УДК 004.852

Баненков Максим Михайлович, магистрант, Саратовский государственный технический университет имени Гагарина Ю.А., г. Саратов

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ЗАДАЧИ ВИРТУАЛЬНОЙ ПРИМЕРКИ ОДЕЖДЫ: ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНЫЕ СЕТИ И ДИФФУЗИОННЫЕ МОДЕЛИ

Аннотация

В статье проведен сравнительный анализ двух подходов к решению задачи виртуальной примерки одежды на основе глубокого обучения: методов на основе генеративно-состязательных сетей и диффузионных моделей. Рассмотрены архитектурные принципы каждого подхода, представлены ключевые модели и их ограничения. Количественное сравнение выполнено по метрикам FID, SSIM и LPIPS на датасете VITON-HD. Показано, что диффузионные модели обеспечивают более высокое качество синтеза и устойчивость к вариациям входных данных, вытесняя GAN-методы в качестве доминирующей архитектурной парадигмы.

Ключевые слова: виртуальная примерка одежды, генеративно-состязательные сети, диффузионные модели, латентная диффузия, синтез изображений, перенос одежды, компьютерное зрение.

Annotation

The article presents a comparative analysis of two deep learning approaches to the virtual clothing try-on task: generative adversarial networks and diffusion models. The architectural principles of each approach are examined, key models and their limitations are presented. Quantitative comparison is performed using FID, SSIM,

and LPIPS metrics on the VITON-HD dataset. It is shown that diffusion models provide superior synthesis quality and robustness to input variations, displacing GAN-based methods as the dominant architectural paradigm.

Keywords: virtual clothing try-on, generative adversarial networks, diffusion models, latent diffusion, image synthesis, garment transfer, computer vision.

Введение

Задача виртуальной примерки одежды формулируется следующим образом: по изображению человека и изображению предмета одежды синтезировать фотореалистичное изображение, на котором данная одежда наложена на тело человека с учётом его позы и пропорций при точном сохранении визуальных свойств изделия — цвета, текстуры и паттернов ткани. Задача относится к классу обусловленной генерации изображений и является нетривиальной: корректное решение требует одновременного выполнения семантической сегментации, оценки позы тела, нелинейного пространственного преобразования и фотореалистичного синтеза.

Классические методы обработки изображений не позволяют решить задачу в полной мере ввиду высокой вариативности входных данных. Начиная с 2018 года задачу решают методами глубокого обучения, и с тех пор сложились два основных направления. В период 2018–2022 годов доминировали методы на основе генеративно-состязательных сетей; с 2023 года в качестве более эффективной альтернативы утвердились диффузионные модели. Настоящая статья систематизирует оба направления, анализирует их архитектурные принципы и обосновывает выбор диффузионной парадигмы для разработки современных систем виртуальной примерки.

Методы на основе генеративно-состязательных сетей

Генеративно-состязательные сети реализуют состязательную схему обучения: генератор производит синтетические изображения, дискриминатор минимизирует вероятность их принятия за реальные [4]. Применительно к задаче виртуальной примерки стандартным стал двухэтапный конвейер. На первом этапе геометрический модуль согласования выполняет параметрическое пространственное преобразование изображения одежды: по предсказанным точкам соответствия применяется деформация, приводящая плоское каталожное изображение к контуру тела на целевом снимке. На втором этапе генеративная сеть синтезирует итоговое изображение, совмещая деформированную одежду с сохраненными участками исходной фотографии.

VITON (2018) [5] был первой работой, сформулировавшей задачу в современном виде и введшей стандартный датасет для оценки. Несмотря на ограниченное разрешение и артефакты при нестандартных позах, предложенная двухэтапная схема определила архитектурный стандарт направления. CP-VTON (2018) [6] усовершенствовал геометрический модуль, добавив явные функции потерь на пространственное соответствие признаков, что снизило число артефактов деформации. VITON-HD (2021) [7] поднял разрешение до 1024×768 пикселей и предложил нормализацию, устойчивую к неточностям автоматической сегментации; данная модель стала де-факто стандартом для количественного сравнения GAN-методов. HR-VITON (2022) [8] ввёл сквозное совместное обучение геометрического и генеративного модулей через механизм условного потока признаков, повысив сохранность деталей при значительных трансформациях.

Архитектура двухэтапного конвейера порождает каскадную зависимость: ошибки геометрического согласования на первом этапе распространяются и

усиливаются на стадии синтеза. Состязательное обучение нестабильно — дисбаланс динамики генератора и дискриминатора требует тщательной настройки и нередко ведет к деградации обучения. При значительных окклюзиях, нестандартных позах и отклонениях телосложения от обучающего распределения качество генерации заметно снижается. К 2022 году стало очевидно, что указанные ограничения носят системный характер и не устранимы в рамках данной парадигмы [9].

