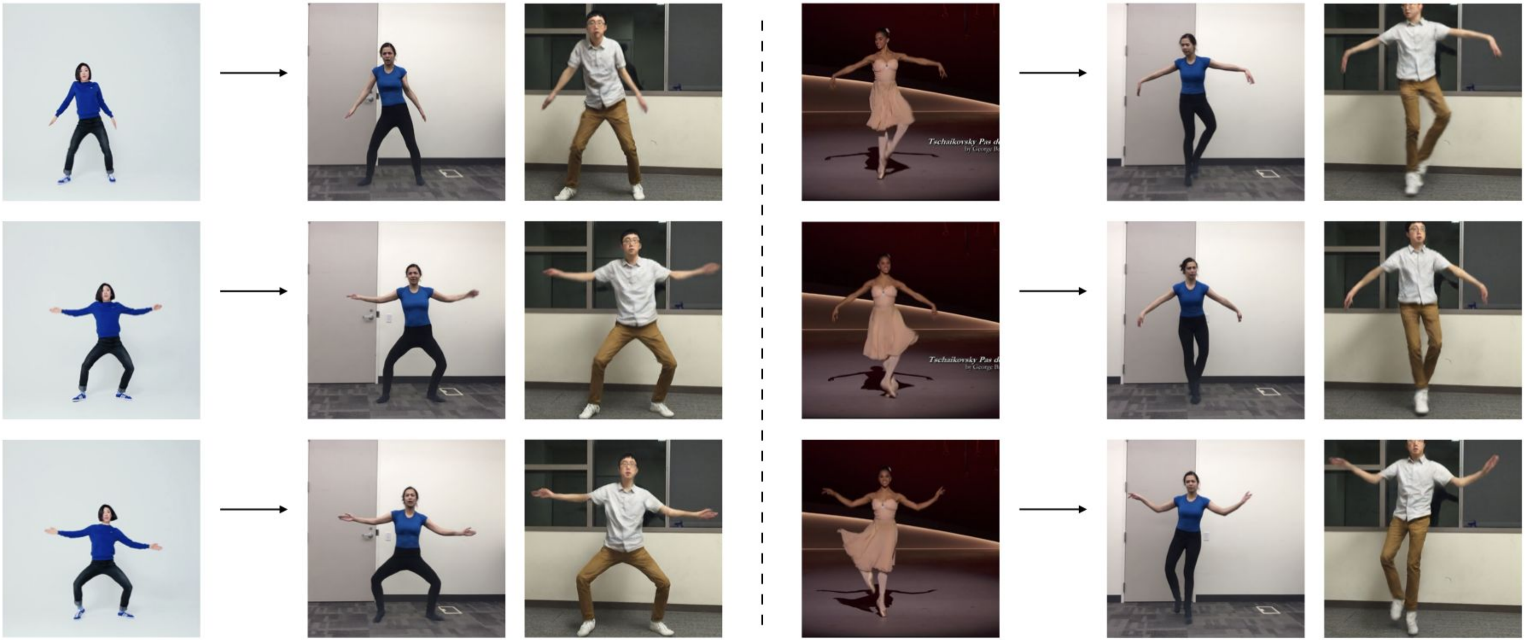
Everybody Dance Now

Caroline Chan∗ Shiry Ginosar Tinghui Zhou† Alexei A. Efros
UC Berkeley

발표자 김연수

Contents



Source Subject | Target Subject 1 | Target Subject 2 | Source Subject | Target Subject 1 | Target Subject 2

Abstract

- This paper presents a simple method for **"do as I do" motion transfer**: **given a source video** of a person dancing, **we can transfer** that performance to a novel (amateur) tar- get after only a few minutes of **the target subject performing standard moves.**
- We approach this problem as **video-to- video translation using pose as an intermediate representation.**
- Approach
  - **extract poses** from the source
  - apply the learned pose-to-appearance **mapping** to generate the target
  - predict two consecutive frames for **temporally coherent** video results
  - separate pipeline for realistic **face synthesis**
- more contribution
  - fake video detection
  - release a dataset

