CoEvoer: Collaborative Evolution Transformer for Upper-Body Expressive Human Pose and Shape Estimation

Yuxiang Zhao[1,2], Wei Huang[1†], Yujie Song[1], Liu Wang[1], Huan Zhao[1]

[1]Shenzhen Campus of Sun Yat-sen University
[2]Alibaba Group
zhaoyao.zyx@alibaba-inc.com, huangwei5@mail.sysu.edu.cn

Abstract

Expressive Human Pose and Shape Estimation (EHPS) plays a crucial role in various AR/VR applications and has witnessed significant progress in recent years. However, current state-of-the-art methods still struggle with accurate parameter estimation for facial and hand regions and exhibit limited generalization to wild images. To address these challenges, we present CoEvoer, a novel one-stage synergistic cross-dependency transformer framework tailored for upper-body EHPS. CoEvoer enables explicit feature-level interaction across different body parts, allowing for mutual enhancement through contextual information exchange. Specifically, larger and more easily estimated regions such as the torso provide global semantics and positional priors to guide the estimation of finer, more complex regions like the face and hands. Conversely, the localized details captured in facial and hand regions help refine and calibrate adjacent body parts. To the best of our knowledge, CoEvoer is the first framework designed specifically for upper-body EHPS, with the goal of capturing the strong coupling and semantic dependencies among the face, hands, and torso through joint parameter regression. Extensive experiments demonstrate that CoEvoer achieves state-of-the-art performance on upper-body benchmarks and exhibits strong generalization capability even on unseen wild images.

Introduction

Expressive human pose and shape estimation (EHPS) (Pang et al. 2023; Liu, Qiu, and Zhang 2024; Tian et al. 2023; Shen et al. 2024) plays a central role in human behaviors understanding and has extensive applications (Zhang et al. 2024, 2023; Cai et al. 2022; Hong et al. 2021, 2022) in virtual reality, motion capture and human-computer interaction. Given that many practical EHPS applications—such as online education and video conference—are inherently upper-body dominated, there has been increasing interest in developing methods specifically tailored to such scenarios, especially following the emergence of pioneering work (Lin et al. 2023).

Despite notable progress, accurately reconstructing expressive regions, particularly the face, hands, and their transition areas, remains a significant challenge (Choutas et al. 2020; Moon, Choi, and Lee 2022a; Lin et al. 2023; Cai et al. 2023). Multi-stage methods (Moon, Choi, and Lee 2022a; Choutas et al. 2020) typically enhance detail recovery by cropping and upsampling specific regions, then feeding them into specialized expert models. While effective, these methods suffer from complex designs and prohibitive computational overhead, limiting their scalability in real-world applications. Frameworks such as OSX (Lin et al. 2023) and SMPLer-X (Cai et al. 2023) do not require separate expert networks for each part. These methods estimate the face, hands, and body jointly within a one-stage framework, offering improved inference efficiency. However, they typically treat different body parts independently in feature space, lacking explicit mechanisms for inter-part interaction. Consequently, they are unable to autonomously model the topological and semantic relationships among different human body components, leading to representational bottlenecks—particularly in the upper body, where fine-grained coordination is critical for expressive motion.

We argue that explicitly modeling feature-level interactions across body parts is crucial for expressive mesh recovery, particularly in upper-body-dominated scenes, for the following reasons: First, in upper-body scenarios, strong spatial and semantic dependencies exist among the body, hands, and face. These parts often move in a highly coordinated manner. For instance, at the spatial level, hand gestures are often conditioned on the configurations of proximal joints such as the shoulder and upper arm, while head and neck pose estimation is heavily influenced by facial orientation. At the semantic level, high-level human intent, such as making a phone call, typically induces characteristic postural patterns, including a downward head tilt and the placement of the hand near the ear. However, these models tokenize input into semantically agnostic patches, hindering their ability to capture structured priors—e.g., that hand motions are guided by arm configuration, or that facial direction should align with torso orientation. As a result, these models struggle to learn the topological and semantic relationships among human body components. Second, to simplify architecture and avoid training multiple expert models, current one-stage methods often crop features around the face and hands from shared feature maps. However, this inevitably introduces information loss—particularly with respect to the spatial context of hands. Incorporating inter-part interactions

†Corresponding author.

