ACCV 2022 Macau

上海科技大学 ShanghaiTech University

Structural Priors for Image inpainting and Synthesis

Shenghua Gao

ShanghaiTech University

立志成才 报国裕民

Digital human modeling and manipulation

Scene modeling and manipulation

Our efforts

上海科技大学 ShanghaiTech University

Human Completion

Raw Image

Output

Audio-driven Gesture Synthesis

speech audio

Human image synthesis

Motion Imitation

Source Image

+

Reference Pose

Synthesized Image

Novel View Synthesis

Source Image

+

Novel View

Synthesized Image

Appearance Transfer

Source Image

+

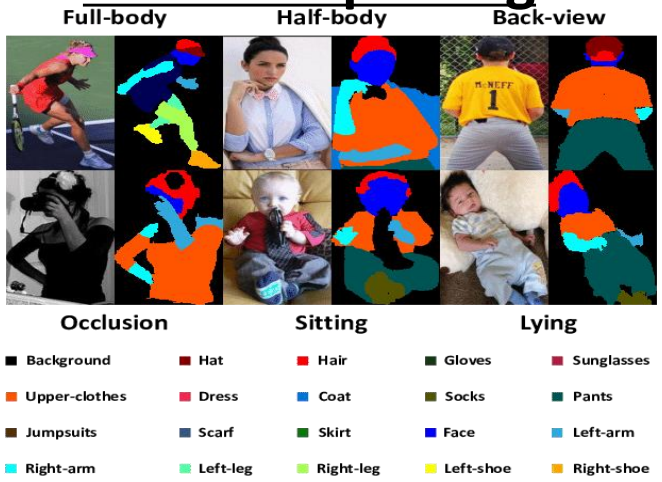
Reference Appearance

Synthesized Image

Indoor Scene Novel View Synthesis

Structural priors

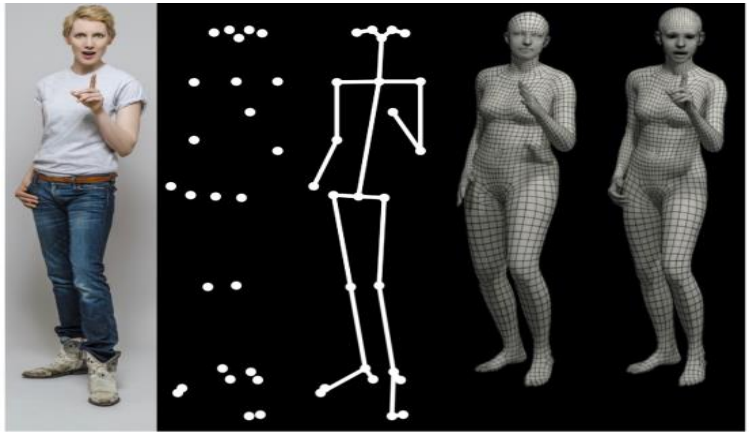
Semantic parsing



Input

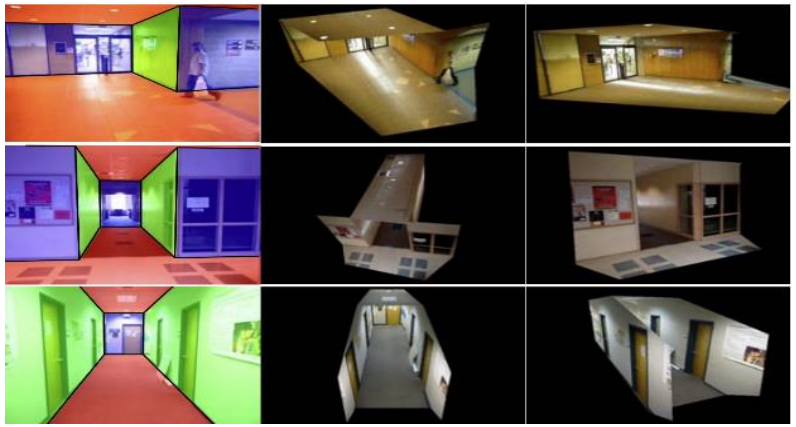
Output

Human shape and pose



speech
audio

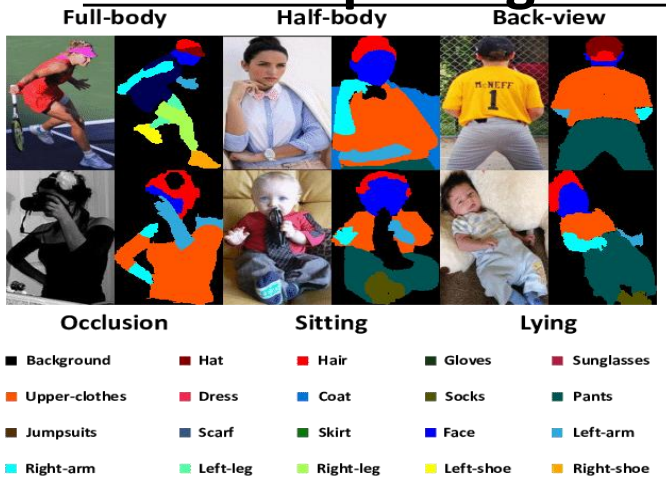
Room layout



Structural priors facilitate image inpainting

上海科技大学 ShanghaiTech University

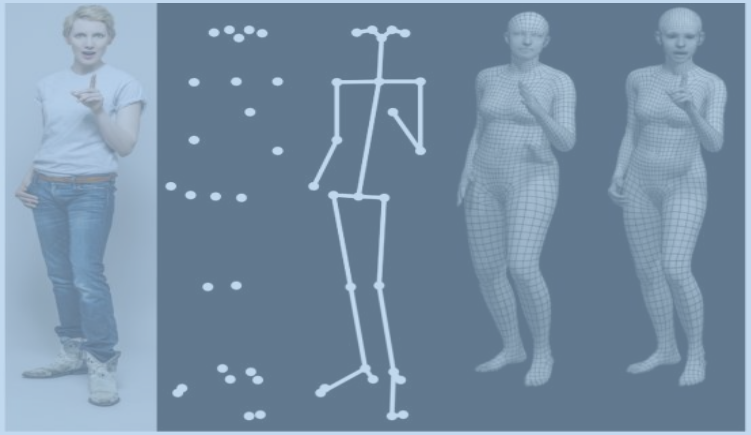
Semantic parsing results

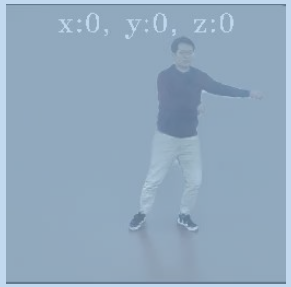


Input

Output

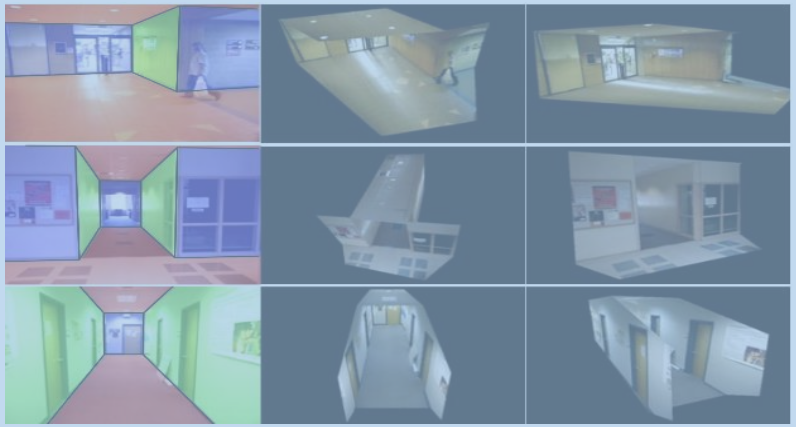
Human shape and pose





speech audio

Room layout



Semantic Aware Human Completion

- **Goal**: Given a **corrupted single person image**, human completion aims to generate a complete image with **reasonable human structure** and **plausible texture**
- It would help the occlusion removal in human modeling.

Raw Image　　Input　　Output

- This was the **first** work for Human Completion

- Zhao, Zibo, et al. "Prior based human completion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Existing Methods

- Only consider scene (natural image or human face) completion
- Less consider the shape prior of in human completion

Natural Scene Completion

Input Output

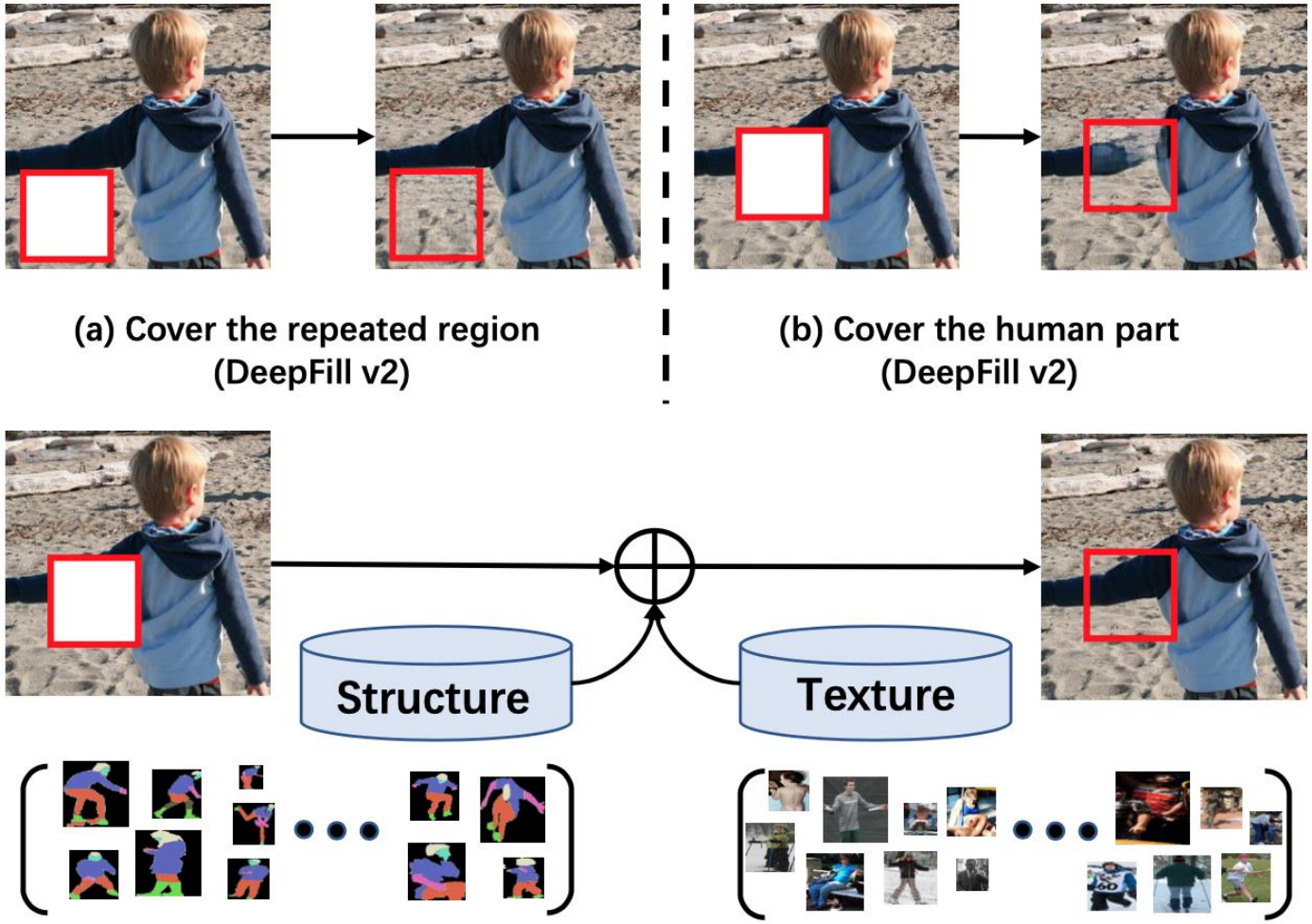
Human Completion

Input Output

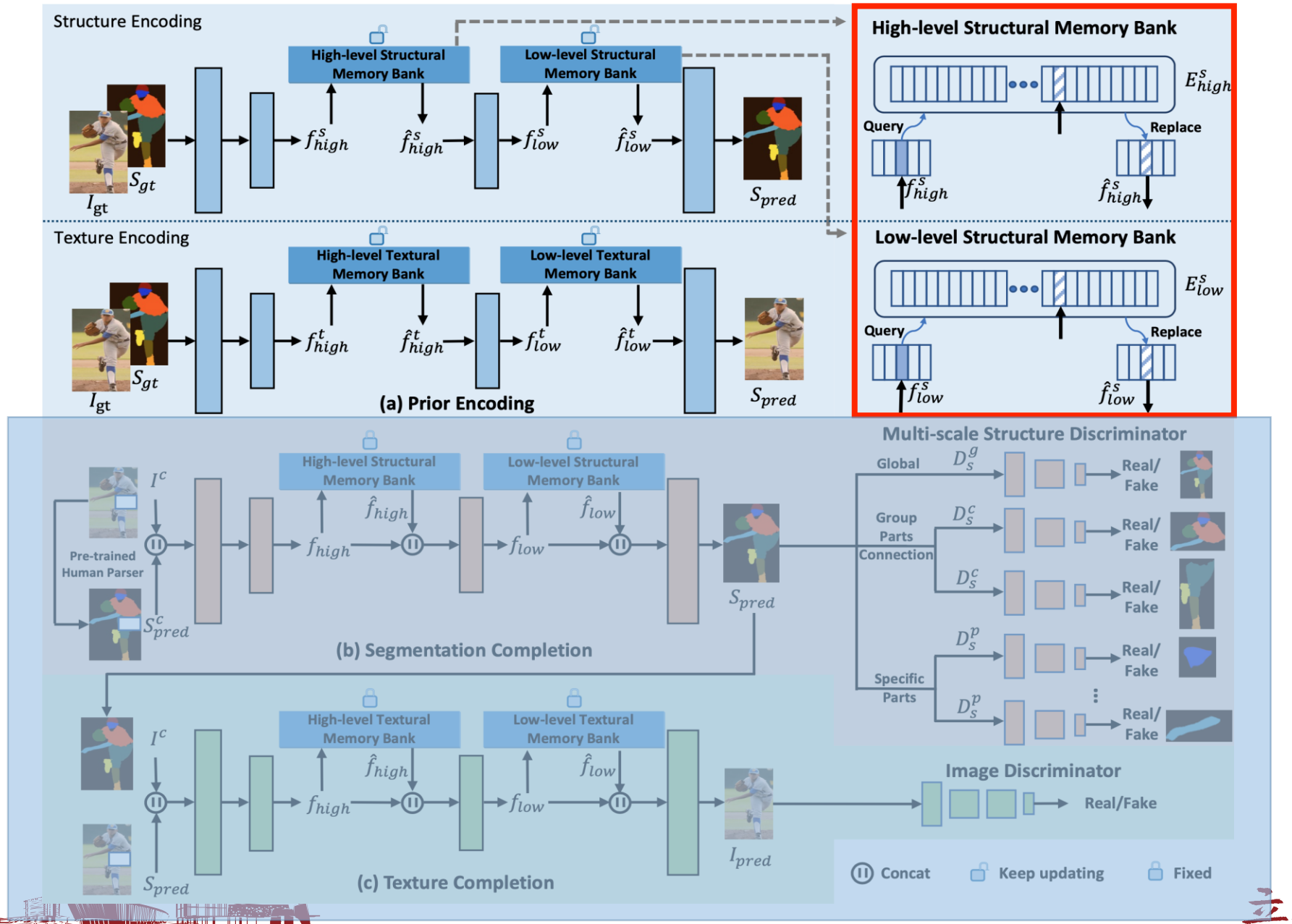
Possible for the failure of existing methods for human completion: Single image lacks references for recovering the lost pixels for human

Solution: We have priors about the possible structure of human body, and such prior should be encoded as side information for human completion.