Abstract

- This paper presents a simple method for **"do as I do" motion transfer**: **given a source video** of a person dancing, **we can transfer** that performance to a novel (amateur) tar- get after only a few minutes of **the target subject performing standard moves.**
- We approach this problem as **video-to- video translation using pose as an intermediate representation.**
- Approach
  - **extract poses** from the source
  - apply the learned pose-to-appearance **mapping** to generate the target
  - predict two consecutive frames for **temporally coherent** video results
  - separate pipeline for realistic **face synthesis**
- more contribution
  - fake video detection
  - release a dataset

Our generator and discriminator architectures follow that presented by Wang et al. [41]. The fake-detector architectures matches that of the discriminator with a final fully connected layer.

Introduction

- GOAL

    → corresponds to frames of the ***target*** subject performing the same motions

    → we transfer motion between these **subjects**(source, target) by learning a simple **video-to-video translation.**

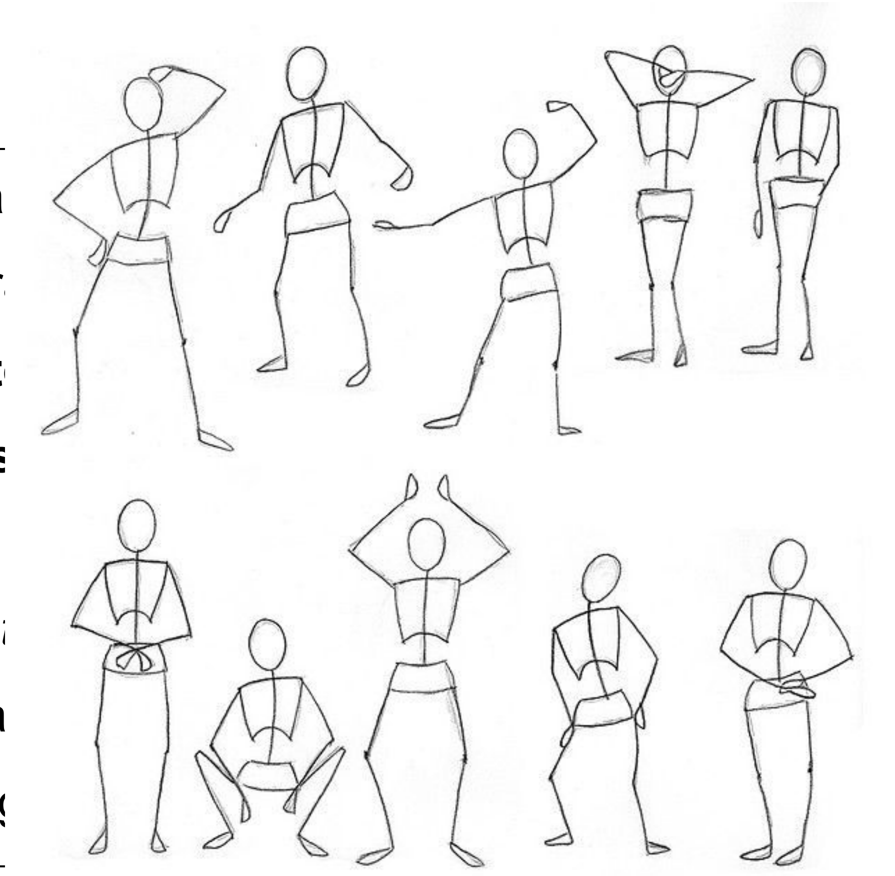
    → to discover an **image-to-image translation** between the source and target **sets. (frame-by-frame manner)**

    BUT, we do not have corresponding pairs of images of the two subjects performing the same motions to supervise learning this translation.

Introduction

- We observe that **keypoint-based pose** preserves motion signatures over time while abstracting away as much subject identity as possible and **can serve as an intermediate representation between any two subjects.**
- We therefore **use pose stick figures** obtained from off-the-shelf human pose detectors, such as **OpenPose**, as an intermediate representation for *frame-to-frame transfer*, as shown in Figure 2.
- We then learn an image-to-image translation model between pose stick figures and images of our target person.

