

Landmark Guided Video Generation

Sijie Song

2018.07.09



Outline

- Every Smile is Unique: Landmark-Guided Diverse Smile Generation
- Skeleton-Aided Articulated Motion Generation

Every Smile is Unique: Landmark-Guided Diverse Smile Generation

CVPR 2018 Poster

Wei Wang^{1,4}, Xavier Alameda-Pineda², Dan Xu¹,
Pascal Fua⁴, Elisa Ricci^{1,3}, and Nicu Sebe¹

¹University of Trento, Italy

²Inria, Grenoble-Alpes University, France

³Fondazione Bruno Kessler (FBK), Italy

⁴Computer Vision Laboratory, EPFL, Lausanne, Switzerland

One-to-Many Video Generation

Neural Face
+
Given Label
(posed/spontaneous smile)

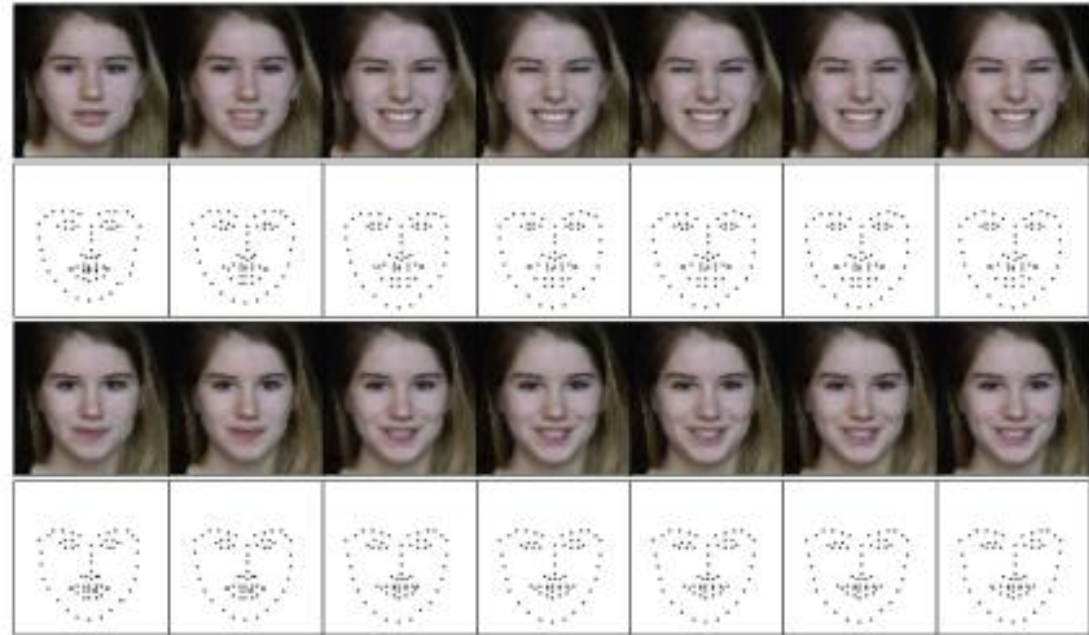



Figure 1. Two different sequences of spontaneous smiles and associated landmarks. While there is a common average pattern, the changes from one sequence to another are clearly visible.

Overview

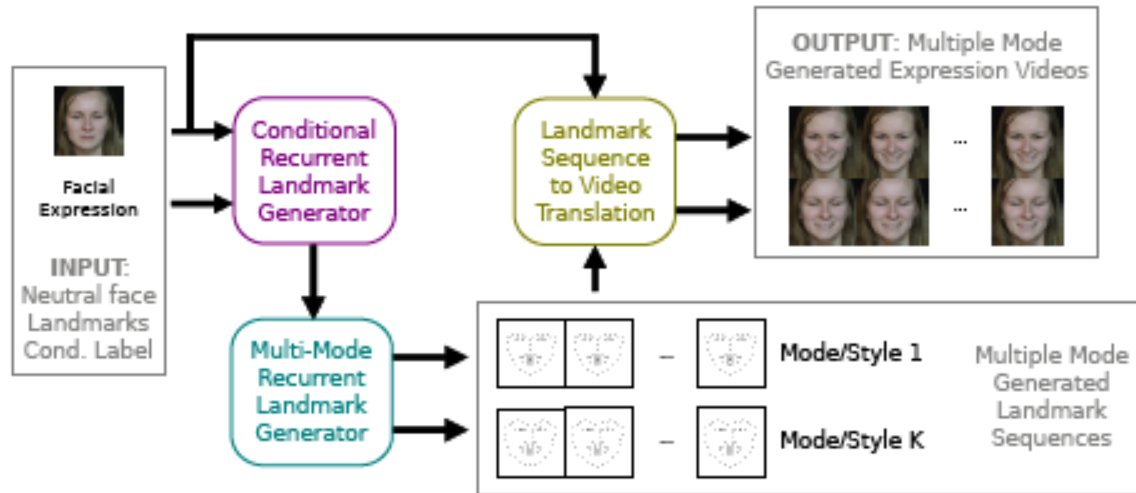


Figure 2. Overview of the proposed framework. The input image is used together with the conditioning label to generate a set of K distinct landmark sequences. These landmark sequences guide the neutral face image to translate into face videos.

- **Conditional Recurrent Landmark Generator**
Image + Label \rightarrow Embedding Sequence
- **Multi-Mode Recurrent Landmark Generator**
Embedding Sequence \rightarrow Landmark Sequence
- **Landmark Sequence to Video Translation**
Landmark Sequence \rightarrow Videos

Conditional Recurrent Landmark Generator

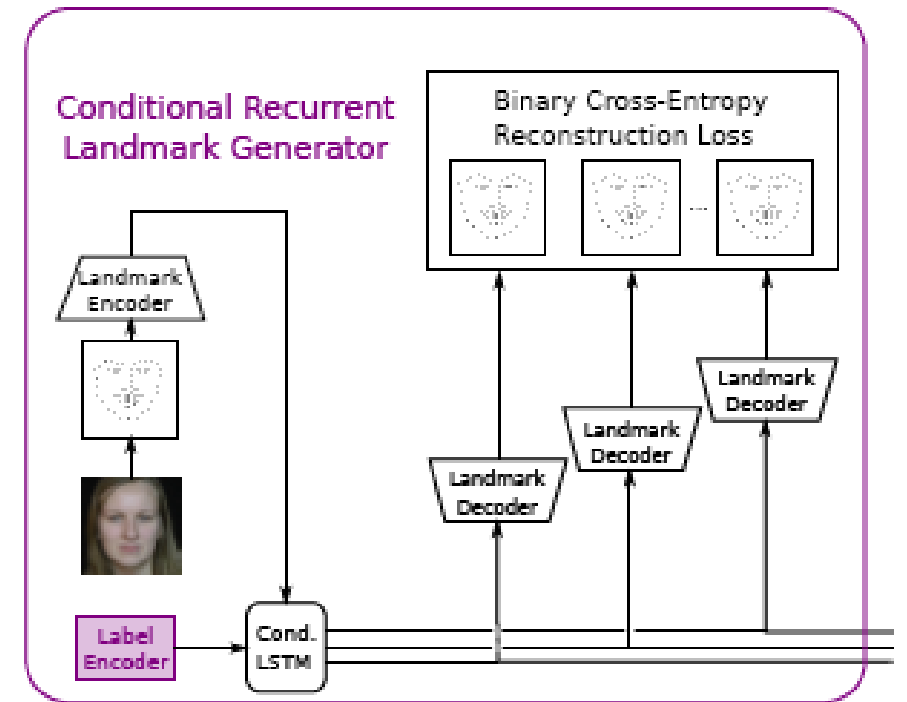
- Face Image \rightarrow **VAE** \rightarrow Compact Embedding h_0
- $h_0 + \text{Label} \rightarrow$ **Cond. LSTM** \rightarrow Face Landmark Embedding Sequence $\mathbf{h} = (h_1, \dots, h_T)$
- $\mathbf{h} = (h_1, \dots, h_T) \rightarrow$ **Decoder**
 \rightarrow Landmark Sequence $\mathbf{x} = (x_1, \dots, x_T)$

