ID: THU-PM-053

https://osx-ubody.github.io

One-Stage 3D Whole-body Mesh Recovery with Component Aware Transformer

Jing Lin[1,2], Ailing Zeng[1] , Haoqian Wang[2] , Lei Zhang[1] , Yu Li[1]

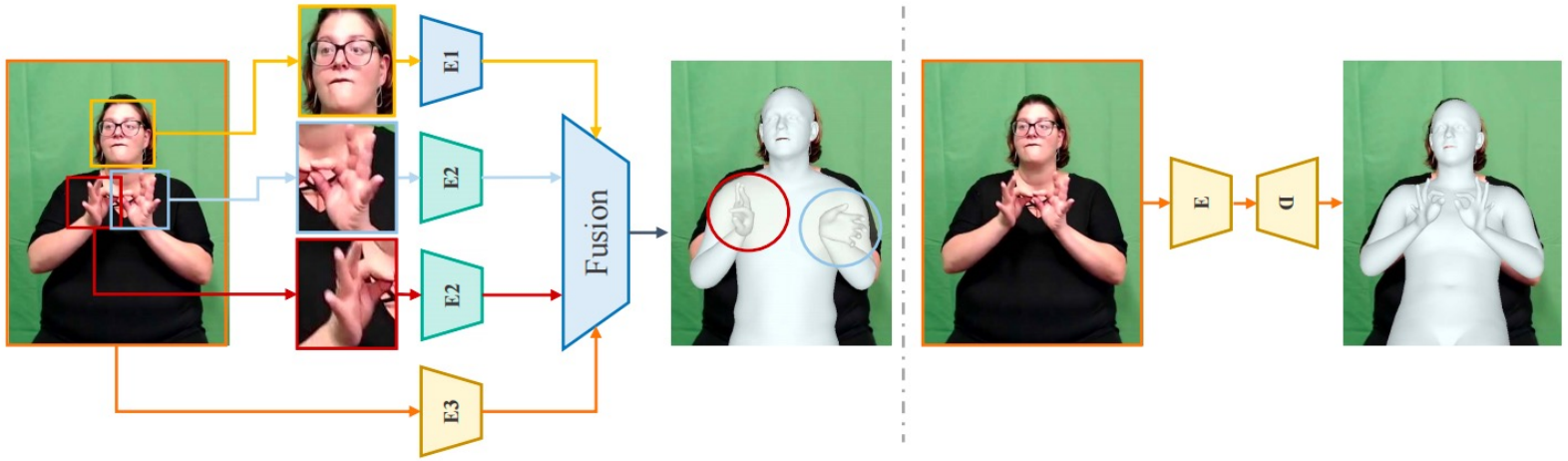
[1]International Digital Economy Academy (IDEA),
[2]The Shenzhen International Graduate School, Tsinghua University

idea INTERNATIONAL DIGITAL ECONOMY ACADEMY
粤港澳大湾区数字经济研究院

清華大學 Tsinghua University

Overview

➢ The first **one-stage** pipeline for expressive 3D whole-body mesh recovery (e.g., SMPLX).



(a) Previous multi-stage pipeline

(b) Our one-stage pipeline

Overview

➢ The first **one-stage** pipeline for Expressive 3D whole-body mesh recovery (e.g., SMPLX).

➢ A large-scale **upper-body dataset**, UBody, bridges the gap between the basic task and downstream applications.

Overview

➢ The first **one-stage** pipeline for Expressive 3D whole-body mesh recovery (e.g., SMPLX).

➢ A large-scale **upper-body dataset**, UBody, bridges the gap between the basic task and downstream applications.

➢ **New state-of-the-art** performance on three datasets.

