This version of the article has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://doi.org/10.1007/s11263-026-03113-6. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms.

BBoxMaskPose v2: Expanding Mutual Conditioning to 3D

Miroslav Purkrabek*, Constantin Kolomiiets and Jiri Matas

Visual Recognition Group, Czech Technical University in Prague,

*Corresponding author(s). E-mail(s): purkrmir@fel.cvut.cz;

Abstract

Most 2D human pose estimation benchmarks are nearly saturated, with the exception of crowded scenes. We introduce PMPose, a top-down 2D pose estimator that incorporates the probabilistic formulation and the mask-conditioning. PMPose improves crowded pose estimation without sacrificing performance on standard scenes. Building on this, we present BBoxMaskPose v2 (BMPv2) integrating PMPose and an enhanced SAM-based mask refinement module. BMPv2 surpasses state-of-the-art by 1.5 average precision (AP) points on COCO and 6 AP points on OCHuman, becoming the first method to exceed 50 AP on OCHuman. We demonstrate that BMP's 2D prompting of 3D model improves 3D pose estimation in crowded scenes and that advances in 2D pose quality directly benefit 3D estimation. Results on the new OCHuman-Pose dataset show that multi-person performance is more affected by pose prediction accuracy than by detection.
The code, models, and data are available on the project website.

Keywords: Human Pose Estimation, Person Detection, Instance Segmentation, Mutual Conditioning



Fig. 1: **BBoxMaskPose v2** establishes a new SOTA detection, segmentation, and 2D pose estimation in crowded scenes (row 2). Combined with SAM-3D-Body [1], it enables 3D human reconstruction even in scenes with complex interactions.

1 Introduction

Human pose estimation (HPE) research is gradually shifting from 2D to 3D. However, state-of-the-art 3D pose estimators [1, 2] rely on accurate 2D input, whether from detections, masks, or points. Thus, improving the 2D HPE remains essential for reliable 3D human body modelling.

Multi-person interaction, where individuals are so close that their bounding boxes are almost the same, is one of the main challenges in 2D and 3D. Real-world data for such cases are limited [3–6]. Existing "crowded" datasets vary in difficulty – datasets such as [3–5] include individuals close to each other, but lack the interactions present in OCHuman [6].

BMPv1 [7] (ICCV 2025) addresses joint detection and pose estimation in multi-person scenes and achieves strong results in crowded scenes. Its core is a self-correcting loop of three lightweight models – a detector, a pose estimator, and a segmentation model – conditioned on each other to iteratively refine their predictions. BMPv1 shows that mutual conditioning between different human-body representations is effective in crowded scenes. However, the prompting of SAM [8] within BMP is complex, often leads to over-segmentation, and makes the mask refinement stage the dominant failure mode.

ProbPose [9] (CVPR 2025) addresses different limitations of top-down human pose estimation through probabilistic keypoint modeling. It predicts visibility, presence probability, and expected OKS[1] for each keypoint, and uses crop augmentation to improve robustness near image borders. These outputs are attractive for BMP because they are useful not only for pose localization, but also for selecting informative prompts for mask refinement.

The work on BMPv1 [7] and ProbPose [9] is complementary. BMPv1 use mask conditioning, which is important for disentangling individuals in crowds, but it lacks rich probabilistic outputs for calibration and prompt selection. ProbPose provides such outputs, but it is not designed for explicit instance-mask conditioning inside the crowded BMP loop. Crowded mutual conditioning requires both properties at once.

[1] Object Keypoint Similarity [10]

This paper introduces **BMPv2**, an upgrade of BMPv1 that consistently improves pose estimation and segmentation performance and builds on BMPv1 [7] (ICCV 2025), ProbPose [9] (CVPR 2025), and SAM-pose2seg [11] (CVWW 2026). Its new contributions are PMPose, the BMPv2 system integration, the higher-precision BMPv2+ variant, OCHuman-Pose, and broader analysis of prompting, runtime, and crowded-scene data.

BMPv2 extends BMPv1 with two new components: PMPose and SAM-pose2seg. By incorporating these changes, BMPv2 sets a new SOTA performance on both standard scenes (i.e. on COCO [10]) and crowds (i.e. on OCHuman [6]) for 2D pose estimation and instance segmentation. The improved mask refinement from SAM-pose2seg further enables an additional PMPose refinement, resulting in **BMPv2+**, which targets high-precision scenarios at the cost of increased computation.

PMPose is a new top-down pose estimation model introduced as part of BMPv2. It combines mask conditioning with probabilistic keypoint modeling in a single architecture. It is suitable both for the BMP loop and as a standalone 2D top-down human pose estimator. PMPose outperforms previous non-iterative pose estimation methods in both standard and crowded scenes.

BMPv2 uses **SAM-pose2seg**, a version of SAM [8] adapted for human pose prompts, for mask refinement. It reduces over-segmentation and replaces the complex prompting strategy of BMPv1. It also makes direct use of the visibility-related information provided by PMPose.

We also analyze OCHuman and CIHP, two crowded-scene benchmarks with different weaknesses. OCHuman contains close interactions but incomplete annotations, while CIHP provides precise masks but no pose annotations. To address the main limitation of OCHuman, we introduce **OCHuman-Pose**, which adds COCO-style keypoint annotations to 2739 previously bounding-box-only instances in the original images. Results on OCHuman-Pose suggest that detection is less dominant than originally implied by OCHuman and that pose estimation remains a major bottleneck.

Finally, we use BMPv2 to prompt SAM-3D-Body [1] and analyze 3D estimation in challenging crowded scenes. This study shows that stronger 2D outputs improve 3D human body estimation

in crowds. This supports the main motivation of the paper: robust 2D understanding remains a necessary part of in-the-wild 3D pipelines.

The main contributions of this paper are:

1. **PMPose**: a family of mask-conditioned probabilistic 2D top-down pose estimators that improve both crowded-scene and standard-scene performance (Sec. 3.1).
2. **BMPv2** and **BMPv2+**: improved crowded-scene methods integrating PMPose and SAM-pose2seg that set a new SOTA for pose estimation and instance segmentation (Secs. 3, 3.3 and 5).
3. **OCHuman-Pose dataset**: an extension of OCHuman [6] with additional keypoint annotations for previously missing instances (Sec. 4.1).
4. **Analysis of BMP components**: ablations of PMPose, SAM-pose2seg prompting, runtime, and crowded-scene data weaknesses (Secs. 4, 5.3.1 and 5.3.5).
5. **Downstream 3D prompting study**: analysis of how improved 2D outputs affect 3D human body estimation in crowds (Sec. 5.4).

2 Related Work

Detection-Based Segmentation: The most direct approach to human instance segmentation uses a detector [12–17, 17, 18]. A detector takes an image as input and outputs instance masks and class labels. Detectors work well in scenes similar to the training data, but struggle when instances overlap heavily. Under severe occlusion, they often merge multiple people into one mask, or assign different body parts to different instances. In crowded scenes, standard non-maximum suppression (NMS) can remove true positives, leading to missed instances. Relaxing NMS (e.g., higher IoU thresholds) can recover these instances, but typically increases false positives in non-crowded scenes. [19] proposed to suppress detection not on bounding box level but after pose estimation, which results in longer inference times since false positive detections have to first be processed by the pose estimator before being suppressed.

Semantics-Aware Prompted Segmentation: Several works [20–23] inject semantics into SAM [8], similarly to our SAM-pose2seg. They take an image and a text prompt as input and detect and segment instances. This setting is often referred to as *open-vocabulary* or *zero-shot* segmentation. While these models capture semantics and segment entire instances, they do not provide localized cues and offer no control over which instance is chosen and, therefore, are unsuitable for iterative loops such as BMP [7].

