles the ground-truth image I_t, the perceptual loss[68] is introduced. This loss function helps to minimize the semantic discrepancy between the two images and is defined as follows:

$$\mathcal{L}_{perc} = \left\| \phi_m \left(\hat{I}_t \right) - \phi_m \left(I_t \right) \right\|_1 \tag{6.17}$$

where function ϕ_m denotes the semantic information extracted from the input image using high-level layers of the VGG16 network. In order to improve the visual quality of the generated image, a feature matching loss is introduced to constrain the generator. This loss function helps to align the statistics of feature maps between the generated image and the ground-truth image and encourages the generator to produce images with realistic texture and style.

$$\mathcal{L}_{fm} = \sum_m \left\| \phi_m \left(\hat{I}_t \right) - \phi_m \left(I_t \right) \right\|_2 \tag{6.18}$$

VGG network $\mathcal{L}_{fm}$ is used to extract feature maps from each activation layer. In addition, the Contextual loss [120] is applied to consider the context of the entire image and the semantic similarity in different regions.

$$\mathcal{L}_{CX} = \sum_m u_m \left[-\log \left(CX \left(\phi_m \left(\hat{I}_t \right), \phi_m \left(I_t \right) \right) \right) \right]. \tag{6.19}$$

where u_m is used to control the importance of the loss in different layers, while ϕ_m denotes the feature maps extracted from each activation layer of the VGG network for both the generated and target images. In addition, a differentiable discriminator is applied to distinguish the output between ground-truth images and images synthesized by the networks. The adversarial loss is defined as follows:

$$\begin{aligned} \mathcal{L}_{adv}^{\mathcal{D}} &= -\mathbb{E}\left[h\left(\mathcal{D}\left(I_t \right) \right) \right] - \mathbb{E}\left[h\left(-\mathcal{D}\left(\mathcal{G}\left(I_r, I_{st} \right) \right) \right) \right] \\ \mathcal{L}_{adv}^{\mathcal{G}} &= -\mathbb{E}\left[\mathcal{D}\left(\mathcal{G}\left(I_r, I_{st} \right) \right) \right] \end{aligned} \tag{6.20}$$

where $h(\cdot)$ denotes the hinge loss to regularize the discriminator.

Total Loss. The full loss function is denoted as:

$$\begin{aligned}\mathcal{L} = \min_{\mathcal{F},\mathcal{G}} \max_{\mathcal{D}} \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{attn}} + \lambda_3 \mathcal{L}_{\text{cycle}} + \lambda_4 \mathcal{L}_{\text{perc}} \\ + \lambda_5 \left(\mathcal{L}_{fm} + \mathcal{L}_{CX}\right) + \lambda_6 \left(\mathcal{L}^{\mathcal{D}}_{\text{adv}} + \mathcal{L}^{\mathcal{G}}_{adv}\right)\end{aligned} \tag{6.21}$$

where λ denotes the importance of each loss.

6.3 Experiment

In this section, our experimental results is presented in detail. Specifically, the datasets and evaluation metrics is firstly introduced in Chapter 6.3.1, followed by a comparison with state-of-the-art methods in Chapter 6.3.2. The implementation details of our approach is described in Chapter 6.3.3. In Chapter 6.3.4, the superiority of our proposed APD-Net is demonstrated over the state-of-the-art methods. Additionally, an ablation study is conducted to investigate the contribution of different components of our approach in Chapter 6.3.5. Finally, the visualizations of the image synthesis process is also provided in Chapter 6.3.5.3.

6.3.1 Datasets and Metrics

Dataset. Person re-identification dataset Market-1501 [210] and the In-shop Clothes Retrieval Benchmark DeepFashion [112] are commonly adopted for evaluating the performance of human pose transfer. Images in the Market-1501 dataset are in-the-wild, with a resolution of 128×64 and significant variations in pose and background. Comparatively, images in DeepFashion are collected from fashion shopping websites with a higher resolution (256×176) with a clean background. Zhu et al. [223] employed the Human Pose Estimator [10] as a pose joints detector and gathered 263,632 training pairs and 12,000 testing pairs for Market-1501 and 101,966 training pairs and 8,570 testing pairs for DeepFashion. In both datasets, the person identities in the testing differ from those in the training set. The same training and testing pairs as PATN [223] is used for both datasets to ensure fair comparison results.

Metrics. A total of nine quantitative evaluations is used following the general

practice of CoCosNet[201, 216]. Firstly, the distance between the Gaussian fitted feature distributions of the generated images and real images using Fréchet Inception Distance (FID) [48] are measured. FID is a metric used to evaluate the similarity between generated images and real images in terms of their statistical distribution of image features and quality. The formula for FID is given by:

$$FID = ||\mu_1 - \mu_2||_2^2 + Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}) \tag{6.22}$$

where μ_1 and μ_2 are the mean feature vectors of real and generated images, respectively, and Σ_1 and Σ_2 are their covariance matrices. Tr denotes the trace of a matrix. Furthermore, Sliced Wasserstein Distance (SWD) [71] is employed to measure the Wasserstein distance between real and synthesized images on a set of random directions. The formula is:

$$SWD(\hat{I}_t, I_t) = \frac{1}{D}\sum_{i=1}^{D} W_2(\hat{I}_t^i, I_t^i) \tag{6.23}$$

where $\hat{I}t$ and It are two sets of images, D is the number of slicing directions, and $\hat{I}^it$ and I^it are the projections along the i-th direction. For both evaluation criteria, lower scores indicate better quality of generated images.

Furthermore, Structural SIMilarity (SSIM) [185] is used to measure the similarity between synthesized and ground-truth images, Learned Perceptual Image Patch Similarity (LPIPS) [203] is used to evaluate the perceptual similarity between two images, and Peak Signal to Noise Ratio (PSNR) is used to measure the difference between the synthesized and ground-truth images in pixel level, respectively.

Finally, high-level semantic features are introduced to evaluate the quality of generated images from the perspective of semantic consistency and texture relevance for the DeepFashion dataset, which has clear images with higher resolution. Specifically, high-level features obtained after $relu3_2$ and $relu4_2$ of a VGG16 model [164] are used to measure the Semantic Consistency. The evaluated distance between generated and referred images and authentic images with low-level features like $relu1_2$ and $relu2_2$ that capture the color and texture information illustrate the Style Relevance. The paired real-ref dataset is used to assess the quality of synthesized images for obtaining the Style relevance between generated and real images.

6.3.2 Comparison Baselines

Our proposed method is compared with several state-of-the-art baselines designed for human pose transfer, including Pose-Attentional Transfer Network (denoted as PATN) [223], Cross-domain Correspondence Network (denoted as CoCosNet) [201], Cross-domain Correspondence Network v2 (denoted as CoCosNetv2) [216], Decoupled GAN (denoted as PISE) [199], and Dual-task Pose Transformer Network (denoted as DPTN) [202].

The weights released by the corresponding authors is used for image generation. Note that the weights of DPTN and PATN were obtained with the dataset in which the image size was 256×176. To achieve the result with the same resolution, the result is obtained with pretrained weight and manually added the background for DPTN, and adjusted the size of input images for PATN. For the Market-1501 dataset, all models were trained with an image size of 128×64 to test their performance.

6.3.3 Implementation Details

Experiment is conducted using images with a resolution of 256×256. The correlation matrix, with a resolution of 64×64, was obtained from feature maps, and the attention warp exemplar was on the same scale. These were used to generate the final output. The model was trained on a single GPU (NVIDIA 3090) for 100 epochs, with a batch size of 4.

