확산 모델을 활용한 텍스트 기반 패션 이미지 편집

권소예, 이재구*
국민대학교 컴퓨터공학과
*jaekoo@kookmin.ac.kr

요 약

본 논문은 확산 모델을 활용한 텍스트 기반 패션 이미지 편집 연구를 제안한다. 기존 연구들은 GAN 을 기반으로 하여 생성된 이미지 품질이 낮은 문제와 의류의 형태에 대한 제어가 불가능한 문제가 존재한다. 본 연구는 이러한 한계를 극복하기 위해 확산 모델 기반의 새로운 편집 방법을 제시한다. 첫 번째로 텍스트 프롬프트에 담긴 형태 정보를 인체 파싱 맵에 반영하고, 두 번째로 이 정보를 기반으로 기존 의류 영역과 새로운 의류 영역을 구분하는 정교한 마스킹 절차를 거쳐 확산 모델에 주입한다. 이 방법을 통해 사용자는 텍스트만으로 의류의 형태는 물론 색상, 질감 등의 내부 요소까지 자유롭게 편집할 수 있다. 정량적 및 정성적 실험 결과, 제안 방법은 기존 연구들 대비 우수한 이미지 품질과 텍스트-이미지 일치도를 달성함을 입증하였다.

1. 서론

패션 이미지 편집은 전자 상거래, 가상 체험 (Virtual Try-On), 디지털 콘텐츠 제작 등 다양한 산업 분야에서 잠재력을 지니고 있어 컴퓨터 비전 분야에서 활발하게 연구되고 있다[1, 2, 3]. 패션 이미지 편집 연구는 크게 두 가지 방향으로 발전해 왔다. 첫 번째는 텍스트 프롬프트와 참조 이미지를 결합하여 의류의 색상이나 질감을 편집하는 방식[4, 5]이며, 두 번째는 텍스트 프롬프트만을 활용하여 사용자 의도에 따라 이미지를 편집하는 방식[6, 7]이다. 본 연구는 사용자가 원하는 편집 방향을 텍스트 프롬프트로 명시하여 높은 접근성과 편의성을 제공하는 텍스트 프롬프트 기반 패션 이미지 편집에 초점을 맞춘다.

텍스트 프롬프트 기반 편집 분야에서 대표적인 선행 연구로는 FICE[6]와 TexFit[7]이 있다. FICE 는 GAN[8] 기반 접근 방식을 사용하여 텍스트 프롬프트에 기반한 의류 편집을 시도하였으나, GAN 의 본질적인 한계[9]로 인해 편집 결과 이미지의 품질이 낮거나 디테일한 표현에 제약이 있다. TexFit 은 비교적 고품질의 편집 결과를 보였지만, 의류의 질감이나 색상과 같은 세부 요소 편집에만 초점을 맞추어, 옷의 형태 정보 자체를 변경하는 데는 근본적인 한계가 존재한다. 결과적으로 기존 연구들은 고품질 이미지를 안정적으로 생성하지 못하고, 의류의 전체적인 형태나 내부 디테일을 동시에 제어하는 기능 역시 부족하다. 이러한 두 가지 제약 때문에 사용자가 의류를 원하는 방식으로 자유롭게 편집할 수 있는 편집 자유도를 제공하지 못하는 문제가 있다.

본 연구는 이러한 기존 연구의 한계를 극복하고자 확산 모델을 활용하여 고품질 이미지 생성이 가능하면서도, 의류의 형태 정보와 색상, 질감 등의 내부 요소까지 텍스트 프롬프트를 통해 자유롭게 편집할 수 있는 새로운 방법을 제안한다. 이를 위해, 우리는 PICTURE[5]에서 제안한 Parsing Editing Module(파싱 편집 모듈)을 활용하는 접근 방식을 채택한다. 구체적으로, 파싱 맵을 기반으로 텍스트 프롬프트에 명시된 새로운 형태 정보를 반영하여 의류 영역의 파싱 맵을 편집한다. 또한, 기존 의류 영역과 편집될 새로운 의류 영역을 구분하는 정교한 마스킹 절차를 제안하고, 이 마스크 정보를 통해 새로운 형태와 내부 요소를 갖춘 의류 이미지를 생성함으로써 패션 이미지 편집을 수행한다.

제안 방법은 기존 텍스트 기반 패션 이미지 편집의 한계를 해소하고 사용자에게 더욱 정교하고 높은 수준의 편집 기능을 제공한다. 정량적 및 정성적 실험 결과를 통해, 본 연구에서 제안한 방법이 기존 연구들에 비해 우수한 텍스트–이미지 일치도와 높은 사실감을 달성함을 확인하였다.

