Context-Aware Person Image Generation

A thesis submitted in fulfilment of the requirements
for the degree of

Doctor of Philosophy
in
Analytics

by

Prasun Roy

Supervisor: Prof Michael Blumenstein
Co-supervisor: Prof Umapada Pal

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW, Australia

February 2025

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Prasun Roy*, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science*, *Faculty of Engineering and Information Technology* at the University of Technology Sydney. This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note:
Signature removed prior to publication.

Prasun Roy

DATE: 7th February, 2025

PLACE: Sydney, Australia

Dedication

To my Mother

Acknowledgments

I would like to express my deepest gratitude to my advisors, Prof Michael Blumenstein (University of Technology Sydney) and Prof Umapada Pal (Indian Statistical Institute Kolkata), for their guidance, support, and feedback in shaping this research.

I am thankful to Dr. Saumik Bhattacharya (IITKGP) for being an incredible mentor and motivator who provided constant encouragement and confidence from the very beginning of my research journey. I will always cherish those insightful research discussions, sleepless nights before submission deadlines, and the button-mashing video game sessions.

I would also like to extend my heartfelt thanks to the panel members during my candidature assessments, Prof Wenjing Jia (UTS) and Prof Wei Liu (UTS), for their time and insightful suggestions, which significantly helped me to improve my research.

I am truly grateful to Dr. Chandan Saha (Calcutta School of Tropical Medicine), whose guidance during my junior years was instrumental in developing my scientific curiosity and analytical thinking. I am also indebted to Prof Manas Ghosh (RCC Institute of Information Technology) for his invaluable mentorship in building my foundation in Computer Science.

I am also thankful to all my collaborators and colleagues at the University of Technology Sydney and Indian Statistical Institute Kolkata for enriching my research with their diverse perspectives. I want to give special thanks to my fellow PhD scholar and friend Subhankar Ghosh (UTS) and my amazing lab mates Alloy Das, Kunal Biswas, Rakesh Dey, Kunal Purkayastha, Surajit Mukherjee, Arnab Halder, Subhajit Maity, and Ankit Lodh.

I would like to acknowledge Anthony Dang (Microsoft) for providing this thesis template, which was adapted from the original design by Dr. Chandranath Adak (IITP).

I want to thank the UTS Graduate Research School (GRS) for providing the International Research Scholarship (IRS) and the Faculty of Engineering & Information Technology Higher Degree Research (FEIT HDR) Scholarship. This research would not be possible without these generous grants and resources.

On a personal note, I express my deepest love and gratitude to my mother for being the source of my strength and inspiration. This thesis would not have been possible without her sacrifices. From my most challenging times to my most memorable day, she has always been there with her boundless love and support. This work is as much hers as it is mine, and I dedicate it to her with all my heart.

Thank you everyone.

Abstract

Human image generation is an intriguing yet fundamentally challenging problem in computer vision. The ability to synthesize high-quality and semantically meaningful novel instances of a person has potential use cases across multiple domains, including academic research and enterprise applications. Visual realism and contextual coherence in such generative frameworks can directly benefit synthetic data generation, scene understanding, creative software, digital media, retail advertisements, animation, and augmented / virtual reality (AR/VR) products. Although the premise is intellectually and commercially appealing, generating realistic novel human instances is a significantly challenging problem. Moreover, imposing semantic constraints on the generative process to achieve contextually coherent visual results introduces additional complexities to the problem. In recent years, both unconditional and conditional generative algorithms have achieved a remarkable uplift in photorealism by adopting Generative Adversarial Networks (GAN) and later Diffusion Models (DM). However, most existing approaches focus on synthesizing an instance of a specific object class. In contrast, a real-world scene generally contains multiple object classes with different inter-object contextual relationships. Therefore, conditioning a generation process on the scene context becomes essential for semantically meaningful visual synthesis.

This thesis explores two foundational aspects of context-aware person image generation. The first phase investigates visually realistic image generation of an isolated human instance from a local input context, such as geometric structure (pose) or textual descriptions. The second phase introduces global semantic constraints in the generative process and learns to blend a human instance into a complex scene while adapting to a contextually valid scene-human interaction. The high degree of appearance diversity and pose variations in human images contribute to the key challenges in the problem, where the aim is to generate semantically consistent novel views of a highly deformable object (human) from a single observation. The proposed research addresses these challenges by introducing generative strategies that achieve state-of-the-art performance on multiple visual and analytical benchmarks for context-aware person image generation.

PUBLICATIONS

RELATED TO THE THESIS

1. **Prasun Roy**, Subhankar Ghosh, Saumik Bhattacharya, Umapada Pal, and Michael Blumenstein. "TIPS: Text-Induced Pose Synthesis." In: *The European Conference on Computer Vision (**ECCV**)*. 2022, pp. 161–178

2. **Prasun Roy**, Subhankar Ghosh, Saumik Bhattacharya, Umapada Pal, and Michael Blumenstein. "Scene Aware Person Image Generation through Global Contextual Conditioning." In: *The International Conference on Pattern Recognition (**ICPR**)*. 2022, pp. 2764–2770

3. **Prasun Roy**, Saumik Bhattacharya, Subhankar Ghosh, and Umapada Pal. "Multi-scale Attention Guided Pose Transfer." In: *Pattern Recognition (**PR**)* 137 (2023), p. 109315

4. **Prasun Roy**, Saumik Bhattacharya, Subhankar Ghosh, Umapada Pal, and Michael Blumenstein. "Semantically Consistent Person Image Generation." In: *The International Conference on Pattern Recognition (**ICPR**)*. 2024, pp. 293–309

5. **Prasun Roy**, Saumik Bhattacharya, Subhankar Ghosh, Umapada Pal, and Michael Blumenstein. "Exploring Mutual Cross-Modal Attention for Context-Aware Human Affordance Generation." In: *The IEEE Transactions on Artificial Intelligence (**TAI**) (**accepted**)* (2025)

OTHERS

1. **Prasun Roy**, Saumik Bhattacharya, Subhankar Ghosh, Umapada Pal, and Michael Blumenstein. "d-Sketch: Improving Visual Fidelity of Sketch-to-Image Translation with Pretrained Latent Diffusion Models without Retraining." In: *The International Conference on Pattern Recognition (**ICPR**)*. 2024, pp. 277–292

2. Alloy Das, Sanket Biswas, **Prasun Roy**, Subhankar Ghosh, Umapada Pal, Michael Blumenstein, Josep Lladós, and Saumik Bhattacharya. “FASTER: A font-agnostic scene text editing and rendering framework.” In: *The IEEE/CVF Winter Conference on Applications of Computer Vision (**WACV**)*. 2025

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE **Page**

LIST OF TABLES

CHAPTER 1

INTRODUCTION

This introductory chapter discusses the essential background and motivation behind person image generation and the significance of incorporating object and semantic contexts in such generative frameworks. The discussion also provides insight into the associated challenges and key research questions. The chapter concludes by summarizing the main contributions of the thesis and outlining an overview of the thesis structure.

1.1 Background and Motivation

Generating realistic human instances is fundamental for building a wide range of modern computer vision application stacks, including but not limited to synthetic data generation, digital media, retail advertisements, animation, and augmented / virtual reality (AR/VR) software. For example, high-quality synthetic human instances can improve person re-identification by providing auxiliary data to the model. Likewise, such generations can streamline the digital advertising of fashion apparel by seamlessly creating different views of a person wearing it. Therefore, the ability to synthesize high-quality and contextually appropriate novel views of a person has potential use cases across academic research and enterprise applications.

The key aim of context-aware person image generation is to generate novel views (images) of a highly deformable object (human) from a single observation and additional contextual input. The input context can provide local or global guidance to the generative

network. A local guidance condition is an object-level context to control the generation of an isolated human instance. For example, keypoint-based target pose representation or textual description of the target pose provides a local structural context to control the intended pose of the generated person. On the other hand, a global guidance condition is a scene-level semantic context that constrains the generated human instance in a complex environment. For example, human-human and human-object interactions in a scene provide a global semantic context for meaningfully blending new human instances. Therefore, practical applications of such generative networks require intricate attention to both local and global contexts.

In recent years, both unconditional and conditional generative algorithms achieved a remarkable uplift in photorealism by adopting Generative Adversarial Networks (GAN) [1, 2] and later Diffusion Models (DM) [3, 4, 5, 6]. However, most existing approaches focus on synthesizing an isolated person instance from keypoint-based pose representation with limited attempts to incorporate semantic constraints. We notice two major limitations in the current methods. First, while the keypoint-based representation provides a structurally accurate target pose, it tends to produce noticeable shape ambiguities when the target pose provider has a significantly different physique than the subject. This problem is not immediately apparent from the initial visual analysis because the datasets consist of source-target image pairs for every individual, ensuring that the subject is the same as the target pose provider. However, the target pose must come from a different person in a real-world inference scenario, leading to the potential geometric bias in the output. The second problem arises from inadequate semantic conditioning for person image generation in a complex scene, resulting in poor and unrealistic generative performance. A complex real-world environment typically contains multiple object classes with different inter-object contextual relationships. Therefore, conditioning the generative process on the scene context becomes essential for semantically meaningful visual synthesis. Due to the potentially appealing application premises, we believe it is worthwhile to investigate and address such limitations for improving the visual quality and generative stability in context-aware person image generation.

This thesis explores two foundational aspects of context-aware person image generation. The first phase investigates visually realistic image generation of an isolated human instance from keypoint-based pose reference, followed by a potential strategy for mitigating structural bias using text-based descriptive pose annotations. The second phase introduces semantic constraints in the generative process and learns to blend a new person into a complex scene while adapting to a contextually valid scene-human interaction.

1.2 Significance and Challenges

At present, the academic and industrial landscapes are going through major transformations with the exponential adoption of generative technologies. Synthetic data generation pipelines allow transformative graphics and simulation tools to be integrated into commercial game engines such as Unreal [7] and Unity [8, 9]. Augmenting real data with samples from procedurally generated large datasets like SURREAL [10], Synscapes [11], and SynBody [12] shows improved model performance for various downstream tasks. Similarly, in the digital media and AR/VR space, recent reports [13, 14] suggest that the global market valuation is expected to expand from USD 22.12 billion to USD 96.32 billion by 2029, growing at a CAGR of 34.2%. Multiple surveys and whitepapers [15, 16] have identified up to 40% return rates in fashion e-commerce, accounting for nearly USD 218 billion loss in revenue globally. Modern generative AI-assisted virtual try-on tools [15] have consistently shown positive impacts on customer behavior, with a 60% reduction in the return rates for online retail purchases. Likewise, in animation and virtual production, motion translation and digital avatars have reduced production costs of studio pipelines while improving user immersion.

Although the premise is intellectually and commercially appealing, generating realistic human instances is a significantly challenging problem. The high degree of appearance diversity and pose variations in human images cause many possible learnable mappings for pose transformation. Moreover, imposing semantic constraints on the generative process introduces additional complexities, such as context representation and ambiguities due to multiple valid solutions. We summarize the main challenges associated with context-aware person image generation as follows.

1. **Visual and structural diversity**: Human appearance can be visually and structurally diverse. Differences in specific physical traits among people from different demographics and wide variations in human attire collectively contribute to visibly diverse person images. Additionally, high deformability of the human body due to wide degrees of freedom results in virtually innumerable complex poses that a person can adopt. Therefore, the generative process must account for such visual and structural variabilities to produce realistic outcomes.

2. **Generalization and dataset bias**: In a pose-conditioned human image generation from a given observation to a novel pose, the existing datasets consist of observed and target image/pose pairs of a specific individual for training supervision. However, the target pose must come from a different person during a real-world inference. In such

cases, a substantial structural inconsistency is likely when the target pose provider has a significantly different physique than the subject. Therefore, a generalized pose representation is required to mitigate such dataset bias.

3. **Context representation**: Context-aware person image generation fundamentally depends on object-level local and scene-level global context representations. An object-level context, such as keypoint-based spatial pose representation or text-based descriptive pose annotation, provides a local geometric context to control the structure of an isolated human instance. Likewise, a scene-level context, such as a person-person or scene-person relationship, provides a global context to constrain the generated human instance within a complex environment for semantically adaptive blending. Therefore, effective mechanisms to represent the object and semantic contexts are crucial for generating expressive human instances.

4. **Semantic ambiguity**: When introducing a new person instance into an existing complex scene, the algorithm attempts to synthesize a novel view of the person that can seamlessly blend into the given scene. However, numerous variations of the target person are feasible in such cases, resulting in multiple semantically valid scene-human compositions. Therefore, the generative process should be able to sample multiple variations of the composed scene while retaining semantic integrity across such variations.

5. **Ethical concerns**: The rapid advance of deep generative models for human imagery has enabled powerful applications but also introduced serious ethical concerns regarding user privacy and data safety. Synthetically generated *DeepFake* images and videos can convincingly fabricate or alter an individual's appearance and actions [17, 18], eroding trust in digital media. Several reports [19, 20] have indicated a massive surge of political disinformation, identity theft, targeted harassment, and audio-visual impersonation using generative technologies in recent years. While researchers are actively exploring computational algorithms [21, 22] for precisely identifying fake information, several recent studies [23, 24] have shown that the majority of such *DeepFake* detectors can be bypassed by simple postprocessing steps. This has prompted calls for legal frameworks such as the **EU AI Act** (European Union) and the **DEEPFAKES Accountability Act** (USA) to mandate watermarking, provenance tracking, and clear liability for misuse. Some ethical mitigation strategies include embedding imperceptible digital watermarks at generation time, developing joint generation-detection architectures to mask malicious outputs, and fostering

interdisciplinary collaboration among technologists, policymakers, content creators, social influencers, and end users to ensure responsible deployment of generative AI.

1.3 Research Objectives and Questions

This research aims to investigate efficient strategies for generating novel human instances from local object-level contexts, followed by imposing semantic constraints for the adaptive composition of a generated person into a complex environment from global scene-level contexts. The main research objectives and associated research questions are summarized as follows.

> ***Objective:*** *To investigate efficient strategies for generating isolated novel views of a specific person from a single observation and local structural context*

- ***Research Question 1:*** How to efficiently improve existing approaches to geometrically guided human pose transformation?
- ***Research Question 2:*** How does strong structural supervision impact the generative process during real-world inference?
- ***Research Question 3:*** How to effectively mitigate the potential structural bias in pose-guided person image generation?

> ***Objective:*** *To design generative strategies for adaptively blending a specific person into a complex scene by imposing global semantic constraints*

- ***Research Question 4:*** How to effectively introduce semantic conditioning in a scene-aware adaptive person image generation pipeline?
- ***Research Question 5:*** How does a data-agnostic approach impact the visual quality and scalability of semantic person image generation and composition?
- ***Research Question 6:*** How does cross-modal information fusion impact human affordance generation in complex scenes and associated downstream tasks?

1.4 Contributions

This research addresses the core challenges in context-aware person image generation by improving the visual quality of generated instances and mitigating probable structural ambiguity during real-world inference, followed by imposing semantic constraints for adaptively blending generated instances into a complex scene. The main contributions of the thesis are summarized as follows.

1. To efficiently improve the visual quality in human pose transformation, an end-to-end network architecture is proposed for generating novel views of a specific person from a single observation and a keypoint-based target pose reference as the local structural context.

2. To address potential structural inconsistencies from keypoint-based pose representation, an alternative network architecture is proposed for human pose transformation using textually descriptive pose annotation as the local context. A new dataset comprising human image and pose description pairs is also introduced to circumvent the lack of existing datasets.

3. To impose semantic constraints on the adaptive blending of a new person into a scene with existing people, a disentangled pipeline is introduced where the collective association of all human poses provides a global semantic context to the generative network.

4. To improve the visual quality, structural stability, and scalability of semantic person instance blending into complex multi-person scenes, a data-agnostic generative architecture is introduced by replacing keypoint-based pose representations with human parsing maps.

5. To mitigate the structural instability of keypoint-based pose representations and space complexity of data-agnostic approaches, a novel cross-attention mechanism is proposed for encoding the global semantic context for expressive human affordance generation in complex scenes.

1.5 Thesis Structure

This thesis consists of an introductory chapter for understanding the premise and objectives, an extensive literature review to identify the limitations and scopes of existing

approaches, five subsequent technical chapters discussing the proposed strategies to address core research questions and a final concluding chapter to summarize the main research findings. The organization of the thesis is as follows.

Chapter 1 introduces the premise by discussing the background, motivation and potential challenges associated with context-aware person image generation. The discussion also sets the core research objectives and identifies the key research questions. The chapter concludes by summarizing the main contributions and organization of the thesis.

Chapter 2 provides a comprehensive literature review by exploring previous works on person image generation to identify the limitations and potential scopes of existing approaches.

Chapter 3 aims to address ***Research Question 1*** by introducing an end-to-end network architecture that uses attention operation at every spatial scale of encoding and decoding branches to increase both low-frequency and high-frequency feature richness. The proposed approach [25] outperforms existing keypoint-based structurally guided human pose transformation methods in multiple visual and analytical benchmarks.

Chapter 4 aims to address ***Research Questions 2 & 3*** by proposing a disentangled human pose transformation strategy [26] that uses textual annotations as the pose descriptor to mitigate potential structural irregularities in a keypoint-based approach.

Chapter 5 introduces a generative architecture for adaptively blending a specific person into a scene with existing people. The proposed method [27] addresses ***Research Questions 4*** by providing the collective association of existing human poses as a global semantic context to the generative network.

Chapter 6 focuses on improving the visual quality, structural stability, and scalability of semantically adaptive person instance blending into complex multi-person scenes. The proposed data-agnostic approach [28] addresses ***Research Questions 5*** by adopting human parsing maps instead of a highly sparse keypoint-based pose representation.

Chapter 7 aims to solve the structural instability of keypoint-based pose representations and space complexity of data-agnostic strategies in an adaptive human pose generation pipeline. Addressing ***Research Questions 6***, the proposed approach [29] introduces a novel cross-attention mechanism to encode the global context that improves sampling semantically valid human actions in a complex scene.

Chapter 8 concludes the thesis by summarizing the research findings and discussing potential scopes for future directions.

CHAPTER 2

LITERATURE REVIEW

This chapter provides an overview of the existing methods related to context-aware person image generation. The discussion includes general techniques for conditional visual synthesis and their adoption into human pose transformation from local and global contextual supervision. Although local structural guidance is well explored in the literature, existing works on global semantic guidance are substantially limited. This extensive literature review enables us to identify and address the shortcomings of current person image generation strategies.

2.1 Generative Modeling for Image-to-Image Translation

GAN: Visual synthesis is a fundamental requirement for many computer vision applications. Since the inception of Generative Adversarial Networks (GAN) [1], several variations of the core architecture achieved remarkable improvements in realistic image synthesis during recent years. The initial proposal of GAN [1] introduced *unconstrained* image generation by enforcing an adversarial learning scheme between a generator and a discriminator. Later, conditional GAN (CGAN) [2] extended this idea to *constrained* image generation. These adversarial learning strategies led to the original foundation of *image-to-image* translation [30] and *cyclic consistency* [31].

Conditional GAN: Image-to-Image translation using conditional GAN is a method of image transformations between two domains. Early architectural improvements introduced a Markovian discriminator [30] for better retention of high-frequency correctness in *paired*

image-to-image translation. A subsequent approach [31] extended the idea to *unpaired* data by enforcing cycle consistency between source and target domains. In [32], the authors used coarse-to-fine generators, multi-scale discriminators, and an additional feature-matching loss for generating higher-resolution images. In [33], the authors achieved generational improvements in semantic image manipulation by introducing *spatially adaptive normalization*. A specific variant of general image-to-image translation focuses on realistic image generation from freehand sketches, where sketches act as rough visual cues to impose structural guidance on the generative process. The initial work exclusively on multi-class sketch-to-image translation proposed a *masked residual unit* [34], accommodating fifty object categories. Another approach proposed a contextual GAN [35] to learn the joint distribution of the sketch and corresponding image. Researchers also explored interactive generation [36] using a gating mechanism to suggest the probable completion of a partial sketch, followed by rendering the final image with a pretrained image-to-image translation model [32]. In [37], the authors proposed a multi-stage class-conditioned approach for *object-level* and *scene-level* image synthesis from freehand sketches, improving the perceptual baseline over direct generations [30], contextual networks [35], and methods based on *scene graphs* [38, 39] or *layouts* [40]. In [41], the authors achieved similar goals with an *unsupervised* approach by introducing a standardization module and disentangled representation learning.

GAN inversion: The main objective of GAN inversion is to find a latent embedding of an image such that the original image can be faithfully reconstructed from the latent code using a pretrained generator. The existing strategies for such inversions can be *learning-based* [42, 43, 44, 45], *optimization-based* [46, 47, 48, 49, 50, 51, 52], or *hybrid* [53, 54]. In a learning-based inversion, an encoder learns to project an image into the latent space, minimizing reconstruction loss between the decoded (reconstructed) and original images. An optimization-based inversion estimates the latent code by directly solving an objective function. In a hybrid approach, an encoder first learns the latent projection, followed by an optimization strategy to refine the latent code. The rich statistical information captured by deep generative networks from large-scale data provides effective *priors* for various downstream tasks, including image-to-image and sketch-to-image translations. In [45], the authors adopted a learning-based GAN inversion strategy using a multi-class deep generative network [55], pretrained on ImageNet dataset [56], as *prior* to achieve sketch-to-image translation for multiple categories. In [57], the authors introduced a framework for generalizing image synthesis to *open-domain* object categories by jointly learning two *in-domain* mappings (image-to-sketch and sketch-to-image) with *random-mixed* strategy.

Diffusion models: A Denoising Diffusion Probabilistic Model (DDPM) [3, 4] is a parameterized Markov chain that learns to generate samples similar to the original data distribution after a finite time. In particular, DDPM uses variational inference to learn to iteratively reverse a stepwise *diffusion* (noising) process. In [58], the authors introduced Denoising Diffusion Implicit Models (DDIM) by generalizing DDPM using non-Markovian diffusion processes with the same learning objective, leading to a deterministic and faster generative process. Recent advances [5, 6] have shown that diffusion models can achieve generational improvements in the visual quality and sampling diversity over GAN while providing a more stable and straightforward optimization objective. The most prolific application of diffusion models in recent literature is *text-conditioned* image generation [59, 60, 61, 62] and modification [63, 64, 65, 66], utilizing a pretrained language-image model [67] to embed the conditioning prompt. In [68], the authors guided the generative process with an iterative latent variable refinement to produce high-quality variations of a reference image. In [69], the authors introduced a class-specific prior preservation loss to finetune an existing text-to-image diffusion model for *personalized* manipulation of a specific subject image from a few observations. Emerging alternative approaches also involved Stochastic Differential Equations (SDE) to guide the generative process following *score-based* [70] or *energy-based* [71, 72] objectives. More recent attempts for sketch-to-image translation involved multiple objectives [73], multi-dimensional control [74], or latent code optimization [75]. In [73], the authors used an additional network to reconstruct the input sketch from the generated image. The denoising process was optimized using a cumulative objective function consisting of the *perceptual similarity* (between the input and reconstructed sketches) and *Cosine Similarity* (between the input and generated images) measures. In [74], the authors provided three-dimensional controls over image synthesis from the strokes and sketches to manipulate the balance between *perceptual realism* and *structural faithfulness* during the conditional denoising process. In [75], the authors introduced a lightweight mapping network for providing structural guidance to a pretrained latent diffusion model [61]. While the method avoided training a dedicated diffusion network, the *differential guidance* made sampling images computationally even more demanding than a large-scale model itself. To alleviate such problems, we proposed dSketch [76] for photorealistic sketch-to-image translation by leveraging the learned feature space of a pretrained *latent diffusion model* (LDM) [61]. We achieved this by using a learnable lightweight feature mapping network to perform latent code translation between *source* (sketch) and *target* (image) domains. By adopting this strategy, we retained the remarkable generative capabilities of the LDM prior without requiring to retrain it.

2.2 Structurally Guided Human Pose Transformation

The key aim of human pose transformation is to generate a novel view of a specific person from a single observation and a given pose. Researchers extensively used keypoints and semantic parsing maps for 2D pose representation alongside body mesh and UV maps for 3D pose representation. Most existing methods adopt GAN or diffusion models as the core architecture for pose transformation in a supervised or semi-supervised manner.

Keypoint-guided approaches: The first human pose transformation method PG^2 [77] introduced a two-stage *coarse-to-fine* generation approach for pose integration and image refinement using U-Net-like architectures [78]. The authors further improved the technique with a *disentangled* person image generation scheme [79] by introducing separate branches for foreground, background, and pose in the network architecture. Siarohin *et al.* [80] proposed *deformable* skip connections in the GAN architecture with a *nearest-neighbor* loss to address pixel-level misalignments in human pose transformation. Esser *et al.* [81] proposed a conditional U-Net architecture for generating different views of an object, conditioned on the latent appearance attributes from a variational autoencoder (VAE) [82]. In [83], the authors introduced a sequence of *pose-attention transfer blocks* for progressive pose transformation. The key idea was to perform the translation on a local manifold at each intermediate step to overcome the difficulties arising from complex structures on the global manifold. In [84], the authors proposed a novel bi-directional feature transformation strategy for better utilizing the guidance image constraints. CoCosNet [85] jointly learned cross-domain correspondence and image translation by mutually improving each other with weak supervision. In [86], the authors proposed a progressive appearance transformation strategy with a region-specific adaptive patch normalization. In PoNA [87], the authors proposed a *pose-guided non-local attention* mechanism with a *long-range dependency* scheme to select important feature regions for pose transformation. Tang *et al.* introduced a generative architecture XingGAN [88] containing two novel blocks to effectively transfer and update the shape and appearance embeddings in a crossing way to improve each other mutually. The authors improved the architecture using a bipartite graph reasoning scheme BiGraphGAN [89] with an attention-based image fusion block for addressing the long-range relations between the source and target pose to mitigate the challenges caused by pose deformation. The attention mechanism works on a single scale of the final layer to generate a one-channel attention mask. In SCA-GAN [90], the authors aimed to alleviate spatial misalignments using edge maps alongside keypoints. In Pot-GAN [91], the authors utilized multi-scale

feature maps for pose-guided person image generation. Ren *et al.* [92] proposed a neural texture extraction and distribution mechanism for controllable person image synthesis. Zhang *et al.* [93] proposed a transformer-based dual-task *Siamese* architecture containing two branches for self-reconstruction and target transformation. The authors showed that the auxiliary task of *source-to-source* generation helps the *source-to-target* generation. In [94], the authors introduced a dynamic sparse attention-based transformer architecture DynaST with dynamic attention in a cascaded multi-layer transformer network.

Parsing / UV map-guided approaches: Men *et al.* [95] used human parsing maps [96] to embed appearance attributes into the latent space as independent codes and a two-branch network for generation. In [97], the authors implicitly represented body pose and shape as a parametric mesh using a learned high-dimensional UV feature map to capture appearance variations across poses, viewpoints, identities, and clothing styles during re-rendering. Zhang *et al.* [98] proposed a two-stage approach by estimating a parsing map aligned with the pose keypoints followed by transferring appearance attributes to render the human instance. The authors later introduced PISE [99], featuring joint global and region-wise local normalization to decouple shape and style with a spatial-aware normalization to retain the spatial context. In SPGNet [100], the authors proposed a two-stage architecture for pose and appearance translation with region-adaptive normalization. In [101], the authors performed source-to-target appearance attributes transfer with a *human body symmetry prior* followed by a pose-conditioned StyleGAN2 [102] generator with *spatial modulation* for photorealistic reposing. Liu *et al.* [103] proposed a *spatial-aware texture transformer* model for source-to-target garment transfer by leveraging the spatial UV prior from DensePose [104]. In StylePoseGAN [105], the authors extended a non-controllable neural re-rendering scheme to controllable person image generation from a single monocular view by imposing pose and appearance conditioning separately. In HumanGAN [106], the authors used a part-based latent appearance encoding in a pose-independent normalized space followed by warping the encoded latent vectors to different poses. Zhou *et al.* [107] introduced a *cross-attention-based style distribution* module, computing between source semantic style and target pose to perform the pose translation. In [108], the authors proposed a *self-driven* approach PT^2 for pose transformation by splitting pose and textures at *patch-level*.

Mesh-guided approaches: Lassner *et al.* [109] proposed a generative scheme to manipulate clothing styles from a given image outline of a projected 3D body mesh. Zanfir *et al.* [110] estimated the 3D body mesh from a single image followed by appearance rendering for pose transformation. Liu *et al.* [111] proposed a pose transformation technique using a

3D mesh recovery module in liquid warping GAN to learn the location and rotation of joints alongside body shape characteristics. The authors later introduced an attention mechanism in the architecture [112] to improve its generative performance. Knoche *et al.* [113] performed pose transformation by intermediate translation into a volumetric space. Specifically, the authors geometrically warped 2D images into dense 3D feature volume, performed pose manipulation, and finally mapped the volumetric representation back to RGB space. In [114], the authors introduced a *lifting-and-projection* network and an *appearance detail compensating* network for 3D mesh-based human pose transformation.

Flow-guided approaches: Li *et al.* [115] proposed a dense and intrinsic *appearance flow* for pose transformation by fitting a 3D model to the given pose pair and then projecting them back to the 2D plane to compute the appearance flow. The authors performed feature warping using the estimated appearance flow to generate photorealistic human instances. Ren *et al.* [116] performed pose transformation by estimating flow fields from the global correlations between source and target domains, followed by computing local attention coefficients and warping source features with a content-aware sampling mechanism. In FDA-GAN [117], the authors applied *deformable local attention* and *flow similarity attention* to perform occlusion and deformation-aware feature fusion for a flow-based pose transformation. VGFlow [118] disentangled the flow into visible and occluded parts of the target with a *visibility-guided flow module*, facilitating simultaneous texture preservation and style manipulation. WaveIPT [119] introduced a pose transformation strategy by fusing the attention and flow in the wavelet domain.

Semi / Self-supervised approaches: Although most existing human pose transformation methods follow a *supervised* learning approach, researchers also attempted *semi-supervised* [120, 121, 122] and *self-supervised* [123, 124] learning strategies in recent years. Pumarola *et al.* [120] proposed a *pose-conditioned bidirectional generator* to map the rendered person image back to the initial pose, thereby dismissing the requirement for supervision during training. Song *et al.* [121] used a *semantic generator* for source-to-target parsing map translation and an *appearance generator* to render the target semantic map. The authors then reconstructed the source by running identical steps in reverse and optimized the networks with a reconstruction loss. Zheng *et al.* [122] proposed a two-stage *coarse-to-fine* generation strategy by learning multi-scale *pose flow* with a *texture-preserving* objective. In [123], the authors formulated self-supervised pose transformation based on *cyclic consistency*. In [124], the authors proposed a self-supervised approach using two collaborative modules to create unaligned pairs in the feature space followed by feature rearrangement. Additionally, the authors introduced a graph-based *body structure*

retaining loss for shape-consistent human pose transformation.

2.3 Textually Supervised Human Pose Transformation

Image generation from textual supervision is an intriguing topic in applied machine vision. In recent years, researchers introduced large-scale multimodal architectures, bridging language understanding and visual attributes. For example, OpenAI's *DALL·E* [59] used a discrete VAE and transformer to translate free-form text prompts into images, while *DALL·E 2* [60] introduced a diffusion-based decoder to improve the resolution and visual fidelity of the generated samples further. Likewise, Google's *Imagen* [62] demonstrated that cascaded diffusion decoders can achieve remarkable photorealism, often rivaling human judgment, when paired with transformer-based text encoders. Conversational agents like *ChatGPT* and *LLaMA* have shown how autoregressive large language models [125, 126] can be steered via natural dialogue to generate descriptive prompts for downstream image or video generation. While such multimodal approaches offer high visual fidelity for general visual synthesis and open-domain creativity, a purpose-specific task such as person image generation requires fine-grained structural control and identity preservation alongside efficient architecture design with lower computational footprints.

However, text-guided person image generation techniques are not widely explored in the literature. The initial approach [127] to this problem used a conditional GAN for *text-to-image* synthesis. In [128], the authors used a VAE to generate human actions from text descriptions. In [129], the authors introduced redescriptions of texts for generating images. Zhou *et al.* [130] proposed a text-guided method for generating human instances by selecting a pose from a set of eight basic poses, followed by controlling appearance attributes of the chosen basic pose. In recent literature, researchers estimated the human appearance [131] and pose [132] from a given text description. Briq *et al.* [132] synthesized 3D human meshes from text using a recurrent GAN and SMPL [133] model.

2.4 Semantically Adaptive Person Image Generation

Unlike isolated human instance generation from local shape contexts, adaptive image generation within a scene requires intricate supervision of the global semantic context. Initially, researchers introduced a contextually relevant *random* person instance into a user-defined location [134] or a probabilistically estimated potential area [135, 136] by performing a background context-conditioned instance-level search followed by image

composition. In contrast, we aim to introduce a *specific* person instance into an optimally estimated scene location such that the new person contextually *blends in* with the existing persons. However, existing methods to address this challenging task are limited in the literature. Gafni *et al.* [137] proposed a conditional GAN-based multi-stage approach for person instance blending in a scene. Specifically, the authors used three dedicated networks for semantic parsing map-based human structure generation, rendering, and face refinement. Kulal *et al.* [138] adopted an end-to-end *conditional inpainting* technique by finetuning a pretrained *latent diffusion model* to achieve similar goals.

2.5 Context-Aware Human Affordance Generation

The original investigation [139] on the relationship between visual perception and human action defined *affordance* as the opportunities for interaction with the surrounding environment. Behavioral studies on regular and cognitively impaired persons indicated evidence that perception results in both visual and motor signals in the human brain. An extended study [140] demonstrated that visual attention to the spatial characteristics of the perceived objects initiates automatic motor signals for different actions. In computer vision, human affordance learning involves novel pose prediction such that the estimated pose represents a valid human action within the scene context. The task is critical for many problems requiring robust semantic reasoning about the environment, such as human motion synthesis [141] and scene-aware human pose generation [142, 143, 144, 145].

