Everybody dance now

CAROLINE CHAN, SHIRY GINOSAR, INGHUI ZHOU, ALEXEI A. EFROS

UC Berkeley

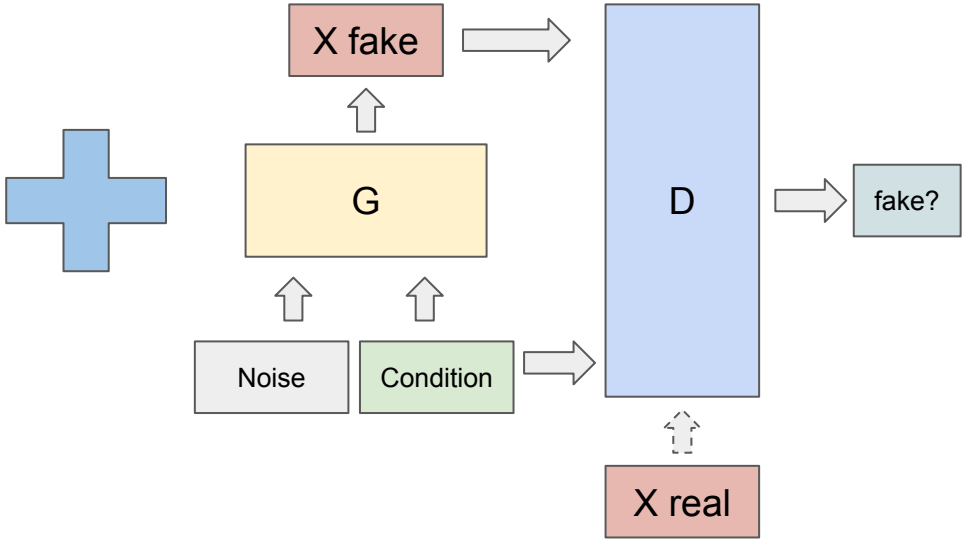
Георгий Кожевников

Что это?

Идея

Pose estimation

Conditional GAN

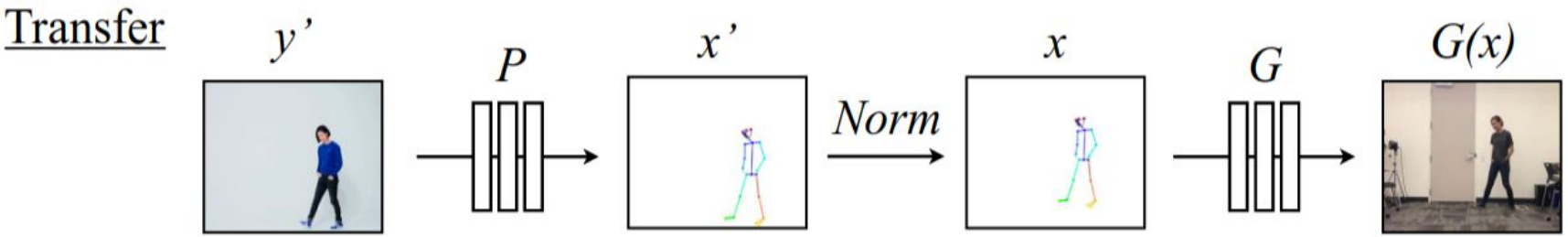


Что это?


Source Video
Detected Pose
Source to Target 1 Result
Source to Target 2 Result

Как это работает?

1. Построение скелета
2. Нормализация скелета
3. Генерация изображения нужного человека по скелету