Pose-Guided Instance Segmentation is the most specific form of prompted segmentation. The model takes an image and a detected or ground-truth human pose and segments the corresponding person. These methods specialize in humans, sacrificing the generality of open-vocabulary or generic segmenters for robustness and precise, localized control. Test-time optimization methods [24, 25] refine predictions at inference without training. While training-free, they are computationally expensive and slow at test time. Standard pose-guided human segmentation methods [6, 26–31] rely on small training datasets and thus generalize poorly, especially in crowded scenes. The closest work related to SAM-pose2seg is CrowdSAM [32], which builds a framework around SAM to automatically annotate bounding boxes in crowds. CrowdSAM uses SAM and DINO [33] as external tools and optimizes prompts, similar to [7]. In contrast, our SAM-pose2seg is end-to-end and, when used in an iterative loop, outperforms CrowdSAM, as shown already in [7].

2D Human Pose Estimation: There are two main approaches: top-down and detector-free. Detector-free can be further divided into single-stage [34–37], bottom-up [38–40] and hybrid [41].

Top-down methods [17, 42–46] use a person detector to detect bounding boxes and estimate one skeleton for each bounding box. They leverage big progress in human detection and specialize in understanding of human structure. Top-down methods are the most successful on datasets such as COCO, MPII or AIC but struggle on crowded datasets (e.g. OCHuman) due to low-quality detections. Most notably, ViTPose [42] combines multiple datasets into one strong backbone and sets a strong baseline, setting up state-of-the-art performance on most datasets. While conditioning pose estimation on bounding boxes (bbox-to-pose; standard top-down approach) is well researched,

conditioning pose on masks (mask-to-pose) was explored only through MaskPose [7].

While the vast majority of top-down approaches are heatmap-based, some models [47–49] predict keypoints through regression or regression-focused heatmaps [50, 51]. Important for this work is ProbPose [9] – a heatmap-based method modelling the underlying probability distribution. ProbPose is the first model to estimate probabilities of keypoints being out of the image along with their visibilities. As will be shown in this work, such estimates are useful for pose-guided instance segmentation in the BMP loop.

Detector-free models do not achieve SOTA performance on COCO but are superior to top-down methods on OCHuman as they are specialized in decoupling close-interaction instances. The most successful model, BUCTD [41], conditions top-down pose estimation with previously estimated keypoints from bottom-up methods. It is a pose-refinement method with state-of-the-art results on OCHuman due to its strong ability to decouple closely interacting people.

Human-Centered Foundational models. The latest directions in human body modeling are foundational models [52–55]. They learn general features describing the human body that could be used for human-related tasks such as segmentation and pose estimation. Most notably, Sapiens 2b [52] was trained on a staggering 2M images and, with 2B parameters, is almost four times bigger than ViTPose-h. Even with this size, foundational models perform comparably or worse than much smaller specialized models.

Iterative Approaches: This work builds on the BBoxMaskPose loop [7]. BMP combines three specialized models into a self-improving loop ideal for crowded and interaction-rich scenarios. While BMP set a new SOTA on all three tasks (detection, segmentation, and 2D pose estimation), it still struggles in cases such as over-segmenting the human body or merging two individuals into the same pose. Other iterative approaches are already mentioned BUCTD [41] and IterDet [56]. IterDet detects people in crowds iteratively, similarly to BMP. This work improves BMPv1 with new pose estimation and pose-guided segmentation models and outperforms BUCTD and BMPv1 by a significant margin.

3D Human Pose and Shape Estimation has gained significant attention in recent years [57]. Most methods rely on SMPL-based body models [58, 59], though several alternatives exist. SMIL [60] extends SMPL to infants. ANNY [61] models population-level variability across race, age, and body type. SKEL [62] augments SMPL with a realistic skeleton, while MHR [63] and ATLAS [64] model body shape directly around the physical skeleton. While some models (e.g. [65]) estimate 3D bodies directly from images, most 3D pose estimation methods require accurate 2D inputs. Earlier approaches [66–68] assume bounding-box–centered crops, whereas recent methods [1, 2] accept prompts such as text or segmentation masks. BMPv2 provides precise and robust 2D predictions, making it well suited for 3D model prompting.

Data: There are various datasets for 2D human pose estimation or instance segmentation. The most notable are COCO [10], MPII [69], and AIC [70]. Datasets such as OCHuman [6], CrowdPose [4], and CIHP [5] focus on multibody problems such as occlusion and self-occlusion. OCHuman is too small for large-scale training and is traditionally used only for evaluation. Furthermore, it contains annotation errors as shown in Sec. 4.1. CrowdPose is big enough for training but is unsuitable for evaluation in a multi-dataset setup as it mixes the train and test sets of COCO, MPII, and AIC. In addition to the pose estimation, other datasets focus on human crowds, e.g. CrowdHuman [3] on detection or CIHP [5] on human parsing. 3D human pose estimation datasets are mostly synthetic [71–73] or lab-captured [74–77], so they do not evaluate performance in the wild. Apart from COCO and OCHuman, there are not enough datasets to evaluate all three tasks (detection, segmentation and pose estimation) on the same dataset. We attempt to improve the situation with OCHuman-Pose in Sec. 4.1. For COCO and related datasets, the evaluation metric is Object Keypoint Similarity (OKS), while Percentage of Correct Keypoints (PCKh) is used for MPII.

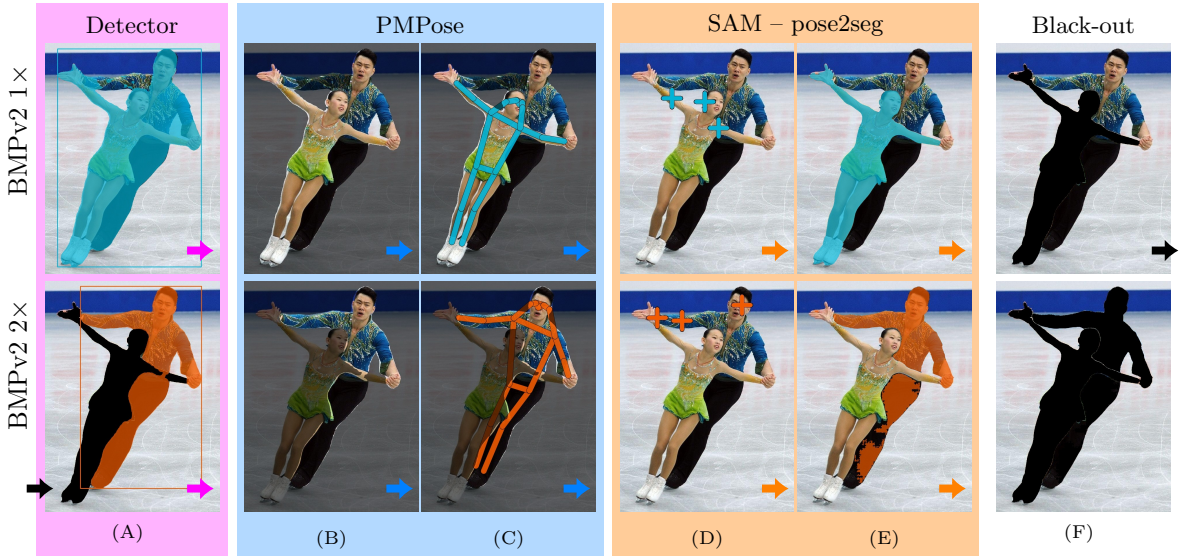


Fig. 2: **BBoxMaskPose v2 (BMPv2) loop visualization.** The method consists of three separate models – detector , PMPose and SAM-pose2seg – each conditioned by the others. A fine-tuned detector detects people in (A) while ignoring already blacked-out instances shown in (F). PMPose estimates pose (C) on person crops with dimmed background (B) to discriminate individuals in crowds. SAM-pose2seg refines masks (E) in full images conditioned by selected keypoints (D) to capture context. The loop ends after a fixed number of iterations or when no people are detected in (A). Image from OCHuman [6].