Earlier methods of affordance learning explored knowledge mining [146] and multimodal feature cues [142] to address the problem. In [146], the authors used a Markov Logic Network for constructing a knowledge base by extracting several object attributes from different image and metadata sources, which can perform various downstream visual inference tasks without any additional classifier, including zero-shot affordance prediction. In [142], the authors used depth map, surface normals, and segmentation map as multimodal cues to train a multi-scale convolutional neural network (CNN) for scene-level semantic label assignment associated with specific human actions. In [147], the authors designed a multi-branch end-to-end CNN with two separate pathways for object detection and affordance label assignment to achieve high real-time inference throughput. Researchers [148] also explored socially imposed constraints for affordance learning. In [148], the authors proposed a graph neural network (GNN) to propagate contextual scene information from egocentric views for action-object affordance reasoning.

Probabilistic modeling of scene-aware human motion generation also involves

semantic reasoning of human interaction with the environment. Initial works on human motion synthesis followed different architectural approaches, such as sequence-to-sequence models [149], generative adversarial networks (GAN) [149, 150, 151], graph convolutional networks (GCN) [152], and variational autoencoders (VAE) [153]. However, these methods mostly ignored the role of environmental semantics. Due to potential uncertainty in human motion, in a recent approach [141], the authors addressed such motion synthesis with a GAN conditioned on scene attributes and motion trajectory to predict probable body pose dynamics.

One key challenge of human affordance generation in 2D scenes is the lack of large-scale datasets with rich pose annotations. In [143], the authors compiled the only public dataset of annotated human body poses in complex 2D indoor scenes by extracting frames from sitcom videos. Aiming to generate a contextually valid human affordance at a user-defined location, the authors proposed sampling the scale and deformation parameters for an existing human pose template using a VAE conditioned on the localized image patches as scene context. In [144], the authors introduced a two-stage GAN architecture for achieving a similar goal by estimating the affine bounding box parameters to localize a probable human in the scene and then generating a potential body pose at that location. The method uses the input scene, corresponding depth, and segmentation maps as semantic guidance. In [145], the authors proposed a transformer-based approach with knowledge distillation for generating human affordances in 2D indoor scenes.

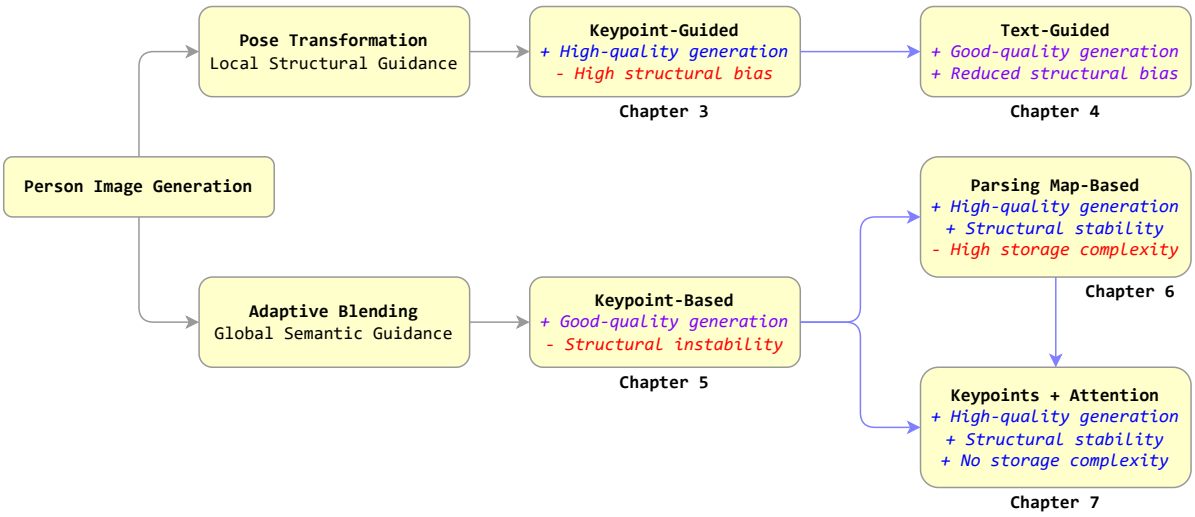

Figure 2.1: An organizational overview of the proposed works in this thesis.

2.6 Chapter Summary

In this chapter, we discussed the recent literature on person image generation. We reviewed the foundational generative architectures, such as GAN and diffusion models, for general image-to-image translation, followed by adopting these techniques into human pose transformation with local or global contextual guidance. The most common methods for providing structural and shape contexts to the generative network include keypoints, parsing maps, UV priors, or 3D mesh. However, existing works on textual supervision and semantic constraints are substantially limited. In the following chapters, we discuss an improved network architecture for structurally guided human pose transformation, explore a technique to alleviate the possible geometric bias with textual supervision, and introduce methods to impose semantic constraints for adaptive person image generation within a given scene. Fig. 2.1 illustrates an overview of the proposed works in this thesis.

CHAPTER 3

STRUCTURALLY GUIDED HUMAN POSE TRANSFORMATION

This chapter introduces an end-to-end network architecture for structurally guided human pose transformation. The proposed approach uses attention operation at multiple scales of the encoding and decoding branches to enhance low-frequency and high-frequency image features, outperforming previous methods on multiple visual and analytical benchmarks.

3.1 Human Pose Transformation

Human pose transformation aims to generate previously unseen novel views (images) of a specific person from a single observation and an additional local geometric context of the intended target pose. Such generative transformation is commonly termed as *Pose Transfer*. The main challenges of human pose transfer arise from large variations in physical traits, clothing, and possible poses.

The initial solution to this problem [77, 79] introduced a coarse-to-fine generation by dividing the problem into individual sub-tasks to handle foreground, background, and pose separately. The complexity of such a multi-stage architecture is later simplified with a unified approach, using deformable GAN [80] and variational U-Net [81]. A more streamlined approach [83] is introduced by leveraging the attention mechanism to transform the pose progressively. The key idea is to perform pose translation on a local manifold at each intermediate step to avoid potential difficulties with multiple complex structures on the global manifold. In this technique, an encoder initially downsamples the

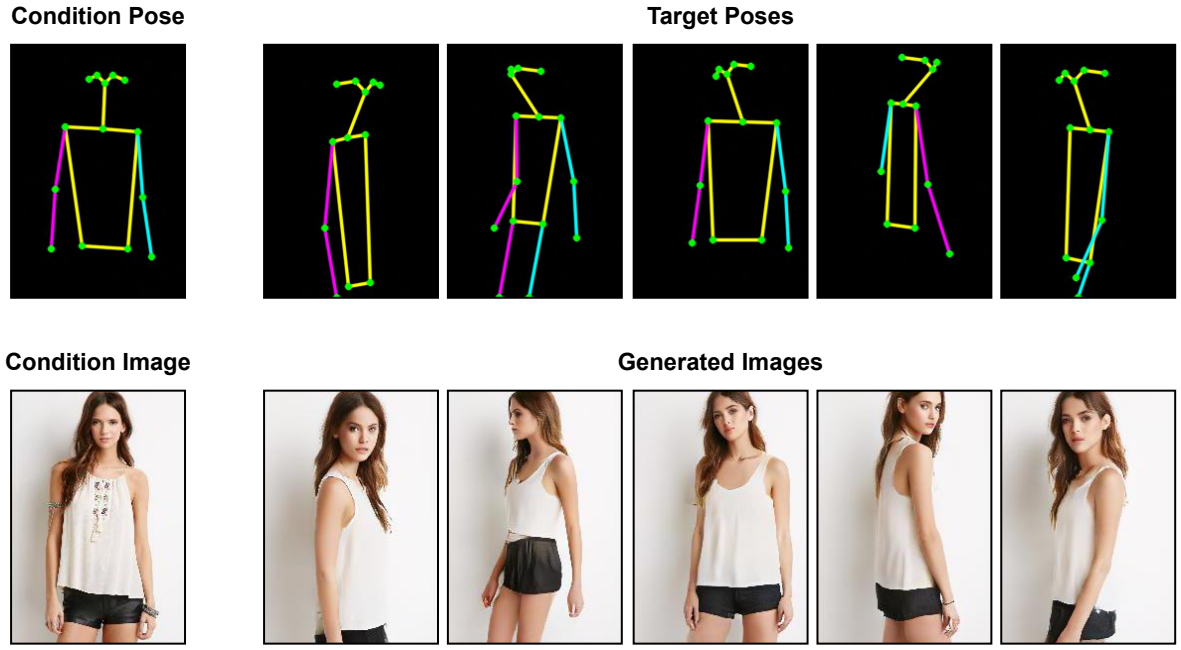

Figure 3.1: A general overview of human pose transformation using the proposed method.

given observation (condition image) and a sparse representation of the target pose to a lower spatial resolution. Next, the attention-guided progressive pose transfer is performed on the encoded feature space. Finally, a decoder upsamples the resulting feature space back to a higher-resolution output image. This elegant end-to-end strategy outperformed all previous methods in visual quality and analytical benchmarks.

Motivated by the efficacy of the attention mechanism on such generative architectures, we explored potential strategies for further improving the output quality of human pose transformation. We hypothesize that only attending to feature space at the lowest spatial resolution causes a significant loss of finer image details. To mitigate this information loss, we propose an improved end-to-end network architecture [25] by incorporating the attention mechanism at every spatial scale of the encoding and decoding branches. This approach helps the network retain additional image details without requiring multiple cascaded *Pose-Attention Transfer Blocks* [83].

Fig. 3.1 illustrates a general overview of human pose transformation using the proposed method. From a given observation (*condition image*), I_A of a person with initial posture (*condition pose*), P_A, the aim is to generate a realistic image, I_B of the same person corresponding to a novel target pose, P_B.

3.2 Human Pose Transformation with Multi-scale Attention

Assuming a given observation (condition image) I_A^j of a person with pose P_A^j, where j denotes the index in the dataset, pose translation aims to generate a realistic image I_B^j of the same person corresponding to an intended target pose P_B^j. Similar to previous approaches [77, 79, 80, 81, 83], we represent human pose as a set of 2D coordinates of 18 body keypoints, estimated using a pretrained *Human Pose Estimator* (HPE) [154]. The HPE estimates each keypoint as a triplet (x_i, y_i, v_i), where (x_i, y_i) denotes 2D coordinates of the i-th keypoint and v_i is a binary state flag to indicate the visibility condition of that keypoint. Specifically, $v_i = 1$ if the keypoint is visible and $v_i = 0$ for an occluded keypoint. To apply spatial convolution, we construct a $h \times w \times 18$ sparse heatmap from the keypoints, where h and w denote the height and width of the corresponding image, respectively. Each of the 18 channels of the heatmap corresponds to one specific keypoint. For a visible keypoint $(x_k, y_k, 1)$, the respective location (x_k, y_k, k) of the heatmap has a value of 1, and the remaining positions of the k-th channel contain zeros. We denote the sparse pose heatmaps as H_A^j and H_B^j corresponding to P_A^j and P_B^j, respectively.

The proposed generator takes the RGB condition image I_A^j of dimension $h \times w \times 3$ and the channel-wise concatenated pose heatmaps H_A^j and H_B^j of dimension $h \times w \times 36$ as inputs and generates an RGB output image $\hat{I}_B^j$. We employ a PatchGAN discriminator [30] to determine the visual correctness of the generated images. The discriminator takes two channel-wise concatenated RGB images, either (I_A^j, I_B^j) or $(I_A^j, \hat{I}_B^j)$, of dimension $h \times w \times 6$ as input and predicts a binary class probability map for the input patches.

3.2.1 Generator Architecture

Fig. 3.2 shows the architecture of the proposed generator. The generator consists of two downsampling paths (encoders) followed by an upsampling path (decoder) with attention links between feature maps at every underlying resolution level. We begin by describing the essential components of the generator, followed by network design specifications for the encoders, decoder, and attention mechanism.

3.2.1.1 Generator Components

Conv1x1. A point-wise 2D convolution operation that preserves the input size. We perform a single 2D convolution with 1×1 kernel, stride = 1, padding = 0, and without adding a bias.

Conv3x3. A 2D convolution operation that preserves the input size. We perform a single 2D convolution with 3×3 kernel, stride = 1, padding = 1, and without adding a bias.

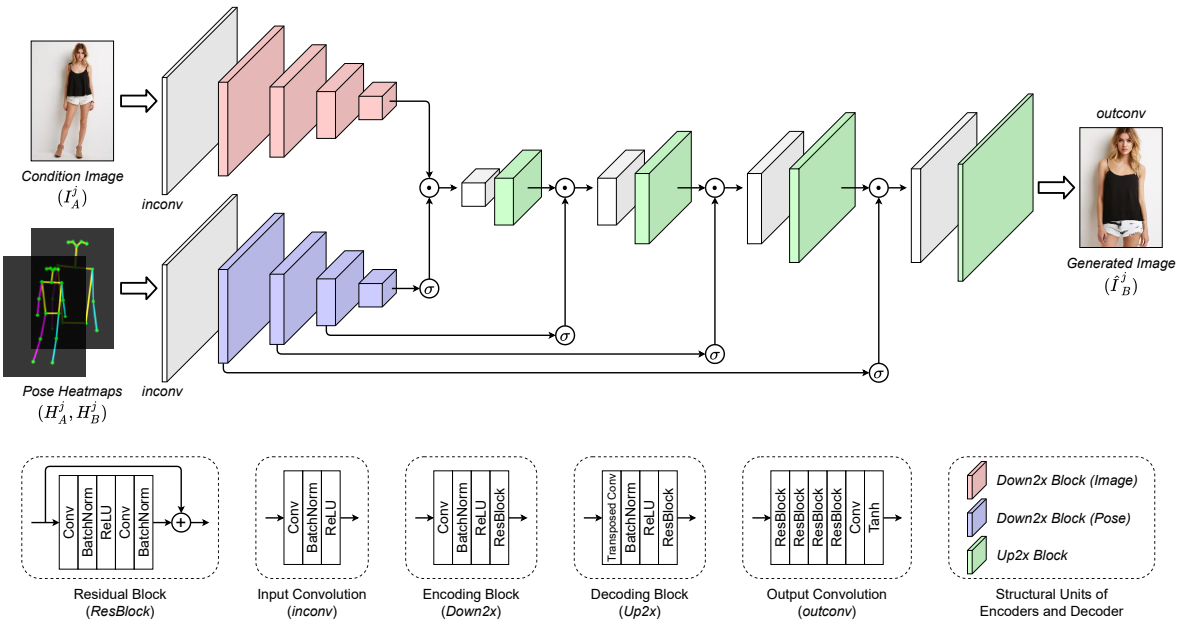

Figure 3.2: Architecture of the proposed generator. The generator takes the condition image I_A^j along with the channel-wise concatenated pose heatmaps (H_A^j, H_B^j) as inputs and generates an estimate $\hat{I}_B^j$ of the target image I_B^j.

Residual Block. A basic residual block [155] that preserves the input size and the number of channels. A residual block is composed of 5 sequential layers – Conv3x3, Batch Normalization [156], ReLU [157], Conv3x3, Batch Normalization. We pass the input through these layers and add the output from the last layer to the original input, producing the final block output.

Down2x Block. A downsampling block in the encoder for compressing the input size by a factor of 2. We perform a 2D convolution with 4×4 kernel, stride = 2, padding = 1, and without adding any bias to downsample the input. The resulting feature maps are passed through sequential layers of Batch Normalization, ReLU, and a Residual Block to produce the final block output.

Up2x Block. An upsampling block in the decoder for expanding the input size by a factor of 2. We perform a 2D transposed convolution with 4×4 kernel, stride = 2, padding = 1, and without adding any bias to upsample the input. The resulting feature maps are passed through sequential layers of Batch Normalization, ReLU, and a Residual Block to produce the final block output.

3.2.1.2 Encoders

The proposed network architecture contains two parallel downstream branches as the encoders. The *image branch* operates on the condition image I_A^j and the *pose branch* operates on the concatenated pose heatmaps $(H_A^j,\ H_B^j)$. Initially, we perform a 2D convolution with *Conv3x3* and project each input to a $h \times w \times N_f$ feature space, where N_f denotes the initial number of feature maps. The convolution is followed by Batch Normalization and ReLU activation to produce the initial input feature maps at each branch. The input feature maps are then passed through N subsequent *Down2x Blocks* at each branch. Each block downscales the input size by a factor of 2 while expanding the number of feature maps by a factor of 2. Therefore, after N consecutive downsampling blocks, we end up with a feature space of dimension $\frac{h}{2^N} \times \frac{w}{2^N} \times N_f \cdot 2^N$.

3.2.1.3 Decoder

The proposed network architecture contains a single upstream branch as the decoder for generating the output image $\hat{I}_B^j$ corresponding to the target pose P_B^j. Starting with a feature space of dimension $\frac{h}{2^N} \times \frac{w}{2^N} \times N_f \cdot 2^N$, we pass the feature maps through N consecutive *Up2x Blocks*. Each block upscales the input size by a factor of 2 while compressing the number of feature maps by a factor of 2. Therefore, after N subsequent upsampling blocks, we end up with a feature space of dimension $h \times w \times N_f$. Finally, the resulting feature maps are passed through 4 consecutive *Residual Blocks* followed by a point-wise 2D convolution operation with *Conv1x1* to project the feature space into an output of dimension $h \times w \times 3$. We apply the hyperbolic tangent activation function $tanh$ on the output tensor to get the normalized output image $\hat{I}_B^j$.

3.2.1.4 Attention Mechanism

The proposed dense multi-scale attention mechanism operates at every spatial scale of the encoding and decoding branches to enhance low-frequency and high-frequency image features in the output. At spatial resolution k, we compute the attention mask M_k by applying an element-wise $sigmoid$ activation function σ on the encoded feature maps of the *pose branch* $H_k^{\mathscr{E}}$. The image feature maps I_k at scale k are updated by performing an element-wise product with the attention mask M_k. The updated image feature maps I_k act as the input to the upsampling block of the decoder at resolution k. We repeat these operations sequentially up to the highest resolution, which results in N such operations.

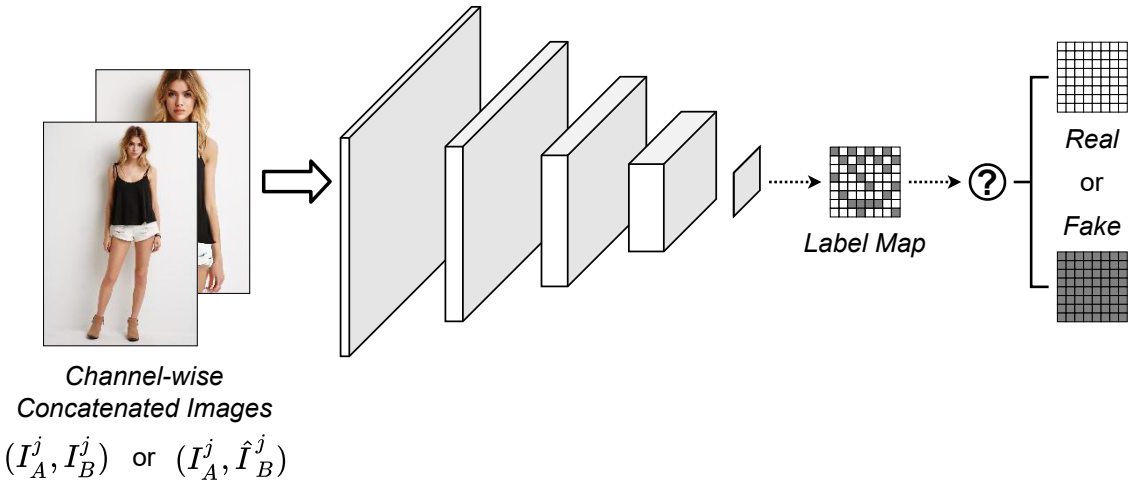

Figure 3.3: Architecture of the PatchGAN discriminator. The discriminator takes two channel-wise concatenated images, either (I_A^j, I_B^j) or $(I_A^j, \hat{I}_B^j)$, as input and estimates a label map, where each label corresponds to the binary class probability of an input patch.

Mathematically, at the lowest resolution level, where $k = N$,

$$I_{N-1}^{\mathscr{D}} = \mathscr{D}_N^{Up2x}\left(I_N^{\mathscr{E}} \odot \sigma\left(H_N^{\mathscr{E}}\right)\right)$$

and for each subsequent higher resolution level where, $k = \{1, ..., N-1\}$,

$$I_{k-1}^{\mathscr{D}} = \mathscr{D}_k^{Up2x}\left(I_k^{\mathscr{D}} \odot \sigma\left(H_k^{\mathscr{E}}\right)\right)$$

Here, we denote downstream encoding as $\mathscr{E}$ and upstream decoding as $\mathscr{D}$.

3.2.2 Discriminator Architecture

We use a Markovian PatchGAN discriminator [30] to estimate high-frequency correctness in the generated images. The discriminator complements the L_1 loss, which assesses low-frequency details only. Such a discriminator operates on $S \times S$ image patches by classifying each patch as *real* or *fake*. As explained by the authors [30], PatchGAN functions as a style/texture loss by modeling the image as a Markov random field, assuming independence between pixels separated by more than a patch diameter.

Our approach enforces adversarial discrimination on the image transition rather than the image itself. We do this by depth-wise concatenating the condition image I_A^j either with the target image I_B^j or with the generated image $\hat{I}_B^j$ where (I_A^j, I_B^j) is labeled as *real* and $(I_A^j, \hat{I}_B^j)$ is labeled as *fake*. We adopt an identical network architecture as [30] that effectively

operates on a 70×70 receptive field (patch) of the input. Fig. 3.3 shows the architecture of the PatchGAN discriminator.

3.2.3 Training Strategy

We train the network in an adversarial scheme with two competing objective functions for generator G and discriminator D. During training, each objective tries to minimize the penalty incurred by self while attempting to maximize the same for the other network.

3.2.3.1 Generator Objective

Mean Absolute Error: We compute point-wise L_1 loss as the Mean Absolute Error (MAE) between the target image I_B^j and the generated image $\hat{I}_B^j$ to ensure low-frequency correctness in the output. Mathematically,

$$\mathscr{L}_1^G = \left\| \hat{I}_B^j \; - \; I_B^j \right\|_1$$

Adversarial Loss: We compute Binary Cross-Entropy (BCE) as an adversarial loss measure using the PatchGAN discriminator to assess high-frequency correctness in the generated images. Mathematically,

$$\mathscr{L}_{GAN}^G = \mathscr{L}_{BCE}\left(D\left(I_A^j, \hat{I}_B^j\right), 1\right)$$

Perceptual Loss: We also include perceptual loss [158] in the generator objective to improve the visual fidelity of the generated images. Mathematically,

$$\mathscr{L}_{P_\rho}^G = \frac{1}{h_\rho w_\rho c_\rho} \sum_{x=1}^{h_\rho} \sum_{y=1}^{w_\rho} \sum_{z=1}^{c_\rho} \left\| \phi_\rho\left(\hat{I}_B^j\right) \; - \; \phi_\rho\left(I_B^j\right) \right\|_1$$

where $\mathscr{L}_{P_\rho}^G$ denotes the perceptual loss computed from the ρ^{th} layer output of a pretrained VGG19 network [159], ϕ_ρ denotes the ρ^{th} layer output with a feature space dimension of $h_\rho \times w_\rho \times c_\rho$. The proposed method retains both coarse and fine details in the generated images by computing perceptual loss at two different layers (4^{th} and 9^{th}) of a VGG19 model pretrained on the ImageNet dataset [56].

The complete generator objective is calculated as a weighted linear combination of the L_1 loss, adversarial loss, and perceptual loss. Mathematically,

$$\mathscr{L}^G = \arg\min_G \max_D \; \lambda_1 \mathscr{L}_1^G \; + \; \lambda_2 \mathscr{L}_{GAN}^G \; + \; \lambda_3 \left(\mathscr{L}_{P_4}^G \; + \; \mathscr{L}_{P_9}^G \right)$$

where λ_1, λ_2 and λ_3 denote the weights for the corresponding loss functions.

3.2.3.2 Discriminator Objective

Adversarial Loss: The discriminator objective has a single adversarial loss component which is calculated as the average BCE loss over a *real* image transition (I_A^j, I_B^j) and a *fake* image transition $(I_A^j, \hat{I}_B^j)$. Mathematically,

$$\mathscr{L}_{GAN}^{D} = \frac{1}{2}\left[\mathscr{L}_{BCE}\left(D\left(I_A^j,\, I_B^j\right),\, 1\right) \;+\; \mathscr{L}_{BCE}\left(D\left(I_A^j,\, \hat{I}_B^j\right),\, 0\right)\right]$$

where we assume the *real* label is 1 and the *fake* label is 0. The complete discriminator objective is given by,

$$\mathscr{L}^{D} = \arg\min_{D}\max_{G}\; \mathscr{L}_{GAN}^{D}$$

3.2.4 Implementation Details

In our implementation, the generator is constructed with 4 downsampling blocks in both the encoders and consequently 4 upsampling blocks in the decoder ($N = 4$). The initial number of feature maps is set to 64 ($N_f = 64$). In the generator objective, we set the weights as, $\lambda_1 = 5$, $\lambda_2 = 1$, and $\lambda_3 = 5$. The generator and discriminator networks consist of 92.2 million and 2.8 million trainable parameters, respectively. We initialize the parameters of both the generator and discriminator before training by sampling from a normal distribution of 0 mean and 0.02 standard deviation. We optimize both generator and discriminator using the stochastic Adam optimizer [160] with learning rate $\eta = 1e^{-3}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$ and weight decay = 0. We train the network on a single NVIDIA TITAN X GPU for 270K iterations with a batch size of 8 and serialize the network weights after every 500 iterations. We select the checkpoint with the best evaluation metrics during inference among the 70 most recent checkpoints.

3.3 Experiments

To evaluate the proposed architecture, we performed extensive qualitative and quantitative comparisons against several previously introduced human pose transformation methods [77, 80, 81, 83] on a fixed dataset [161]. Our method outperforms the previous approaches on most evaluation metrics. We also conducted an opinion-based user study for subjective visual quality assessment. Additionally, we performed an exhaustive ablation study on the network design and explored immediate adoptions of the proposed architecture for other potential use cases as a *drop-in* solution.

3.3.1 Dataset

We perform all experiments on the DeepFashion *In-shop Clothes Retrieval* dataset [161] that contains 176 × 256 high-quality isolated person images centered on 256 × 256 square grids. The dataset features wide variations in physical traits, outfits, and poses, making it suitable for evaluating and comparing human pose transformation techniques. For a direct and fair comparison, we adopt the exact train-test split provided by Zhu *et al.* [83], where 101,966 image pairs are selected randomly for training and 8,570 image pairs for testing. The identities of persons in the training set do not overlap with those in the testing set to ensure better generalization.

3.3.2 Evaluation Metrics

Currently, a quantifiable generalized metric for visual image quality assessment is an open problem in computer vision. Previous authors [77, 80, 81, 83] assessed visual quality using Structural Similarity Index (SSIM) [162], Inception Score (IS) [163], Detection Score (DS) [164], and Percentage of Correct Keypoints (PCKh) [165]. **SSIM** measures the perceived quality of generated images by comparing them with real images and considering image degradation as the perceived change in structural information. **IS** uses the Inception architecture [166] as an image classifier to estimate the Kullback-Leibler (KL) divergence [167] between label and marginal distributions for a large set of images. **DS** uses a pretrained object detector to estimate target class recognition confidence of the object detection model as a measure of perceptual quality. **PCKh** aims to quantify the shape consistency between generated and real human images by estimating the percentage of correctly aligned keypoints. Additionally, we measured the Learned Perceptual Image Patch Similarity (LPIPS) [168] as a modern standard for image quality assessment. **LPIPS** quantifies the perceptual similarity between real and generated images using spatial feature maps obtained from a pretrained backbone network, such as VGG19 [159] and SqueezeNet [169] in our experiments.

3.3.3 Qualitative and Quantitative Comparison

Fig. 3.4 shows a qualitative comparison among previously proposed major human pose transformation algorithms [77, 80, 81, 83] and our method. The visual comparison demonstrates that images generated by the proposed strategy retain better skin tone, hairstyle, facial hair, and limb structure. Additionally, our approach works better than the previous methods in preserving the garment styles and textures. From an apparent visual

Figure 3.4: Qualitative comparison among different human pose transformation methods. I_A^j denotes the *condition image* (observation), I_B^j denotes the *target image* (ground truth), and subsequent columns show the generated images by PG^2 [77], Deformable GAN [80], VUNet [81], PATN [83] and the proposed method.

inspection, images generated by the proposed method also look more realistic than other techniques.

The apparent visual superiority of our method is further reflected in the quantitative evaluation of multiple perceptual metrics. Table 3.1 summarizes the SSIM, IS, DS, PCKh, and LPIPS scores for analytically benchmarking different human pose transformation techniques. To ensure cycle consistency, we generated images in both directions – $\hat{I}_B^j$ from the image pairs (I_A^j, I_B^j) and $\hat{I}_A^j$ from the reverse image pairs (I_B^j, I_A^j). We evaluated each metric on both $\hat{I}_B^j$ and $\hat{I}_A^j$, computing their mean as the final score. For SSIM and IS, we

Table 3.1: Quantitative comparison among different human pose transformation methods. The best scores are in **bold**, and the second-best scores are underlined.

Method	SSIM ↑	IS ↑	DS ↑	PCKh ↑	LPIPS (VGG19) ↓	LPIPS (SqueezeNet) ↓
PG^2 [77]	**0.773**	3.163	0.951	0.89	0.523	0.416
Deform [80]	0.760	3.362	0.967	0.94	-	-
VUNet [81]	0.763	**3.440**	0.972	0.93	-	-
PATN [83]	**0.773**	3.209	**0.976**	0.96	0.299	0.170
Ours	0.769	3.379	**0.976**	**0.98**	**0.200**	**0.111**
Ground Truth	1.000	3.864	0.974	1.00	0.000	0.000

obtained slightly lower scores than the best results. Our method achieved the best scores for DS, PCKh, and LPIPS. We improved the PCKh score over PATN [83] by 2%, indicating superior shape consistency due to better keypoint alignment. We estimated LPIPS using two different backbones – VGG16 [159] and SqueezeNet [169], achieving significantly better scores in each case, indicating superior perceptual quality over other methods.

3.3.4 User Study

Although the aforementioned metrics for quantitative evaluation of visual quality are widely adopted in the literature, a noticeable amount of outliers show the fundamental limitations of such benchmarks in quantifying perceptual quality. This problem leaves human perception as the most reliable way to assess image quality. Therefore, we performed an opinion-based subjective user study to analyze the visual quality of generated images as perceived by the human vision system. The study consists of two manual tasks – a constrained test and an unconstrained test. In the constrained test, users discriminate between *real* and *fake* images by visual inspection within a fixed amount of time. In the unconstrained test, users perform similar discrimination without any time limit. We followed similar protocols as [77, 80, 83] for the constrained test except for the allowed observation time for each image. We increased the observation time from 1 to 5 seconds, allowing users additional time for better visual inspection before making a decision. This protocol explicitly puts the proposed method in a more challenging position than the previous approaches [77, 80, 83].

We selected 130 *real* and 130 *fake* (generated) images for our user study. 10 images from each set are used as the fixed practice samples. A user is shown 20 images during a test, where 10 images are drawn randomly from each set of the remaining 120 images. For every anonymous user submission, the fraction of *real* images identified as *generated* (**R2G**) and the fraction of *generated* images identified as *real* (**G2R**) are recorded. The final

Table 3.2: Evaluation scores of the user study on different human pose transformation methods. The best scores are in **bold**, and the second-best scores are underlined.

Method	Exposure Time (second)	R2G (%) ↑	G2R (%) ↑	Accuracy (%) ↓
PG^2 [77]	1.0	9.20	14.90	87.95
Deform [80]	1.0	12.42	24.61	81.49
PATN [83]	1.0	19.14	31.78	74.54
Ours	5.0	25.90	**54.26**	**59.92**
Ours	∞	**30.37**	46.30	61.67

score is calculated as the mean of the global aggregate of all submissions. In Table 3.2, we show the evaluation scores of our user study conducted with 61 individuals along with the scores reported in previous works [77, 80, 83]. Even with higher observation time for better visual inspection, our method exhibits significantly higher R2G and G2R scores and, consequently, much lower recognition accuracy among the users. These results further imply that the images generated by our method are visually more realistic than other methods [77, 80, 83], leading to higher confusion among the users during the study.

3.3.5 Ablation Study

The core design characteristic of the proposed network architecture is focused on the attention mechanism at different underlying resolution levels of the generator, as shown in Fig. 3.2. To study the efficacy of such architecture, we performed an ablation study with 4 generator variants. The first variant (**A0**) is a generic encoder-decoder architecture that does not use any attention operation. The second variant (**A1-LR**) uses a single attention operation at the lowest feature resolution. The third variant (**A1-HR**) also uses a single attention operation but at the highest feature resolution. The fourth variant (**FULL**) is equivalent to the proposed architecture and performs attention operations at every feature resolution, as illustrated in Fig. 3.2. We trained each network on the same training data for 125K iterations, keeping the discriminator, training mechanism, and all other implementation conditions the same, as discussed in Sec. 3.2.

Fig. 3.5 shows a qualitative comparison among the generated images by different network variants. Model **A0** generates blurry and visually inconsistent images. Model **A1-LR** generates visually and structurally consistent images but lacks realistic detail in the face and limbs. Model **A1-HR** performs poorly, resulting in blurry and incomplete images. Model **FULL** performs significantly better than other variants, producing realistic images with finer detail while preserving visual and structural consistencies.