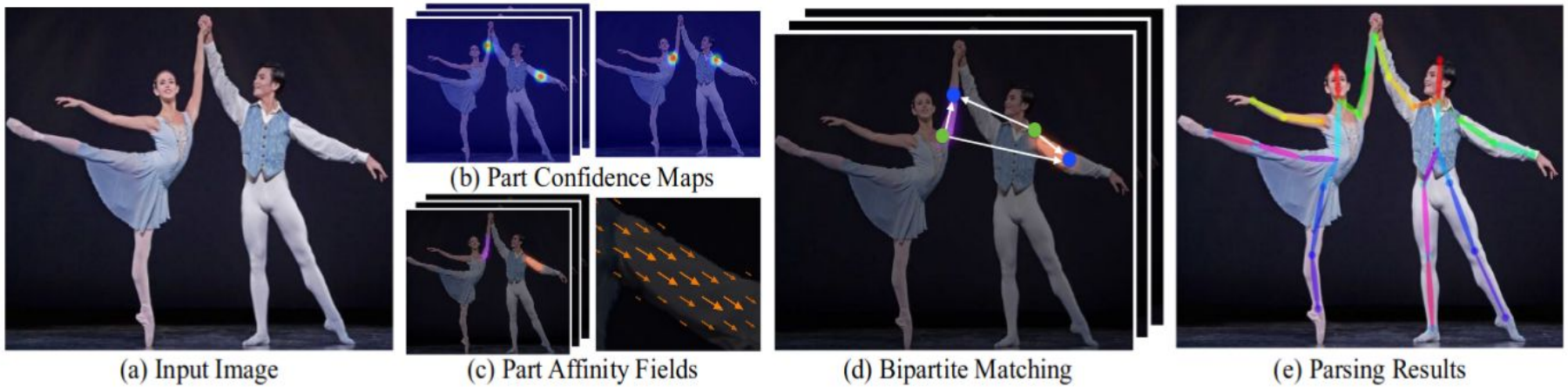


Детали

Построение скелета (pose estimation)

OpenPose: Realtime Multi-Person 2D PoseEstimation using Part Affinity Fields

Berkeley/CMU/Facebook Research, 2017

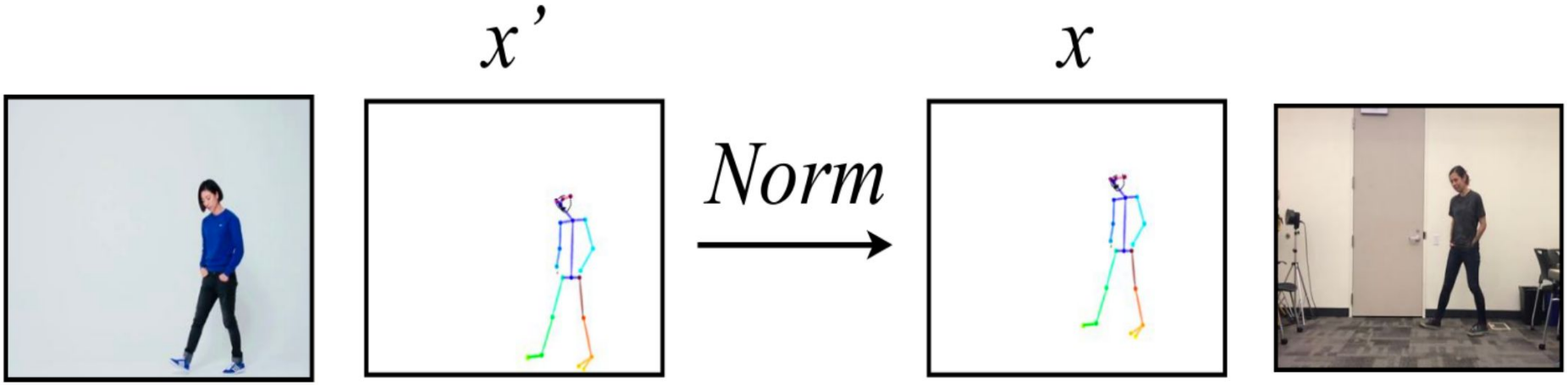


Детали

Нормализация скелета

$$b = t_{min} + \frac{a_{source} - s_{min}}{s_{max} - s_{min}}(t_{max} - t_{min}) - f_{source}$$

$$scale = \frac{t_{far}}{s_{far}} + \frac{a_{source} - s_{min}}{s_{max} - s_{min}}\left(\frac{t_{close}}{s_{close}} - \frac{t_{far}}{s_{far}}\right)$$