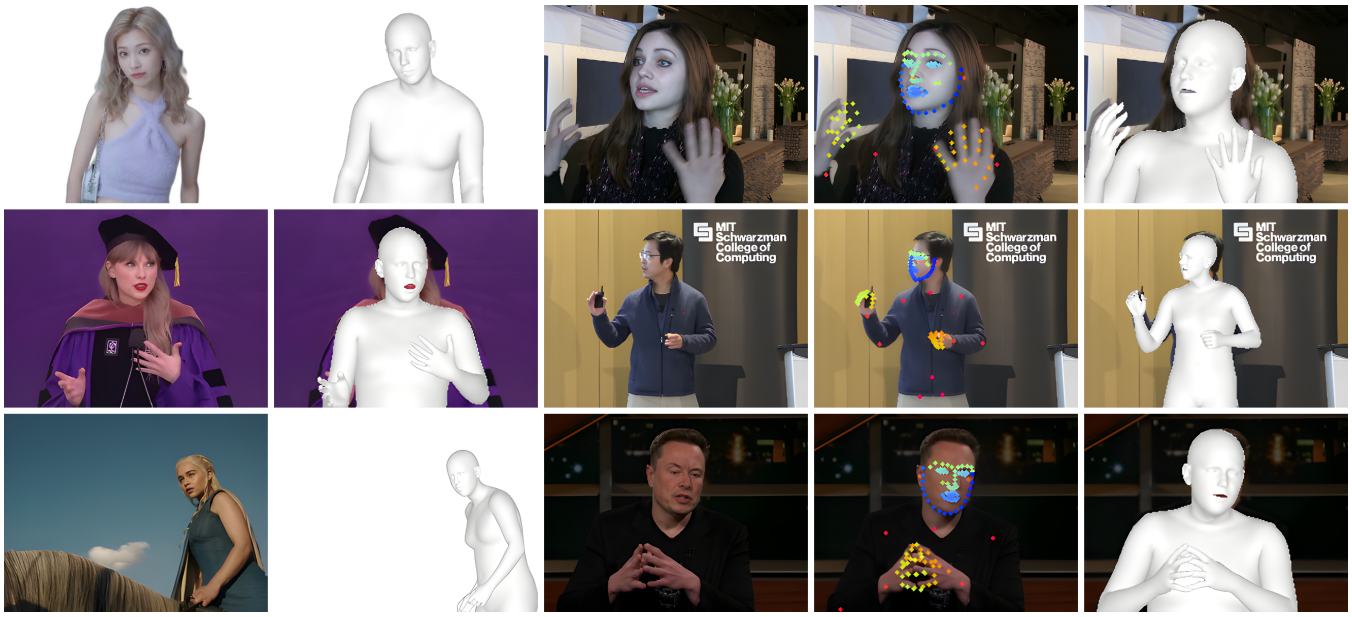

Figure 1: Our proposed CoEvoer achieves the mutual adaptation among different body parts through the mutual complementation and correction of global semantics and local features. When it is difficult to estimate some challenging regions like face and hands, it can rely on the semantic information provided by other parts for correction, which effectively improves the expressiveness and robustness of the method.

allows the model to recover lost spatial cues using global body posture, thereby enabling more accurate localization of hands and face. Moreover, such interactions improve robustness and generalization in challenging conditions such as occlusion or in-the-wild imagery: when either the body or hands/face is estimated with higher confidence, it can provide informative context to guide the estimation of the other. These observations motivate the need for a unified framework that can dynamically model semantic and spatial interactions across expressive body regions.

In this work, we propose CoEvoer, a simple yet effective one-stage framework that facilitates token-level cross-part interaction, tailored for upper-body scenarios where the face, body, and hands are highly interdependent. Unlike existing one-stage transformer-based architectures that lack explicit modeling of inter-part semantics, our method incorporates structured cross-part communication to enhance representation learning. Specifically, facial expression tokens attend to upper-body pose embeddings to refine head orientation estimation, while hand-related tokens are contextually enriched via interactions with the shoulder and upper-arm features. Conversely, body representations are refined through feedback from facial and hand tokens, enabling a more holistic understanding of the full upper-body configuration. This explicit modeling of semantic dependencies across parts leads to more accurate human pose estimation, especially in challenging regions such as the face and hands, and helps correct anatomically implausible or incoherent hand poses. Moreover, the proposed interaction mechanism allows information from one part to compensate for occlusions or ambiguities in another. For instance, when the face is partially occluded, the head pose can still be reliably estimated using cues from the torso orientation or shoulder alignment. The comparison between our proposed framework and existing methods is illustrated in Figure 2. Our contributions can be summarized as follows.

- We propose a novel one-stage framework with explicit token-level interaction, tailored for upper-body scenarios where body parts exhibit strong semantic and spatial coupling. To the best of our knowledge, this is the first work to explicitly optimize cross-part interaction mechanisms for upper-body-focused EHPS.
- By enabling semantic collaboration and contextual enhancement across body parts, our approach effectively captures inter-part dependencies and significantly improves model robustness and generalization, especially under occlusions and in-the-wild conditions.
- Extensive experiments on public benchmarks demonstrate that our method achieves state-of-the-art performance on upper-body EHPS tasks, with notable improvements in pose estimation of challenging regions such as the face and hands.