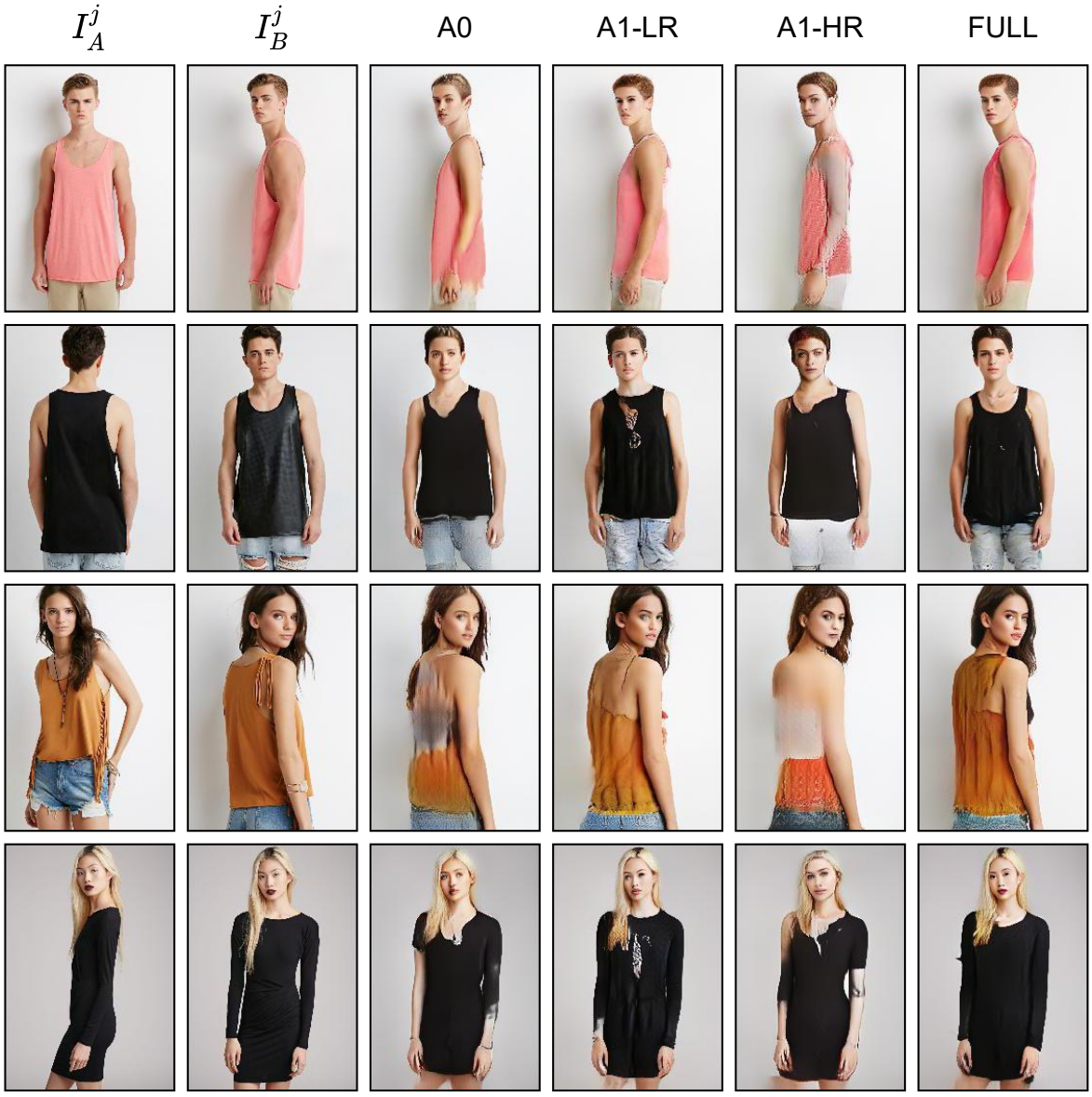

Figure 3.5: Qualitative ablation analysis among different network variants. I_A^j denotes the *condition image* (observation), I_B^j denotes the *target image* (ground truth), and subsequent columns show the generated images by different network architectures.

We conducted a quantitative comparison among different model variants to back the qualitative results of the visual ablation analysis. Table 3.3 summarizes evaluated SSIM, IS, DS, PCKh, and LPIPS scores for all model variants. As expected, model **FULL** achieves the best evaluation scores on most metrics. Interestingly, model **A1-LR** achieves marginally better IS and DS scores than model **FULL**, even with visually inferior generation results, as

Table 3.3: Quantitative ablation analysis among different network variants. The best scores are in **bold**, and the second-best scores are underlined.

Model	SSIM ↑	IS ↑	DS ↑	PCKh ↑	LPIPS (VGG19) ↓	LPIPS (SqueezeNet) ↓
A0	0.760	3.055	0.969	**0.97**	0.221	0.124
A1-LR	0.758	**3.211**	**0.976**	**0.97**	0.210	0.117
A1-HR	0.755	2.859	0.965	0.95	0.229	0.134
FULL ***(proposed)***	**0.764**	3.171	0.975	**0.97**	**0.204**	**0.113**
Ground Truth	1.000	3.864	0.974	1.00	0.000	0.000

evident from Fig. 3.5. This anomaly reiterates the limitations of current perceptual metrics in generalizing human perception. For this reason, such comparative studies use multiple metrics to determine the optimal solution.

From the visual and analytical ablation studies, we conclude that the proposed attention mechanism at multiple scales of the generator architecture significantly improves the visual quality in generated images. Consequently, these experiments justify the network design philosophy behind the proposed architecture.

3.3.6 Extended Applications

The proposed architecture can function as a potential *drop-in* solution for multiple applications other than human pose transformation. We extended our experiments to address a few such problems by directly integrating the proposed architecture without any fundamental modification. While task-specific modifications of the base architecture can further improve the generative performance, such analyses are beyond the scope of this thesis. Regardless, we show that even without any modification, the proposed network performs remarkably well in many practical applications other than human pose transformation, providing a general *drop-in* solution to these problems.

3.3.6.1 Semantic Reconstruction

In novel view synthesis, one typical solution is image reconstruction from semantic maps. Unlike an unconditional approach, a conditional reconstruction uses a known observation (*condition image*) as a reference to transfer image attributes to the target view. We used finely annotated parsing masks from the DeepFashion *In-shop Clothes Retrieval* dataset [161], where each mask contains up to 16 semantic labels for different body or garment parts. Analogous to the keypoint-based approach, a semantic map is represented as a 16-channel heatmap, where each channel corresponds to the binary mask of one specific

Figure 3.6: Qualitative results of semantic reconstruction using the proposed method.

body or garment part. We selected 5,161 training image pairs and 795 testing image pairs in our experiments. The network is trained for 65K iterations while keeping all other implementation conditions the same, as discussed in Sec. 3.2. After training, the generative network can successfully reconstruct a target semantic mask by transferring visual attributes from the observed reference image. Fig. 3.6 demonstrates a few qualitative results of such conditional semantic reconstruction by the proposed method.

3.3.6.2 Virtual Try-On

Although the proposed network is not originally intended for virtual try-on, the ability of conditional semantic reconstruction facilitates extending the architecture to such applications. The main objective of a virtual try-on task is to replace selected parts of the attire of a target person with that of a reference person. We address this in two sequential steps. First, we reconstruct the target semantic map using the reference image. Second, the initially reconstructed image is refined by bitwise operations using the target image and a partial binary mask with selected parts of the attire. Mathematically,

$$I_{fine} = \left[M_{parts} \odot I_{coarse}\right] \oplus \left[\left(1 - M_{parts}\right) \odot I_{target}\right]$$

Figure 3.7: Qualitative results of virtual try-on using the proposed method.

where I_{fine} denotes the refined target person image with replaced attire, I_{coarse} denotes the reconstructed target person image, I_{target} denotes the target person image with original attire, and M_{parts} denotes a binary mask of selected parts of the attire. Fig. 3.7 demonstrates a few qualitative results of virtual try-on using this approach.

3.3.6.3 Font Style Transfer

As a part of our earlier work [170] on the *Scene Text Editing* (STE) problem, we used two independent networks for preserving structural and color consistencies in a character-to-character transformation. Although the proposed architecture aims toward human pose transformation, the technique can remarkably improve the generative performance of such two-stage font style transfer schemes by leveraging multi-scale attention and the end-to-end network with skeleton supervision. The learning model receives the structural reference of a character as a single-channel binary image tensor of its skeleton. As the font properties are unknown during inference, we trained the model using target character

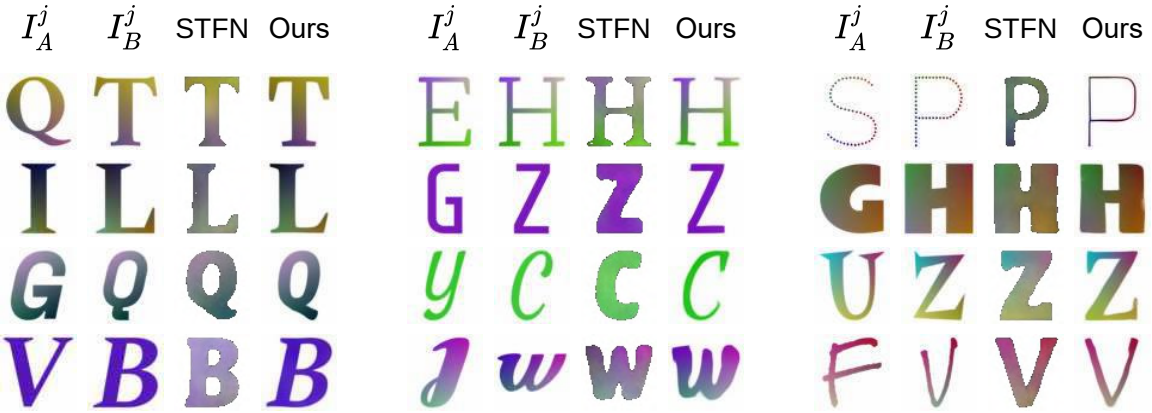

Figure 3.8: Qualitative comparison of the proposed method against a dedicated font style transfer technique. I_A^j denotes the *condition image* (observation), I_B^j denotes the *target image* (ground truth), and subsequent columns show the generated images by STEFANN [170] (abbreviated as STFN) and our method with *skeleton supervision*.

Table 3.4: Quantitative comparison of the proposed method against a dedicated font style transfer technique, STEFANN [170]. The best scores are in **bold**.

Method ↑	**SSIM ↑**	**PSNR (dB) ↑**	**LPIPS (VGG19) ↓**	**LPIPS (SqueezeNet) ↓**
STEFANN [170]	0.450	13.347	0.397	0.273
Ours	**0.638**	**16.876**	**0.208**	**0.089**
Ground Truth	1.000	∞	0.000	0.000

skeletons from a fixed font, providing a coarse geometric context of the target character to the generator. In our experiments, we trained the network on a small subset of STEFANN [170]. We randomly selected 200 image pairs of uppercase characters from both training and testing sets. This led to 203000 image pairs from 1015 fonts for training and 60000 image pairs from 300 fonts for testing, with a random color applied for each font. The skeletons are estimated from the binarized character image by applying Gaussian blur with a 3×3 kernel followed by a parallel thinning algorithm [171]. We trained the network for 65K iterations, keeping all other implementation conditions the same, as discussed in Sec. 3.2. Fig. 3.8 illustrates a qualitative comparison between STEFANN [170] and the proposed approach for font style transfer. Even with over 70% lesser training data, the end-to-end skeleton-guided approach can preserve structural and color consistencies significantly better than the two-stage pipeline. This visual analysis is backed by evaluation scores on multiple analytical metrics, as summarized in Table 3.4. In another recent technique [172], the authors directly adopted the proposed architecture for word-level text transformation.

3.3.7 Limitations

Impact of inaccurate pose estimation: Similar to previous keypoints-based human pose transformation methods [77, 79, 80, 81, 83], the generative performance of the proposed method directly depends on the HPE model [154]. Inaccurately predicted keypoints lead to significant visual deformations in the generated images. Fig. 3.9 shows a few examples where the proposed method fails to generate realistic outputs.

Figure 3.9: Limitations of the proposed method due to inaccurate pose estimation.

Impact of physical stature: A keypoint-based pose representation allows structurally accurate person image generation corresponding to the intended target pose. To facilitate such geometrically conditioned pose transformations, the dataset [161] consists of source-target image pairs for every individual, ensuring that the *condition image* and *target image* correspond to a single person. Such data samples guarantee structural consistency between the *condition pose* and *target pose* as they correspond to one specific human body. However, in a realistic inference scenario, individual identities of *condition image* and *target image* are different. This often leads to noticeable shape ambiguities in generated images when the *target pose* has a substantially different physique than the *condition pose*. The problem is not immediately apparent from the initial visual analysis using test

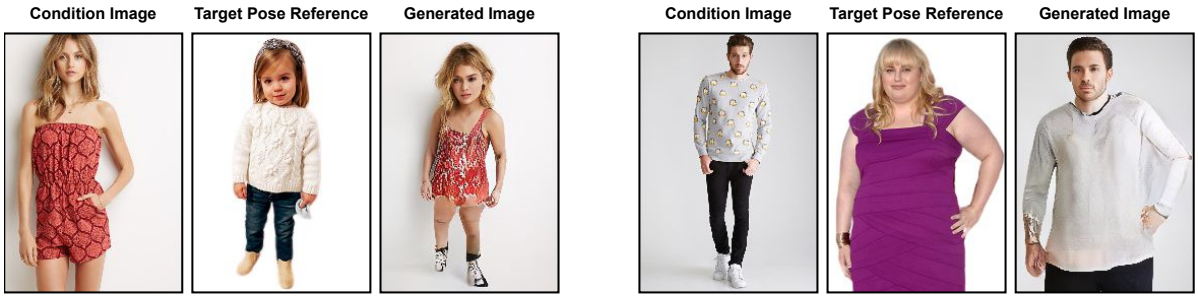

Figure 3.10: Limitations of the proposed method due to different physical statures of *condition pose* and *target pose* references.

set samples of the dataset because of the aforementioned identity preservation. Fig. 3.10 illustrates this problem by drawing the target reference from an individual with a visibly different physique.

3.4 Chapter Summary

In this chapter, we introduced a multi-scale attention strategy with an end-to-end network architecture for structurally guided human pose transformation. At the conceptual level, the proposed method performs a single attention operation at each feature resolution to improve both low-frequency and high-frequency details in the generated images. Experimental studies show that this approach outperforms the previous pose transformation methods on multiple visual and analytical benchmarks. Additionally, the proposed architecture works remarkably well for several other generative tasks, providing a general *drop-in* solution to these problems. However, similar to other keypoint-guided pose transformation approaches, the proposed method is also associated with structural ambiguity due to a high geometric bias towards the target pose. This problem is particularly noticeable when the physical statures of the *observed subject* and *target pose reference* are widely different. In the next chapter, we explore a potential strategy to mitigate such structural bias by adopting a more general pose representation instead of keypoint-based strict geometric guidance.

CHAPTER 4

TEXTUALLY SUPERVISED HUMAN POSE TRANSFORMATION

In Chapter 3, we have shown that the keypoint-based sparse pose representations work incredibly well for human pose transformation. However, the approach has a significant limitation when the physical statures of the *observed subject* and *target pose reference* are substantially different. This chapter investigates the shortcomings of a keypoint-guided pose transformation strategy and explores a more general pose representation to mitigate these issues. In particular, we used textual descriptions to describe a human pose. The proposed architecture consists of three independent networks for (a) text-to-pose translation, (b) pose refinement, and (c) pose transformation. Additionally, we compiled a new dataset for benchmarking text-guided person image generation techniques. Experimental studies show promising visual and analytical performance by the proposed method.

4.1 Addressing Structural Ambiguities in Human Pose Transformation

The goal of human pose transformation is to generate a *target image* I_B of a person with an intended *target pose* P_B from a given *condition image* (observation) I_A of that person having a *condition pose* P_A. A human pose P is commonly expressed as a set of body-joint locations (*keypoints*) K. As the geometric locations of the keypoints can vary significantly from person to person depending on the physical stature, two spatially different sets of

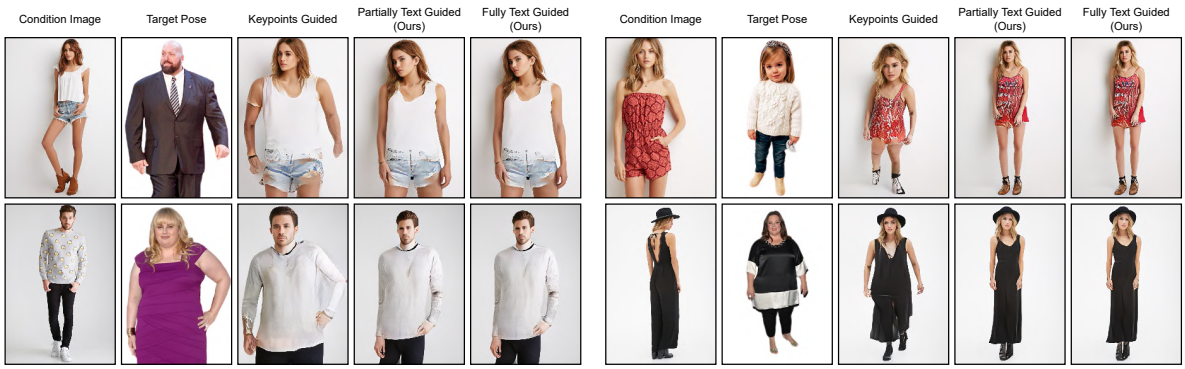

Figure 4.1: An overview of the proposed approach. Keypoint-guided methods tend to produce structurally inconsistent images when the physical stature of the *target pose reference* significantly differs from the *condition image* (observed subject). The proposed text-guided technique successfully addresses this issue while retaining the ability to generate visually appealing results close to the keypoint-guided baseline.

keypoints K and K' may represent the same pose P. Generally, we estimate K_B directly from I_B to represent P_B for training and evaluation. However, as I_B is unknown during inference, P_B has to be estimated from K'_B of a different person image I'_B, eventually creating a dilemma. One way to circumvent the problem is training the model to adapt to *target pose* P_B, represented by keypoints K'_B, estimated from the image I'_B of some other person. However, constructing such datasets is significantly challenging.

To address this problem, we adopted an alternative representation of the human pose using descriptive textual annotations. Initially, we estimated the target keypoints K_B from the textual description T_B of the target pose P_B. The estimated keypoints K_B are then used to generate the output image $\tilde{I}_B$ through a similar pose transformation technique discussed in Chapter 3. As the estimation of K_B is directly conditioned on T_B, computing P_B does not directly depend on I_B anymore, and so the supervision becomes free from structural bias.

Fig. 4.1 illustrates the limitations of existing keypoint-guided methods of human pose transformation. The keypoint-guided techniques tend to follow the body structure of the *target pose reference* rather than the intended *condition image subject*. Thus, they occasionally fail to produce convincing results when the physical statures of I_A and I_B are significantly different. In contrast, the proposed method adopts descriptive textual annotations for a more generalized pose representation, mitigating the strong geometric bias towards the *target pose reference*.

The main contributions of the proposed method are summarized as follows.

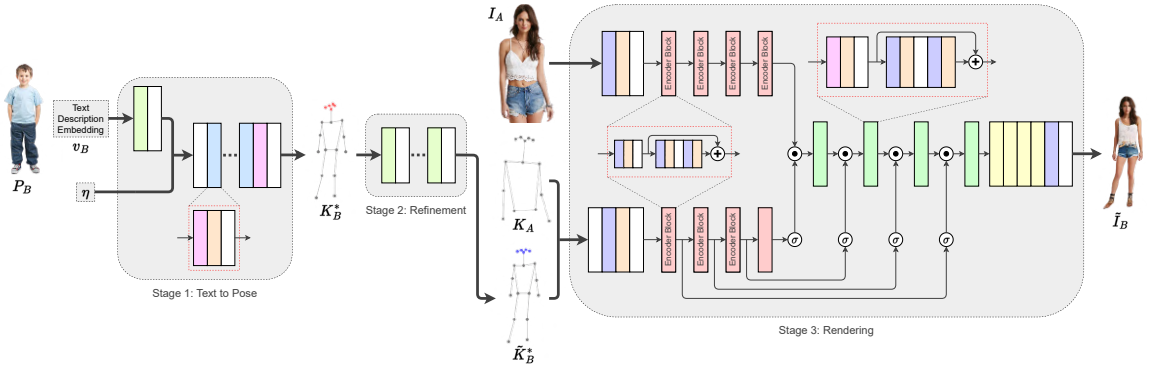

Figure 4.2: Architecture of the proposed pipeline. The workflow consists of three stages. **Stage 1** estimates a spatial representation K_B^* for the target pose P_B from the corresponding text description embedding v_B. **Stage 2** regressively refines the initial estimation of the face keypoints to obtain more accurate target keypoints $\tilde{K}_B^*$. **Stage 3** performs human pose transformation on the observed condition image I_A having pose P_A represented by keypoints K_A to generate the final output image $\tilde{I}_B$.

- We proposed a human pose transformation strategy that takes an image of a person and the textual description of an intended pose to generate a novel instance of that person at the intended pose. To the best of our knowledge, this is one of the earliest attempts of conditional person image generation from textual supervision to address the structural bias in strict keypoint-guided approaches.

- We compiled a new dataset for benchmarking text-guided person image generation methods, containing descriptive pose annotations for 40488 human images.

4.2 Textual Supervision for Human Pose Transformation

The proposed architecture consists of three independently learnable sequential stages for (a) text-to-pose translation, (b) pose refinement, and (c) pose transformation. In the first stage, an initial coarse estimation of the target pose is derived from the corresponding text description embedding. The coarse pose is then refined through regression in the next stage. Lastly, the final output is generated by performing a pose transformation conditioned on the appearance of the observed subject. Fig. 4.2 shows the architecture of the proposed generative pipeline.

4.2.1 Text-to-Pose Translation

For a given person image I_A, the aim is to generate the output image I_B of the same person, where the target pose P_B is represented by a textual description T_B. We encode T_B into an embedded vector v_B either by many-hot encoding or using a pre-trained language model, such as Word2Vec [173], FastText [174], or BERT [175]. First, we estimate the keypoints K_B from v_B using a generative model to provide structural guidance to the pose transformation network in a later stage. To train such a generative model, we represent the keypoints as $k_j \in \mathbb{R}^{m \times n}$, where $k_j \in K; \forall\, j$, and the domain of spatial dimensions for both I_A and I_B is $\mathbb{R}^{m \times n}$. As small spatial perturbations in k_j do not change P_B substantially, we express the pose with a Gaussian distribution $\mathcal{N}(k_j, \sigma_j); \forall\, j$ for reducing the high sparsity in the pose representation. Although for different k_j, the invariance of the pose is valid for different amounts of spatial deviations, we can assume $\sigma_j = \sigma$, a constant, $\forall\, j$, if σ_j is small. This way, we represent keypoint-based pose representations as Gaussian *heatmaps*.

Taking motivation from [131], we designed a *Wasserstein Generative Adversarial Network* (WGAN) to estimate the target keypoints K_B from the text embedding v_B. In our generator G_T, we first project v_B into a 128-dimensional latent space ϕ_B using a linear layer with leaky ReLU activation. To allow some structural variations in the generated poses, we sample a 128-dimensional noise vector $\eta \sim \mathcal{N}(\mathbf{0}, I)$, where I is a 128×128 identity matrix. Both ϕ_B and η are linearly concatenated and passed through 4 up-convolution blocks. At each block, we perform a transposed convolution followed by batch normalization [156] and ReLU activation [157]. The four transposed convolutions use 256, 128, 64, and 32 filters, respectively. We produce the final output from G_T by passing the output of the last up-convolution block through another transposed convolution layer with 18 filters and $tanh$ activation. The final generator output $G_T(v_B, \eta)$ has a spatial dimension of $64 \times 64 \times 18$, where each channel represents one of the 18 keypoints $k_j,\ j \in \{1, 2, \ldots, 18\}$. In our discriminator (*critic*) D_T, we first perform 4 successive convolutions, each followed by leaky ReLU activation, on the 18-channel heatmap. The four convolutions use 32, 64, 128, and 256 filters, respectively. The output of the last convolution layer is concatenated with 16 copies of ϕ_B arranged in a 4×4 tile. The concatenated feature map is then passed through a point convolution layer with 256 filters and leaky ReLU activation. We estimate the final scalar output from D_T by passing the feature map through another convolution layer with a single filter. We mathematically define the objective function for D_T as follows.

$$L_D \;=\; -\,\mathbb{E}_{(x, v_B) \sim p_t, \eta \sim p_\eta} \left[D_T(x, v_B) \;-\; D_T\big(G_T(\eta, v_B), v_B\big) \right]$$

where $(x, v_B) \sim p_t$ is the heatmap and text embedding pair sampled from the training

set, $\eta \sim p_\eta$ is the noise vector sampled from a Gaussian distribution, and $G_T(\eta, v_B)$ is the generated heatmap for the given text embedding v_B. Researchers [176] have shown that the WGAN training is more stable if D_T is Lipschitz continuous, which mitigates the undesired behavior due to gradient clipping. To enforce the Lipschitz constraint, we compute the gradient penalty as follows.

$$\mathscr{C}_T \;=\; \mathbb{E}_{(\tilde{x},v_B)\sim p_{\tilde{x},v_B}}\left[\left(\left\|\nabla_{\tilde{x},v_B} D_T\left(\tilde{x},v_B\right)\right\|_2 \;-\; 1\right)^2\right]$$

where $\|.\|_2$ indicates the L_2-norm and $\tilde{x}$ is an interpolated sample between a real sample x and a generated sample $G_T(\eta, v_B)$, i.e., $\tilde{x} = \alpha x + (1-\alpha)G_T(\eta, v_B)$, where α is a random number, selected from a uniform distribution between 0 and 1. The above equation enforces the Lipschitz constraint by restricting the gradient magnitude to 1. We define the overall objective of D_T by combining the above two equations for L_D and $\mathscr{C}_T$ as follows.

$$L_{D_T} \;=\; L_D \;+\; \lambda \mathscr{C}_T$$

where λ is a regularization constant. We keep $\lambda = 10$ in all of our experiments. We mathematically define the objective function for G_T as follows.

$$\begin{aligned} L_{G_T} \;=\; & -\mathbb{E}_{\eta\sim p_\eta, v_B\sim p_{v_B}}\left[D_T\left(G_T\left(\eta, v_B\right), v_B\right)\right] \\ & -\mathbb{E}_{\eta\sim p_\eta, v_B^1, v_B^2\sim p_{v_B}}\left[D_T\left(G_T\left(\eta, \frac{v_B^1+v_B^2}{2}\right), \frac{v_B^1+v_B^2}{2}\right)\right] \end{aligned}$$

where $v_B^1, v_B^2 \sim p_{v_B}$ are text encodings sampled from the training set. The second term in the equation helps the generator learn from the interpolated text encodings, which are not originally present in the training set. We estimate the target keypoints K_B^* from the 18-channel heatmap generated from G_T by computing the maximum activation ψ_j^{max}, $j \in \{1,2,\ldots,18\}$ for every channel. The spatial location of the maximum activation for the j-th channel determines the coordinates of the j-th keypoint if $\psi_j^{max} \geq 0.2$. Otherwise, the j-th keypoint is considered occluded if $\psi_j^{max} < 0.2$.

4.2.2 Face Keypoints Refinement

While G_T produces a reasonable estimate of the target keypoints from the corresponding textual description, the estimation K_B^* is often noisy. The spatial perturbation is most prominent for the face keypoints (nose, two eyes, and two ears) due to their proximity. Slight positional variations for other keypoints generally do not drastically affect the pose

representation. Therefore, we refine the initial estimate of the face keypoints by regression using a linear fully connected network *RefineNet* N_R. At first, the five face keypoints k_i^f, $i \in \{1, 2, \ldots, 5\}$ are translated by $(k_i^f - k_n)$, where k_n is the spatial location of the nose. In this way, we align the nose with the origin of the coordinate system. Then, we normalize the translated face keypoints such that the scaled keypoints k_i^s are within a square of span ± 1 and the scaled nose is at the origin (0, 0). Next, we flatten the coordinates of the five normalized keypoints to a 10-dimensional vector v_f and pass it through three linear fully connected layers, where each layer has 128 nodes and ReLU activation. The final output layer of the network consists of 10 nodes and $tanh$ activation. While training, we augment k_i^s with small amounts of random 2D spatial perturbations and try to predict the original values of k_i^s. We optimize the parameters of N_R by minimizing the mean squared error (MSE) between the actual and the predicted coordinates. Finally, we denormalize and retranslate the predicted face keypoints. The refined set of keypoints $\tilde{K}_B^*$ is obtained by updating the coordinates of the face keypoints of K_B^* with the predictions from *RefineNet*.

4.2.3 Pose Transformation

We use the previously proposed multi-scale attention-guided pose transformation technique [25] to generate the output image, as discussed in Sec. 3.2. The network takes an existing observation of the person I_A as the *condition image* and the channel-wise concatenated pose heatmaps $(H_A, \tilde{H}_B^*)$ derived from respective keypoints K_A and $\tilde{K}_B^*$, to produce a final output image $\tilde{I}_B$.

4.2.4 Implementation Details

In stage 1, the text-to-pose conversion network uses the stochastic Adam optimizer [160] to train both G_T and D_T, containing 1.8 million and 0.8 million trainable parameters, respectively. We keep learning rate $\eta_1 = 1e^{-4}$, $\beta_1 = 0$, $\beta_2 = 0.9$, $\epsilon = 1e^{-8}$, and weight decay = 0 for the optimizer. While training, we update G_T once after every 5 updates of D_T. In stage 2, we optimize 35722 trainable parameters of the face keypoints refinement network (*RefineNet*) N_R using stochastic gradient descent, keeping learning rate $\eta_2 = 1e^{-2}$. In stage 3, the pose transformation network also uses the Adam optimizer to train both G_S and D_S, containing 92.2 million and 2.8 million trainable parameters, respectively. In this case, we keep learning rate $\eta_3 = 1e^{-3}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$, and weight decay = 0. Before training, the parameters of G_T, D_T, G_S, and D_S are initialized by sampling from a normal distribution of 0 mean and 0.02 standard deviation. We train the individual networks on

a single NVIDIA TITAN X GPU. The text-to-pose conversion network is trained for 100K iterations with a batch size of 8. The face keypoints refinement network is trained for 100 epochs with a batch size of 128. The pose transformation network is trained for 270K iterations with a batch size of 8.

4.3 Experiments

We performed exhaustive qualitative and quantitative analyses of the proposed technique against existing keypoint-guided [83] and text-guided [130] baselines. Additionally, we conducted extensive ablation studies on the network architecture to justify the proposed network design specifications.

4.3.1 Dataset

As this is one of the earliest attempts to perform a text-guided human pose transformation, we introduced a new dataset named *DeepFashion Pose Annotations and Semantics* (DF-PASS) to compensate for the lack of similar public datasets. DF-PASS contains text-based descriptive pose annotations for 40488 human images from the DeepFashion dataset [161]. Each text annotation contains brief descriptions of (1) gender, (2) visibility states of each keypoint, (3) head and face orientations, (4) body orientation, (5) hand and wrist pose, and (6) leg pose. Fig. 4.3 illustrates the format of the many-hot pose encoding recorded during annotation. We recruited five in-house annotators to acquire the text descriptions, which two independent verifiers have validated. Each annotator described a pose during data acquisition by selecting options from a predefined set of possible attribute states. In this way, we collected many-hot embedding vectors alongside the text descriptions. In our experiments, we used 37344 samples for training and 3144 samples for testing out of 40488 annotated images, following the same data split provided by [83].

4.3.2 Evaluation Metrics

Currently, a quantifiable generalized metric for visual image quality assessment is an open problem in computer vision. Following previous authors, we assessed the visual quality of the generated images using Structural Similarity Index (SSIM) [162], Inception Score (IS) [163], Detection Score (DS) [164], Percentage of Correct Keypoints (PCKh) [165], and Learned Perceptual Image Patch Similarity (LPIPS) [168]. In our evaluation, we computed LPIPS with two different backbones – VGG19 [159] and SqueezeNet [169]. Additionally,

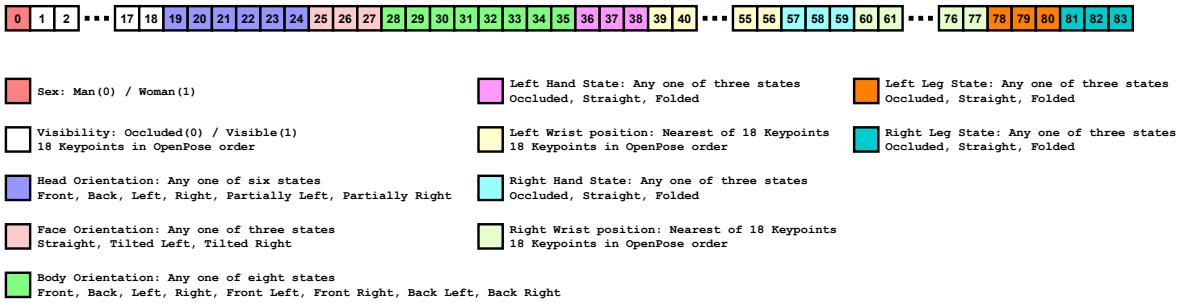

Figure 4.3: The layout of the many-hot encoding vector in the proposed DF-PASS dataset.

we introduced another metric named Gender Consistency Rate (GCR), which evaluates whether the generated image $\tilde{I}_B$ can be identified to be of the same gender as the subject I_A by a pretrained classifier. GCR serves two purposes – first, it ensures that the gender-specific features are present in the generated image, and second, it ensures that the generated target image is consistent with the source image. To calculate GCR, we removed the last layer of the VGG19 network and added a single neuron with sigmoid activation to obtain a binary classifier. Then, we trained the classifier with person images from the DeepFashion dataset [161] by setting binary class labels to male (0) and female (1) samples. After training, the model is used to compute the gender recognition rate for generated images.

4.3.3 Qualitative and Quantitative Comparison

Fig. 4.4 demonstrates the initially estimated keypoints K_B^* from the text description T_B by the text-to-pose generator G_T in stage 1. Although K_B^* captures the pose P_B and closely resembles K_B, a precise observation shows that the face keypoints of K_B^* significantly differ from that of K_B. Fig. 4.5 shows the efficacy of regressive refinement in stage 2, rectifying the spatial perturbations of face keypoints in K_B^* to obtain a more accurate representation $\tilde{K}_B^*$, having a closer resemblance with K_B and consequently P_B. Fig 4.6 illustrates a qualitative comparison of the proposed method against existing keypoint-guided [83] and text-guided [130] baselines. Keypoint-guided methods generate structurally inconsistent results when physical statures of I_A and I_B are widely different, leading to large deviations between K_A/P_A and K_B/P_B. On the other hand, the existing text-guided method often misinterprets the target pose due to a limited set of basic poses used for pose representation. In contrast, the proposed approach does not directly utilize structural information from the images, thereby retaining the intended physical constitution of the subject in the generated images.

