Отчет 1 по Face/Hair swap'y

Игнатов Максим

Введение

В данном отчете я напишу отчет про прочитанные про статьи [1], [2], [3], [4]. В целом, первые 3 статьи, основанные на GAN, имеют много похожего, например, используют одинаковые компоненты в лоссе, так что я сосредоточусь на вкладе и особенностях в подходах.

1 One Shot Face Swapping on Megapixels [1]

Contribution

1. Первый подход, позволяющий делать face swap на мегапиксельном уровне

2. Новая архитектура кодировщика HieRFE: Кодировщик иерархически организует представление лица в расширенном латентном пространстве W^{++}, сохраняя более детализированную информацию по сравнению с предыдущими методами. Также новый способ манипуляции латентным кодом, позволяющий одновременно управлять несколькими атрибутами без явного разделения признаков.

3. Датасет для детекции дипфейков.

Основные методы

GAN Inversion. Авторы используют латентное пространство $W^+ \in \mathbb{R}^{18\times512}$ вместо ранее предложенного $W \in \mathbb{R}^{1\times512}$, чтобы обеспечить более точное восстановление изображения. Энкодер обучается либо для предсказания латентного представления, либо для минимизации ошибки между представлениями при случайных инициализациях.

Latent Code Manipulations. Операции семантического редактирования реализуются через добавление высокоразмерных направлений.

Архитектура модели

Модель состоит из трёх компонентов:

1. Face Encoder (HieRFE), состоящий из ResNet-like бэкбоуна, проецирующий в латентное пространство ($l \in \mathbb{R}^{1\times 512}$). Эту часть учат на следующем лоссе (x изначальное изображение, $\hat{x}$ восстановаленное StyleGAN'ом):

$$L_{\text{rec}} = \|x - \hat{x}\|_2$$
$$L_{\text{LPIPS}} = \|F(x) - F(\hat{x})\|_2$$
$$L_{\text{id}} = 1 - \cos(R(x), R(\hat{x}))$$
$$L_{\text{ldm}} = \|P(x) - P(\hat{x})\|_2$$
$$L_{\text{inv}} = \lambda_1 L_{\text{rec}} + \lambda_2 L_{\text{LPIPS}} + \lambda_3 L_{\text{id}} + \lambda_4 L_{\text{ldm}}$$

   где $F(\cdot)$ это экстрактор фич лица, $R(\cdot)$ – модель, обученная на ArcFace лоссе, $P(\cdot)$ – модель для определения точек лица.

2. Latent code manipulation. состоит из Face Transfer Module состоит из 14 Face transfer Blocks (которые похожи на LSTM). Учат на примерно таком же лоссе, только для выходов FTM.

3. **StyleGAN2.** берется генератор для получения картинки.

Комментарии

Кажется, статья – старый бейзлайн по GAN-based, subject-agnostic face swap. Плохая работа со светом, качество общей реконструкции нормальное, есть проблемы с поворотами лица.
Не смог запустить пока из-за размера StyleGAN2.

2 Generative High fidelity One Shot Transfer [2]

Contribution

1. Eye-based loss, помогающий синхронизовать направление взгляда и в целом глаза.

2. Более плавное встраивание source лица.

3. Техника для стабилизации лица для соседних кадров видео

Архитектура модели

Авторы сохранили архитектуру из более старой статьи FaceShifter, в крадце: в качестве Identity Encoder'a выступает модель, обученная на ArcFace лоссе, с помощью U-net-like архитектуры извлекаем из target'a признаки, далее последовательно вмешиваем в них вектор, полученный из Identity Encoder'a.
Стоит добавить, что в экспериментах авторы пробуют разные вариации архитектур для генератора и source энкодера, но прирост в метриках там не очень большой.