Model Hyperparameter. The weight parameters are set for each loss function as follows: $\mathcal{L}\text{align} : \lambda_1 = 10$, $\mathcal{L}\text{attn} : \lambda_2 = 10$, $\mathcal{L}\text{cycle} : \lambda_3 = 1$, $\mathcal{L}\text{perc} : \lambda_4 = 0.001$, $\mathcal{L}\text{fm} + \mathcal{L}\text{CX} : \lambda_5 = 1$, and $\mathcal{L}\text{adv} : \lambda_6 = 10$. There are six hyperparameters in the formula, all of which use a default weight of 1. Accordingly, values is tested at different magnitudes—such as 0.1, 1, 10, and 100—to measure their impact on the FID metric, ultimately determining the current set of hyperparameters. In practice, it is observed during experiments that, with the exception of the affine loss used in the early stages, the attention loss and cycle loss during training had relatively minor effects on the quality of the generated results. The implementation is based

Table 6.1: Comparison with state-of-the-art on DeepFashion.

	PATN	CoCosnet	CoCosnetv2	PISE	DPTN	Ours
$FID \downarrow$	27.923	27.212	24.345	13.571	**12.117**	24.711
$SSIM \uparrow$	0.772	0.775	0.773	0.768	0.778	**0.780**
$LPIPS \downarrow$	0.254	0.221	0.201	0.208	0.199	**0.197**
$PSNR \uparrow$	17.712	18.891	19.101	18.521	19.149	**19.823**
$SWD \downarrow$	17.713	14.892	12.523	11.091	14.152	**9.551**
$SC \uparrow$	0.953	0.922	0.944	0.948	0.961	**0.963**
$Color(ref) \uparrow$	0.886	0.896	**0.931**	0.874	0.908	0.921
$Texture(ref) \uparrow$	0.923	0.925	0.928	0.894	0.921	**0.930**
$Texture(real) \uparrow$	0.948	0.923	0.948	0.913	0.952	**0.958**

on the Pytorch framework. Adaptive Moment Estimation (Adam) [78] optimization algorithm, which has been shown to outperform other algorithms in deep learning, is used to reach the global minima with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate for both the generator and discriminator was set to 1×10^{-4}. Instance normalization [177] was used as the normalization method in both the generator and the discriminators. Leaky ReLU [118] was applied after all normalization layer in the discriminators, with a negative slope coefficient of 0.2.

6.3.4 Benchmark Results

6.3.4.1 Quantitative and qualitative comparison

Quantitative Results. To ensure the reliability of our results, the evaluation of pre-trained models from previous keypoint-based methods [223, 202] is conducted on our testing set.Results for PISE [199]is unable to obtained on the Market-1501 dataset due to additional human parsing label requirements. The quantitative comparisons of our proposed method with previous works are presented in Table 6.1 and Table 6.2. The results suggest that our proposed method generally outperforms [223, 201, 216, 199, 202] on most metrics, with some steady numeric improve-

Table 6.2: Comparison with state-of-the-art on Market-1501.

	PATN	CoCosnet	CoCosnetv2	PISE	DPTN	Ours
$FID \downarrow$	22.681	27.597	19.635	-	**18.995**	19.601
$SSIM \uparrow$	0.282	0.223	0.278	-	0.285	**0.303**
$LPIPS \downarrow$	0.319	0.330	0.285	-	0.271	**0.267**
$PSNR \uparrow$	14.262	13.337	13.385	-	14.521	**15.753**
$SWD \downarrow$	24.532	19.863	17.984	-	18.352	**16.971**
$SC \uparrow$	0.565	0.628	0.694	-	0.703	**0.729**
$Color(ref) \uparrow$	0.720	0.715	0.753	-	**0.774**	0.767
$Texture(ref) \uparrow$	0.606	0.705	0.732	-	0.684	**0.743**
$Texture(real) \uparrow$	0.668	0.723	0.744	-	0.757	**0.762**

ments observed for both datasets. Specifically, for the FID metric, DPTN achieved the lowest FID score, meaning it performed best in terms of quality and diversity of generated images. However, it is worth noting that although our APD-Net is slightly inferior on the FID metric, it achieved higher rankings on multiple metrics including SSIM, LPIPS, PSNR, SWD, SC, Texture(ref), and Texture(real). These findings suggest that our method may have potential advantages in computer vision applications. However, some images generated by our APD-Net had blurry backgrounds, making it difficult to evaluate them accurately using the FID score [48]and Human Pose Estimator[10]. Therefore, it is important to consider these limitations when interpreting our results. Nonetheless, our study provides insights into the performance of our proposed method compared to state-of-the-art methods.

Qualitative Results. Figure 6.2 presents a qualitative comparison of the proposed APD-Net with other state-of-the-art methods on the DeepFashion [112] dataset. The results demonstrate that APD-Net generates visually appealing images with natural poses and detailed appearances while preserving the semantic information of the input images. Compared to PATN [223], APD-Net produces more detailed appearances with accurate textures, especially in complex patterns. DPTN [202] also produces competitive results with realistic textures, but it struggles to generate com-

Figure 6.2: Qualitative comparisons on DeepFashion dataset.

plex textures in local regions due to the lack of spatial transformations in low-level feature maps. CoCosnet [201] and CoCosnetv2 [216] generate realistic images with complex patterns, but they may suffer from artifacts when dealing with patterns like logos and words. In contrast, APD-Net combines the affine-part with attention warp examplar to generate accurate body figures and informative textures, which avoids the limitations of other methods. The proposed method explicitly aligns the condition poses and target poses using the aligned pose-attentional mechanism, which is critical for maintaining appearance consistency under large pose variance. Additionally, the network combines the warp examplar and affine-part features in an efficient way, as demonstrated in the inter-process visualization. Overall, Figure 6.2 demonstrates that APD-Net is capable of generating high-quality images that preserve the semantic information of the input images and maintain appearance consistency under large pose variance.