Диффузионные модели в задаче виртуальной примерки

Диффузионные модели основаны на вероятностном формализме цепи Маркова: прямой процесс итеративно добавляет гауссовский шум к изображению вплоть до его полного разрушения, обратный процесс — обучаемая нейронная сеть — поэтапно восстанавливает изображение из шума, руководствуясь управляющими условиями [10]. Переход к латентным диффузионным моделям, в которых денойзинг выполняется в компактном латентном пространстве предобученного автоэнкодера, кратно снизил вычислительные требования и открыл возможность работы с изображениями высокого разрешения [11]. Архитектура Stable Diffusion, реализующая данный принцип, стала базовой платформой для большинства современных VTON-систем.

TryOnDiffusion (2023) [12] — первая работа, применившая диффузионные модели к VTON. Предложенная архитектура Parallel-UNet обрабатывает изображения человека и одежды двумя независимыми ветвями, объединёнными посредством перекрестного внимания; отказ от явного геометрического варпинга позволяет модели самостоятельно обучить необходимые пространственные соответствия. Данная система впервые превзошла лучшие GAN-решения по всем ключевым метрикам. OOTDiffusion

(2024) [13] предложил механизм целевого слияния признаков одежды и тела внутри денойзирующей сети при поддержке примерки как отдельных элементов гардероба, так и полного образа. OutfitAnyone (2024), разработанный Alibaba, является наиболее функционально полным открытым решением на текущий момент. Ключевой новацией служит сеть ReferenceNet, архитектурно воспроизводящая U-Net денойзера и встраивающая признаки одежды через механизм пространственного внимания; это обеспечивает точную передачу текстур и принтов при произвольных деформациях. Система поддерживает явное управление позой через скелетные карты или параметрическую модель тела SMPL, генерацию в разрешении до 1080×1920 и постпроцессинговый рефайнер для восстановления высокочастотных деталей.

Организация системы хранения

Оценка проведена на датасете VITON-HD (11 647 пар изображений) по трём метрикам: FID характеризует близость распределений реальных и синтезированных изображений в признаковом пространстве; SSIM оценивает структурное сходство; LPIPS — перцептивное качество на основе активаций глубоких нейронных сетей. По всем трем показателям диффузионные модели демонстрируют устойчивое превосходство; значение FID лучшей диффузионной системы более чем вдвое ниже аналогичного показателя лучшей GAN-модели (таблица 1).

Метод	Год	FID	SSIM	LPIPS
VITON-HD	2021	26,7	0,861	0,062

HR-VITON	2022	21,4	0,878	0,054
TryOnDiffusion	2023	13,6	0,902	0,038
OOTDiffusion	2024	12,1	0,911	0,034
OutfitAnyone	2024	10,8	0,923	0,029

Таблица 1. Сравнение методов на датасете VITON-HD

Параметрическое геометрическое преобразование, применяемое в GAN-методах, обеспечивает приемлемое качество для однородных изделий, однако систематически вносит артефакты при сложных принтах и нестандартных позах. Диффузионные модели передают визуальные свойства одежды через механизмы внимания, воспроизводя детали в процессе итеративной денойзинговой генерации, что принципиально устойчивее к вариативности входных данных. Помимо этого, GAN-методы, обученные на датасетах с ограниченным диапазоном поз, деградируют при значительных отклонениях от обучающего распределения; диффузионные системы с модулем явного управления позой обеспечивают стабильную генерацию для произвольных конфигураций тела. Единственное направление, в котором GAN-методы сохраняют преимущество, — скорость инференса: 0,1–0,5 с против 1–5 с у диффузионных моделей с ускоренным сэмплированием. Сводное сравнение по ключевым критериям приведено в таблице 2.

Критерий	GAN-методы	Диффузионные модели
Качество синтеза	Среднее	Высокое

Сохранность текстур	Умеренная	Высокое
Устойчивость к позам	Ограниченная	Высокое
Поддержка полного образа	Ограниченная	Полная
Скорость инференса	0,1–0,5 с	1–5 с
Гибкость управления генерацией	Опосредованная	Прямая (поза, текст)
Стабильность обучения	Нестабильное	Стабильное

Таблица 1. Сравнение методов на датасете VITON-HD

Текущий вектор развития направлен на сокращение разрыва в скорости инференса диффузионных моделей. Применение дистилляции знаний, методов последовательной консистентности и адаптивного числа шагов денойзинга позволяет существенно снизить латентность без значимой потери качества. Параллельно исследуются гибридные архитектуры, сочетающие геометрический модуль согласования с диффузионным генератором: предварительная деформация повышает точность пространственных соответствий, диффузионный синтез обеспечивает фотореалистичность результата.