- **Training:**

Binary Cross-Entropy Loss (BCE) with training sequences

$$\{\mathbf{y}^n = (y_1^n, \dots, y_T^n)\}_{n=1}^N$$

$$\mathcal{L}_{\text{BCE}} = \sum_{n,t=1}^{N,T} y_t^n \odot \log x_t^n + (1 - y_t^n) \odot \log(1 - x_t^n)$$



Multi-Mode Recurrent Landmark Generator

- $\mathbf{h} = (h_1, \dots, h_T) \rightarrow \mathbf{K \text{ LSTMs}} \rightarrow \{\mathbf{h}_k = (h_{1k}, \dots, h_{Tk})\}_{k=1}^K$.
- **Training:**

Push & Pull loss

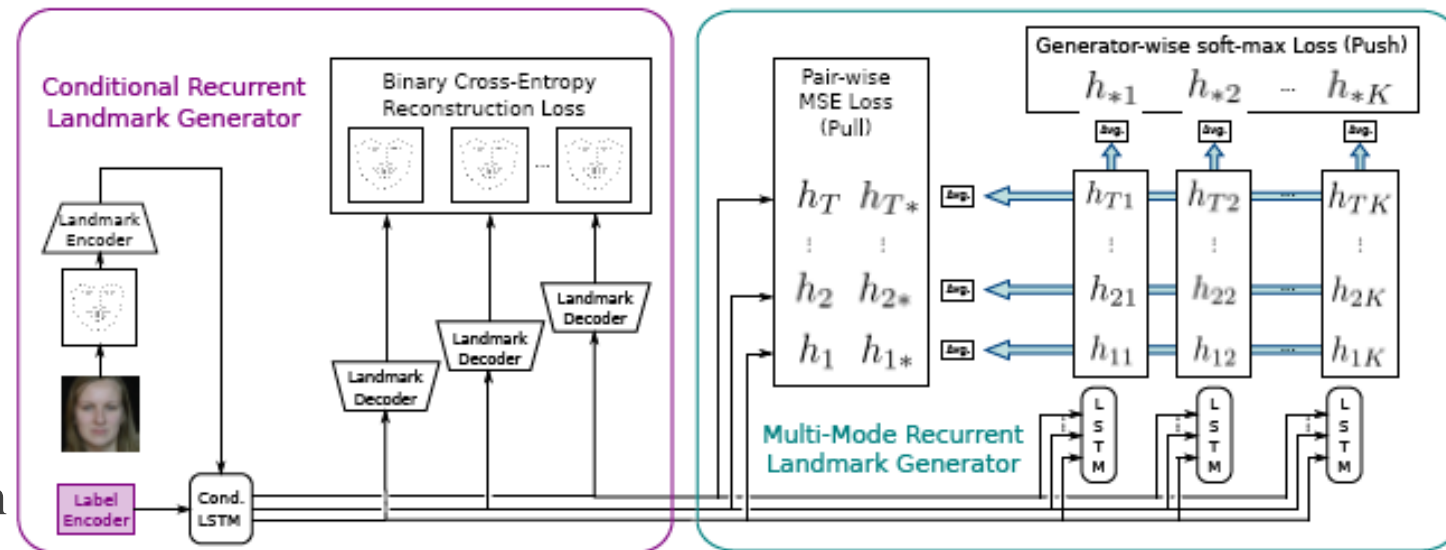
- Push the sequences to be distinct

$$\mathcal{L}_{\text{Push}} = - \sum_{n,k=1}^{N,K} \log \phi_k(h_{*k}^n)$$

ϕ_k : FC + softmax

- Pull them towards a common pattern

$$\mathcal{L}_{\text{Pull}} = \sum_{n,t=1}^{N,T} \|h_t^n - h_{t*}^n\|_2$$



Landmark Sequence to Video Translation

- U-Net Structure with

Face landmark images $\{\mathbf{y}^n = (y_1^n, \dots, y_T^n)\}_{n=1}^N$

Face images $\{\mathbf{z}^n = (z_1^n, \dots, z_T^n)\}_{n=1}^N$

- **Training:**

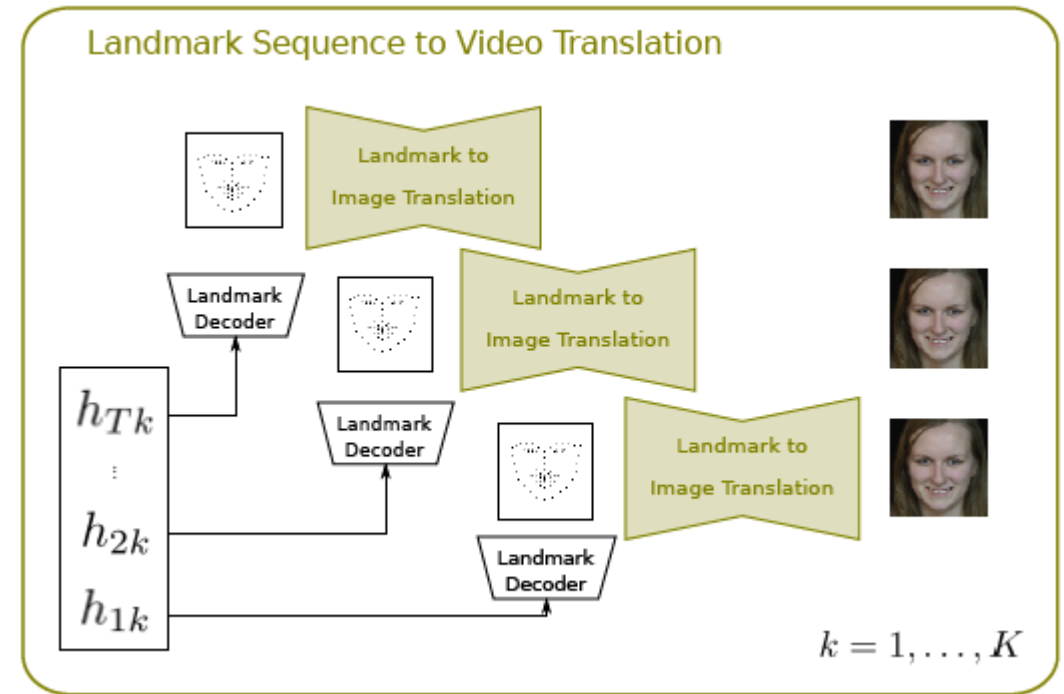
Reconstruction loss

$$\mathcal{L}_{\text{Rec}} = \sum_{n,t=1}^{N,T} \|z_t^n - w_t^n(\theta_{\mathcal{G}})\|_1$$

$$w_t^n(\theta_{\mathcal{G}}) = \mathcal{G}(y_t^n, z_0^n; \theta_{\mathcal{G}})$$