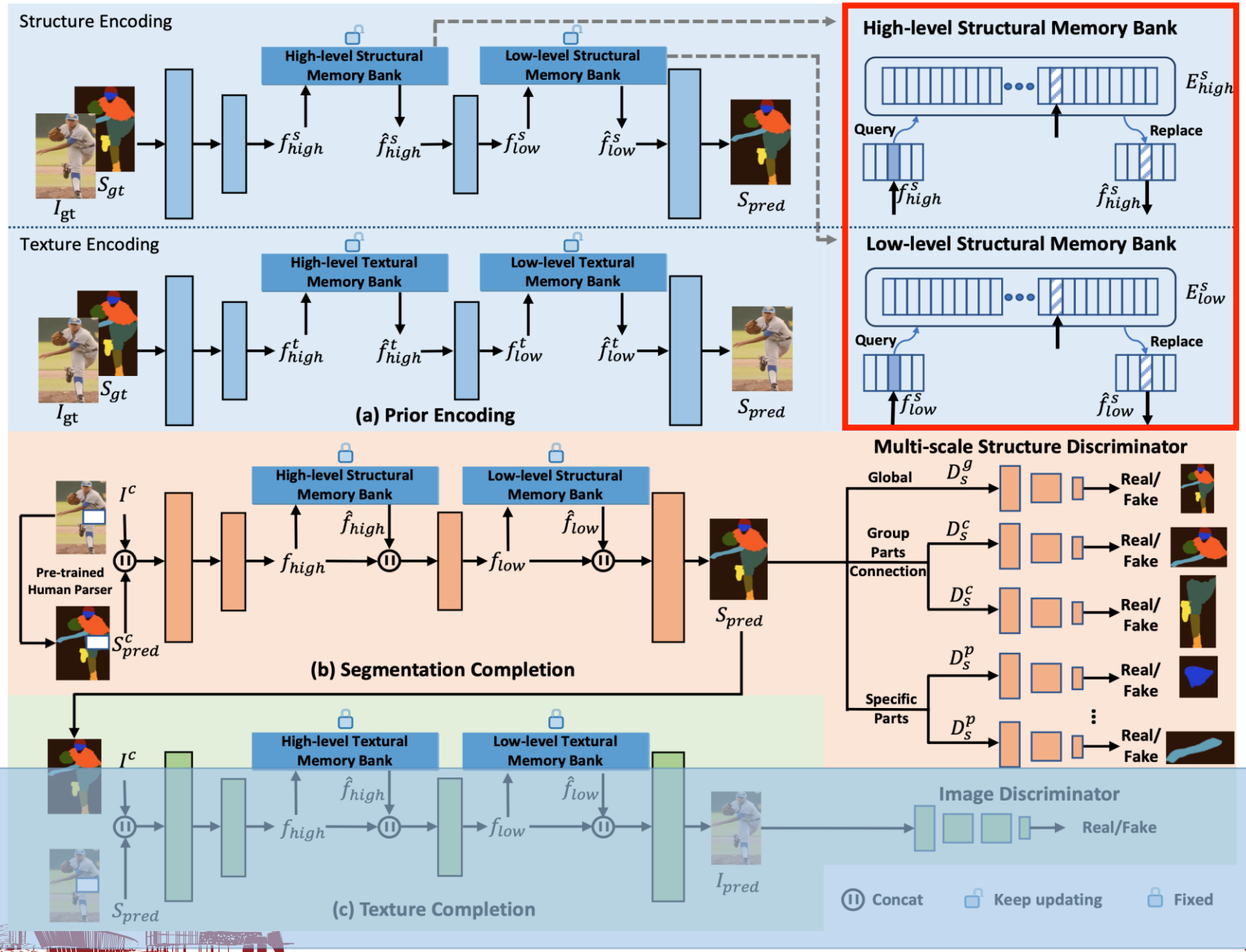
Key Idea — Utilization of Priors



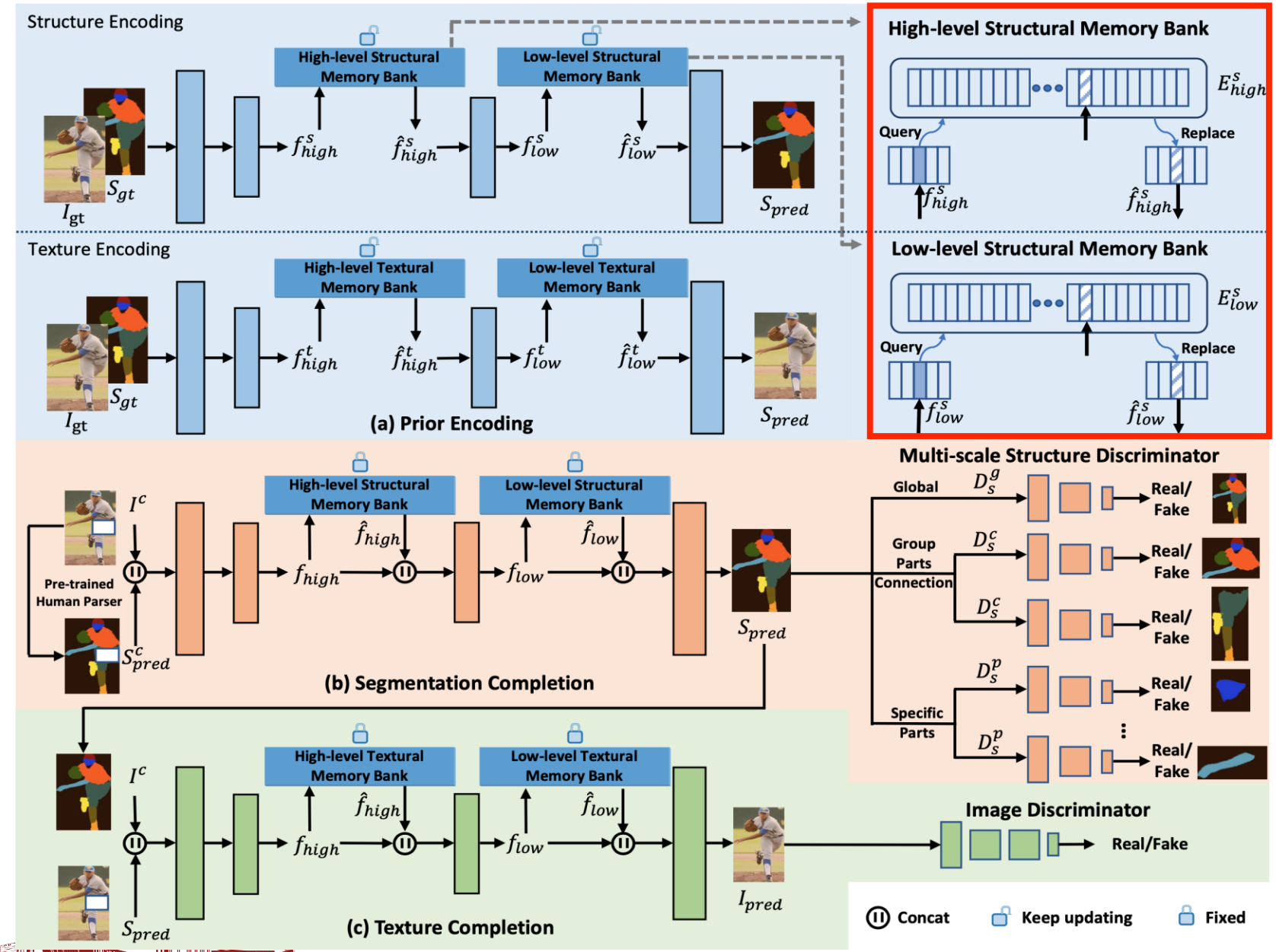
Our solution



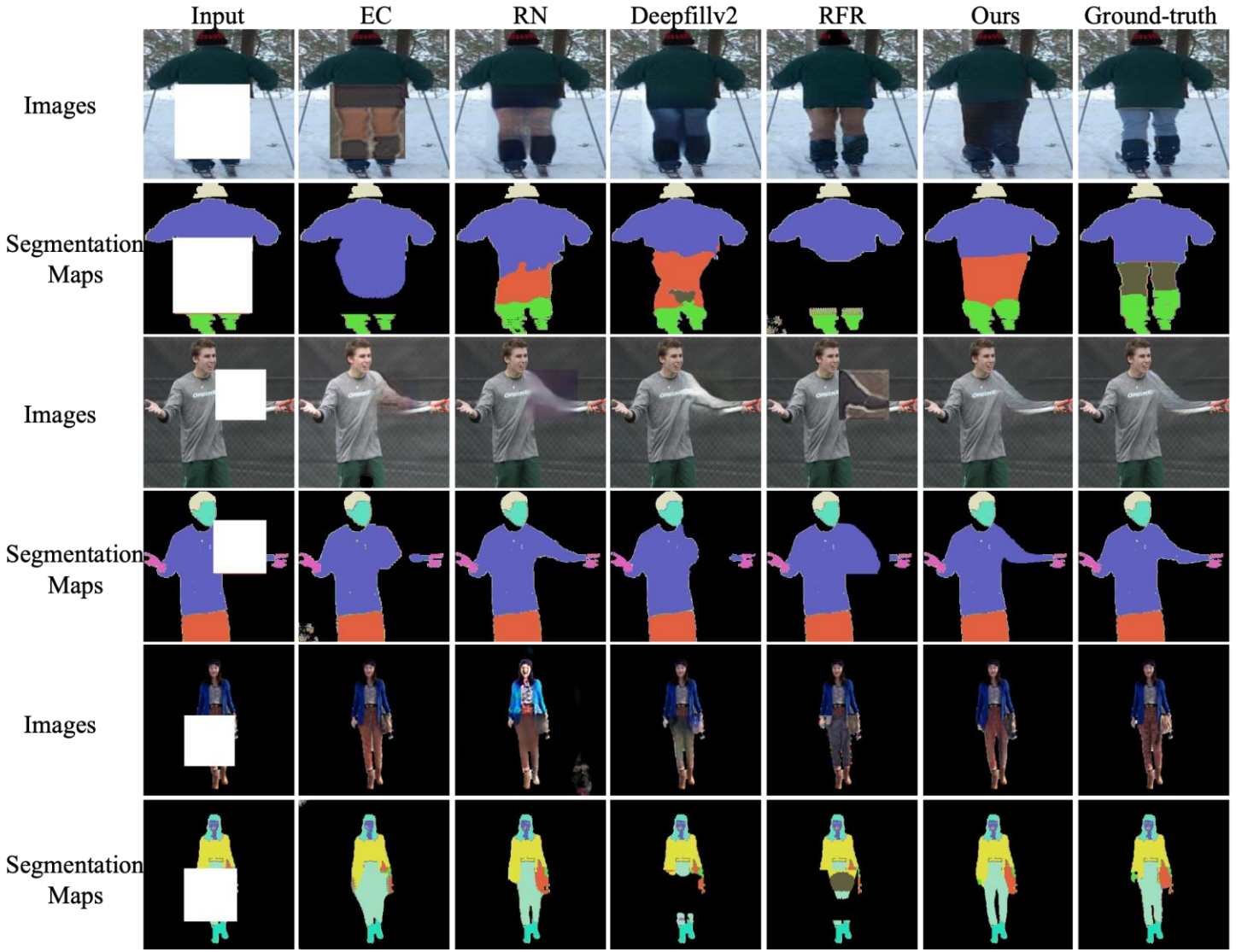
Our solution



Our solution



Visualization



Free-Form Occlusions

Structural priors facilitate human manipulation

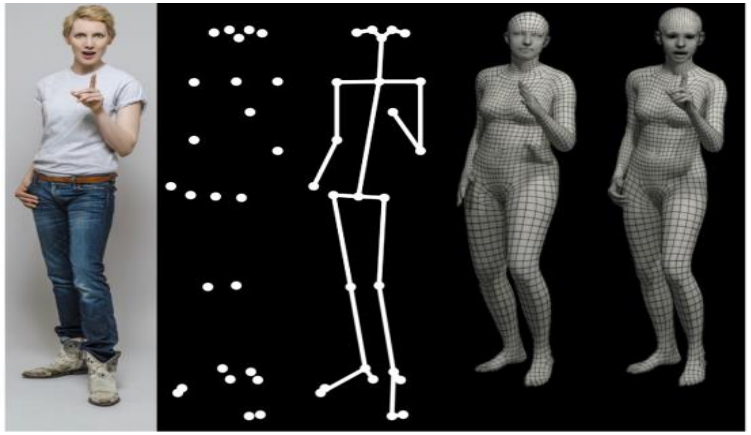
上海科技大学 ShanghaiTech University

Semantic parsing results

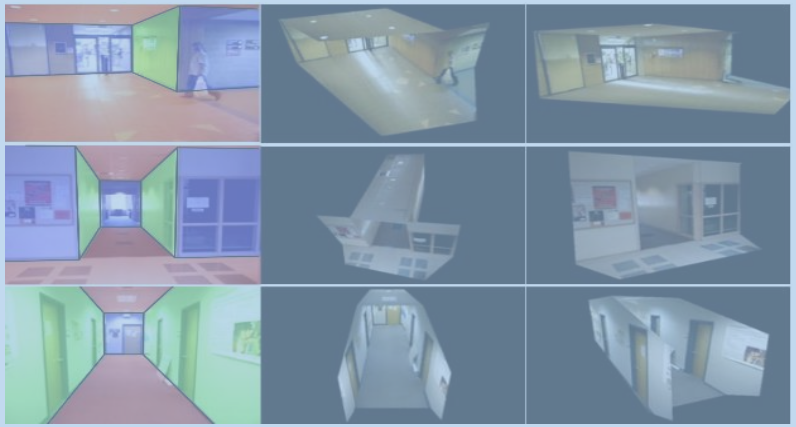
Input

Output

Human shape and pose

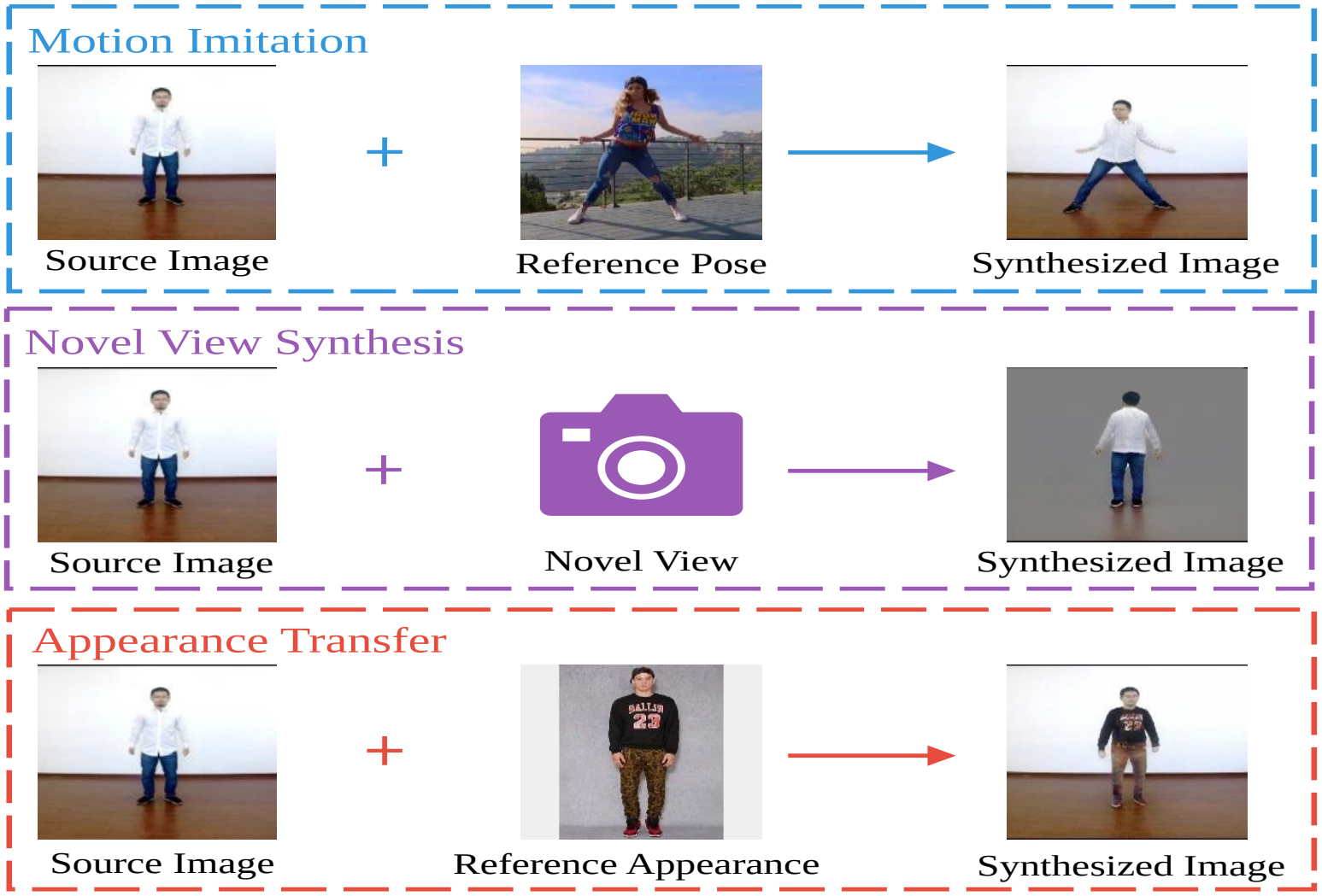


Room layout



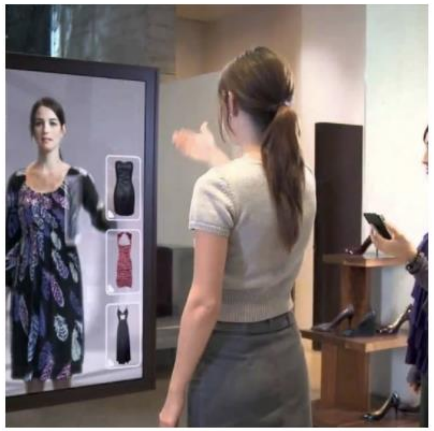
立志成才报国裕民

SMPL Guided Human Image Analysis



Wen Liu, et al, "Liquid Warping GAN with Attention: A Unified Framework for Human Image Synthesis", IEEE TPAMI, 2020
Wen Liu, et al, Liquid Warping GAN: A Unified Framework for Human Image Synthesis. ICCV, 2019