Заключение

Методы на основе генеративно-состязательных сетей сформировали формальный фундамент задачи виртуальной примерки: ввели стандартную постановку, базовые датасеты и метрики. Вместе с тем каскадная зависимость этапов конвейера, нестабильность состязательного обучения и ограниченная

обобщаемость на нестандартные позы и телосложения являются структурными недостатками, не устранимыми в рамках данной архитектуры.

Диффузионные модели устранили указанные ограничения, перейдя к вероятностной обусловленной генерации в латентном пространстве. Передача визуальных свойств одежды через механизмы внимания обеспечивает высокую сохранность деталей; явное управление позой повышает устойчивость к вариациям входных данных. По совокупности метрик ведущие диффузионные системы превосходят лучшие GAN-решения на 40–60% по показателю FID. Диффузионная парадигма является архитектурно обоснованным выбором при проектировании современных информационных систем виртуальной примерки одежды.

Литература

1. Гудфеллоу, Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль. — М.: ДМК Пресс, 2018. — 652 с.
2. Форсайт, Д. Компьютерное зрение. Современный подход / Д. Форсайт, Ж. Понс. — М.: Вильямс, 2019. — 928 с.
3. Паттерсон, Дж. Глубокое обучение с точки зрения практика / Дж. Паттерсон, А. Гибсон. — СПб.: БХВ-Петербург, 2018. — 418 с.
4. Goodfellow, I. Generative adversarial nets / I. Goodfellow et al. // NeurIPS. — 2014. — Vol. 27.
5. Han, X. VITON: An image-based virtual try-on network / X. Han et al. // CVPR. — 2018. — С. 7543–7552.
6. Wang, B. CP-VTON: Cloth-Preserving Virtual Try-On Network / B. Wang et al. // ECCV. — 2018.
7. Choi, S. VITON-HD: High-resolution virtual try-on via misalignment-aware normalization / S. Choi et al. // CVPR. — 2021. — С. 14131–14140.

8. Lee, S. High-resolution virtual try-on with misalignment and occlusion-handled conditions / S. Lee et al. // ECCV. — 2022. — C. 204–219.

9. Dhariwal, P. Diffusion models beat GANs on image synthesis / P. Dhariwal, A. Nichol // NeurIPS. — 2021. — Vol. 34.

10. Ho, J. Denoising diffusion probabilistic models / J. Ho, A. Jain, P. Abbeel // NeurIPS. — 2020. — Vol. 33.

11. Rombach, R. High-resolution image synthesis with latent diffusion models / R. Rombach et al. // CVPR. — 2022. — C. 10684–10695.

12. Zhu, L. TryOnDiffusion: A tale of two UNets / L. Zhu et al. // CVPR. — 2023. — C. 4606–4615.

13. Xu, Y. OOTDiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on / Y. Xu et al. // arXiv:2403.01779. — 2024.

Literature

1. Goodfellow, I., Bengio, Y., & Courville, A. Deep Learning. MIT Press, 2016. (Russian edition: DMK Press, 2018).

2. Forsyth, D. A., & Ponce, J. Computer Vision: A Modern Approach. Pearson, 2011. (Russian edition: Williams, 2019).

3. Patterson, J., & Gibson, A. Deep Learning: A Practitioner's Approach. O'Reilly Media, 2017. (Russian edition: BHV-Petersburg, 2018).

4. Goodfellow, I. et al. Generative adversarial nets. NeurIPS, 2014. Vol. 27.

5. Han, X. et al. VITON: An image-based virtual try-on network. CVPR, 2018. pp. 7543–7552.

6. Wang, B. et al. CP-VTON: Cloth-Preserving Virtual Try-On Network. ECCV, 2018.

7. Choi, S. et al. VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. CVPR, 2021. pp. 14131–14140.

8. Lee, S. et al. High-resolution virtual try-on with misalignment and occlusion-handled conditions. ECCV, 2022. pp. 204–219.

9. Dhariwal, P., Nichol, A. Diffusion models beat GANs on image synthesis. NeurIPS, 2021. Vol. 34.

10. Ho, J., Jain, A., Abbeel, P. Denoising diffusion probabilistic models. NeurIPS, 2020. Vol. 33.

11. Rombach, R. et al. High-resolution image synthesis with latent diffusion models. CVPR, 2022. pp. 10684–10695.

12. Zhu, L. et al. TryOnDiffusion: A tale of two UNets. CVPR, 2023. pp. 4606–4615.

13. Xu, Y. et al. OOTDiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. arXiv:2403.01779, 2024.