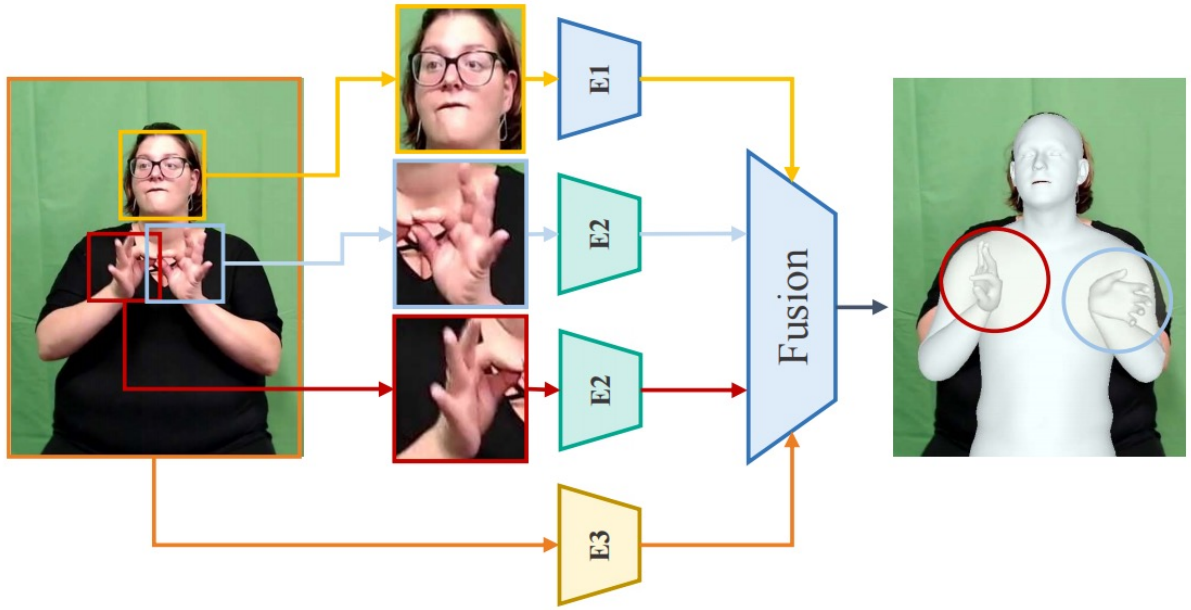
Method	AGORA-test					EHF						3DPW	
	MPVPE ↓			N-MPVPE ↓		MPVPE ↓			PA-MPVPE ↓			MPJPE ↓	PA-MPJPE ↓
	All	Hands	Face	All	Body	All	Hands	Face	All	Hands	Face	Body	Body
ExPose [36]	217.3	73.1	51.1	265.0	184.8	77.1	51.6	35.0	54.5	12.8	5.8	93.4	60.7
FrankMocap [41]	-	55.2	-	-	207.8	107.6	42.8	-	57.5	12.6	-	96.7	61.9
PIXIE [13]	191.8	49.3	50.2	233.9	173.4	89.2	42.8	32.7	55.0	11.1	**4.6**	91.0	61.3
Hand4Whole [29]	-	-	-	-	-	79.2	43.2	**25.0**	53.1	12.1	5.8	-	-
Hand4Whole [29]×	135.5	47.2	41.6	144.1	96.0	76.8	**39.8**	26.1	**50.3**	**10.8**	5.8	86.6	54.4
OSX (Ours)	**122.8**$\downarrow_{9.5\%}$	**45.7**	**36.2**	**130.6**	**85.3**	**70.8**$\downarrow_{7.8\%}$	53.7	26.4	48.7	15.9	6.0	**74.7**$\downarrow_{13.4\%}$	**45.1**

Background

Existing Problems

➢ **Multi-stage methods:** use three separate network for the body pose, facial expression and hand pose estimation.

heavy computational cost and unnatural connection

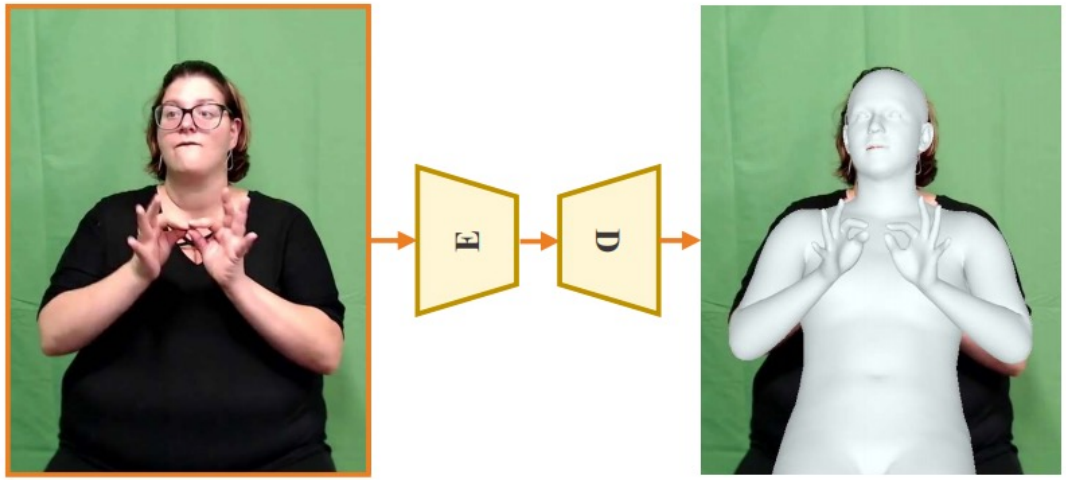


(a) Previous multi-stage pipeline

Existing Problems

➢ **Multi-stage methods:** use three separate network for the body pose, facial expression and hand pose estimation.

One Stage Method?

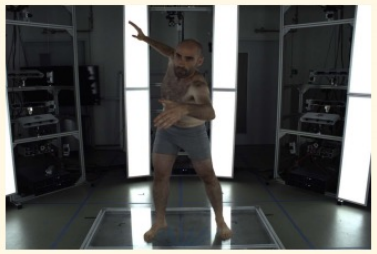


Existing Problems

➢ **Data discrepancy:** Existing datasets are whole-body scenes. But in many daily life scenes, upper body is a major focus.

Model can not perform well in real-life scenes


(a) AGORA


(b) EHF


(c) Multi-Shot-AVA


(d) MSCOCO


(e) 3DPW

Existing Datasets

Existing Problems

➢ **Data discrepancy:** Existing datasets are whole-body scenes. But in many daily life scenes, upper body is a major focus.

Upper-Body Dataset?