3 Method

BMPv1 [7] connects a detector, pose estimator, and segmenter in a self-improving loop as shown in Fig. 2. The core idea is mutual consistency between different human-body representations. Each model is conditioned on the others and iteratively refines its predictions. The loop ends when no people are detected.

The detector is conditioned on previously processed instances. After an instance is processed, its mask is blacked out, forcing the detector to ignore it in subsequent iterations. This enables discovery of missed instances, typically in the background, and recovers cases where multiple people collapse into a single detection.

The pose estimator is conditioned on a segmentation mask from the detector or the mask refinement stage. Masks are more precise than bounding boxes and help disentangle individuals under heavy overlap. The pose estimator in BMPv1, MaskPose, is a standard heatmap-based top-down model with mask conditioning.

The final component of BMPv1 is SAM, which refines segmentation masks using previous estimates and predicted poses. BMPv1 proposes a complex strategy for selecting pose-based point prompts. Despite this, mask refinement remains the dominant failure mode.

In this work, we revisit BMP and introduce two improvements. We replace MaskPose with PMPose and SAM2 with SAM-pose2seg. Both are introduced in the following sections. We further introduce a more expensive but more accurate variant, BMP+, and show how BMP integrates with 3D pose estimation.

3.1 PMPose

Human pose estimation is a central element of the BMP loop. It resolves detection ambiguity in crowds by enforcing anatomical constraints (e.g., each person has two arms) and validates detections before mask refinement. Accurate poses are critical for the final refinement stage; incorrect joints prevent the separation of overlapping individuals.

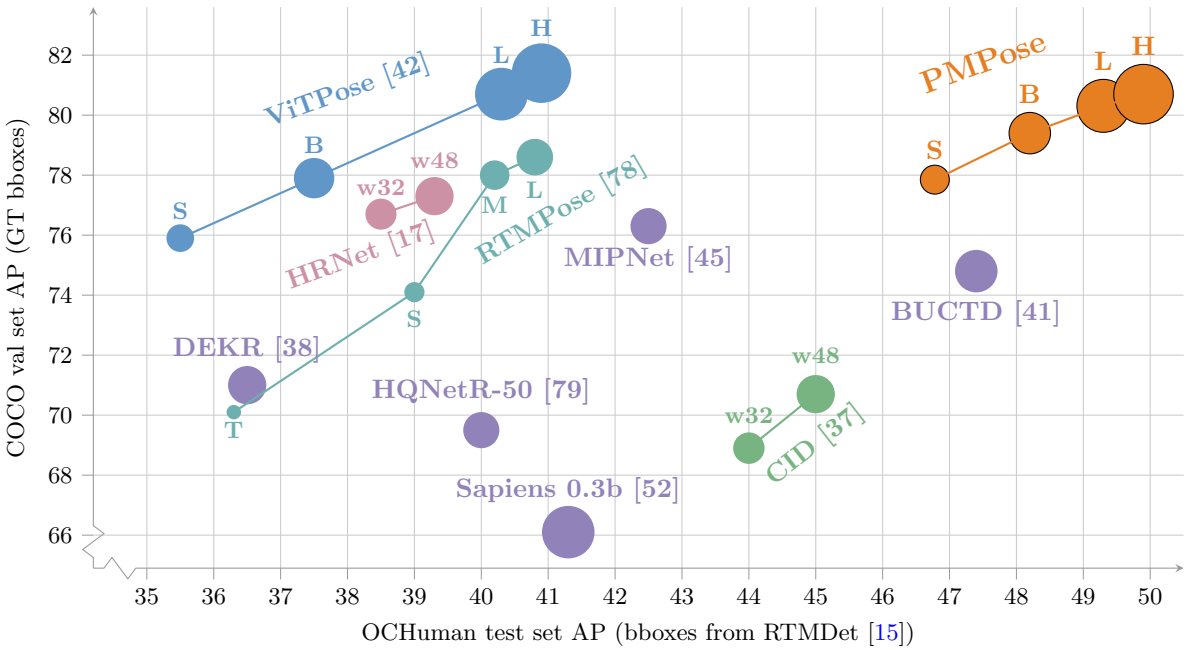


Fig. 3: **Comparison of PMPose with state-of-the-art models** on COCO and OCHuman. The bubble diameter is proportional to the logarithm of the number of model parameters. PMPose generalizes best to the unseen OCHuman dataset while keeping a SOTA-level performance on COCO.

The dominant failure modes are (1) merging people when predicted joints belong to multiple individuals and (2) missing disconnected parts, such as the wrists or ankles that are visually detached due to occlusion. BMPv1 uses MaskPose [7], which benefits from mask conditioning and reduces merges between individuals. However, MaskPose inherits limitations common in top-down pose estimators. First, because it is trained only on in-image keypoints, it tends to place joints inside the image even when they lie outside, causing instabilities along the bounding box border. Second, its confidence estimates are poorly calibrated. Confidence is crucial for selecting prompts in the mask refinement stage, and overconfident errors lead to incorrect prompt selection that propagates through the pipeline. We address these issues with PMPose (ProbMaskPose), a unified model for pose estimation in BMPv2.

PMPose combines the strengths of MaskPose and ProbPose [9]. It keeps mask conditioning from MaskPose and adopts the presence probability, visibility estimation, expected OKS prediction, and the resulting border stability from ProbPose. The ViT backbone remains unchanged; only the prediction head and the loss differ.

PMPose follows the same COCO+AIC+MPII training setup as MaskPose and is further fine-tuned with CropAugmentation on the same datasets. CropAugmentation enables the localization of keypoints outside the visible region and improves the presence and visibility predictions. For the semi-transparent visibility masking introduced by MaskPose, we use $\alpha = 0.25$ (compared to 0.2 in MaskPose), which is bit-shift friendly and speeds up augmentation.

PMPose has the same outputs as ProbPose [9]: keypoints with calibrated probability maps ($\mathcal{K}$), presence probability (p_p) indicating whether each joint is present, visibility estimates (v), and expected OKS ($\mathbb{E}_{OKS}$) as a self-quality measure. All of them are conditioned on an instance segmentation mask and bounding box as in MaskPose [7]. In the BMP loop, visibility is used to select prompts in the pose refinement stage (see Tab. 7) and $\mathbb{E}_{OKS}$ is used as a confidence measure for COCO-style AP evaluation.

Probability maps are calibrated with the same process (temperature scaling) as in ProbPose.

We quantize the probabilities in 5% increments and ensure each range contains approximately 5% ground truth keypoints of the CropCOCO validation set.

We introduce a whole family of models: PMPose-S/B/L/H, paired with ViT-S/B/L/H backbones [80]. The design is consistent across scales; capacity is the only difference. PMPose outperforms ViTPose [42] and improves over MaskPose, especially in crowded scenes (Fig. 3). On COCO, performance matches state-of-the-art levels. PMPose-S provides practical speed for real-time or near-real-time use, while PMPose-H targets the highest accuracy.

3.2 SAM-pose2seg

SAM-pose2seg builds on SAM 2.1 [8] and we first review the baseline to clarify our modifications.

SAM is trained on a large and diverse dataset, which gives it strong generalization to unseen data. However, SAM lacks explicit semantic understanding. Objects are loosely defined and may correspond to almost arbitrary regions.

Our goal is to preserve SAM's generalization while specializing it for pose-guided human instance segmentation. We therefore introduce minimal changes and keep most of the original architecture unchanged.

During training, SAM uses point prompts dynamically sampled from the ground-truth mask. For each mask, a single point is uniformly sampled and used to predict a mask. To simulate refinement, additional points are sampled from the error region, defined as the area where the prediction disagrees with the ground truth. This process is repeated for eight iterations, which results in the behavior observed in BMPv1 [7]: increasing the number of prompt points can degrade performance rather than improve it.