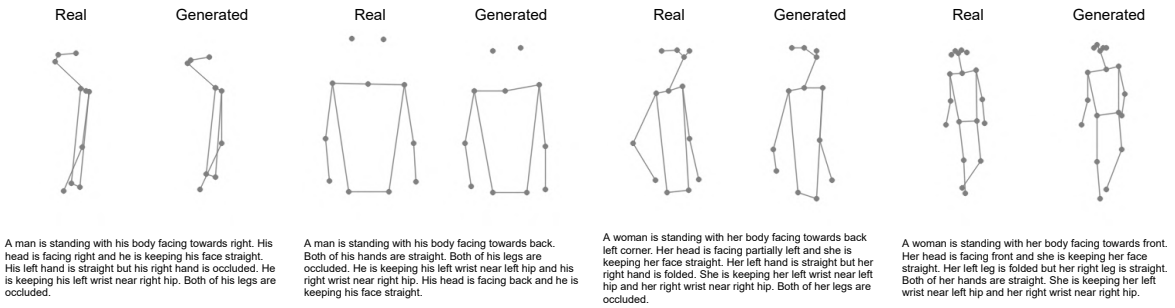

Figure 4.4: Qualitative results of text-to-pose generation using G_T in stage 1.

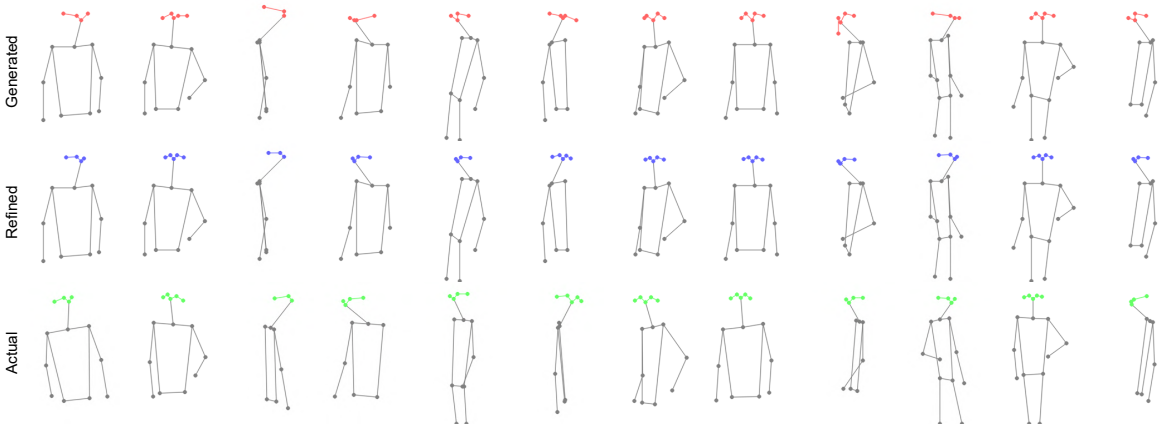

Figure 4.5: Qualitative results of regressive refinement using N_R in stage 2.

Our method performs well, irrespective of the pose representation of the observed subject. A *partially* text-guided approach uses keypoints K_A while a *fully* text-guided approach uses text description T_A to represent observed pose P_A. In both cases, the target pose P_B is represented by the text description T_B.

As the proposed architecture has three major components – (a) text-to-pose translation, (b) face keypoints refinement, and (c) pose transformation, analyzing individual stages is important. Table 4.1 summarizes the evaluation scores of different pose transformation methods on the DeepFashion dataset [161]. The keypoint-guided baseline [83] performs well for such *within-distribution* target poses. However, the proposed text-guided approach also performs satisfactorily, as reflected in SSIM, IS, DS, and LPIPS scores. As PCKh uses keypoint coordinates, our method achieves a low PCKh score compared to the keypoint-based method, which uses precise keypoints for the target image generation. For evaluating *out-of-distribution* target poses, we selected 50 pairs of source and target images from DeepFashion. However, we estimated the target keypoints from random *real-world*

Figure 4.6: Qualitative comparison of the proposed method against existing keypoint-guided [83] and text-guided [130] baselines. I_A denotes the *condition image* (observed subject) and I_B denotes the *target pose reference*. Keypoint-guided methods generate structurally inconsistent results when physical statures of I_A and I_B are widely different. On the other hand, the existing text-guided method often misinterprets the target pose due to a limited set of basic poses used for pose representation. The proposed text-guided technique successfully addresses these issues and generates visually realistic human images with intended physical appearances.

person images (outside DeepFashion), having identical poses as the chosen target images. Table 4.2 shows that the proposed technique achieves significantly higher SSIM and PCKh scores in such cases, indicating substantially better structural alignment.

4.3.4 User Study

Although analytical metrics, such as SSIM, IS, DS, PCKh, and LPIPS, are widely adopted for quantitatively comparing human pose transformation techniques, these metrics do not always agree with human perception. A quantifiable metric for evaluating image quality is an open challenge in computer vision. Therefore, we conducted an additional opinion-based user assessment as a subjective analysis of the generated image quality. Following a

Table 4.1: Quantitative comparison of different human pose transformation methods on ***within-distribution*** target pose references from the **DeepFashion** dataset [161].

Method	**SSIM ↑**	**IS ↑**	**DS ↑**	**PCKh ↑**	**GCR ↑**	**LPIPS ↓ (VGG)**	**LPIPS ↓ (SqzNet)**
PATN [83]	0.773	3.209	0.976	0.96	0.983	0.299	0.170
Zhou et al. [130]	0.373	2.320	0.864	0.62	0.979	0.310	0.215
Partially Text Guided **(ours)**	0.549	3.269	0.950	0.53	0.963	0.402	0.290
Fully Text Guided **(ours)**	0.549	3.296	0.950	0.53	0.963	0.402	0.289
Ground Truth	1.000	3.790	0.948	1.00	0.995	0.000	0.000

Table 4.2: Quantitative comparison of different human pose transformation methods on ***out-of-distribution*** target pose references from the **real-world**.

Method	**SSIM ↑**	**IS ↑**	**DS ↑**	**PCKh ↑**	**GCR ↑**	**LPIPS ↓ (VGG)**	**LPIPS ↓ (SqzNet)**
PATN [83]	0.677	2.779	0.996	0.64	1.000	0.294	0.183
Zhou et al. [130]	0.615	2.891	0.931	0.52	1.000	0.271	0.182
Partially Text Guided **(ours)**	0.696	2.093	0.990	0.84	1.000	0.262	0.155
Fully Text Guided **(ours)**	0.695	2.171	0.991	0.85	1.000	0.263	0.157
Ground Truth	1.000	2.431	0.984	1.00	1.000	0.000	0.000

similar protocol as the previous authors [77, 80, 83], the study records an instant decision from an individual on whether a displayed image is *real* or *generated*. We created a subset of 260 *real* and 260 *generated* images, with 10 images of each type used as a practice set for warm-up. During the test, 20 random images (10 real + 10 generated) are drawn from the remaining images and shown to the examiner. We compute the **R2G** (the fraction of *real* images identified as *generated*) and **G2R** (the fraction of *generated* images identified as *real*) scores from the user submissions. Our method achieved a mean R2G score of 0.5404 and a mean G2R score of 0.6968 for submissions from 156 individual volunteers.

4.3.5 Ablation Study

We performed a set of ablation experiments to analyze the impact of different components on the generative performance of the proposed network architecture.

Impact of text encoder: Table 4.3 summarizes a quantitative comparison of the proposed architecture with different text encoding schemes. The analysis shows that with different text encoders, such as Word2Vec [173], FastText [174], and BERT [175], the model achieves similar benchmark scores as with the baseline many-hot representation. The results indicate that the proposed text-to-pose translation network is robust to most text encoders.

Table 4.3: Quantitative ablation analysis on different text encoding methods.

Text Encoding Method	**SSIM ↑**	**IS ↑**	**DS ↑**	**PCKh ↑**	**GCR ↑**	**LPIPS ↓ (VGG)**	**LPIPS ↓ (SqzNet)**
Word2Vec [173]	0.550	3.251	0.949	0.52	0.973	0.401	0.289
FastText [174]	0.548	3.275	0.949	0.52	0.968	0.399	0.285
BERT [175]	0.549	3.269	0.950	0.53	0.963	0.402	0.290
Many-hot	0.558	3.228	0.953	0.60	0.970	0.388	0.274
Ground Truth	1.000	3.790	0.948	1.00	0.995	0.000	0.000

Table 4.4: Quantitative ablation analysis on face keypoints refinement.

Target Pose	**Refinement**	**SSIM ↑**	**IS ↑**	**DS ↑**	**PCKh ↑**	**GCR ↑**	**LPIPS ↓ (VGG)**	**LPIPS ↓ (SqzNet)**
Keypoints	✘	0.545	3.221	0.952	0.53	0.960	0.404	0.290
Keypoints	✔	0.549	3.269	0.950	0.53	0.963	0.402	0.290
Text description	✘	0.545	3.261	0.952	0.53	0.960	0.404	0.290
Text description	✔	0.549	3.296	0.950	0.53	0.963	0.402	0.289
Ground Truth	**n/a**	1.000	3.790	0.948	1.00	0.995	0.000	0.000

Table 4.5: Quantitative ablation analysis on the attention mechanism.

Attention Mechanism	**SSIM ↑**	**IS ↑**	**DS ↑**	**PCKh ↑**	**GCR ↑**	**LPIPS ↓ (VGG)**	**LPIPS ↓ (SqzNet)**
Single-scale attention	0.540	3.170	0.921	0.54	0.954	0.415	0.298
Multi-scale attention	0.549	3.269	0.950	0.53	0.963	0.402	0.290
Ground Truth	1.000	3.790	0.948	1.00	0.995	0.000	0.000

Impact of refinement network: Table 4.4 summarizes a quantitative comparison of the proposed architecture *with* and *without* face keypoints refinement. Although the improvement seems marginal from the evaluation scores, the visual overhaul is remarkable, as illustrated in Fig. 4.7. Without a dedicated refinement mechanism, borderline variations in spatial positions of the initially estimated face keypoints often lead to extreme distortions. Therefore, refinement is desirable in the proposed architecture.

Impact of attention mechanism: We also analyzed the efficacy of *multi-scale attention* in the pose transformation network [25]. Unlike *multi-scale attention*, where the attention operations are performed at every feature resolution, for *single-scale attention*, a single attention operation is performed only at the lowest feature resolution. Table 4.5 summarizes the evaluation scores with both strategies. The analysis shows substantially superior performance with *multi-scale attention*, retaining both low-frequency and high-frequency details in the generated images.

Figure 4.7: Qualitative ablation analysis of the proposed method *with* and *without* face keypoints refinement. Without a dedicated refinement mechanism, marginal variations in spatial positions of the estimated keypoints often lead to undesired distortions.

4.3.6 Limitations

To the best of our knowledge, the proposed work is one of the earliest attempts to mitigate structural bias in human pose transformation using textual supervision. Although our approach generates visually compelling results in our experiments, the quality of the generated images degrades in some specific scenarios. One potential reason behind the problem is the lack of sufficient fine-grained details in the textual description of the target pose, leading to incorrect interpretation by the text-to-pose generator. Another probable reason for performance bottlenecks is error accumulation in the proposed multi-stage architecture. Therefore, any erroneous estimation in an earlier stage can propagate to later stages, impacting the overall generative performance. Fig. 4.8 shows a few limiting cases of the proposed method.

4.4 Chapter Summary

In this chapter, we addressed the common structural inconsistency problem in keypoint-guided human pose transformation methods, introducing a multi-stage generative architecture with textual supervision. This structural inconsistency problem is particularly relevant when the physical statures of the *observed subject* and *target pose reference* are substantially different, creating a geometric bias towards the physique of the reference

Figure 4.8: Limitations of the proposed method.

person. We mitigated such structural bias by replacing keypoint-based target pose representation with textually descriptive pose annotations. The proposed architecture consists of three independent networks for (a) text-to-pose translation, (b) pose refinement, and (c) pose transformation. First, a WGAN generator estimates the target keypoints from a textual pose description as an initial coarse pose representation. Next, the initial face keypoints are refined by regression to obtain a more structurally accurate target pose for reducing possible distortions in the generated image. Lastly, a multi-scale attention-guided conditional GAN performs the human pose transformation to generate the final output. Due to the lack of public datasets, we compiled a new dataset for benchmarking text-guided person image generation techniques by extending the DeepFashion dataset with human-annotated pose descriptions. Experimental studies show that the proposed method performs remarkably well for conditional person image generation using textual guidance as the local context. In the next chapter, we adopt a similar architecture to introduce global semantic contexts in person image generation.

CHAPTER 5

SCENE-AWARE PERSON IMAGE GENERATION

In the previous chapters, we explored conditional image generation of an isolated person instance from local structural contexts. We initially introduced an end-to-end network architecture with a multi-scale attention mechanism to improve output image quality. Then, we addressed the potential structural inconsistencies in such techniques using a multi-stage strategy with textual supervision. This chapter focuses on imposing semantic constraints on person image generation. In particular, we adopted a similar multi-stage architecture for the adaptive blending of a new person into a given scene with existing people, where the collective association of all existing human poses provides a global semantic context to the generative network.

5.1 Semantic Context from Observed Humans

In this work, the main objective is to introduce a new instance of a specific person into a scene with existing people. Therefore, the generative network needs to infer the location, scale, pose, and identity of the new instance such that the generated person adaptively *blends into* the scene while retaining the global semantics. To address the overall complexity of the task, we divided the problem into three independently learnable subtasks. First, assuming the global context as the collective association of all human poses, we interpreted the semantic context as the geometric information expressed by a set of skeletal structures corresponding to the existing persons in the scene. Each skeleton is a set of spatial coordinates of 18 body keypoints. Therefore, we represented the global context

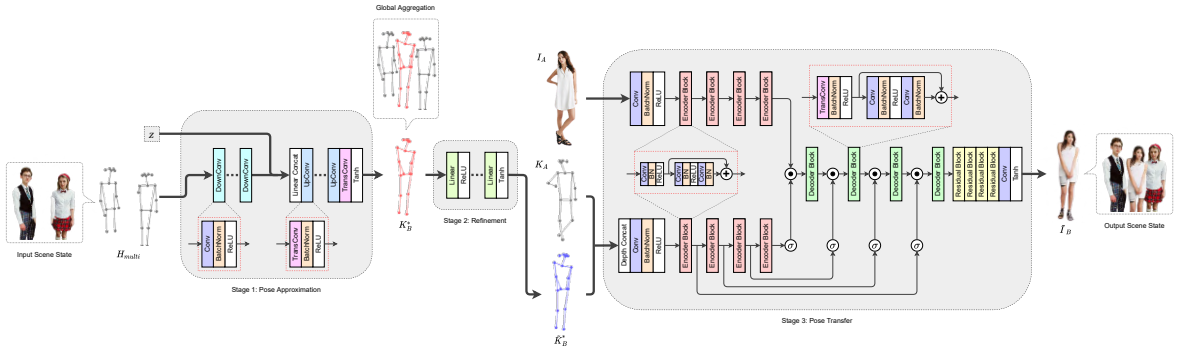

Figure 5.1: Architecture of the proposed pipeline. The workflow consists of three stages. **Stage 1** estimates the location, scale, and a spatial representation K_B^* for the target pose P_B from the global semantic context H_{multi}. **Stage 2** regressively refines the initial estimation of the face keypoints to obtain more accurate target keypoints $\tilde{K}_B^*$. **Stage 3** performs human pose transformation on the observed condition image I_A having pose P_A represented by keypoints K_A to generate the final output image $\tilde{I}_B$. Finally, the generated person instance $\tilde{I}_B$ is blended into the given scene using the location and scale predicted from Stage 1.

by encoding the existing human skeletal structures as an 18-channel *many-hot* heatmap, where each channel corresponds to one specific body keypoint. We estimated the initial pose of the target person from the global context encoding using a WGAN model. Next, we used a fully connected linear network to refine deviations in face keypoints of the initial coarse pose. This refinement strategy substantially improves the realism of the generated samples. Finally, a pose transformation network generates the target instance from the refined target pose and a given observation of the subject. We inserted the target instance into the scene using the inferred location and scale of the pose predicted in the first stage.

5.2 Scene-Aware Human Instance Generation

The proposed architecture consists of three independently learnable networks for (a) pose estimation, (b) pose refinement, and (c) pose transformation. In the first stage, an initial coarse estimation of the target pose is derived from the global semantic context. In the second stage, the coarse pose is refined by regression. In the third stage, the target instance is generated by a pose transformation conditioned on the appearance of the observed subject. Finally, the generated instance is blended into the given scene using the location and scale of the initial pose inferred from the first stage. Fig. 5.1 illustrates the architecture of the proposed generative pipeline.

5.2.1 Context Preserving Pose Estimation

Assuming a given scene that contains n persons, the proposed method aims to introduce another person into the scene, such that the new $(n+1)$-th person *blends in* with the existing n individuals while preserving the global semantic context. We represent the global geometric context of such a scene by a set P_n of n spatial pose representations corresponding to all n existing persons. Every individual pose P_i is expressed as another set of k body joint locations (keypoints). For a scene image of dimension $w \times h$, we represent every individual pose P_i as a $w \times h \times k$ binary heatmap H_i, where each channel of the heatmap corresponds to one specific keypoint. The spatial location of every visible keypoint is encoded with 1 in the respective position of H_i and 0 everywhere else (*one-hot*). We assume an aggregate of all such n heatmaps as the global geometric scene context. This results in a binary heatmap H_n of dimension $w \times h \times k$, where each channel contains spatial encoding for a specific keypoint of all n existing persons (*many-hot*). To reduce the high sparsity in H_n, we use an isotropic representation for each channel by estimating the Gaussian of the Euclidean distance from the keypoint. Such representation of the global geometric scene context is denoted by H_{multi}.

We use a WGAN model to estimate a potential target pose by sampling from a normal distribution, conditioned on the global geometric scene context H_{multi}. The network consists of a generator network G_T and a discriminator network D_T. In G_T, we first encode H_{multi} to a 512-dimension vector v_B by downscaling through consecutive convolution layers. To allow variations in the generated poses, a noise vector $z \sim \mathcal{N}(\mathbf{0}, I)$, where I is the identity matrix of size 512×512, is used as another input for G_T. Both v_B and z are linearly concatenated and passed through 4 up-convolution layers. Each of the up-convolution layers contains a transposed convolution followed by batch normalization [156] and ReLU activation [157]. The transposed convolution layers in the upscaling branch contain 256, 128, 64, and 32 filters, respectively. The output of the final up-convolution layer is passed through another transposed convolution layer followed by $tanh$ activation to produce the final output of G_T. The final transposed convolution layer contains 18 filters, resulting in a generated output $G_T(v_B, z)$ of dimension $64 \times 64 \times 18$, where each channel represents one of the 18 keypoints k_j, $j \in \{1, 2, \ldots, 18\}$. To facilitate adversarial learning, a discriminator D_T evaluates the validity of the generated keypoints. In D_T, the predicted heatmap $G_T(v_B, z)$ is passed through 4 consecutive convolution layers, where each layer has a stride of 2 and is followed by leaky ReLU activation. The convolution layers contain 32, 64, 128, and 256 filters, respectively. Finally, the scalar output of the discriminator is estimated using another convolution layer with a single channel and stride 4.

The optimization objective L_D for the discriminator D_T is given by

$$L_D = -\mathbb{E}_{(x,v_B)\sim p_t, z\sim p_z}\left[D_T(x, v_B) - D_T(G_T(z, v_B), v_B)\right]$$

where $(x, v_B) \sim p_t$ is the heatmap and context encoding pair sampled from the training set, and $z \sim p_z$ is a noise vector sampled from the Gaussian distribution. Researchers [176] have shown that the WGAN training is more stable if D_T is Lipschitz continuous, which mitigates the undesired behavior due to gradient clipping. To enforce the Lipschitz constraint, we compute the gradient penalty as follows.

$$\mathscr{P}_T = \mathbb{E}_{(\tilde{x},v_B)\sim p_{\tilde{x},v_B}}\left[\left(\left\|\nabla_{\tilde{x},v_B} D_T(\tilde{x}, v_B)\right\|_2 - 1\right)^2\right]$$

where $\|.\|_2$ indicates the L_2 norm and $\tilde{x}$ is a synthetic sample obtained from the weighted sum of a real sample x and a generated sample $G_T(z, v_B)$, mathematically, $\tilde{x} = \alpha G_T(z, v_B) + (1-\alpha)x$, where α is a random number, selected from a uniform distribution between 0 and 1. As indicated by [176], $\mathscr{P}_T$ tries to restrict the gradient magnitude to 1 and helps to enforce the Lipschitz constraint. After including the Gradient Penalty term, the final loss function for D_T is given by

$$L_{D_T} = L_D + \lambda \mathscr{P}_T$$

where λ is the regularizing constant that controls the effect of gradient penalty $\mathscr{P}_T$ on the overall discriminator loss L_{D_T}. In all our experiments, we set $\lambda = 10$. To optimize the generator network G_T, we define the loss function L_{G_T} as follows.

$$L_{G_T} = -\mathbb{E}_{v_B\sim p_{v_B}, z\sim p_z}\left[D_T(G_T(z, v_B), v_B)\right]$$

The generator G_T estimates the probable location, scale, and pose of the target person as an 18-channel heatmap. We estimate the precise spatial locations of individual keypoints from this heatmap. To represent the target pose as a set of keypoints K_B^*, we compute the maximum activation ψ_j^{max}, $j \in \{1, 2, \ldots, 18\}$ for every channel. The position of the maximum activation for j-th channel is assumed to be the spatial location of j-th keypoint if $\psi_j^{max} \geq 0.2$. The j-th keypoint is considered occluded if $\psi_j^{max} < 0.2$.

5.2.2 Face Keypoints Refinement

G_T produces a reasonable initial estimate K_B^* of the target pose from the global scene context. However, due to potential uncertainty in the position and location of the target

person, the keypoints K_B^* often have spatial perturbations. Such spatial perturbations are most prominent for the facial keypoints (nose, two eyes, and two ears). As the next stage follows a keypoint-guided pose transformation scheme, even slight fluctuations in the coordinates of the facial keypoints result in a visibly distorted output. To mitigate the effects of such perturbations, we use a fully connected linear network ω_R for refining the facial keypoints by regression. First, the five facial keypoints k_i^f, $i \in \{1,2,\ldots,5\}$ are translated by $(k_i^f - k_n)$, where k_n is the spatial location of the nose. Next, the translated facial keypoints are normalized, such that the scaled keypoints k_i^s are bounded within a unit square with the scaled nose keypoint at the origin (0, 0). Then, we flatten the scaled facial keypoints into a 10-dimensional vector v_f and pass it through three fully connected linear layers, each having 128 nodes followed by ReLU activation. The output layer contains 10 nodes, followed by $tanh$ activation. The network is optimized by minimizing the mean squared error (MSE) between the actual and the predicted keypoints. While training, we augment k_i^s with small random 2D spatial perturbations and try to predict the original values of k_i^s. Finally, the predicted keypoints are denormalized and retranslated by the same amount. The refined keypoints $\tilde{K}_B^*$ are obtained by updating the coordinates of the facial keypoints in $K*_B$ with the predictions from ω_R.

5.2.3 Pose Transformation

We use the previously proposed multi-scale attention-guided pose transformation technique [25] to generate the output image, as discussed in Sec. 3.2. The network takes an existing observation of the person I_A as the *condition image* and the channel-wise concatenated pose heatmaps $(H_A, \tilde{H}_B^*)$ derived from respective keypoints K_A and $\tilde{K}_B^*$, to produce a final output image $\tilde{I}_B$.

5.2.4 Implementation Details

In stage 1, we optimize both generator and discriminator of the WGAN using the stochastic Adam optimizer [160] with learning rate $\eta_1 = 1e^{-4}$, $\beta_1 = 0$, $\beta_2 = 0.9$, $\epsilon = 1e^{-8}$, and weight decay = 0. While training, we update the generator once for every 5 updates of the discriminator to prevent mode collapse. In stage 2, the linear pose refinement network is optimized using stochastic gradient descent with learning rate $\eta_2 = 1e^{-2}$. In stage 3, the multi-scale attention-guided pose transformation network is also optimized with the stochastic Adam optimizer [160], but with learning rate $\eta_3 = 1e^{-3}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$, and weight decay = 0. We initialize the parameters of every generator and

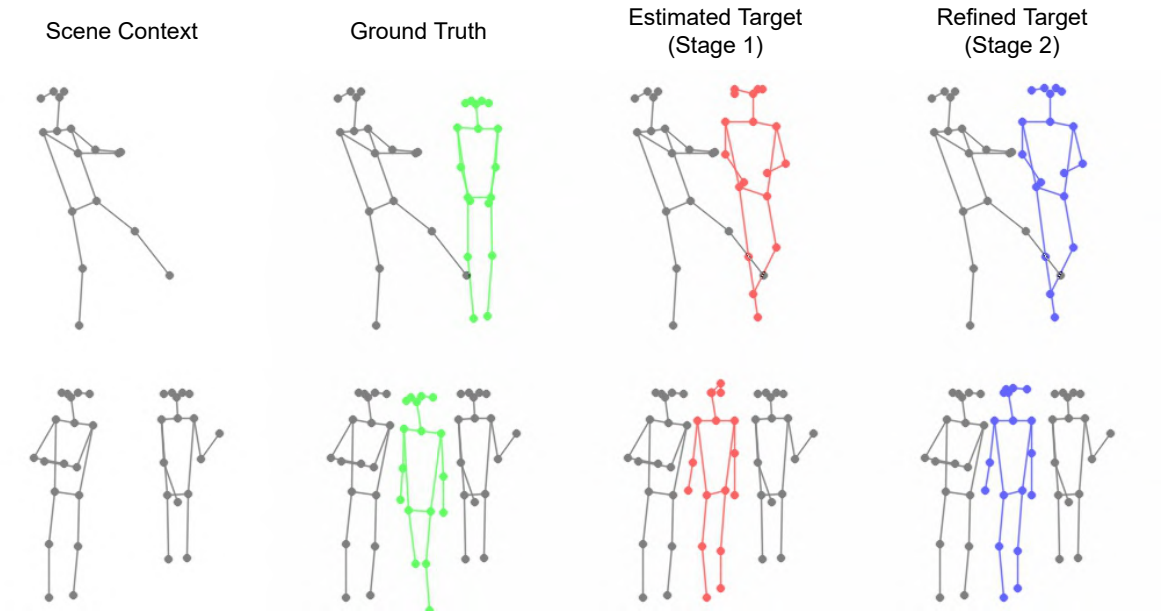

Figure 5.2: Qualitative results of *pose estimation* (stage 1) followed by *pose refinement* (stage 2). Due to the uncertainty in the inferred location and pose, the target pose may look substantially different from the ground truth. However, it does not affect the generation performance as long as the global geometric context is preserved and the target person *blends in* with the existing persons in the scene.

discriminator before training by sampling from a normal distribution of 0 mean and 0.02 standard deviation. We train the individual networks on a single NVIDIA TITAN X GPU. The pose estimation network is trained for 100K iterations with a batch size of 8. The face keypoints refinement network is trained for 100 epochs with a batch size of 128. The pose transformation network is trained for 270K iterations with a batch size of 8.

5.3 Experiments

In this section, we discuss the experimental studies performed on the proposed approach, including dataset specifications, result analyses, and limitations.

5.3.1 Datasets

We used a multi-human parsing dataset MHP-v1 [177] to train the pose estimation model. The dataset contains 4980 images where multiple persons appear in contextually correlated poses for each image. We first use a pre-trained human pose estimator [154] to derive

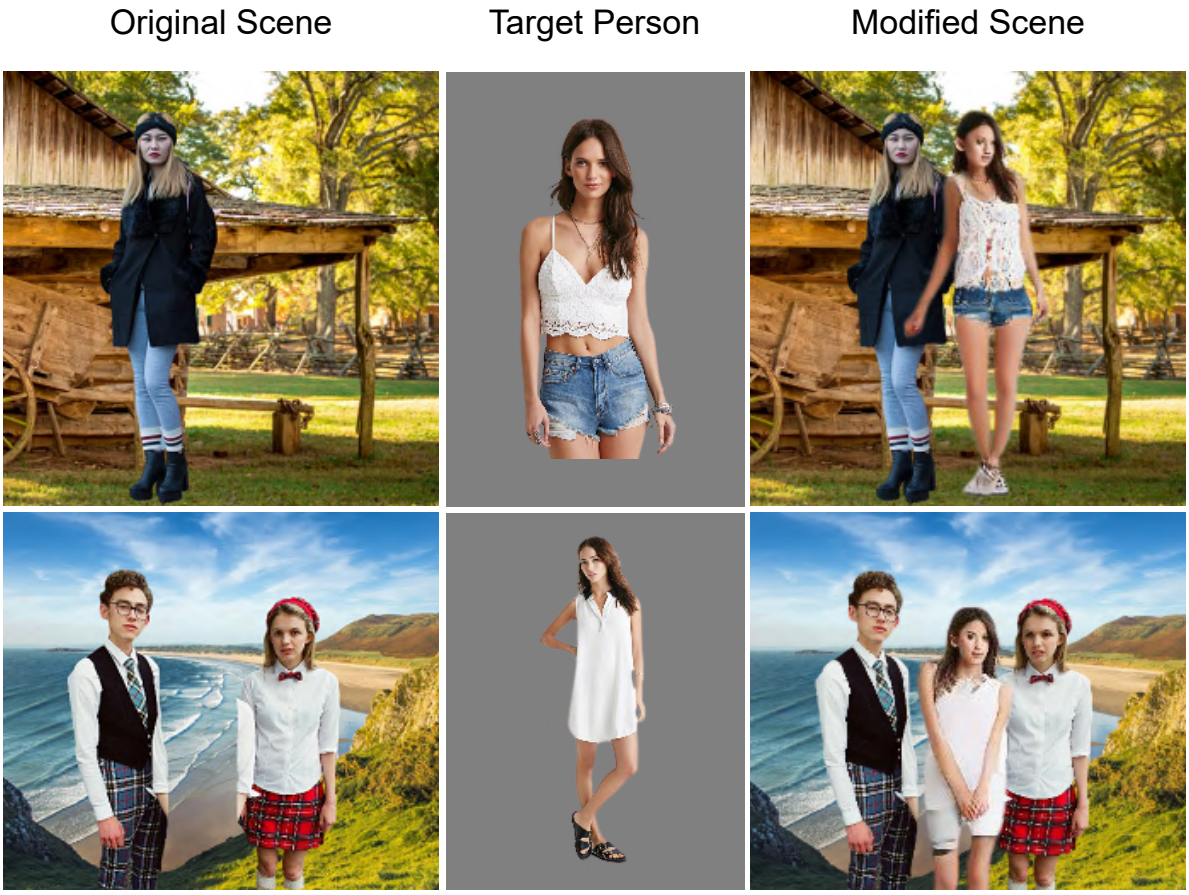

Figure 5.3: Qualitative results generated by the proposed method.

the keypoints of individual persons in each image. While training, one random person is selected as the target, and the remaining persons are used as existing individuals. To train both pose refinement and pose transfer networks, we used the DeepFashion dataset [161]. Out of 40488 images, 37344 samples are used for training and 3144 are kept for testing.

5.3.2 Result Analysis

Fig. 5.2 shows the qualitative results of the pose estimation and refinement. The proposed method can estimate realistic target poses for non-existent persons while maintaining global contextual consistency with the existing individuals in the scene. An important point to note is that there is uncertainty in the geometry of the target person as the target can adopt a wide range of valid location and pose orientations. This natural ambiguity is introduced by sampling the target pose from a Gaussian distribution. Therefore, the predicted geometry of the target may be substantially different from the ground truth, but

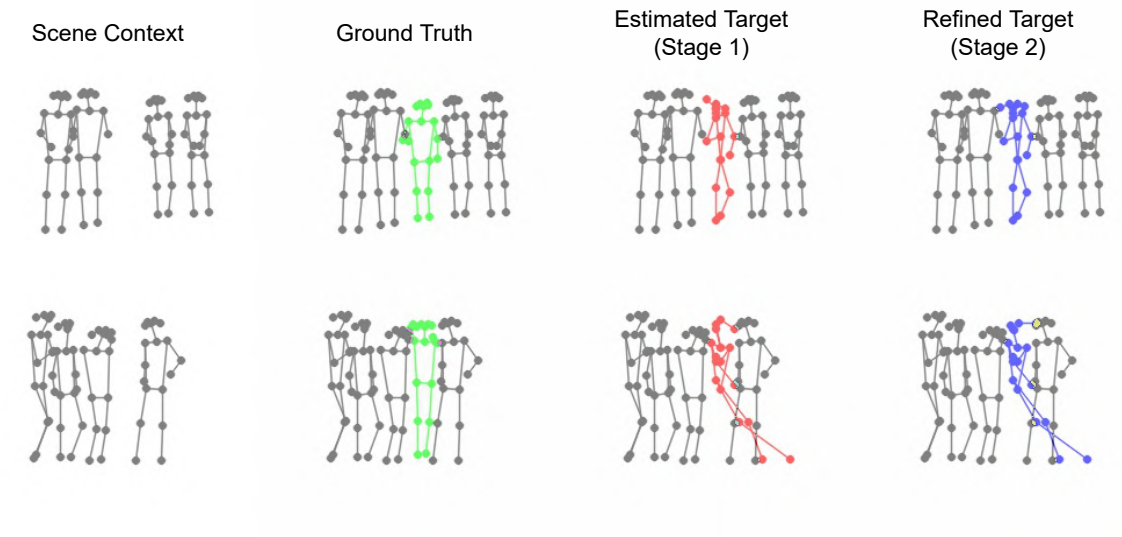

Figure 5.4: Limitations of the proposed method.

this does not affect the generation performance as long as the target *blends in* with the existing persons in the scene, preserving the global semantic context. In Fig. 5.3, we show the qualitative results of the entire pipeline consisting of all three stages. The proposed method can generate high-quality and contextually relevant target person instances that adaptively blend well with the existing individuals in the scene.