Лосс

Добавили и изменили (X_s – source face, X_t – target face):

1. L_{rec} взяли из SimSwap, теперь ненулевой при разных изображениях одного человека P^i: $L_{rec} = \begin{cases} \|\hat{Y}_{s,t} - X_t\|_2^2, & X_s, X_t \in P^i \\ 0, & \text{otherwise} \end{cases}$
2. $L_{eye} = \|hm(\hat{Y}) - hm(X_t)\|_2^2$, где $hm(\cdot)$ отвечает за heatmap'у глаз на изображениях.

Комментарии

Получилось запустить демо, проверял только на статейных датасетах, в целом, результаты похожие.
Не очень понял, почему статья пошла в журнал, а не на конференцию. Могу предположить, что результаты не очень хорошие и в целом, новых вещей представили мало, т.к. больше объединили подходы с других статьей.
Сначала было не понятно, почему надо было сначала сжимать изображение, а потом использовать Super Resolution модель, а не, например, использовать VAE какой-нибудь, но потом увидел, что статья до LDM вышла, может из-за этого.

3 High-resolution Face Swapping via Latent Semantics Disentanglement [3]

Contribution

Авторы раскрывают семантику скрытого представления в StyleGAN, помогая переносить личность и структуру объекта.

Два новых ограничения для улучшения согласования лиц в видео, включая ограничение code trajectory, которое ограничивает смещение между скрытыми кодами соседних кадров, и ограничение траектории потока, которое работает в пространстве RGB, гарантируя плавность видео.

тесты на разных датасетах

Основные методы

Метод начинает различать атрибуты, такие как поза, выражение лица (структура лица, structure attributes) и цвет и освещение (структура внешнего вида, appearance attributes) и работает с ними по разному.
Авторы декомпозирует структуру и внешние атрибуты, используя иерархическое скрытое пространство StyleGAN: Первые K векторов латентного кода используются для энкодинга для структурных атрибутов, остальные (более глубокие) для appearance attributes.
Предыдущие методы встраивали получиное лицо с помощью пуассоновского смешивания, в итоге получалось изображение с артифактами на краях, в этой работе предлагается подход, в котором фичи из энкодера StyleGAN последовательно подмешиваются в итоговый энкодер, заменяя фичи из энкодера для target.

Для работы с видео в лосс добавляется компонента (латентный код в предсказанных кадрах должен изменяться так же, как и в настоящих): $L_{ct} = \sum_{k=1}^{M} \|(\hat{g}_s^k - \hat{g}_s^{k-1}) - (g_s^k - g_s^{k-1})\|$
Также используется предсказание оптического потока, пусть $f_{t\Rightarrow j} = \Phi(y_f^i, y_f^j)$, где $\Phi(\cdot,\cdot)$ это претрейн сетки для optical flow (PWC-Net), тогда делается предположение о линейности изменения направления потока (причем авторы замечают, что это не всегда верно, например для объектов с ускорением): $L_{ft} = \sum_k \|(f_{k\Rightarrow k+2} + f_{k+2\Rightarrow k})/2 - f_{k\Rightarrow k+1}\|_2$

4 DiffSwap (High-Fidelity and Controllable Face Swapping via 3D-Aware Masked Diffusion) [4]

Contribution

1. Face swap as conditional impainting (conditional masked diffusion on the latent space).

2. Высокая точность (midpoint estimation method для эффективного восстановления приемлемого результата диффузии для лиц).

3. Сохранение формы (управляемость метода позволяет использовать 3D ориентиры в качестве conditioning'а при генерации для сохранения формы лица из source изображения)

Метод

Авторы формулируют задачу Face-swap как conditional inpainting, атрибуты сгенерированных лиц и исходного контролируются через вектора состояния.
Для начала авторы тренеруют VQ-GAN, проецирующий в латентное пространство, затем LDM, причем с сохранением identity сгенерированной картинки. Во время инференса конструируется бинарная сегментационная маска в латентном пространстве, для контроля в рабочем регионе.
Conditioning выполняется через эмбеддинг лица, который опять получается через претрейн ArcFace, далее проецируются в нужную размерность через MLP. Еще добавляется вектор из landmark features и отдельно маски для регионов носа, глаз, рта. Затем маски проецируются в нужное пространство и к ним применяется multi head self attention, чтобы уловить связь между регионами.