Applications



virtual fitting

short video editing

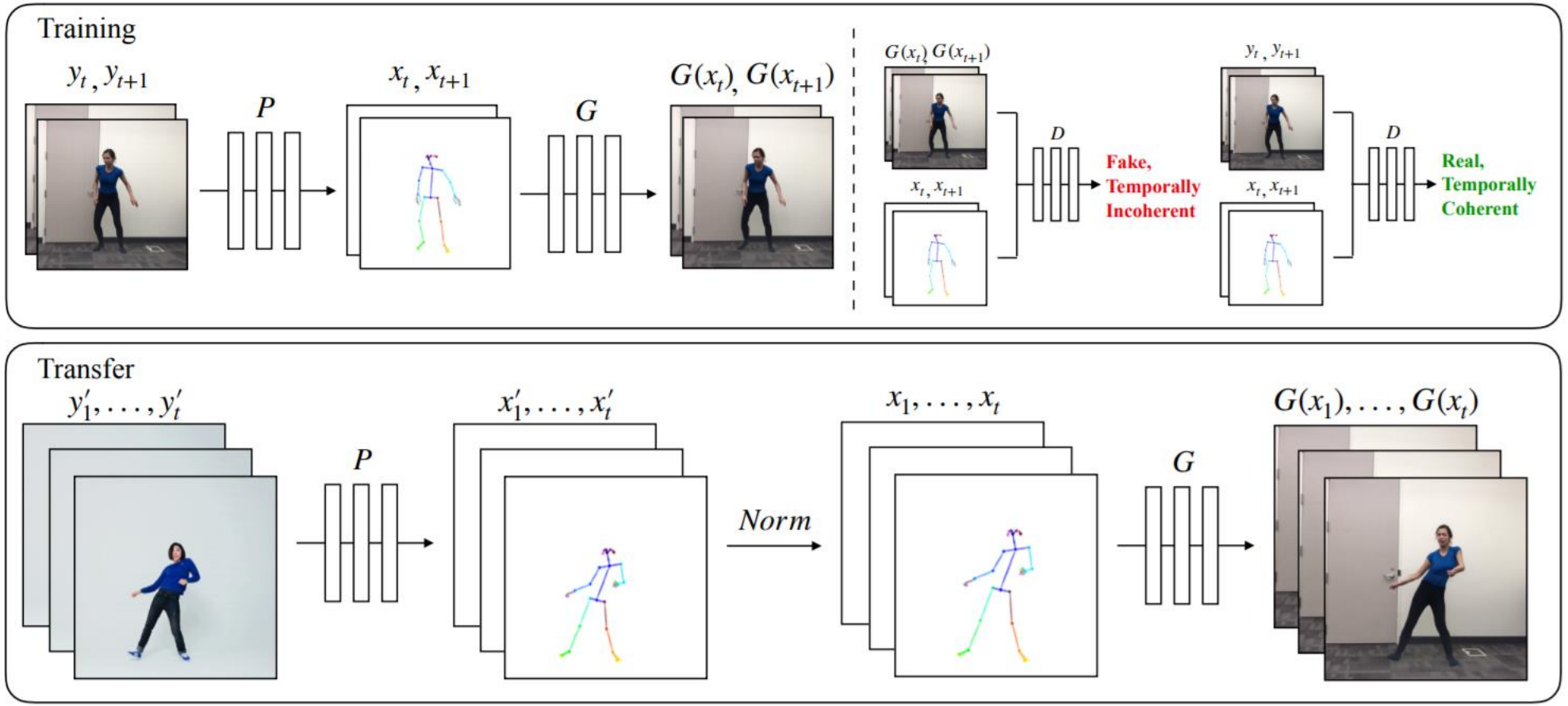
Entertainment

Virtual presenter

VR Games

Intelligent video Editing

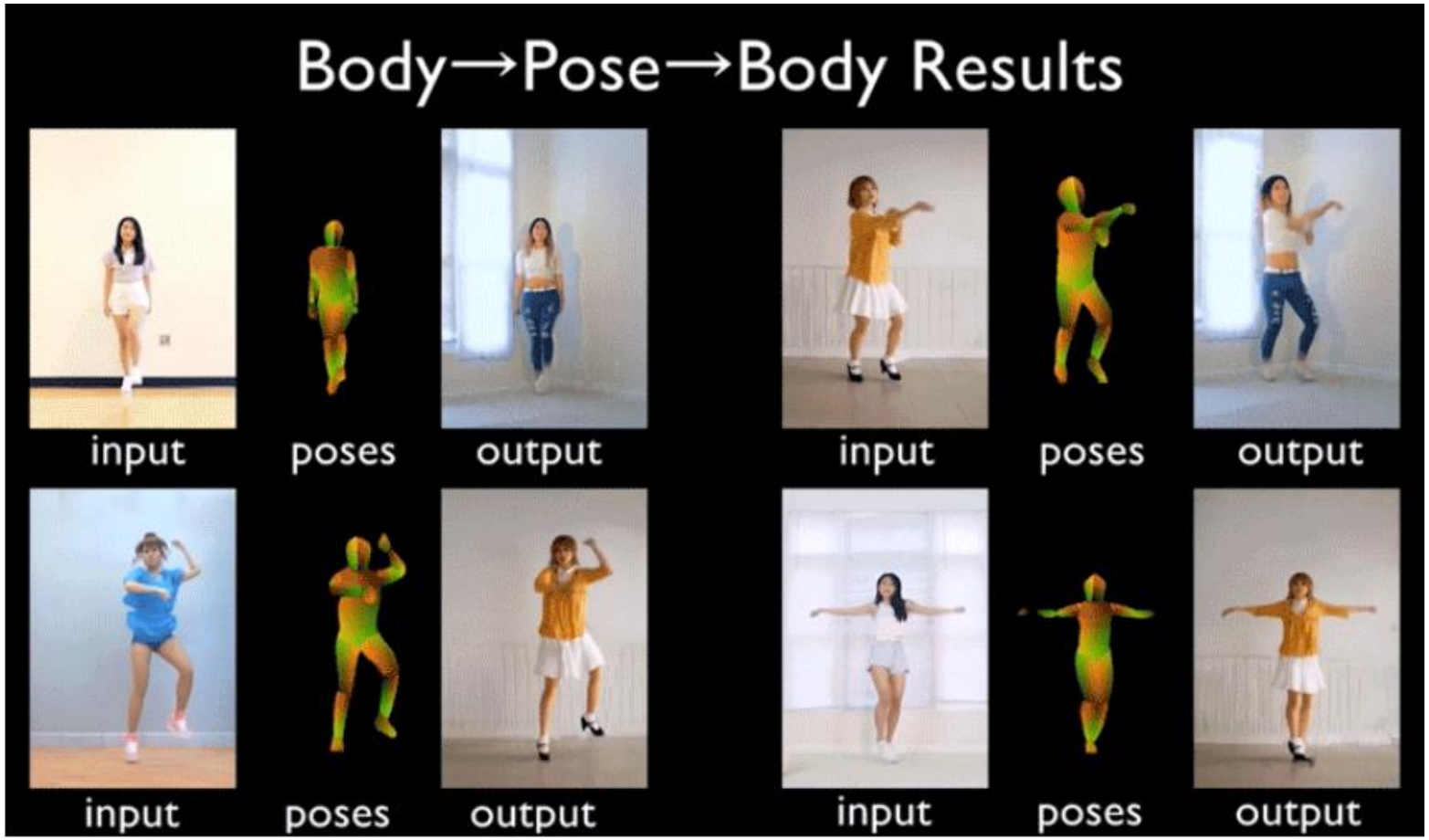
Related work



Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody Dance Now, in ICCV 2019

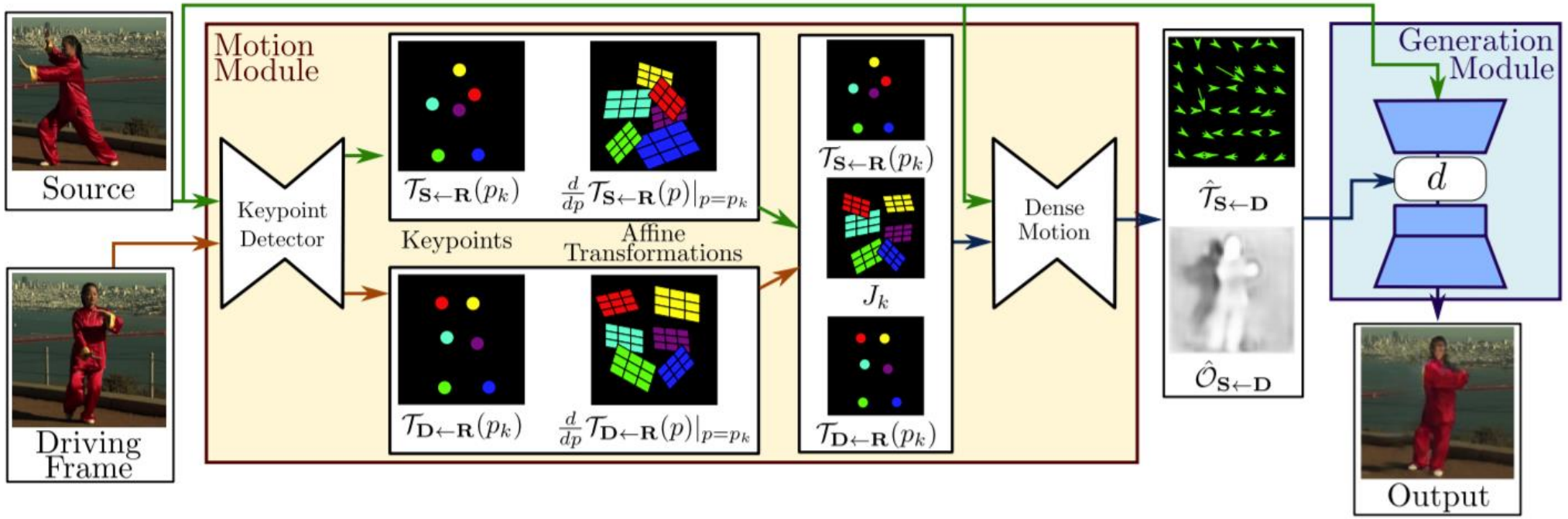
Related work

Image-to-Image translation



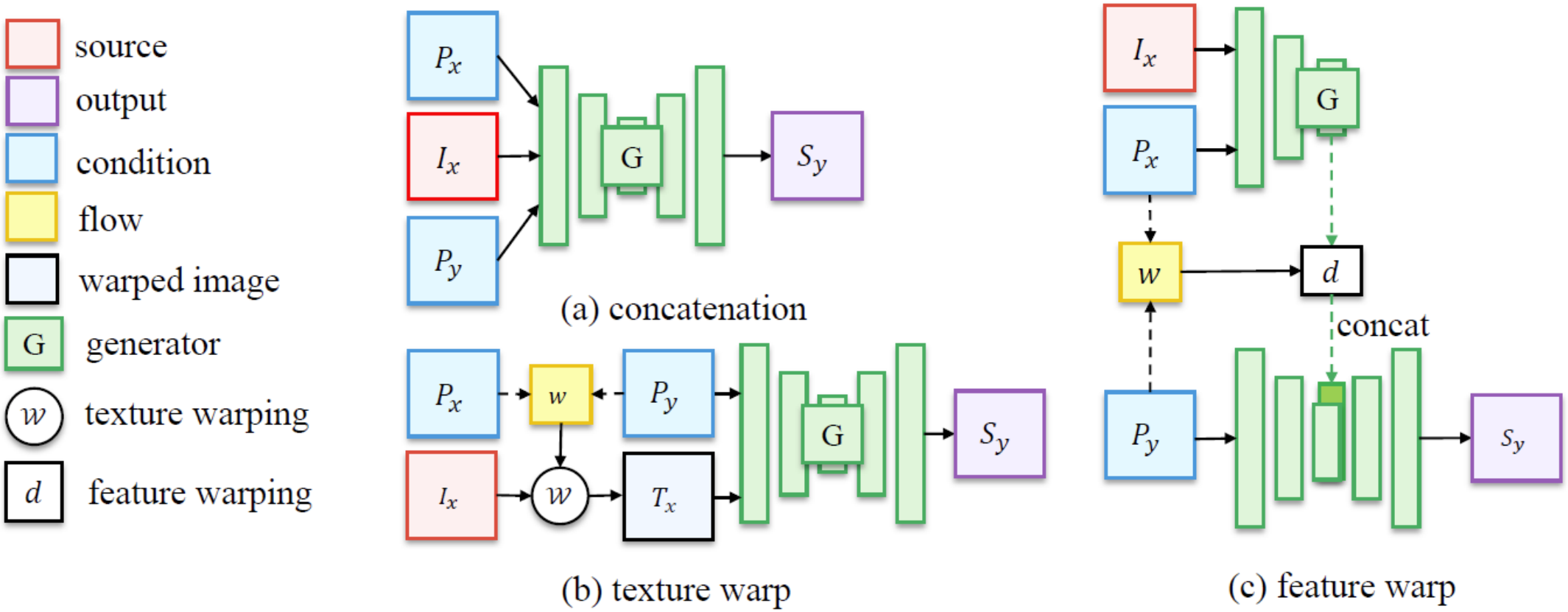
Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, Bryan Catanzaro, Video-to-Video Synthesis, in NeurIPS 2018

Related work



Siarohin, Aliaksandr and Lathuilière, Stéphane and Tulyakov, Sergey and Ricci, Elisa and Sebe, Nicu, First Order Motion Model for Image Animation, Conference on Neural Information Processing Systems (NeurIPS) 2019.

A review of existing methods



Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao, Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis, ICCV 2019.

Existing work:

- Sparse keypoints based methods may change the shape of the target person
- Cannot generalize well to novel persons
- Details of faces and clothes are lost

source

Reference

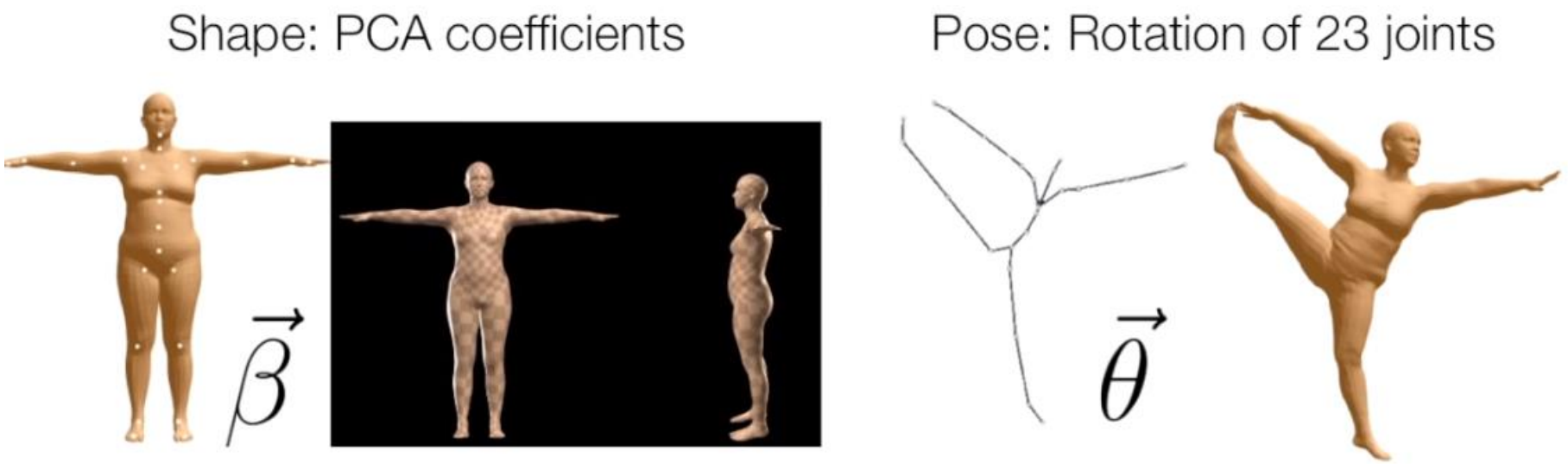
Target

G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, “Synthesizing images of humans in unseen poses,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

Our solution

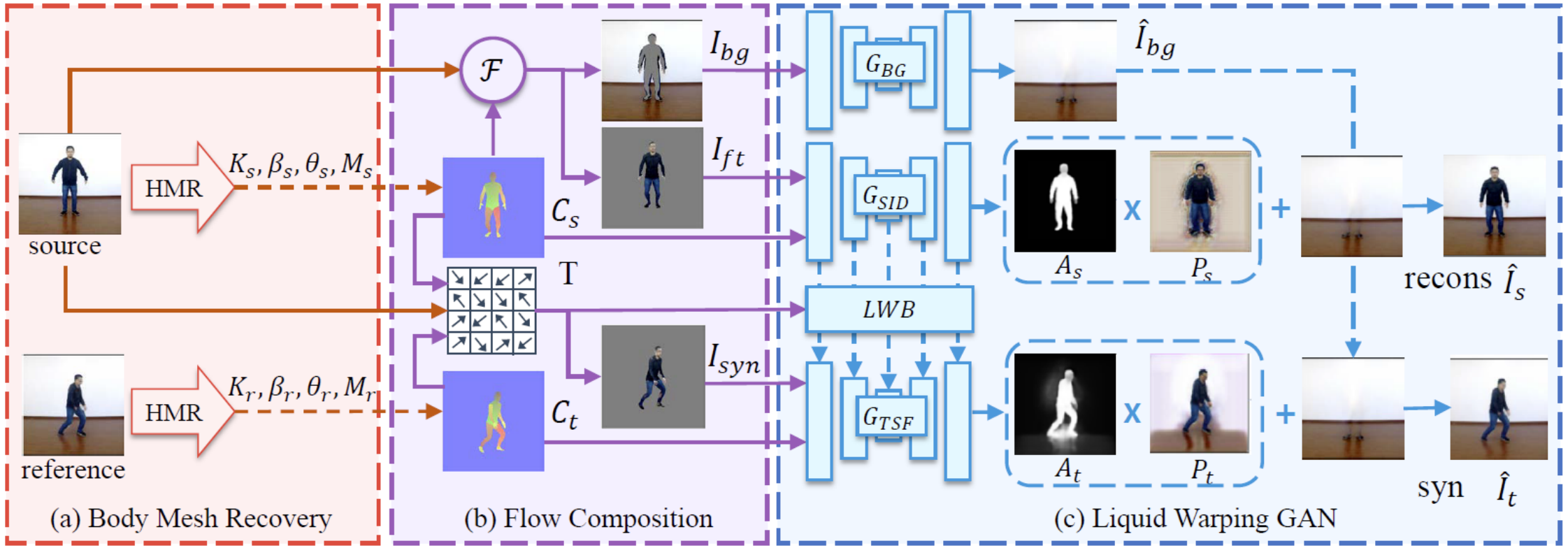
Model Human with SMPL model: decouples the human pose and human shape

- SMPL= $M(\theta, \beta)$,θ (pose) β (shape);

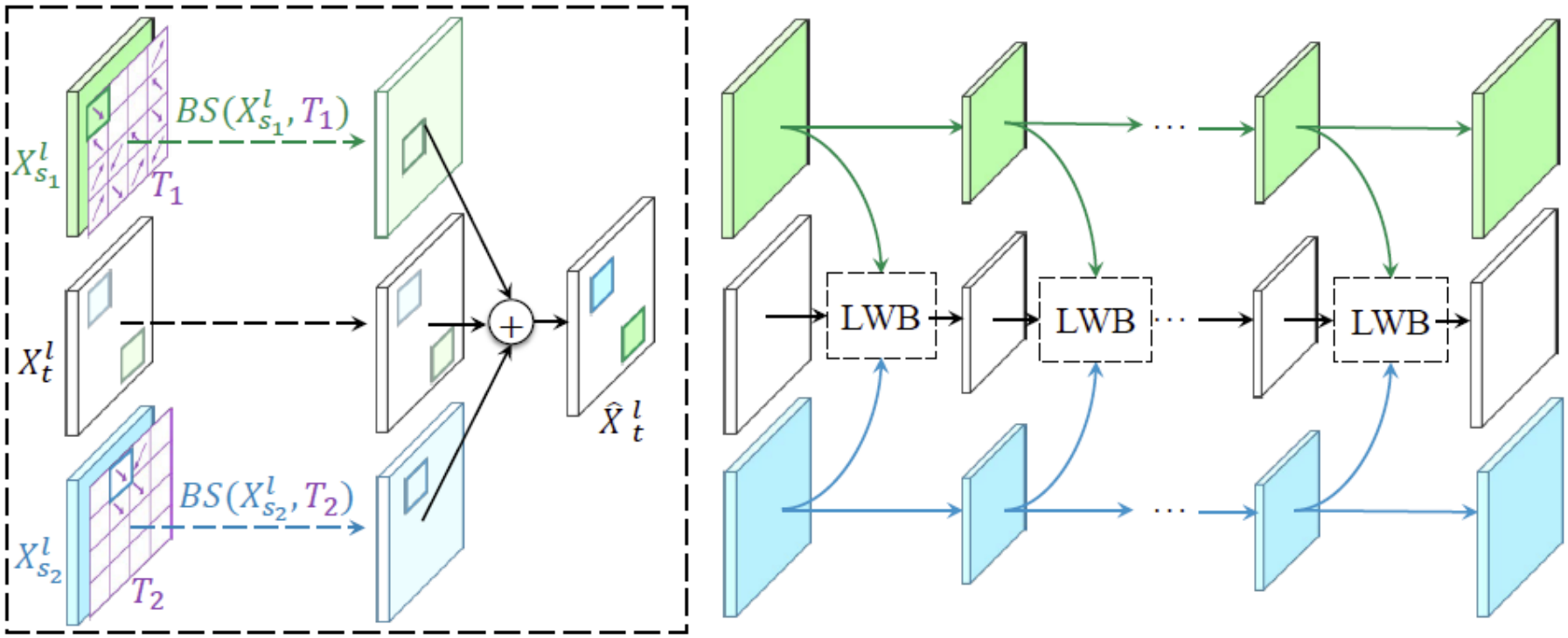


Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. SIGGRAPH Asia 2015.