Introduction



- We observe tha… signatures over time while abstr…ossible and **can serve as an int… two subjects.**
- We therefore **us**… e-shelf human pose detectors, such … entation for *frame-to-frame* …
- We then learn a … een pose stick figures and ima…

Introduction

- We ob … s over time w … nd **can serve … cts.**
- We th … an pose detect … frame …
- We th … tick figure…

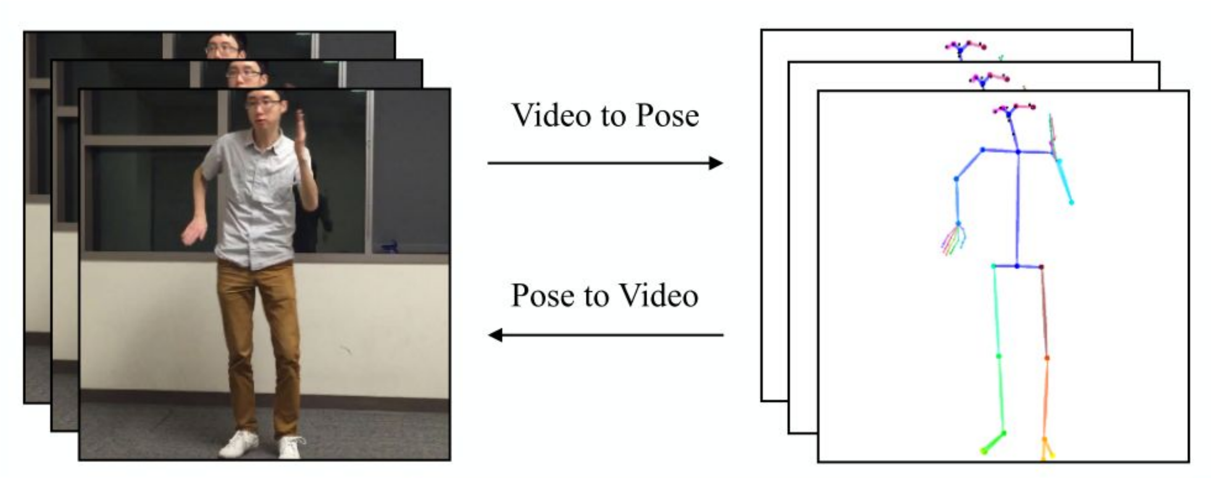


Figure 2: Our method creates correspondences by detecting poses in video frames (Video to Pose) and then learns to generate images of the target subject from the estimated pose (Pose to Video).

Introduction

- We observe that **keypoint-based pose** preserves motion signatures over time while abstracting away as much subject identity as possible and **can serve as an intermediate representation between any two subjects.**
- We therefore use pose stick figures obtained from off-the-shelf human pose detectors, such as OpenPose, as an intermediate representation for frame-to-frame transfer, as shown in Figure 2.
- We then learn an **image-to-image translation model** between pose stick figures and images of our target person.

Related Work - OpenPose, DensePose

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields *

Zhe Cao Tomas Simon Shih-En Wei Yaser Sheikh
The Robotics Institute, Carnegie Mellon University
{zhecao,shihenw}@cmu.edu {tsimon,yaser}@cs.cmu.edu

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

Zhe Cao, *Student Member, IEEE,* Gines Hidalgo, *Student Member, IEEE,*
Tomas Simon, Shih-En Wei, and Yaser Sheikh

Related Work - OpenPose

Pose Estimation

- Top-Down Approach
- **Bottom-up Approach**
  - 한 장의 사진에서 먼저 각각의 관절에 대한 정보를 찾고 이 관절이 어떤 관절과 연결되는지 찾아 하나의 사람으로 만들어 주는 것
  - robustness를 보장해주면서, 많은 수 의 사람들이 등장해도 한 명일때와 다르지 않게 처리가 가능

https://reading-cv-paper.tistory.com/entry/TPAMI-OpenPose-Realtime-Multi-Person-2D-Pose-Estimation-using-Part-Affinity-Fields