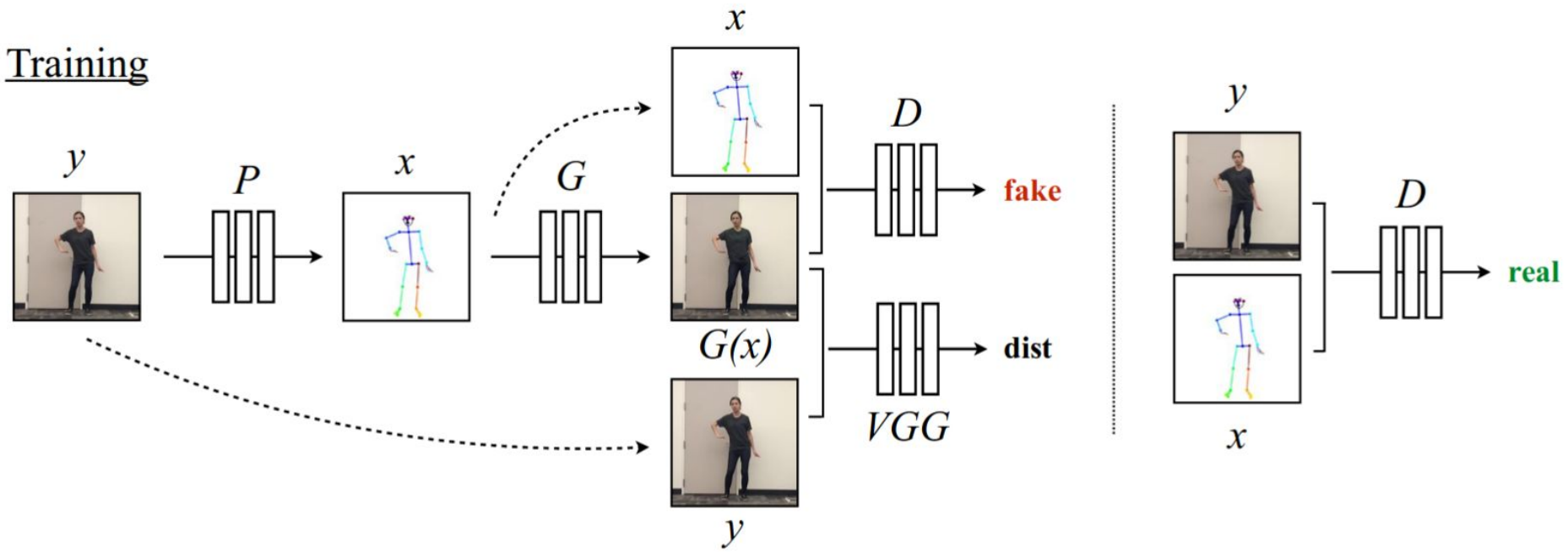
Детали

Генерация изображения по скелету

Особенности:

1. Учёт предыдущего кадра для связности последовательных кадров
2. Отдельная проработка лица

Как это учится?



Как это учится?

High-Resolution Image Synthesis and Semantic Manipulation with Conditional GAN [pix2pixHD]

NVIDIA/Berkeley, 2017

$$\min_G \left(\left(\max_{D_1, D_2, D_3} \sum_{k=1,2,3} \mathcal{L}_{\text{GAN}}(G, D_k) \right) + \lambda_{FM} \sum_{k=1,2,3} \mathcal{L}_{\text{FM}}(G, D_k) + \lambda_{VGG} \mathcal{L}_{VGG}(G(\mathbf{s}), \mathbf{x}) \right)$$

Как это учится?

High-Resolution Image Synthesis and Semantic Manipulation with Conditional GAN [pix2pixHD]

NVIDIA/Berkeley, 2017

$$\min_G \left(\left(\max_{D_1, D_2, D_3} \sum_{k=1,2,3} \mathcal{L}_{\text{GAN}}(G, D_k) \right) + \lambda_{FM} \sum_{k=1,2,3} \mathcal{L}_{\text{FM}}(G, D_k) + \lambda_{VGG} \mathcal{L}_{VGG}(G(\mathbf{s}), \mathbf{x}) \right)$$

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{(\mathbf{s},\mathbf{x})}[\log D(\mathbf{s}, \mathbf{x})] + \mathbb{E}_{\mathbf{s}}[\log(1 - D(\mathbf{s}, G(\mathbf{s}))]$$

Conditional GAN loss

Как это учится?

High-Resolution Image Synthesis and Semantic Manipulation with Conditional GAN [pix2pixHD]

NVIDIA/Berkeley, 2017

$$\min_G \left(\left(\max_{D_1, D_2, D_3} \sum_{k=1,2,3} \mathcal{L}_{\text{GAN}}(G, D_k) \right) + \lambda_{FM} \sum_{k=1,2,3} \mathcal{L}_{\text{FM}}(G, D_k) + \lambda_{VGG} \mathcal{L}_{VGG}(G(\mathbf{s}), \mathbf{x}) \right)$$

$$\mathcal{L}_{\text{FM}}(G, D_k) = \mathbb{E}_{(\mathbf{s},\mathbf{x})} \sum_{i=1}^{T} \frac{1}{N_i} [||D_k^{(i)}(\mathbf{s}, \mathbf{x}) - D_k^{(i)}(\mathbf{s}, G(\mathbf{s}))||_1]$$

Discriminator feature-matching loss

Как это учится?