During both training and inference, SAM predicts three candidate masks per prompt. A mask is selected based on the predicted confidence score. This design partially addresses object ambiguity that is not explicitly resolved during the original training procedure.

3.2.1 Adapting SAM for the pose2seg task

We fine-tune from the original SAM 2.1. The backbone is frozen to preserve generalization. The architecture remains unchanged; only the training procedure is modified.

We adapt the model in two ways.

Human-focused decoder fine-tuning. We fine-tune the decoder to predict only human instances, including clothing as defined by the training annotations. This removes task ambiguity and directly addresses the main failure mode of BMPv1: over-segmentation of skin regions.

Pose-guided prompting during training. We replace random point prompting with pose-aware prompting to inject human keypoint information into the model. The first prompt is the most visible keypoint. In subsequent iterations, we select a keypoint located in the current error region. If no keypoint lies in the error region, we fall back to sampling a random point from that region. As a result, the model learns to operate on pose-driven prompts, which have a different spatial distribution than the random points used in the original SAM training.

This sampling strategy incorporates pose cues directly into training and better aligns the model with pose-guided human segmentation. Training uses human keypoints predicted with ProbPose [9], allowing the use of datasets without keypoint ground truth (CIHP [5]).

Fine-tuning also enables rapid specialization to the human class. Consequently, the complex prompting strategy identified as optimal in BMPv1 is no longer necessary. Selecting a small number of high-visibility keypoints yields comparable or better performance while being significantly simpler. Experiments show that two keypoints are sufficient to characterize an instance; we use three keypoints in the final model for slightly improved performance.

Sec. 5 analyzes how SAM-pose2seg improves the BMP loop and its connection to PMPose through visibility and presence probability prediction. Additional ablations are provided in the technical report [11].

3.3 BMP+

BMPv1 [7] ablations showed that re-running the pose estimator after mask refinement brought only marginal gains, not justifying the extra computation. This no longer holds for BMPv2. With stronger components—SAM-pose2seg for masks and PMPose for pose—looping them

yields significant gains. We therefore propose BMPv2+, a higher-performance, higher-cost variant of BMPv2.

Like BMPv2, BMPv2+ starts with instance detection, followed by pose estimation and mask refinement. The key difference is that pose estimation and mask refinement are looped until convergence. In practice, one additional PMPose pass is sufficient, as the second SAM-pose2seg step brings no further gains.

This loop increases computational cost but improves accuracy, particularly for human pose estimation. BMPv2+ is suited for use cases where performance outweighs runtime and resource constraints.

3.4 From 2D to 3D

BMPv1 [7] demonstrated strong robustness in crowded scenes. In BMPv2, we extend this capability to 3D by leveraging reliable 2D masks, boxes, and poses to reason about human bodies under dense interactions. Our goal is to recover accurate 3D poses and shapes despite extreme overlap and occlusion.

BMPv2 uses masks, bboxes, and 2D poses as prompts for a 3D human pose estimator. The 3D step is executed only after the BMP loop converges, as 3D models are computationally heavy. Although 3D estimation could be integrated into the loop similarly to the 2D pose, its cost would make iterative refinement impractical.

The 3D pose predictor encodes the full image once to obtain the image features. For each detected individual, it predicts 3D pose and shape by prompting with the estimated instance's bbox and mask. This avoids repeated feature extraction and keeps the per-person overhead manageable.

Predicted 3D bodies are filtered using 3D non-maximum suppression. Instance confidence for NMS is inherited from the 2D predictions produced by the BMP loop.

Quantitative results are in Sec. 5.4, with qualitative examples in Fig. 1.

4 Data

BMP targets scenes with very close human interactions where bounding boxes alone cannot separate individuals. Evaluating such models is difficult due to the lack of suitable data. To our knowledge, only CrowdPose [4] and OCHuman [6] focus on human pose estimation in such scenes. CrowdPose, however, is problematic in multi-dataset training due to data leakage, while Fig. 3 shows that multi-dataset training is essential for strong generalization. This leaves OCHuman as the only viable option, yet we show below that it is insufficient for in-the-wild detection and pose estimation.

The following two sections analyze the OCHuman and CIHP datasets, their statistics, how they differ, and what their weaknesses are. We show that annotations in OCHuman are incomplete and, therefore, that detection results on the dataset are misleading. To address this gap, we introduce an improved variant of OCHuman, OCHuman-Pose. The dataset is available on the project website.

Table 1: **The "crowded" datasets:** OCHuman, CIHP and the new OCHuman-Pose. $\overline{IoU_{max}}$ is the mean of IoU_{max} over all instances. In OCHuman-Pose, more than 50% of previously un-annotated instances are added to OCHuman images. CIHP is less crowded (see Fig. 4) but larger and its masks are more precise.

	OCH		OCH-Pose		CIHP
	val	test	val	test	val
# imgs	2 500	2 231	2 500	2 231	5 000
# kpts	4 291	3 819	6 546	5 863	0
# masks	4 291	3 819	4 291	3 819	17 520
$\overline{IoU_{max}}$	0.545	0.546	0.527	0.527	0.209

Table 2: **ViTPose*-b** [42] **performance** on OCHuman and new OCHuman-Pose using GT bounding boxes (top) and RTMDet [15] (bottom) bounding boxes as input. Missing annotations in OCHuman are evaluated as detection errors (false positives), masking pose estimation errors. On OCHuman-Pose, performance with GT and detected bboxes is much closer, which aligns with the qualitative analysis.

GT bboxes	OCHuman		OCHuman-Pose	
	val AP	test AP	val AP	test AP
✓	90.9	91.0	86.4	86.2
✗	44.5	44.1	75.3	76.1

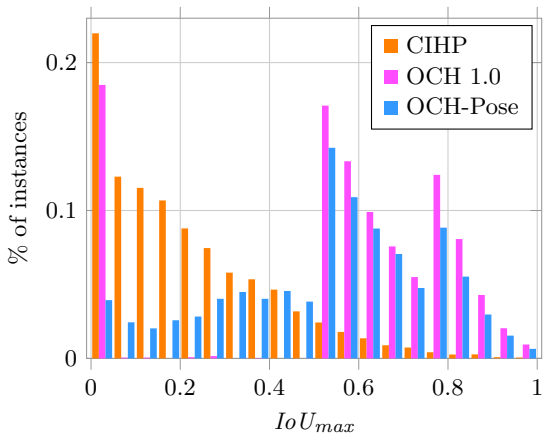


Fig. 4: **Instance density** measured by IoU_{max} in OCHuman, OCHuman-Pose and CIHP datasets. Despite having almost the same mean of IoU_{max} as OCHuman (Tab. 1), OCHuman-Pose has a more realistic distribution of in-the-wild images, free from artificial instance selections. CIHP has a different distribution because it focuses on crowds instead of close human interactions.



Fig. 5: **Missing annotations** in OCHuman. The originally annotated instances are often in the background, which causes problems in COCO-style AP evaluation. OCHuman-Pose expands the original dataset with new instances and previously ignored annotations.

4.1 OCHuman-Pose

OCHuman was originally created for the pose-to-segmentation task, where the goal is to estimate an instance mask given a human pose. Although not designed for general-purpose pose estimation, it became the de-facto benchmark for in-the-wild crowded human pose estimation due to the absence of better alternatives.

The main issue with OCHuman is that many individuals are not annotated. For each person, the dataset defines IoU_{max} as the maximum IoU with any other instance in the image. The authors kept only cases with $IoU_{max} = 0$ or $IoU_{max} > 0.5$, excluding all instances with intermediate overlap ($0 < IoU_{max} \leq 0.5$), as shown in Fig. 4. This filtering leaves numerous unannotated people, often placing the annotated subjects in the background, as shown in Fig. 5.