Our solution



Liquid Warping Block

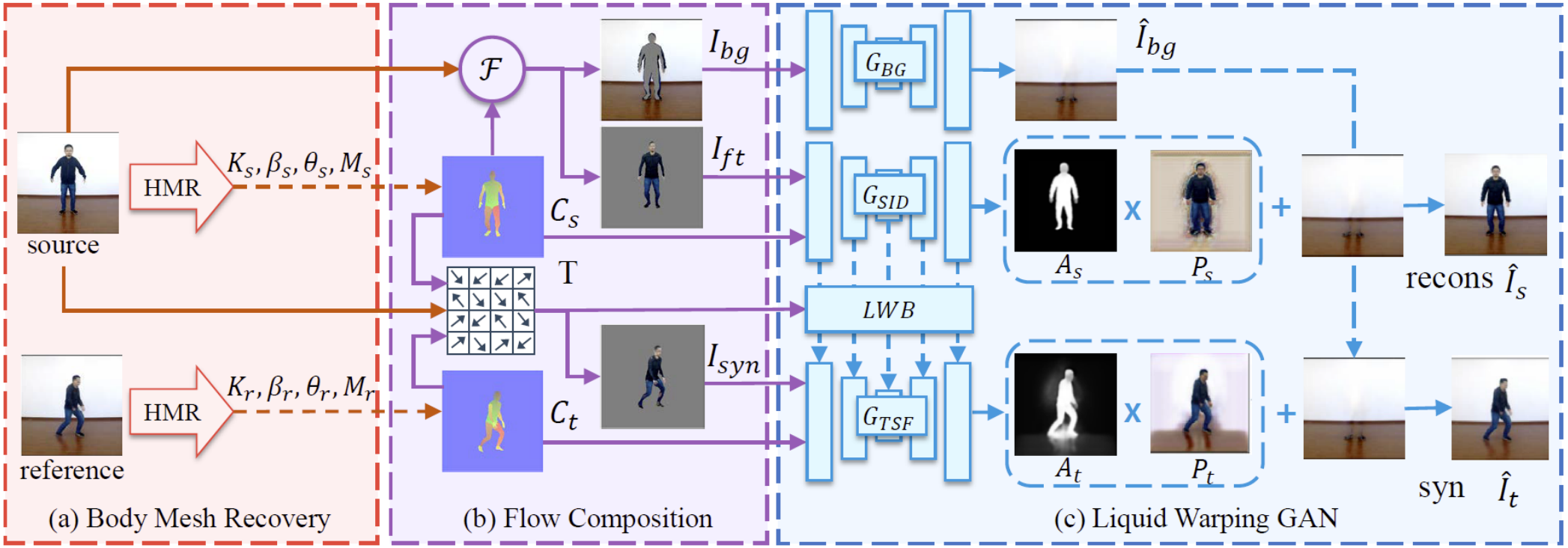


(a) Liquid Warping Block (LWB)

(b) Liquid Warping Generator

$$\widehat{X}^l_t = BS(X^l_{s_1}, T_1) + BS(X^l_{s_2}, T_2) + X^l_t.$$

Loss functions

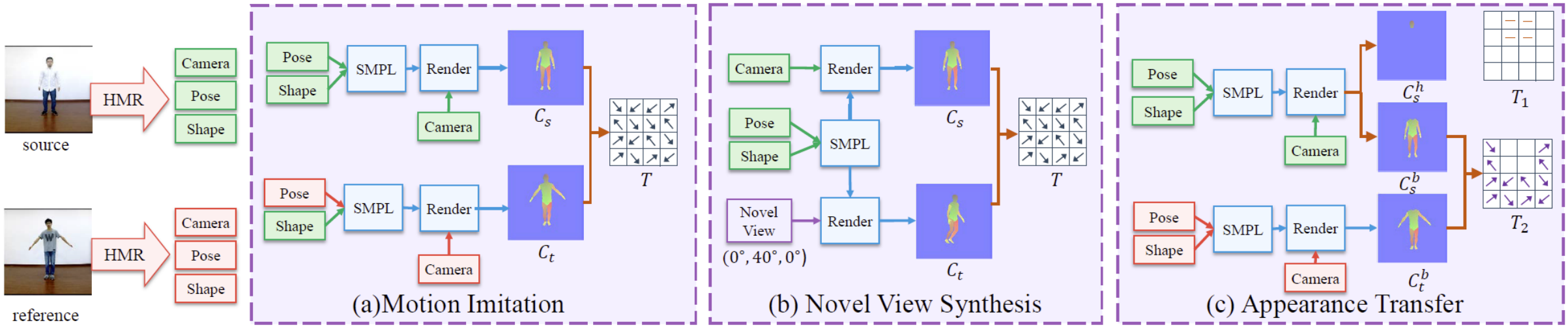


Loss function for generator: $\mathcal{L}^G = \lambda_p \mathcal{L}_p + \lambda_f \mathcal{L}_f + \lambda_a \mathcal{L}_a + \mathcal{L}^G_{adv}$

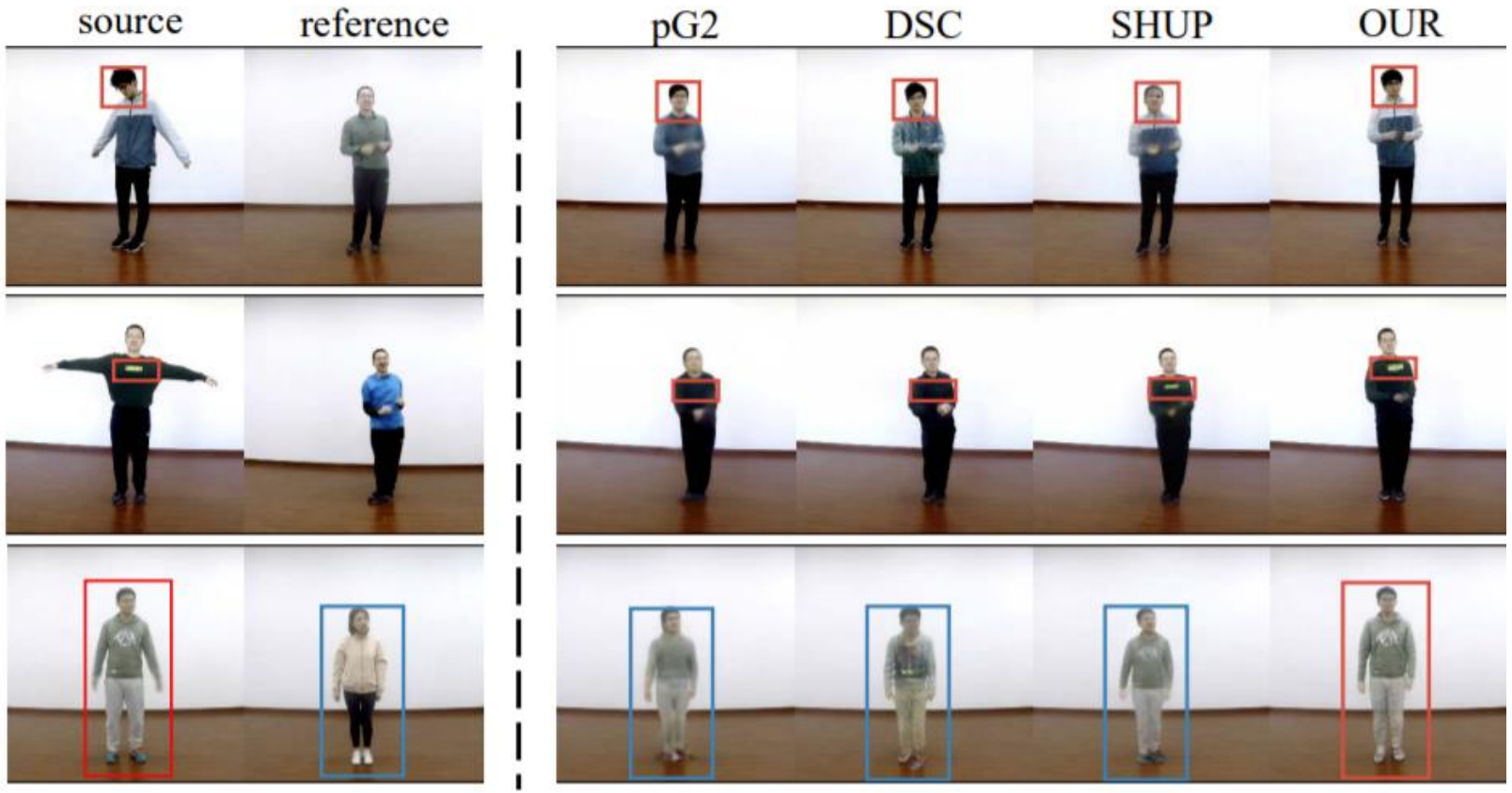
- Perceptual Loss: $\mathcal{L}_p = \|\hat{I}_s - I_s\|_1 + \|f(\hat{I}_t) - f(I_r)\|_1$, here f is a pre-trained VGG-19;
- Face Identity Loss: $\mathcal{L}_f = \|g(\hat{I}_t) - g(I_r)\|_1$, here, g is a pre-trained SphereFaceNet;
- Adversarial Loss: $\mathcal{L}^G_{adv} = \sum D(\hat{I}_t, C_t)^2$, here, D is the discrminator network;
- Attention Regularization Loss, $\mathcal{L}_a = \|A_s - S_s\|_2^2 + \|A_t - S_t\|_2^2 + TV(A_s) + TV(A_t)$.

Loss function for discriminator: $\mathcal{L}^D = \sum[D(\hat{I}_t, C_t) + 1]^2 + \sum[D(I_r, C_t) - 1]^2$