High-Resolution Image Synthesis and Semantic Manipulation with Conditional GAN [pix2pixHD]

NVIDIA/Berkeley, 2017

$$\min_G \left(\left(\max_{D_1, D_2, D_3} \sum_{k=1,2,3} \mathcal{L}_{\text{GAN}}(G, D_k) \right) + \lambda_{FM} \sum_{k=1,2,3} \mathcal{L}_{\text{FM}}(G, D_k) \right.$$
$$\left. + \lambda_{VGG} \mathcal{L}_{VGG}(G(\mathbf{s}), \mathbf{x}) \right)$$

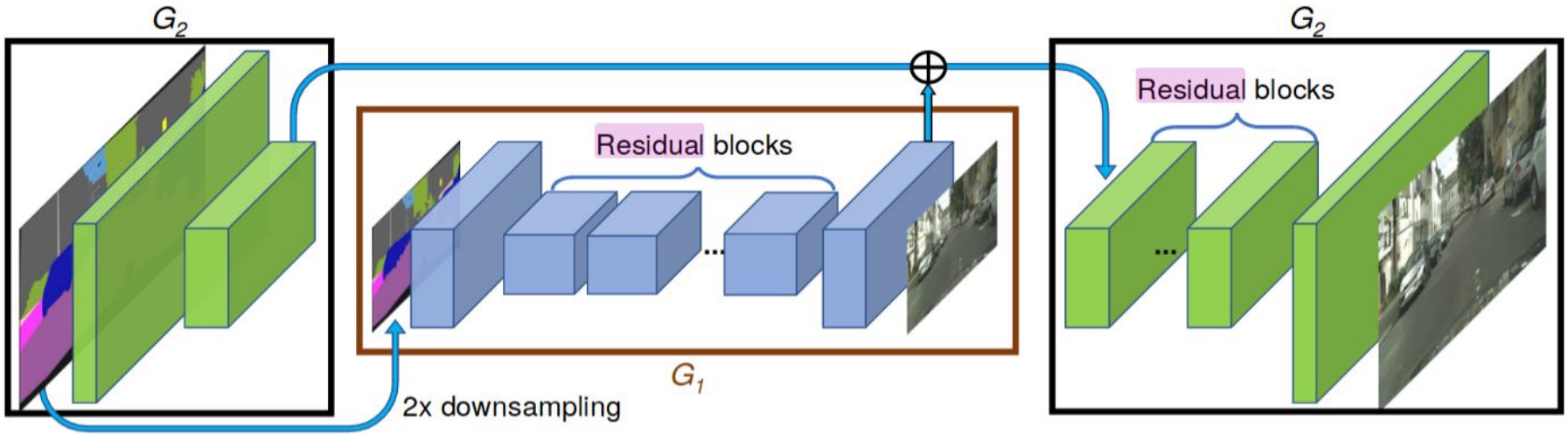
$$\mathcal{L}_{VGG}(G(\mathbf{s}), \mathbf{x}) = \sum_{i=1}^{N} \frac{1}{M_i} \left[\left| \left| F^{(i)}(\mathbf{x}) - F^{(i)}(G(\mathbf{s})) \right| \right|_1 \right]$$

VGG perceptual loss

Как это учится?

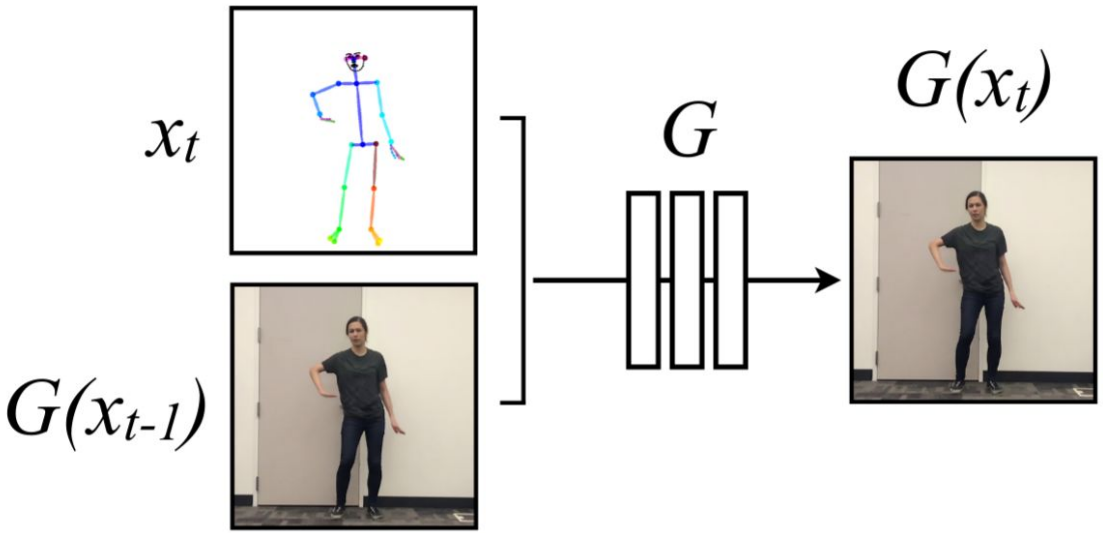
High-Resolution Image Synthesis and Semantic Manipulation with Conditional GAN [pix2pixHD]