Related Work

Expressive Human Pose and Shape Estimation

EHPS not only needs to capture human postures and shapes (Li et al. 2022b; Wang et al. 2023b,a; Kocabas, Athanasiou, and Black 2020; Kocabas et al. 2021a; Kolotouros et al. 2019), but also to estimate facial expressions (Deng et al. 2019; Tewari et al. 2017) and hand gestures (Sun et al. 2022; Zhou et al. 2021; Xiang, Joo, and Sheikh 2019). In recent

years, there has been a notable upsurge in research interests (Cai et al. 2023; Sun et al. 2024; Baradel et al. 2024) focused on whole-body mesh recovery from monocular images. This trend is, to a certain extent, propelled by the ground-breaking research (Pavlakos et al. 2019) in the domain of enhancing whole-body parametric models. Different from existing studies (Moon, Choi, and Lee 2022a; Zhou et al. 2021; Feng et al. 2021) that separately recover the face, hands, or body, expressive human mesh recovery focuses on the joint estimation of the human face (Aldrian and Smith 2012; Tewari et al. 2017; Deng et al. 2019; Egger et al. 2020), hands (Boukhayma, Bem, and Torr 2019; Chatzis et al. 2020; Huang et al. 2021), and body (Choi, Moon, and Lee 2020; Kanazawa et al. 2018; Kolotouros et al. 2019; Kocabas et al. 2021a,b; Zeng et al. 2022a,b). Since the face and hands regions are small but have a large number of details, many existing multi-stage methods crop out different body parts, scale them up to a higher resolution, and then feed them into their respective expert models for parameter estimation. ExPose (Choutas et al. 2020) extracts high resolution cropped regions of the face and hands through the body driven attention mechanism and feeds them into their respective expert networks to utilize the specific knowledge from the face and hand datasets. FrankMocap (Rong, Shiratori, and Joo 2021) operates by separately conducting 3D mesh recovery for the face, hands, and body. Afterward, the outputs are combined via an integration module. PIXIE (Feng et al. 2021) combines separate estimations by leveraging the shared shape space of SMPL-X that encompasses all body parts. Hand4Whole (Moon, Choi, and Lee 2022a) predicts the 3D wrists by exploiting MCP features, generating more accurate rotations and improving hand estimation. Such multi-stage frameworks are complex and have relatively long inference time, making it difficult to be applied in practical scenarios.

One-stage Human Pose and Shape Estimation

In recent years, with the continuous development of whole-body datasets, many one-stage frameworks (Lin et al. 2023; Cai et al. 2023; Baradel et al. 2024; Sun et al. 2024) have been proposed for more concise and efficient expressive whole-body recovery. OSX (Lin et al. 2023) is a pioneering work in one-stage frameworks. It proposes a simple yet effective transformer architecture, which significantly improves the model's inference efficiency. SMPLer-X (Cai et al. 2023) leverages the large-scale model and further expands the dataset, endowing the model with stronger expressive power and transferability. Multi-HMR (Baradel et al. 2024) detects people by predicting 2D heatmaps of human positions, enabling the recovery of 3D human body meshes of multi-person from a single RGB image. Built upon DETR, AiOS (Sun et al. 2024) formulates the multi-person whole-body mesh recovery task as a progressive set prediction problem with diverse sequential detections. It is particularly suitable for scenarios that are crowded and feature significant occlusion. Despite their efficiency, most one-stage methods still treat each body part independently during encoding and decoding, lacking mechanisms for inter-part communication. This leads to spatial misalignment and limits the model's ability to capture semantic correlations across regions—especially problematic in expressive upper-body scenarios where face, hands, and torso often move in a coordinated manner. Our work addresses these limitations by introducing a token-level interaction mechanism that enables contextual refinement across body parts within a one-stage framework.

Preliminaries

Our model is constructed based on the 3D parametric model SMPL-X (Pavlakos et al. 2019). Given the image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, our proposed method estimates pose parameters $\theta \in \mathbb{R}^{53\times 3}$ including body poses $\theta_{body} \in \mathbb{R}^{22\times 3}$, left hand poses $\theta_{lhand} \in \mathbb{R}^{15\times 3}$, right hand poses $\theta_{rhand} \in \mathbb{R}^{15\times 3}$, jaw poses $\theta_{jaw} \in \mathbb{R}^{1\times 3}$, as well as shape parameters $\beta \in \mathbb{R}^{10}$ and facial expression parameters $\phi \in \mathbb{R}^{10}$. Finally, these parameters are fed into a SMPL-X layer to obtain the final 3D human mesh.

Proposed Method

Portrait Foreground Extraction

In expressive upper-body pose estimation, performance often degrades when the subject appears distant from the camera. In such cases, detailed regions such as the face and hands occupy fewer pixels and are more vulnerable to background interference. This introduces irrelevant contextual information into body features, thereby hindering effective feature interaction across body parts. To mitigate this issue, we propose Portrait Foreground Extraction, a module designed to explicitly enhance the model's focus on foreground regions. Inspired by recent advances in semantic segmentation (Hou et al. 2020; Guo et al. 2022; Ni et al. 2024), Portrait Foreground Extraction isolates the human foreground by combining structured global context with refined local details. Given an input image $\mathcal{I}_{img}$, we first apply a convolutional encoder to extract initial features $\mathcal{Z}_{img}$ and adjust their channel dimensions. To capture the spatial distribution of the foreground, we perform adaptive pooling along the height and width dimensions, producing directional context features $\mathcal{Z}_{horizontal}$ and $\mathcal{Z}_{vertical}$, respectively. These are fused via element-wise addition to generate a coarse foreground activation map $\mathcal{Z}_{rec}$. To further refine this map, we apply strip convolutions in both horizontal and vertical directions. These operations enhance linear structures and boundary information, which are critical for distinguishing human silhouettes from background clutter. Finally, the refined foreground features are aggregated using a set of initialized queries, which act as foreground-aware tokens. Since these queries are subsequently used to inject pose priors into the interaction process for facial and hand pose estimation, we refer to them as Prior Interaction Tokens. The resulting foreground representation $\mathcal{Z}_{human}$ provides cleaner semantics and serves as a more robust foundation for downstream token-level feature interactions.