Calculation of Transformation T for different tasks

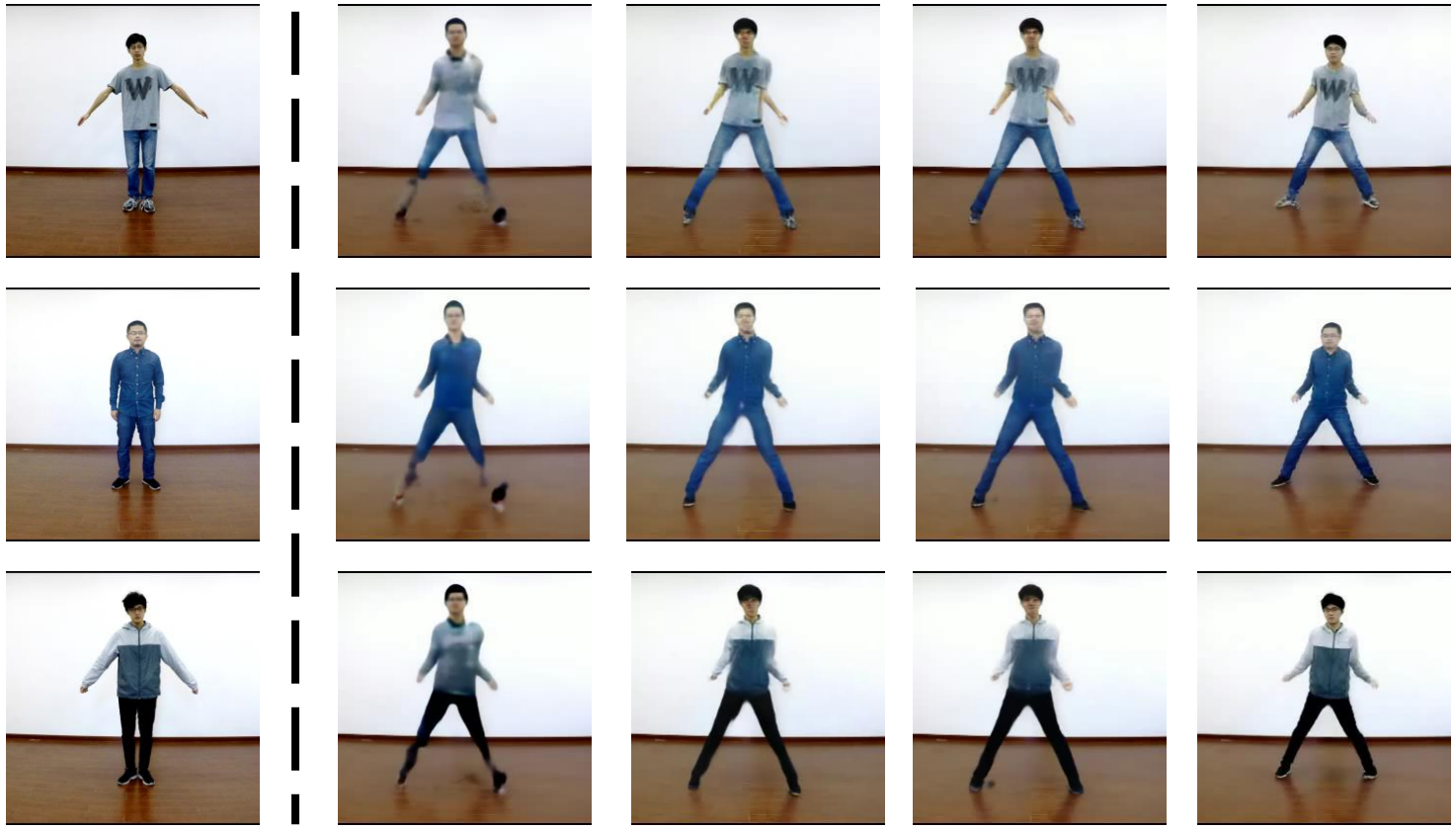


Experimental Results

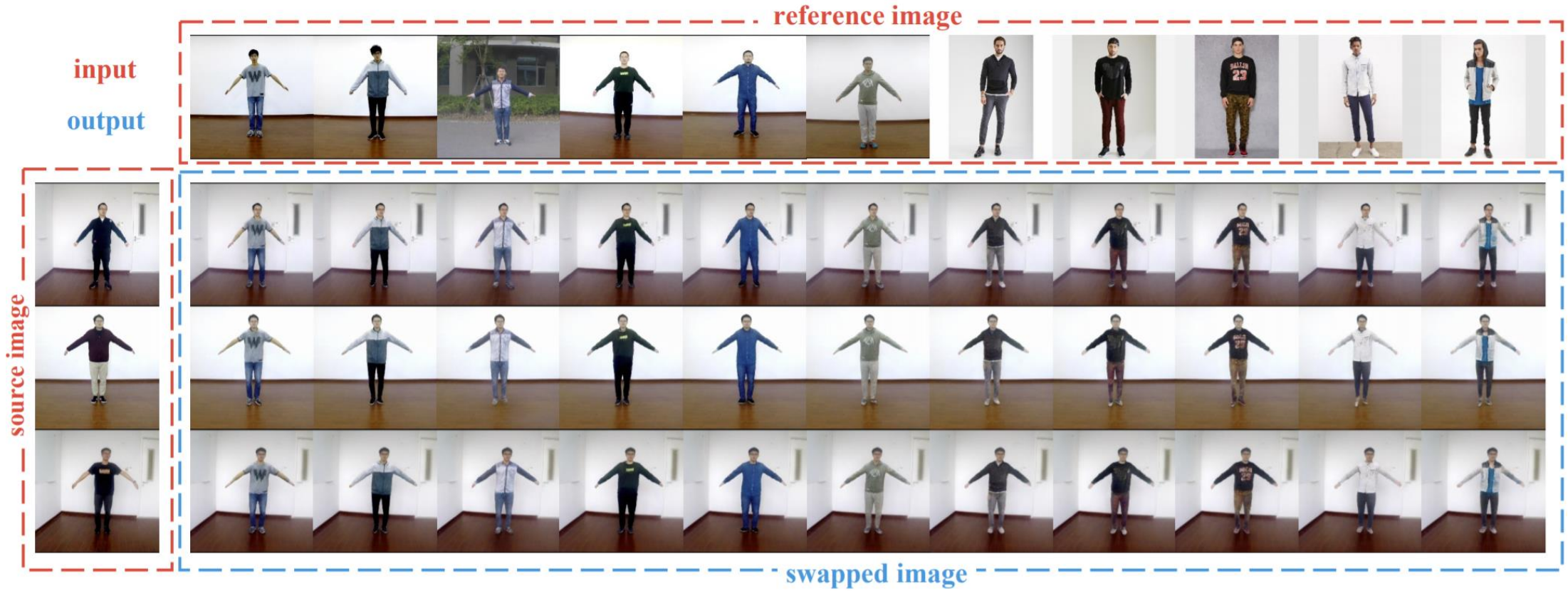


Experimental Results: Motion Imitation

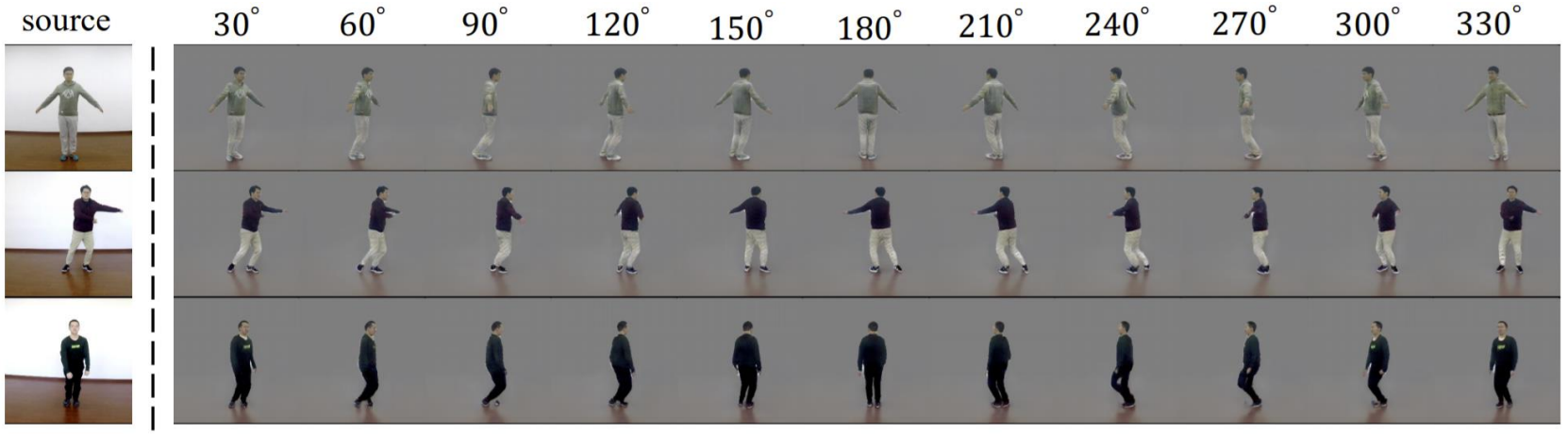
Reference Pose



Experimental Results: Appearance Transfer

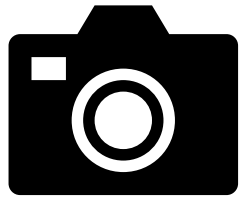


Experimental Results:Novel View Synthesis



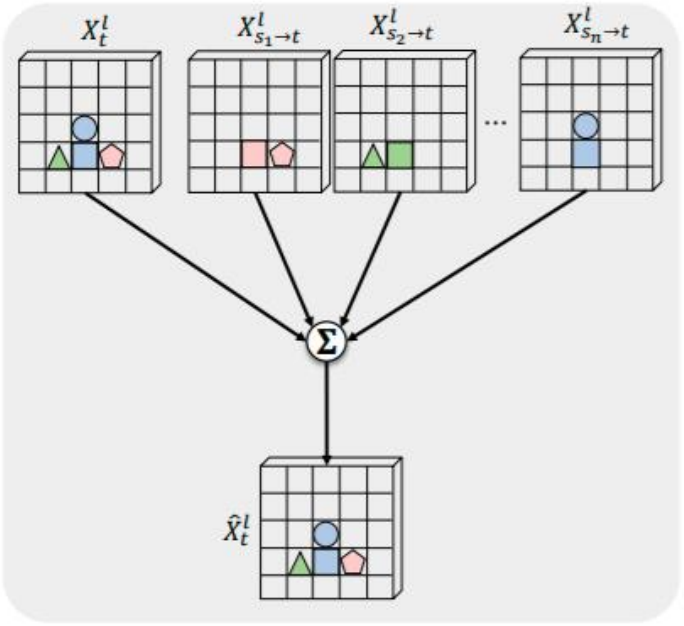
Source

+

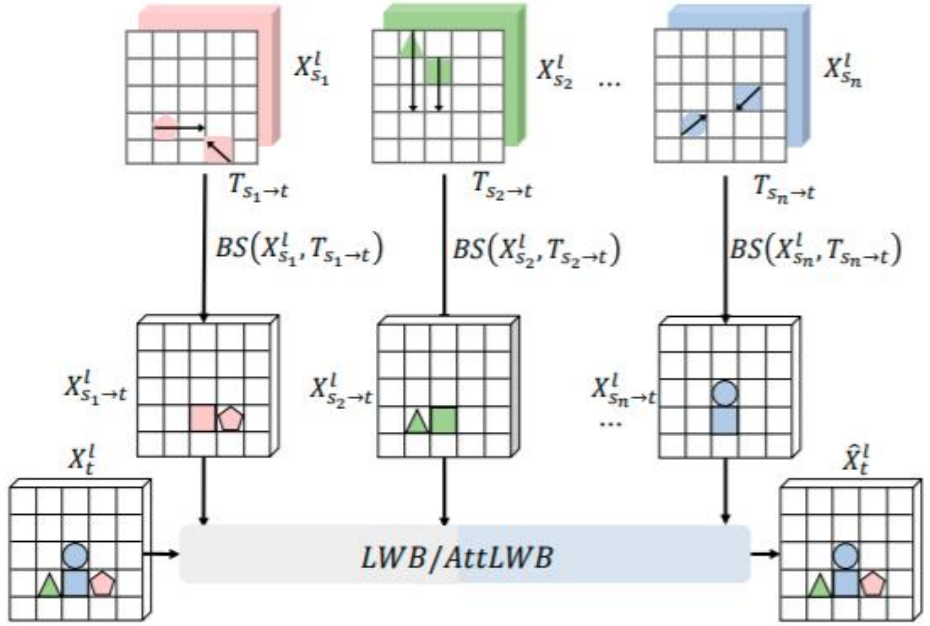


Novel view

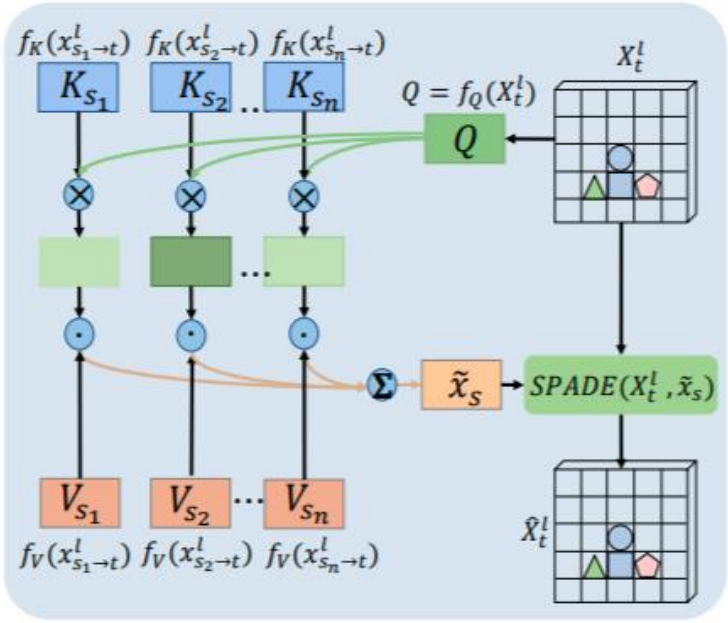
Attentional Liquid Warping GAN



(a) Add Warping Block (LWB)

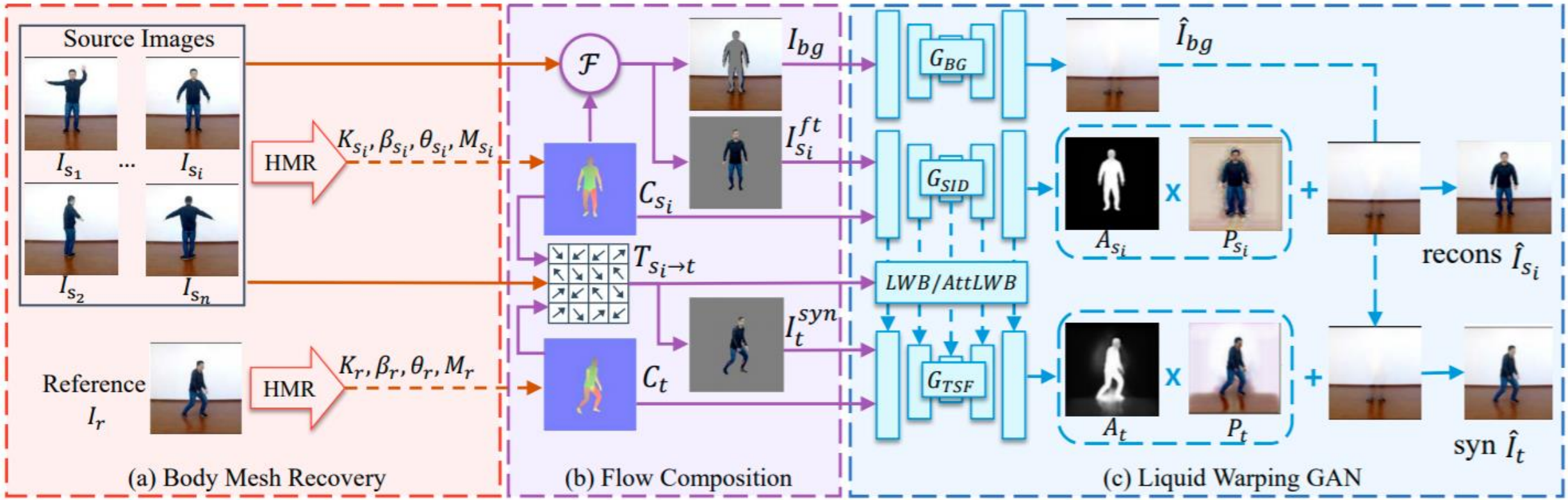


(b) (Attentional) Liquid Warping Block



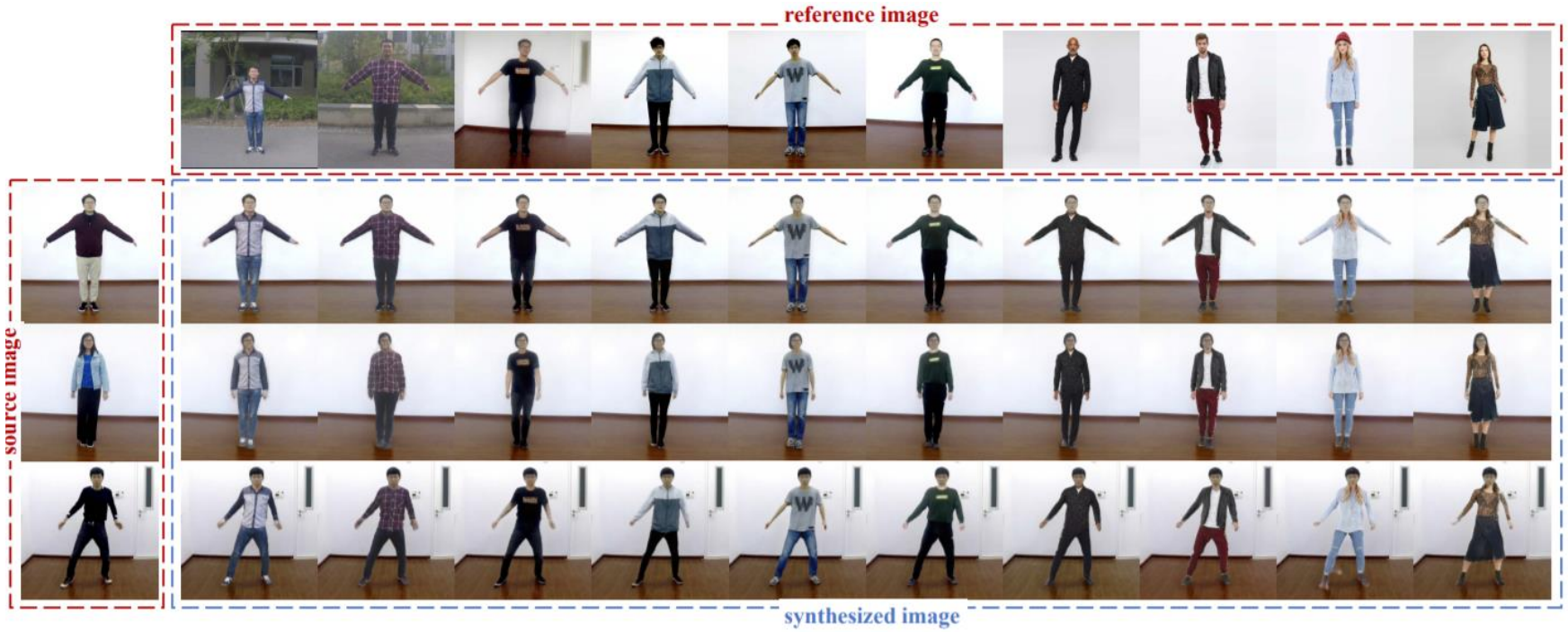
(c) Attentional Warping Block (AttLWB)

Attentional Liquid Warping GAN



Experimental Results

	PSRN↑	SSIM↑	LPIPS↓	Body-CS↑	Face-CS↑
LWB	17.707	**0.734**	0.225	0.891	0.642
AttLWB	**17.783**	0.726	**0.220**	**0.896**	**0.706**



Experimental Results: Motion Imitation

input reference Synthesize video (512 x 512)