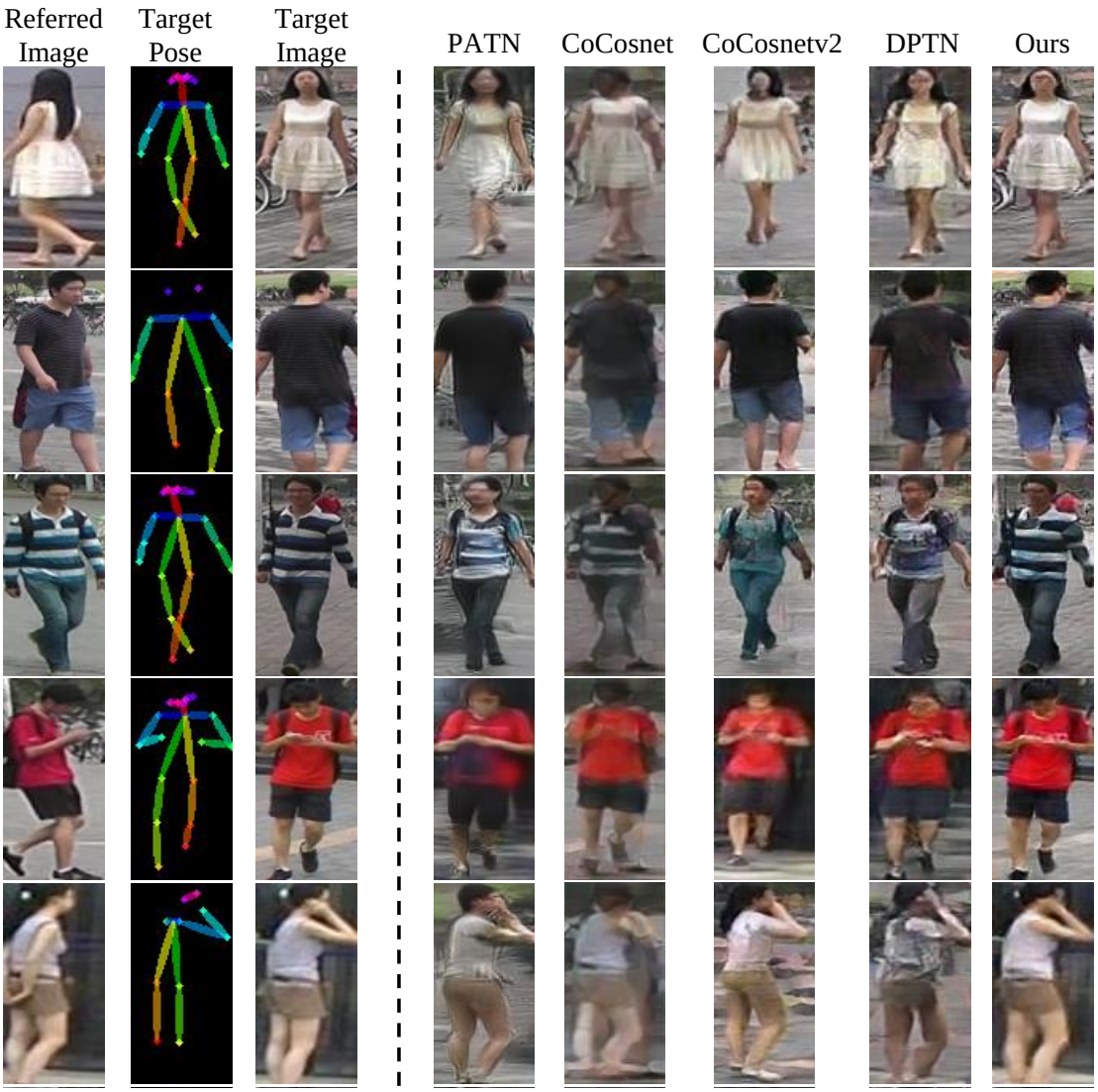


Figure 6.3: Qualitative comparisons on Market-1501 Dataset.

The performance of our method is also evaluated on Market-1501 [210], a dataset with poor image quality. Some examples are shown in Figure 6.3. Similar to the phenomenon observed in the DeepFashion dataset, APD-Net outperforms other methods in terms of body appearance and texture details. Specifically, the images generated from our method retain shape consistency, while the results of CoCosnet[201] and PATN[223] lose some details and appear blurry. Moreover, our generated images preserve clothing pattern details. For the unseen regions, images synthesized by our model generate plausible results, while other methods produce unsatisfactory images that are blurry and/or have incorrect appearances.

Table 6.3: Comparison of model size and testing speed on DeepFashion dataset and Market-1501 dataset. "M" denotes millions and "fps" denotes Frames Per Second.

	PATN	CoCosnet	CoCosnetv2	PISE	DPTN	Ours
Params ↓	41.36 M	143.6 M	46.4 M	64.01 M	**9.79 M**	49.1 M
SpeedF ↑	86.85 fps	43.47 fps	16.13 fps	41.02 fps	**71.43 fps**	70.79 fps

Table 6.4: User study (%). ***R2G*** means the percentage of real images rated as generated w.r.t. all real images. ***G2R*** means the percentage of generated images rated as real w.r.t. all generated images. The results of other methods are drawn from their papers.

Dataset	Measures	PATN	CoCosnet	CoCosnetv2	PISE	DPTN	Ours
DeepFashion	*R2G* ↑	19.14	19.17	20.04	20.07	21.98	**22.04**
	G2R ↑	31.78	31.82	33.21	33.23	35.02	**35.98**
Market-1501	*R2G* ↑	32.23	31.07	33.25	-	34.21	**36.45**
	G2R ↑	63.47	61.26	64.21	-	65.12	**70.22**

6.3.4.2 Model and computation complexity comparison

Table 6.3 presents a comparison of the model and computation complexity between our methods and previous approaches. These methods is tested using a single NVIDIA 3090 graphics card on the same workstation. To compute the speed, only GPU time is took into account when generating all the testing pairs of DeepFashion. Notably, the APD-Net enjoys better performance on image generation while having a similar level compared with previous methods in terms of the number of parameters and the computation complexity, owing to the simple and neat structure of the building blocks of our network.

6.3.4.3 User study

Human perception is better suited for assessing the authenticity of generated images. To evaluate the objective realism of our generated images, 100 volunteers

	Baseline	Attn-model	Affine module	Full
$FID \downarrow$	29.6	27.6	27.4	**24.711**
$SSIM \uparrow$	0.707	0.735	0.742	**0.780**
$LPIPS \downarrow$	0.242	0.210	0.205	**0.197**
$PSNR \uparrow$	16.724	17.986	17.385	**19.823**
$SWD \downarrow$	11.87	12.50	9.72	**9.551**
$SC \uparrow$	0.922	0.944	0.961	**0.963**
$Color(ref) \uparrow$	0.896	**0.922**	0.908	0.921
$Texture(ref) \uparrow$	0.925	0.928	0.921	**0.930**
$Texture(real) \uparrow$	0.923	0.948	0.952	**0.958**

Table 6.5: Quantitative results of ablation study on DeepFashion.

were enliseted to assess each image's authenticity (real or fake) within a second. Following the protocol outlined in [223], shuffled 55 real and 55 generated images are randomly selected , using the first ten for practice and the remaining 100 for evaluation. Two evaluation measures were adopted: ***R2G*** which means the percentage of real images rated as generated with respect to all real images, and ***G2R*** which means the percentage of generated images rated as real with respect to all generated images. As shown in Table 6.4, our APD-Net model demonstrated significant performance improvements over state-of-the-art methods across all metrics, confirming that our generated images are more natural, realistic, and sharp. Notably, for DeepFashion, our APD-Net model achieved the highest scores in both the *R2G* and *G2R* measures, with scores of 22.04% and 35.98%, respectively. Additionally, our method outperforms in handling condition images of poor quality, with 36.45% and 70.22% of APD-Net-generated images deemed real by volunteers, as reflected in the *R2G* and *G2R* measures in Table 6.4.

6.3.5 Ablation Study

Our APD-Net network incorporates two essential design characteristics. The first leverages the benefits of attention and affine transformation operations, while

the second introduces a confidence map in the attention operation to enhance the image synthesis module's performance. As demonstrated in the previous section, our results clearly highlight the advantages of APD-Net. Therefore, our ablation experiments focus on verifying the effectiveness of these two critical design features by reporting both the quantitative and qualitative results with and without these modules, respectively.

6.3.5.1 Methods with different modules

Baseline. To prove the effectiveness of the modules, the model is trained with the vanilla correspondence matrix and used the generator as the baseline model. The same set of loss functions was used during training.