Collaborative Evolution Enhancer

Retrieval of different body parts. To ensure inference efficiency, we adopt a streamlined one-stage architecture with

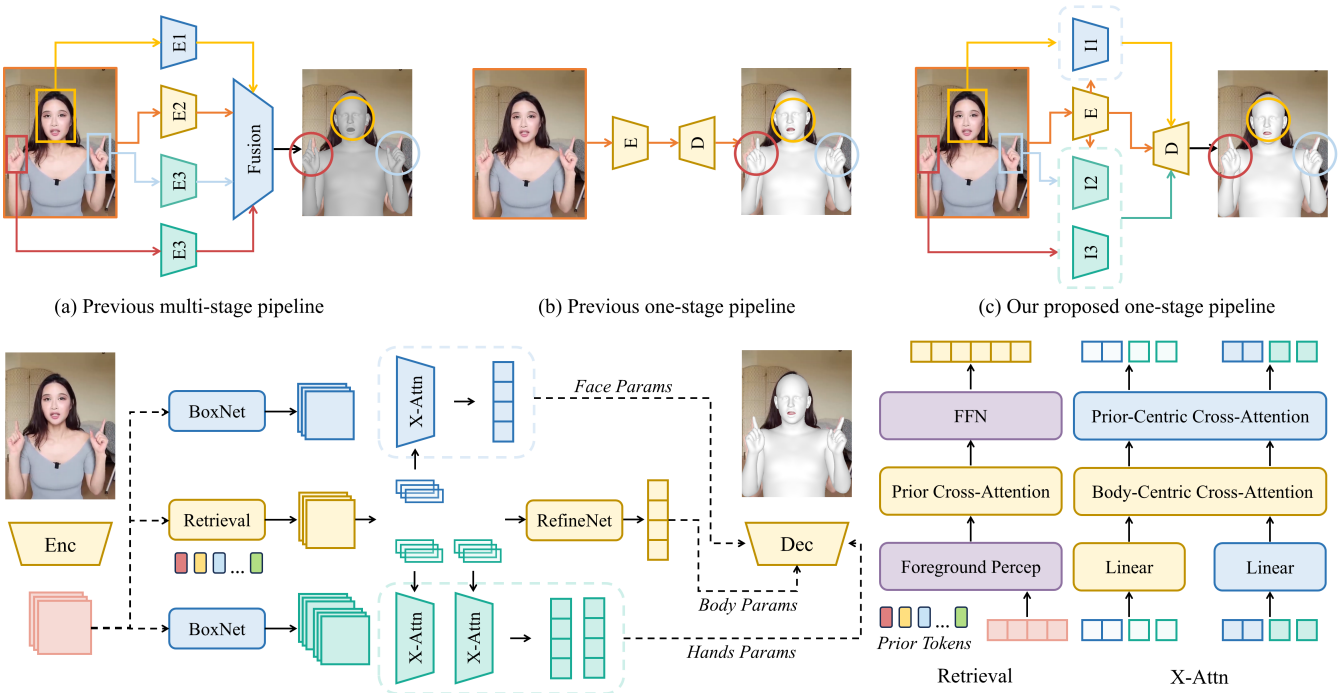

Figure 2: Comparison of existing mesh recovery methods and ours: multi-stage frameworks use part-specific experts (face E1, body E2, hands E3) for independent processing; one-stage frameworks adopt unified encoding-decoding, but still perform separate regressions for different parts without interaction. Our method preserves one-stage efficiency and adds explicit cross-part token-level interactions (I1: body-head, I2: body-left hand, I3: body-right hand).

a single shared encoder (Dosovitskiy et al. 2020; Li et al. 2022a; Zhu et al. 2020). The feature representations for the face ($\mathcal{Z}_{face}$), left hand ($\mathcal{Z}_{lhand}$), and right hand ($\mathcal{Z}_{rhand}$) are extracted using BoxNet, which first predicts bounding boxes via multi-layer perceptron, followed by RoIAlign applied to the shared feature map to retrieve the corresponding local features. In upper-body scenarios, different body parts exhibit strong spatial and semantic correlations. Spatially, head movement is closely linked to the neck and shoulder regions, while hand positions are largely governed by arm configurations. Semantically, actions such as answering a phone call or gesturing during a conversation involve coordinated movements among the head, hands, and torso. These patterns suggest the existence of consistent inter-part dependencies. To model these relationships, we divide the torso features $\mathcal{Z}_{human}$—enhanced by the Portrait Foreground Extraction—into three tokens: $\mathcal{T}_{face}$, $\mathcal{T}_{lhand}$, and $\mathcal{T}_{rhand}$. These tokens serve to model the spatial and semantic dependencies between the torso and each corresponding body part, providing contextual cues that enhance the effectiveness of subsequent token-level feature interactions.