The COCO-style mAP evaluation [10] sorts the predictions by confidence and assigns the ground truth greedily. Any prediction that ranks first without a matching annotation becomes a false positive. On OCHuman, this forces models to assign the highest confidence to the annotated instance, even when it is not the most prominent person in the image. As illustrated in Fig. 5, there is no reason for these annotated individuals – often in the background – to receive the highest confidence, yet the evaluation protocol penalizes any alternative.

We address this issue by annotating all missing instances with keypoints. As shown in Fig. 5, some individuals had keypoints in the original OCHuman annotations but were excluded due to the dataset's filtering scheme. We reinstate these annotations and add manual annotations for all instances that lacked them.

Tab. 1 summarizes the statistics of the revised dataset. OCHuman-Pose adds over 50% new instances to the original images. Although the average IoU_{max} remains similar, the distribution now better reflects in-the-wild conditions (Fig. 4): the number of isolated cases with $IoU_{max} = 0$ is reduced and the previously missing range $0 < IoU_{max} \leq 0.5$ is now represented. Measuring pose estimation performance on OCHuman-Pose reflects reality more accurately than the original OCHuman as shown in Tab. 2. Annotation details are described at Suppl A.

Note that OCHuman-Pose does not include masks for the newly added persons, and thus cannot be used to evaluate segmentation mAP.

4.2 CIHP

Because OCHuman-Pose lacks masks and the original OCHuman annotates only a subset of instances, we require an additional dataset for evaluating instance segmentation in crowded scenes. CIHP contains dense crowds rather than close-contact interactions, leading to different bounding-box overlap patterns (see IoU_{max} distribution in Fig. 4). It is substantially larger than

OCHuman and provides more precise masks, but it has no pose annotations and includes some borderline cases (e.g., tiny head fragments).

Since CIHP is a different domain than OCHuman, we report segmentation results on both.

5 Experiments

Unless stated otherwise, we use the following models. RTMDet-L from [7] is used for detection.

PMPose-B is used for human pose estimation. It is trained on a mixture of COCO [10], MPII [69], and AIC [70]. For MPII and AIC, segmentation masks are generated using SAM2.1 [8] with ground-truth bounding boxes and poses as prompts. Mask augmentation is critical for generalization and robustness to detected masks.

SAM-pose2seg is built on a SAM-hiera-B+ backbone with our adaptation (see Sec. 3.2). It is fine-tuned on COCO and CIHP [5]. Since CIHP training data lacks pose annotations, pseudo ground-truth poses are generated using an earlier PMPose-B variant together with ground-truth segmentation masks.

Training details (e.g., schedules and learning rates), together with training scripts for reproducibility, are provided in Suppl. D and on the project website

5.1 2D Pose Estimation

Tab. 3 compares BMPv2 with the state-of-the-art in crowded human pose estimation. We show a comparison on the OCHuman dataset for comparison with previous approaches. More reliable comparison is on OCHuman-Pose with fixed annotations as explained in Sec. 4.1. Lastly, performance on COCO is shown for comparison in non-crowded scenarios.

Our new top-down method, PMPose-B, significantly outperforms the state-of-the-art among top-down methods in crowded scenes while keeping performance on COCO. PMPose performs comparably to BUCTD [41], which is an iterative method. As shown in Fig. 3, PMPose is superior to the compared pose models, setting a new state-of-the-art.

BMPv2 outperforms BMPv1 [7] in all measures on all datasets. BMPv2+ is even better and is the first method to exceed 50 AP and 80 AP performance on the OCHuman test and the OCHuman-Pose test, respectively.

BMPv2+ has slightly worse results on COCO compared to BMPv2 for two reasons. One, COCO contains many small instances (more than 30% of instances have a bbox size smaller than 100px). Mask refinement is not suitable for such small instances (most pose-guided segmentation methods such as [6, 29] ignore small instances in evaluation), and looping SAM-pose2seg with PMPose amplifies the error. Second, COCO contains annotation errors along the bounding box edges as shown in [9]. Therefore, ProbPose-like trained PMPose performs poorly on such keypoints.

The last trend worth mentioning is the performance difference between OCHuman and OCHuman-Pose. By annotating missing instances, the performance measure better corresponds to reality and shows the gap between detection error and pose estimation error. BMPv2+ has almost the same performance as ViTPose with GT bboxes (85.8 vs. 86.4) on val set and even outperforms ViTPose with GT bboxes on the test set (86.8 vs. 86.2). The results of ViTPose with GT bboxes are in Tab. 2. These results show that, with the BMP loop, the gap between GT and detected boxes is much smaller than on the original OCHuman. This suggests that detection is no longer the dominant bottleneck, and that errors in OCHuman-Pose arise to a large extent from pose estimation rather than detection.

5.2 Detection and Segmentation

Comparison with SOTA on detection and instance segmentation is in Tab. 4. The first part of the table shows standard detectors, the middle part shows pose-guided instance segmentation models, and the bottom part compares against BMPv1.

BMPv2 outperforms BMPv1 in all measures. The most dramatic increase is in instance segmentation thanks to the improved SAM-pose2seg.

The only measure where BMP is not SOTA is CIHP bounding-box AP. Here BMPv2 even underperforms BMPv1. The improved SAM-pose2seg is built for images with higher resolution and instances of people. CIHP contains small instances similarly to COCO. Further, CIHP contains many instances where only a very small part of the person is visible (e.g. the top of the head and no keypoints). For masks on CIHP, BMPv2 is

Table 3: **Pose estimation – comparison with state-of-the-art**. Best results in bold; second best underlined. ‡ results not reported. † uses a different detector. ¶ ignores *small* instances in COCO. Summary: PMPose improves MaskPose and sets the new SOTA for top-down approaches. BMPv2 improves BMPv1 [7] and BMPv2+ significantly outperforms the previous SOTA.

Model	OCHuman		OCHuman-Pose		COCO
	val AP	test AP	val AP	test AP	val AP
Methods without explicit detector (single stage + bottom-up)					
DEKR [38]	37.9	36.5	—‡	—‡	71.0
HQNet R-50¶ [79]	—‡	40.0	—‡	—‡	69.5
CID-w48 [37]	46.1	45.0	—‡	—‡	68.9
BUCTD [41]	48.3	47.4	—‡	—‡	74.8
Top-down methods; bounding boxes from RTMDet-l [15]					
Sapiens 0.3b [52]	42.0	41.3	69.8	70.5	66.1
MIPNet† [45]	42.0	42.5	—‡	—‡	76.3
ViTPose-B [42]	37.6	37.5	66.9	67.8	76.4
ViTPose*-B [42]	44.5	44.1	75.3	76.1	77.3
MaskPose-B [7]	46.6	46.6	77.1	78.0	76.8
PMPose-B	47.9	48.2	78.9	80.0	76.9
Iterative methods					
BUCTD 2× [41]	48.8	48.3	—‡	—‡	—‡
BMPv1 2× [7]	48.7	49.2	79.0	79.9	76.5
BMPv2 2×	<u>51.3</u>	<u>51.5</u>	<u>82.4</u>	<u>83.4</u>	**78.8**
BMPv2+ 2×	**55.8**	**55.8**	**85.8**	**86.8**	<u>78.1</u>

the best because the mask improvement for bigger instances outweighs wrongly detected small bounding boxes.

Since BMPv2+ converges after second run on PMPose (see Sec. 3.3), its results for detection and instance segmentation are the same as BMPv2.

5.3 Ablation Study

Tab. 5 shows contribution of SAM-pose2seg, PMPose and the 'plus' version. As expected, SAM-pose2seg improves instance segmentation even for BMPv1 due to its human-focused fine-tuning. Similarly, PMPose improves pose estimation even in BMPv1.

The most dramatic improvement is BMPv2+ which loops SAM-pose2seg and PMPose. Improvements in both models add more than 4 AP points on pose estimation.