Attention Correspondence Matrix. To evaluate the performance gain, the model is trained with the attention correspondence matrix. The other module remained the same as the baseline.

Affine Information. To evaluate the effectiveness of sub-part deformation, the affine information is added to the model. The attention correspondence matrix was removed, as in the attention model.

Full Model. The full model, represents the APD-Net, which is composed of both the attention correlation matrix and the affine transformation module.

6.3.5.2 Results and analysis

Quantitative Results. Table 6.5 and Table 6.6 present the quantitative results of the methods used in the ablation study. The baseline model achieved the worst score in the quantitative comparison because it struggles to obtain spatial information from the input image, resulting in a lack of details in the final image generation. In contrast, the attention model and affine transformation module achieved good results, with a steady gain compared to the baseline. These results demonstrate that spatial information improves the generator's ability to synthesize the final image. Additionally, the affine transformation module outperformed the attention model in each score, indicating that it provides texture details that benefit the final image generation. When combining the attention correspondence matrix and affine trans-

	Baseline	Attn-model	Affine module	Full
$FID \downarrow$	28.472	25.968	24.673	**19.601**
$SSIM \uparrow$	0.231	0.243	0.278	**0.303**
$LPIPS \downarrow$	0.342	0.330	0.285	**0.267**
$PSNR \uparrow$	13.724	14.986	15.385	**15.753**
$SWD \downarrow$	21.738	19.832	18.473	**16.971**
$SC \uparrow$	0.682	0.694	0.711	**0.729**
$Color(ref) \uparrow$	0.696	**0.771**	0.713	0.767
$Texture(ref) \uparrow$	0.697	0.728	0.721	**0.743**
$Texture(real) \uparrow$	0.681	0.703	0.698	**0.762**

Table 6.6: Quantitative results of ablation study on Market-1501.

formation module, the full model achieved significant gains, demonstrating that the combination improves image quality and texture details significantly.

Qualitative Results. Figure 6.4 and Figure 6.5 illustrates the qualitative comparisons in the ablation models. From Figure 6.4 The baseline model was able to preserve the body shape but lacked the ability to handle textures and details during image synthesis. With the attention correspondence matrix and affine transformation module, both the attention model and affine model preserved the details and body shapes when deforming the referred image to the target pose, although different types of artifacts were present in the final results. The model based on the attention correspondence matrix was able to extract long-term deformation and synthesize results with accurate body structure, but the dense connections were unable to transfer patterns like logos. As a result, the complex texture-like patterns in the second and third row of the figure were not appropriately generated. In contrast, the model with only the affine transformation module was able to generate images with complex textures and special patterns by extracting whole patterns from the target pose. However, the affine sub-parts were unable to estimate the real pose deformation with desired semantic information, resulting in blur in local parts. Our full model takes advantage of both the attention correspondence matrix and affine

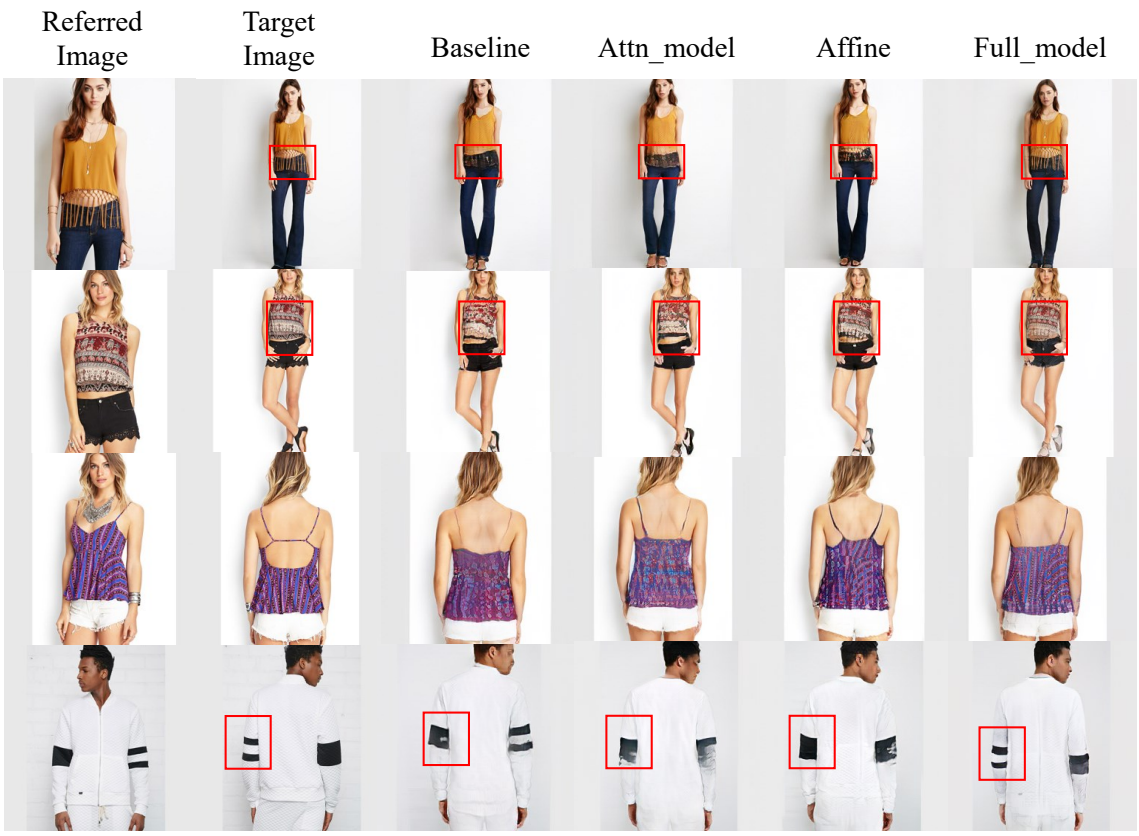


Figure 6.4: Qualitative results of ablation study on DeepFashion dataset. The red rectangles in it illustrate the discrepancy of generated results.

transformation module to synthesize results in the target pose with complex textures and accurate structures.

6.3.5.3 Visualization of the process

Intermediate Results. To demonstrate how each module works in our framework, the fusion map, and the impact of the extracted new features on the generated results, the intermediate results is presented in Figure 6.6. From the figure, the affine warp guidance selects the complex texture, while the attention warp exemplar chooses the body structure. The red rectangles in the fusion maps indicate the weight between the warp exemplar and affine warp guidance. In contrast, the yellow rectangles represent the features the fusion map selected. Because of the low resolution and complex background in the Market-1501 [210] dataset, the attention warp image can only identify a vague body shape and the main color. However, the affine warp guidance can capture more texture details and assist in generating

Figure 6.5: Qualitative results of ablation study on Market-1501 dataset.

a more accurate warp exemplar. Our method leverages the details from the affine warp guidance and the body appearance from the Pose-guided Attention Estimator, enabling the synthesis of images with realistic structures and patterns. Figure 6.7 shows additional results of the generated attention warp images. It is evident that the attention warp image can effectively capture the shape of the target pose, indicating that it preserves the appearance of the referred image and guides the final generation. Even in datasets with low image quality, the attention warp images can still recognize characters' body appearances and primary colors.