Related Work - OpenPose, DensePose

Pose Estimation

- Top-Down Approach
- **Bottom-up Approach**

(a) Input Image (b) Part Confidence Maps (c) Part Affinity Fields (d) Bipartite Matching (e) Parsing Results

https://reading-cv-paper.tistory.com/entry/TPAMI-OpenPose-Realtime-Multi-Person-2D-Pose-Estimation-using-Part-Affinity-Fields

Related Work - Pix2PixHD

High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs

Ting-Chun Wang[1] Ming-Yu Liu[1] Jun-Yan Zhu[2] Andrew Tao[1] Jan Kautz[1] Bryan Catanzaro[1]
[1] NVIDIA Corporation [2] UC Berkeley

https://haawron.tistory.com/26

Method

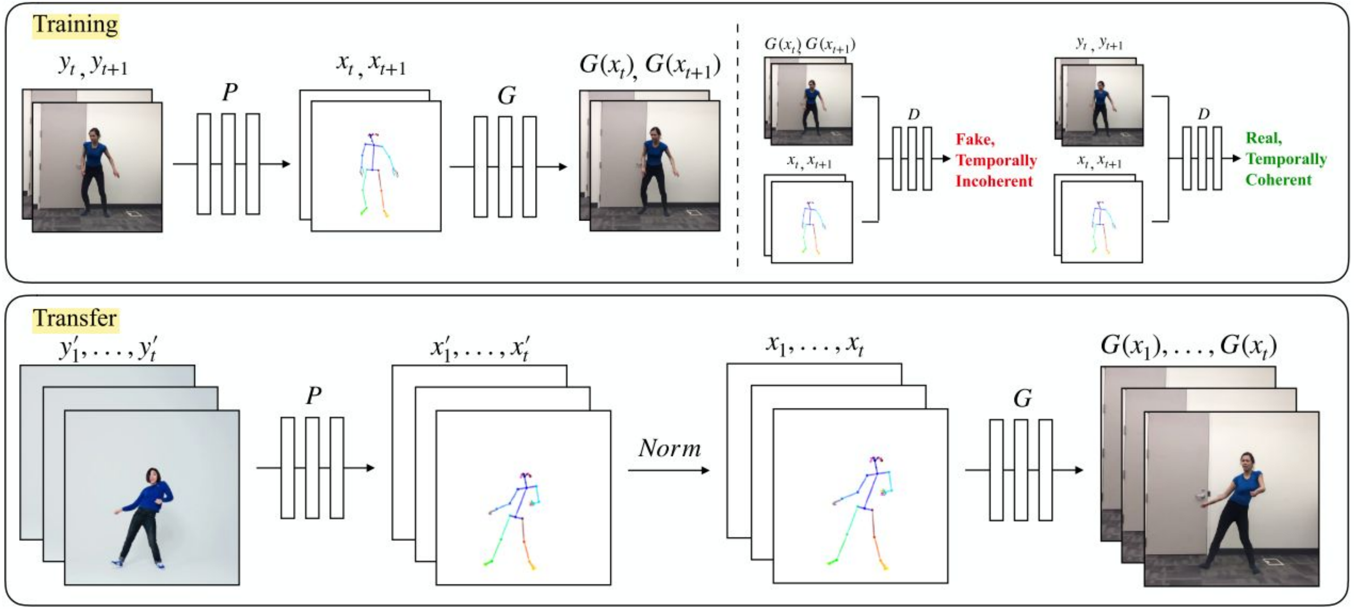


Figure 3: (Top) **Training**: Our model uses a pose detector P to create pose stick figures from video frames of the target subject. We learn the mapping G alongside an adversarial discriminator D which attempts to distinguish between the "real" correspondences $(x_t, x_{t+1}), (y_t, y_{t+1})$ and the "fake" sequence $(x_t, x_{t+1}), (G(x_t), G(x_{t+1}))$. (Bottom) **Transfer**: We use a pose detector P to obtain pose joints for the source person that are transformed by our normalization process $Norm$ into joints for the target person for which pose stick figures are created. Then we apply the trained mapping G.