NVIDIA/Berkeley, 2017



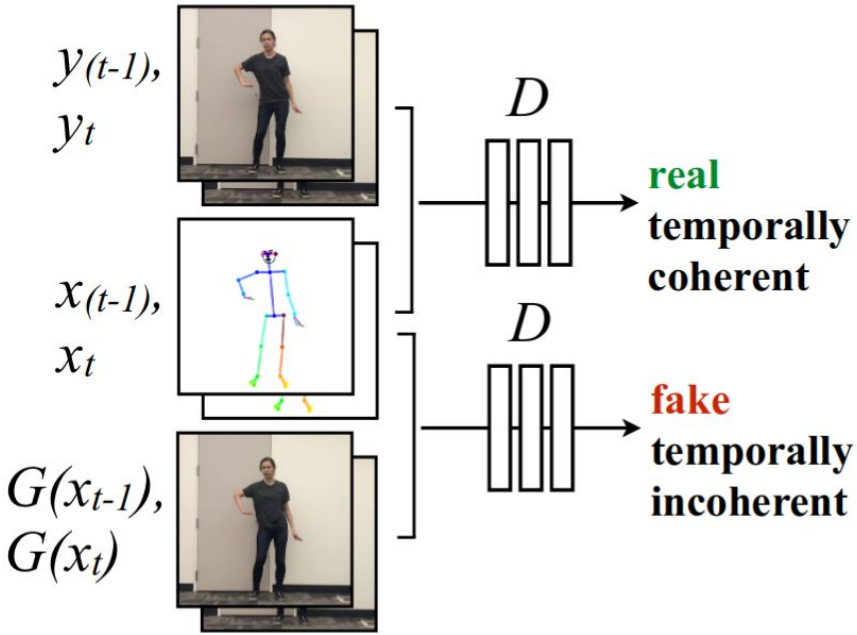
Как это учится?

Учёт предыдущих кадров для связности



Как это учится?

Учёт предыдущих кадров для связности



Как это учится?