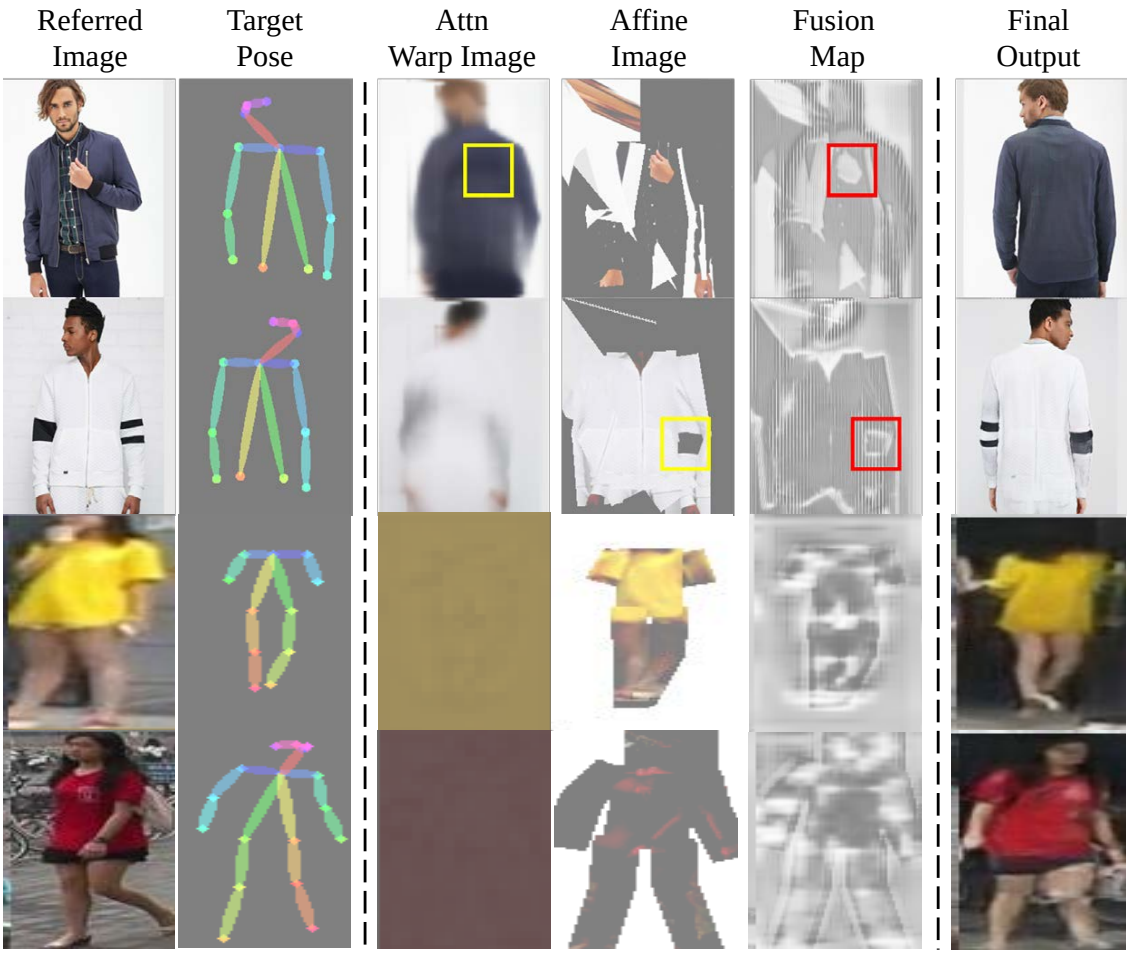


Figure 6.6: The detail process of image generation. Attention warp image and affine warp guidance are combined with the fusion map to obtain final image synthesis. The above two rows are results of the DeepFashion dataset and the below two rows are results of the Market-1501 dataset.

6.4 Chapter Summary

In this chapter, a novel network for pose-guided image synthesis that utilizes an attention correspondence matrix and an affine-based operation is firstly proposed. Through empirical analysis, the advantages and drawbacks of these two modules for spatial transformation and detail preservation are evaluated. Our ablation study shows that the model, which combines both the attention correspondence matrix and the affine transformation module, performs better in generating accurate results from the target pose while preserving the details of the texture.

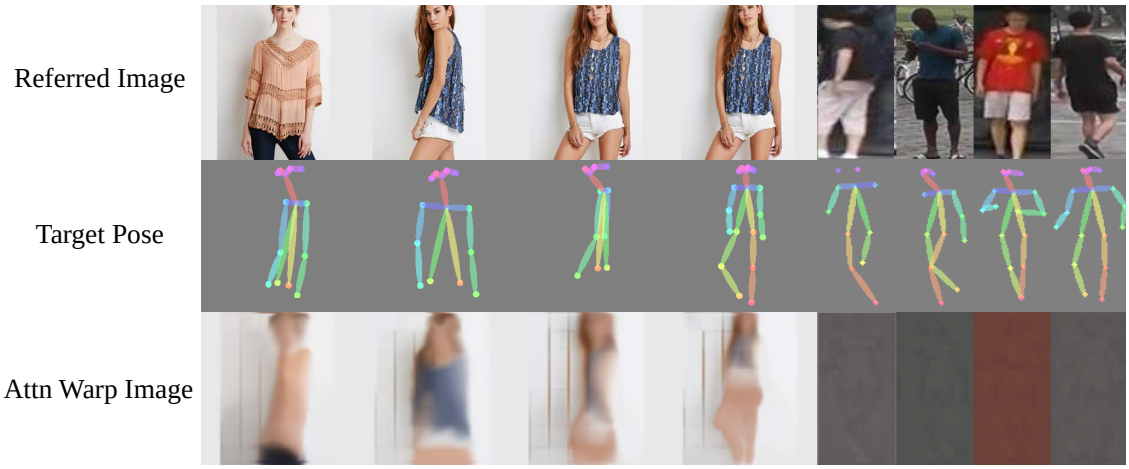


Figure 6.7: Visualization of the attention warp images. The left four columns are results of the DeepFashion dataset and the right four columns are results of the Market-1501 dataset.

6.5 Ethical Considerations and Responsible Use

While the pose transfer technology presented in this chapter demonstrates significant technical innovation in enabling pose transformation between different individuals, we acknowledge the important ethical implications that accompany such capabilities. The ability to synthesize realistic human poses raises legitimate concerns regarding potential misuse, particularly in the creation of non-consensual synthetic imagery.

Our research strictly adheres to established ethical guidelines for computer vision research involving human subjects. All datasets used in this work consist of publicly available images or images collected with explicit informed consent from participants. It is emphasized that any practical deployment of this technology should implement robust consent mechanisms and user authentication protocols. The training data employed in our experiments has been sourced from authorized datasets commonly used in academic research, including Market1501, DeepFashion. We have ensured compliance with the respective data usage licenses and terms of service. We advocate for the responsible development and deployment of pose transfer technologies. This includes: (1) implementing technical safeguards to prevent malicious use, (2) establishing clear usage policies and guidelines, (3) promoting transparency in synthetic content generation, and (4) supporting the development

of detection methods for synthetic media.

This research is conducted with the intention of advancing scientific understanding in computer vision and contributing to beneficial applications such as virtual try-on systems, animation, and accessibility tools for individuals with mobility limitations.

Chapter 7

Conclusions and Suggestions for Future Research

This chapter begins by presenting the conclusions of this thesis, summarizing the main findings and contributions of the research. Following the conclusions, the limitations of the study are discussed. Finally, the chapter outlines prospects for future work.