Explicit token-level cross-part interaction. The suboptimal performance of one-stage frameworks in facial and hand pose estimation largely stems from their design for high inference efficiency. Unlike multi-stage approaches that crop high-resolution patches and process them with specialized branches, one-stage methods retrieve features directly from shared feature maps. As a result, the representations for the face, left hand, and right hand suffer from degraded spatial localization and become decoupled from the holistic body context. To address this limitation, we introduce a bidirectional cross-attention mechanism that explicitly models spatial and semantic dependencies among different body parts. The globally contextualized torso features—enhanced by the Portrait Foreground Extraction module—are partitioned into three segments: $\mathcal{T}_{face}$, $\mathcal{T}_{lhand}$, and $\mathcal{T}_{rhand}$, corresponding to the face, left hand, and right hand, respectively. These segments serve as keys and values to augment the queries from the facial and hand tokens. Conversely, the local part tokens also act as queries to retrieve complementary contextual cues from their respective torso segments, enabling mutual refinement between local and global representations. Specifically, the bidirectional interaction is defined as follows:

$$\widetilde{\mathcal{Z}}_{face} = \mathcal{F}^{corr}(\mathcal{Z}_{face}, \mathcal{T}_{face}, \mathcal{T}_{face}) \quad (1)$$

$$\widetilde{\mathcal{T}}_{face} = \mathcal{F}^{corr}(\mathcal{T}_{face}, \mathcal{Z}_{face}, \mathcal{Z}_{face}) \quad (2)$$

where $\widetilde{\mathcal{Z}}_{face}$ denotes the corrected facial feature, and $\widetilde{\mathcal{T}}_{face}$ is the refined torso segment associated with the face. The general form of the attention-based correction function is:

$$\mathcal{F}^{corr}(Q, K, V) = \mathcal{F}^{fc}(\mathcal{F}^{attn}(Q, K, V)) \quad (3)$$

The same interaction procedure is applied to the left hand and right hand tokens and their corresponding torso segments, namely $(\mathcal{Z}_{lhand}, \mathcal{T}_{lhand})$ and $(\mathcal{Z}_{rhand}, \mathcal{T}_{rhand})$. To consolidate the enhanced global context, we further fuse the

Method	MPVPE ↓			PA-MPVPE ↓		
	All	Hand	Face	All	Hand	Face
ExPose	171.5	83.7	45.1	66.9	12.0	3.9
PIXIE	168.4	55.6	45.2	61.7	12.2	4.2
H4W	104.1	45.7	27.0	44.8	8.9	2.8
OSX	81.9	41.5	21.2	42.2	8.6	2.0
AiOS	58.6	39.0	19.6	32.5	7.3	2.8
SMPLer-X	57.4	40.2	21.6	31.9	10.3	2.8
Ours	51.6	27.8	16.2	26.5	7.1	1.8

Table 1: Reconstruction errors on UBody. Red background indicates best results, yellow background indicates second best results.

Method	MPVPE ↓			PA-MPVPE ↓		
	All	Hand	Face	All	Hand	Face
ExPose	219.8	115.4	103.5	88.0	12.1	4.8
FrankMocap	218.0	95.2	105.4	90.6	11.2	4.9
PIXIE	203.0	89.9	95.4	82.7	12.8	5.4
H4W	183.9	72.8	81.6	73.2	9.7	4.7
OSX	168.6	70.6	77.2	69.4	11.5	4.8
Ours	160.8	68.7	75.7	65.9	10.1	4.7

Table 2: Reconstruction errors on AGORA. Red background indicates best results, yellow background indicates second best results.

refined torso segments $\widetilde{\mathcal{T}}_{face}$, $\widetilde{\mathcal{T}}_{lhand}$, and $\widetilde{\mathcal{T}}_{rhand}$ using a multi-layer perceptron, referred to as the RefineNet module. This fusion propagates part-aware semantics into the global body representation, reinforcing anatomical coherence. By leveraging bidirectional cross-attention for token-level feature refinement, our method yields more accurate and anatomically plausible full-body pose estimations. This mechanism also enhances the localization of facial and hand keypoints and effectively mitigates issues such as unnatural hand configurations. Finally, we feed the refined part-specific features into separate MLP heads to predict the output parameters: $\mathcal{O}_{body}$, $\mathcal{O}_{face}$, $\mathcal{O}_{lhand}$, and $\mathcal{O}_{rhand}$.

Loss function. Our proposed CoEvoer is trained in an end-to-end manner by minimizing the loss function $\mathcal{L}$, defined as follows:

$$\mathcal{L} = \mathcal{L}_{param} + \mathcal{L}_{kpts} + \mathcal{L}_{bbox} \tag{4}$$

where $\mathcal{L}_{param}$ is the L1 distance between the estimated SMPL-X parameters and the ground truth, and $\mathcal{L}_{kpts}$ is the L1 distance between the predicted keypoints and the ground truth, which includes the 3D keypoints of the human body and the projected 2D keypoints. $\mathcal{L}_{param}$ and $\mathcal{L}_{kpts}$ are used to supervise the human body posture and shape. $\mathcal{L}_{bbox}$ refers to the estimated coordinates of the facial and hand bounding boxes, which are used to supervise the detection of the facial and hand regions.