Функция потерь

В предыдущих работах показывается важность identity лосса (cosine distance between ArcFace model embeddings), т.к. во время деноизинга нам надо делать многоэтого проходов моделью, возникает сложность с нахождением id. Для авторы вводят midpoint estimation.
Суть метода в том, чтобы сократить число шагов для расшумления изображения во время обучения:

1. $t_1 = \lfloor t/2 \rfloor$ итоге получаем промежуточное приближение $\hat{z}_{t_1} = \frac{z_t - \sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t_1}}\epsilon_\theta(z_t,t,C)}{\sqrt{\bar{\alpha}_t/\bar{\alpha}_{t_1}}}$
2. Получаем итоговое изображение с помощью промежуточного результата $\hat{z}_0^{\text{midpoint}} = \frac{\hat{z}_{t_1} - \sqrt{1-\bar{\alpha}_{t_1}}\epsilon_\theta}{\sqrt{\bar{\alpha}_{t_1}}}$

Инференс

Во время инференса данное изображение x^{tgt} с помощью энкодера $\mathcal{E}$ проецируется в латентное пространство как z_0^{tgt}. Затем выполняется conditional impainting в латентном пространстве:

$$z_t^{\text{tgt}} \leftarrow \sqrt{\alpha_t} z_0^{\text{tgt}} + \sqrt{1-\alpha_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I)$$

$$z_t \leftarrow M \odot z_t + (1 - \mathcal{M}) \odot z_t^{\text{tgt}}$$

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(z_t, t; C_{\text{swap}}) \right) + \sigma_t \cdot \epsilon_t$$

$$z_T, n_t, \epsilon_t \in \mathcal{N}(0, I), t = T, ..., 1$$

Для сохранения пространственной информации о лице, используется фреймворк для 3D реконструкции лица, из которого можно получить информацию о форме, выражении, позе лица. В таком случае, можно заменить информацию о форме target лица с помощью source лица. Итоговая маска $\mathcal{M}$ конструируется как выпуклая оболочка (convex hull) из точек обоих лиц.

Комментарии

Картинки получаются лучше, чем у [3], [2], лучше работа со светом, формой и тд, переносятся атрибуты, как усы и брови. Из минусов, не подходит для работы с видео.
Кода нет, но очень интересно

5 Face Adapter for Pre-Trained Diffusion Models with Fine-Grained ID and Attribute Control

The authors introduce the Face-Adapter, an efficient adapter designed to enhance the performance of pre-trained diffusion models in tasks such as face reenactment and swapping.

5.1 Spatial Condition Generator

The Spatial Condition Generator (SCG) is designed to enhance precision and provide spatial guidance for face reenactment and swapping tasks. It predicts 3D facial landmarks and generates masks for areas that require modification, addressing challenges such as background inconsistencies, facial shape variations, and fine-grained motion control.