7.1 Conclusions

This thesis has made a series of advancements in deep learning-based intelligent fashion image generation, addressing some of the complex issues in this area. Specifically:

Chapter 4 focuses on how to quickly and automatically generate fashion hand-drawing sketch images. It includes fashion landmark detection, fashion clothing segmentation, and ultimately, automatic mapping to efficiently and quickly generate a complete outfit. The effective of the automatic fashion sketch generation reveals the follwing task: illustartive image transfer in Chapter 4 and Chapter 5 and human image pose transfer in Chapter 6.

Chapter 4 presents a novel network that improve the performance of image style transferring in informative and uninformative image dataset. The generated

sketch image can be transferred to images with target style, contributing to the development of intelligent fashion image generation systems.

Chapter 5 Leverages the existing challenges within illustrative transformation and introduce a novel network that address these challenges. The proposed model incorporates the concept of disentanglement, utilizing a shared image extractor and distinct style adaption modules to learn the content and style of images, and converts these into an illustrative style. This chapter additionally create a high-resolution real-to-illustration dataset for the future research.

Chapter 6 addresses the task of pose-guided human image transfer through a novel network that utilizes an attention correspondence matrix and an affine-based operation. Simultaneously, given that there are still many challenging objectives within this task, a comprehensive study that includes the problem definition and methodologies and the application based on these task is presented . Additionally, challenges of pose-guided human image transfer are discussed while and potential directions are suggested for future research in this field.

7.2 limitations

Since the intelligent system has successfully managed to automatically generate fashion images under certain conditions, there are still some limitations that need improvement.

The first limitation lies in the task of sketch-to-illustrative image style transfer. Although the model can perform style transfer, it still faces challenges in accurately transforming images into the target style that reflects the designers' unique aesthetics.

In the task of real-to-illustrative image transfer, precise alignment remains an area for improvement, especially in complex domains like fashion. The generated illustrative images often show small differences from the originals, particularly in aspects such as clothing texture and garment shape. The sketch images may lack sufficient detail, which can lead to inadequate constraints during the final image

synthesis.

For the task of pose-guided human image transfer, while the system can handle the task, there is still room for enhancing pixel-wise details. Additionally, the generated results are still 2D images, lacking the third dimension that would allow for rotation and more intuitive display.

The fourth limitation is that while the current intelligent fashion image generation system can perform sketch-to-illustrative and real-to-illustrative translations, it is still limited in performing arbitrary domain-to-domain translations. Furthermore, due to the limitations of existing datasets in the fashion field, there is a need to build a high-resolution dataset that captures the specific styles of individual designers. This is crucial for effectively addressing style transfer from the source domain to the target domain.

7.3 Suggestions for Future Research

Future work should address these limitations to develop a more comprehensive intelligent fashion image generation system.

Firstly, it is crucial to expand the available datasets to achieve arbitrary domain-to-domain style image transfer. By creating a more comprehensive paired dataset of sketches, illustrative images, and real images, existing powerful model architectures can learn the artistic styles of different designers, enabling the transferred images to embody the unique artistic styles of the designers. More varied and labeled data will make deep learning models more stable, so it is necessary to collect more extensive and diverse datasets that reflect the richness and complexity of fashion styles.

Secondly, the current multi-view presentation of designs is based on pose-guided human image transfer in 2D. In the future, by exploring rapid 2D-to-3D image generation, multi-view presentations can be achieved through 3D, enhancing object consistency while providing a more intuitive display of designers' and practitioners' ideas. This approach not only improves visibility but also streamlines the design process for designers and practitioners, inspiring their creativity.

By overcoming these limitations and exploring future directions, the intelligent fashion image generation system can continue to evolve, producing more accurate images that align with the preferences of designers and practitioners. This will allow the system to cater to diverse fashion domains and individual tastes more effectively.

References

[1] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7708–7717, 2018.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[3] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4502–4511, 2019.

[4] Gaurab Bhattacharya, Nikhil Kilari, Jayavardhana Gubbi, V Bagya Lakshmi, and P Balamuralidhar. F-attnet: Towards multi-scale feature fusion for fashion attribute prediction. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[5] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5968–5976, 2023.

[6] McKinsey BOF. The state of fashion report 2022. `https://www.businessoffashion.com/reports/news-analysis/`

the-state-of-fashion-2022-industry-report-bof-mckinsey/, 2022. Accessed: (11 May 2023).

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[8] Shidong Cao, Wenhao Chai, Shengyu Hao, and Gaoang Wang. Image reference-guided fashion design with structure-aware transfer by diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3524–3528, 2023.

[9] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. Unsupervised diverse colorization via generative adversarial networks. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 151–166. Springer, 2017.

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019.

[11] Haibo Chen, Lei Zhao, Jun Li, and Jian Yang. Tssat: Two-stage statistics-aware transformation for artistic style transfer. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6878–6887, 2023.

[12] Ming Chen, Yingjie Qin, Lizhe Qi, and Yunquan Sun. Improving fashion landmark detection by dual attention feature enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[13] Yilin Chen, Fan Zhou, and Zhuo Su. Cross-rolling attention network for fashion landmark detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4837–4843. IEEE, 2022.

[14] Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati,

and Jiaying Liu. Fashion meets computer vision: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–41, 2021.

[15] Wonwoong Cho, Sungha Choi, David Keetae Park, Inkyu Shin, and Jaegul Choo. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10639–10647, 2019.

[16] Charles Corbiere, Hedi Ben-Younes, Alexandre Ramé, and Charles Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2268–2274, 2017.

[17] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[18] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14638–14647, 2021.

[19] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[21] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022.

[22] Pragati Ashok Deole and Rushi Longadge. Content based image retrieval

using color feature extraction with knn classification. *IJCSMC*, 3(5):1274–1280, 2014.

[23] Wei Di, Catherine Wah, Anurag Bhardwaj, Robinson Piramuthu, and Neel Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *Proceedings of the IEEE Conference on computer vision and pattern recognition workshops*, pages 8–13, 2013.

[24] Antonio D'Innocente, Nikhil Garg, Yuan Zhang, Loris Bazzani, and Michael Donoser. Localized triplet loss for fine-grained fashion image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3910–3915, 2021.

[25] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[26] 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.

[27] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1161–1170, 2019.

[28] Jian Dong, Qiang Chen, Zhongyang Huang, Jianchao Yang, and Shuicheng Yan. Parsing based on parselets: A unified deformable mixture model for human parsing. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):88–101, 2015.

[29] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[30] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.

[31] Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. Insetgan for full-body image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7723–7732, 2022.

[32] 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 *European Conference on Computer Vision*, pages 1–19. Springer, 2022.

[33] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[34] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[35] Yuying Ge, Ruimao Zhang, and Ping Luo. Metacloth: Learning unseen tasks of dense fashion landmark detection from a few samples. *IEEE Transactions on Image Processing*, 31:1120–1133, 2021.

[36] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5337–5345, 2019.

[37] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[38] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[39] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7450–7459, 2019.

[40] 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.

[41] Xiaoling Gu, Jie Huang, Yongkang Wong, Jun Yu, Jianping Fan, Pai Peng, and Mohan S Kankanhalli. Paint: Photo-realistic fashion design synthesis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(2):1–23, 2023.

[42] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[43] Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2669–2680, 2023.

[44] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10471–10480, 2019.

[45] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018.