(f) Conduct Music

(g) Talk Show

(h) Entertainment

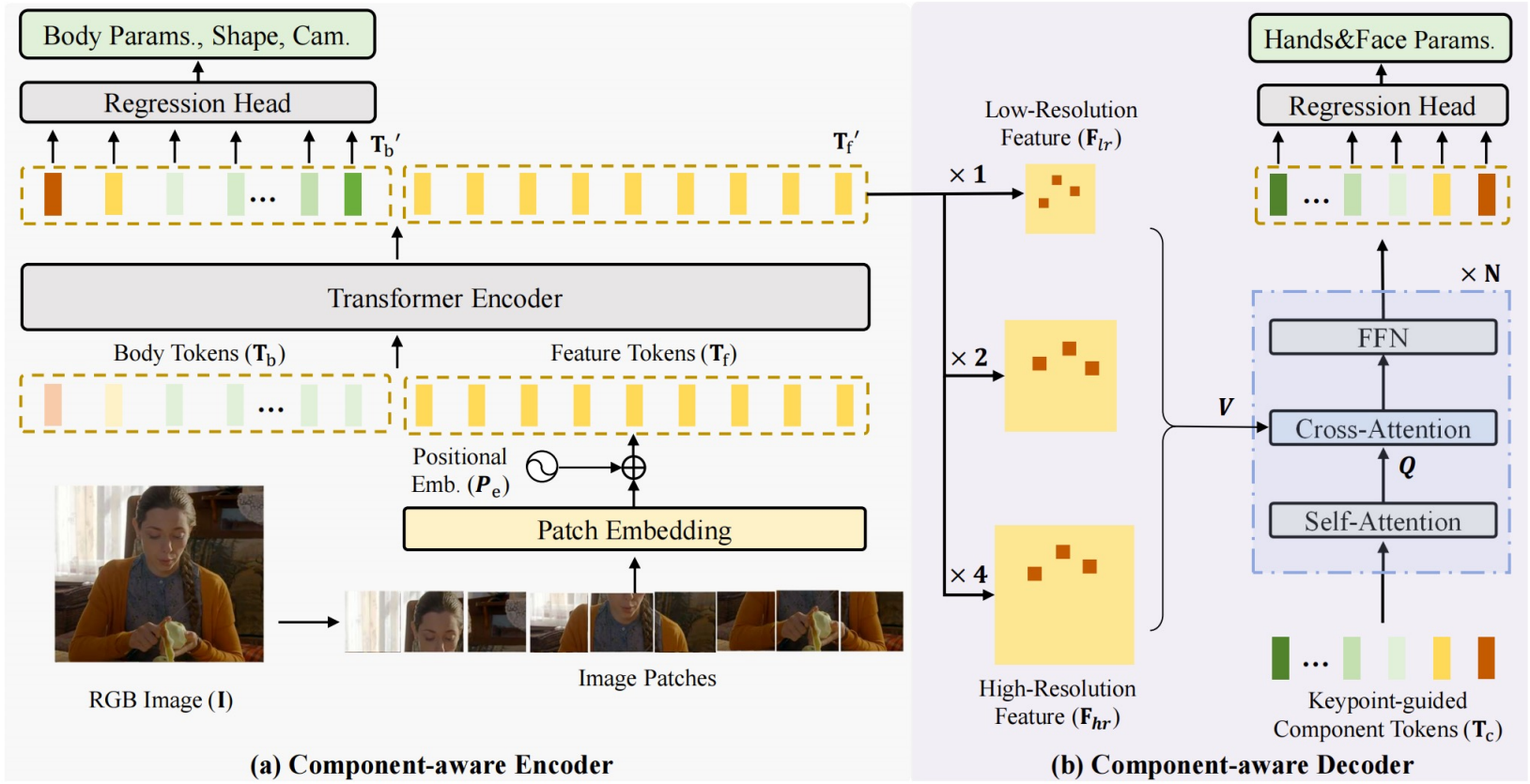
(i) Sign Language

(j) Magic Show

Real-life Scenes

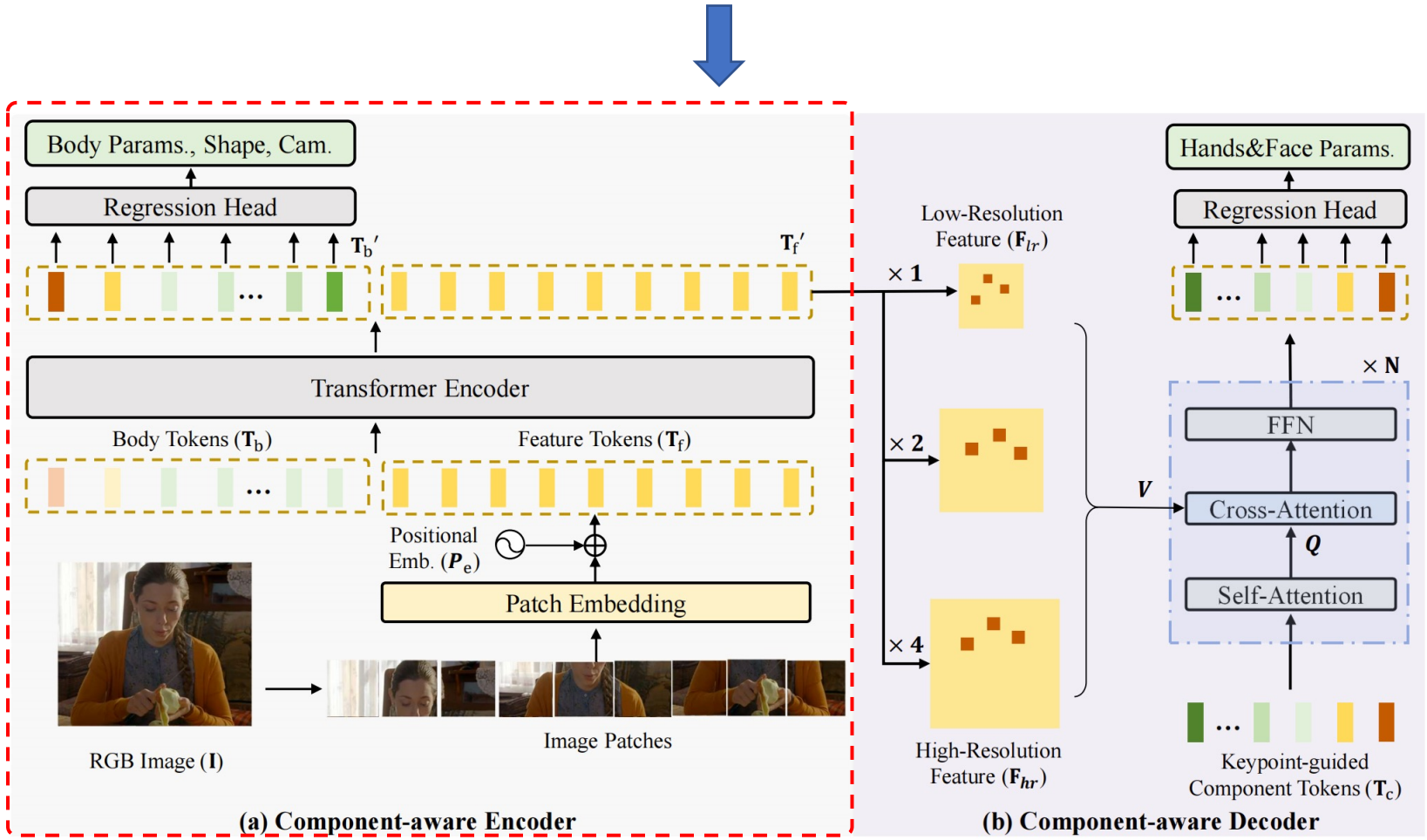
Methods

Body pose estimation needs the **global** dependencies while the **facial** expression and **hand** pose concentrate more on the **local** feature.