Adversarial loss

$$\mathcal{L}_{\text{Adv}} = \sum_{n,t=1}^{N,T} \log \mathcal{D}([z_0^n, z_t^n]; \theta_{\mathcal{D}}) + \sum_{n,t=1}^{N,T} \log(1 - \mathcal{D}([z_0^n, w_t^n(\theta_{\mathcal{G}})]; \theta_{\mathcal{D}}))$$



Training Strategy

- Three Training Phases:

First, train VAE

Second, finetune VAE + train Cond. LSTM

Third, finetune VAE + finetune Cond. LSTM + train K LSTMs

- Video Translation Module: training apart

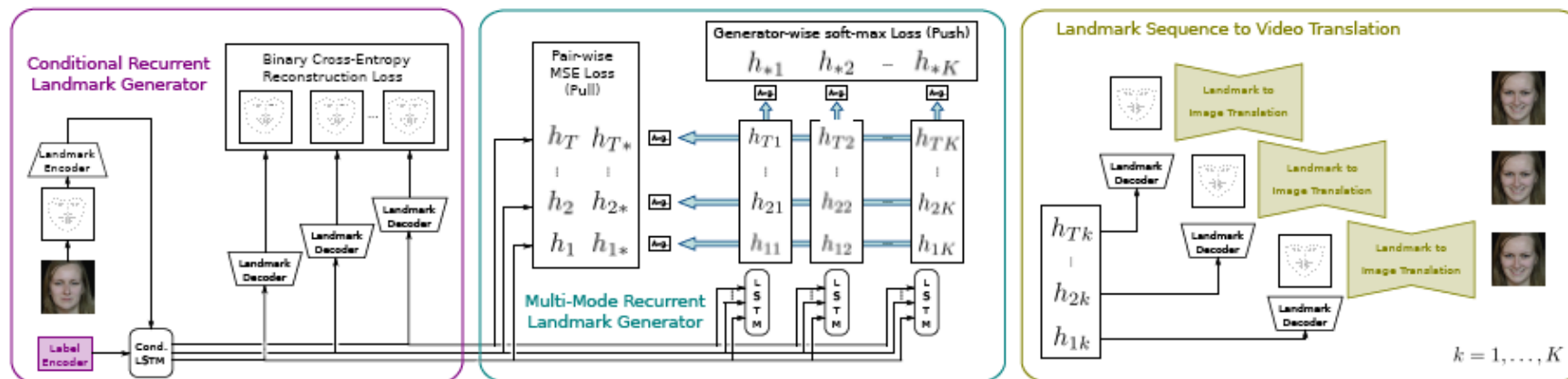


Figure 3. Detail of the conditional multi-mode recurrent network. The left block (magenta) encodes the landmark image and generates a sequence of landmark embeddings according to the conditioning label. The second block (turquoise) generates K different landmark embedding sequences. Finally, the third block (ocher) translates each of the sequences into a face video.

Experiments

- Datasets: $T = 32, 64 \times 64$
 - UvA-NEMO Smile
 - DISFA
 - DISFA+
- Preprocess
 - Utilize AU (action unit) to select smile patterns

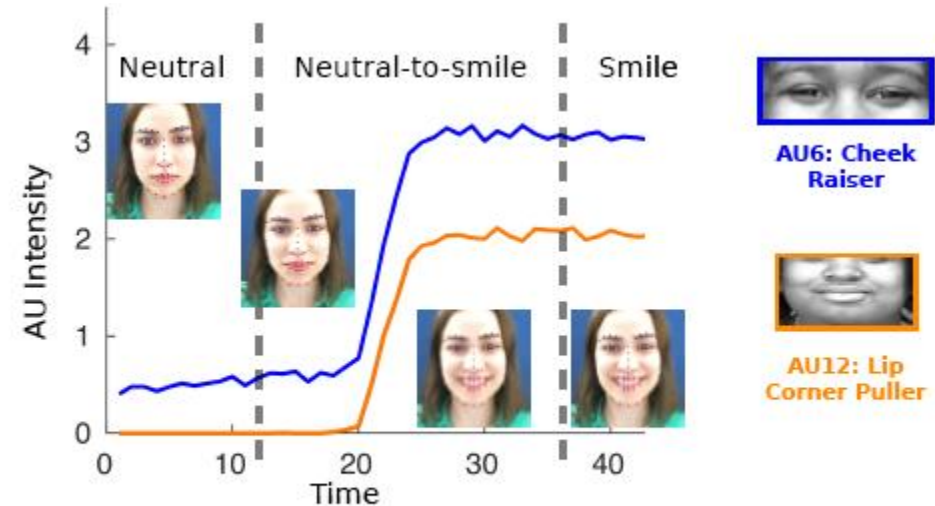


Figure 4. Action unit dynamics in neutral-to-smile transitions: *cheek raiser* and *lip corner puller*.

Experiments

- Qualitative Evaluation (Cond. LSTM)



Figure 5. Landmark sequences generated with the first block of our CMM-Net. The associated face images are obtained using the landmark sequence to video translation block. The left block corresponds to generated spontaneous smiles, while the right block to posed smiles. The three row pairs correspond to the UvA-NEMO, DISFA & DISFA+ datasets respectively. Images better seen at magnification.

Experiments

- Qualitative Evaluation (Multi-Mode, $K = 3$)

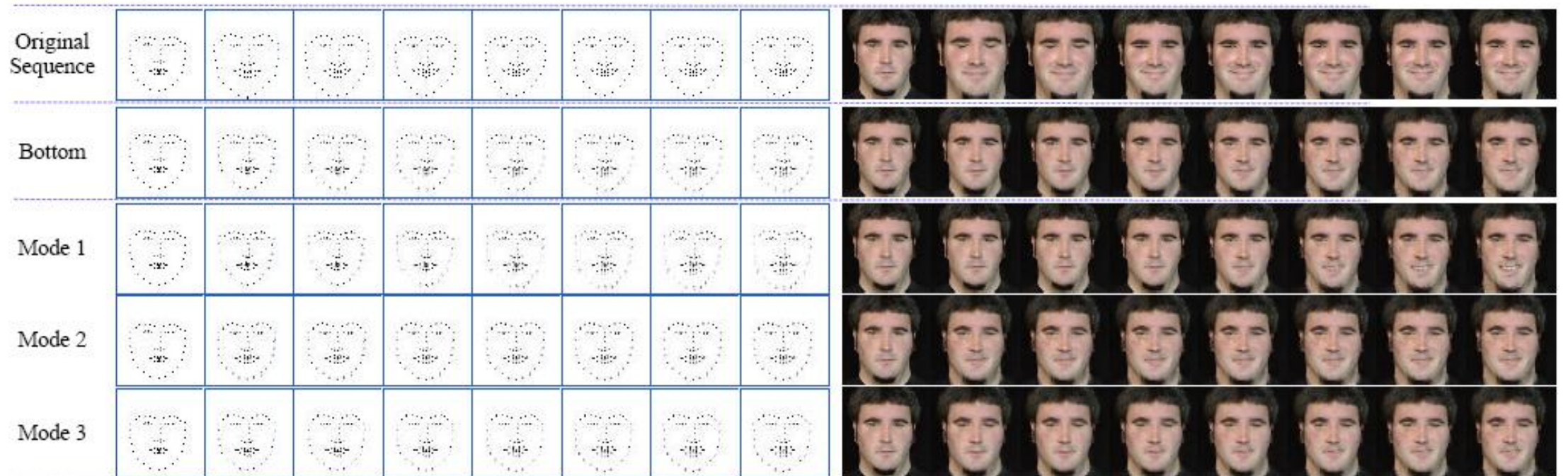


Figure 6. Multi-mode generation example with a sequence of the UvA-NEMO dataset: landmarks (left) and associated face images (right). The rows correspond to the original sequence, output of the Conditional LSTM, and output of the Multi-Mode LSTM (last three rows).