Reference Pose

Reference Appearance

Structural priors facilitated human editing

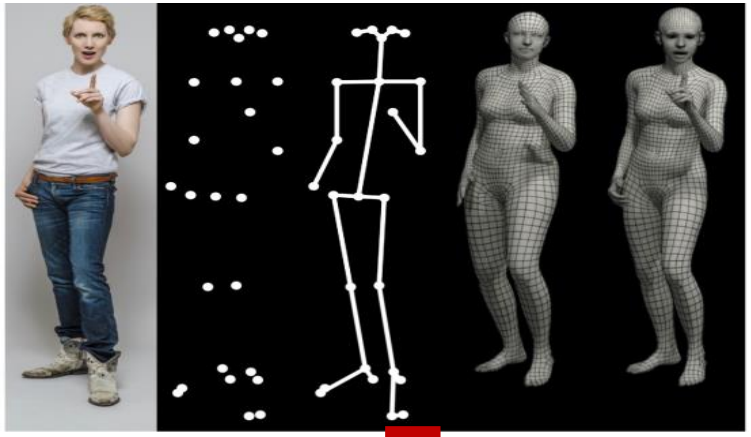
上海科技大学 ShanghaiTech University

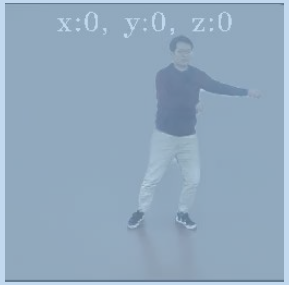
Semantic parsing results

Input

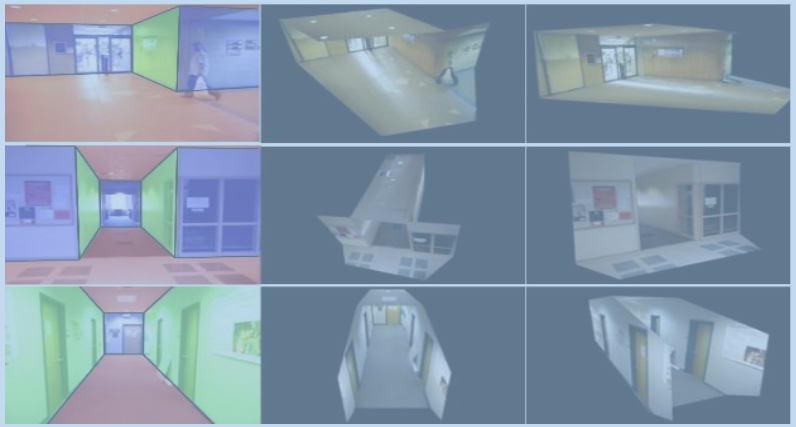
Output

Human shape and pose



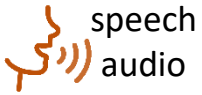


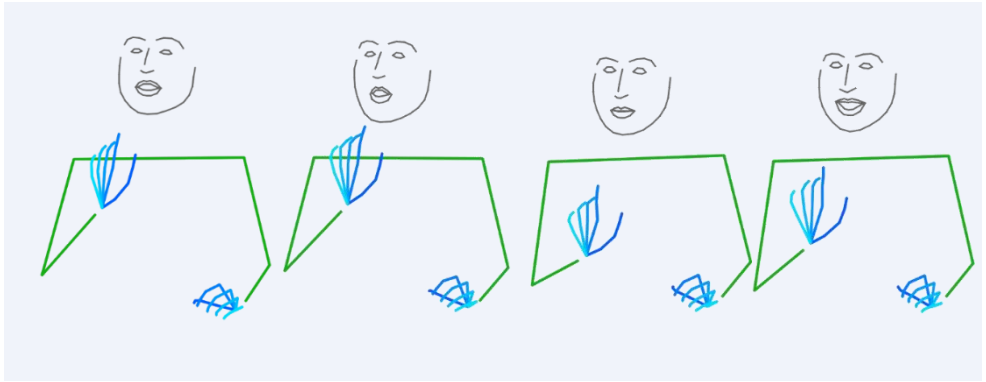
Room layout

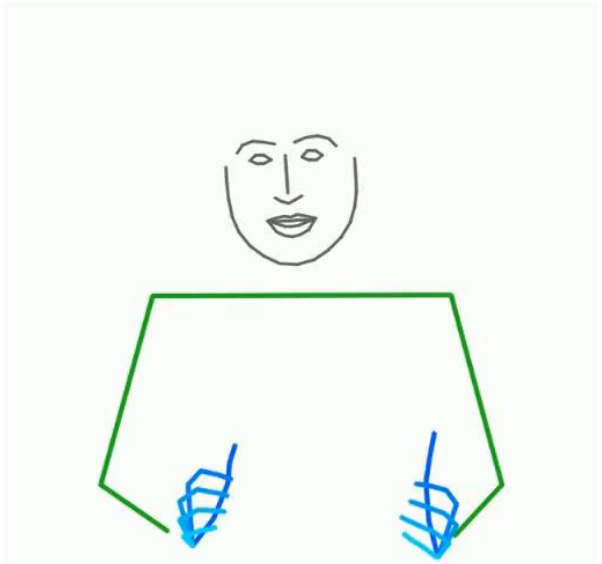


立志成才报国裕民

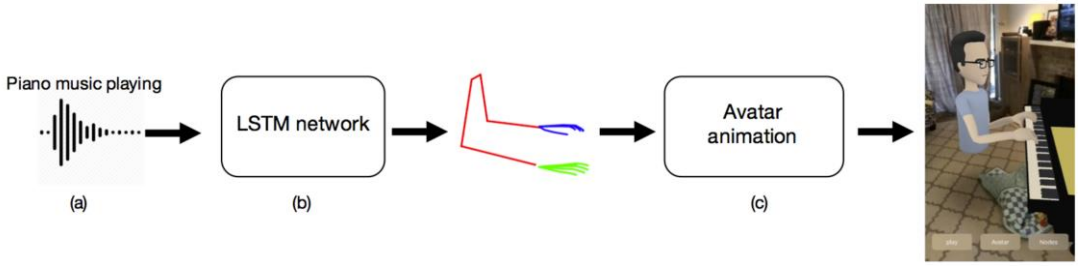
Audio-Driven Gesture Synthesis



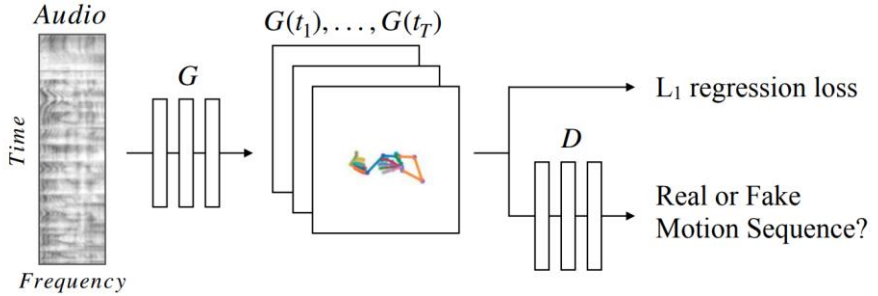




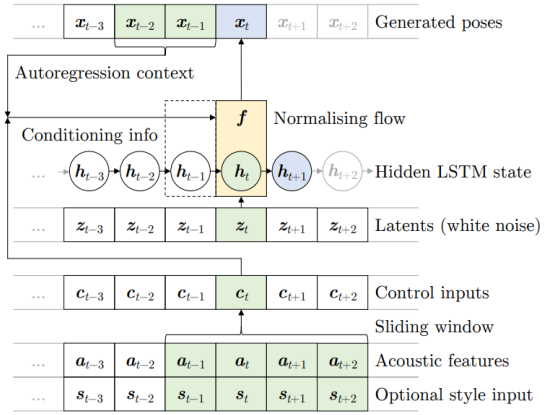
Audio-Driven Gesture Synthesis



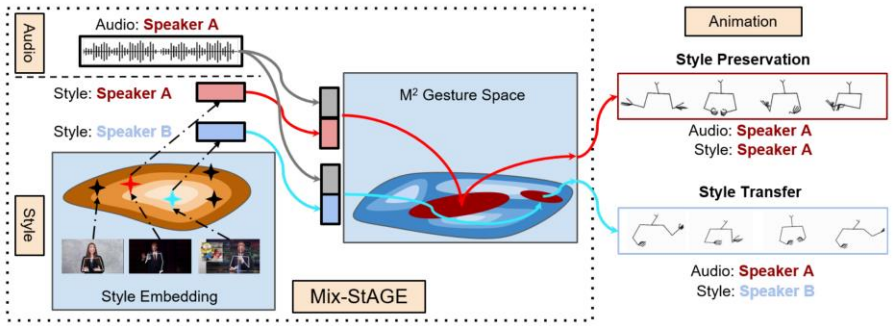
- LSTM Regression
- [Shlizerman *et al.*]



- CNN regression with an adversarial loss
- [Ginosar *et al.*]

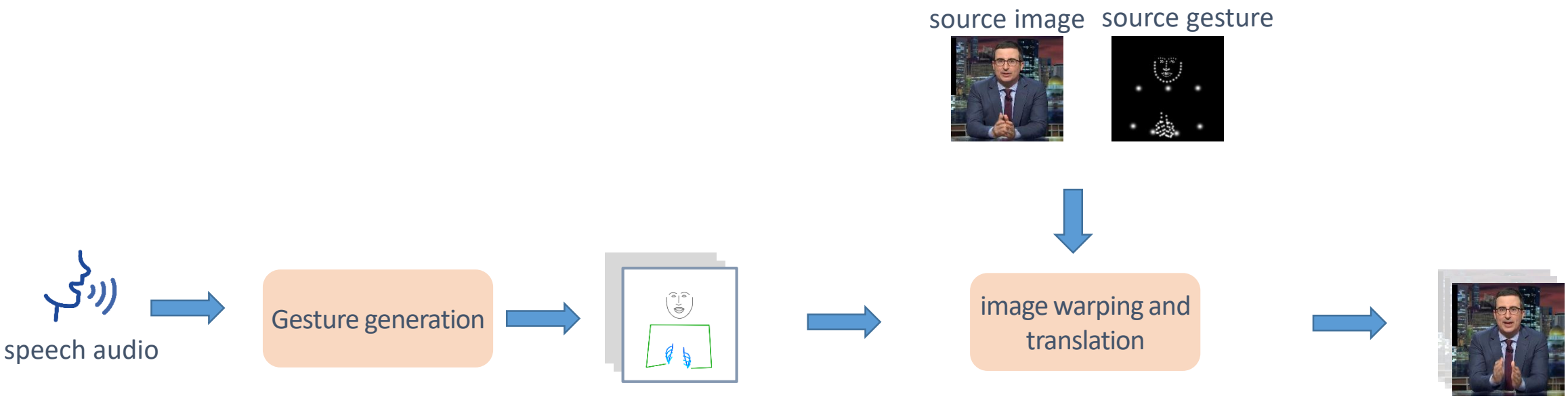


- probabilistic modeling with normalizing flows
- [Alexanderson *et al.*]



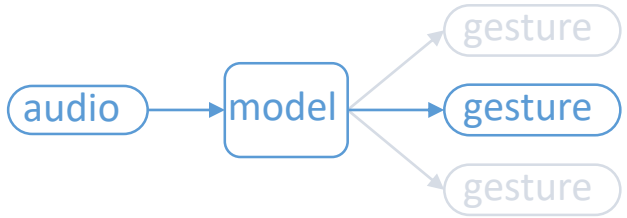
- style transfer and preserving across subjects
- [Ahuja *et al.*]

Pose guided Audio-Driven Gesture Synthesis

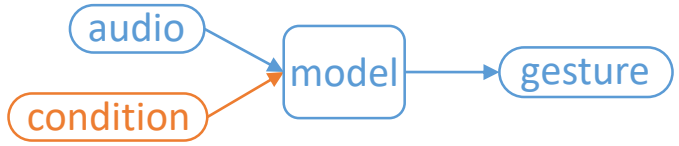


Zhi et al, Speech Drives Templates: Co-Speech Gesture Synthesis with Learned Templates, ICCV 2021

Motivation



one-to-many mapping



conditional one-to-one mapping

Learning:

- Previous regression-based methods suffers the underfitting issue
- Our solution introduces the conditions to relieve ambiguity.

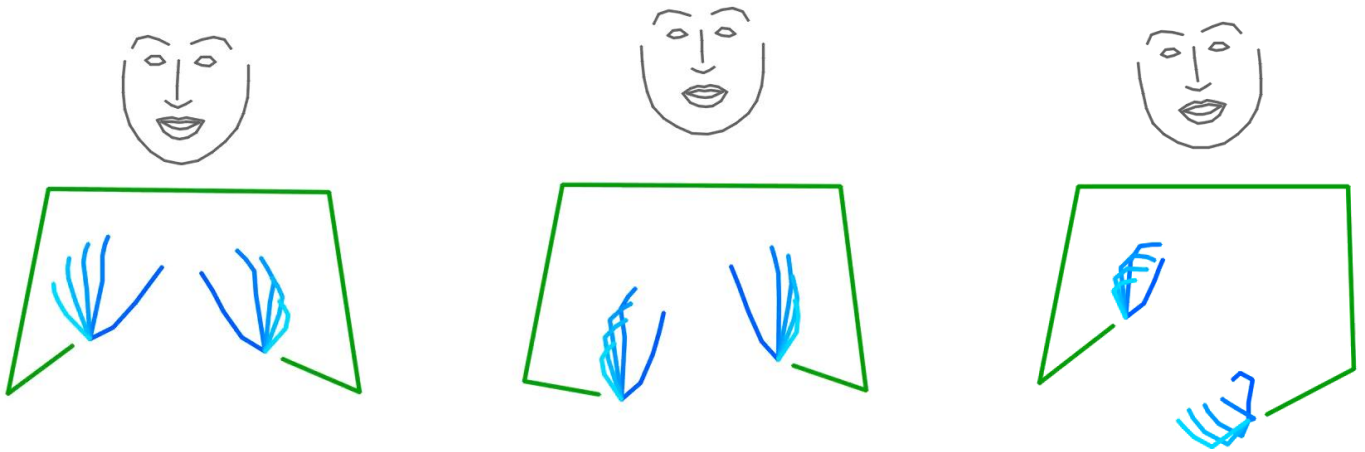
Evaluation:

- L2 distance is not suitable for the evaluation of the one-to-many mapping
- Use distribution distance instead of point-wise distance to measure fidelity.
- We propose a lip-audio synchronization as a metric for synchronization evaluation.

Speech Drives Templates

speech
Audio
(input)

generated
Gestures
(output)



template
Vectors
(input)

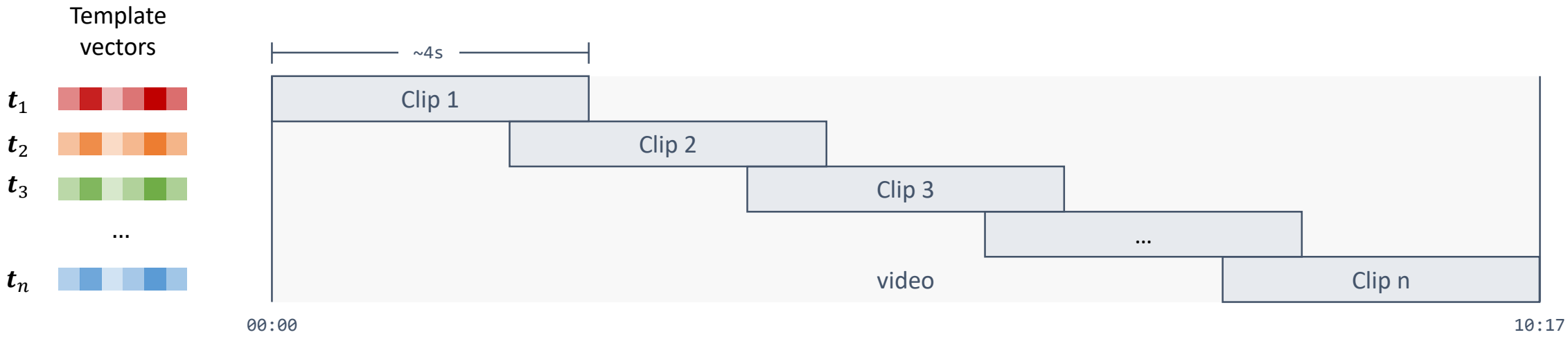
t_1

t_2

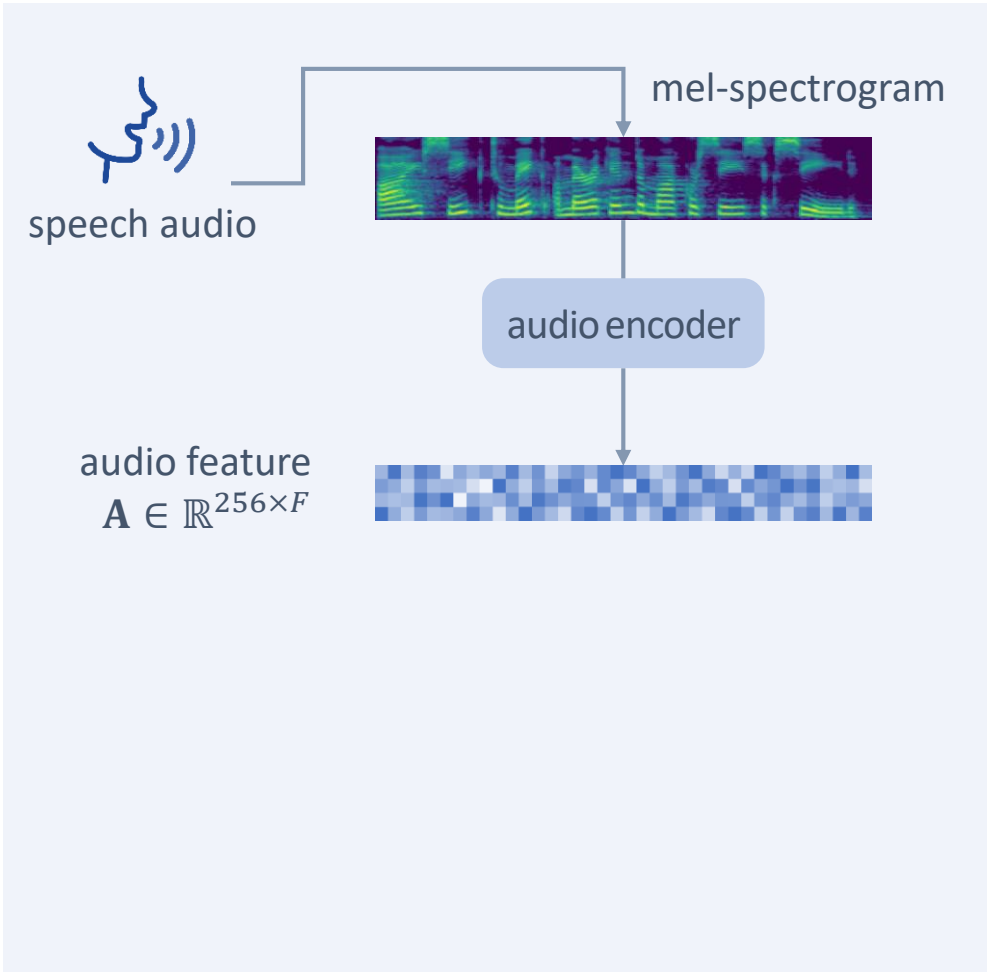
t_3

Template Vector Learning

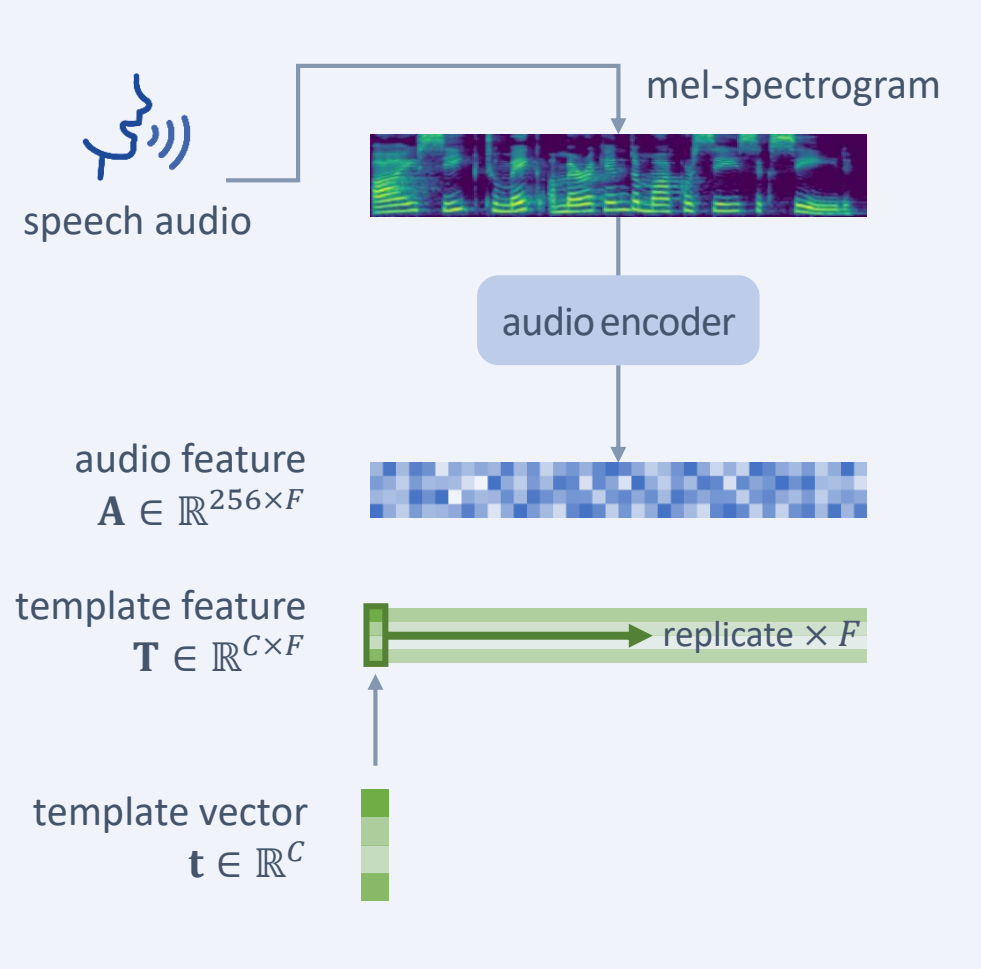
1. BP: optimize template vectors with the back-propagated gradients of the regression loss.
2. VAE: train a VAE to reconstruct all gesture clips and take the encoding of each clip as its template vector.