Note that when SAM-pose2seg is used in BMPv1, it is prompted with BMPv1's complex prompt selection method for fair comparison. Since MaskPose [7] does not predict visibility, using BMPv2 prompting method would bring unfairly bad results.

5.3.1 PMPose Performance

The Tab. 6 compares PMPose with ViTPose [42], ProbPose [9], and MaskPose [7].

MaskPose introduced mask conditioning, which greatly improved pose estimation in crowded OCHuman scenes. ProbPose introduced presence probability (p_p) and improved visibility (v) estimation. Its probabilistic keypoint representation – presence probability, visibility, and calibrated probmaps – provides richer information for downstream processing and SAM prompting. As shown in Tab. 7, visibility estimation is especially useful for keypoint selection.

Mask conditioning, however, is a core component of the BMP loop. It is necessary for reliable instance separation in crowded scenes, which makes ProbPose harder to use in the loop. PMPose combines the strengths of both methods. It predicts visibility, presence probability, and calibrated probmaps on mask-conditioned images. These two components are complementary rather than conflicting. As shown in Tab. 6, PMPose achieves SOTA performance on both crowded OCHuman images and standard COCO images.

Table 4: Detection and segmentation – comparison with state-of-the-art. Best results in bold; second best underlined. ‡ results not reported. BMPv2 improves BMPv1 [7] especially in instance segmentation thanks to SAM-pose2seg. BMPv2+ results are the same as BMPv2, as the 'plus' version differs only by an additional PMPose iteration.

Model	OCHuman val AP bbox	OCHuman val AP mask	OCHuman test AP bbox	OCHuman test AP mask	OCHuman-Pose val AP bbox	OCHuman-Pose test AP bbox	CIHP val AP bbox	CIHP val AP mask
Detectors								
HRNet [17]	27.4	19.9	27.1	19.4	39.1	39.8	65.2	58.7
ConvNeXt [18]	29.2	19.8	29.4	20.4	41.4	42.2	69.1	60.5
HQNet R-50 [79]	30.6	31.5	29.5	31.1	—‡	—‡	—‡	—‡
CoDETR SWIN-L [12]	29.0	—‡	29.6	—‡	41.6	42.5	**73.5**	—‡
RTMDet-l [15]	31.1	27.1	30.0	26.2	45.7	45.7	69.5	63.9
Pose-guided segmentation methods								
Occlusion C&P [81]	—‡	28.9	—‡	28.3	—‡	—‡	—‡	—‡
ExPoSeg [29]	—‡	26.4	—‡	26.8	—‡	—‡	—‡	—‡
Crowd-SAM [32]	—‡	31.4	—‡	—‡	—‡	—‡	—‡	—‡
Iterative methods								
BMPv1 2× [7]	36.3	33.7	35.9	34.0	49.0	48.8	69.7	65.9
BMPv2 2×	**37.6**	**40.9**	**37.5**	**41.1**	**50.4**	**51.3**	68.8	**69.5**

Table 5: Ablation of BMP components. SAM-pose2seg improves mask AP, PMPose helps pose AP. SAM-pose2seg and PMPose in BMPv2+ bring dramatic pose improvements. BMPv1 with PMPose gives worse segmentation than standalone BMPv1, as the original prompting is optimized for MaskPose's confidence.

BMP version	OCHuman pose AP	OCHuman mask AP
BMPv1	49.2	34.0
BMPv1 + SAM-pose2seg	49.5	40.0
BMPv1 + PMPose	50.2	30.8
BMPv2	51.5	**41.1**
BMPv2+	**55.8**	**41.1**

The small drop from re-trained MaskPose-B to PMPose-B on COCO (79.8 → 79.4) is caused by incorrect border-point annotations in COCO. As explained in [9], COCO often places annotations inside the image even when the keypoint lies outside it. See [9], especially the supplementary material, for details.

Table 6: PMPose vs. similar models using ViT backbone [80] but different decoders and training schemes. Bounding boxes from RTMDet-L [15]. MaskPose† was re-trained with an improved training schedule compared to the [7]. PMPose is the only model combining mask conditioning with probabilistic outputs, and it keeps SOTA performance both in crowds and in standard images.

Model	mask input	$v + p_p$	COCO val AP	OCH test AP
ViTPose-S [42]	✗	✗	75.9	35.5
ViTPose-B [42]	✗	✗	77.9	37.5
MaskPose-B [7]	✓	✗	76.5	45.0
ProbPose-S [9]	✗	✓	76.6	37.7
MaskPose-S†	✓	✗	78.3	45.7
MaskPose-B†	✓	✗	**79.8**	46.9
PMPose-S	✓	✓	77.9	46.8
PMPose-B	✓	✓	79.4	**48.2**

5.3.2 Visibility Estimation for SAM-pose2seg

Tab. 7 shows how SAM 2.1 [8] and SAM-pose2seg behave when prompted with different keypoints. PMPose predicts (1) presence probability, (2) expected OKS and (3) visibility for each keypoint.

Table 7: SAM prompting with keypoints selected by different metrics. For SAM [8], prompts are chosen using the prompting strategy from [7], whereas SAM-pose2seg uses the top-3 points per metric. These configurations correspond to the optimal settings for each model. For confidence†, MaskPose [7] was used as PMPose does not return confidence. Visibility is the best predictor of usefulness for keypoint prompting.

Metric	COCO val AP	OCH test AP	CIHP val AP
SAM 2.1			
conf.†	32.7	24.6	58.6
pres. prob.	38.2	23.0	61.8
exp. OKS	35.4	25.3	64.5
vis	**41.2**	**29.5**	**71.6**
SAM-pose2seg			
conf.†	38.3	33.6	66.6
pres. prob.	43.8	31.0	69.3
exp. OKS	43.9	**34.7**	72.0
vis.	**44.6**	**34.7**	**72.7**

All three variables, along with MaskPose's confidence [7] can be used to select keypoints for prompting. In the experiment, SAM-pose2seg was fine-tuned for each variable for fair comparison. As shown in Tab. 7, PMPose's visibility is the best predictor of keypoint usefulness. PMPose is useful not only for improved localization but also because the information it provides about each keypoint can be used to select prompts more reliably.

Suppl. B expands on this ablation and addresses the follow-up question of why visibility works the best.

5.3.3 SAM-pose2seg qualitative examples

Fig. 6 illustrates a qualitative effect behind the mask AP gains of SAM-pose2seg in Tabs. 4 and 5. When prompted in crowded scenes, the original SAM 2.1 [8] often oversegments adjacent people or merges them with nearby human-shaped regions. After our human-focused fine-tuning described in Sec. 3.2, the decoder becomes substantially better at separating nearby instances under pose guidance. Because SAM-pose2seg predicts only the human class, it is less prone to the object ambiguity of the generic SAM decoder and therefore benefits from a much simpler prompting rule based on only the top-3 visible keypoints.



Fig. 6: Examples of reduced oversegmentation on OCHuman. Left: mask refined with SAM 2.1 [8] using up to 6 keypoints (optimal strategy from [7]). Right: mask refined with our SAM-pose2seg using up to 3 most visible keypoints.

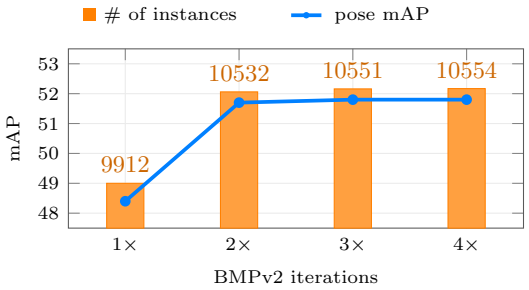


Fig. 7: Impact of BMP iterations on the number of predictions and pose estimation performance on OCHuman val [6]. After second iteration, additional iterations bring only marginal instance gains and just a few new instances.