Method

Pose Encoding and Normalization

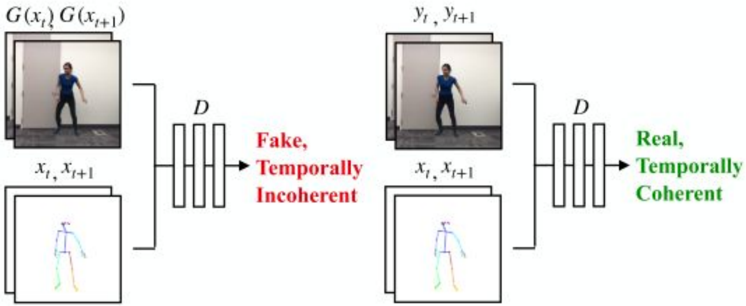
- Encoding body poses : use pose detector ***P*** **(OpenPose)**
  - estimates 2D x,y joint coordinates
- Global pose normalization
  - Why? → subjects may **have different limb proportions**
  - **it may be necessary to transform the pose keypoints of the source** person so that they appear in accordance with the target person's body shape and location as in the **Transfer section**
  - We find this transformation by analyzing **the heights and ankle positions for the poses** of each subject and use a linear mapping between the closest and farthest ankle po- sitions in both videos. After gathering these positions, we calculate **the scale and translation for each frame** based on its corresponding pose detection

Method

Pose to Video Translation

- Our video synthesis method is based off of an **adversarial single frame generation** process
- For our purposes, **G synthesizes** images of a person given a **pose stick figure**
- Such **single-frame image-to-image translation methods**
- Add a learned model of **temporal coherence** as well as a module for **high resolution face generation**

Method



Pose to Video Translation

- Temporal Smoothing
  - the discriminator is now tasked with determining both the difference in realism and temporal coherence between the “fake” sequence (x_{t−1}, x_t, G(x_{t−1}), G(x_t)) and “real” sequence (xt−1, xt, yt−1, yt).
  - $$\mathcal{L}_{\text{smooth}}(G, D) = \mathbb{E}_{(x,y)}[\log D(x_t, x_{t+1}, y_t, y_{t+1})] + \mathbb{E}_x[\log(1 - D(x_t, x_{t+1}, G(x_t), G(x_{t+1}))] \qquad (1)$$
- Face GAN
  - We add a specialized GAN setup to **add more detail and realism to the face region**
  - $$\mathcal{L}_{\text{face}}(G_f, D_f) = \mathbb{E}_{(x_F, y_F)}[\log D_f(x_F, y_F)] + \mathbb{E}_{x_F}[\log(1 - D_f(x_F, G(x)_F + r))].$$

Method

Full Objective (To train)

- We employ training in stages where the full image GAN is optimized separately from the specialized face GAN
-

$$\min_G((\max_{D_i}\sum_{k_i}\mathcal{L}_{\text{smooth}}(G,D_k))+\lambda_{FM}\sum_{k_i}\mathcal{L}_{\text{FM}}(G,D_k)$$
$$+\lambda_P(\mathcal{L}_P(G(x_{t-1}),y_{t-1})+\mathcal{L}_P(G(x_t),y_t))) \quad (3)$$

ting. We follow the progressive learning schedule from pix2pixHD and learn to synthesize at 512×256 at the first (global) stage, and then upsample to 1024×512 at the second (local) stage. For predicting face residuals, we use the global generator of pix2pixHD and a single 70×70 Patch-GAN discriminator [16]. We set hyperparameters $\lambda_P = 5$ and $\lambda_{VGG} = 10$ during the global and local training stages respectively. For the dataset collected in Section 4.1, we trained the global stage for 5 epochs, the local stage for 30 epochs, and the face GAN for 5 epochs.