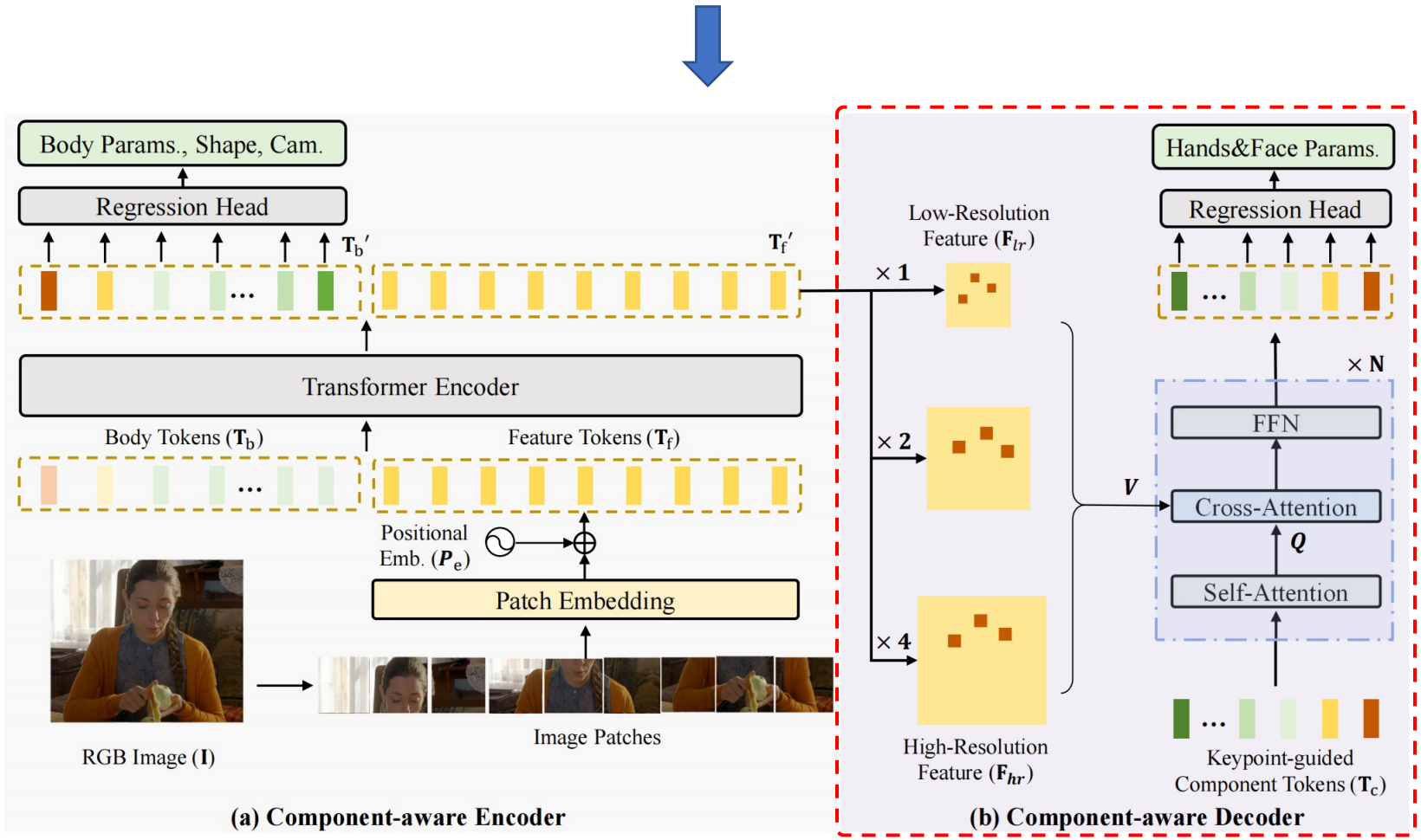
Methods

Body pose estimation needs the **global** dependencies while the **facial** expression and **hand** pose concentrate more on the **local** feature.



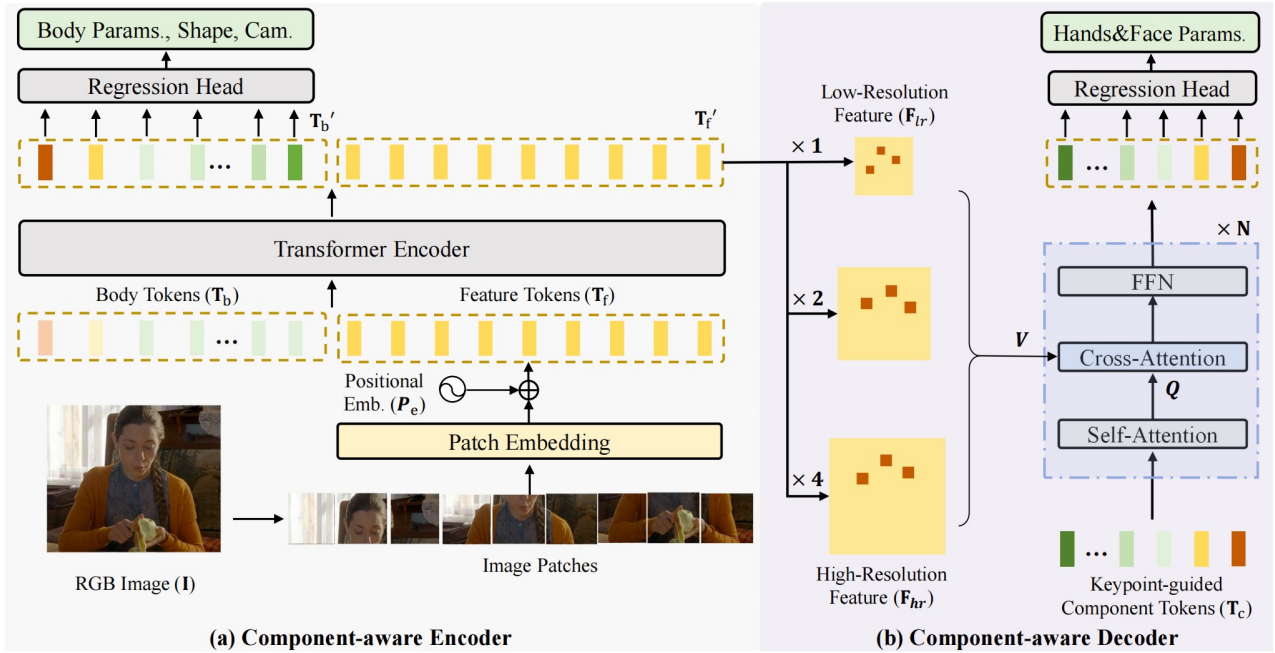
Methods

Body pose estimation needs the **global** dependencies while the **facial** expression and **hand** pose concentrate more on the **local** feature.



Methods

Body pose estimation needs the **global** dependencies while the **facial** expression and **hand** pose concentrate more on the **local** feature.



Cross-Attention: $\mathrm{CA}(\mathbf{Q}, \mathbf{V}, p_q) = \sum_{l=1}^{L} \sum_{k=1}^{K} A_{lqk} W \mathbf{V}_l(\phi_l(p_q) + \Delta p_{lqk}),$

Upper-Body Dataset

➢ Dataset Characteristics

1. **Upper-body**: with partial observation and heavy truncation
2. **Expressive:** rich hand gestures and facial expressions
3. **Large-scale:** fifteen real-life scenarios, more than **1M** frames
4. **Challenging:** severe truncation, dynamic camera view, etc.