[46] Kota Hara, Vignesh Jagadeesh, and Robinson Piramuthu. Fashion apparel detection: the role of deep convolutional neural network and pose-dependent priors. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[48] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[49] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[50] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[51] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the 27th ACM international conference on multimedia*, pages 275–283, 2019.

[52] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[53] Ying Hu, Liqiang Xiao, Yongkun Wang, and Yaohui Jin. Fashion pose machine for fashion landmark detection. In *2018 International Conference on Image and Video Processing, and Artificial Intelligence*, volume 10836, pages 186–190. SPIE, 2018.

[54] Zhiyuan Hu, Jia Jia, Bei Liu, Yaohua Bu, and Jianlong Fu. Aesthetic-aware

image style transfer. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3320–3329, 2020.

[55] Chang-Qin Huang, Ji-Kai Chen, Yan Pan, Han-Jiang Lai, Jian Yin, and Qiong-Hao Huang. Clothing landmark detection using deep networks with prior of key point associations. *IEEE transactions on cybernetics*, 49(10):3744–3754, 2018.

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

[57] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[58] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[59] A Mustafa Ihsan, Chu Kiong Loo, Sinan A Naji, and Manjeevan Seera. Superpixels features extractor network (sp-fen) for clothing parsing enhancement. *Neural Processing Letters*, 51:2245–2263, 2020.

[60] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[61] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[62] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image

translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[63] Nataraj Jammalamadaka, Ayush Minocha, Digvijay Singh, and CV Jawahar. Parsing clothes in unrestricted images. In *BMVC*, volume 1, page 2, 2013.

[64] Wei Ji, Xi Li, Yueting Zhuang, Omar El Farouk Bourahla, Yixin Ji, Shihao Li, and Jiabao Cui. Semantic locality-aware deformable network for clothing segmentation. In *IJCAI*, pages 764–770, 2018.

[65] Shuhui Jiang, Jun Li, and Yun Fu. Deep learning for fashion style generation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4538–4550, 2021.

[66] 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.

[67] Yang Jiao, Yan Gao, Jingjing Meng, Jin Shang, and Yi Sun. Learning attribute and class-specific representation duet for fine-grained fashion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2023.

[68] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[69] Arnab Karmakar and Deepak Mishra. A robust pose transformational gan for pose guided person image synthesis. In *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*, pages 89–99. Springer, 2019.

[70] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22623–22633. IEEE, 2023.

[71] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[72] 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.

[73] 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.

[74] 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.

[75] Amena Khatun, Simon Denman, Sridha Sridharan, and Clinton Fookes. Pose-driven attention-guided image generation for person re-identification. *Pattern Recognition*, 137:109246, 2023.

[76] Tarasha Khurana, Kushagra Mahajan, Chetan Arora, and Atul Rai. Exploiting texture cues for clothing parsing in fashion images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2102–2106. IEEE, 2018.

[77] Bo-Kyeong Kim, Geonmin Kim, and Soo-Young Lee. Style-controlled synthesis of clothing segments for fashion image manipulation. *IEEE Transactions on Multimedia*, 22(2):298–310, 2019.

[78] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[79] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[80] Yasuyo Kita, Toshio Ueshiba, Ee Sian Neo, and Nobuyuki Kita. Clothes state recognition using 3d observed data. In *2009 IEEE International Conference on Robotics and Automation*, pages 1220–1225. IEEE, 2009.

[81] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4422–4431, 2019.

[82] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[83] Brian Lao and Karthik Jagadeesh. Convolutional neural networks for fashion classification and object detection. *CCCV 2015 Comput. Vis*, 546:120–129, 2015.

[84] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.

[85] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128(10):2402–2417, 2020.

[86] Sumin Lee, Sungchan Oh, Chanho Jung, and Changick Kim. A global-local embedding module for fashion landmark detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[87] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021.

[88] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1952–1961, 2023.

[89] Junnan Li and et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 2022.

[90] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

[91] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468, 2018.

[92] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019.

[93] Yixin Li, Shengqin Tang, Yun Ye, and Jinwen Ma. Spatial-aware non-local attention for fashion landmark detection. In *2019 IEEE international conference on multimedia and expo (ICME)*, pages 820–825. IEEE, 2019.

[94] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[95] Xiaodan Liang, Liang Lin, Wei Yang, Ping Luo, Junshi Huang, and Shuicheng Yan. Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. *IEEE Transactions on Multimedia*, 18(6):1175–1186, 2016.

[96] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized

convolutional neural network. In *Proceedings of the IEEE international conference on computer vision*, pages 1386–1394, 2015.

[97] Fangjian Liao, Xingxing Zou, and Wai Keung Wong. Attentional pixel-wise deformation for pose-based human image generation. *Expert Systems with Applications*, 246:123073, 2024.

[98] Fangjian Liao, Xingxing Zou, and Waikeung Wong. Deep fabric prints generation for fashion. *Design and Semantics of Form and Movement*, 88, 2023.

[99] Fangjian Liao, Xingxing Zou, and Waikeung Wong. Minigan: Toward informative and uninformative image transferring. *AATCC Journal of Research*, page 24723444221136635, 2023.

[100] Fangjian Liao, Xingxing Zou, and Waikeung Wong. Appearance and pose-guided human generation: A survey. *ACM Computing Surveys*, 56(5):1–35, 2024.

[101] Fangjian Liao, Xingxing Zou, and Waikeung Wong. Uni-dllora: Style fine-tuning for fashion image translation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6404–6413, 2024.

[102] Jianxin Lin, Yingxue Pang, Yingce Xia, Zhibo Chen, and Jiebo Luo. Tuigan: Learning versatile image-to-image translation with two unpaired images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 18–35. Springer, 2020.

[103] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[104] Ji Liu, Heshan Liu, Mang-Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Pose-guided high-resolution appearance transfer via progressive training. *arXiv preprint arXiv:2008.11898*, 2020.

[105] Ji Liu, Zhenyu Weng, and Yuesheng Zhu. Precise region semantics-assisted gan for pose-guided person image generation. *CAAI Transactions on Intelligence Technology*, 2023.

[106] Kuan-Hsien Liu, Ting-Yen Chen, and Chu-Song Chen. Mvc: A dataset for view-invariant clothing retrieval and attribute prediction. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*, pages 313–316, 2016.

[107] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

[108] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6649–6658, 2021.

[109] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[110] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019.

[111] Yun Liu, Guolei Sun, Yu Qiu, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. Transformer in convolutional neural networks. *arXiv preprint arXiv:2106.03180*, 3, 2021.

[112] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[113] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 229–245. Springer, 2016.

[114] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[115] Cewu Lu, Li Xu, and Jiaya Jia. Combining sketch and tone for pencil drawing production. In *Proceedings of the symposium on non-photorealistic animation and rendering*, pages 65–73. Citeseer, 2012.

[116] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017.

[117] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[118] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, page 3, 2013.

[119] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[120] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 768–783, 2018.

[121] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5084–5093, 2020.

[122] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[123] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[124] Mohammadreza Mostajabi, Payman Yadollahpour, and Gregory Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3376–3385, 2015.

[125] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[126] Kamyar Nazeri, Eric Ng, and Mehran Ebrahimi. Image colorization using generative adversarial networks. In *International conference on articulated motion and deformable objects*, pages 85–94. Springer, 2018.

[127] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[128] Ruru Pan, Weidong Gao, Jihong Liu, and Hongbo Wang. Automatic recognition of woven fabric pattern based on image processing and bp neural network. *The Journal of the Textile Institute*, 102(1):19–30, 2011.