Method

[Transfer] we divide our pipeline into three(3) stages

1. **pose detection** → OpenPose
2. global pose normalization → accounts for differences between the source and target **body shapes and locations within the frame**
3. target person의 합성된 video 생성 → mapping from normalized pose stick figures to the target subject

Experiments

Ablation conditions

- Frame-by-frame synthesis (FBF)
- Temporal smoothing (FBF+TS)
- Our model (FBF+TS+FG)

Evaluation metrics

- SSIM. Struc- tural Similarity
- LPIPS Learned Perceptual Im- age Patch Similarity

Experiments

Region	Metric	FBF	FBF+TS	FBF+TS+FG
Face	SSIM	0.784	0.811	**0.816**
	LPIPS	0.045	0.039	**0.036**
Body	SSIM	0.828	**0.838**	**0.838**
	LPIPS	0.057	0.051	**0.050**

(a) Metric comparison for synthesized face (top) and full-body (bottom) regions. Metrics are averaged over the 5 subjects. For SSIM higher is better. For LPIPS lower is better.

Condition	1	2	3	4	5	Total
FBF	54.1%	69.7%	62.4%	53.8%	60.0%	58.8%
FBF+TS	59.6%	56.4%	50.3%	53.0%	53.1%	53.9%

(b) Perceptual study results for subjects 1 through 5 and in total average. We report the percentage of time participants chose **our** method as more realistic than the ablated conditions.

Table 3: Ablation studies. We compare frame-by-frame synthesis (FBF), adding temporal smoothing (FBF+TS) and our final model with temporal smoothing and Face GAN modules (FBF+TS+FG).

Condition	1	2	3	4	5	Total
Prefer FBF+TS	60.5%	62%	57.5%	50%	62.5%	58.5%

Table 4: Comparison of our method without Face GAN (FBF+TS) to the FBF ablation for subjects 1 through 5 and in total average. We report the percentage of time participants chose the FBF+TS ablation over the FBF ablation.

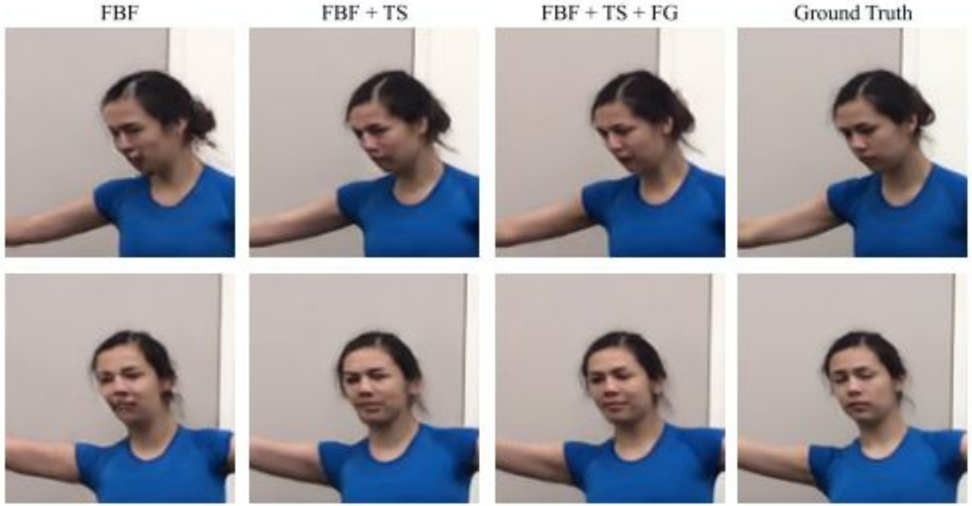


Figure 6: Face image comparison on held-out data. We compare frame-by-frame synthesis (FBF), adding temporal smoothing (FBF+TS) and our full model (FBF+TS+FG).

Limitations and Discussion

- Further work could focus on improving results **by combining target videos with different clothing or scene lighting**, improving pose detection sys- tems, and **mitigating the artifacts** caused by high frequency textures in loose/wrinkled clothing or hair.
- Future work could focus on the **train- ing data**, i.e. what poses and how many are needed to learn a effective model. This area relates to work on understanding which training examples are most influential