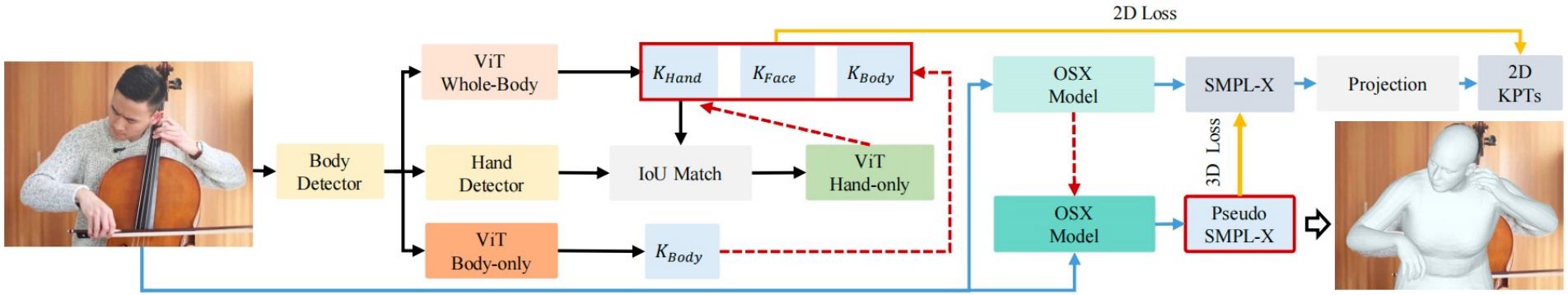
Upper-Body Dataset

➢ 15 Real-Life Scenes

UBody contains 15 human-centric real-life scenes, which mainly focus on the upper-body parts.

Upper-Body Dataset

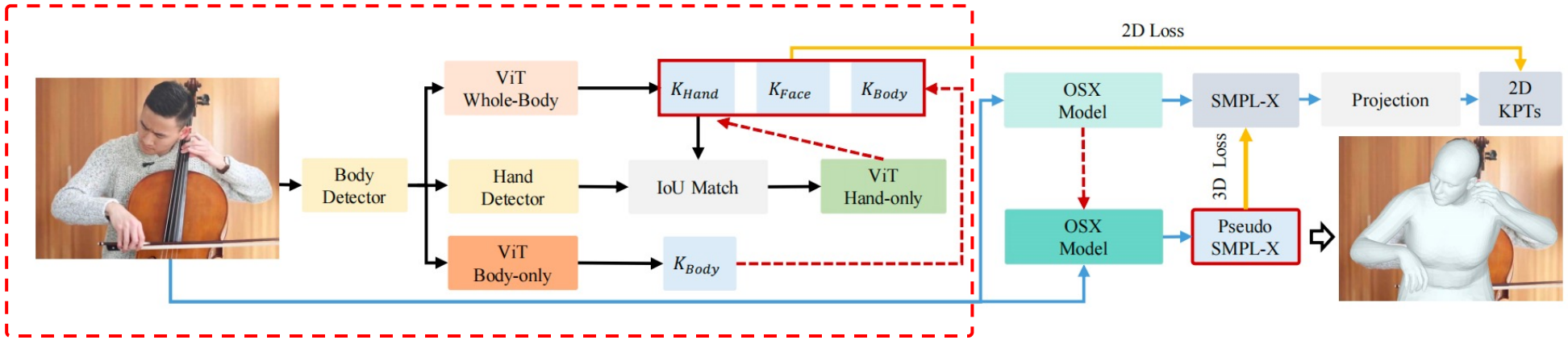
➢ Annotation Pipeline



We annotate 2d whole-body keypoints and 3d whole-body mesh with an automatic annotation pipeline.

Upper-Body Dataset

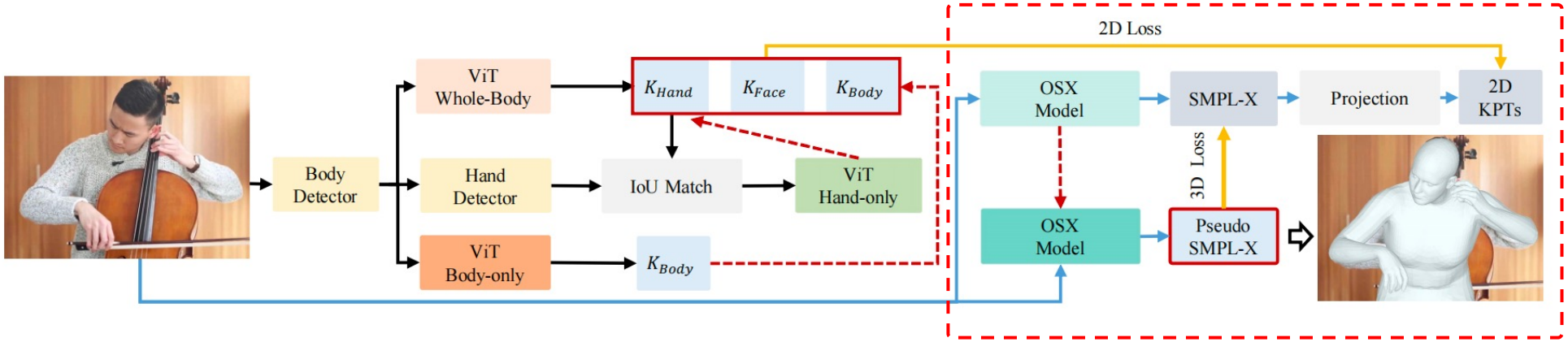
➢ Annotation Pipeline



2D whole-body **keypoints** annotation: BodyHands detects bounding box, 4 ViT-based models predict the whole-body, body, face and hand keypoints.