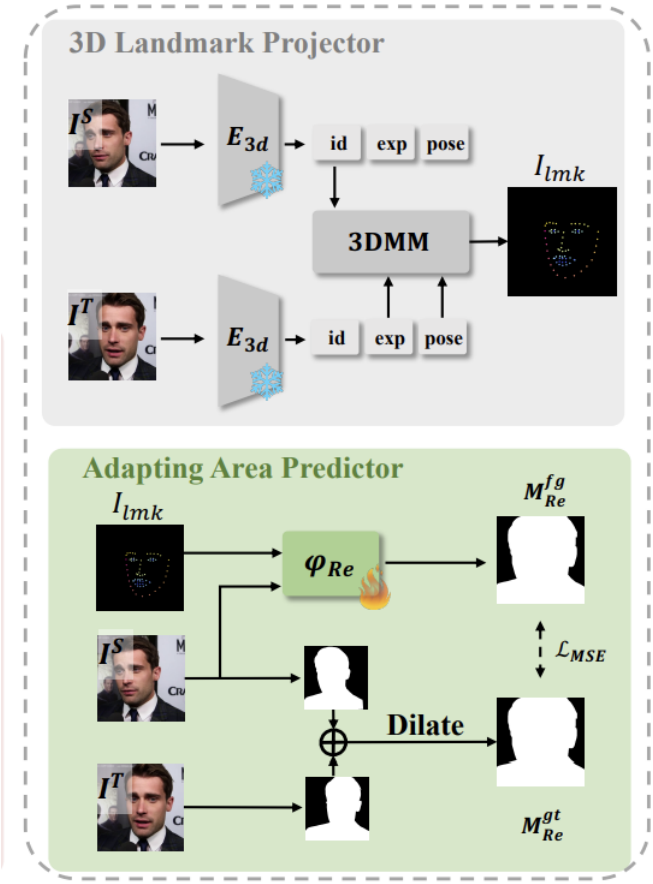

Рис. 1: Enter Caption

Components of the SCG:

1. 3D Landmark Projector: - Uses a 3D reconstruction method E_{3d} to separately extract identity, expression, and pose coefficients for the source and target images.
   - Combines the source identity with the target's expression and pose to reconstruct a new 3D face.

- Projects this reconstructed face to generate precise motion landmarks I_{lmk}.
- This allows the SCG to provide more accurate guidance for facial movements, even under large pose variations.

2. Adapting Area Predictor:
- Identifies the ROI in the target images that requires editing. - Predicts masks that cover both the source and target's head regions (e.g., hair, face, neck), ensuring seamless blending.
- In face reenactment, it accounts for background motion caused by camera or object movements, ensuring consistency and avoiding artifacts like blurry backgrounds.
- For face swapping, supplying the target background can also give the model clues about environmental lighting and spatial references.
- Mask predictor φ_{Re} that accepts the target image I^T and motion landmarks I_{lmk} is trained to predict the adapting area mask M_{Re}^{fg}. The mask ground truth M_{Re}^{gt} is generated by taking the union of the head regions, followed by dilation. Head regions are obtained using a pre-trained face parsing model.

Рис. 2: Enter Caption

5.2 Identity Encoder

The Identity Encoder ensures robust identity preservation in the generated images by integrating identity embeddings into the pre-trained diffusion model.

Workflow:

1. A pre-trained face recognition model (ArcFace, E_{id}) extracts a high-level face embedding (f_{id}) from the source image.

2. A lightweight three-layer transformer decoder ϕ_{dec} maps the face embedding into the text semantic space of the pre-trained diffusion model, obtaining

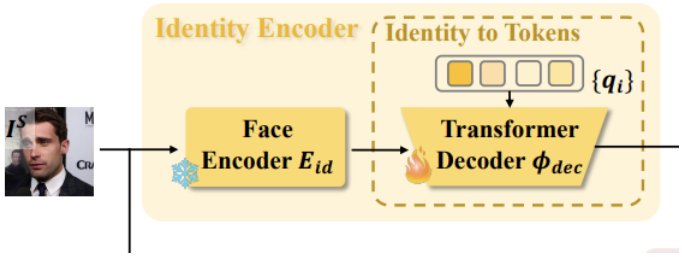

Рис. 3: Enter Caption

the identity tokens. This mapping uses learnable queries ($q_{id} = \{q_1, ..., q_N\}$, here $N = 77$) to ensure the identity embedding sequence length matches the text input requirement of the diffusion model.

Plug-and-Play Design:

- The Identity Encoder does not require fine-tuning of the diffusion model's parameters. Instead, it seamlessly integrates identity embeddings into the diffusion model's U-Net via cross-attention.
- This approach avoids overfitting and preserves the pre-trained model's generative priors.

5.3 Attribute Controller

In line with ControlNet, authors create a copy of U-Net ϕ_{Ctr} and add spatial control as the conditioning input.

