

SIAM: Synchronous Interaction Attention for Human Mesh Recovery (Supplementary)

Niaz Ahmad¹ Saif Ullah² Youngmoon Lee² Guanghui Wang¹

¹Toronto Metropolitan University, ²Hanyang University

{niazahmad89}{wangcs}@torontomu.ca {fahad7878}{youngmoonlee}@hanyang.ac.kr

The supplementary document includes the further description of Spatio-temporal pooling function, real-time analysis, limitations, future work, and the code related to this study.

1. Spatio-temporal Pooling Function

Design and operation of $\mathcal{P}_{\text{sync}}$. Let $R_{n+t'}^{(i)} = (x_{t'}^{(i)}, y_{t'}^{(i)})$ be the center of instance i (from the IIM module) and $s_{t'}^{(i)}$ its scale (estimated from keypoints or center-heatmap spread). $\mathcal{P}_{\text{sync}}$ proceeds in three steps:

(1) Spatial pooling (per frame). We crop a local region around $R_{n+t'}^{(i)}$ using RoIAlign with a square box

$$B_{t'}^{(i)} = \left[x_{t'}^{(i)} - \frac{\kappa s_{t'}^{(i)}}{2}, x_{t'}^{(i)} + \frac{\kappa s_{t'}^{(i)}}{2} \right] \times \left[y_{t'}^{(i)} - \frac{\kappa s_{t'}^{(i)}}{2}, y_{t'}^{(i)} + \frac{\kappa s_{t'}^{(i)}}{2} \right],$$

resampled to a $K \times K$ grid (we use $K=7$ and $\kappa=2.0$). The pooled patch is passed through a 3×3 conv \rightarrow BN \rightarrow ReLU and global average pooling (GAP) to yield a frame-level instance descriptor

$$g_{t'}^{(i)} = \text{GAP} \left(\phi \left(\text{RoIAlign} \left(F_{n+t'}, B_{t'}^{(i)} \right) \right) \right) \in \mathbb{R}^d.$$

We apply L_2 -normalization to stabilize subsequent attention:

$$\bar{g}_{t'}^{(i)} = \frac{g_{t'}^{(i)}}{\|g_{t'}^{(i)}\|_2}.$$

(2) Temporal attention (within window). For the query time t , we form keys/values from the window $\mathcal{T} = \{t' \mid t - \delta \leq t' \leq t + \delta\}$:

$$q_t^{(i)} = W_q \bar{g}_t^{(i)}, \quad k_{t'}^{(i)} = W_k \bar{g}_{t'}^{(i)}, \quad v_{t'}^{(i)} = W_v \bar{g}_{t'}^{(i)}.$$

We compute attention scores with an optional visibility bias $b_{t'}^{(i)} \in [0, 1]$ (from detection confidence/occlusion flags):

$$e_{t,t'}^{(i)} = \frac{\langle q_t^{(i)}, k_{t'}^{(i)} \rangle}{\sqrt{d}} + \lambda \text{PE}(t' - t) + \mu \log(\varepsilon + b_{t'}^{(i)}),$$

where $\text{PE}(\cdot)$ is a sinusoidal relative-time encoding, and $\lambda, \mu, \varepsilon$ are constants. We obtain weights

$$\alpha_{t'}^{(i)} = \text{softmax}_{t' \in \mathcal{T}} \left(e_{t,t'}^{(i)} \right),$$

masking invalid frames (no detection) by setting $e_{t,t'}^{(i)} = -\infty$.

(3) Temporal aggregation and projection. The synchronized instance embedding is

$$F_i^{n+t} = \sigma \left(W_o \sum_{t' \in \mathcal{T}} \alpha_{t'}^{(i)} v_{t'}^{(i)} + \gamma \bar{g}_t^{(i)} \right),$$

where γ adds a residual from the current frame and σ denotes LayerNorm. If $\delta=0$ or neighbors are masked, $\mathcal{P}_{\text{sync}}$ reduces to the spatial descriptor of the current frame.

This pooling operation leverages temporal continuity and spatial locality to produce consistent instance-level embeddings across the sequence. It is effective for short-term interactions (e.g., coordinated motion, identity persistence under occlusion) and feeds the SIA module to build a cross-frame interaction graph, where each instance is influenced by intra-frame neighbors and temporally adjacent counterparts—facilitating robust pose estimation and mesh reconstruction in complex video scenarios.

2. Real-time Visuals

To evaluate the real-time capabilities of the proposed system, we perform human mesh recovery using the 3DPW [1] dataset. Visual examples are provided in Figure 4 of the main paper. To better assess the real-time performance of the SIAM model, we convert a set of images from the 3DPW dataset into a sequence of videos and perform a series of tests. The videos are available attached in the supplementary folder and also on our web page. ^{*}

^{*}<https://sites.google.com/view/niazahmad/projects/siam-wacv-26>

3. Limitations and Future Work.

Although SIAM brings new contributions to the field, certain aspects require further investigation in future research. One key challenge lies in decomposing individual instances within the feature space, as the instance interaction attention mechanism introduces strong bonds between the instances; isolating them in the feature space is a challenging task, specifically regressing the keypoint coordinates in dense scenes. In future work, we aim to conduct an in-depth exploration of the Feature Decomposition module to improve the capability of isolating instances in the feature space for better representation of individual features.

References

- [1] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. [1](#)