Method	MPVPE ↓			PA-MPVPE ↓		
	All	Hand	Face	All	Hand	Face
ExPose	77.1	51.6	35.0	54.5	12.8	5.8
FrankMocap	107.6	42.8	-	57.5	12.6	-
PIXIE	89.2	42.8	32.7	55.0	11.1	4.6
H4W	76.8	39.8	26.1	50.3	10.8	5.8
OSX	70.8	53.7	26.4	48.7	15.9	6.0
Ours	66.3	48.9	25.0	46.4	14.8	5.8

Table 3: Reconstruction errors on EHF. Red background indicates best results, yellow background indicates second best results.

Experiments

Experimental Setup

Following existing methods, we use Human3.6M (Ionescu et al. 2013), COCO-Wholebody (Lin et al. 2014), and MPII (Andriluka et al. 2014) as the training set. The SMPL/SMPL-X pseudo-GTs are obtained from EFT (Joo, Neverova, and Vedaldi 2021) and NeuralAnnot (Moon, Choi, and Lee 2022b). Since UBody (Lin et al. 2023) is the most recent large-scale dataset for human pose and shape estimation, offering significantly more data and a broader coverage of real-world scenarios than previous datasets, it has been widely adopted as a representative benchmark for evaluating the downstream applicability of human mesh recovery methods. In our experiments, we fine-tune our model on the UBody and evaluate its performance accordingly.

Implementation. PyTorch is used for implementation. We use the Adam optimizer to conduct the training with an initial learning rate of 1×10^{-4}. Meanwhile, we adopt data augmentation methods including scaling, rotation, random horizontal flip, and color jittering. All the hyperparameter settings are kept consistent with comparative methods.

Evaluation metrics. For expressive human pose and shape estimation, we adopt the mean per-vertex position error (MPVPE) as the main evaluation metric. Meanwhile, we also report the procrustes analysis mean per-vertex position error (PA-MPVPE) after rigid alignment to further analyze the performance of the model. Regarding the 3D hand error, we report the average 3D errors of the left hand and the right hand. The units of all the reported metrics are millimeters.

Quantitative Comparison with SOTA

Table 1 shows the detailed comparison between our proposed CoEvoer and existing human pose and shape estimation methods. We achieved improvements of 10.1% and 16.9% in MPVPE and PA-MPVPE respectively on UBody, among which the MPVPE of the hand achieved an improvement of 28.7%. It is worth noting that even though SMPLer-X and AiOS used additional datasets, CoEvoer still outperforms the existing SOTA methods. As a unified one-stage framework, our method outperforms even the two-stage pipelines without relying on any supplementary face-only or hand-only datasets. As shown in Table 2 and Table

Figure 3: Qualitative visualization comparison on the UBody dataset. Each row shows, from left to right: the input image, results from Hand4Whole, AiOS, SMPLer-X, and our proposed CoEvoer. Best viewed in color with zoom-in for details.

3, we also achieved performance improvements of 4.6% and 6.4% on AGORA and EHF respectively, and the estimation results of the hand and face on the EHF dataset were improved by 8.9% and 5.3% respectively.

Qualitative Comparison with SOTA

As depicted in Figure 3, we present a qualitative comparison between our proposed CoEvoer, and existing state-of-the-art approaches, including the two-stage pipeline Hand4Whole and the one-stage methods SMPLer-X and AiOS. Notably, both SMPLer-X and AiOS are trained with additional external datasets. In the first example, both Hand4Whole and SMPLer-X suffer from noticeable finger interpenetration artifacts, while AiOS fails to accurately estimate the facial and hand poses. In the second example, Hand4Whole and SMPLer-X exhibit suboptimal hand recovery, and AiOS produces anatomically implausible hand configurations. For the remaining examples, we omit detailed discussion and encourage readers to evaluate the results based on global body consistency, self-intersections, and unnatural pose artifacts. It is particularly worth noting that in the fifth and most challenging case, all baseline methods exhibit substantial estimation failures, whereas CoEvoer maintains robust and accurate performance.

Ablation Study and Discussion

As shown in Table 4, we design two variants to evaluate the effectiveness of our proposed modules. Variant1 is designed to assess the impact of explicitly modeling token-level interactions across different body parts. It performs feature fusion via self-attention on concatenated part features without using cross-attention for inter-part information exchange, thereby isolating the contribution of explicit cross-part communication in our framework.

To further demonstrate the effectiveness of our approach, we present a visualization of keypoint estimation results for CoEvoer and the variant in Figure 4, focusing on facial and partial torso regions, including the neck and shoulders. As

illustrated, the subject in the image is looking downward, causing the facial keypoints to appear significantly lower than in typical upright poses. The variant tends to predict relatively accurate torso keypoints, which are less affected by the head orientation, while its facial estimation results are biased upwards. In contrast, our method captures the dependency between the torso and face through token-level interaction. By recognizing the subtle deviation in the neck position, CoEvoer successfully infers the downward head pose and accordingly corrects the facial keypoints, leading to a more accurate and coherent estimation.