5.3.4 Number of Iterations

Fig. 7 shows that the main benefit of looping appears already in the second BMP iteration. Pose mAP increases from 48.4 to 51.7, while the number of predictions rises from 9912 to 10532. These additional detections are typically previously ignored instances which become recoverable once earlier predictions are blacked out. Further

Table 8: Runtime of BMP components measured on a single P100 with 16 GB memory and batch size 1, time in ms. The second column shows if the component is per-image or per-instance. The most expensive components run only once per image.

Model	img	p05	p50	avg	p95
RTMDet-L [15]	✓	43.3	47.8	49.4	60.3
PMPose-B	✗	27.5	32.9	34.0	44.6
SAM bck.	✓	139.8	140.1	140.1	140.6
SAM dec.	✗	6.8	7.0	7.7	10.3
SAM-3DB [1]	✗	1347.3	1486.5	1471.2	1561.7
Sapiens 0.3b [52]	✗	548.3	562.0	565.4	583.1

iterations bring only diminishing returns: the third pass adds only 19 predictions and yields no meaningful mAP improvement, while the fourth adds only 3 predictions on the whole OCHuman validation set. By that point, nearly all foreground pixels have already been blacked-out and the detector does not discover any new instances.

5.3.5 Runtime Analysis

Tab. 8 breaks down the inference cost of individual BMP components. The most expensive part is the SAM backbone, followed by the detector, but both are executed only once per image. In contrast, PMPose and the SAM decoder are evaluated per instance, yet each call is relatively cheap, which makes iterative refinement practical. This explains the near-linear scaling in Fig. 8: runtime grows with the number of people, but the slope is controlled mainly by the lightweight per-instance stages rather than by repeatedly recomputing the heavy image-level features.

The broader comparison in Fig. 8 and Suppl. C also clarifies why we do not place large foundation models inside the loop. Even Sapiens 0.3b [52] is slower than two BMP iterations, while full SAM-3D-Body [1] is substantially more expensive than the entire BMP pipeline. For that reason, SAM-3D-Body is used only once as a downstream 3D prediction stage rather than as a recurrent component of the refinement loop.

5.4 3D Pose Estimation

Tab. 9 shows that BMP is suitable for SAM-3D-Body [1] prompting. SAM-3D-Body needs a prompt for mesh reconstruction. The prompt

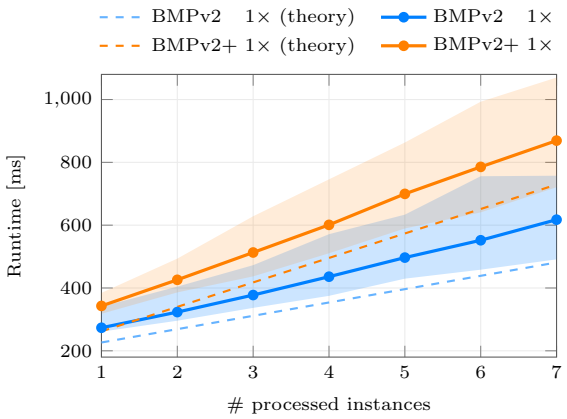


Fig. 8: BMP runtime scaling with the number of processed instances. Solid lines: median, shaded regions: the p05 to p95 range. Dashed lines: the theoretical runtime as a sum of components from Tab. 8. Measured on a P100 GPU with 16GB memory and batch size 1 on OCHuman. Images with >7 detected instances are too rare for statistical evaluation. Comparison with other models is in the Fig. C5 in Suppl. C.

could be a bounding box, a mask, or a 2D pose. Our experiments showed that pose does not help nearly as much as segmentation masks. When SAM-3D-Body is prompted with BMPv2 masks, predicted 3D meshes are of comparable quality to 2D poses. This suggests that precise masks simplify both 2D and 3D pose estimation in crowds.

Note that Tab. 9 measures the reprojection of 3D keypoints back to the 2D image, as there is no 3D ground truth for extremely crowded scenarios such as OCHuman. The reprojection measure is good enough to measure the crude correctness of the pose but cannot measure the depth ambiguity. However, depth ambiguity is usually not as large a problem in multi-body interactions as the correct assignment of limbs to individuals. Qualitative examples are in Figs. 1 and 10.

5.5 Robustness to Domain Shift

We evaluate BMP generalization on an unseen private dataset from [82]. The most challenging subset involves close adult-infant interactions, where standard pose estimators struggle to separate individuals. The dataset is further challenging because (1) adults are often only partially visible,

Table 9: **Prompting methods for SAM-3D-Body** [1] on OCHuman test. Pose AP is the reprojection score of the 3D pose. Prompting with BMPv2 gives significantly better 3D poses thanks to much better instance masks.

Detector	masks	OCHuman 3D pose AP	OCH-Pose 3D pose AP
ViTDet-H [16]	✗	36.8	59.6
RTMDet-L [15]	✗	38.5	68.0
BMP [7]	✗	39.9	68.3
BMPv2	✗	**46.4**	**75.8**
RTMDet-L [15]	✓	40.0	69.5
BMP [7]	✓	45.6	74.9
BMPv2	✓	**54.8**	**85.2**

which degrades performance of top-down methods, and (2) infants are absent from standard training sets, causing a strong domain shift. Due to privacy constraints, infant datasets are rarely shared, and models are typically not trained on infants. As the dataset is private, none of the evaluated models – including SAM – were trained on its images, resulting in substantial domain shift.

Tab. 10 and Fig. 9 report quantitative and qualitative results on the challenging lap subset. Both BMPv1 and BMPv2 significantly outperform off-the-shelf SOTA (RTMDet [15] + ViTPose [42]). PMPose, through probabilistic modeling, correctly predicts keypoints outside the image and avoids mixing individuals. The dominant failure mode is missed detections in some frames rather than incorrect association. Since evaluation is frame-based, such failures could be mitigated with temporal consistency in video.

BMPv2 achieves strong performance in complex multi-body scenes and is suitable when accuracy is the primary goal. For time-constrained applications, PMPose, especially PMPose-s, offers a favorable trade-off between accuracy and speed.

6 Conclusions

We present BBoxMaskPose v2 (BMPv2), an improved version of the BBoxMaskPose loop [7]. Performance gains come from a stronger 2D pose estimator, PMPose, and a pose-guided adaptation of SAM for human instance segmentation, SAM-pose2seg. BMPv2 outperforms BMPv1 across all metrics. It is the first method to exceed 50 AP on OCHuman and also maintains state-of-the-art performance on COCO.

Table 10: **Robustness to domain shift** on a non-public application dataset from [82]. PMPose performs similarly to ViTPose* [42] while providing more information about each keypoint. Both BMPv1 [7] and BMPv2 significantly improve over top-down methods. The private dataset guarantees that the data are from an unseen domain.

Method	pose AP
RTMDet-l [15] + ViTPose-B [42]	73.3
RTMDet-l [15] + ViTPose*-B [42]	76.9
RTMDet-l [15] + PMPose-B	76.4
BMPv1 2× [7]	83.2
BMPv2 2×	**83.3**

Key findings are:

1. Unifying MaskPose [7] and ProbPose [9] yields PMPose, a more accurate and expressive model improving performance in both standard and crowded scenes.
2. Adapting SAM-pose2seg [11] for pose-guided human instance segmentation significantly improves the BMP loop, beyond mask precision alone.
3. PMPose and SAM-pose2seg enable an additional pose–mask refinement loop, forming BMPv2+. Pose estimation improves substantially, as the second PMPose iteration operates on masks refined by SAM-pose2seg. BMPv2+ sets a new SOTA on OCHuman.
4. Poor performance on OCHuman [6] is caused by missing annotations. The proposed OCHuman-Pose dataset adds previously ignored instances to better reflect real scenes. On OCHuman-Pose, models using ground-truth and detected boxes have similar performance, indicating that pose estimation, not detection, is the main bottleneck.
5. Robust 2D prompts are key for accurate 3D human pose estimation in crowded scenes. With BMPv2, state-of-the-art prompted 3D models remain effective on OCHuman even under severe instance overlap.