[129] Kaicheng Pang, Xingxing Zou, Fangjian Liao, and Waikeung Wong. M-vton: Multi-layer virtual try-on system. *Design and Semantics of Form and Movement*, 266, 2023.

[130] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015.

[131] Viral Parekh, Karimulla Shaik, Soma Biswas, and Muthusamy Chelliah. Fine-grained visual attribute extraction from fashion wear. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3973–3977, 2021.

[132] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.

[133] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[134] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023.

[135] Shenhan Qian, Dongze Lian, Binqiang Zhao, Tong Liu, Bohui Zhu, Hai Li, and Shenghua Gao. Kgdet: Keypoint-guided fashion detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2449–2457, 2021.

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

[137] 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.

[138] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[139] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.

[140] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[141] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018.

[142] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[143] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[144] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13535–13544, 2022.

[145] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020.

[146] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.

[147] Muhamad Fajar Rivaldy, Febryanti Sthevanie, and Anditya Arifianto. Classification fashion image using local binary pattern and artificial neural network multi layer perceptron. *eProceedings of Engineering*, 7(2), 2020.

[148] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[149] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[150] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4814–4821, 2019.

[151] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[152] Ludger Rüschendorf. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1):117–129, 1985.

[153] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[154] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[155] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. A style-aware content loss for real-time hd style transfer. In *proceedings of the European conference on computer vision (ECCV)*, pages 698–714, 2018.

[156] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[157] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2025.

[158] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8242–8250, 2018.

[159] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[160] Ryotaro Shimizu, Takuma Nakamura, and Masayuki Goto. Fashion-specific ambiguous expression interpretation with partial visual-semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3496–3501, 2023.

[161] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416, 2018.

[162] Alexey Sidnev, Alexey Trushkov, Maxim Kazakov, Ivan Korolev, and Vladislav Sorokin. Deepmark: One-shot clothing detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[163] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. A high performance crf model for clothes parsing. In *Asian conference on computer vision*, pages 64–81. Springer, 2014.

[164] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[165] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[166] Asa Cooper Stickland and Iain Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995. PMLR, 2019.

[167] Guang-Lu Sun, Xiao Wu, Hong-Han Chen, and Qiang Peng. Clothing style recognition using fashion attribute detection. *EAI Endorsed Transactions on Ambient Systems*, 2(5), 2015.

[168] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[169] Zhengwentai Sun, Yanghong Zhou, Honghong He, and PY Mok. Sgdiff: A style guided diffusion model for fashion synthesis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8433–8442, 2023.

[170] Wisarut Surakarin and Prabhas Chongstitvatana. Classification of clothing with weighted surf and local binary patterns. In *2015 international computer science and engineering conference (ICSEC)*, pages 1–4. IEEE, 2015.

[171] Jan Svoboda, Asha Anoosheh, Christian Osendorfer, and Jonathan Masci. Two-stage peer-regularized feature recombination for arbitrary image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13816–13825, 2020.

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

[173] Jilin Tang, Yi Yuan, Tianjia Shao, Yong Liu, Mengmeng Wang, and Kun Zhou. Structure-aware person image generation with pose decomposition and semantic correlation. *arXiv preprint arXiv:2102.02972*, 2021.

[174] Pongsate Tangseng, Zhipeng Wu, and Kota Yamaguchi. Looking at outfit to parse clothing. *arXiv preprint arXiv:1703.01386*, 2017.

[175] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

[176] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[177] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normal-

ization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016.

[178] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[179] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[180] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[181] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4271–4280, 2018.

[182] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5703–5713, 2019.

[183] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[184] Xinrui Wang and Jinze Yu. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8090–8099, 2020.

[185] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[186] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[187] Jiacong Xu, Zixiang Xiong, and Shankar P Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19529–19539, 2023.

[188] Lumin Xu, Yingda Guan, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Vipnas: Efficient video pose estimation via neural architecture search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16072–16081, 2021.

[189] Shilin Xu, Xiangtai Li, Jingbo Wang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. Fashionformer: A simple, effective and unified baseline for human fashion segmentation and recognition. In *European Conference on Computer Vision*, pages 545–563. Springer, 2022.

[190] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *2012 IEEE Conference on Computer vision and pattern recognition*, pages 3570–3577. IEEE, 2012.

[191] Sijie Yan, Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 172–180, 2017.

[192] Fan Yang and Guosheng Lin. Ct-net: Complementary transfering network for garment transfer with arbitrary geometric changes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9899–9908, 2021.

[193] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1467–1475, 2019.

[194] Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *European conference on computer vision*, pages 517–532. Springer, 2016.

[195] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019.

[196] Weijiang Yu, Xiaodan Liang, Ke Gong, Chenhan Jiang, Nong Xiao, and Liang Lin. Layout-graph reasoning for fashion landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2937–2945, 2019.

[197] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019.

[198] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[199] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. PISE: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7982–7990, June 2021.

[200] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[201] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020.

[202] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7713–7722, 2022.

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

[204] Sanyi Zhang, Si Liu, Xiaochun Cao, Zhanjie Song, and Jie Zhou. Watch fashion shows to tell clothing attributes. *Neurocomputing*, 282:98–110, 2018.

[205] Sanyi Zhang, Guo-Jun Qi, Xiaochun Cao, Zhanjie Song, and Jie Zhou. Human parsing with pyramidical gather-excite context. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3):1016–1030, 2020.

[206] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 2248–2255, 2013.

[207] Xujie Zhang, Yu Sha, Michael C Kampffmeyer, Zhenyu Xie, Zequn Jie, Chengwen Huang, Jianqing Peng, and Xiaodan Liang. Armani: Part-level garment-text alignment for unified cross-modal fashion design. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4525–4535, 2022.

[208] Yuwei Zhang, Peng Zhang, Chun Yuan, and Zhi Wang. Texture and shape biased two-stream networks for clothing classification and attribute recognition. In *Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13538–13547, 2020.

[209] Bo Zhao, Xiao Wu, Qiang Peng, and Shuicheng Yan. Clothing cosegmentation for shopping images with cluttered background. *IEEE Transactions on Multimedia*, 18(6):1111–1123, 2016.

[210] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[211] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27th ACM international conference on multimedia*, pages 266–274, 2019.

[212] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[213] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[214] Dongliang Zhou, Haijun Zhang, Qun Li, Jianghong Ma, and Xiaofei Xu. Coutfitgan: learning to synthesize compatible outfits supervised by silhouette masks and fashion styles. *IEEE transactions on multimedia*, 2022.

[215] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016.

[216] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11465–11475, 2021.

[217] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *European Conference on Computer Vision*, pages 161–178. Springer, 2022.

[218] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In

Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017.

[219] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.

[220] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615, 2023.

[221] 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.

[222] Shumin Zhu, Xingxing Zou, Jianjun Qian, and Wai Keung Wong. Learning structured relation embeddings for fine-grained fashion attribute recognition. *IEEE Transactions on Multimedia*, 26:1652–1664, 2023.

[223] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019.

[224] Xingxing Zou and Waikeung Wong. fashion after fashion: A report of ai in fashion. *arXiv preprint arXiv:2105.03050*, 2021.

[225] Xingxing Zou and Waikeung Wong. Stylishgan: Toward fashion illustration generation. *AATCC Journal of Research*, page 24723444221147972, 2023.