Upper-Body Dataset

➢ Annotation Pipeline



3D SMPLX fitting: Iterative training-labeling-revision loop to fit the projected 2D keypoints into the annotated 2D keypoints.

Upper-Body Dataset

➢ **Annotation Visualization**

Experiment

➢ Comparison with existing methods on three datasets.

Method	AGORA-test					EHF						3DPW	
	MPVPE ↓			N-MPVPE ↓		MPVPE ↓			PA-MPVPE ↓			MPJPE ↓	PA-MPJPE ↓
	All	Hands	Face	All	Body	All	Hands	Face	All	Hands	Face	Body	Body
ExPose [48]	217.3	73.1	51.1	265.0	184.8	77.1	51.6	35.0	54.5	12.8	5.8	93.4	60.7
FrankMocap [56]	-	55.2	-	-	207.8	107.6	42.8	-	57.5	12.6	-	96.7	61.9
PIXIE [16]	191.8	49.3	50.2	233.9	173.4	89.2	42.8	32.7	55.0	11.1	**4.6**	91.0	61.3
Hand4Whole [39]	-	-	-	-	-	79.2	43.2	**25.0**	53.1	12.1	5.8	-	-
Hand4Whole [39]×	135.5	47.2	41.6	144.1	96.0	76.8	**39.8**	26.1	**50.3**	**10.8**	5.8	86.6	54.4
OSX (Ours)	**122.8**↓9.5%	**45.7**	**36.2**	**130.6**	**85.3**	**70.8**↓7.8%	53.7	26.4	48.7	15.9	6.0	**74.7**↓13.4%	**45.1**

Experiment

➢ Comparison with existing methods on three datasets.

Method	AGORA-test					EHF						3DPW	
	MPVPE ↓			N-MPVPE ↓		MPVPE ↓			PA-MPVPE ↓			MPJPE ↓	PA-MPJPE ↓
	All	Hands	Face	All	Body	All	Hands	Face	All	Hands	Face	Body	Body
ExPose [48]	217.3	73.1	51.1	265.0	184.8	77.1	51.6	35.0	54.5	12.8	5.8	93.4	60.7
FrankMocap [56]	-	55.2	-	-	207.8	107.6	42.8	-	57.5	12.6	-	96.7	61.9
PIXIE [16]	191.8	49.3	50.2	233.9	173.4	89.2	42.8	32.7	55.0	11.1	**4.6**	91.0	61.3
Hand4Whole [39]	-	-	-	-	-	79.2	43.2	**25.0**	53.1	12.1	5.8	-	-
Hand4Whole [39]×	135.5	47.2	41.6	144.1	96.0	76.8	**39.8**	26.1	**50.3**	**10.8**	5.8	86.6	54.4
OSX (Ours)	**122.8**↓9.5%	**45.7**	**36.2**	**130.6**	**85.3**	**70.8**↓7.8%	53.7	26.4	48.7	15.9	6.0	**74.7**↓13.4%	**45.1**

➢ Benchmark on the proposed UBody datasets.

Method	MPVPE ↓			PA-MPVPE ↓			PA-MPJPE ↓	
	All	Hand	Face	All	Hand	Face	Body	Hand
ExPose [48]	171.5	83.7	45.1	66.9	12.0	3.9	70.7	12.3
PIXIE [16]	168.4	55.6	45.2	61.7	12.2	4.2	66.8	12.3
Hand4Whole [39]	104.1	45.7	27.0	44.8	8.9	2.8	45.5	9.0
Hand4Whole [39]×	157.4	62.2	49.8	82.2	9.8	3.9	92.8	10.0
OSX (Ours)	92.4	47.7	24.9	42.4	10.8	2.4	42.9	11.0
OSX (Ours)†	**81.9**	**41.5**	**21.2**	**42.2**	**8.6**	**2.0**	**48.4**	**8.8**

Experiment

➢ Comparison with existing methods on three datasets.

Method	AGORA-test					EHF						3DPW	
	MPVPE ↓			N-MPVPE ↓		MPVPE ↓			PA-MPVPE ↓			MPJPE ↓	PA-MPJPE ↓
	All	Hands	Face	All	Body	All	Hands	Face	All	Hands	Face	Body	Body
ExPose [48]	217.3	73.1	51.1	265.0	184.8	77.1	51.6	35.0	54.5	12.8	5.8	93.4	60.7
FrankMocap [56]	-	55.2	-	-	207.8	107.6	42.8	-	57.5	12.6	-	96.7	61.9
PIXIE [16]	191.8	49.3	50.2	233.9	173.4	89.2	42.8	32.7	55.0	11.1	**4.6**	91.0	61.3
Hand4Whole [39]	-	-	-	-	-	79.2	43.2	**25.0**	53.1	12.1	5.8	-	-
Hand4Whole [39]×	135.5	47.2	41.6	144.1	96.0	76.8	**39.8**	26.1	**50.3**	**10.8**	5.8	86.6	54.4
OSX (Ours)	**122.8**↓9.5%	**45.7**	**36.2**	**130.6**	**85.3**	**70.8**↓7.8%	53.7	26.4	48.7	15.9	6.0	**74.7**↓13.4%	**45.1**

➢ Benchmark on the proposed UBody datasets.