Components of the Attribute Controller

1. Spatial Control:
   - Combines target motion landmarks (from the Spatial Condition Generator) with the background extracted from the target image.
   - Both tasks can be viewed as processes of performing conditional inpainting utilizing the given identity and other missing attribute content, following the provided spatial control I_{Sp}
   - Formally:

$$I_{Sp} = \begin{cases} I^s * (1 - M_{Re}^{fg}) + I_{lmk}^T & \text{for face reenactment,} \\ I^T * (1 - M_{Sw}^{fg}) + I_{lmk}^T & \text{for face swapping} \end{cases}$$

2. Attribute Template:
   - Designed to supplement missing information including lightning, parts

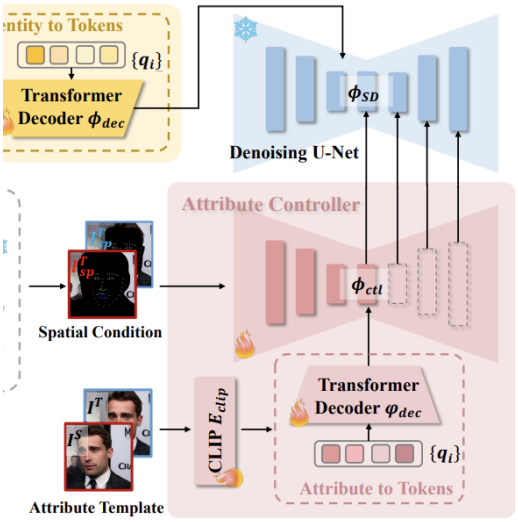

Рис. 4: Enter Caption

of the background and hair.
- Uses the CLIP model to extract attribute embeddings $f_{attr} \in \mathbb{R}^{257*d}$ from the source (for reenactment) or target (for swapping) image.
- Uses patch tokens and global (?) tokens to obtain local and global features
- A three-layer transformer maps these features into the text space with learnable queries ($q = \{q_1, ..., q_k\}$, here $K = 77$) making them compatible with the diffusion model.

5.4 Metrics and experiments

Evaluated at VoxCeleb1/2, FF+. Method achieves comparable or superior

Methods	ID ↑	Pose↓	Exp↓	Gaze↓
FaceShifter [19]†	87.99	0.0342	6.32	0.072
SimSwap [4]*	96.78	**0.0261**	**5.94**	**0.0549**
HifiFace [33]†	94.26	0.0382	6.50	0.0573
InfoSwap [12]*	**99.26**	0.0371	7.25	0.0617
BlendFace [27]*	89.91	0.0286	6.15	0.0556
DiffSwap [49]*	19.16	0.0237	4.94	0.0665
Ours	96.47	0.0319	6.66	0.0607

Рис. 5: Metrics

results in FID, PSNR, and LPIPS compared to SoTA methods.

Methods	Face Reenactmnet					Face Swapping				
	FID↓	Pose↓	Exp↓	Gaze↓	ID↑	FID↓	Pose↓	Exp↓	Gaze↓	ID↑
w/o AAP	33.61	0.0281	3.72	0.045	0.6355	33.97	0.0395	6.13	0.0548	0.4530
w/o CLIP FT	33.09	0.0287	3.74	0.0435	0.6474	31.97	0.0396	6.21	0.0540	0.4696
Full Model	31.18	0.0266	3.61	0.0422	0.6616	30.78	0.0406	6.14	0.0547	0.4688

Рис. 6: Quantitative comparison of the Face Adapter under different ablative configurations

5.5 Ablation study

5.5.1 Adapting Area Predictor

Without AAP, the model struggles with background consistency, resulting in blurry or inconsistent results. It also enables seamless blending with the surrounding due to the fact that the model is not trained on the inpainting task.

5.5.2 CLIP Fine-Tuning

Freezing CLIP parameters results in a decline in attribute accuracy and image quality. Fine-tuning improves extraction of detailed attributes (hair, lighting, etc.) and enhances identity preservation.