Учёт предыдущих кадров для связности

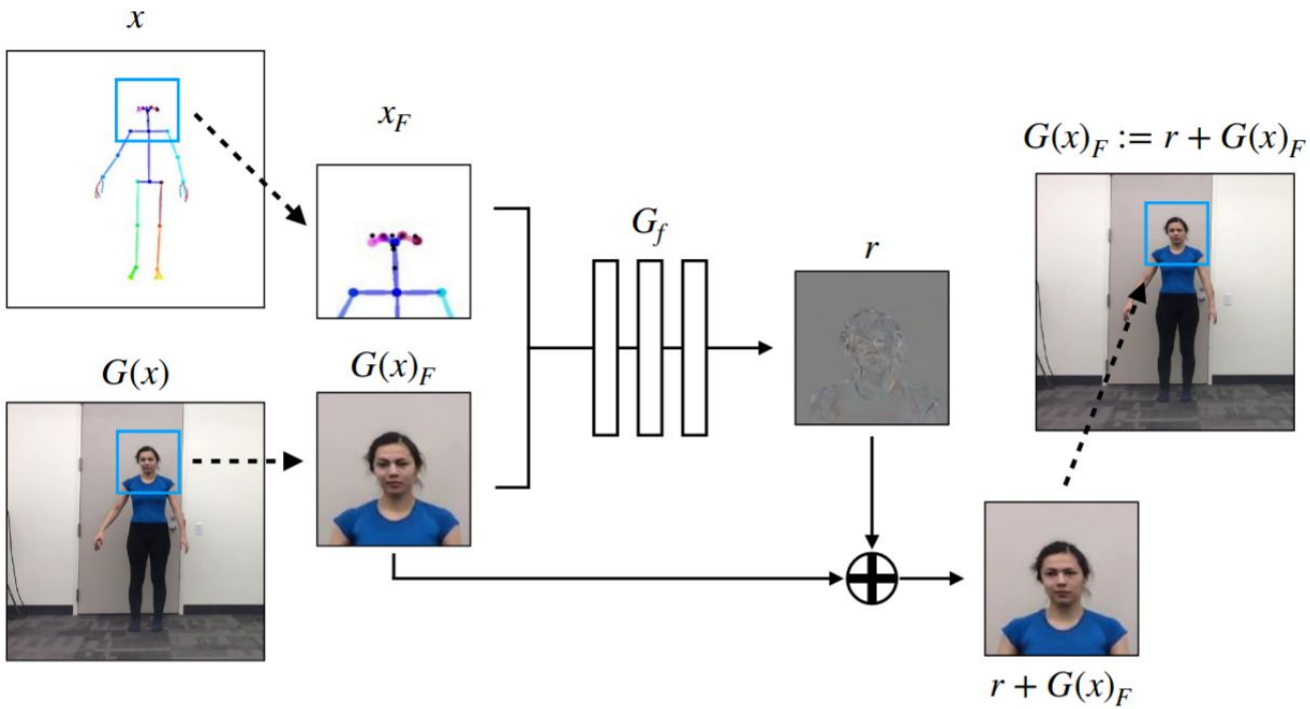
$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{(x,y)}[\log D(x, y)] + \mathbb{E}_x[\log(1 - D(x, G(x))]$$

$$\mathcal{L}_{\text{smooth}}(G, D) = \mathbb{E}_{(x,y)}[\log D(x_{t-1}, x_t, y_{t-1}, y_t)]$$
$$+\mathbb{E}_x[\log(1 - D(x_{t-1}, x_t, G(x_{t-1}), G(x_t))]$$

Как это учится?

Основная идея - добавление к региону лица исходного изображения некоторой маски (residual), которая улучшит его вид.

Доработка лица



Как это учится?

Нужная маска генерируется отдельной моделью (FaceGAN).

Доработка лица

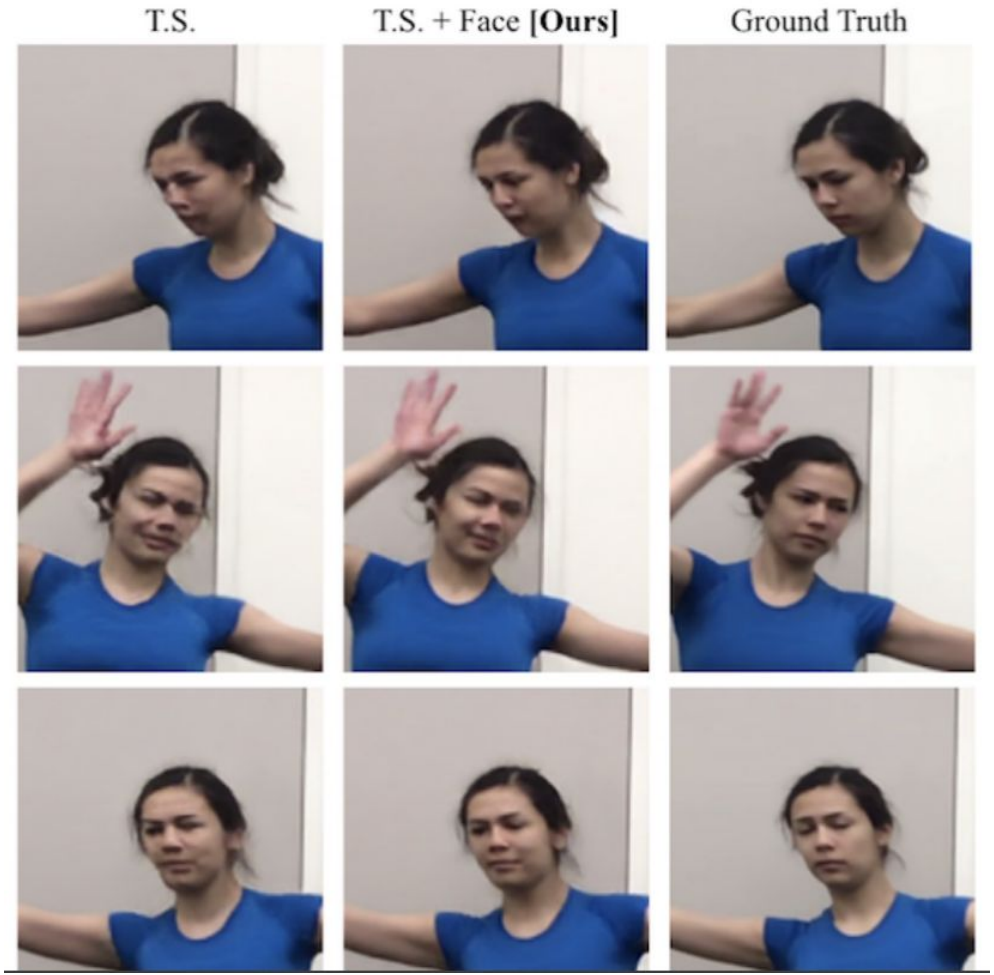
$$\min_{G_f}\left(\left(\max_{D_f}\mathcal{L}_{\text{face}}(G_f, D_f)\right) + \lambda_{VGG}\mathcal{L}_{VGG}(r + G(x)_F, y_F)\right)$$