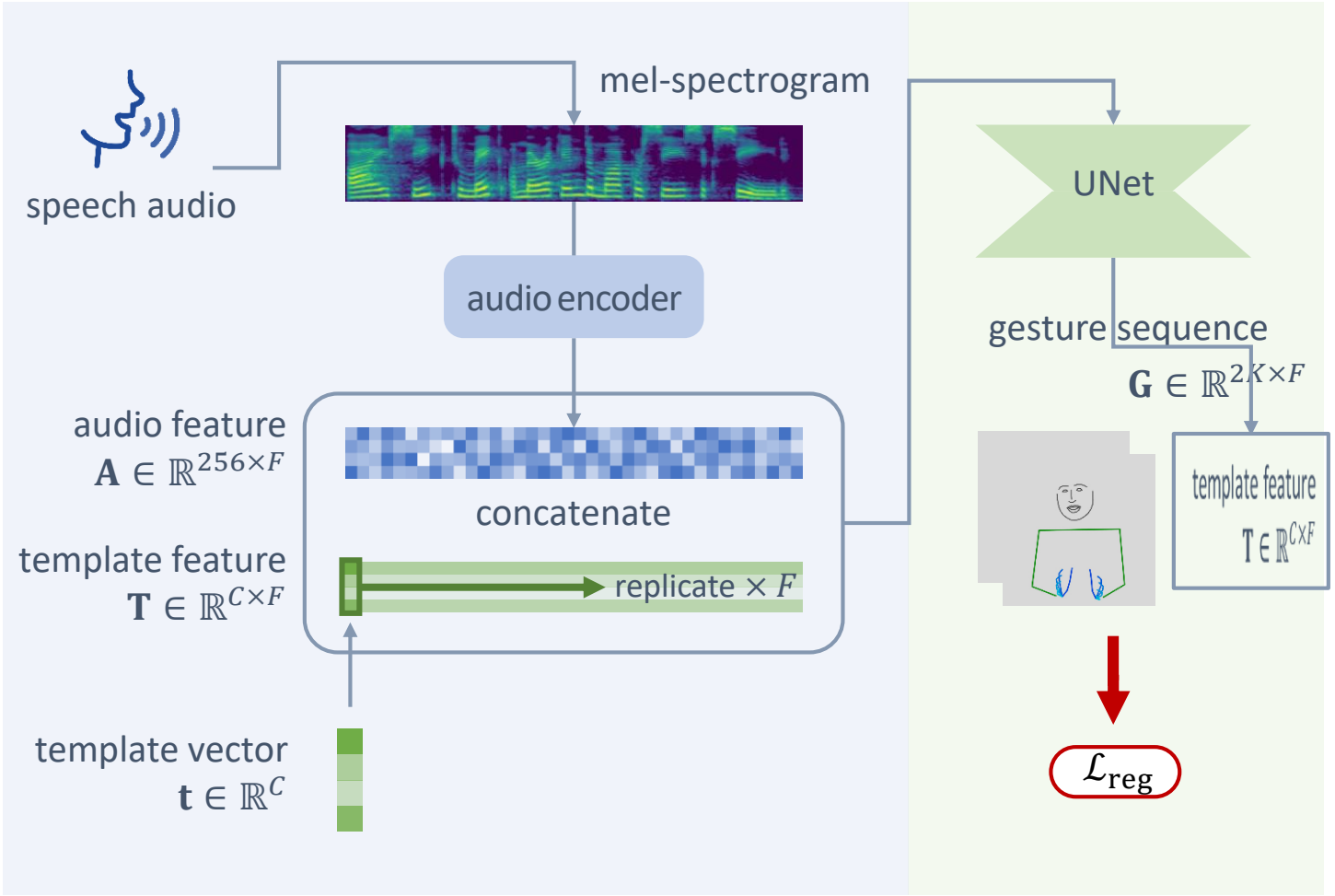
Pipeline(skeleton generation)



Pipeline(skeleton generation)

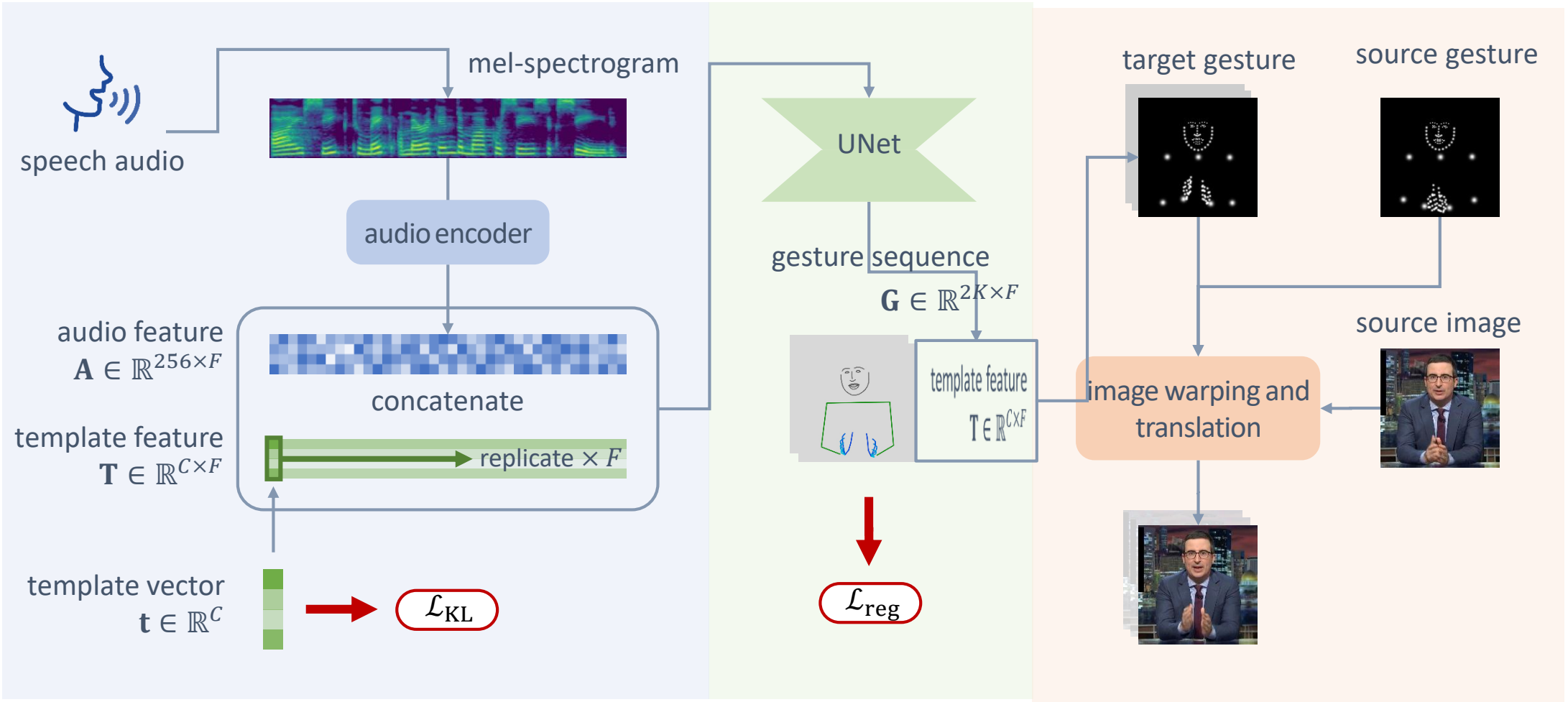


Pipeline (skeleton generation)



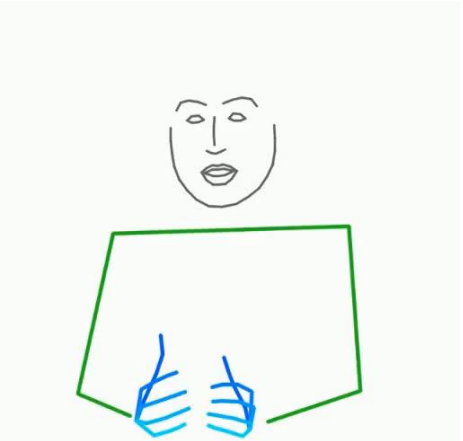
$$\mathcal{L}_{\text{reg}} = \frac{1}{F}\sum_{i=1}^{F} \|\mathbf{G}^{(i)} - \hat{\mathbf{G}}^{(i)}\|_1$$

Pipeline (image synthesis)

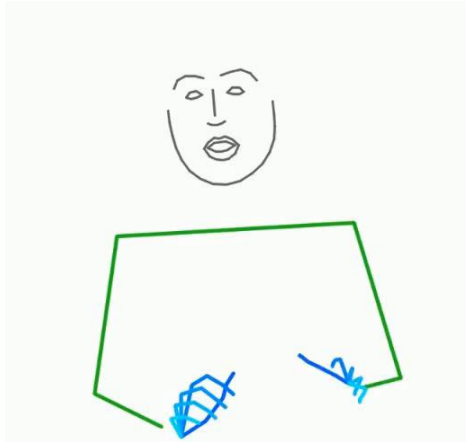


Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8340– 8348, 2018

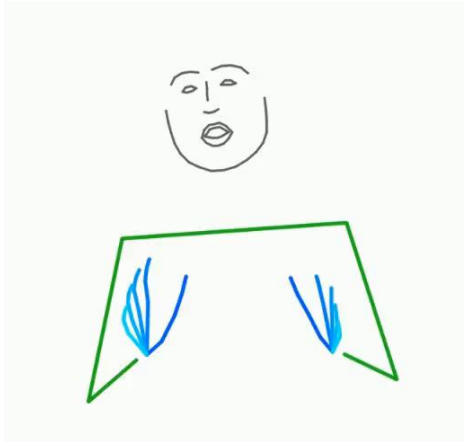
Different Templates Driven by the Same Speech Audio



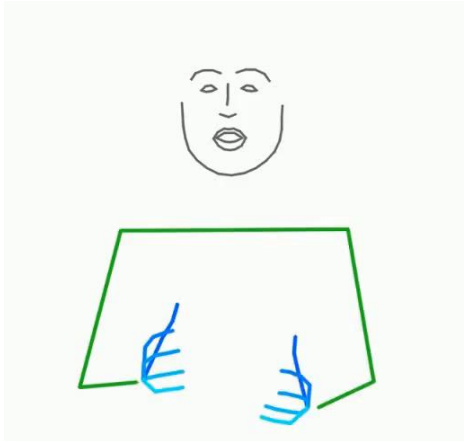
- Template A



- Template B



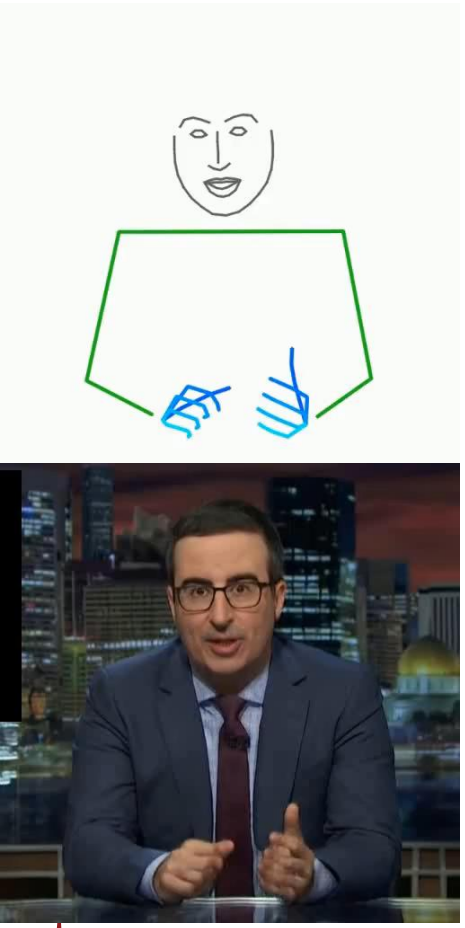
- Template C



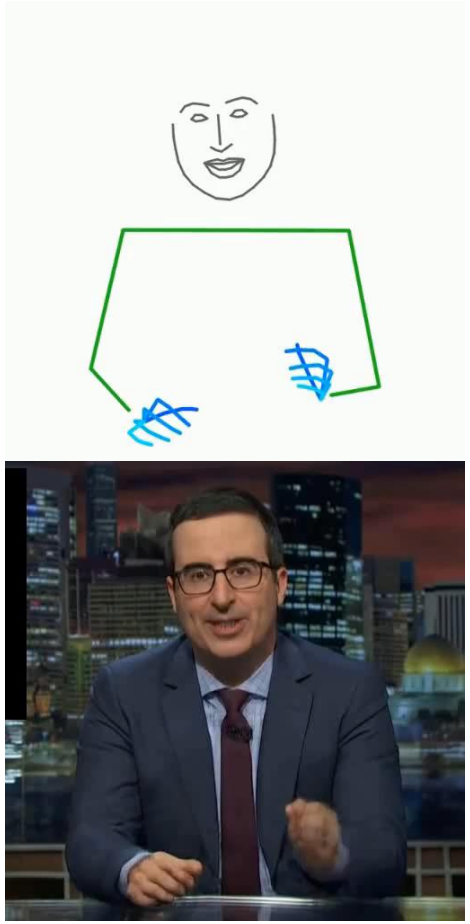
- Template D

right hand | left hand | altering hands | both hands

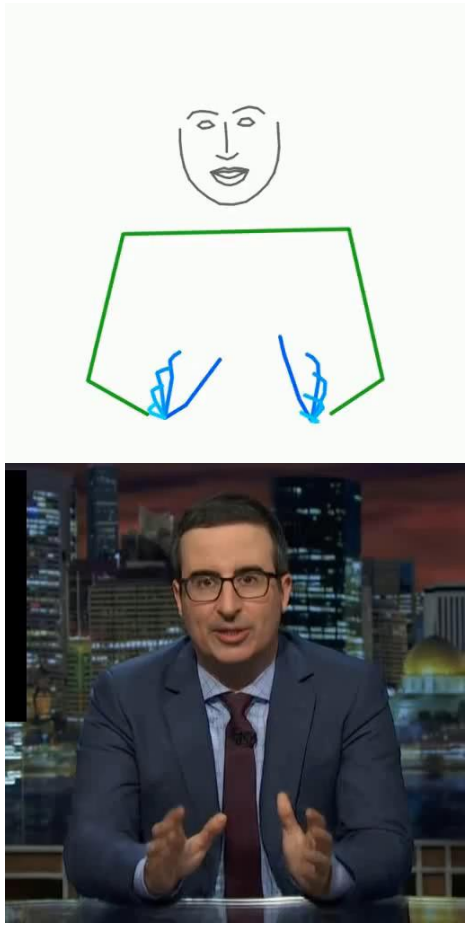
The Same Template Driven by Different Audio Clips



- Audio A



- Audio B

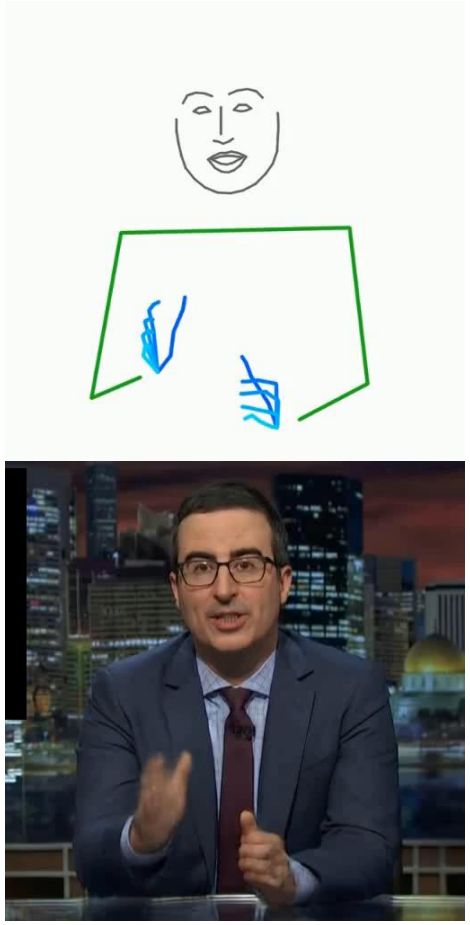


- Audio C

- Audio D

Comparison with Baselines (Oliver)

- Audio2Body
- [Shlizerman et al.]

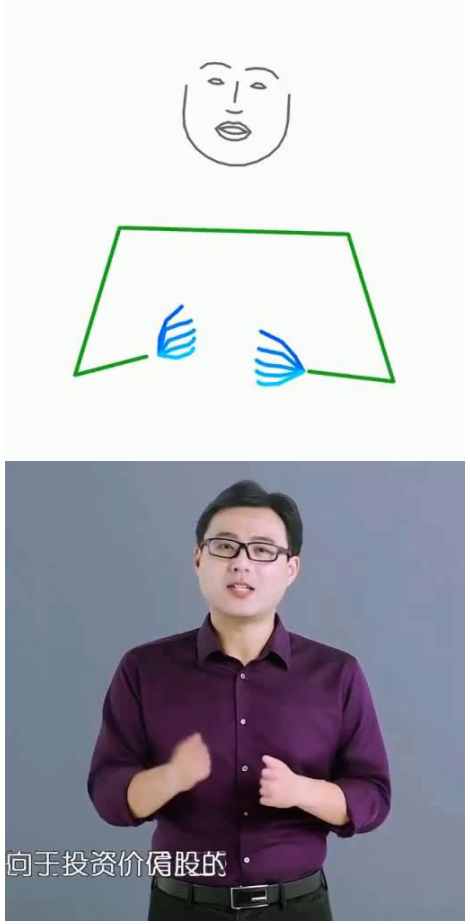


- Speech2Gesture
- [Ginosar et al.]