2. 관련 연구

2.1 패션 이미지 편집

패션 이미지 편집 연구는 크게 텍스트 및 참조

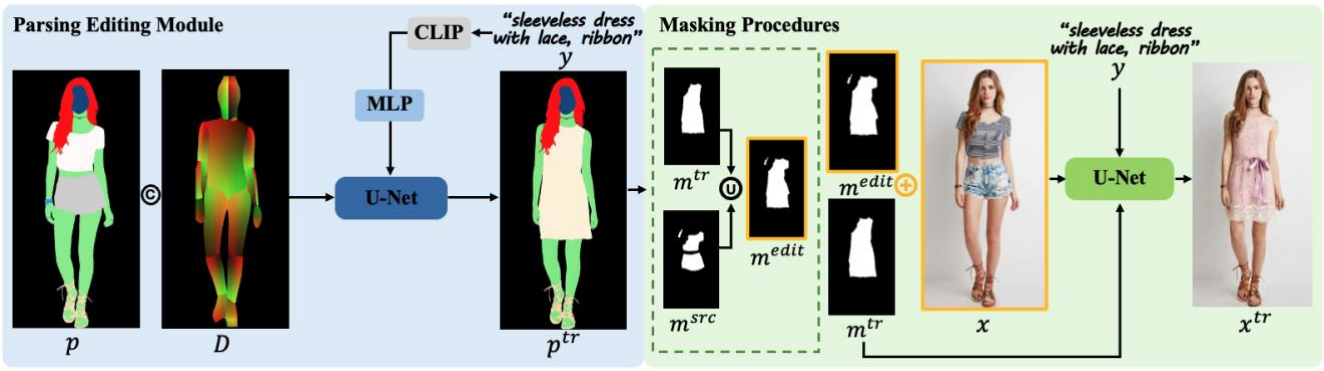

그림 1. 제안 방법의 구조

이미지 기반 편집과 텍스트 조건부 편집의 두 가지 주요 접근 방식으로 발전해 왔다[4, 5, 6, 7]. 텍스트 및 참조 이미지 기반 편집 방식은 PICTURE[5]나 FashionTex[4]와 같이 참조 이미지의 질감을 입력 이미지에 전이시키고 텍스트 프롬프트를 이용해 구조적 제어를 수행하여 편집 결과를 얻을 수 있다. 그러나 이 방식은 편집 과정에서 사용자가 참조 이미지를 반드시 준비해야 한다는 점에서 사용자의 아이디어를 즉각적으로 구현하는 데 제약이 따른다. 반면, 텍스트 조건부 편집 방식은 자연어 명령만으로 직관적인 편집이 가능하다는 높은 접근성의 장점을 가진다. FICE 와 같은 초기 GAN 기반 연구는 텍스트를 이용해 색상이나 패턴을 변경하려 했으나 이미지 품질의 한계가 명확했으며, TexFit 과 같은 후속 연구는 텍스처 편집에 집중하여 품질을 높였지만, 의류의 형태를 변경하는 편집 기능은 지원하지 않는다. 따라서 본 연구는 텍스트 조건부 편집의 직관적인 장점을 유지하면서, 기존 연구들의 품질 문제와 편집 범위의 한계를 해결하고자 한다.

2.2 확산 모델 기반 이미지 편집

확산 모델은 이미지 합성 분야에서 GAN 을 능가하는 고품질의 이미지 생성 능력과 다양성을 입증하며 새로운 표준으로 자리 잡았다. DALL-E 2[10], LDM[11] 등의 대규모 모델들은 CLIP[12]과 결합하여 텍스트 프롬프트에 기반한 범용적인 이미지 생성 및 편집을 가능하게 했다. 또한 ControlNet[13]과 같은 연구는 에지 맵, 깊이 맵 등의 구조적 정보를 확산 모델에 조건으로 주입하여 일반 이미지 편집의 제어 가능성을 향상시켰다. 그러나 이러한 확산 모델 기반의 이미지 편집 연구들은 대부분 일반적인 도메인의 이미지 편집에 초점을 맞추고 있어, 사람의 포즈, 복잡한 의류의 주름 및 쉐이딩, 인체 파싱 영역 등 패션 도메인에 특화된 정교한 제어 메커니즘을 제공하는 데는 한계가 있다. 특히, 옷의 형태를 미세하게 변경하고 그 변경된 영역에 새로운 속성을 자연스럽게 합성하는 복합적인 패션 편집 요구사항을 충족시키기 어렵다. 따라서 본 연구는 패션 도메인에 특화된 편집 방법을 제공하는 것을 목표로 한다.

3. 본론

[그림 1]은 본 연구에서 제안하는 텍스트 기반 패션 이미지 편집 방법론의 전체 개요를 보여준다. 본 방법론은 입력으로 사람 이미지 x 와 사용자의 편집 의도를 담은 텍스트 프롬프트 y 가 주어졌을 때, 텍스트 프롬프트 y에 따라 의류 영역이 편집된 최종 결과 이미지 x^{tr}을 생성하는 것을 목표로 한다.