$$\mathcal{L}_{\text{face}}(G_f, D_f) = \mathbb{E}_{(x_F, y_F)}[\log D_f(x_F, y_F)]$$
$$+\mathbb{E}_{x_F}[\log\left(1 - D_f(x_F, G(x)_F + r)\right)].$$

Как это учится?

Доработка лица

Это работает (местами)



Как это учится?

Итоговый процесс обучения

1. Обучаем основной генератор

$$\min_G \left(\left(\max_{D_1, D_2, D_3} \sum_{k=1,2,3} \mathcal{L}_{\text{smooth}}(G, D_k) \right) + \lambda_{FM} \sum_{k=1,2,3} \mathcal{L}_{\text{FM}}(G, D_k) \right.$$
$$\left. + \lambda_{VGG} \left(\mathcal{L}_{VGG}(G(x_{t-1}), y_{t-1}) + \mathcal{L}_{VGG}(G(x_t), y_t) \right) \right)$$

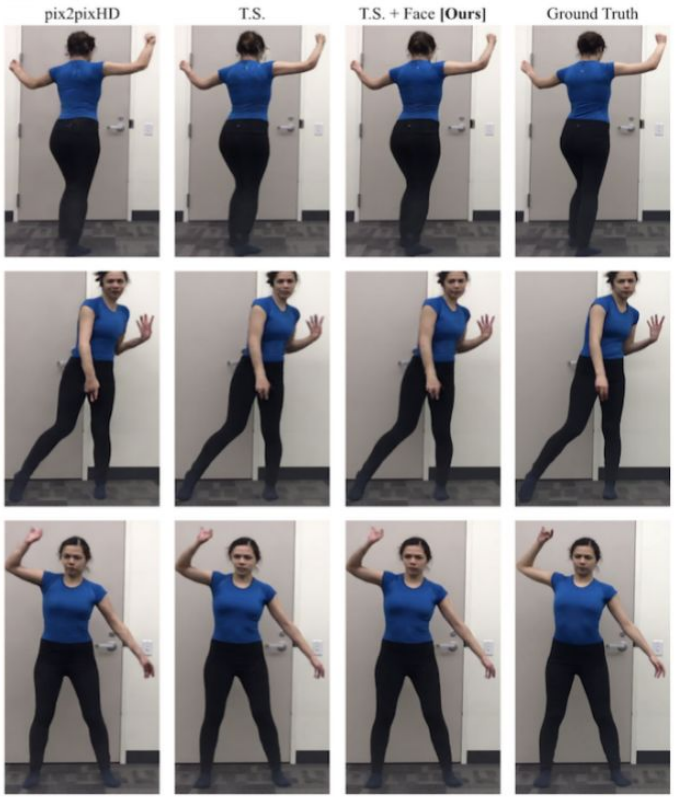
Как это учится?

Итоговый процесс обучения

1. Обучаем основной генератор
2. Обучаем генератор лиц

$$\min_{G_f}\left(\left(\max_{D_f}\mathcal{L}_{\text{face}}(G_f, D_f)\right) + \lambda_{VGG}\mathcal{L}_{VGG}(r + G(x)_F, y_F)\right)$$

Результаты



Loss	SSIM mean	LPIPS mean
pix2pixHD	0.89564	0.03189
T.S.	0.89597	0.03137
T.S. + Face **[Ours]**	**0.89807**	**0.03066**

Table 1. Body output image comparisons - result cropped to bounding box around input pose. For all tables, T.S. denotes a model with our temporal smoothing setup, and T.S. + Face is our full model with both the temporal smoothing setup and Face GAN.