Figure 4: Comparison of facial keypoint estimation results. The first column shows the input image, the second column visualizes the output of Variant1, and the third column shows the result produced by our proposed CoEvoer. Best viewed in color and zoom-in for more clarity.

We further compare our proposed method with the variant under extreme conditions. As shown in Figure 5, the input is a wild image where the subject's right arm is completely occluded. Our method accurately infers the approximate location of the invisible right arm and even captures the specific hand gesture of gripping the saddle while riding. In contrast, the variant model predicts a generic downward-facing hand pose, failing to reflect the contextual semantics of the scene.

Figure 5: Comparison of mesh recovery results. The first column shows the input image, the second column visualizes the output of Variant1, and the third column shows the result produced by our proposed CoEvoer. Best viewed in color and zoom-in for more clarity.

Method	MPVPE ↓			PA-MPVPE ↓		
	All	Hand	Face	All	Hand	Face
Variant1-w/o C.E.E.	66.5	36.9	23.8	33.0	8.9	2.1
Variant2-w/o P.F.E.	54.4	29.7	17.7	27.5	9.4	2.1
Ours	51.6	27.8	16.2	26.5	7.1	1.8

Table 4: The ablation study on UBody. C.E.E. and P.F.E. are abbreviations for Collaborative Evolution Enhancer and Portrait Foreground Extraction respectively.

While the proposed CoEvoer framework is primarily tailored to improve expressive upper-body pose and shape estimation, the quantitative results in Table 2 and Table 3 indicate consistent improvements across full-body benchmarks as well. The qualitative visualizations in Figure 6 further highlight the robustness of our approach under challenging scenarios, such as severe hand occlusions (left) and uncommon pose configurations (right). It is worth noting that, in the left example, the reconstructed lower limbs exhibit reduced accuracy. We attribute this phenomenon to the relatively weak correlation between the legs and the face/hands in full-body settings, which limits the amount of useful information that can be inferred for lower-limb estimation from upper-body cues. We believe this observation, together with our experimental findings, can motivate future studies on inter-region dependencies in human pose and shape estimation.

Figure 6: Qualitative visualization of human mesh recovery results on the AGORA dataset. The first and third columns show the input image, the second and fourth columns show the human mesh recovery produced by our proposed CoEvoer. Best viewed in color and zoom-in for more clarity.

Conclusion

In this work, we introduced CoEvoer, a simple yet effective one-stage framework for expressive human pose and shape estimation in upper-body-dominated scenarios. Unlike existing approaches that treat body parts independently, CoEvoer explicitly models token-level interactions across the face, hands, and body, capturing rich semantic and spatial dependencies inherent to expressive motion. Through structured cross-part communication, CoEvoer enables each region to leverage contextual cues from others, resulting in more coherent and anatomically plausible mesh recovery, especially in challenging regions such as the hands and face. Our method preserves the efficiency of one-stage architectures while significantly enhancing representational capacity and generalization, particularly under occlusion and in-the-wild conditions. Extensive experiments on public benchmarks demonstrate that CoEvoer achieves state-of-the-art performance, validating the efficacy of explicit inter-part interaction in upper-body EHPS tasks.

Acknowledgments

The work was supported by the Shenzhen Science and Technology Program (Grant No. JCYJ20230807110807015), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515011757), and Baidu, Inc. We thank the anonymous reviewers for their constructive comments and suggestions.

References

Aldrian, O.; and Smith, W. A. 2012. Inverse rendering of faces with a 3D morphable model. *IEEE transactions on pattern analysis and machine intelligence*, 35(5): 1080–1093.

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 3686–3693.

Baradel, F.; Armando, M.; Galaaoui, S.; Brégier, R.; Weinzaepfel, P.; Rogez, G.; and Lucas, T. 2024. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, 202–218. Springer.

Boukhayma, A.; Bem, R. d.; and Torr, P. H. 2019. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10843–10852.

Cai, Z.; Ren, D.; Zeng, A.; Lin, Z.; Yu, T.; Wang, W.; Fan, X.; Gao, Y.; Yu, Y.; Pan, L.; et al. 2022. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, 557–577. Springer.

Cai, Z.; Yin, W.; Zeng, A.; Wei, C.; Sun, Q.; Yanjun, W.; Pang, H. E.; Mei, H.; Zhang, M.; Zhang, L.; et al. 2023. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36: 11454–11468.

Chatzis, T.; Stergioulas, A.; Konstantinidis, D.; Dimitropoulos, K.; and Daras, P. 2020. A comprehensive study on deep learning-based 3D hand pose estimation methods. *Applied Sciences*, 10(19): 6850.

Choi, H.; Moon, G.; and Lee, K. M. 2020. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, 769–787. Springer.

Choutas, V.; Pavlakos, G.; Bolkart, T.; Tzionas, D.; and Black, M. J. 2020. Monocular expressive body regression through body-driven attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, 20–40. Springer.

Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; and Tong, X. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 0–0.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Egger, B.; Smith, W. A.; Tewari, A.; Wuhrer, S.; Zollhoefer, M.; Beeler, T.; Bernard, F.; Bolkart, T.; Kortylewski, A.; Romdhani, S.; et al. 2020. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5): 1–38.