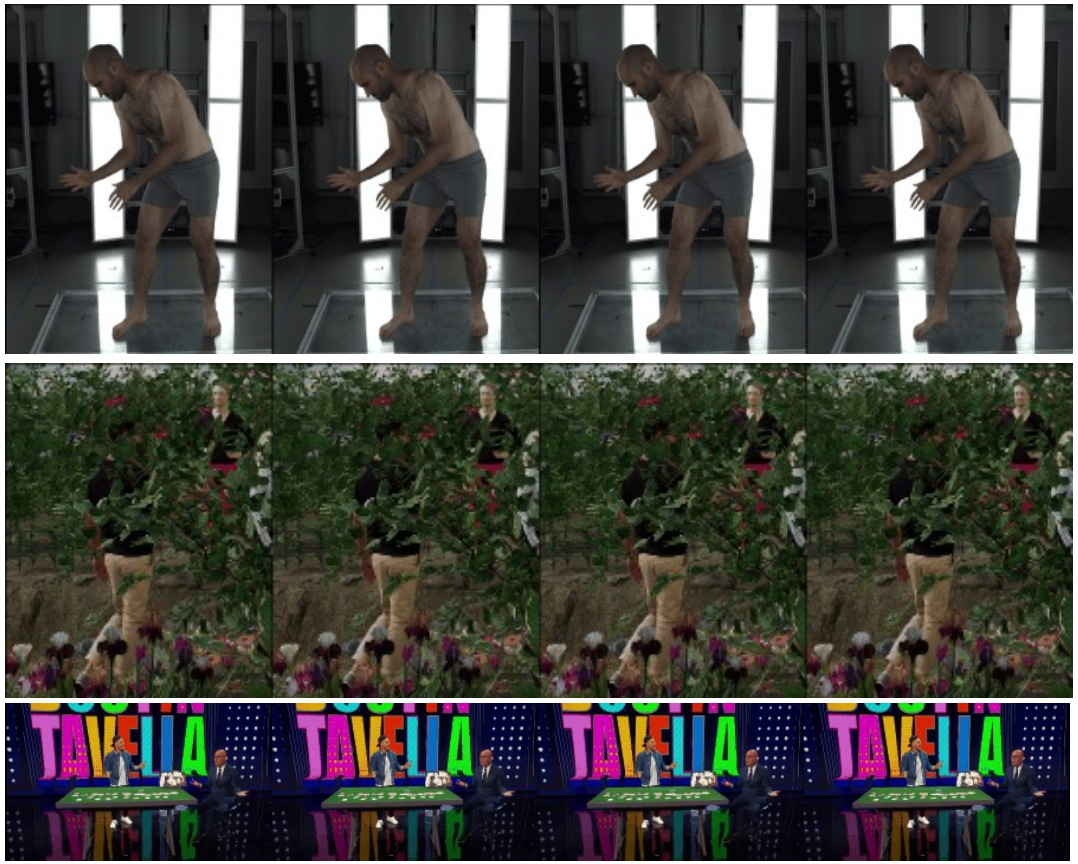
Method	MPVPE ↓			PA-MPVPE ↓			PA-MPJPE ↓	
	All	Hand	Face	All	Hand	Face	Body	Hand
ExPose [48]	171.5	83.7	45.1	66.9	12.0	3.9	70.7	12.3
PIXIE [16]	168.4	55.6	45.2	61.7	12.2	4.2	66.8	12.3
Hand4Whole [39]	104.1	45.7	27.0	44.8	8.9	2.8	45.5	9.0
Hand4Whole [39]×	157.4	62.2	49.8	82.2	9.8	3.9	92.8	10.0
OSX (Ours)	92.4	47.7	24.9	42.4	10.8	2.4	42.9	11.0
OSX (Ours)†	**81.9**	**41.5**	**21.2**	**42.2**	**8.6**	**2.0**	**38.4**	**8.8**

➢ Ablation studies of the keypoint-guided attention and upsample scale.

Hand	Ours	w/o *H.D.*	w/o *K.G*	w/o both
MPVPE	**53.7**	55.3	55.1	56.4
PA-MPVPE	**15.9**	17.7	17.6	18.1
Face	**Ours**	**w/o *F.D.***	**w/o *K.G***	**w/o both**
MPVPE	**26.4**	27.2	26.4	26.8
PA-MPVPE	6.0	5.9	**5.8**	6.0
Upsampling	**× 1**	**× 2**	**× 4**	**× 8**
MPVPE	54.9	54.3	**53.7**	54.1

Experiment

➢ Visual Comparison with existing methods on three datasets.



Input Image ExPose Hand4Whole Ours

Thanks

Code & Paper & Data