Loss	SSIM mean	LPIPS mean
pix2pixHD	0.81374	0.03731
T.S.	0.8177	0.03662
T.S. + Face **[Ours]**	**0.83046**	**0.03304**

Table 2. Face output image comparisons - result cropped to bounding box around input face

Loss	Body (23)	Face (70)	Hands (21)	Overall (135)
pix2pixHD	**2.39352**	1.1872	3.86359	2.0781
T.S.	2.63446	1.14348	3.76056	2.06884
T.S. + Face **[Ours]**	2.56743	**0.91636**	**3.29771**	**1.92704**

Table 3. Mean pose distances, using the pose distance metric described in Section 7. Lower pose distance is more favorable.

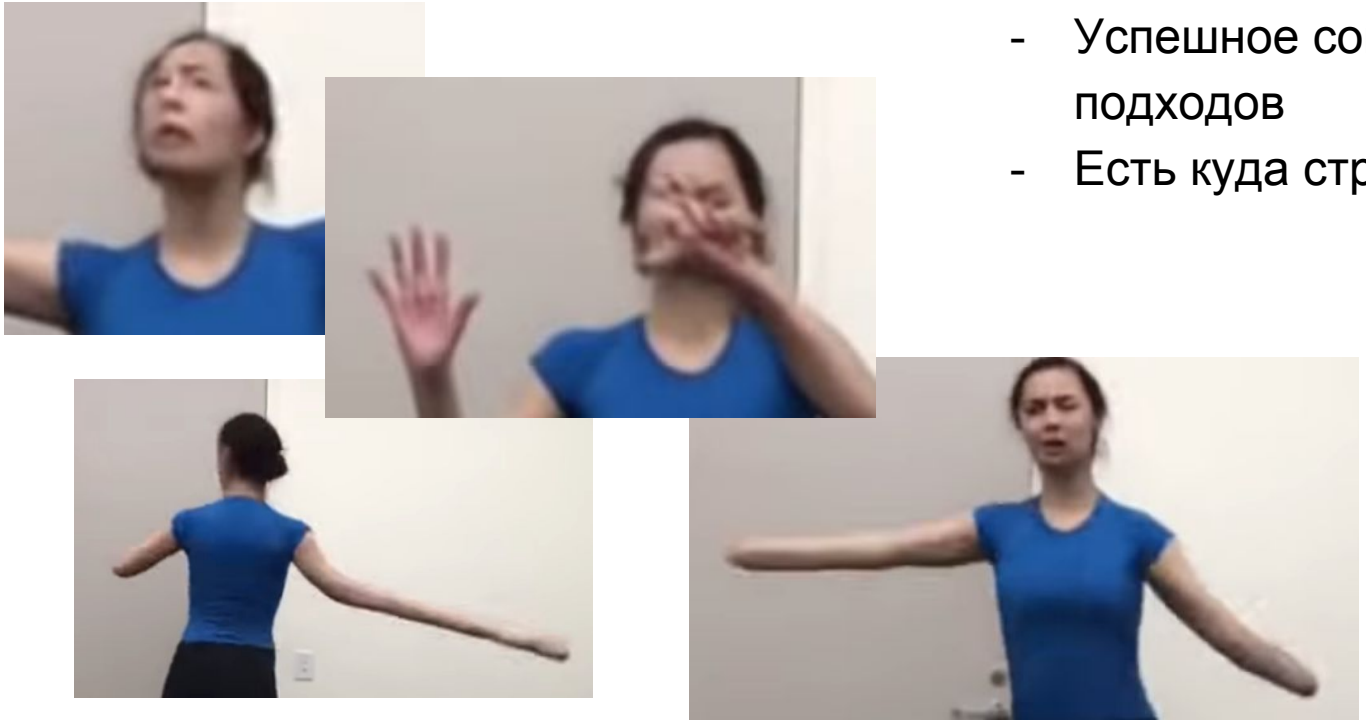
Результаты


Source Video
Detected Pose
Source to Target 1 Result
Source to Target 2 Result

Выводы и перспективы

- Успешное совмещение разных подходов

Выводы и перспективы



- Успешное совмещение разных подходов
- Есть куда стремиться

Выводы и перспективы

- Успешное совмещение разных подходов
- Есть куда стремиться
- Real-time?