Feng, Y.; Choutas, V.; Bolkart, T.; Tzionas, D.; and Black, M. J. 2021. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, 792–804. IEEE.

Guo, M.-H.; Lu, C.-Z.; Hou, Q.; Liu, Z.; Cheng, M.-M.; and Hu, S.-M. 2022. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in neural information processing systems*, 35: 1140–1156.

Hong, F.; Pan, L.; Cai, Z.; and Liu, Z. 2021. Garment4d: Garment reconstruction from point cloud sequences. *Advances in Neural Information Processing Systems*, 34: 27940–27951.

Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; and Liu, Z. 2022. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*.

Hou, Q.; Zhang, L.; Cheng, M.-M.; and Feng, J. 2020. Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4003–4012.

Huang, L.; Zhang, B.; Guo, Z.; Xiao, Y.; Cao, Z.; and Yuan, J. 2021. Survey on depth and RGB image-based 3D hand shape and pose estimation. *Virtual Reality & Intelligent Hardware*, 3(3): 207–234.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339.

Joo, H.; Neverova, N.; and Vedaldi, A. 2021. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, 42–52. IEEE.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5253–5263.

Kocabas, M.; Huang, C.-H. P.; Hilliges, O.; and Black, M. J. 2021a. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11127–11137.

Kocabas, M.; Huang, C.-H. P.; Tesch, J.; Müller, L.; Hilliges, O.; and Black, M. J. 2021b. SPEC: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11035–11045.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2252–2261.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022a. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, 280–296. Springer.

Li, Z.; Liu, J.; Zhang, Z.; Xu, S.; and Yan, Y. 2022b. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, 590–606. Springer.

Lin, J.; Zeng, A.; Wang, H.; Zhang, L.; and Li, Y. 2023. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21159–21168.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, Y.; Qiu, C.; and Zhang, Z. 2024. Deep learning for 3d human pose estimation and mesh recovery: A survey. *Neurocomputing*, 128049.

Moon, G.; Choi, H.; and Lee, K. M. 2022a. Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2308–2317.

Moon, G.; Choi, H.; and Lee, K. M. 2022b. Neuralannot: Neural annotator for 3d human mesh training sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2299–2307.

Ni, Z.; Chen, X.; Zhai, Y.; Tang, Y.; and Wang, Y. 2024. Context-guided spatial feature reconstruction for efficient semantic segmentation. In *European Conference on Computer Vision*, 239–255. Springer.

Pang, H. E.; Cai, Z.; Yang, L.; Tao, Q.; Wu, Z.; Zhang, T.; and Liu, Z. 2023. Towards robust and expressive whole-body human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36: 17330–17344.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10975–10985.

Rong, Y.; Shiratori, T.; and Joo, H. 2021. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1749–1759.

Shen, W.; Yin, W.; Wang, H.; Wei, C.; Cai, Z.; Yang, L.; and Lin, G. 2024. HMR-Adapter: A Lightweight Adapter with Dual-Path Cross Augmentation for Expressive Human Mesh Recovery. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6093–6102.

Sun, Q.; Wang, Y.; Zeng, A.; Yin, W.; Wei, C.; Wang, W.; Mei, H.; Leung, C.-S.; Liu, Z.; Yang, L.; et al. 2024. Aios: All-in-one-stage expressive human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1834–1843.

Sun, Y.; Huang, T.; Bao, Q.; Liu, W.; Gao, W.; and Fu, Y. 2022. Learning monocular mesh recovery of multiple body parts via synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2669–2673. IEEE.

Tewari, A.; Zollhofer, M.; Kim, H.; Garrido, P.; Bernard, F.; Perez, P.; and Theobalt, C. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE international conference on computer vision workshops*, 1274–1283.

Tian, Y.; Zhang, H.; Liu, Y.; and Wang, L. 2023. Recovering 3d human mesh from monocular images: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(12): 15406–15425.

Wang, W.; Ge, Y.; Mei, H.; Cai, Z.; Sun, Q.; Wang, Y.; Shen, C.; Yang, L.; and Komura, T. 2023a. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3925–3935.

Wang, Y.; Sun, Q.; Wang, W.; Ling, J.; Cai, Z.; Xie, R.; and Song, L. 2023b. Learning dense uv completion for human mesh recovery. *arXiv preprint arXiv:2307.11074*.

Xiang, D.; Joo, H.; and Sheikh, Y. 2019. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10965–10974.

Zeng, A.; Ju, X.; Yang, L.; Gao, R.; Zhu, X.; Dai, B.; and Xu, Q. 2022a. Deciwatch: A simple baseline for $10\times$ efficient 2d and 3d pose estimation. In *European Conference on Computer Vision*, 607–624. Springer.

Zeng, A.; Yang, L.; Ju, X.; Li, J.; Wang, J.; and Xu, Q. 2022b. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*, 625–642. Springer.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6): 4115–4128.

Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; and Liu, Z. 2023. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 364–373.

Zhou, Y.; Habermann, M.; Habibie, I.; Tewari, A.; Theobalt, C.; and Xu, F. 2021. Monocular real-time full body capture with inter-part correlations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4811–4822.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.