5.3.3 Limitations

Unlike conventional human pose estimators (HPE), where the keypoints are estimated for a given person, the proposed method attempts to infer an appropriate pose for a *non-existent* person using contextual information from existing individuals. The problem is challenging due to the absence of a target figure and the ambiguity in inferring a potentially *valid* pose. Therefore, on some occasions, the proposed technique struggles to estimate a correct pose, impacting the overall generative performance. This problem is particularly dominant in crowded scenes with many existing individuals. In such situations, the initial estimation of the target pose is often *too noisy to refine*, which means that the refinement strategy fails to account for the large spatial deviations. The highly perturbed pose eventually impacts pose transformation, causing visible distortions in the generated images. Fig. 5.4 shows a few such limiting cases.

5.4 Chapter Summary

In this short chapter, we explored a strategy to introduce semantic constraints in an adaptive person image generation process. Our main objective was to insert an additional person into a scene with existing individuals, such that the new person instance adopts a contextually valid pose. The *contextual validity* is enforced with a general assumption that all existing persons in the scene follow a common contextual *theme*. So, the collective association of all existing human poses can provide a global semantic context to infer probable poses for non-existent persons. However, there are two key challenges in the problem. First, unlike a conventional human pose estimator (HPE) [154], where the keypoints are estimated from an observed human image, here the generative network needs to infer conceivable postures of a non-existent person by observing the poses of surrounding individuals. Second, there is substantial ambiguity in the solution. Due to multiple possible outcomes for the location, scale, and pose, a generated instance can be *contextually valid* despite being significantly different from the ground truth.

The proposed architecture contains three independently learnable networks for (a) pose estimation, (b) pose refinement, and (c) pose transformation. The first stage derives an initial coarse estimation of the target pose from the global semantic context using a WGAN. The next stage reduces the spatial deviations of face keypoints in the coarse pose by regressing through a linear network. The final stage uses the previously proposed multi-scale attention mechanism for pose transformation using a conditional GAN. Finally, the generated instance is blended into the given scene using the location and scale of the initial pose inferred from the first stage. Although the proposed architecture performs well in a wide range of scenes, the method starts showing its limitations as the number of existing persons increases. As shown in Fig. 5.4, the uncertainty of estimating a semantically *valid* pose for a non-existent person in *crowded* scenes often leads to a highly perturbed target pose that is *too noisy to refine* for the regressor. We can address such limitations in two ways.

- One possible option to reduce structural perturbations is replacing keypoints with a less sparse pose representation, such as *human parsing maps*. In Chapter 6, we discussed this strategy in detail.
- Another probable approach is improving the global semantic context representation by *mutually attending* two different feature modalities, such as the image and corresponding segmentation map. In Chapter 7, we explored this strategy in detail.

CHAPTER 6

SEMANTICALLY ADAPTIVE PERSON IMAGE GENERATION

In Chapter 5, we introduced an initial proposal for imposing semantic constraints in an adaptive person image generation scheme, where the collective association of all existing human poses represented the global scene context. Although the method showed promising results, the ambiguity of inferring a contextually valid pose for a non-existent person often caused substantial instabilities in the estimated pose. This chapter focuses on improving the generative performance for the same objective by replacing keypoint-based pose representation with parsing maps in a data-agnostic approach.

6.1 Deriving Local Attributes from Global Scene Context

The core objective of this work is to blend an observed person into a scene with existing individuals, such that the inserted instance adopts semantically consistent location, scale, and pose with the existing persons. Most human pose transformation strategies follow a conditional image-to-image translation approach using appearance attributes of an observed exemplar and a target pose. The target pose supervision is generally provided as keypoints [77, 79, 80, 81, 83, 88, 89], parsing maps [95, 99, 107], or text descriptions [26, 130]. Therefore, such methods are limited to isolated person instance generation from local object-level contextual guidance. In contrast, the objective of this work requires additional supervision from existing persons in the scene as a global contextual guidance. Specifically, we aim to estimate the best possible local attributes for pose transformation by deriving them from the global semantic representation. However, due to multiple semantically

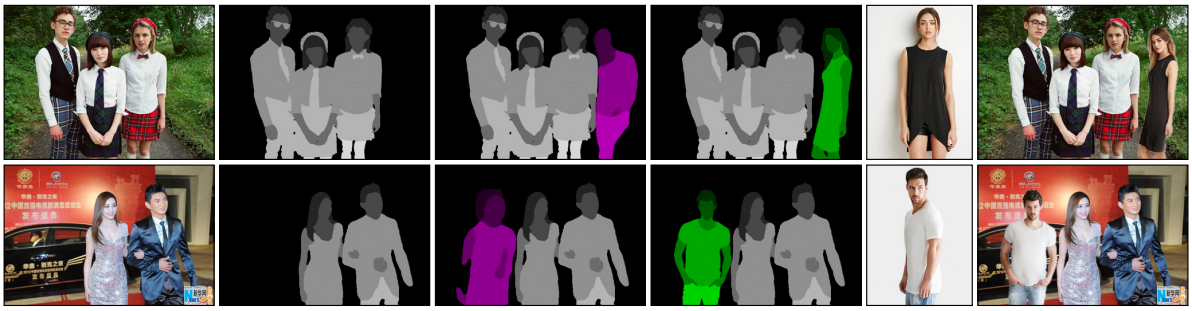

Figure 6.1: An overview of the proposed method. **(a)** Original scene. **(b)** Semantic maps of existing persons in the scene. **(c)** Coarse estimation of the potential target semantic map. **(d)** Data-driven refinement of the coarse semantic map. **(e)** An exemplar of the target person. **(f)** Generated scene with the rendered target person.

feasible outcomes, determining a set of local descriptors for a non-existent person is challenging and often requires additional refinement of the initial estimation.

In our previous approach, as discussed in Chapter 5, we represented the global scene context as a collective association of all existing human poses. However, the method shows frequent instabilities in the estimated poses due to wide deviations in the spatial locations of predicted keypoints, virtually making the pose transformation infeasible. To mitigate such instabilities, in the proposed approach, we replaced keypoints with *human parsing maps* to reduce the sparsity in the local descriptors. The initially estimated coarse masks are refined by a *data-agnostic* approach with a clustered database of precomputed fine human parsing maps. The fine masks allow highly accurate visual attribute transfer from the observed exemplar to the target instance, rendering high-quality realistic human instances. Fig. 6.1 illustrates an overview of the proposed method.

6.2 Semantically Adaptive Human Instance Generation

The proposed architecture consists of three independently learnable networks for (a) coarse generation, (b) refinement, and (c) rendering. The first stage uses an *image-to-image* translation to estimate the local parsing map for a probable person from the global semantic context, expressed as an image containing the parsing maps of all existing individuals in the scene. The generated coarse parsing map provides an initial estimate of the target location and scale. However, the coarse map performs poorly during visual attribute transfer for rendering due to inaccuracies in the label boundaries. Therefore, in the second stage, we introduced a data-agnostic refinement strategy to retrieve a

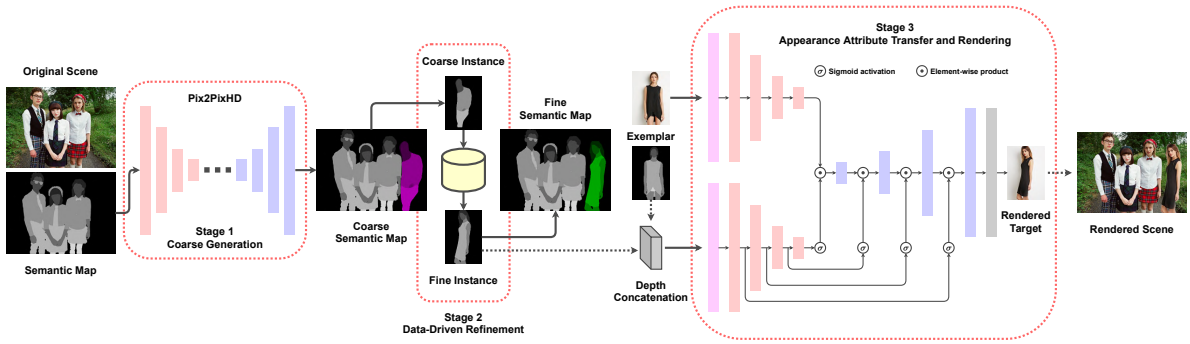

Figure 6.2: Architecture of the proposed pipeline. The workflow consists of three stages. **(a)** Coarse semantic map estimation from the global scene context in stage 1. **(b)** Data-driven refinement of the initially estimated coarse semantic map in stage 2. **(c)** Rendering the fine semantic map by transferring appearance attributes from an exemplar in stage 3.

representative target parsing map from an existing set of precomputed fine parsing maps. Finally, the third stage uses a conditional GAN to render the target person by transferring appearance attributes from a given exemplar to the target parsing map. Fig. 6.2 shows the architecture of the proposed method. Additionally, we can use image harmonization [178, 179] as an optional post-processing step to reduce blending inconsistencies between foreground and background.

6.2.1 Coarse Generation Network

We use an encoder-decoder architecture to generate an initial estimate of the target location, scale, and pose. This network performs an *image-to-image* translation from a semantic map S containing N persons to another semantic map T having the $(N+1)$-th person. The network aims to generate a coarse semantic map for a new person, so the new person is contextually relevant to the existing persons in the scene.

Both S and T are single-channel semantic maps containing eight labels corresponding to eight regions of a human body. This reduced set of label groups simplifies the semantic map generation while retaining sufficient information for high-quality image synthesis in the following stages. The reduced set of semantic label groups contains – background (0), hair (1), face (2), torso and upper limbs (3), upper body wear (4), lower body wear (5), lower limbs (6), and shoes (7). In [137], the authors also provide one channel for the face and another optional channel to specify the region boundary for the target. In contrast, we do not consider these additional channels due to different strategies for refinement and rendering in later stages.

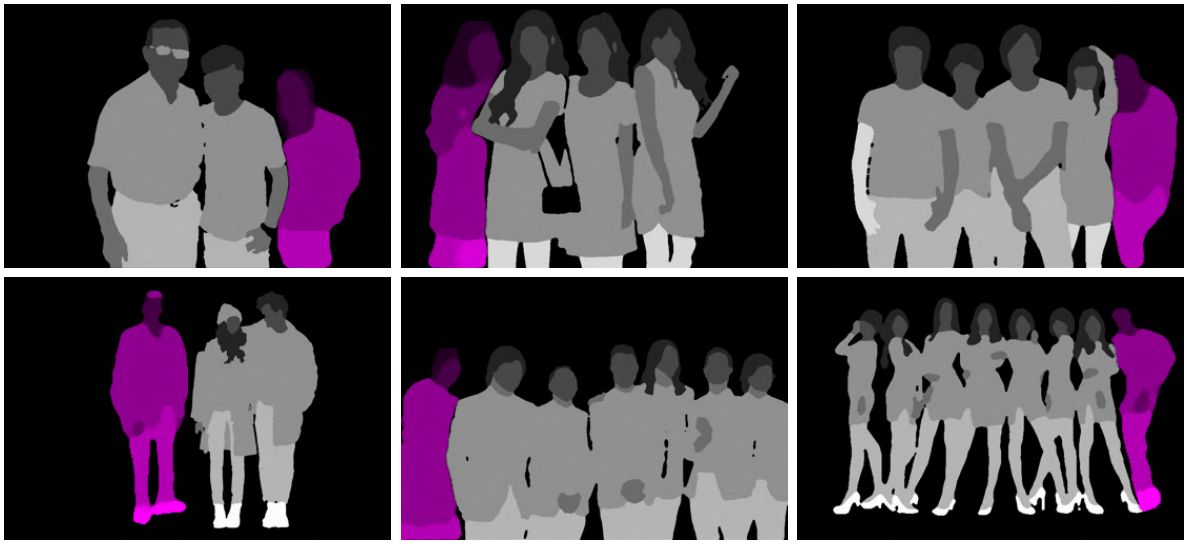

Figure 6.3: Qualitative results of the initial target semantic map generation in stage 1. Semantic maps of existing persons are marked in *gray*, and the initial estimation of the target semantic map is marked in *purple*.

The coarse generation network adopts the default encoder-decoder architecture of Pix2PixHD [33]. We use a spatial dimension of 368×368 for the semantic maps. The original semantic maps are resized while maintaining the aspect ratio and then padded with zeroes to have the desired square dimension. We use nearest-neighbor interpolation when resizing to preserve the number of label groups in the semantic maps. The only modification we apply to the default Pix2PixHD architecture is disabling the VGG feature-matching loss because it is possible to have wide variations in the target location, scale, and pose, which leads to significant uncertainty in the generated semantic map. Fig. 6.3 shows a few examples of the coarse generation.

6.2.2 Data-Agnostic Refinement Strategy

The coarse semantic map provides a reasonable estimate for the target person, which is contextually coherent with the global semantics of the scene. While the spatial location and scale of the target are immediately usable to localize a new person into the scene, the semantic map itself is not sufficiently viable to produce realistic results. In [137], the authors use a *multi-conditional rendering network* (MCRN) on the roughly estimated semantic map, followed by a *face refinement network* (FRN) on the rendered target. While this approach produces some decent results, it is limited in scope due to solely relying on the initially generated rough semantic map from the *essence generation network* (EGN).

We notice two crucial issues in this regard. Firstly, the coarse semantic map heavily affects the visual realism of the generated image. Secondly, it is not easy to achieve control over the appearance of the generated target with a fixed semantic representation. For example, if EGN produces a semantic map that appears to be a man while the intended exemplar is a woman. The subtle difference in core appearance attributes between the estimated semantic map and exemplar poses a significant challenge in practically usable generation results. We attempt to improve visual quality and appearance diversity in the generated results by introducing a data-driven refinement strategy with a clustered *knowledge base*.

We collect a set of finely annotated semantic maps of high-quality human images to construct a small database with a diverse range of natural poses. This database works as a *knowledge base* for our method. To optimally split the knowledge base into several clusters, we first encode the individual semantic maps using a VGG-19 [159] model pretrained on ImageNet [56]. The semantic maps are resized to a square grid of size 128×128, maintaining the aspect ratio and using zero padding. The resampling uses nearest-neighbor interpolation. After passing the resized image through the VGG-19 network, the final feature extraction layer produces an output of dimension $4 \times 4 \times 512$. To avoid too many features during clustering, we apply adaptive average pooling to map the feature space into a dimension of $1 \times 1 \times 512$. The pooled feature space is flattened to a 512-dimensional feature vector. We perform K-means clustering on the encoded feature vectors corresponding to the samples in the knowledge base. From our ablation study, we have found 8 clusters work best for our case. After the learning converges, we split the knowledge base by the predicted class labels.

During refinement, the coarse semantic map is center-cropped and resized to dimension 128×128, maintaining the aspect ratio. The resampling uses the same nearest-neighbor interpolation as earlier. The resized coarse semantic map is then similarly encoded and passed to the pretrained K-means model for inference. After receiving a cluster assignment, we measure the *Cosine Similarity* between the encoded coarse semantic map and every sample previously classified as a cluster member. The refinement returns one or more existing samples by the similarity score-based ranking. The retrieved selection acts as the refined semantic map of the target person.

As we perform a nearest-neighbor search in the semantic feature space of samples in precomputed clusters, given a coarse semantic map, we can dynamically select a refined candidate for either *women* or *men* as required. This step can be automated if the gender of the exemplar is either known or estimated using a pretrained classifier. Fig. 6.4 shows *Top-5* matches for both *women* and *men* samples given a coarse semantic map as the query.

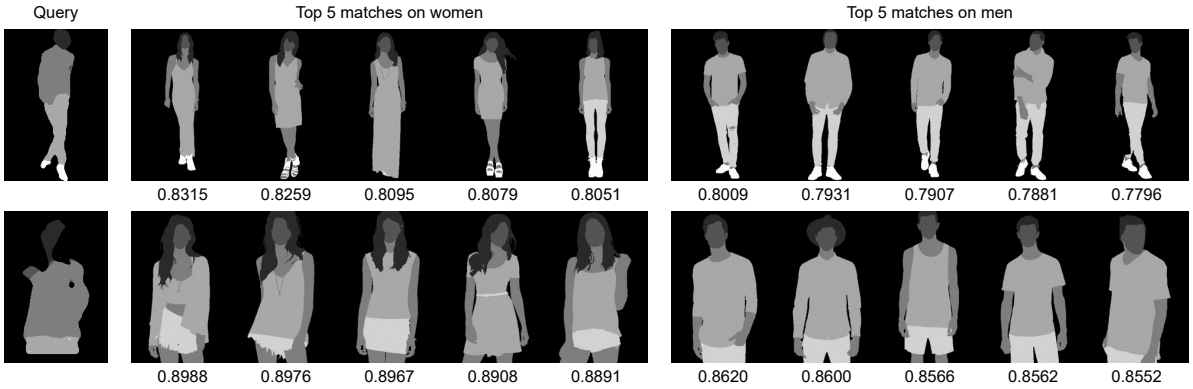

Figure 6.4: Qualitative results of refinement in stage 2. The first column shows a coarse semantic map as the *query*, and the following columns show the *Top-5* fine semantic maps retrieved for both genders. The *Cosine Similarity* score for each retrieval is shown below the respective sample.

6.2.3 Appearance Attribute Transfer and Rendering

In [137], the authors trained the rendering network on single instances extracted from multi-person images. In contrast, we impose the rendering task as a *human pose transformation* problem to transfer appearance attributes conditioned on the pose translation. Let us assume a pair of images I_A and I_B of the same person but with different poses P_A and P_B, respectively. We aim to train the network such that it renders a realistic approximation $\hat{I}_B$ (generated) of I_B (target) by conditioning the pose translation (P_A, P_B) on the appearance attributes of I_A (exemplar). We represent each pose with a semantic map consisting of 7 label groups – background (0), hair (1), face (2), skin (3), upper body wear (4), lower body wear (5), and shoes (6). For effective attribute transfer on different body regions, the semantic map P is converted into a 6-channel binary heatmap (0 for the background and 1 for the body part) H, where each channel indicates one specific body region. We use a spatial dimension of $256 \times 256 \times 3$ for I_A, I_B, and $\hat{I}_B$. Consequently, the same for H_A and H_B is $256 \times 256 \times 6$. We adopted a multi-scale attention-guided generative network [25] for rendering. The generator $\mathscr{G}$ takes the exemplar I_A and the depth-wise concatenated heatmaps (H_A, H_B) as inputs to produce an estimate $\hat{I}_B$ for the target I_B. The discriminator $\mathscr{D}$ takes the channel-wise concatenated image pairs, either (I_A, I_B) (*real*) or $(I_A, \hat{I}_B)$ (*fake*), to estimate a binary class probability map for 70×70 receptive fields (*patches*).

The generator $\mathscr{G}$ has two separate but identical encoding pathways for I_A and (H_A, H_B). At each branch, the input is first mapped to a $256 \times 256 \times 64$ feature space by convolution (3×3 kernel, stride=1, padding=1, bias=0), batch normalization, and ReLU activation.

The feature space is then passed through 4 consecutive downsampling blocks, where each block reduces the spatial dimension by half while doubling the number of feature maps. Each block consists of convolution (4×4 kernel, stride=2, padding=1, bias=0), batch normalization, and ReLU activation, followed by a basic *residual block* [155]. The network has a single decoding path that upsamples the combined feature space from both the encoding branches. We have 4 consecutive upsampling blocks in the decoder, where each block doubles the spatial dimension while compressing the number of feature maps by half. Each block consists of transposed convolution (4×4 kernel, stride=2, padding=1, bias=0), batch normalization, and ReLU activation, followed by a basic *residual block*. We apply an attention operation at every spatial dimension to preserve both coarse and fine appearance attributes in the generated image. Mathematically, for the first decoder block at the lowest resolution, $k = 1$,

$$I_1^D = D_1\left(I_4^E \odot \sigma\left(H_4^E\right)\right)$$

and for the subsequent decoder blocks at higher resolutions, $k = \{2,3,4\}$,

$$I_k^D = D_k\left(I_{k-1}^D \odot \sigma\left(H_{5-k}^E\right)\right)$$

where I_k^D is the output from the k-th decoder block, I_k^E and H_k^E are the outputs from the k-th encoder blocks of image branch and pose branch respectively, σ denotes the *sigmoid* activation function, and $\odot$ denotes the Hadamard product. Finally, the resulting feature space goes through 4 consecutive basic *residual blocks*, followed by a convolution (1×1 kernel, stride=1, padding=0, bias=0) and *tanh* activation to project the feature maps into the final output image $\hat{I}_B$ of size $256 \times 256 \times 3$.

The generator objective function $\mathscr{L}_{\mathscr{G}}$ is a linear combination of three loss terms. It includes a pixel-wise L_1 loss $\mathscr{L}_{MAE}^{\mathscr{G}}$, an adversarial discrimination loss $\mathscr{L}_{GAN}^{\mathscr{G}}$ estimated using the discriminator $\mathscr{D}$, and a perceptual loss $\mathscr{L}_{VGG_\rho}^{\mathscr{G}}$ estimated using a VGG-19 network pretrained on ImageNet. Mathematically,

$$\mathscr{L}_{MAE}^{\mathscr{G}} = \left\|\hat{I}_B - I_B\right\|_1$$

where $\|.\|_1$ denotes the L_1 norm or the mean absolute error (MAE).

$$\mathscr{L}_{GAN}^{\mathscr{G}} = \mathscr{L}_{BCE}\left(\mathscr{D}\left(I_A, \hat{I}_B\right), 1\right)$$

where $\mathscr{L}_{BCE}$ denotes the binary cross-entropy loss.

$$\mathscr{L}^{\mathscr{G}}_{VGG_\rho} \;=\; \frac{1}{h_\rho w_\rho c_\rho} \sum_{i=1}^{h_\rho} \sum_{j=1}^{w_\rho} \sum_{k=1}^{c_\rho} \left\| \phi_\rho\left(\hat{I}_B\right) \;-\; \phi_\rho\left(I_B\right) \right\|_1$$

where ϕ_ρ denotes the output of dimension $h_\rho \times w_\rho \times c_\rho$ from the ρ-th layer of a VGG-19 network pretrained on ImageNet. We incorporate two perceptual loss terms for $\rho = 4$ and $\rho = 9$ into the cumulative generator objective. Therefore, the final generator objective is given by

$$\mathscr{L}_{\mathscr{G}} \;=\; \arg\min_G \max_D \;\; \lambda_1 \mathscr{L}^{\mathscr{G}}_{MAE} \;+\; \lambda_2 \mathscr{L}^{\mathscr{G}}_{GAN} \;+\; \lambda_3 \left(\mathscr{L}^{\mathscr{G}}_{VGG_4} \;+\; \mathscr{L}^{\mathscr{G}}_{VGG_9} \right)$$

where λ_1, λ_2, and λ_3 are the tunable weights for the corresponding loss components.

The discriminator $\mathscr{D}$ is a generic PatchGAN [30] that operates on 70×70 receptive fields (*patches*) over the input. It takes the depth-wise concatenated image pairs, either (I_A, I_B) or $(I_A, \hat{I}_B)$, as a *real* (1) or *fake* (0) image transition, respectively. $\mathscr{D}$ predicts a binary class probability map for the input patches.

The discriminator objective function $\mathscr{L}_{\mathscr{D}}$ has only a single component $\mathscr{L}^{\mathscr{D}}_{GAN}$, calculated as the average BCE loss over *real* and *fake* transitions. Mathematically,

$$\mathscr{L}^{\mathscr{D}}_{GAN} \;=\; \frac{1}{2} \left[\mathscr{L}_{BCE}\left(\mathscr{D}\left(I_A, I_B\right), 1\right) \;+\; \mathscr{L}_{BCE}\left(\mathscr{D}\left(I_A, \hat{I}_B\right), 0\right) \right]$$

Therefore, the final discriminator objective is given by

$$\mathscr{L}_{\mathscr{D}} \;=\; \arg\min_D \max_G \;\; \mathscr{L}^{\mathscr{D}}_{GAN}$$

6.2.4 Implementation Details

Stage 1: We train the coarse generation network with batch size 16 and VGG feature-matching loss disabled. All other training parameters are kept to defaults as specified by the authors of Pix2PixHD [33].

Stage 2: The clustering follows Lloyd's K-means algorithm with 8 clusters, a relative tolerance of $1e^{-4}$, 1000 maximum iterations, and 10 random centroid initializations.

Stage 3: For the rendering network, we set $\lambda_1 = 5$, $\lambda_2 = 1$, and $\lambda_3 = 5$ in the generator objective. The parameters of both the generator and discriminator networks are initialized before optimization by sampling from a normal distribution of mean = 0 and standard deviation = 0.02. We use the stochastic Adam optimizer [160] to update the parameters of both networks. We set the learning rate $\eta = 1e^{-3}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$, and weight decay = 0 for both optimizers. The network is trained with batch size 4.

We train the individual networks on a single NVIDIA TITAN X GPU. The coarse generation network is trained for 200K iterations with a batch size of 4 and the appearance rendering network is trained for 300K iterations with a batch size of 8.

6.3 Experiments

In this section, we discuss the experimental studies performed on the proposed approach, including datasets, evaluation metrics, comparative analyses, ablations, and limitations.

6.3.1 Datasets

We use the multi-human parsing dataset LV-MHP-v1 [177] to train the coarse generation network in stage 1. The dataset contains 4980 high-quality images, each having at least two persons (average is three), and the respective semantic annotations for every individual in the scene. The annotation includes 19 label groups – background (0), hat (1), hair (2), sunglasses (3), upper clothes (4), skirt (5), pants (6), dress (7), belt (8), left shoe (9), right shoe (10), face (11), left leg (12), right leg (13), left arm (14), right arm (15), bag (16), scarf (17), and torso skin (18). As discussed in Sec. 6.2.1, we reduce the original 19 label groups to 8 by merging as – background + bag (0), hair (1), face (2), both arms + torso skin (3), hat + sunglasses + upper clothes + dress + scarf (4), skirt + pants + belt (5), both legs (6), both shoes (7). While training the coarse generation network, we select one random instance of a scene as the target person and the remaining instances as the input context. We prepare 14854 training pairs from 4945 images and 115 test pairs from the remaining 35 images.

For data-driven refinement in stage 2 and rendering network in stage 3, we use the DeepFashion dataset [161]. The dataset contains high-quality isolated person instances with wide pose and attire variations. A subset of the samples has color annotations for 16 semantic label groups. We reduce the number of label groups to 7 by merging multiple semantic regions as – background + bag (0), hair + headwear (1), face + eyeglass (2), neckwear + skin (3), top + dress + outer (4), skirt + belt + pants (5), leggings + footwear (6). We prepare 9866 images and corresponding semantic maps for creating our clustered database. We select 9278 image pairs for training and 786 image pairs for testing the rendering network.

6.3.2 Evaluation metrics

Similar to our previous analyses in Chapters 3, 4, and 5, we use the Structural Similarity Index (SSIM) [162], Inception Score (IS) [163], Detection Score (DS) [164], PCKh [165], and

Figure 6.5: Qualitative comparison of the proposed method against existing approaches by Lee *et al.* [136], Gafni *et al.* [137], and Kulal *et al.* [138].

Learned Perceptual Image Patch Similarity (LPIPS) [168] for quantitatively benchmarking the proposed rendering technique against existing human pose transformation methods [77, 80, 81, 83, 88, 89, 95, 99, 107, 137].

6.3.3 Result Analysis

Qualitative and quantitative comparisons: In Fig. 6.5, we compare our approach qualitatively with existing person image insertion techniques [136, 137, 138]. The visual analysis shows unrealistic persons for [136] and inadequate rendering for [137]. In [138], the authors have assumed the objective as a conditional inpainting problem, improving the overall visual quality of image blending over [136, 137]. However, in our experiments, the technique [138] often fails to insert a new person into multi-person scenes, and the method lacks a faithful appearance attribute transfer to retain the exemplar identity. In contrast,

Table 6.1: Quantitative comparison among different human pose transformation methods for rendering the target person instance. The best scores are in **bold**, and the second-best scores are underlined.

Method	SSIM ↑	IS ↑	DS ↑	PCKh ↑	LPIPS ↓ (VGG)	LPIPS ↓ (SqzNet)
PG^2 [77]	0.773	3.163	0.951	0.89	0.523	0.416
Deform [80]	0.760	3.362	0.967	0.94	-	-
VUNet [81]	0.763	3.440	0.972	0.93	-	-
PATN [83]	0.773	3.209	0.976	0.96	0.299	0.170
XingGAN [88]	0.762	3.060	0.917	0.95	0.224	0.144
BiGraphGAN [89]	0.779	3.012	0.954	**0.97**	0.187	0.114
ADGAN [95]	0.677	3.116	0.938	0.96	0.256	0.144
PISE [99]	0.759	3.210	0.974	0.96	0.201	0.109
CASD [107]	0.724	**3.460**	**0.987**	**0.97**	0.222	0.120
WYWH (KP) [137]	0.788	3.189	-	-	0.271	0.156
WYWH (DP) [137]	0.793	3.346	-	-	0.264	0.149
Ours	**0.845**	3.351	0.968	**0.97**	**0.124**	**0.064**
Ground Truth	1.000	3.687	0.970	1.00	0.000	0.000

the proposed method can produce realistic and visually appealing results for person insertion into a complex scene, with a semantically consistent pose while preserving the appearance and identity of the exemplar. To analyze the overall generation quality of the rendering network, we perform a quantitative comparison against ten previous person image generation methods [77, 80, 81, 83, 88, 89, 95, 99, 107, 137]. As shown in Table 6.1, the proposed rendering method outperforms previous approaches in most evaluation metrics.

Subjective evaluation: Additionally, we have conducted an opinion-based user study with 72 volunteers to rate the final generated scenes as *real* or *fake*. Following the protocols in [137], we have kept the allowed observation time unrestricted during the study. The proposed method has received a mean opinion score of **64.4%** against 59.2% by [138], 51.8% by [137], and 32.1% by [136], respectively.

6.3.4 Ablation Study

Impact of feature representation on refinement: As discussed in Sec. 6.2.2, we use 512-dimensional VGG-encoded features to guide the refinement process. To evaluate the significance of feature representation in the proposed refinement strategy, we compare VGG-encoded features with native pixel-based features in our ablation analysis by converting the input image into a feature vector. The conversion process downscales

Figure 6.6: Qualitative ablation analysis on the feature representation for refinement. The *Cosine Similarity* score for each retrieval is shown below the respective sample.

Table 6.2: Quantitative ablation analysis on refinement with pixel-based features. The best scores are in **bold**, and the second-best scores are underlined.

Number of clusters	Average Cosine Similarity of Top match ↑			Average Cosine Similarity of Top-5 matches ↑		
	Men	Women	Overall	Men	Women	Overall
K = 8	0.7127	**0.7562**	**0.7608**	0.6912	**0.7366**	**0.7402**
K = 16	**0.7146**	0.7539	0.7598	**0.6933**	0.7357	**0.7402**
K = 32	0.7014	0.7449	0.7492	0.6768	0.7270	0.7302
K = 64	0.5852	0.6767	0.6810	0.5580	0.6301	0.6346

Table 6.3: Quantitative ablation analysis on refinement with VGG-encoded features. The best scores are in **bold**, and the second-best scores are underlined.

Number of clusters	Average Cosine Similarity of Top match ↑			Average Cosine Similarity of Top-5 matches ↑		
	Men	Women	Overall	Men	Women	Overall
K = 8	**0.8212**	**0.8319**	**0.8390**	0.7933	**0.8171**	**0.8245**
K = 16	0.8184	0.8307	0.8371	**0.7941**	0.8146	0.8227
K = 32	0.8073	0.8313	0.8379	0.7824	0.8140	0.8225
K = 64	0.7995	0.8290	0.8368	0.7715	0.8109	0.8208

(nearest-neighbor interpolation) the original 176×256 images to 22×32, keeping the aspect ratio intact, followed by flattening to a 704-dimensional feature vector. We evaluate both the feature representation techniques for different numbers of clusters ($K = 8, 16, 32, 64$). As shown in Tables 6.2 and 6.3, for any K, the VGG-encoded feature representation outperforms the native pixel-based representation in average *Cosine Similarity* score of top retrievals. Fig. 6.6 illustrates the similarity score-based ranking of retrieved samples with each feature encoding type for both genders. The VGG feature-based clustering provides a better resemblance between the query and retrieved semantic maps. From our ablation study, we find $K = 8$ works best for our data.