Experiments

- Qualitative Evaluation (State-of-the-Art)



Figure 7. Qualitative comparison. From top to bottom: original sequence, Video-GAN, CRA-Net and CMM-Net. Video-GAN introduces many artifacts compared to the other two. CRA-Net learn the smile dynamics, but fail to preserve the identity, as opposed to CMM-Net which produces realistic smiling image sequences.

Experiments

- Quantitative Analysis

Table 1. Quantitative Analysis. The SSIM and Inception Score.

	UvA-NEMO Spont.			UvA-NEMO Posed			DISFA Spont.			DISFA+ Posed		
Model	IS	Δ IS	SSIM	IS	Δ IS	SSIM	IS	Δ IS	SSIM	IS	Δ IS	SSIM
Original	1.419	-	-	1.437	-	-	1.426	-	-	1.595	-	-
Video GAN	1.576	0.157	0.466	1.499	0.062	0.450	1.777	0.351	0.243	1.547	0.048	0.434
CRA-Net	1.311	0.108	0.553	1.310	0.127	0.471	1.833	0.407	0.749	1.534	0.061	0.839
CMM-Net	1.354	0.065	0.854	1.435	0.002	0.827	1.447	0.021	0.747	1.533	0.062	0.810

Table 2. CMM-Net vs Video-GAN and CMM-Net vs CRA-Net: percentage (%) of the preferences of the generated videos.

Models	Spontaneous Smile	Posed Smile
Video-GAN [41]	10.14	7.24
CMM-Net	85.14	83.68
~	4.72	9.08
CRA-Net	17.76	11.94
CMM-Net	54.87	59.72
~	27.37	28.33

Experiments

- Analyzing the Dynamics of AUs

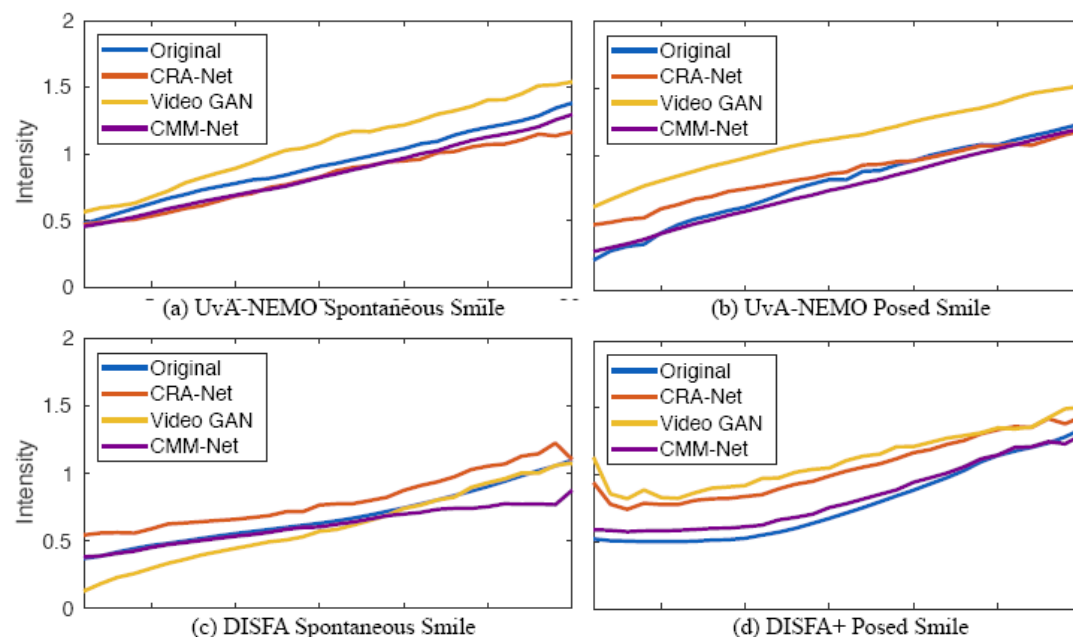


Figure 8. Dynamics of the action units in N2S sequences.

Table 3. Distance between the AU curves of different methods and those of the original sequences.

Model	UvA-NEMO Spont.	UvA-NEMO Posed	DISFA	DISFA+
Video GAN	2.976	2.618	3.775	7.979
CRA-Net	4.452	9.783	2.400	9.931
CMM-Net	2.234	1.472	2.035	1.812

Every Smile is Unique

- Distinct Smile Generation
 - Conditional Recurrent Landmark Generator
 - Multi-Mode Recurrent Landmark Generator
 - Landmark Sequence to Video Translation

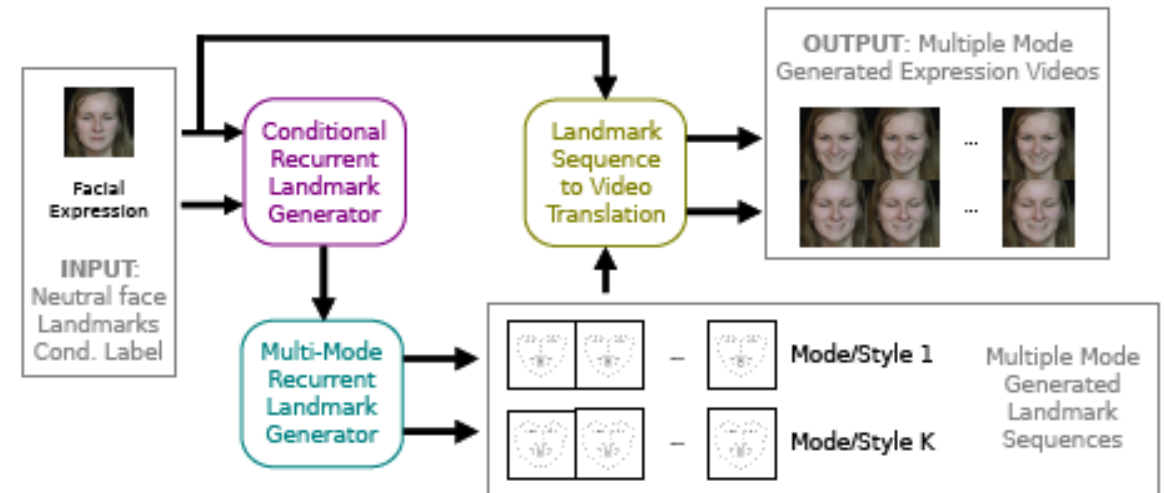


Figure 2. Overview of the proposed framework. The input image is used together with the conditioning label to generate a set of K distinct landmark sequences. These landmark sequences guide the neutral face image to translate into face videos.