3.1 파싱 편집 모듈

본 섹션은 텍스트 프롬프트의 형태 정보를 인체 파싱 맵에 반영하는 방법, 즉 파싱 편집 모듈의 작동 원리를 설명한다. 이 방법의 목표는 입력 이미지의 의류 영역을 사용자가 원하는 구조적 형태로 변화시킨 편집된 파싱 맵 P^{tr}을 생성하는 것이다. 주어진 입력으로는 인체 파싱 이미지 p와 포즈 이미지 D, 그리고 사용자의 편집 의도를 나타내는 텍스트 프롬프트 y가 사용된다. 이 모듈은 LDM[11]의 구조를 기반으로 한다. 텍스트 프롬프트 y는 CLIP 인코더를 거쳐 특징 벡터로 추출된다. 추출된 텍스트 특징을 인체 파싱 이미지 공간에 맞게 매핑하기 위해 MLP layer 를 거친 후, U-Net 의 중간 계층에 Cross-Attention 연산을 통해 조건 정보로 입력된다. 인체 파싱 이미지 p는 입력 사람 이미지의 인체 정보를 제공하며, U-Net 의 입력 레이어에 직접 주입된다. 이로써 U-Net 은 입력 사람 이미지의 인체 정보를 인지하면서도 텍스트 조건에 맞는 새로운 형태로 의류 영역을 편집하는 것이 가능해진다. 이러한 조건부 확산 과정을 통해 모듈은 텍스트에 명시된 형태 정보를 반영한 P^{tr}을 효과적으로 생성한다.

3.2 마스킹 절차

이 섹션에서는 이 P^{tr}을 활용하여 원본 이미지의 의류 영역과 새로운 의류 영역을 구분하는 마스킹 과정과 이 마스크를 활용한 편집 방법을 설명한다. 패션 이미지 편집을 정교하게 수행하기 위해, 우리는 세 가지 핵심 마스크 정보를 정의하고 추출한다. 첫 번째, 소스 마스크 m^{src}는 인체 파싱 이미지 p에

서 추출된 기존 의류 영역을 나타낸다. 두 번째, 타겟 마스크 m^{tr}는 편집된 파싱 맵 P^{tr}에서 추출된 새로운 의류 형태 영역을 나타낸다. 세 번째, 편집 영역 마스크 m^{edit}는 실제 편집이 필요한 영역을 정의하며, m^{edit} 추출 과정은 아래 수식과 같다.

$$m^{edit} = m^{src} \cup m^{tr} \quad (1)$$

m^{edit} 은 원본 의류와 새로운 의류 형태 전체 영역을 포괄한다. 이후 m^{edit} 은 StableDiffusion-Inpaint[11]의 마스크 이미지로 활용되어 편집이 필요한 영역을 지정한다. 동시에 m^{tr} 은 ControlNet 의 조건 정보로 활용되어, 새로운 의류 형태(m^{tr})의 경계를 확산 모델에 더욱 정교하게 전달한다. 텍스트 프롬프트 y는 Cross-Attention 연산을 통해 U-Net 에 조건 정보로 주입된다. 이 조건부 과정을 통해 모델은 새로운 의류 형태와 속성을 생성하고, m^{edit} 의 외부 영역(배경, 인체 등)에서는 원본 이미지의 정보를 유지하는 패션 편집 결과 이미지 x^{tr} 을 최종적으로 생성한다.

4. 실험

제안 방법의 성능을 평가하기 위해 기존 방법과 성능을 비교하는 실험을 진행한다. 기존 방법은 SDXL-Inpaint[11]와 TexFit[7]으로 각 텍스트를 조건으로 하는 일반 이미지 편집 모델과 패션 이미지 편집 모델이다. 공정한 평가를 위해 50 스텝의 DDIM 샘플러를 사용하고, 가이드 스케일을 7.5 로 설정하여 추론을 진행하였다. 실험에 사용된 사람 이미지 데이터는 DeepFasion-MultiModal[15]로부터 랜덤하게 300 개를 선택하였다. 텍스트 프롬프트 데이터는 GPT-4 를 통해 상의, 하의, 드레스를 나타내는 텍스트 프롬프트 각 100 개씩 총 300 개를 생성하였다.

제안 방법의 우수성을 입증하기 위해, 편집된 이미지와 텍스트 프롬프트와의 일치도를 CLIP-Text 점수[11]를 통해 유사성을 평가하였다. 추가로 GPT-4V 를 활용하여 프롬프트 일치도와 생성 이미지의 자연스러움을 평가하였다. 각 입력에 대해 세 방법의 생성 결과를 GPT-4V 에 제시하고, 두 평가 기준을 종합하여 가장 우수한 결과를 하나만 선택하도록 하였다. 이를 통해 세 모델의 선택 비율이 100%가 되도록 구성하여, 각 방법의 전반적인 우수성을 정량적으로 비교하였다.