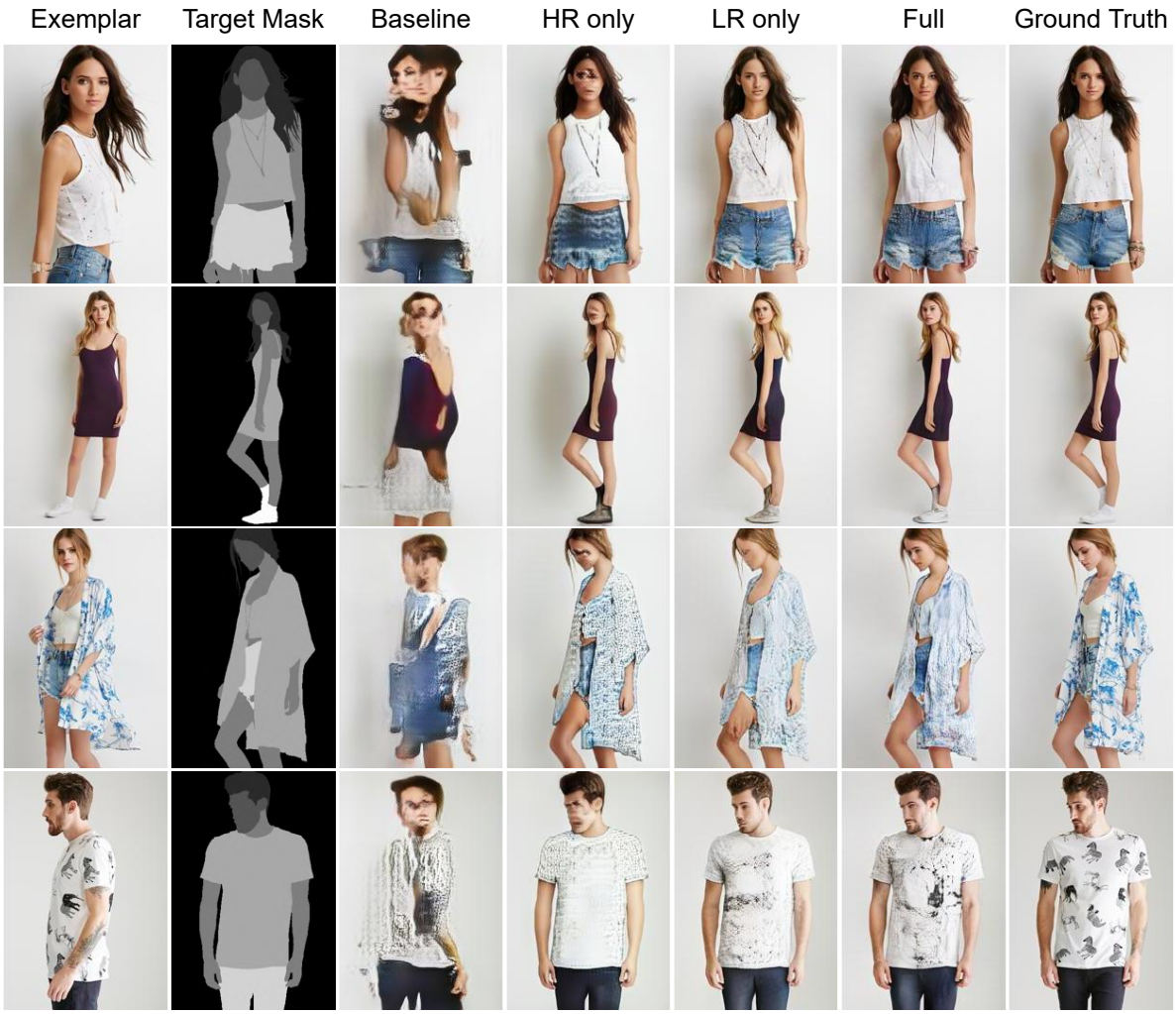

Figure 6.7: Qualitative ablation analysis among different variants of the rendering network.

Impact of attention mechanism on rendering: The multi-scale attention mechanism in the rendering network plays a crucial role in the generated image quality. We explore four different configurations to validate and select the optimal attention strategy. In the first configuration (**Baseline**), we remove all attention operations, concatenate I_4^E with H_4^E depth-wise, and pass the concatenated feature space through the decoder block. We consider only one attention operation in the rendering network for the second and third configurations. In the second variant (**HR only**), the attention operation is performed at the highest feature resolution only (just before the decoder block D_4). Similarly, in the third variant (**LR only**), the attention operation is performed at the lowest feature resolution only (just before the decoder block D_1). In the final configuration (**Full**), we use the proposed attention mechanism as shown in Fig. 6.2 and described in Sec. 6.2.3. We train and evaluate all four network variants on the same dataset splits while keeping all experimental

Table 6.4: Quantitative ablation analysis among different variants of the rendering network. The best scores are in **bold**, and the second-best scores are underlined.

Model	SSIM ↑	IS ↑	DS ↑	PCKh ↑	LPIPS ↓ (VGG)	LPIPS ↓ (SqzNet)
`Baseline`	0.657	**3.667**	0.902	0.46	0.338	0.260
`HR only`	0.825	3.271	0.954	0.96	0.154	0.088
`LR only`	0.840	3.326	0.966	0.96	0.131	0.068
`Full` ***(proposed)***	**0.845**	3.351	**0.968**	**0.97**	**0.124**	**0.064**
Ground Truth	1.000	3.687	0.970	1.00	0.000	0.000

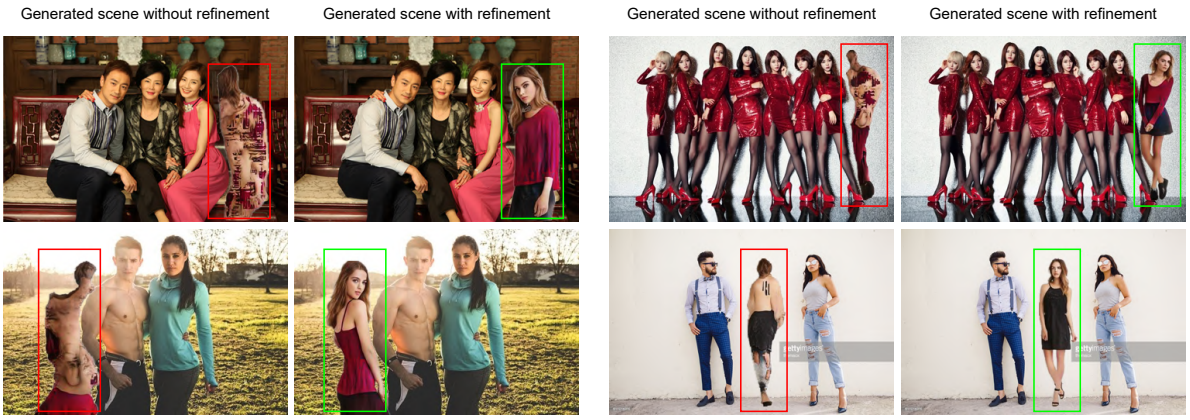

Figure 6.8: Qualitative ablation analysis on the impact of refinement on rendering. Each pair of examples shows a rendered person instance in the modified scene *without* and *with* refinement, marked with *red* and *green* bounding boxes, respectively.

conditions the same, as noted in Sec. 6.2.4. Table 6.4 summarizes the evaluated metrics and Fig. 6.7 shows the qualitative results for all four configurations. We conclude from these analytical and visual results that the proposed attention mechanism provides the best generative performance, retaining both the *low-frequency* and *high-frequency* details in the generated images.

Impact of refinement on rendering: Fig. 6.8 demonstrates the necessity of the refinement mechanism on the rendered person instance by comparing the final modified scene *without* and *with* refinement applied.

6.3.5 Generative Performance

6.3.5.1 Visual Results *In the Wild*

Fig. 6.9 shows the visual results generated by the proposed method on a few randomly collected scenes *outside* the LV-MHP-v1 dataset [177]. We estimate the semantic parsing

Figure 6.9: Generative performance of the proposed method *in the wild* using random scenes *outside* the dataset. Each set of examples shows – the original scene (**left**), an exemplar of the target person (**middle**), and the final generated scene (**right**).

maps corresponding to existing persons using a *self-correcting human parser* [180] pretrained on the ATR dataset [181]. The remaining steps to generate the final rendered scene follow an identical pipeline in sequence, as discussed in Sec. 6.2. These results demonstrate that the proposed approach extends well beyond the experimental setup and can adapt to various complex natural scenes.

6.3.5.2 Appearance Control

The proposed method renders the target person by transferring appearance attributes from an exemplar to the fine target semantic parsing map. Different label groups in the semantic parsing map can define precise segmentation masks for respective body parts. Therefore, it provides an implicit way to manipulate the appearance of various body regions of the target person by using *bitwise* operations. Mathematically, the modified target image $\hat{I}$ is given by

$$\hat{I} = \left[M \odot \mathscr{G}\left(I_{style}, \left(H_{style}, H_{target}\right)\right)\right] \oplus \left[(1 - M) \odot I_{target}\right]$$

where I_{style} and I_{target} denote the style reference and target person, respectively; H_{style} and H_{target} denote the heatmap representations for the semantic parsing maps of the

Figure 6.10: Qualitative results of appearance control in rendered person. (From top to bottom) **First row:** A person as the style reference. **Second row:** Target person. **Third row:** Target person with replaced hair as the reference. **Fourth row:** Target person with replaced upper body wear as the reference. **Fifth row:** Target person with replaced lower body wear as the reference.

style reference and target person, respectively; M defines a binary segmentation mask for the body part being manipulated; $\mathscr{G}$ is the generator; $\odot$ and $\oplus$ are the *element-wise* multiplication and addition operations, respectively. Figs. 6.10 and 6.11 illustrate this idea of appearance control using samples from the DeepFashion dataset [161].

6.3.5.3 Pose Variations

We refine the initial coarse estimation of the target semantic map by retrieving closely aligned samples from the pre-partitioned clusters using a *Cosine Similarity* score-based ranking in the encoded latent space. Besides refining the coarse target semantic map, this data-driven strategy provides an explicit way to achieve unconstrained pose variations for the synthesized person. Fig. 6.12 demonstrates a few examples to illustrate this idea by rendering the target person with *Top-5* retrieved semantic maps from the refinement stage.

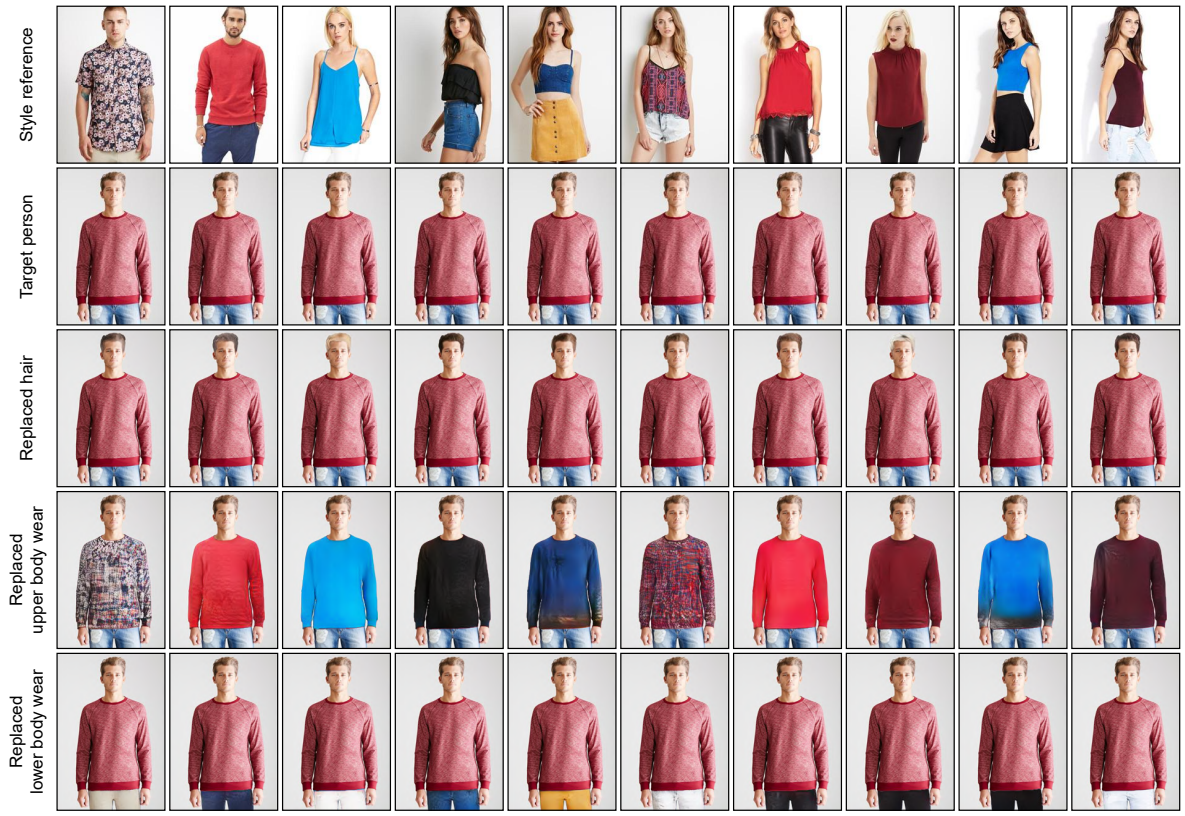

Figure 6.11: Qualitative results of appearance control in rendered person. (From top to bottom) **First row:** A person as the style reference. **Second row:** Target person. **Third row:** Target person with replaced hair as the reference. **Fourth row:** Target person with replaced upper body wear as the reference. **Fifth row:** Target person with replaced lower body wear as the reference.

6.3.6 Limitations

Although the proposed method can produce high-quality, visually appealing results for a wide range of complex scenes, there are a few occasions when the technique fails to generate a realistic outcome. Due to a disentangled multi-stage approach, these limiting cases may occur from different pipeline components. Coarse generation in stage 1 provides the spatial location and scale of the target person. Therefore, wrong inference in this step leads to a misinterpretation of the position and scale in the final target. The refined semantic target map is retrieved from the pre-partitioned clusters based on encoded features of the coarse semantic map in stage 2. Consequently, an extremely rough generation in stage 1 or a misclassified outlier during clustering in stage 2 can lead to a generated person that does not blend well with the existing persons in the scene. Finally, due to a supervised approach of training the renderer in stage 3, the appearance attribute

Figure 6.12: Generating pose variations with the proposed method. **Top row:** Coarse estimation (*in purple*) followed by *Top-5* refined estimations (*in green*). **Bottom row:** Original scene followed by generated scenes from respective fine semantic maps. An exemplar of the target person provides gender information in the refinement query and appearance attributes to the renderer.

Figure 6.13: Limitations of the proposed method due to probable misinterpretation of the **(a)** *position*, **(b)** *scale*, **(c)** *context*, or **(d)** *rendering*.

transfer may struggle to generate high-quality outputs for imbalanced or unconventional target poses. We show a few such cases in Fig. 6.13.

6.4 Chapter Summary

In this chapter, we proposed a method for adaptive person image generation and blending into a scene. The main objective of the work is to introduce a new person into a given scene

with existing individuals, such that the new instance adopts a contextually coherent pose and meaningfully blends with the existing persons. In the previous chapter, we used the collective association of all existing human poses to represent the global semantic context. However, multiple possibilities of a contextually valid pose for a *non-existent* person often led to substantial ambiguities in the keypoint-based sparse pose estimation. In this work, we addressed such instabilities by replacing keypoint-based poses with *semantic parsing maps* and adopting a *data-agnostic* approach.

The proposed architecture consists of three independently learnable networks for (a) coarse generation, (b) refinement, and (c) rendering. First, we use an *image-to-image* translation method to estimate a probable target parsing map from the global context. While the spatial characteristics of the generated coarse mask provide sufficient geometric information about the potential target, it does not preserve enough label group correctness for a proper attribute transfer in the rendering stage. We mitigate this issue through a *data-agnostic* distillation of the coarse semantic map by selecting fine precomputed candidate maps from a clustered *knowledge base* using a similarity score-based ranking. Finally, the appearance attributes from an observed exemplar are transferred to the chosen candidate semantic map using a generative renderer. The rendered instance is then injected into the original scene using the geometric information obtained during coarse generation.

Although the proposed method produces visually intriguing results, there are many scopes for improvements. The refinement strategy requires maintaining a database of precomputed fine semantic mask images. Therefore, the existing pose diversity in the database limits the realistically attainable pose variations in the generated images. While we can easily address this problem by expanding the database with additional sample variations and hierarchically reclustering it, such horizontal scaling requires gradually increasing storage capacity. Another crucial aspect is the representation of global semantic context as a collection of existing human poses in the scene, assuming all existing individuals follow similar postures. However, such assumptions only hold for a group activity, such as photo shoots or dance. Therefore, generating random independent human activities requires more intricate contextual supervision involving other non-human elements of the scene. In the next chapter, we explore an improved semantic representation to address these limitations.

CHAPTER 7

CONTEXT-AWARE HUMAN AFFORDANCE GENERATION

In the previous two chapters, we explored adaptive person instance generation in a complex scene by expressing the global semantic context as a collective association of all existing human poses within the scene. The structural instabilities in the initial approach are later addressed with a data-agnostic strategy, producing remarkable visual results. However, such techniques function under a fundamental assumption that all existing individuals in the scene follow a common group activity. Consequently, this limits the applicability of these methods to generate independent human interactions that require intricate contextual supervision from other non-human elements of the scene. This chapter introduces an improved global semantic representation for complex human affordance generation by mutually cross-attending two feature modalities, achieving substantial performance gains over the existing approaches.

7.1 Semantic Context for Human Affordance Generation

The original conception [139] of relating perception with action describes affordance as the *opportunities for interaction with the environment*. In computer vision, human affordance prediction involves probabilistic modeling of novel human actions, such that the estimated pose interprets a semantically meaningful interaction with the environment. The task is fundamental to many vision problems, such as machine perception, robot navigation, scene understanding, contextually sound novel human pose generation, and content creation. However, predicting a contextually relevant valid pose for a non-existent

Figure 7.1: An overview of the proposed method. **Left:** Predicted locations for a new person in the scene. **Middle:** Estimated scale at each predicted location. **Right:** Final human pose estimated after scaling and deformation at each predicted location.

person is extremely challenging because, unlike the generic pose estimation task, we do not have an actual human body for supervision. So, in this case, the generator has to rely exclusively on the environmental semantics, requiring an intricate focus on the scene context representation.

Earlier works on affordance-aware human pose generation have explored knowledge base representation [146], social reasoning constraints [148], variational autoencoder [143], adversarial learning [141, 144], and transformer [145]. However, the majority of existing methods investigate different network design philosophies while putting less emphasis on the contextual representation of the scene. Unlike 3D, the lack of rich information about the surrounding environment in a 2D scene makes the problem particularly challenging and requires a robust representation of the scene context.

In this work, we proposed an affordance-aware human pose generation method in complex 2D indoor scenes by introducing a novel context representation technique leveraging cross-attention between two spatial modalities. The key idea is to mutually attend the convolution feature spaces from the scene image and the corresponding semantic segmentation map to get a modulated encoding of the scene context. Initially, we estimate a probable location where a person can be centered using a VAE conditioned on the global context encoding of the entire scene. After determining a potential center, local context embeddings are computed from square patches centered around that location at multiple scales. Next, we use a classifier on the multi-scale context vectors to predict the most likely pose as the template class from a set of existing human pose candidates. Finally, we use two VAEs conditioned on the context embeddings and the template class to sample the scale and linear deformations separately. The estimated scale and deformations are applied to the predicted pose template to get the target pose. Fig. 7.1 illustrates an overview of the proposed method.

The main contributions of the proposed work are summarized as follows.

- We proposed an affordance-aware human pose generation method in complex 2D indoor scenes by introducing a novel scene context representation technique that leverages mutual cross-attention between two spatial modalities for robust semantic encoding.

- Unlike the existing methods where the target position is user-defined, the proposed method utilizes global scene context to sample probable locations for the target, which provides additional flexibility for constructing a fully automated human affordance generation pipeline.

7.2 Context-Aware Human Affordance Generation with Mutual Attention

Generating a semantically meaningful body pose of a non-existent person is challenging. Because, unlike a traditional pose estimation approach, an actual human is not present for estimating the body joints, which forces the generator to depend exclusively on the scene context. However, directly using the spatial feature maps from one [143, 147, 145] or multiple [148, 141, 144] modalities does not provide a comprehensive encoding of the scene context, thereby limiting the sampling quality of the generator. Our key idea is a *bidirectional cross-attention* mechanism between feature spaces of two modalities for a modulated scene context representation. In particular, we use an ImageNet-pretrained [56] VGG-19 model [159] to estimate the convolution feature maps from the scene image and the corresponding segmentation mask, followed by mutually cross-attending the two feature spaces. The proposed affordance generation pipeline consists of four stages. In the first stage, we use a conditional VAE to estimate a probable location within the scene where a person can be centered. In the second stage, a classifier predicts the most likely template pose for the estimated location from a set of existing human pose candidates. In the subsequent stages, we use two dedicated conditional VAEs to sample the scale and linear deformation parameters for the predicted template. Fig. 7.3 illustrates the proposed architecture.

7.2.1 Context Representation

Conceptually, the attention mechanism [182] dynamically increases the receptive field in a network architecture [183]. Unlike *self-attention*, which combines two similar embedding

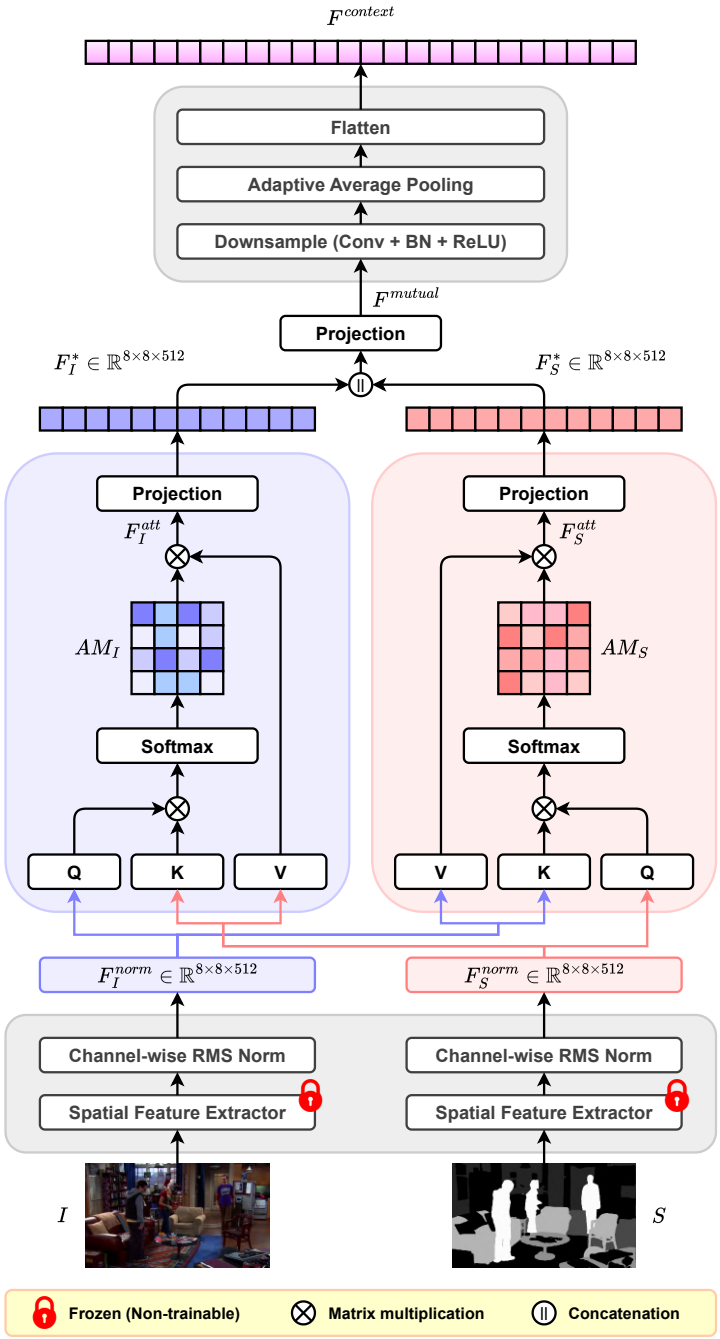

Figure 7.2: Architecture of the proposed *Mutual Cross-Modal Attention* (**MCMA**) block.

spaces, *cross-attention* asymmetrically combines two separate embedding spaces. More specifically, self-attention computes the attention maps from queries Q, keys K, and values V projected from the same token space, while cross-attention uses K and V projected from a different token space as the contextual guidance. However, the underlying computation

steps are identical in both cases and involve a similarity measure between Q and K to form an attention matrix AM, followed by weighing V with AM to obtain the updated query tokens.

In the proposed cross-attention mechanism, we use an image I and the corresponding semantic segmentation map S as inputs from two different modalities for mutually attending to each other. The segmentation maps are estimated from respective images using OneFormer [184] with DiNAT-L [185] backbone pretrained on the ADE20K dataset [186]. We reduce the initial 150 semantic labels to 8 most significant categories to represent indoor scenes as grayscale semantic maps with the following values – wall (36), floor (72), stairs (108), table (144), chair (180), bed (216), person (252), and background/everything else (0). Initially, we resize I and S with a rescaling function $f : \mathbb{R}^{h\times w\times 3} \rightarrow \mathbb{R}^{256\times 256\times 3}$. The resized images are passed through the convolutional backbone of an ImageNet-pretrained [56] frozen VGG-19 model [159] to extract the respective spatial feature maps F_I^{conv} and F_S^{conv}, $g : \mathbb{R}^{256\times 256\times 3} \rightarrow \mathbb{R}^{8\times 8\times 512}$. The resulting feature maps F_I^{conv} and F_S^{conv} are normalized with channel-wise root mean square layer normalization (RMSNorm) [187] to obtain the normalized feature maps F_I^{norm} and F_S^{norm}. Mathematically, the normalization operation defines a function $\phi : \mathbb{R}^{8\times 8\times 512} \rightarrow \mathbb{R}^{8\times 8\times 512}$ as

$$F^{norm} = \phi\left(F^{conv}\right) = \gamma \frac{F^{conv}\sqrt{N_c}}{max\left(\|F^{conv}\|_2, \epsilon\right)}$$

where N_c denotes the number of channels in feature map F^{conv}, $\epsilon \in \mathbb{R}^n$ is a very small constant ($\approx 1e^{-12}$) to provide numerical stability, and $\gamma \in \mathbb{R}^n$ is a learnable gain parameter set to 1 at the beginning.

For computing the multi-head attention maps, we first define three projection functions q, k, and v to map the normalized feature maps F^{norm} into queries Q, keys K, and values V as

$$Q = q\left(F^{norm}\right) : \mathbb{R}^{8\times 8\times 512} \rightarrow \mathbb{R}^{8\times 8\times N_{embed}}$$

$$K = k\left(F^{norm}\right) : \mathbb{R}^{8\times 8\times 512} \rightarrow \mathbb{R}^{8\times 8\times N_{embed}}$$

$$V = v\left(F^{norm}\right) : \mathbb{R}^{8\times 8\times 512} \rightarrow \mathbb{R}^{8\times 8\times N_{embed}}$$

where the projection functions use point-wise convolution with 1×1 kernels and zero bias, and N_{embed} is the length of the cumulative embedding space of all attention heads. In the proposed architecture, we use 8 attention heads of dimension 64 each, resulting $N_{embed} = 8 \times 64 = 512$. Next, we update the feature maps from both modalities I and S

by computing the respective cross-attention matrices, taking Q from one modality and K & V from the other. Mathematically,

$$F_I^{att} = softmax\left(\frac{Q_I K_S^T}{\sqrt{d}}\right) V_S$$

$$F_S^{att} = softmax\left(\frac{Q_S K_I^T}{\sqrt{d}}\right) V_I$$

where d is a scaling parameter. In the proposed method, we take the attention head dimension as d.

The final updated cross-modal feature maps are estimated as $F_I^* = p(F_I^{att})$ and $F_S^* = p(F_S^{att})$, where the function $p : \mathbb{R}^{8\times8\times N_{embed}} \rightarrow \mathbb{R}^{8\times8\times512}$ defines a projection that uses convolution with 1×1 kernels and zero bias.

Finally, we compute the mutual cross-modal feature maps as $F^{mutual} = t(c(F_I^*, F_S^*))$ by a channel-wise concatenation $c : \mathbb{R}^{8\times8\times512} \times \mathbb{R}^{8\times8\times512} \rightarrow \mathbb{R}^{8\times8\times1024}$ of F_I^* and F_S^*, followed by a projection $t : \mathbb{R}^{8\times8\times1024} \rightarrow \mathbb{R}^{8\times8\times512}$ using convolution with 1×1 kernels and zero bias.

The vectorized representation of the contextual embedding $F^{context} \in \mathbb{R}^{8192}$ is obtained by first downsampling the 8×8 feature maps of F^{mutual} into a spatial resolution of 4×4 with strided convolution (4×4 kernel, stride = 2, padding = 1, bias = 0), then following a set of sequential operations – batch normalization, ReLU activation, 4×4 adaptive average pooling, and flattening. We use $F^{context}$ as the semantic condition in the proposed architecture. The architecture of the proposed *mutual cross-modal attention* (**MCMA**) block is illustrated in Fig. 7.2.

7.2.2 Estimating Locations of Non-existent Persons

Unlike the existing approaches, where the location of the target person is user-specified, we attempt to estimate the probable locations of non-existent persons automatically to minimize user intervention. However, the problem is challenging as potential human candidates may appear at several locations with varying scales and poses. Therefore, inferring a location directly from spatial feature space is difficult. In the proposed method, we use a VAE to sample a possible location by conditioning the network on the global context embedding $F^{context} \in \mathbb{R}^{2048}$, computed over the entire scene. The encoder and decoder in the VAE architecture use a shared feature space $F^{shared} \in \mathbb{R}^{128}$, obtained by projecting $F^{context}$ into a 128-dimensional vector with a fully connected (FC) layer and ReLU activation.

In the encoder network, the 2D location coordinates $o(x, y) \in \mathbb{R}^2$ is first mapped into a 128-dimensional vector using two consecutive FC layers having ReLU activations, and the output is linearly concatenated with F^{shared}. We compute the mean $\mu \in \mathbb{R}^{32}$ and variance $\sigma \in \mathbb{R}^{32}$ of the latent distribution $P(\mu, \sigma)$ by projecting the concatenated feature vector through two separate FC layers. The latent embedding vector z is obtained using the reparameterization technique [82] as $z = \mu + \sigma \odot \epsilon$, where the values of $z \in \mathbb{R}^{32}$ are sampled from the estimated distribution $P(z \mid \mu, \sigma)$ and the values of $\epsilon \in \mathbb{R}^{32}$ are sampled from the normal distribution $\mathcal{N}(0, 1)$.

In the decoder network, the latent embedding z is projected into a 128-dimensional space using two consecutive FC-ReLU layers. The shared feature vector F^{shared} also receives an additional projection into a 128-dimensional space through a single FC-ReLU layer. The two projected vectors are linearly concatenated and passed over another 128-dimensional FC-ReLU layer. Finally, we use one more FC layer to project the feature space into the predicted 2D location coordinates $\overline{o}(x, y) \in \mathbb{R}^2$.

The optimization objective for the network consists of two loss components. To measure the spatial deviation, we compute the mean squared error ($\mathscr{L}_{MSE}$) between the 2D target and predicted locations $o(x, y)$ & $\overline{o}(x, y)$ as follows.

$$\mathscr{L}_{MSE} \;=\; \left\| \,\overline{o}\,(x, y) \;-\; o\,(x, y) \,\right\|_2$$

To evaluate the statistical difference between the estimated probability distribution of the embedding space $P(z \mid \mu, \sigma)$ and the normal distribution $\mathcal{N}(0, 1)$, we compute the Kullback–Leibler divergence [167] ($\mathscr{L}_{KLD}$) between the two distributions as follows.

$$\mathscr{L}_{KLD} \;=\; KL\big(P\left(z \mid \mu, \sigma\right) \parallel \mathcal{N}\,(0, 1)\big)$$

We update the network parameters by minimizing the cumulative objective $\mathscr{L} = \mathscr{L}_{MSE} + \mathscr{L}_{KLD}$ using stochastic Adam optimizer [160], keeping a fixed learning rate of $1e^{-3}$ and β-coefficients at (0.5, 0.999). During inference, we estimate a probable center $o^*(x, y)$ for a non-existent person by using a random noise $\eta \in \mathbb{R}^{32}$, with values sampled from the normal distribution $\mathcal{N}(0, 1)$, and the shared context embedding $F^{shared} \in \mathbb{R}^{128}$ as inputs to the VAE decoder $\mathscr{D}^{vae}$. Formally, $o^*(x, y) = \mathscr{D}^{vae}(\eta, F^{shared})$, $\eta_i \sim \mathcal{N}(0, 1)$ for $\eta_i \in \eta$.

7.2.3 Finding Pose Templates for Non-existent Persons

Unlike the conventional human pose estimation techniques [154, 188, 189, 190], directly inferring the valid pose of a non-existent person is difficult [27] due to the unavailability

of an actual human body for supervision. Thus, after sampling a probable location $o^*(x, y)$ within the scene where a person can be centered, we select a potential candidate from an existing set of m valid human poses as the initial guess (template) at that position. In the next stage, the template pose is scaled and deformed to estimate the target pose. The candidate pool is constructed using the K-medoids algorithm [191] to select m representative pose templates from all the available human poses in the training data. Each template class is represented as a m-dimensional one-hot embedding vector $y \in \mathbb{R}^m$. To estimate the class probabilities $\overline{y} \in \mathbb{R}^m$ for an expected pose, we use a multi-class classifier on the context embedding.

The global cross-modal feature maps $F^{mutual} \in \mathbb{R}^{8\times8\times512}$ are estimated over the entire scene as discussed in Sec. 7.2.1. In addition to the global features, we compute two local feature representations $F_A^{mutual} \in \mathbb{R}^{8\times8\times512}$ and $F_B^{mutual} \in \mathbb{R}^{8\times8\times512}$ over two square patches A and B, centered at the initially sampled location $o^*(x, y)$. The size of patch A is equal to the scene height, and the size of patch B is half of the scene height. A channel-wise concatenation of the global and local feature maps represents the cumulative feature space $F_*^{mutual} \in \mathbb{R}^{8\times8\times1536}$. We derive the combined context embedding vector $F_*^{context} \in \mathbb{R}^{8192}$ by first downsampling F_*^{mutual} into a spatial dimension of $\mathbb{R}^{4\times4\times512}$ with strided convolution (4×4 kernel, stride = 2, padding = 1, bias = 0), then following a set of sequential operations similar to Sec. 7.2.1 – batch normalization, ReLU activation, 4×4 adaptive average pooling, and flattening. Finally, the classifier predicts the multi-class probabilities by projecting the context embedding $F_*^{context}$ into a m-dimensional output vector $\overline{y} \in \mathbb{R}^m$ using an FC-layer, followed by $softmax$ activation.

The optimization objective of the classifier is a multi-class (categorical) cross-entropy loss ($\mathscr{L}_{CCE}$), which is formally defined using the negative log-likelihood as follows.

$$\mathscr{L}_{CCE} = -\sum_{i=1}^{m} y_i \, log\left(\overline{y}_i\right)$$

where $y_i \in y$ and $\overline{y}_i \in \overline{y}$ denote the probabilities of i-th class, $i \in \{1, ..., m\}$, in the target (y) and predicted ($\overline{y}$) label vectors, respectively.