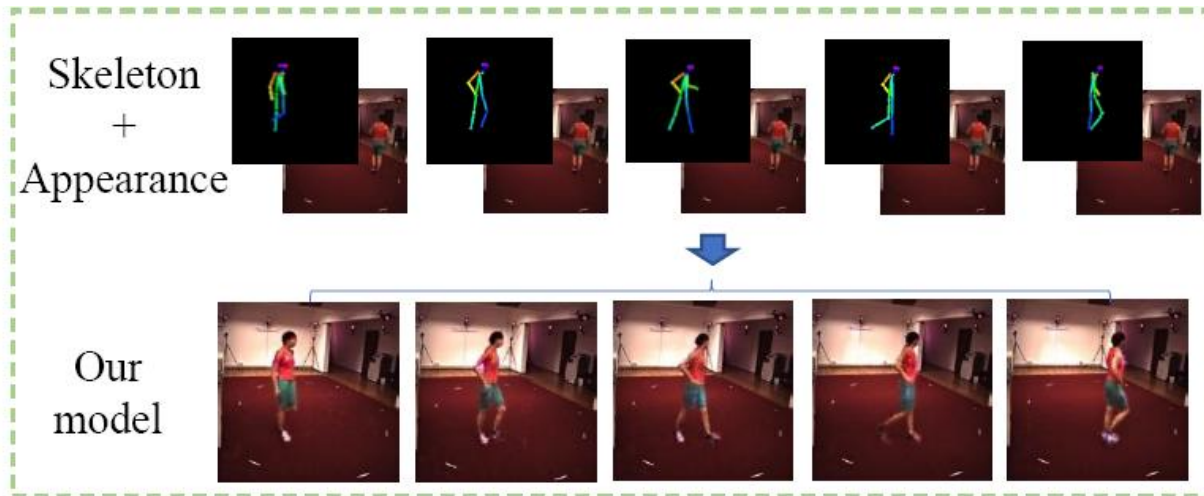
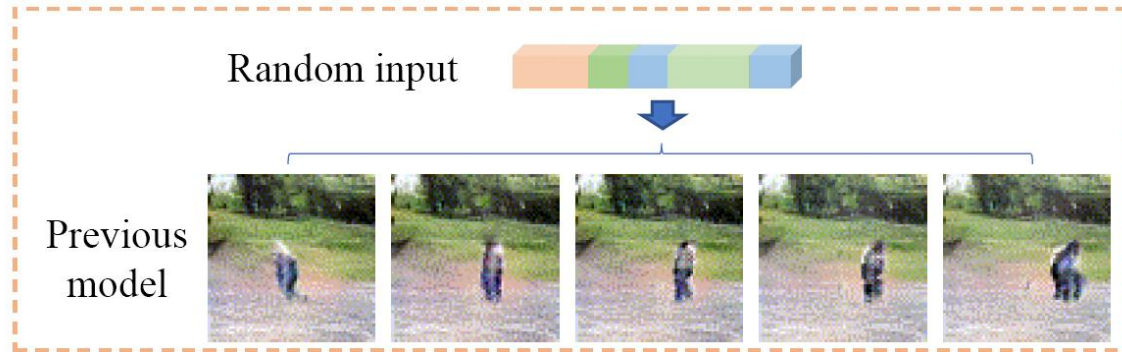
Skeleton-aided Articulated Motion Generation

MM 2017

Yichao Yan, Jingwei Xu, Bingbing Ni, Xiaokang Yang

Shanghai Jiaotong University

Overview



Skeleton Conditional GAN

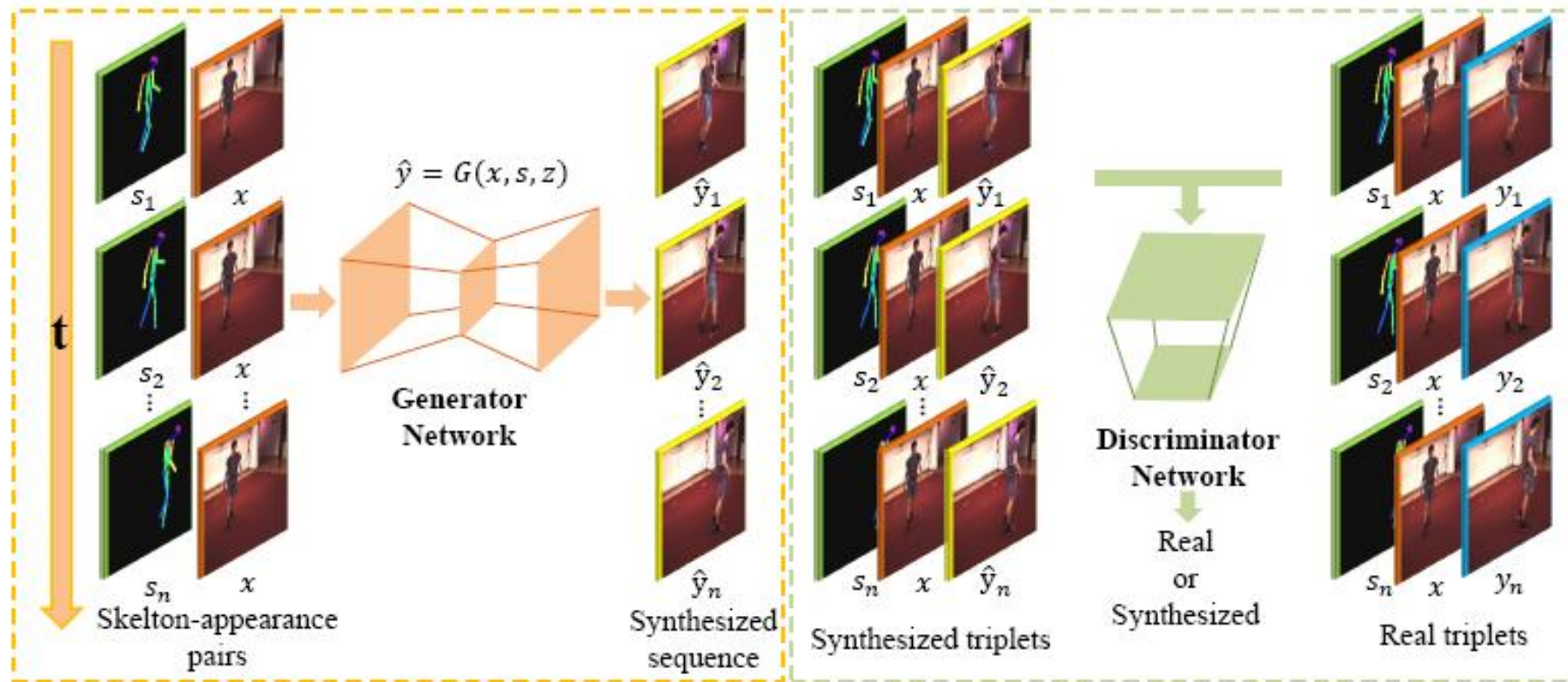


Fig. 2. Architecture of the generation and discrimination network. The inputs for the generator are the skeleton-appearance pairs (s, x) and generate the synthesized sequence $\hat{Y} = \{\hat{y}_1, \dots, \hat{y}_n\}$. The discriminator D tries to distinguish real triplets (x, s, y) and synthesized triplets (x, s, \hat{y}) .