6 ReFace (Realistic and Efficient Face Swapping A Unified Approach with Diffusion Models)

6.1 Conditional Inpainting Diffusion Training

6.1.1 Face Shape Augmentation

Target face mask is slightly shifted and transformed during training. This prevents the model from learning to trivially paste the reconstructed target image from x^{tar}.

6.1.2 Reference Augmentation

In the paper $x^{ref} = \mathcal{A}(x^{tar} \otimes m^{tar})$ is defined during inpaining diffusion training and $x^{ref} = \mathcal{A}(x^{src} \otimes m^{src})$ during step (b) (see full pipeline bellow). Where $\mathcal{A}$ is an augment operator comprising of random resize, horizontal flip, rotate, blur and elastic transform operations.
- x^{ref} acts as the reference or "source" image during inpatining training, mimicking the source image used at inference.

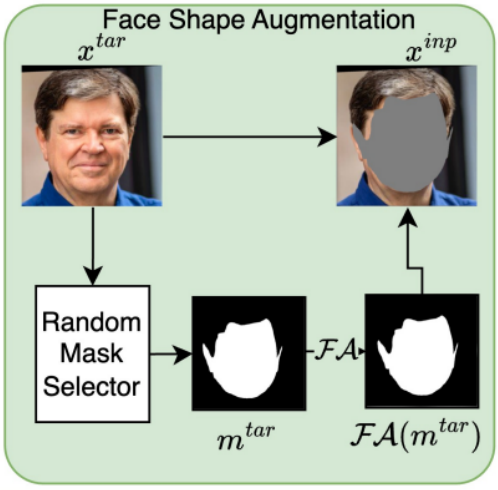

Рис. 7: Enter Caption

- By generating x^{ref} from x^{tar}, the pipeline creates a controlled scenario where the source and target are derived from the same image. This helps the model focus on learning the inpainting task.

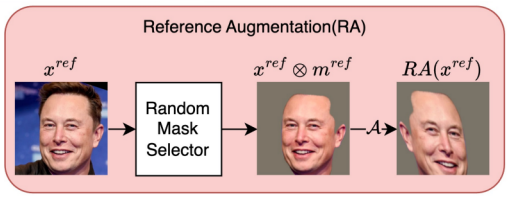

Рис. 8: Enter Caption

6.1.3 Condition Generation

Conditional feature is calculated as weighted average of CLIP, ArcFace, ID, Landmarks embeddings. CLIP was used to extract additional information from the images in addition to previously proposed ArcFace and ID features. All features are projected into same dimension space $\mathbb{R}^D$ using linear layers. The conditional feature is used as the key in each cross-attention layer in the diffusion U-Net. This stage works as a self-supervision where the diffusion model takes the input as $\{z_t, z^{inp}, m^{tar}\}$ (z_t, z^{inp} are latents from $\mathcal{E}$), and using the condition f, outputs the the predicted noise incurred in the diffusion pipeline from step $t-1$ to t. Thus, the loss for this stage is formulated as,

$$L_{Diff} = \mathbb{R}_{t,z_0,\epsilon} \| \epsilon_0(z_t, z^{inp}, m^{tar}, f, t) - \epsilon \|_2^2$$

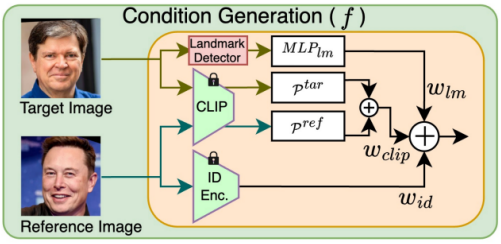

Рис. 9: Enter Caption

6.2 Multistep swapping enhancement loss

This section addresses the need to improve ID transfer and perceptual similarity in face-swapping by introducing additional loss functions and using a multistep DDIM sampler during training: **Challenge:**
Traditional diffusion training predicts noise at one timestep only, which is insufficient to compute identity or perceptual losses effectively. **Solution:**
- A differentiable DDIM sampler is used during training, producing intermediate denoised outputs at multiple steps (NNN-steps).
- Identity and perceptual losses are computed at each denoising step to reinforce the identity transfer and target fidelity.
Total loss can be calculated using following equaions.