We update the network parameters by minimizing $\mathscr{L}_{CCE}$ using stochastic Adam optimizer [160], keeping a fixed learning rate of $1e^{-3}$ and β-coefficients at (0.5, 0.999). During inference, we obtain the one-hot pose template class embedding $y^* \in \mathbb{R}^m$ by selecting the predicted class with the highest probability. Formally, $y^* = argmax(\overline{y})$.

7.2.4 Scaling the Selected Pose Template

After inferring a probable location $o^*(x, y)$ and one-hot pose template class embedding $y^* \in \mathbb{R}^m$, we use a conditional VAE to sample the expected scaling factors (height and width) of the target person. The estimated parameters are used to rescale the normalized pose template of unit height and width into the target dimensions. The encoder and decoder networks of the VAE use a shared feature space $F_*^{shared} \in \mathbb{R}^{128}$ as the condition, which is derived from the cumulative context vector $F_*^{context} \in \mathbb{R}^{2048}$ and pose template class embedding $y^* \in \mathbb{R}^m$. As discussed in Sec. 7.2.3, $F_*^{context}$ encodes both global (over entire scene) and local (over localized patches) cross-modal scene context representations. To compute the shared feature space, we first project y^* into a 128-dimensional vector by two consecutive FC-ReLU layers and then linearly concatenate the projected output with $F_*^{context}$. The concatenated vector is passed through another FC-ReLU layer to obtain the 128-dimensional shared feature representation F_*^{shared}.

The encoder and decoder of the VAE adopt a similar architecture as the location estimator discussed in Sec. 7.2.2. The encoder takes the scaling parameters $s(\Delta x, \Delta y) \in \mathbb{R}^2$ and shared feature vector $F_*^{shared} \in \mathbb{R}^{128}$ as inputs to predict the mean $\mu_s \in \mathbb{R}^{32}$ and variance $\sigma_s \in \mathbb{R}^{32}$ of the latent distribution $P_s(\mu_s, \sigma_s)$. The decoder takes the latent embedding $z_s \in \mathbb{R}^{32}$ and shared feature vector $F_*^{shared} \in \mathbb{R}^{128}$ as inputs to estimate the probable scaling parameters $\overline{s}(\Delta x, \Delta y) \in \mathbb{R}^2$, where z_s is computed using the reparameterization technique as $z_s = \mu_s + \sigma_s \odot \epsilon$, $z_s \sim P_s(\mu_s, \sigma_s)$, and $\epsilon_i \sim \mathcal{N}(0, 1)$ for $\epsilon_i \in \epsilon$.

The optimization objective for the network consists of two loss components. The first term measures the spatial deviation $\mathscr{L}_{MSE}$ between $s(\Delta x, \Delta y)$ and $\overline{s}(\Delta x, \Delta y)$ as the L_2 norm. The second term estimates the statistical difference $\mathscr{L}_{KLD}$ between $P_s(z_s \mid \mu_s, \sigma_s)$ and $\mathcal{N}(0, 1)$ as the KL divergence. Mathematically, the loss terms can be represented as follows.

$$\mathscr{L}_{MSE} = \left\| \overline{s}\left(\Delta x, \Delta y\right) - s\left(\Delta x, \Delta y\right) \right\|_2$$

$$\mathscr{L}_{KLD} = KL\left(P_s\left(z_s \mid \mu_s, \sigma_s\right) \parallel \mathcal{N}(0, 1)\right)$$

We update the network parameters by minimizing the cumulative objective $\mathscr{L} = \mathscr{L}_{MSE} + \mathscr{L}_{KLD}$ using stochastic Adam optimizer, with a fixed learning rate of $1e^{-3}$ and β-coefficients of (0.5, 0.999). During inference, we estimate the probable scaling factors $s^*(\Delta x, \Delta y)$ for the selected pose template by using a random noise $\eta \in \mathbb{R}^{32}$, with values sampled from $\mathcal{N}(0, 1)$, and the shared scene context embedding $F_*^{shared} \in \mathbb{R}^{128}$ as inputs to the VAE decoder $\mathscr{D}_s^{vae}$. Formally, $s^*(\Delta x, \Delta y) = \mathscr{D}_s^{vae}(\eta, F_*^{shared})$, $\eta_i \sim \mathcal{N}(0, 1)$ for $\eta_i \in \eta$.

7.2.5 Deforming the Selected Pose Template

The potential human pose at the sampled location $o^*(x, y)$ is estimated by applying a linear deformation on the chosen pose template. A linear deformation vector d is the set of distances between the cartesian coordinates of each body keypoint. Assuming a human pose is represented with r keypoints, we define $d = \{dx_1, dy_1, ..., dx_r, dy_r\} \in \mathbb{R}^{2r}$, where (dx_j, dy_j) denotes the differences between the template and target coordinates along x and y axes for the j-th keypoint, $1 \leqslant j \leqslant r$. In the proposed method, we represent a human pose with 16 major body joints ($r = 16, d \in \mathbb{R}^{32}$), following the MPII [165] keypoint format.

The deformation parameters are estimated using an identical VAE architecture as discussed in Sec. 7.2.4. The encoder takes the context embedding $F_*^{shared} \in \mathbb{R}^{512}$ and linear deformations $d \in \mathbb{R}^{32}$ as inputs to predict the mean $\mu_d \in \mathbb{R}^{32}$ and variance $\sigma_d \in \mathbb{R}^{32}$ of the latent distribution $P_d(\mu_d, \sigma_d)$. The decoder takes $F_*^{shared} \in \mathbb{R}^{512}$ and the latent vector $z_d \in \mathbb{R}^{32}$ as inputs to estimate the probable deformation parameters $\overline{d} \in \mathbb{R}^{32}$, where $z_d = \mu_d + \sigma_d \odot \epsilon$, $z_d \sim P_d(\mu_d, \sigma_d)$, and $\epsilon_i \sim \mathcal{N}(0, 1)$ for $\epsilon_i \in \epsilon$.

The network parameters are optimized by minimizing the cumulative objective $\mathcal{L} = \mathcal{L}_{MSE} + \mathcal{L}_{KLD}$ using stochastic Adam optimizer, with a fixed learning rate of $1e^{-3}$ and β-coefficients of (0.5, 0.999). Mathematically,

$$\begin{aligned} \mathcal{L}_{MSE} &= \left\| \overline{d} - d \right\|_2 \\ \mathcal{L}_{KLD} &= KL\big(P_d\left(z_d \mid \mu_d, \sigma_d\right) \parallel \mathcal{N}(0, 1)\big) \end{aligned}$$

The probable deformation parameters $d^* \in \mathbb{R}^{32}$ for the selected pose template is inferred by using a random noise $\eta \in \mathbb{R}^{32}$, with values sampled from $\mathcal{N}(0, 1)$, and the shared scene context embedding $F_*^{shared} \in \mathbb{R}^{512}$ as inputs to the VAE decoder $\mathcal{D}_d^{vae}$. Formally, $d^* = \mathcal{D}_d^{vae}(\eta, F_*^{shared})$, $\eta_i \sim \mathcal{N}(0, 1)$ for $\eta_i \in \eta$.

7.2.6 Target Transformation

Assuming a normalized human pose template of unit scale $h^*(x_1, y_1, ..., x_r, y_r)$ corresponding to the predicted pose template class y^*, we compute the target pose $\overline{h}(\overline{x}_1, \overline{y}_1, ..., \overline{x}_r, \overline{y}_r)$ from the estimated center $o^*(x_0, y_0)$, scaling factors $s^*(\Delta x, \Delta y)$, and linear deformations $d^*(dx_1, dx_2, ..., dx_r, dy_r)$ as follows.

$$\begin{aligned} \overline{x}_i &= \frac{w}{w_0}\left[(x_i \Delta x + dx_i) + \left(x_0 - \frac{\Delta x}{2}\right)\right] \\ \overline{y}_i &= \frac{h}{h_0}\left[\left(y_i \Delta y + dy_i\right) + \left(y_0 - \frac{\Delta y}{2}\right)\right] \end{aligned}$$

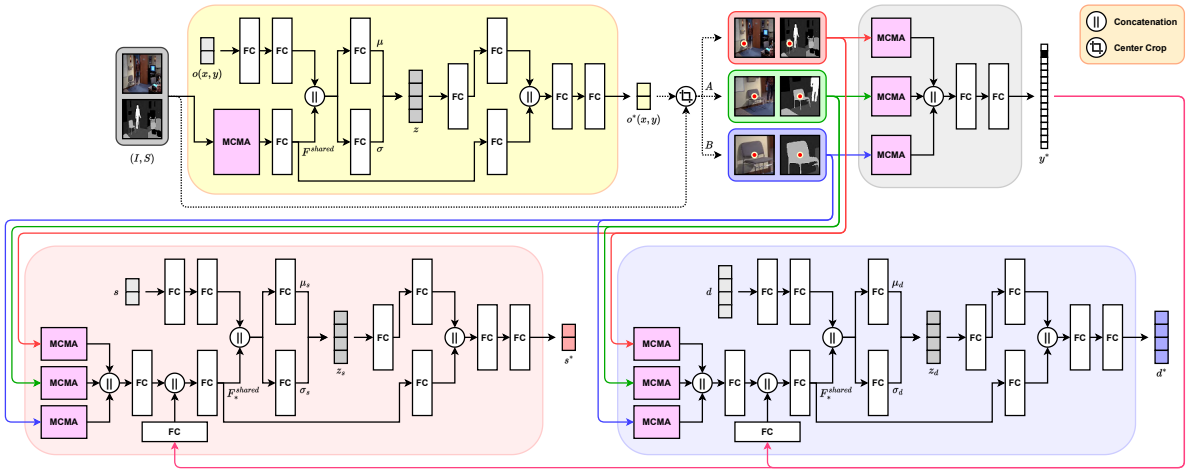

Figure 7.3: Architecture of the proposed pipeline. The workflow is divided into four subnetworks to estimate the probable location o^*, pose template class y^*, scaling parameters s^*, and linear deformations d^* of a potential target pose. Every subnetwork exclusively uses the proposed *Mutual Cross-Modal Attention* (**MCMA**) block to encode global and local scene contexts as shown in Fig. 7.2.

where $i \in \{1, 2, ..., r\}$, (h_0, w_0) is the rescaled image patch size for network input, and (h, w) is the size of the original scene. In the proposed method, we use $h_0 = w_0 = 256$.

7.3 Experiments

In this section, we discuss the experimental studies performed on the proposed method, including dataset, visual results, evaluation metrics, ablation studies, and limitations. The location, template, scale, and deformation estimation networks consist of 2.9, 24.7, 7.9, and 26.4 million trainable parameters, respectively. We train each network for 100 epochs with a batch size of 8 on a single NVIDIA TITAN X GPU.

7.3.1 Dataset

The number of large-scale annotated public datasets for complex 2D affordance generation is significantly limited in the literature. The main challenge is to obtain a specific frame in two different states – *with* and *without* populated with random persons. Researchers [144, 192] have attempted to resolve the issue by randomly removing and inpainting existing person instances from the scene. However, these datasets [144, 192] are not publicly available, and our attempt to remove persons from a complex scene leaves significant visual artifacts even with state-of-the-art inpainting techniques [193], making such data generation methods [144, 192] insufficient for our purpose.

Following the recent works [143, 145], we train and evaluate the proposed method

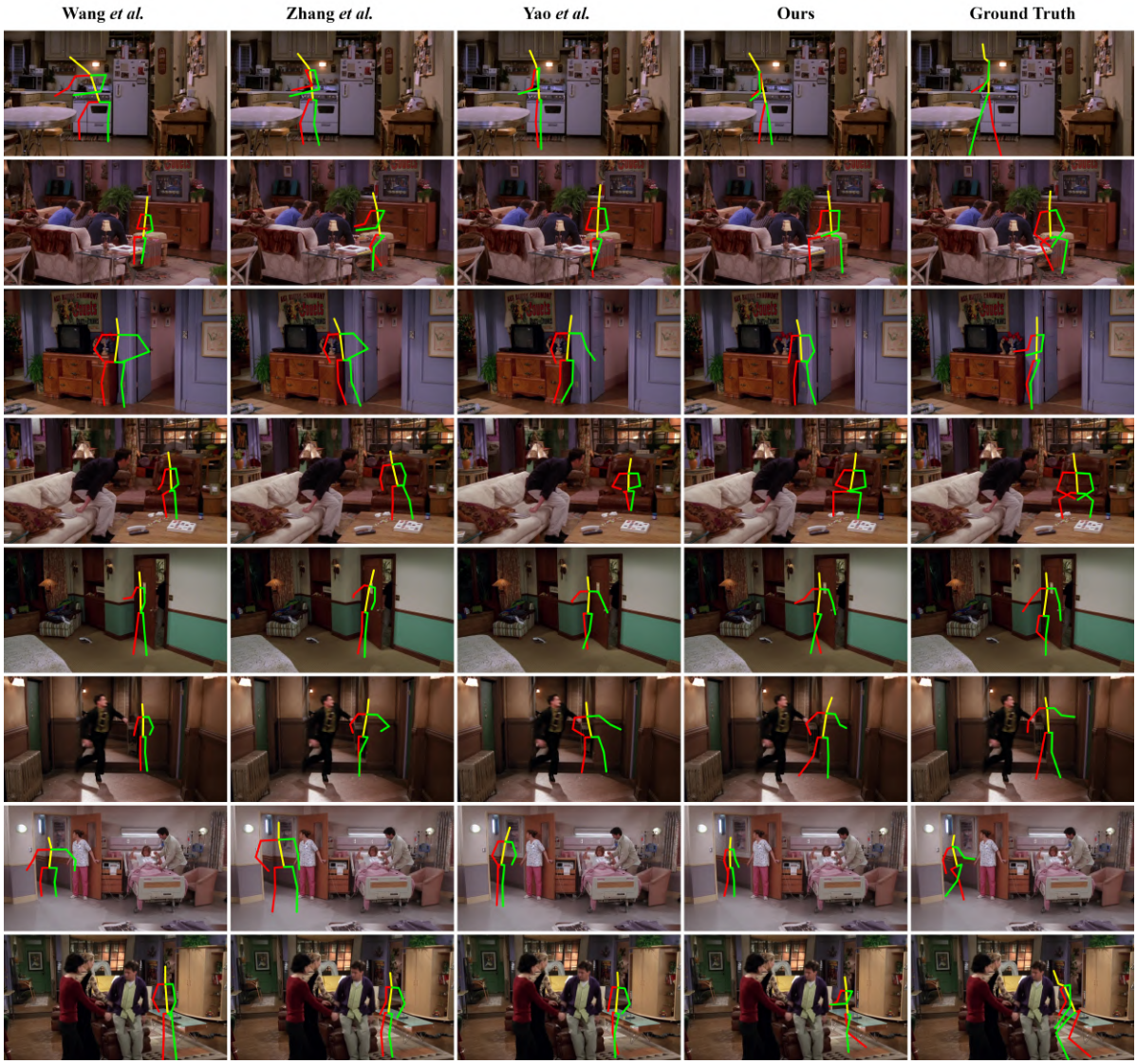

Figure 7.4: Qualitative comparison of the proposed method against existing affordance generation techniques by Wang *et al.* [143], Zhang *et al.* [144], and Yao *et al.* [145].

on an openly available large-scale sitcom dataset [143] for fair comparisons. The dataset comprises 28837 human interaction samples over 11499 video frames extracted from seven sitcoms. The training set consists of 24967 poses over 10009 frames from six sitcoms, while the evaluation set contains 3870 poses over 1490 frames from one sitcom. Each human pose is represented with 16 keypoints following the standard MPII format [165].

7.3.2 Visual Results and Evaluation Metrics

Qualitative analysis: To demonstrate the efficacy of the proposed method, we compare the visual results against major existing works [143, 145, 144] on 2D human affordance

generation. In [143, 145], the networks directly learn from a single modality of image features. In [144], the authors introduce an adversarial learning mechanism with two additional modalities of segmentation and depth maps alongside image features. We notice that adopting a transformer-based architecture [145] or combining multiple modalities in an adversarial learning method [144] provide only marginal improvements over the baseline approach [143]. In contrast, the proposed method focuses on imposing a better semantic constraint in the learning strategy by introducing a novel cross-attention mechanism. Fig. 7.4 illustrates a qualitative comparison of the proposed method with existing human affordance generation techniques [143, 145, 144]. The visual results show that our approach produces more realistic human interactions in complex scenes than previous methods.

Quantitative evaluation: We evaluate the alignment of the generated pose with the ground truth for analytically comparing the proposed method against existing human affordance generation techniques [143, 145, 144]. In particular, we use the percentage of correct keypoints (PCK/PCKh) [194], average keypoint distance (AKD), mean absolute error (MAE), mean squared error (MSE), and cosine similarity (SIM) as the evaluation metrics for pose alignment. To estimate the correctness of the estimated scale, we measure the intersection over union (IOU) between the target and predicted pose bounding boxes. Following [145], our comparative study also includes evaluation results from additional baseline methods (heatmap and regression), pose estimation techniques [195, 196], and object placement algorithms [197, 198]. Additionally, we train and evaluate another variation of the proposed architecture by replacing ***semantic segmentation maps*** with ***depth maps***. **PCK** and **PCKh** measure the similarity between two poses by computing the fraction of correctly aligned keypoints, where a valid alignment denotes that the distance between two respective keypoints is within a predetermined tolerance. PCK uses $\alpha * torso\ width$, while PCKh uses $\beta * head\ size$ as the tolerance threshold, where $0 < \alpha, \beta \leqslant 1$. **AKD** is the average Euclidean distance between respective pairs of target and predicted keypoints. **MAE** and **MSE** measure the average absolute and squared linear deviations, respectively, along both coordinate axes between all respective pairs of actual and inferred keypoints. **SIM** evaluates the average cosine similarity between the positional vectors corresponding to every target and predicted keypoint pair. **IOU** enumerates the intersection over union ratio between the bounding rectangles of predicted and target poses. We summarize the quantitative evaluation scores for each competing method in Table 7.1. The proposed method exhibits significantly better evaluation scores, reflecting the apparent visual superiority of the qualitative analysis.

Table 7.1: Quantitative comparison of the proposed method against existing pose estimation [195, 196], object placement [197, 198], and affordance generation [143, 145, 144] techniques. The best scores are in **bold**, and the second-best scores are underlined.

Method	PCK ↑	PCKh ↑	AKD ↓	MAE ↓	MSE ↓	SIM ↑	IOU ↑	User Score ↑
Heatmap	0.363	0.422	11.298	7.112	53.45	0.9864	0.402	0.000
Regression	0.386	0.451	10.929	6.840	51.29	0.9899	0.426	0.000
UniPose [195]	0.387	0.447	9.966	6.241	46.78	0.9918	0.471	0.012
PRTR [196]	0.408	0.474	9.724	6.088	45.72	0.9934	0.489	0.025
PlaceNet [197]	0.060	0.072	79.978	50.325	377.78	0.9476	0.118	0.000
GracoNet [198]	0.380	0.441	10.576	6.614	49.60	0.9912	0.427	0.000
Wang *et al.* [143]	0.401	0.462	9.940	6.208	46.65	0.9928	0.482	0.022
Zhang *et al.* [144]	0.372	0.428	10.252	6.409	48.14	0.9906	0.405	0.005
Yao *et al.* [145]	0.414	0.479	9.514	5.918	44.86	0.9954	0.494	0.104
Ours *(Depth)*	0.407	0.472	6.680	4.163	32.78	0.9966	**0.566**	0.205
Ours *(Semantic)*	**0.433**	**0.503**	**6.352**	**3.972**	**29.81**	**0.9969**	**0.566**	**0.299**
Ground Truth	1.000	1.000	0.000	0.000	0.00	1.0000	1.000	0.328

Subjective evaluation (User study): While the evaluation metrics are analogous to visual rationality for most cases in our experiments, the quantitative scores alone may not be sufficient to claim the superiority of a generation scheme. The reason is the potential uncertainty of a generated human pose. Within the given scene context, it is possible to obtain multiple valid human interactions with the environment. Therefore, the evaluation scores against a specific ground truth pose may not always provide a rational judgment of superiority. We have conducted an opinion-based user study with 42 volunteers to select the most visually realistic sample from a pool of images generated by the competing methods, including ground truth. The mean opinion score **(MOS)** is estimated as the average fraction of times a method is preferred over the others. As shown in Table 7.1, the proposed approach achieves a significantly higher user preference over other existing methods, indicating the best semantic integrity in the generated samples similar to the ground truth.

Visualization of the learned distribution: Due to many feasible outcomes, there are substantial ambiguities when inferring a possible pose at a random location within the scene, depending on the surrounding context. The ground truth is only one of the many possibilities. For example, a selected position in front of a chair can lead to either *standing* or *sitting* postures. Therefore, the association between an estimated pose and scene objects is an ideal way to visualize the sampled pose distribution. We discover two broad pose categories, *standing* and *sitting*, in the dataset by manually inspecting the given templates. To visualize the learned distribution, we randomly sample 10000 poses for a

Figure 7.5: Visualization of the learned distribution for each pose category. **(Left)** Input scene. **(Middle)** Distribution of the *standing* poses. **(Right)** Distribution of the *sitting* poses.

scene and assign a pose category (*standing* or *sitting*) to each sampled location based on the predicted pose template at that position. Then, we visualize the bivariate distribution of the sampled coordinates for each pose category. Fig. 7.5 shows a few such visualizations, exhibiting ideal associations between a pos type and the scene objects.

7.3.3 Ablation Study

In the ablation analysis, we evaluate ten different network configurations by varying attention mechanisms in the proposed MCMA block and altering contextual modalities (*depth map* or *semantic map*) to determine an optimal network architecture. The `Baseline` architecture entirely excludes contextual supervision by removing all MCMA blocks from the proposed architecture. Configurations `Self-I`, `Self-D`, and `Self-S` also drop all MCMA blocks but introduce *self-attention* on a single modality. In particular, `Self-I` uses self-attention on the image I itself, while `Self-D` and `Self-S` use self-attention on the *depth map* D and *semantic segmentation map* S, respectively. The next four architectures `Cross-D/I`, `Cross-I/D`, `Cross-S/I`, and `Cross-I/S` introduce *cross-attention* from an additional input stream. Specifically, for configurations `Cross-D/I` and `Cross-S/I`, we compute query (Q) from D and S, respectively, while retrieving key (K) and value (V) from

Figure 7.6: Qualitative ablation analysis on the impact of different auxiliary inputs. **(a)** Depth map. **(b)** Estimated bounding region from depth context. **(c)** Estimated pose from depth context. **(d)** Semantic map. **(e)** Estimated bounding region from semantic context. **(f)** Estimated pose from semantic context.

Table 7.2: Quantitative ablation analysis among different **MCMA** block configurations.

Model	Context	PCK ↑	PCKh ↑	AKD ↓	MAE ↓	MSE ↓	SIM ↑	IOU ↑
`Baseline`	None	0.274 ±0.006	0.322 ±0.008	9.998 ±0.042	6.248 ±0.034	48.920 ±0.408	0.9845 ±0.0005	0.398 ±0.004
`Self-I`	Image	0.346 ±0.004	0.399 ±0.005	7.899 ±0.035	4.925 ±0.025	38.714 ±0.406	0.9947 ±0.0004	0.458 ±0.012
`Self-D`	Depth	0.317 ±0.004	0.372 ±0.006	8.582 ±0.025	5.625 ±0.028	42.104 ±0.344	0.9888 ±0.0002	0.406 ±0.002
`Cross-D/I`	Image	0.341 ±0.004	0.387 ±0.008	8.146 ±0.032	5.077 ±0.028	39.972 ±0.228	0.9927 ±0.0005	0.447 ±0.006
`Cross-I/D`	Depth	0.374 ±0.005	0.429 ±0.005	7.260 ±0.047	4.424 ±0.025	35.649 ±0.301	0.9952 ±0.0002	0.498 ±0.002
`Mutual-I+D`	Image + Depth	0.407 ±0.003	0.472 ±0.002	6.680 ±0.031	4.163 ±0.020	32.782 ±0.294	0.9966 ±0.0001	0.566 ±0.004
`Self-S`	Semantic	0.336 ±0.006	0.391 ±0.009	8.174 ±0.038	5.112 ±0.024	38.452 ±0.404	0.9895 ±0.0001	0.418 ±0.014
`Cross-S/I`	Image	0.355 ±0.005	0.407 ±0.008	7.746 ±0.039	4.904 ±0.018	36.354 ±0.477	0.9950 ±0.0002	0.478 ±0.008
`Cross-I/S`	Semantic	0.398 ±0.006	0.458 ±0.004	6.904 ±0.045	4.325 ±0.029	32.402 ±0.344	0.9958 ±0.0002	0.515 ±0.005
`Mutual-I+S`	Image + Semantic	**0.433** ±0.004	**0.503** ±0.004	**6.352** ±0.036	**3.972** ±0.022	**29.810** ±0.326	**0.9969** ±0.0001	**0.566** ±0.002

I. Likewise, for configurations `Cross-I/D` and `Cross-I/S`, we compute Q from I, while estimating K and V from the contextual input D or S, respectively. The final two network variants `Mutual-D` and `Mutual-S` use the proposed *mutual cross-modal attention* (MCMA) mechanism utilizing D or S as the auxiliary contextual supervision, respectively, alongside I. In all our experiments, we estimate the *depth maps* using a recent monocular depth estimation technique *Depth Anything* [199] and the *semantic segmentation maps* using a recent transformer-based universal image segmentation method *OneFormer* [184].

To measure the spatial alignment of the estimated pose against the ground truth, we evaluate PCK, PCKh, AKD, MAE, MSE, and SIM. Additionally, we compute the IOU between the predicted and target pose bounding boxes to measure the correctness of inferred scales. Table 7.2 summarizes the evaluation scores of every network variant in our ablation study. The analysis shows that additional supervision from other modalities generally results in better performance. Introducing the MCMA block into the architecture further improves this performance gain by a significant margin, reflecting the efficacy of the proposed approach for robust scene context representation. Interestingly, we observe that *semantic segmentation maps* generally perform better than *depth maps* as auxiliary contextual input. We hypothesize that the object labels in a segmentation map provide additional semantic context to the model. Fig. 7.6 shows a visual comparison of the predicted pose from *depth*-context against *semantic*-context.

To investigate the impact of semantic label granularity, pose templates, and dedicated VAEs for scale and deformation on the proposed method, we analyze six additional configurations `A - F` of the architecture. The first four models `A - D` use different numbers of semantic labels in the segmentation maps. Specifically, configuration `A` uses 2 labels for *foreground* (all objects merged) and *background*. Configuration `B` uses 3 labels by dividing the foreground objects into *non-human objects* and *humans*. Configuration `C` uses 4 labels by further splitting the non-human object labels into *fixed* (wall, floor, stairs) and *movable* (table, chair, bed) object categories. Configuration `D` retains all 150 initially estimated semantic labels unaltered. To verify the requirements of multiple pose templates, Configuration `E` drops the classifier from the architecture and uses the first template as a fixed predefined pose. To validate the necessity of 2 dedicated VAEs for separately estimating scale and deformation parameters, Configuration `F` uses a single unified VAE to predict both scale and deformation parameters as a single vector. Finally, Configuration `G` denotes the proposed architecture that uses 8 semantic labels, 30 pose templates with a template classifier, and 2 dedicated VAEs for estimating scale and deformation parameters.

Table 7.3 summarizes evaluation scores for all the network configurations. The results show that using *too few* or *too many* semantic labels does not contribute towards performance improvements. Also, with a fixed template, the linear deviations between the target and template keypoints can vary more randomly. For example, the deviation of a *standing* pose template from a *standing* target pose is often much smaller than from a *sitting* target pose. However, with multiple pose templates, the model estimates a probable pose (template) first and then samples linear deformation parameters from a more predictable range to translate the template into the target pose. Likewise, the

Table 7.3: Quantitative ablation analysis among different network configurations.

Model	Configuration Notes	PCK ↑	PCKh ↑	AKD ↓	MAE ↓	MSE ↓	SIM ↑	IOU ↑
A	2 semantic labels	0.217	0.267	11.941	7.223	52.88	0.9821	0.409
B	3 semantic labels	0.292	0.351	9.112	5.917	45.11	0.9901	0.477
C	4 semantic labels	0.377	0.441	7.019	4.508	37.79	0.9964	0.564
D	150 semantic labels	0.376	0.436	7.313	4.573	40.12	0.9959	0.557
E	fixed template / no classifier	0.352	0.400	7.915	4.841	47.69	0.9951	0.541
F	single VAE for scale + deform	0.369	0.423	7.206	4.523	38.02	0.9961	0.563
G	***proposed***	**0.433**	**0.503**	**6.352**	**3.972**	**29.81**	**0.9969**	**0.566**

possible ranges of scaling and deformation parameters are widely different. The scaling parameters are the height and width of the minimal bounding box around a human pose. These values are much larger than the deformation parameters comprising small linear deviations between two sets of pose keypoints. So, forcing the architecture to infer the scaling and deformation parameters with a single unified VAE causes noticeable instability due to poor normalization. These observations further justify the proposed network design.

7.3.4 Downstream Applications

The ability to sample semantically valid scene-aware complex human poses directly facilitates downstream tasks such as novel person instance generation using off-the-

Figure 7.7: Qualitative results of downstream rendering of human instances. **(Left)** Input scene. **(Middle)** Estimated pose by the proposed method. **(Right)** Rendered person.

shelf pose transfer or pose rendering techniques. Such downstream rendering to inject novel person instances into complex scenes is critical in various application domains, including augmented / virtual reality, digital media, and synthetic data generation. While the state-of-the-art keypoint-based person generation techniques provide high-quality photorealistic rendering, the algorithms also demand precise supervision of the target pose. Therefore, the proposed method must sample a valid and accurate human pose for successful downstream rendering. Fig. 7.7 shows a few examples of rendered persons using an off-the-shelf pose transfer technique PIDM [200] on sampled poses from our method, demonstrating the geometric correctness of the predicted keypoints.

7.3.5 Limitations

Estimating a valid pose for a non-existent person in complex scenes is a fundamentally challenging problem with multiple acceptable solutions other than the ground truth. For example, for a scene containing a bed, a probable pose can be *standing*, *sitting*, or *lying down*, with many feasible variations for each case. While on most occasions, the proposed method generally infers realistic poses with affordance-aware human interactions, there are a few instances when the technique fails to sample an acceptable posture. Alongside

Figure 7.8: Limitations of the proposed method. **(Left)** Auxiliary semantic context. **(Middle)** Estimated bounding region. **(Right)** Estimated human pose.

the advantages of the modularity and flexibility in the disentangled multi-stage approach, a potential error propagation problem also exists. More precisely, inferential error in an earlier stage may propagate through later stages, negatively impacting the overall predictive performance of the pipeline. We show a few examples of such limiting cases in Fig. 7.8, illustrating the misinterpretation in sampled location **(top row)**, scale **(middle row)** or deformation **(bottom row)** by the proposed method.

7.4 Chapter Summary

In this chapter, we investigated semantically constrained human affordance generation in complex environments. Due to many probable outcomes, estimating contextually valid poses for non-existent persons is ambiguous and extremely challenging. Most existing approaches focus on architectural innovations of the network without any significant emphasis on the semantic understanding of the scene. By introducing a novel cross-attention mechanism, we showed that a robust semantic representation of the global scene context can substantially improve the generative performance. In particular, the key idea is to mutually attend the convolution feature spaces of two spatial modalities to obtain a modulated representation of the scene context. The proposed architecture consists of four independently learnable dedicated networks to estimate the probable location, pose template, scale, and liner deformations, providing an automated human affordance generation pipeline. We demonstrated that by integrating with an *off-the-shelf* human pose transformation method, the proposed method achieves remarkable results for semantically adaptive person image generation, addressing structural instabilities and storage overheads in our previous approaches.

CHAPTER 8

CONCLUSIONS AND FUTURE SCOPES

In this concluding chapter, we summarize the key findings and contributions of our research on context-aware person image generation. We also discuss the potential scopes of future research on such generative technologies.

8.1 Key Research Findings and Contributions

In this thesis, we discussed five technical works, where Chapters 3 and 4 focused on human pose transformation from local contexts, followed by Chapters 5, 6, and 7 addressing adaptive person image generation from global semantic contexts. The key research findings are summarized as follows.

- In **Chapter 3**, we introduced an *end-to-end* network architecture [25] for human pose transformation using keypoint-based *local* geometric guidance. The core idea is a *multi-scale attention* mechanism that enhances both *low-frequency* and *high-freaquency* details in the generated images. Our method outperforms the previous techniques in most visual and analytical comparisons. Additionally, we demonstrated that the proposed architecture could be a general *drop-in* solution to many applications other than human pose transformation, such as semantic reconstruction, virtual try-on, font style manipulation, and scene text editing.

- In **Chapter 4**, we proposed a *multi-stage* strategy [26] for human pose transformation using text descriptions as the *local* shape context. The proposed approach addresses

the probable shape irregularities in a keypoint-guided technique when the physical stature of the *target person* widely differs from that of the *target pose provider*. In addition, we compiled a new dataset containing descriptive pose annotations for 40488 human images to alleviate the lack of similar public datasets for benchmarking text-guided human pose transformation techniques.

- In **Chapter 5**, we introduced *global semantic constraints* in a scene-aware person image generation technique [27], aiming to contextually adapt a new person into an environment with multiple existing individuals. We represented the global semantic context as the collective association of all existing human poses in the scene to estimate a probable location, scale, and pose for the new person. Although the proposed strategy shows some promising results, the ambiguity of estimating a *semantically valid* pose for a *non-existent* person often leads to high structural perturbations in the estimated pose. To reduce such instabilities, we explored two different strategies – (a) reducing sparsity in the pose representation (Chapter 6) or (b) improving the global semantic context representation (Chapter 7).

- In **Chapter 6**, we proposed a *semantically adaptive* person image generation strategy [28], improving the visual quality and structural stability over our previous approach in Chapter 5. The proposed method adopts *human parsing maps* instead of a highly sparse keypoint-based pose representation alongside a *data-agnostic* refinement strategy, achieving remarkable visual quality in the generated samples.