Skeleton Conditional GAN

- Reference image x , skeleton s , ground truth y
- Training:** $\mathcal{L}(G, D) = \mathcal{L}_c(G, D) + \lambda \mathcal{L}_{L_1}(G) + \beta \mathcal{L}_{tri}(G)$

Adversarial Loss

$$\mathcal{L}_c(G, D) = \mathbb{E}_{x, s, y \sim p_{data}(x, s, y)} [\log D(x, s, y)] + \mathbb{E}_{x, s \sim p_{data}(x, s), z \sim p_z(z)} [\log(1 - D(x, s, G(x, s, z)))]$$

L_1 Loss

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{x, s \sim p_{data}(x, s), z \sim p_z(z)} [\|y - G(x, s, z)\|_1]$$

Triplet Loss

$$\mathcal{L}_{tri}(G) = \sum_{i=1}^m [\|t_i^a - t_i^p\|_2^2 - \|t_i^a - t_i^n\|_2^2 + \alpha]_+$$

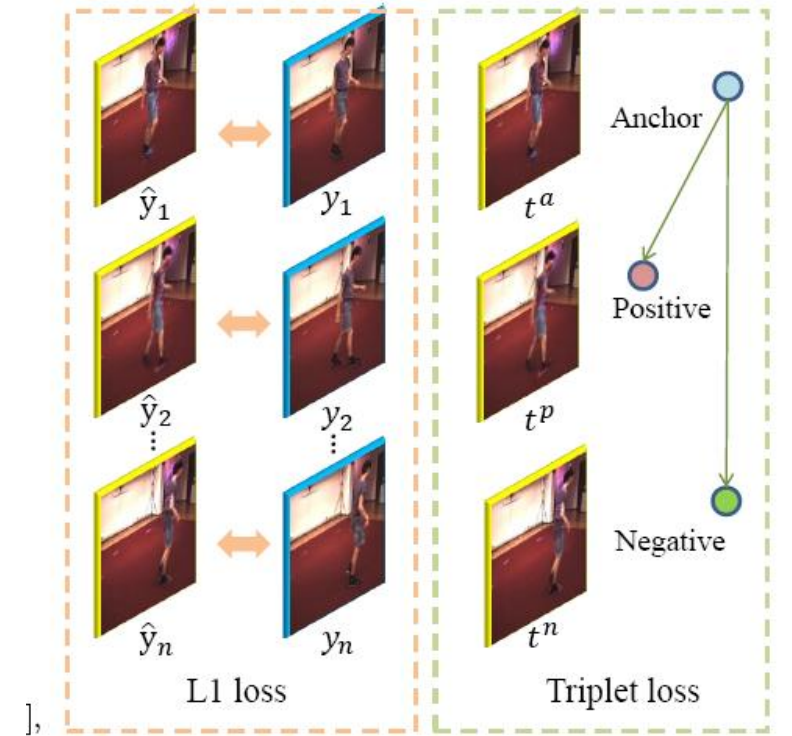
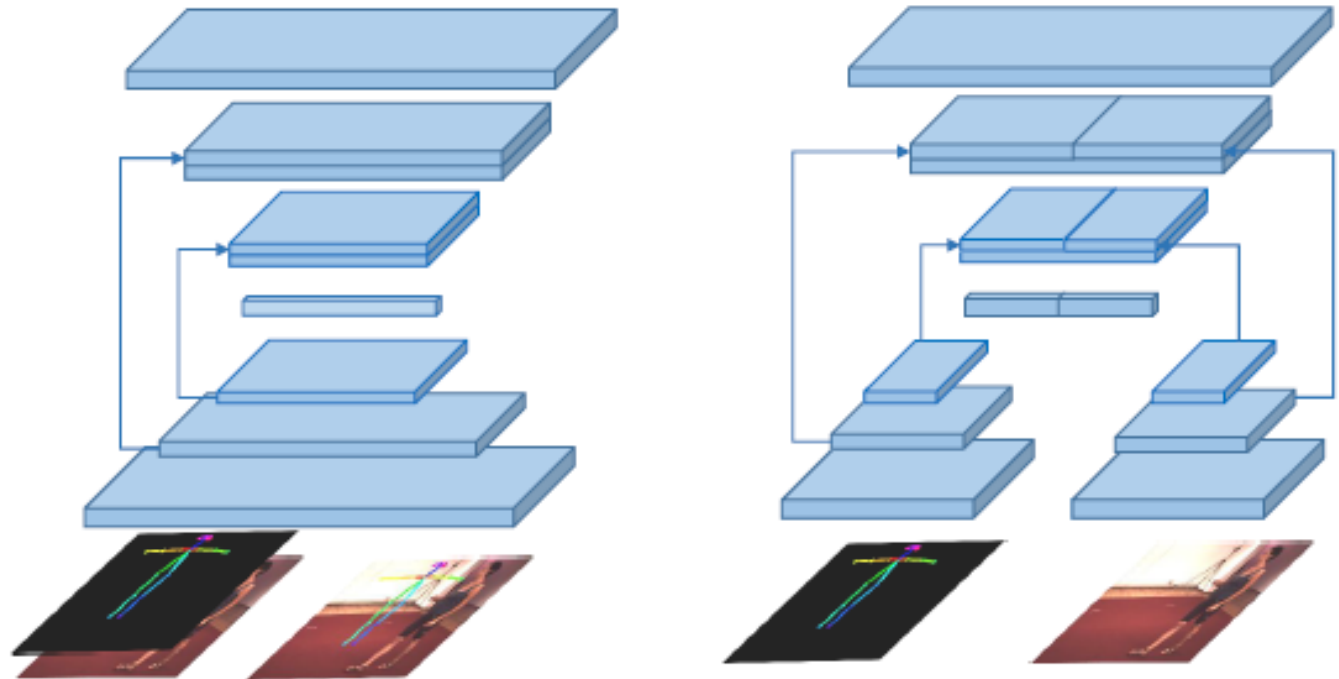


Fig. 3. Loss terms of the model. Despite of the GAN loss, we take two additional loss terms, i.e., the L1 loss to enhance the image-to-image translation quality, and the triplet loss to guarantee the continuity of the generated motion sequence.

Generator Architecture

- Stacked Generator
 - Stacked Generator 1
Concat skeleton & RGB
 - Stacked Generator 2
Draw skeleton on RGB
- Siamese Generator



(a) Stacked Generator.

(b) Siamese Generator.

Fig. 4. Structures of the generator. For both the stacked generator and the Siamese generator, we take the U-Net structure for both generators.

Detailed Architecture

- Generator

TABLE I
DETAILED STRUCTURE OF THE GENERATOR.

Encoder	
Layer	Input size: $256 \times 256 \times 6(3)$
1	Conv-(64,K4 \times 4,S2), lReLU-(0.2)
2	Conv-(128,K4 \times 4,S2), BN, lReLU-(0.2)
3	Conv-(256,K4 \times 4,S2), BN, lReLU-(0.2)
4-8	Conv-(512,K4 \times 4,S2), BN, lReLU-(0.2)
Decoder	
Layer	Input size: $60 \times 60 \times 1(2)$
1	FConv-(512,K4 \times 4,S2), BN, Dropout-(0.5), ReLU
2-5	FConv-(1024,K4 \times 4,S2), BN, Dropout-(0.5), ReLU
6	FConv-(512,K4 \times 4,S2), BN, Dropout-(0.5), ReLU
7	FConv-(256,K4 \times 4,S2), BN, Dropout-(0.5), ReLU
8	FConv-(128,K4 \times 4,S2), BN, Dropout-(0.5), ReLU

- Discriminator

TABLE II
DETAILED STRUCTURE OF THE DISCRIMINATOR.