[표 1]을 보면 알 수 있듯이, 제안 방법이 CLIP-Text 점수와 GPT-4V 기반의 선호도 모두에서 기존 방법들보다 월등히 높은 수치를 기록하여, 편집된 이미지와 텍스트 프롬프트 간의 일치도 및 이미지의 전체적인 자연스러움 측면에서 제안 방법의 우수성을 정량적으로 입증하였다. [그림 2]는 제안 방법과 기존 방법의 정성적 성능 비교 결과를 보여준다. 제안 방법은 텍스트 프롬프트의 형태 정보뿐만 아니라 의류의 내부 속성까지 충실히 반영된 자연스러운 결과 이미지를 생성하는 것을 확인할 수 있다. 반면에 TexFit 의 결과는 의류의 내부 속성 요소

표 1. 모델 간의 정량적 성능 평가

	CLIP-Text(↑)	GPT-4V(↑)
SDXL-Inpaint	28.92	6.63%
TexFit	28.71	16.92%
Ours	**30.19**	**76.45%**

는 반영되지만, 형태 정보는 아예 변하지 않아 프롬프트와의 일치도가 떨어진다. SDXL-Inpaint 결과 역시 마지막 행에서 볼 수 있듯이 "long sleeve"와 같은 형태 정보가 반영되지 못했을 뿐만 아니라, 전반적으로 부자연스러운 품질을 보인다.

5. 결론

본 연구는 기존 텍스트 기반 패션 이미지 편집 방법론의 한계였던 이미지 품질 저하 및 의류 형태 제어 부족 문제를 해결하고자 확산 모델을 활용한 새로운 편집 방법론을 제안하였다. 제안된 방법은 텍스트 프롬프트의 형태 정보를 Parsing Editing Module 을 통해 편집된 인체 파싱 맵으로 변환하고, 이를 기반으로 정의된 편집 영역 마스크 및 타겟 마스크를 SDXL-Inpaint 파이프라인과 ControlNet 의 다중 조건으로 주입하여 편집을 수행한다. 이 독창적인 접근 방식을 통해 우리는 사용자가 텍스트만으로 의류의 형태적 변화 내부 속성을 동시에, 그리고 정교하게 제어하는 것이 가능함을 입증하였다. 정량적 및 정성적 실험 결과는 본 연구의 방법론이 기존의 선행 연구들 대비 우수한 이미지 품질과 텍스트 의도와의 높은 일치도를 달성함을 보여준다.

감사의 글

이 논문은 과학기술정보통신부 및 정보통신기획평가원의 생성 AI 선도인재양성사업(IITP-2024-RS-2024-00397085)과 2024 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.RS-2024-00357879, 지능형 개인 맞춤 만성 질환 관리를 위한 AI 기반 생체데이터 융합 및 생성 기술)을 받아 수행된 연구 결과임

참고문헌

[1] Han, Xintong, et al. "Viton: An image-based virtual try-on network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[2] Baldrati, Alberto, et al. "Multimodal garment designer: Human-centric latent diffusion models for fashion image editing." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

[3] Wu, Menghua, et al. "High-fidelity 3d face generation from natural language descriptions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[4] Lin, Anran, et al. "Fashiontex: Controllable virtual try-on with text and texture." ACM SIGGRAPH 2023 conference proceedings. 2023.

[5] Ning, Shuliang, et al. "Picture: Photorealistic virtual try-on from unconstrained designs." Proceedings of the

그림 2. 방법 간의 정성적 성능 평가

IEEE/CVF conference on computer vision and pattern recognition. 2024.

[6] Pernuš, Martin, et al. "FICE: Text-conditioned fashion-image editing with guided GAN inversion." Pattern Recognition 158 (2025): 111022.

[7] Wang, Tongxin, and Mang Ye. "Texfit: Text-driven

fashion image editing with diffusion models." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 9. 2024.
[8] Goodfellow, Ian, et al. "Generative adversarial networks." Communications of the ACM 63.11 (2020): 139-144.
[9] Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." Advances in neural information processing systems 34 (2021): 8780-8794.
[10] Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 1.2 (2022): 3.
[11] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[12] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.
[13] Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.
[14] Li, Peike, et al. "Self-correction for human parsing." IEEE Transactions on Pattern Analysis and Machine Intelligence 44.6 (2020): 3260-3271.
[15] Jiang, Yuming, et al. "Text2human: Text-driven controllable human image generation." ACM Transactions on Graphics (TOG) 41.4 (2022): 1-11.