- In **Chapter 7**, we explored a *cross-attention* mechanism to improve the global semantic encoding for structurally robust human *affordance generation* [29]. The core idea is a *modulated* context representation by *mutually* attending two different feature modalities. The proposed method demonstrates structurally and contextually robust results, addressing both geometric instabilities (Chapter 5) and storage complexities (Chapter 6) in our earlier approaches.

In this thesis, we have explored strategies for generating novel human instances from local object-level contexts and also investigated semantic constraints to adaptively compose a generated person instance into a complex environment using global scene-level contexts. To bring this study to a close, we now revisit the main research objectives and the associated research questions as outlined in Section 1.3 and summarize our contributions that address each of these questions. By reflecting on each outcome, we demonstrate the

extent to which our research has fulfilled its objectives and offer insights into the broader implications of our findings.

Objective: *To investigate efficient strategies for generating isolated novel views of a specific person from a single observation and local structural context*

- ***Research Question 1:*** How to efficiently improve existing approaches to geometrically guided human pose transformation?
  - We introduced an *end-to-end* network architecture [25] with a cascaded *multi-scale attention* mechanism for human pose transformation using keypoint-based *local* geometric guidance, outperforming the previous techniques in the visual and analytical comparisons.
- ***Research Question 2:*** How does strong structural supervision impact the generative process during real-world inference?
  - We demonstrated [26] that strong structural supervision in the keypoint-guided techniques causes noticeable shape irregularities when the physical stature of the *target person* widely differs from that of the *target pose provider*, leading to unrealistic generation.
- ***Research Question 3:*** How to effectively mitigate the potential structural bias in pose-guided person image generation?
  - We proposed a *multi-stage* strategy [26] for human pose transformation that uses text descriptions as the *local* shape context, effectively mitigating the potential structural bias in pose-guided techniques.

Objective: *To design generative strategies for adaptively blending a specific person into a complex scene by imposing global semantic constraints*

- ***Research Question 4:*** How to effectively introduce semantic conditioning in a scene-aware adaptive person image generation pipeline?
  - We introduced a *global contextual conditioning* strategy in a scene-aware person image generation technique [27], aiming to semantically adapt a new person into an environment with multiple existing individuals.

- ***Research Question 5:*** How does a data-agnostic approach impact the visual quality and scalability of semantic person image generation and composition?
  - We demonstrated that replacing sparse keypoints with *parsing maps* for human pose representation alongside a *data-agnostic* refinement strategy can achieve remarkable visual results and structural stability for *semantically adaptive* person image generation and composition.
- ***Research Question 6:*** How does cross-modal information fusion impact human affordance generation in complex scenes and associated downstream tasks?
  - We proposed a *cross-attention* mechanism by *mutually* attending two spatial feature modalities. This approach effectively improves the global semantic representation for structurally robust human *affordance generation* [29] and associated downstream tasks such as rendering novel human instances.

8.2 Future Scopes

This thesis primarily focuses on generating human *images* from local geometric or global semantic contextual guidance. The immediate next iteration should extend the proposed techniques into *videos* and 3D. Although such attempts need to accommodate additional complexities of *temporal consistency* and polygon mesh, these extensions can facilitate better adoption in many downstream generative tasks across academic research and enterprise applications, including but not limited to synthetic data generation, digital media, retail advertisements, animation, and augmented / virtual reality (AR/VR) software.

We summarize some emerging directions for future research in this domain as follows.

- Person image generation in 3D with adaptive contextual and spatial dynamics to model realistic human poses and gestures.
- Human activity simulation in videos with temporal consistency across dynamic interactions and physical expressions.
- Personality and emotion-aware human image and video generation by integrating behavioral traits into the generative process.
- Human motion modeling for articulated animation generation, path planning, and virtual navigation.

- Exploring improved semantic understanding for dynamic and multimodal context integration into the generative process.

These directions combine advances in machine learning, computer vision, graphics, and human-computer interaction, influencing diverse applications across multiple domains. Finally, we believe the proposed pieces of work and their potential applications will motivate further exploration of context-aware person image generation.

Disclaimer and Ethical Statement: *This research focuses on the generative modeling of human images. All experiments in this thesis use images from open datasets in the public domain. We acknowledge the ethical challenges involving generative models, including critical concerns about transparency, privacy, security, and accountability. We strongly advocate mitigating negative societal impacts and collaborative development of generative technologies with diverse stakeholders to ensure scientific innovation with ethical responsibilities.*

BIBLIOGRAPHY

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative adversarial nets.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 27 (2014).

[2] Mehdi Mirza and Simon Osindero. “Conditional generative adversarial nets.” In: *arXiv preprint arXiv:1411.1784* (2014).

[3] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. “Deep unsupervised learning using nonequilibrium thermodynamics.” In: *The International Conference on Machine Learning (ICML)*. 2015, pp. 2256–2265.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), pp. 6840–6851.

[5] Alexander Quinn Nichol and Prafulla Dhariwal. “Improved denoising diffusion probabilistic models.” In: *The International Conference on Machine Learning (ICML)*. 2021, pp. 8162–8171.

[6] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat GANs on image synthesis.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021), pp. 8780–8794.

[7] Thomas Pollok, Lorenz Junglas, Boitumelo Ruf, and Arne Schumann. “UnrealGT: Using Unreal Engine to generate ground truth datasets.” In: *The International Symposium on Visual Computing (ISVC)*. 2019, pp. 670–682.

[8] You-Cyuan Jhang, Adam Palmar, Bowen Li, Saurav Dhakad, Sanjay Kumar Vishwakarma, Jonathan Hogins, Adam Crespi, Chris Kerr, Sharmila Chockalingam, Cesar Romero, Alex Thaman, and Sujoy Ganguly. *Training a performant object detection ML model on synthetic data using Unity Perception tools*. 2020.

[9] Aqsa Sabir, Rahat Hussain, Akeem Pedro, Mehrtash Soltani, Dongmin Lee, Chansik Park, and Jae-Ho Pyeon. “Synthetic data generation with Unity 3D and Unreal Engine for construction hazard scenarios: A comparative analysis.” In: *The International Conference on Construction Engineering and Project Management (ICCEPM)*. 2024, pp. 1286–1288.

[10] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. “Learning from synthetic humans.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4627–4635.

[11] Magnus Wrenninge and Jonas Unger. “Synscapes: A photorealistic synthetic dataset for street scene parsing.” In: *arXiv preprint arXiv:1810.08705* (2018).

[12] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. “SynBody: Synthetic dataset with layered human models for 3D human perception and modeling.” In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 20282–20292.

[13] MarketsandMarkets. *Augmented and virtual reality market size, share and growth*. 2024.

[14] Amy Webb, Melanie Subin, Victoria Chaitof, Nick Bartlett, Sam Jordan, Mark Bryan, Sam Guzik, Marc Palatucci, Andrew McDermott, Andrew Hornstra, Emily Caufield, Candice Rhea, Erica Peterson, and Sarah Johnson. *2025 Tech trends report*. 2025.

[15] Mirrorsize. *How AI is slashing fashion ecommerce returns by 60%: The secret to happier shoppers and higher profits*. 2024.

[16] Christina Sol. *Benefits of virtual fitting rooms for ecommerce CX*. 2024.

[17] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. “Deepfakes and beyond: A survey of face manipulation and fake detection.” In: *Information Fusion* 64 (2020), pp. 131–148.

[18] Yisroel Mirsky and Wenke Lee. “The creation and detection of deepfakes: A survey.” In: *ACM Computing Surveys (CSUR)* 54.1 (2021), pp. 1–41.

[19] Danielle K Citron and Robert Chesney. “Deep fakes: A looming challenge for privacy, democracy, and national security.” In: *California Law Review (CLR)* 107.6 (2019), pp. 1753–1820.

[20] Teresa Weikmann, Hannah Greber, and Alina Nikolaou. “After deception: how falling for a deepfake affects the way we see, hear, and experience media.” In: *The International Journal of Press/Politics (IJPP)* 30.1 (2025), pp. 187–210.

[21] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. “Deepfake detection based on discrepancies between faces and their context.” In: *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 44.10 (2021), pp. 6111–6121.

[22] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. “SIDA: Social media image deepfake detection, localization and explanation with large multimodal model.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025, pp. 28831–28841.

[23] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. “Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples.” In: *The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2021, pp. 3348–3357.

[24] Chi Liu, Huajie Chen, Tianqing Zhu, Jun Zhang, and Wanlei Zhou. “Making deepfakes more spurious: Evading deep face forgery detection via trace removal attack.” In: *The IEEE Transactions on Dependable and Secure Computing (TDSC)* 20.6 (2023), pp. 5182–5196.

[25] **Prasun Roy**, Saumik Bhattacharya, Subhankar Ghosh, and Umapada Pal. “Multi-scale Attention Guided Pose Transfer.” In: *Pattern Recognition (**PR**)* 137 (2023), p. 109315.

[26] **Prasun Roy**, Subhankar Ghosh, Saumik Bhattacharya, Umapada Pal, and Michael Blumenstein. “TIPS: Text-Induced Pose Synthesis.” In: *The European Conference on Computer Vision (**ECCV**)*. 2022, pp. 161–178.

[27] **Prasun Roy**, Subhankar Ghosh, Saumik Bhattacharya, Umapada Pal, and Michael Blumenstein. “Scene Aware Person Image Generation through Global Contextual

Conditioning." In: *The International Conference on Pattern Recognition (**ICPR**)*. 2022, pp. 2764–2770.

[28] **Prasun Roy**, Saumik Bhattacharya, Subhankar Ghosh, Umapada Pal, and Michael Blumenstein. "Semantically Consistent Person Image Generation." In: *The International Conference on Pattern Recognition (**ICPR**)*. 2024, pp. 293–309.

[29] **Prasun Roy**, Saumik Bhattacharya, Subhankar Ghosh, Umapada Pal, and Michael Blumenstein. "Exploring Mutual Cross-Modal Attention for Context-Aware Human Affordance Generation." In: *The IEEE Transactions on Artificial Intelligence (**TAI**) (**accepted**)* (2025).

[30] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. "Image-to-Image translation with conditional adversarial networks." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1125–1134.

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2017, pp. 2223–2232.

[32] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. "High-resolution image synthesis and semantic manipulation with conditional GANs." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8798–8807.

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. "Semantic image synthesis with spatially-adaptive normalization." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2337–2346.

[34] Wengling Chen and James Hays. "SketchyGAN: Towards diverse and realistic sketch to image synthesis." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 9416–9425.

[35] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. "Image generation from sketch constraint using contextual GAN." In: *The European Conference on Computer Vision (ECCV)*. 2018, pp. 205–220.

[36] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. "Interactive sketch & fill: Multiclass sketch-to-image

translation." In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 1171–1180.

[37] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. "SketchyCOCO: Image generation from freehand scene sketches." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5174–5183.

[38] Justin Johnson, Agrim Gupta, and Li Fei-Fei. "Image generation from scene graphs." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 1219–1228.

[39] Oron Ashual and Lior Wolf. "Specifying object attributes and relations in interactive scene generation." In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 4561–4569.

[40] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. "Image generation from layout." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 8584–8593.

[41] Jiayun Wang, Sangryul Jeon, Stella X Yu, Xi Zhang, Himanshu Arora, and Yu Lou. "Unsupervised scene sketch to photo synthesis." In: *The European Conference on Computer Vision (ECCV)*. 2022, pp. 273–289.

[42] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. "Invertible conditional GANs for image editing." In: *The Conference on Neural Information Processing Systems (NeurIPS) Workshops*. 2016.

[43] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. "Generative visual manipulation on the natural image manifold." In: *The European Conference on Computer Vision (ECCV)*. 2016, pp. 597–613.

[44] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. "Inverting layers of a large generator." In: *The International Conference on Learning Representations (ICLR) Workshops*. 2019.

[45] Zirui An, Jingbo Yu, Runtao Liu, Chuang Wang, and Qian Yu. "SketchInverter: Multi-class sketch-based image generation via GAN inversion." In: *The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 4319–4329.

[46] Zachary C Lipton and Subarna Tripathi. “Precise recovery of latent vectors from generative adversarial networks.” In: *The International Conference on Learning Representations (ICLR) Workshops*. 2017.

[47] Antonia Creswell and Anil Anthony Bharath. “Inverting the generator of a generative adversarial network.” In: *The IEEE Transactions on Neural Networks and Learning Systems (TNNLS)* 30.7 (2018), pp. 1967–1974.

[48] Fangchang Ma, Ulas Ayaz, and Sertac Karaman. “Invertibility of convolutional generative networks from partial measurements.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 31 (2018).

[49] Aditya Ramesh, Youngduck Choi, and Yann LeCun. “A spectral regularizer for unsupervised disentanglement.” In: *arXiv preprint arXiv:1812.01161* (2018).

[50] Rameen Abdal, Yipeng Qin, and Peter Wonka. “Image2StyleGAN: How to embed images into the StyleGAN latent space?” In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 4432–4441.

[51] Rameen Abdal, Yipeng Qin, and Peter Wonka. “Image2StyleGAN++: How to edit the embedded images?” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 8296–8305.

[52] Andrey Voynov and Artem Babenko. “Unsupervised discovery of interpretable directions in the GAN latent space.” In: *The International Conference on Machine Learning (ICML)*. 2020, pp. 9786–9796.

[53] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. “Seeing what a GAN cannot generate.” In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 4502–4511.

[54] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. “In-domain GAN inversion for real image editing.” In: *The European Conference on Computer Vision (ECCV)*. 2020, pp. 592–608.

[55] Andrew Brock, Jeff Donahue, and Karen Simonyan. “Large scale GAN training for high fidelity natural image synthesis.” In: *The International Conference on Learning Representations (ICLR)*. 2019.

[56] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “ImageNet: A large-scale hierarchical image database.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 248–255.

[57] Xiaoyu Xiang, Ding Liu, Xiao Yang, Yiheng Zhu, Xiaohui Shen, and Jan P Allebach. “Adversarial open domain adaptation for sketch-to-photo synthesis.” In: *The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022, pp. 1434–1444.

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. “Denoising diffusion implicit models.” In: *The International Conference on Learning Representations (ICLR)*. 2021.

[59] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. “Zero-shot text-to-image generation.” In: *The International Conference on Machine Learning (ICML)*. 2021, pp. 8821–8831.

[60] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. “Hierarchical text-conditional image generation with CLIP latents.” In: *arXiv preprint arXiv:2204.06125* (2022).

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. “High-resolution image synthesis with latent diffusion models.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10684–10695.

[62] 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.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), pp. 36479–36494.

[63] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. “Text2LIVE: Text-driven layered image and video editing.” In: *The European Conference on Computer Vision (ECCV)*. 2022, pp. 707–723.

[64] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. “Prompt-to-Prompt image editing with cross-attention control.” In: *The International Conference on Learning Representations (ICLR)*. 2023.

[65] Tim Brooks, Aleksander Holynski, and Alexei A Efros. “InstructPix2Pix: Learning to follow image editing instructions.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 18392–18402.

[66] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. “Null-text inversion for editing real images using guided diffusion models.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 6038–6047.

[67] 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: *The International Conference on Machine Learning (ICML)*. 2021, pp. 8748–8763.

[68] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. “ILVR: Conditioning method for denoising diffusion probabilistic models.” In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 14347–14356.

[69] 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: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 22500–22510.

[70] 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.” In: *The International Conference on Learning Representations (ICLR)*. 2021.

[71] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. “EGSDE: Unpaired image-to-image translation via energy-guided stochastic differential equations.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), pp. 3609–3623.

[72] Ximing Xing, Chuang Wang, Haitao Zhou, Zhihao Hu, Chongxuan Li, Dong Xu, and Qian Yu. “Inversion-by-Inversion: Exemplar-based sketch-to-photo synthesis via stochastic differential equations.” In: *arXiv preprint arXiv:2308.07665* (2023).

[73] Qiang Wang, Di Kong, Fengyin Lin, and Yonggang Qi. "DiffSketching: Sketch control image synthesis with diffusion models." In: *The British Machine Vision Conference (BMVC)*. 2022.

[74] Shin-I Cheng, Yu-Jie Chen, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. "Adaptively-realistic image generation from stroke and sketch with diffusion model." In: *The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 4054–4062.

[75] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. "Sketch-guided text-to-image diffusion models." In: *The ACM SIGGRAPH Conference Proceedings*. 2023, pp. 1–11.

[76] **Prasun Roy**, Saumik Bhattacharya, Subhankar Ghosh, Umapada Pal, and Michael Blumenstein. "d-Sketch: Improving Visual Fidelity of Sketch-to-Image Translation with Pretrained Latent Diffusion Models without Retraining." In: *The International Conference on Pattern Recognition (**ICPR**)*. 2024, pp. 277–292.

[77] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. "Pose guided person image generation." In: *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

[78] 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 (MICCAI)*. 2015, pp. 234–241.

[79] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. "Disentangled person image generation." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 99–108.

[80] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. "Deformable GANs for pose-based human image generation." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 3408–3416.

[81] Patrick Esser, Ekaterina Sutter, and Björn Ommer. "A variational U-Net for conditional appearance and shape generation." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8857–8866.

[82] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes.” In: *The International Conference on Learning Representations (ICLR)*. 2014.

[83] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. “Progressive pose attention transfer for person image generation.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2347–2356.

[84] Badour AlBahar and Jia-Bin Huang. “Guided image-to-image translation with bi-directional feature transformation.” In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 9016–9025.

[85] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. “Cross-domain correspondence learning for exemplar-based image translation.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5143–5153.

[86] Siyu Huang, Haoyi Xiong, Zhi-Qi Cheng, Qingzhong Wang, Xingran Zhou, Bihan Wen, Jun Huan, and Dejing Dou. “Generating person images with appearance-aware pose stylizer.” In: *International Joint Conference on Artificial Intelligence (IJCAI)*. 2020, pp. 623–629.

[87] Kun Li, Jinsong Zhang, Yebin Liu, Yu-Kun Lai, and Qionghai Dai. “PoNA: Pose-guided non-local attention for human pose transfer.” In: *The IEEE Transactions on Image Processing (TIP)* 29 (2020), pp. 9584–9599.

[88] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. “XingGAN for person image generation.” In: *The European Conference on Computer Vision (ECCV)*. 2020, pp. 717–734.

[89] Hao Tang, Song Bai, Philip HS Torr, and Nicu Sebe. “Bipartite graph reasoning GANs for person image generation.” In: *The British Machine Vision Conference (BMVC)*. 2020.

[90] Wing-Yin Yu, Lai-Man Po, Yuzhi Zhao, Jingjing Xiong, and Kin-Wai Lau. “Spatial content alignment for pose transfer.” In: *The IEEE International Conference on Multimedia and Expo (ICME)*. 2021, pp. 1–6.

[91] Tianjiao Li, Wei Zhang, Ran Song, Zhiheng Li, Jun Liu, Xiaolei Li, and Shijian Lu. "PoT-GAN: Pose transform GAN for person image synthesis." In: *The IEEE Transactions on Image Processing (TIP)* 30 (2021), pp. 7677–7688.

[92] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. "Neural texture extraction and distribution for controllable person image synthesis." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 13535–13544.

[93] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. "Exploring dual-task correlation for pose guided person image generation." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 7713–7722.

[94] Songhua Liu, Jingwen Ye, Sucheng Ren, and Xinchao Wang. "Dynast: Dynamic sparse transformer for exemplar-guided image generation." In: *The European Conference on Computer Vision (ECCV)*. 2022, pp. 72–90.

[95] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. "Controllable person image synthesis with attribute-decomposed GAN." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5084–5093.

[96] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 932–940.

[97] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. "Neural re-rendering of humans from a single image." In: *The European Conference on Computer Vision (ECCV)*. 2020, pp. 596–613.

[98] Jinsong Zhang, Xingzi Liu, and Kun Li. "Human pose transfer by adaptive hierarchical deformation." In: *Computer Graphics Forum* 39.7 (2020), pp. 325–337.

[99] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. "PISE: Person image synthesis and editing with decoupled GAN." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 7982–7990.

[100] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. “Learning semantic person image generation by region-adaptive normalization.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 10806–10815.

[101] Badour Albahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. “Pose with style: Detail-preserving pose-guided image synthesis with conditional StyleGAN.” In: *The ACM Transactions on Graphics (TOG)* 40.6 (2021), pp. 1–11.

[102] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. “Analyzing and improving the image quality of StyleGAN.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 8110–8119.

[103] Ting Liu, Jianfeng Zhang, Xuecheng Nie, Yunchao Wei, Shikui Wei, Yao Zhao, and Jiashi Feng. “Spatial-aware texture transformer for high-fidelity garment transfer.” In: *The IEEE Transactions on Image Processing (TIP)* 30 (2021), pp. 7499–7510.

[104] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. “DensePose: Dense human pose estimation in the wild.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7297–7306.

[105] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. “Style and pose control for image synthesis of humans from a single monocular view.” In: *arXiv preprint arXiv:2102.11263* (2021).

[106] Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. “HumanGAN: A generative model of human images.” In: *The International Conference on 3D Vision (3DV)*. 2021, pp. 258–267.

[107] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. “Cross attention based style distribution for controllable person image synthesis.” In: *The European Conference on Computer Vision (ECCV)*. 2022, pp. 161–178.

[108] Nannan Li, Kevin J Shih, and Bryan A Plummer. “Collecting the puzzle pieces: Disentangled self-driven human pose transfer by permuting textures.” In: *The*

IEEE/CVF International Conference on Computer Vision (ICCV). 2023, pp. 7126–7137.

[109] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. "A generative model of people in clothing." In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2017, pp. 853–862.

[110] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. "Human appearance transfer." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 5391–5399.

[111] 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: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 5904–5913.

[112] Wen Liu, Zhixin Piao, Zhi Tu, Wenhan Luo, Lin Ma, and Shenghua Gao. "Liquid warping GAN with attention: A unified framework for human image synthesis." In: *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 44.9 (2021), pp. 5114–5132.

[113] Markus Knoche, István Sárándi, and Bastian Leibe. "Reposing humans by warping 3D features." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2020, pp. 1044–1045.

[114] Jinxiang Liu, Yangheng Zhao, Siheng Chen, and Ya Zhang. "A 3D mesh-based lifting-and-projection network for human pose transfer." In: *The IEEE Transactions on Multimedia (TMM)* 24 (2021), pp. 4314–4327.

[115] Yining Li, Chen Huang, and Chen Change Loy. "Dense intrinsic appearance flow for human pose transfer." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 3693–3702.

[116] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. "Deep image spatial transformation for person image generation." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 7690–7699.

[117] Liyuan Ma, Kejie Huang, Dongxu Wei, Zhaoyan Ming, and Haibin Shen. "FDA-GAN: Flow-based dual attention GAN for human pose transfer." In: *The IEEE Transactions on Multimedia (TMM)* 25 (2021), pp. 930–941.

[118] Rishabh Jain, Krishna Kumar Singh, Mayur Hemani, Jingwan Lu, Mausoom Sarkar, Duygu Ceylan, and Balaji Krishnamurthy. "VGFlow: Visibility guided flow network for human reposing." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 21088–21097.

[119] Liyuan Ma, Tingwei Gao, Haitian Jiang, Haibin Shen, and Kejie Huang. "WaveIPT: Joint attention and flow alignment in the wavelet domain for pose transfer." In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 7215–7225.

[120] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. "Unsupervised person image synthesis in arbitrary poses." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8620–8628.

[121] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. "Unsupervised person image generation with semantic parsing transformation." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2357–2366.

[122] Haitian Zheng, Lele Chen, Chenliang Xu, and Jiebo Luo. "Unsupervised pose flow learning for pose guided synthesis." In: *arXiv preprint arXiv:1909.13819* (2019).

[123] Soubhik Sanyal, Alex Vorobiov, Timo Bolkart, Matthew Loper, Betty Mohler, Larry S Davis, Javier Romero, and Michael J Black. "Learning realistic human reposing using cyclic self-supervision with 3D shape, pose, and appearance consistency." In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 11138–11147.

[124] Zijian Wang, Xingqun Qi, Kun Yuan, and Muyi Sun. "Self-supervised correlation mining network for person image generation." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 7703–7712.

[125] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et

al. “Language models are few-shot learners.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), pp. 1877–1901.

[126] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. “LLaMA: Open and efficient foundation language models.” In: *arXiv preprint arXiv:2302.13971* (2023).

[127] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. “Generative adversarial text to image synthesis.” In: *The International Conference on Machine Learning (ICML)*. 2016, pp. 1060–1069.

[128] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. “Video generation from text.” In: *The AAAI Conference on Artificial Intelligence*. 2018, pp. 7065–7072.

[129] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. “MirrorGAN: Learning text-to-image generation by redescription.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1505–1514.

[130] Xingran Zhou, Siyu Huang, Bin Li, Yingming Li, Jiachen Li, and Zhongfei Zhang. “Text guided person image synthesis.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 3663–3672.

[131] Yifei Zhang, Rania Briq, Julian Tanke, and Juergen Gall. “Adversarial synthesis of human pose from text.” In: *The DAGM German Conference on Pattern Recognition (GCPR)*. 2020, pp. 145–158.

[132] Rania Briq, Pratika Kochar, and Juergen Gall. “Towards better adversarial synthesis of human images from text.” In: *arXiv preprint arXiv:2107.01869* (2021).

[133] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. “SMPL: A skinned multi-person linear model.” In: *The ACM Transactions on Graphics (TOG)* 34.6 (2015), pp. 1–16.

[134] Hengshuang Zhao, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Brian Price, and Jiaya Jia. “Compositing-aware image search.” In: *The European Conference on Computer Vision (ECCV)*. 2018, pp. 502–516.

[135] Fuwen Tan, Crispin Bernier, Benjamin Cohen, Vicente Ordonez, and Connelly Barnes. “Where and who? automatic semantic-aware person composition.” In: *The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2018, pp. 1519–1528.

[136] Donghoon Lee, Sifei Liu, Jinwei Gu, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. “Context-aware synthesis and placement of object instances.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 31 (2018).

[137] Oran Gafni and Lior Wolf. “Wish you were here: Context-aware human generation.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 7840–7849.

[138] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A. Efros, and Krishna Kumar Singh. “Putting people in their place: Affordance-aware human insertion into scenes.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 17089–17099.

[139] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

[140] Stephen J Anderson, Noriko Yamagishi, and Vivian Karavia. “Attentional processes link perception and action.” In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269.1497 (2002), pp. 1225–1232.

[141] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. “Scene-aware generative network for human motion synthesis.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 12206–12215.

[142] Anirban Roy and Sinisa Todorovic. “A multi-scale cnn for affordance segmentation in rgb images.” In: *The European Conference on Computer Vision (ECCV)*. 2016, pp. 186–201.

[143] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. “Binge watching: Scaling affordance learning from sitcoms.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2596–2605.

[144] Lingzhi Zhang, Weiyu Du, Shenghao Zhou, Jiancong Wang, and Jianbo Shi. “Inpaint2Learn: A self-supervised framework for affordance learning.” In: *The*

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2022, pp. 2665–2674.

[145] Jieteng Yao, Junjie Chen, Li Niu, and Bin Sheng. "Scene-aware human pose generation using transformer." In: *The ACM International Conference on Multimedia (MM)*. 2023, pp. 2847–2855.

[146] Yuke Zhu, Alireza Fathi, and Fei-Fei Li. "Reasoning about object affordances in a knowledge base representation." In: *The European Conference on Computer Vision (ECCV)*. 2014, pp. 408–424.

[147] Thanh-Toan Do, Anh Nguyen, and Ian Reid. "AffordanceNet: An end-to-end deep learning approach for object affordance detection." In: *The IEEE International Conference on Robotics and Automation (ICRA)*. 2018, pp. 5882–5889.

[148] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. "Learning to act properly: Predicting and explaining affordances from images." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 975–983.

[149] Emad Barsoum, John Kender, and Zicheng Liu. "HP-GAN: Probabilistic 3D human motion prediction via GAN." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2018, pp. 1418–1427.

[150] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. "Deep video generation, prediction and completion of human action sequences." In: *The European Conference on Computer Vision (ECCV)*. 2018, pp. 366–382.

[151] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. "Pose guided human video generation." In: *The European Conference on Computer Vision (ECCV)*. 2018, pp. 201–216.

[152] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. "Convolutional sequence generation for skeleton-based action synthesis." In: *The IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 4394–4402.

[153] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. "Action2Motion: Conditioned generation of 3D human motions." In: *The ACM International Conference on Multimedia (MM)*. 2020, pp. 2021–2029.

[154] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. “Realtime multi-person 2D pose estimation using part affinity fields.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 7291–7299.

[155] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[156] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating deep network training by reducing internal covariate shift.” In: *The International Conference on Machine Learning (ICML)*. 2015, pp. 448–456.

[157] Vinod Nair and Geoffrey E Hinton. “Rectified linear units improve Restricted Boltzmann Machines.” In: *The International Conference on Machine Learning (ICML)*. 2010, pp. 807–814.

[158] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. “Perceptual losses for real-time style transfer and super-resolution.” In: *The European Conference on Computer Vision (ECCV)*. 2016, pp. 694–711.

[159] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition.” In: *The International Conference on Learning Representations (ICLR)*. 2015.

[160] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization.” In: *The International Conference on Learning Representations (ICLR)*. 2015.

[161] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. “DeepFashion: powering robust clothes recognition and retrieval with rich annotations.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1096–1104.

[162] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. “Image quality assessment: From error visibility to structural similarity.” In: *The IEEE Transactions on Image Processing (TIP)* 13.4 (2004), pp. 600–612.

[163] Tim Salimans, Ian J Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. “Improved techniques for training GANs.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 29 (2016).

[164] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. "SSD: Single shot multibox detector." In: *The European Conference on Computer Vision (ECCV)*. 2016, pp. 21–37.

[165] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. "2D Human pose estimation: New benchmark and state of the art analysis." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 3686–3693.

[166] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9.

[167] Solomon Kullback and Richard A Leibler. "On information and sufficiency." In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86.

[168] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. "The unreasonable effectiveness of deep features as a perceptual metric." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 586–595.

[169] 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 <0.5MB model size." In: *arXiv preprint arXiv:1602.07360* (2016).

[170] **Prasun Roy**, Saumik Bhattacharya, Subhankar Ghosh, and Umapada Pal. "STEFANN: Scene text editor using font adaptive neural network." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**)*. 2020, pp. 13228–13237.

[171] TY Zhang and Ching Y. Suen. "A fast parallel algorithm for thinning digital patterns." In: *Communications of the ACM* 27.3 (1984), pp. 236–239.

[172] Alloy Das, Sanket Biswas, **Prasun Roy**, Subhankar Ghosh, Umapada Pal, Michael Blumenstein, Josep Lladós, and Saumik Bhattacharya. "FASTER: A font-agnostic scene text editing and rendering framework." In: *The IEEE/CVF Winter Conference on Applications of Computer Vision (**WACV**)*. 2025.

[173] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient estimation of word representations in vector space.” In: *arXiv preprint arXiv:1301.3781* (2013).

[174] Ben Athiwaratkun, Andrew Gordon Wilson, and Anima Anandkumar. “Probabilistic fasttext for multi-sense word embeddings.” In: *arXiv preprint arXiv:1806.02901* (2018).

[175] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of deep bidirectional transformers for language understanding.” In: *arXiv preprint arXiv:1810.04805* (2018).

[176] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. “Improved training of Wasserstein GANs.” In: *arXiv preprint arXiv:1704.00028* (2017).

[177] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. “Multiple-human parsing in the wild.” In: *arXiv preprint arXiv:1705.07206* (2017).

[178] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson W.H. Lau. “Harmonizer: Learning to perform white-box image and video harmonization.” In: *The European Conference on Computer Vision (ECCV)*. 2022, pp. 690–706.

[179] Jianqi Chen, Yilan Zhang, Zhengxia Zou, Keyan Chen, and Zhenwei Shi. “Dense pixel-to-pixel harmonization via continuous image representation.” In: *arXiv preprint arXiv:2303.01681* (2023).

[180] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. “Self-correction for human parsing.” In: *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 44.6 (2020), pp. 3260–3271.

[181] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. “Deep human parsing with active template regression.” In: *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 37.12 (2015), pp. 2402–2414.

[182] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

[183] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. “Non-local neural networks.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7794–7803.

[184] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. “OneFormer: One transformer to rule universal image segmentation.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 2989–2998.

[185] Ali Hassani and Humphrey Shi. “Dilated neighborhood attention transformer.” In: *arXiv preprint arXiv:2209.15001* (2022).

[186] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. “Semantic understanding of scenes through the ADE20K dataset.” In: *International Journal of Computer Vision (IJCV)* 127 (2019), pp. 302–321.

[187] Biao Zhang and Rico Sennrich. “Root mean square layer normalization.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019).

[188] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. “Deep high-resolution representation learning for human pose estimation.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5693–5703.

[189] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. “ViTPose: Simple vision transformer baselines for human pose estimation.” In: *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), pp. 38571–38584.

[190] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. “Human Pose as compositional tokens.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 660–671.

[191] Hae-Sang Park and Chi-Hyuck Jun. “A simple and fast algorithm for K-medoids clustering.” In: *Expert Systems with Applications* 36.2 (2009), pp. 3336–3341.

[192] Sijie Zhu, Zhe Lin, Scott Cohen, Jason Kuen, Zhifei Zhang, and Chen Chen. “TopNet: Transformer-based object placement network for image compositing.” In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 1838–1847.

[193] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. "Resolution-robust large mask inpainting with fourier convolutions." In: *The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022, pp. 2149–2159.

[194] Yi Yang and Deva Ramanan. "Articulated human detection with flexible mixtures of parts." In: *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35.12 (2012), pp. 2878–2890.

[195] Bruno Artacho and Andreas Savakis. "UniPose: Unified human pose estimation in single images and videos." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 7035–7044.

[196] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. "Pose recognition with cascade transformers." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 1944–1953.

[197] Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. "Learning object placement by inpainting for compositional data augmentation." In: *The European Conference on Computer Vision (ECCV)*. 2020, pp. 566–581.

[198] Siyuan Zhou, Liu Liu, Li Niu, and Liqing Zhang. "Learning object placement via dual-path graph completion." In: *The European Conference on Computer Vision (ECCV)*. 2022, pp. 373–389.

[199] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. "Depth anything: Unleashing the power of large-scale unlabeled data." In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 10371–10381.

[200] 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: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 5968–5976.