Discriminator	
Layer	Input size: $256 \times 256 \times 9$
1	Conv-(64,K4 \times 4,S2), lReLU-(0.2)
2	Conv-(128,K4 \times 4,S2), BN, lReLU-(0.2)
3	Conv-(256,K4 \times 4,S2), BN, lReLU-(0.2)
4-6	Conv-(512,K4 \times 4,S2), BN, lReLU-(0.2)

Experiments

- Datasets
 - KTH
 - Human3.6M
- Structure Analysis
 - Stacked generator: Lose identity
 - Siamese generator: Keep identity
 - Disentangle pose and appearance

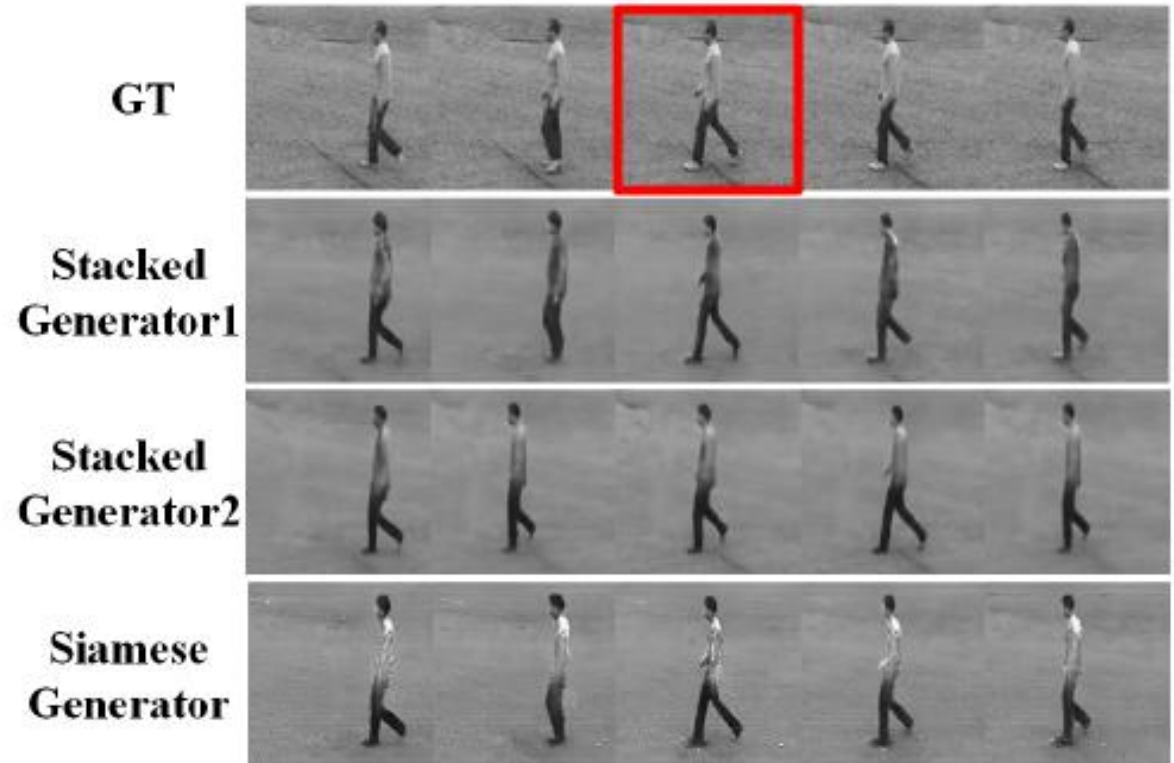
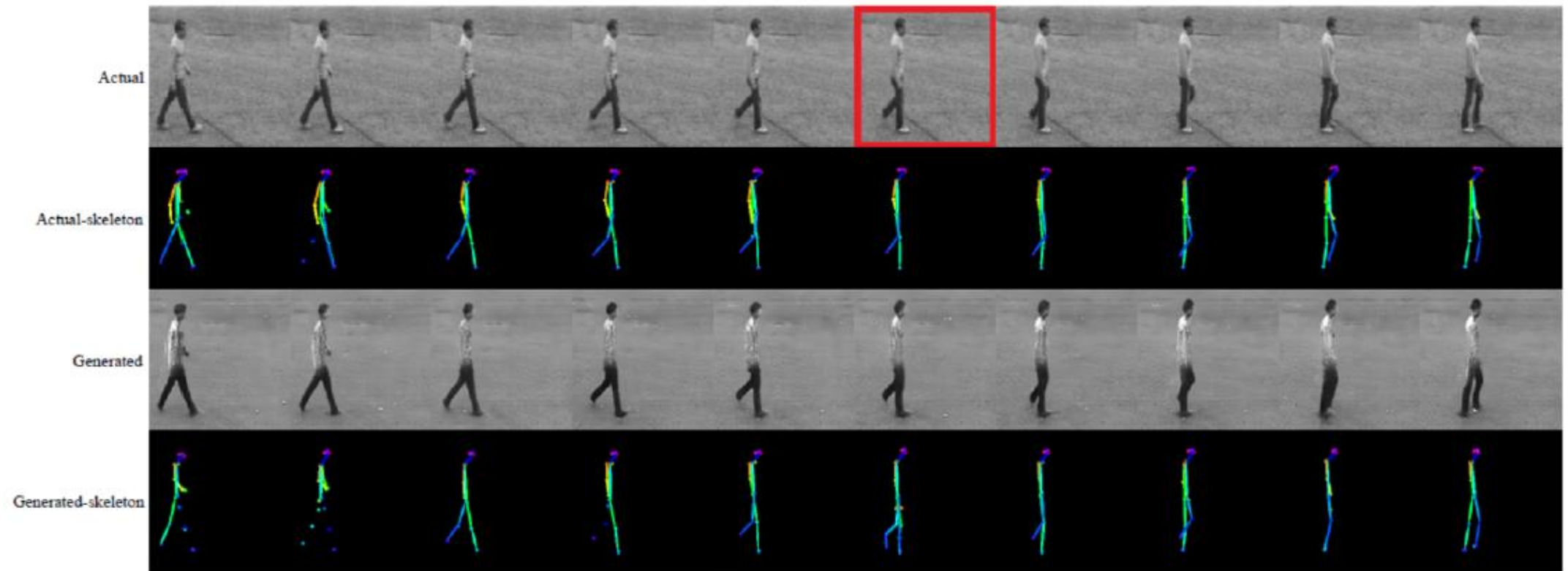


Fig. 5. Examples generated by different generator structure. The first row contains the ground truth motion sequence, the image with red bounding box is the appearance reference image.

Experiments

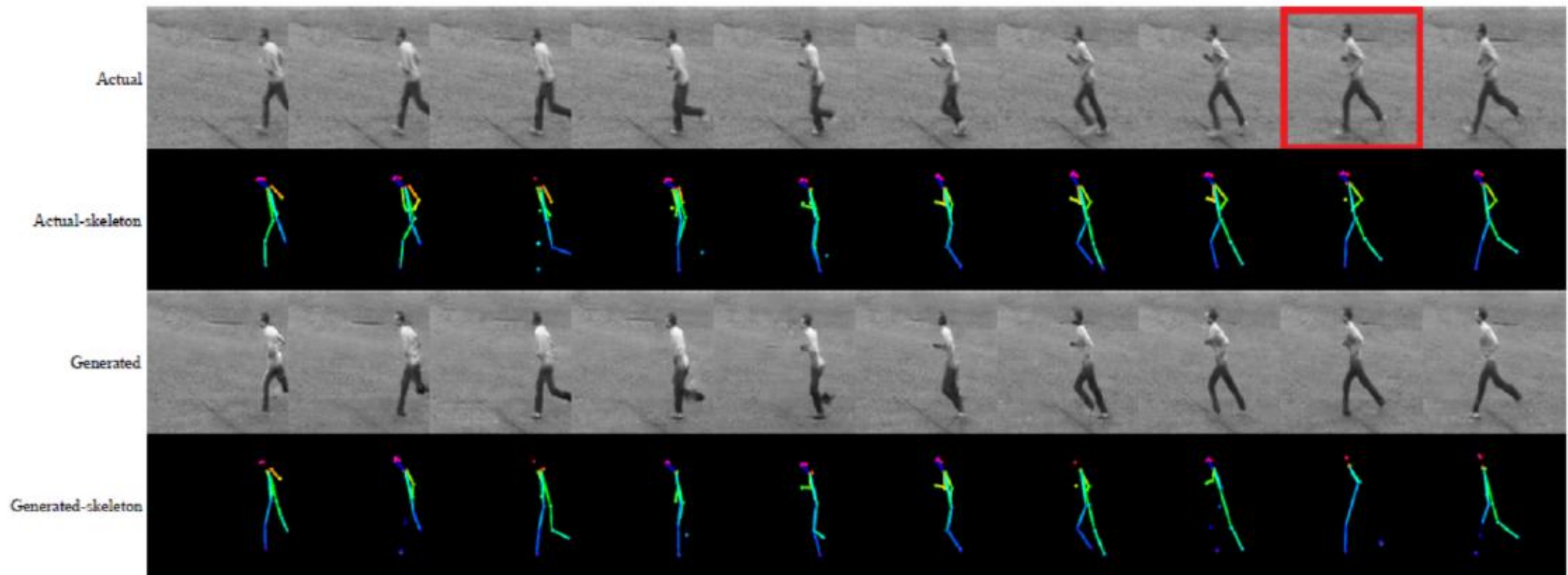
- Motion Generation (KTH)