$$L_{id} = \sum_{i=1}^{N} (1 - \langle \mathcal{D}(\hat{z}_{0,t_i}) \otimes m^{tar}, x^{src} \otimes m^{src} \rangle)$$

$$L_{ps} = L_{PIPS}(\mathcal{D}(\hat{z}_{0,t_i}), x^{tar})$$

Where $\hat{z}_{0,t_i}$ is obtained using $\hat{z}_0(z_t, t) = \frac{z_t - (\sqrt{1-\overline{\alpha}_t})\epsilon_0(z_t,t)}{\sqrt{\overline{\alpha}_t}}$

$$L_{Total} = L_{Diff} + w_{ID} L_{id} + w_{PS} L_{PS}$$

As mentioned above, $x^{ref} = \mathcal{A}(x^{src} \otimes m^{src})$ used for condition generation (f).

6.3 Mask shuffling

Authors propose augmentation that is used in Face Shape Augmentation and Reference Augmentation.
Technique:

- Randomly select subsets of facial region masks (eyes, skin, background, etc.) from 17 predefined categories.
- Use the selected masks as both reference m^{ref} and target m^{tar} masks during training. This:

1. Prevents overfitting by diversifying the training process, enable the model to generalize beyond standard face-swapping to more complex tasks, such as head swapping (including hair and etc.).

2. Enables the model to generalize beyond standard face-swapping to more complex tasks, such as **head swapping** (including hair and accessories).

3. Improves realism of the generated hairstyles in the head swap task

6.4 Metrics and experiments

Model is evaluated on Celeba-HQ and FFHQ datasets. It shows outperforming FID, ID retrival metrics. Also inference time and training cost is much lower comparing to other diffusion-based methods.

Method	**FID**↓	**ID retrieval** ↑ Top-1	Top-5	**Pose**↓	**Expr.**↓
MegaFS [39]	12.0	59.6%	74.1%	3.33	1.11
HifiFace [33]	11.58	75.3%	87.1%	3.28	1.41
SimSwap [4]	13.8	90.6%	96.4%	2.98	1.07
E4S [20]	12.38	70.2%	82.73%	4.50	1.31
FaceDancer [26]	18.34	0.20%	0.90%	**2.6**	**0.96**
DiffFace [14]	8.59	87.2%	94.4%	3.80	2.28
DiffSwap [38]	8.58	78.2 %	93.6 %	2.92	1.10
Ours	**5.53**	**95.4%**	**98.7%**	3.74	1.04

Рис. 10: Comparison on FFHQ dataset.

Датасеты

Датасетов с лицами много, основные, на которых авторы сравнивают методы это FFHQ, FaceForensics++, CelebA(-HQ).

Метрики

Помимо метрик для качества генерации (FID, PRD) есть специфические метрики для FaceSwap, такие как Pose, Expression Estimation (с помощью точек лица), ID Retrival (Rank@1 классификации)

Общая информация о проделанной работе

Посмотрел все вспомогательные теоретические материалы, прочитал статьи. Статьи я читал полностью, но решил информацию об экспериментах авторов не влючать в обзор, а то он бы слишком большой получился. До выдачи доступа к кластеру смог потестить только GHOST в режиме работы с фото.

Список литературы

[1] https://arxiv.org/pdf/2105.04932

[2] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9851423&tag=1

[3] https://arxiv.org/pdf/2203.15958

[4] https://openaccess.thecvf.com/content/CVPR2023/papers/Zhao_DiffSwap_High-Fidelity_and_Controllable_Face_Swapping_via_3D-Aware_Masked_Diffusion_CVPR_2023_paper.pdf