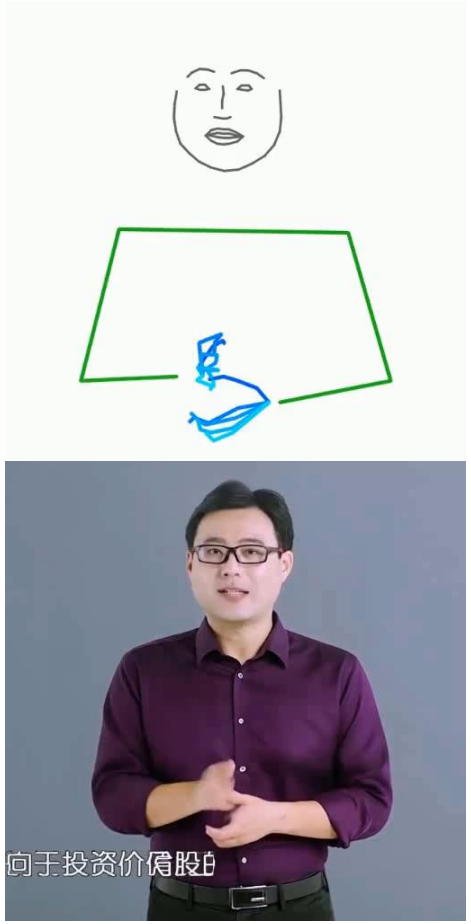
- MoGlow
- [Alexanderson et al.]

- **Ours**

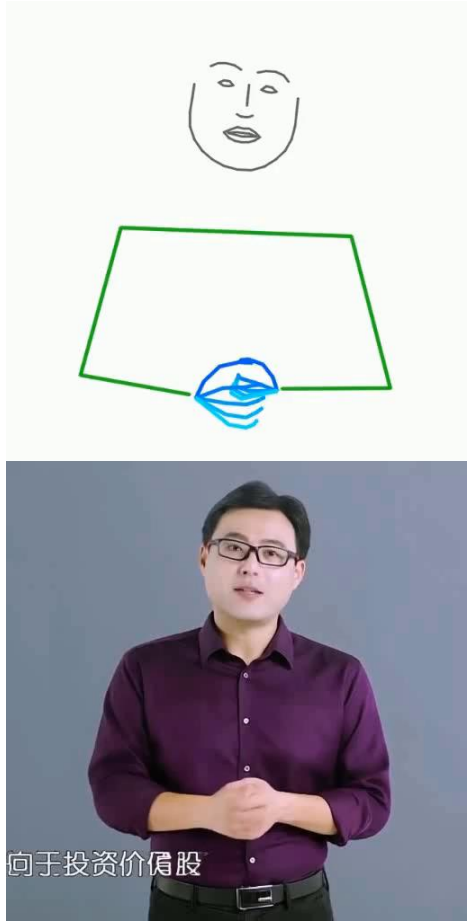
Comparison with Baselines (Xing)



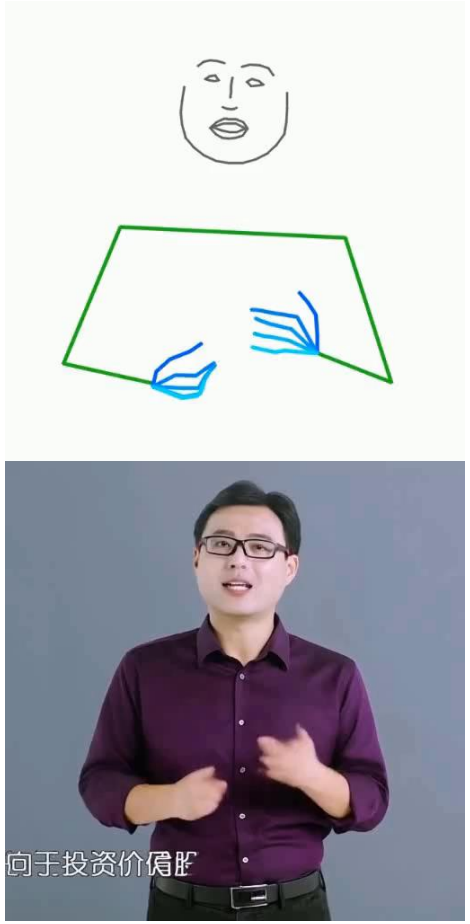
- Audio2Body
- [Shlizerman et al.]



- Speech2Gesture
- [Ginosar et al.]



- MoGlow
- [Alexanderson et al.]



- **Ours**

Structural priors facilitate scene novel view synthesis

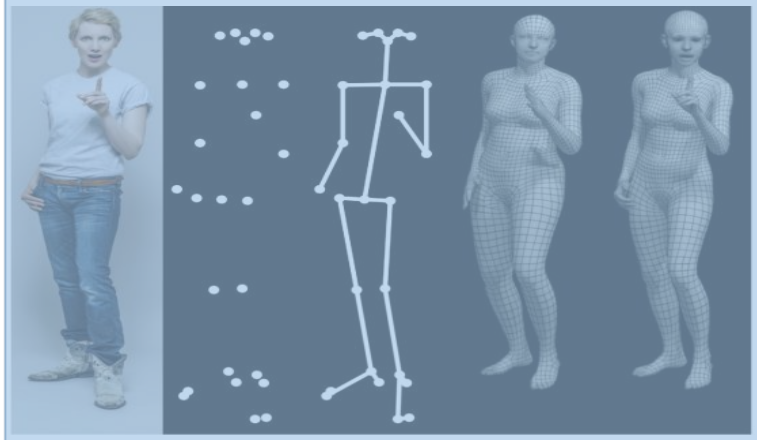
上海科技大学 ShanghaiTech University

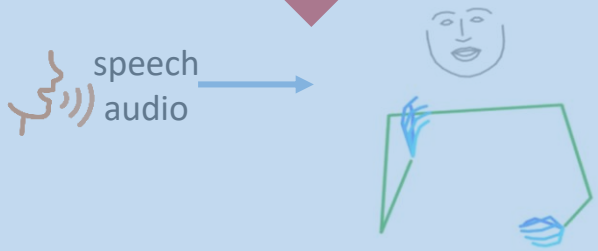
Semantic parsing results

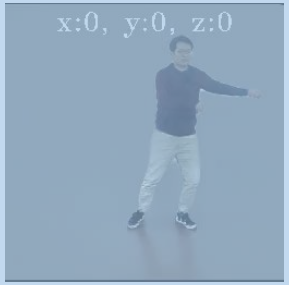
Input

Output

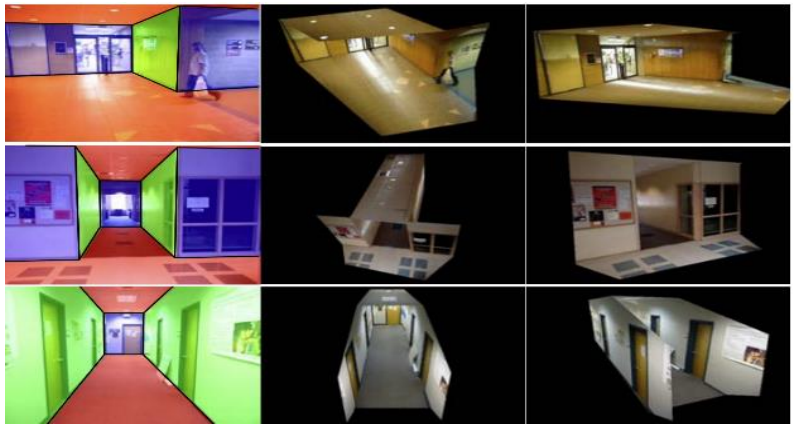
Human shape and pose







Room layout



立志成才 报国裕民

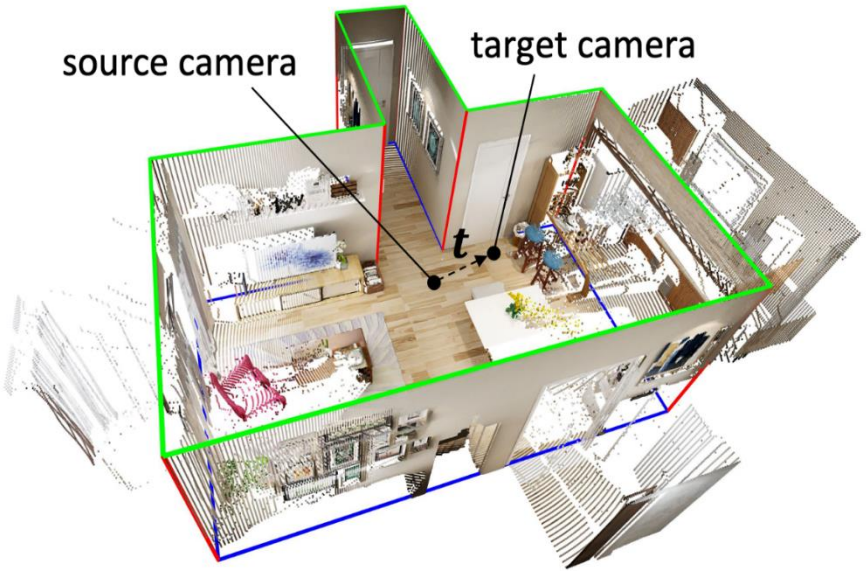
Layout Guided Novel View Synthesis

➢Task

- Panoramic novel view synthesis from a single indoor panorama.

➢Applications

- Virtual Reality (VR), such as virtual house tour.
- Provide a 6-DoF scene viewing experience.

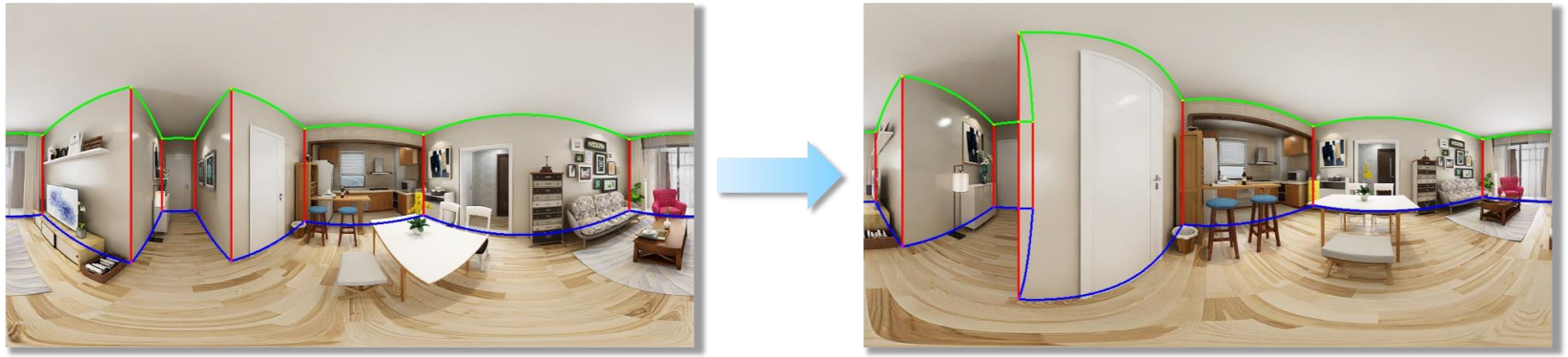


source view

target view

Challenges

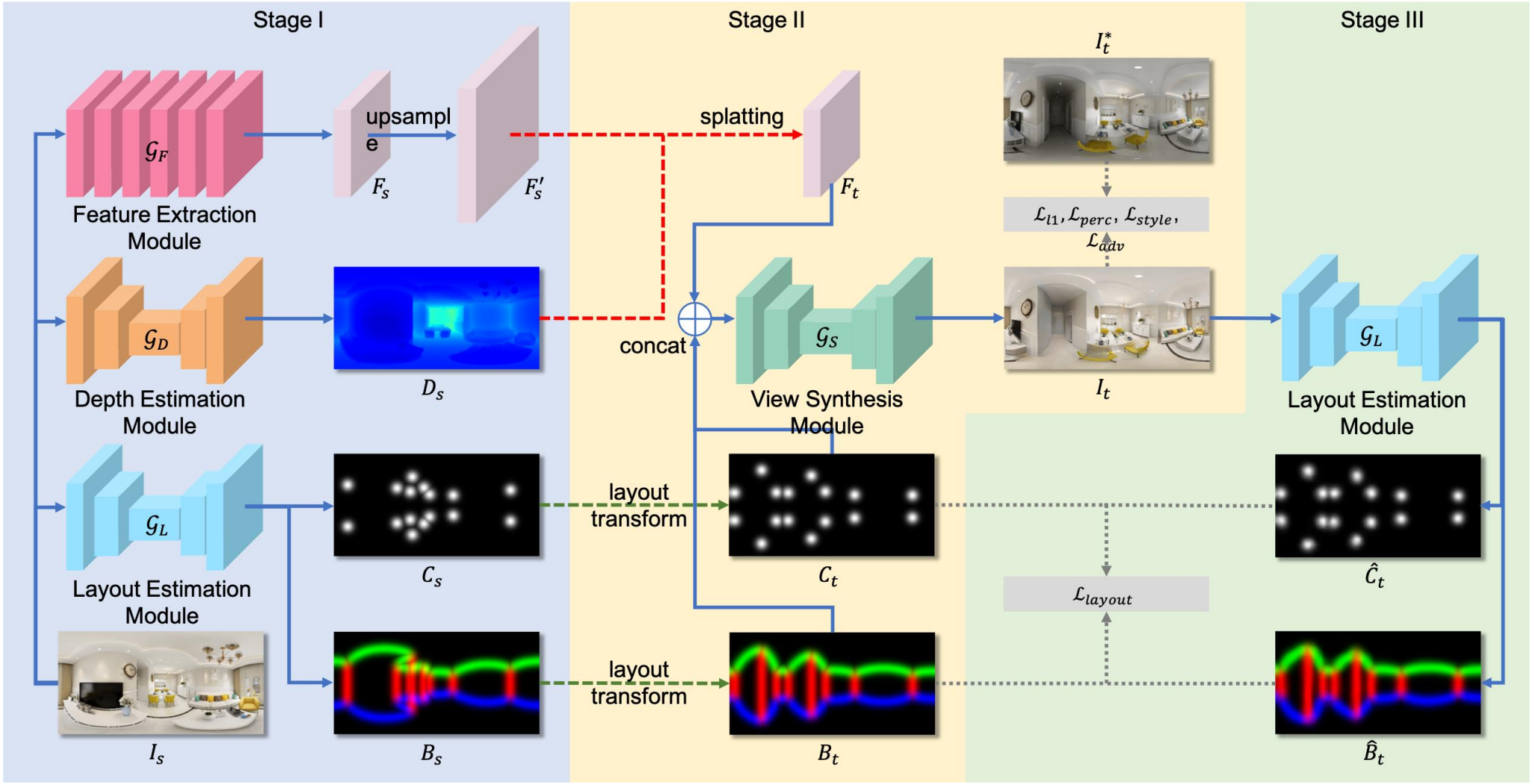
- Previous novel view synthesis work often considers camera translation from 0.2m to 0.3m.
- We consider large camera translations from 1.0m to 2.0m.
- The contents of panoramas are more complex than perspective images.



source view

target view

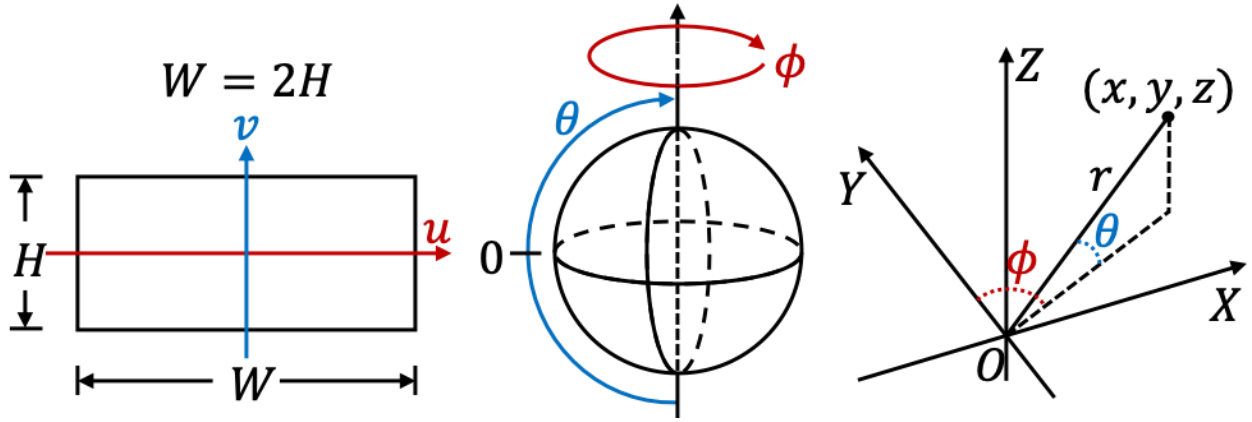
Method



Overview of our proposed method.

Method

➢How to transform between views?



- Three types of coordinate systems:
  - Panoramic pixel grid coordinate system $\mathcal{P}$
  - Spherical polar coordinate system $\mathcal{S}$
  - 3D Cartesian camera coordinate system $\mathcal{C}$
- View transformation process: from $\mathcal{P}_s$ to $\mathcal{P}_t$

$$g = g_{\mathcal{S}_t \mapsto \mathcal{P}_t} \circ g_{\mathcal{C}_t \mapsto \mathcal{S}_t} \circ g_{\mathcal{C}_s \mapsto \mathcal{C}_t} \circ g_{\mathcal{S}_s \mapsto \mathcal{C}_s} \circ g_{\mathcal{P}_s \mapsto \mathcal{S}_s}$$

Experiments

Qualitative results on our dataset.

Experiments

Source View | Ours (without layout) | Ours (with layout) | Target View (Ground Truth)

The influence of room layout guidance

Layout-Guided Novel View Synthesis from a Single Indoor Panorama

Jiale Xu, Jia Zheng, Yanyu Xu, Rui Tang, Shenghua Gao

CVPR 2021

Summary

- Structural priors: human shape, room layout, template, etc.
- How to leverage priors for more realistic image/video generation.

Acknowledgements

My collaborator: Shenghan Qian, Jiale Xu, Yihao Zhi, Zhixin Piao, and Wen Liu, Zhi Tu, Zehao Yu, Lei Jin, Yanyu Xu, Jia Zheng, …

Organizations:
Kujiale, Tencent, Alibaba, …

Thank You!

Email:gaoshh@shanghaitech.edu.cn