(a) Walking.

Experiments

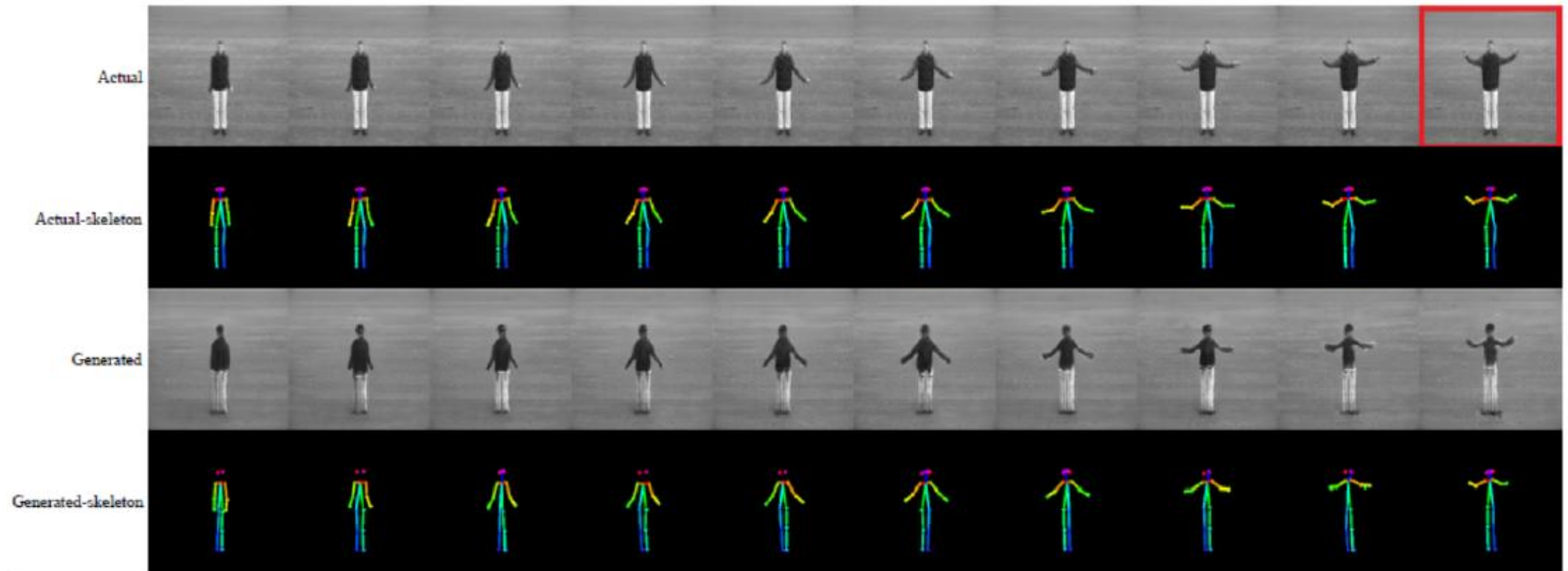
- Motion Generation (KTH)



(b) Running.

Experiments

- Motion Generation (KTH)



(c) Hand waving.

Experiments

- Motion Generation (Human3.6M)



(a) Walking.



(b) Walking without background.

Fig. 7. Generation results on Human3.6M dataset. (a) Walking sequence with background. (b) The same sequence without background.

Experiments

- Component Analysis

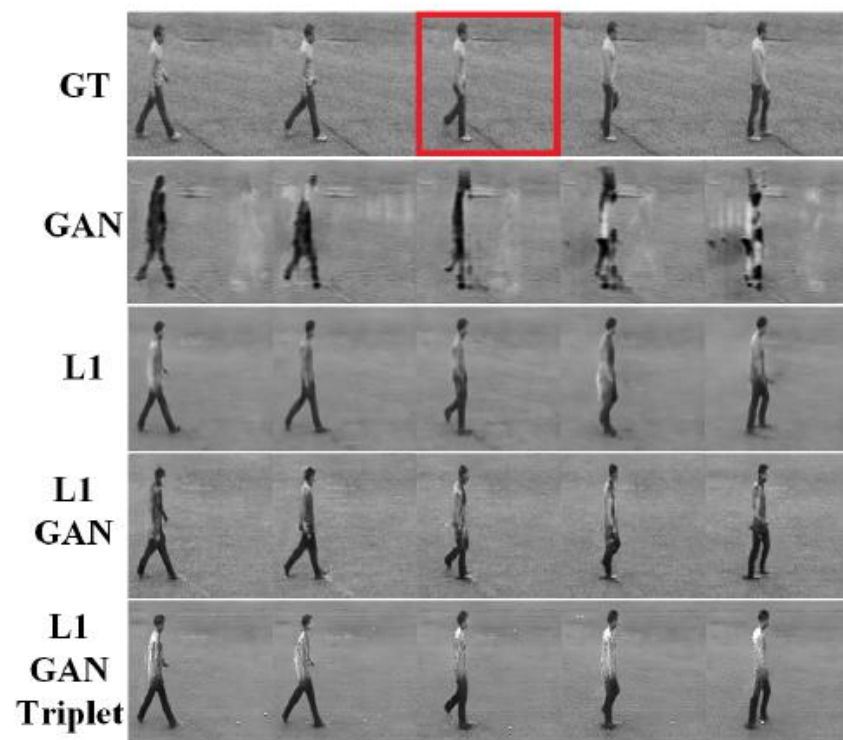


Fig. 8. Examples generated with different loss terms.

TABLE III
RECOGNITION ACCURACIES (%) ON THE GENERATED SEQUENCES. GT DENOTES THE GROUND-TRUTH RESULTS. L1 DENOTES L1 LOSS, G DENOTES GAN LOSS, T DENOTES TRIPLET LOSS.

Action	GT	L1	G	L1+G	L1+G+T
Walking	100	83.1	43.1	93.8	93.8
Running	98.5	76.9	41.5	92.3	92.3
Hand waving	100	95.4	47.7	100	100

Skeleton-aided Articulated Motion Generation

- First skeleton-guided video generation
- Triplet loss for temporal relation

Thanks

Q & A
Sijie Song