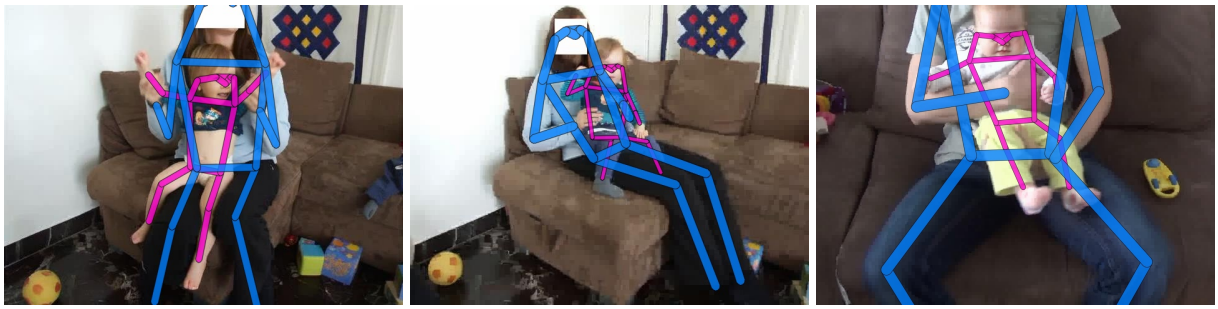


Fig. 9: **BMPv2 results on a non-public dataset** from [82]. BMPv2 is able to correctly disentangle individuals even in an unseen domain. Thanks to PMPose, keypoints outside of the image are not predicted inside. Results are shown only for authorized scenes; the dataset itself is more diverse than these examples. Infants anonymized with BLANKET [83].



Fig. 10: **Additional qualitative results** from OCHuman [6].

Appendix A OCHuman-Pose Annotations

We provide extended annotations to the OCHuman dataset [6] on the project website[2]. This section details the new dataset statistics, the annotation process, and the intended usage.

As mentioned in Sec. 4, OCHuman (OCHuman 1.0 or OCH 1.0 in this section) was originally created for the pose-to-segmentation task and contains many unannotated instances. Such instances are not problematic in the original intended usage, but they cause problems in evaluation protocols that include detection, as many instances are incorrectly counted as false positives.

A.1 OCHuman-Pose Statistics

OCHuman 1.0 contains three types of instance annotations, each with different annotation types. First are annotations with all three annotation types – bounding box, instance segmentation mask and COCO-style keypoints. There are $4291 + 3819 = 8110$ such instances (see Tab. 1).

Second, we have instances with annotated bounding boxes and skeletons but without masks. Such annotations are ignored in OCHuman 1.0 as they are useless for the pose-to-segmentation task. There are $821 + 739 = 1557$ such instances in OCHuman 1.0.

Lastly, there are instances with annotated bounding boxes but without mask and keypoints. OCHuman 1.0 contains $1434 + 1305 = 2739$ bounding-box-only instances.

OCHuman-Pose merges these three types and contains bounding box and COCO-style keypoint annotations for $4291 + 821 + 1434 = 6546$ annotations for the validation set and $3819+739+1305 = 5863$ annotations for the test set.

OCHuman 1.0 contains 5081 images of which only 4731 are used. There are 350 ignored images with bounding-box-only annotations. Since we do not know the reason why these images are ignored (at least some are duplicates), we ignore them in OCHuman-Pose as well to keep the same image distribution.

[2] https://MiraPurkrabek.github.io/BBox-Mask-Pose/

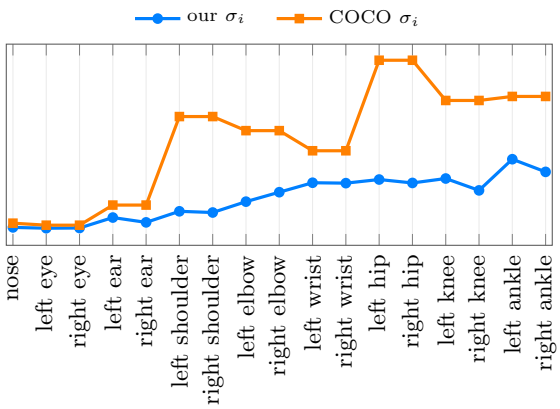


Fig. A1: Annotation quality of our new annotations vs. COCO annotations. Estimated per-keypoint sigmas are consistently lower than the original COCO values for most body keypoints, especially shoulders, hips, knees, and ankles.

A.2 Annotation Process

For instances with bounding-box-only annotations, we manually annotated COCO-style keypoints. The annotation was done by two professional full-time in-house annotators. An annotator annotated all 2739 new instances. The other annotator annotated 975 of the instances to measure variance. Additionally, a researcher with extensive experience in human pose estimation visually checked a random subset of ∼8% of annotations. If an error was found in the random subset, the annotator checked the entire dataset for similar errors.

Annotation error is compared to COCO annotation quality through per-keypoint σ_i in Fig. A1. Our annotations are more precise for two reasons. First, our annotators were trained on a small set of images to annotate the reprojection of the anatomical joint into the image. The careful training reduced the error especially for shoulders and hips, which are not well-defined joints and have high COCO annotation error. Second, our annotation tool visualizes left and right keypoints, which helps reduce confusion during the annotation process.

When using one set of annotations as GT and the other as prediction, we obtained somewhat surprising results. First, the annotators did not agree if the keypoint location is clear enough

for annotation (one annotated $v = 0$ while the other $v > 0$) in almost 25% of keypoints. Even though our two annotators agreed on keypoint location better than COCO annotators, human performance is only 85.4 mAP (98.4 AP@0.5) after filtering for visibility inconsistencies. The lower score is caused by long-tail errors in the challenging crowd data of OCHuman. As mentioned in [84], mAP above 85.0 is considered human annotation error so our annotations are at least on-par with COCO annotations even in more challenging crowded scenes.

For annotation, a dedicated GUI annotation tool was used. The tool was designed specifically for 2D human pose annotation and its design minimizes annotation errors such as left-right flip. The tool is available on GitHub[3].

Our annotation process did not add any new bounding boxes to OCHuman 1.0, we added annotations to existing bounding boxes. As a consequence, OCHuman-Pose still contains small instances in the background with no annotation. The problem is especially prevalent in sports images with blurry audience in the background (e.g. Fig. 10, rows 1 and 2). OCHuman-Pose mitigates the problem of incorrect false positives compared to OCHuman 1.0 but does not solve it completely. However, the COCO evaluation protocol is robust to such missing annotations as long as predicted background instances have smaller confidence than foreground instances. OCHuman-Pose is in line with this assumption since the missing annotations are only very small and insignificant instances, which was not true for OCHuman 1.0 (see Fig. 5).

A.3 Intended Usage

The annotation errors in OCHuman 1.0 were caused by using the dataset for unintended usage. The dataset was introduced in [6] as a crowded pose-to-segmentation task and only later was used for in-the-wild pose estimation. However, the original task did not include detection of a person and therefore did not need to contain all instances in an image. To prevent similar misunderstandings in the future, we lay out the intended usage of the OCHuman-Pose dataset. The list will not be complete as we are not able to predict future research directions, but a researcher should always keep in mind the limitations of the dataset.

First and foremost, the dataset is not intended for training. The OCHuman 1.0 dataset does not contain a training set, and similarly OCHuman-Pose is intended only for evaluation due to its small size. Our experiments suggested that even hyperparameter search on the validation set could lead to overfitting, especially with big models and sensitive parameters.

Second, OCHuman-Pose annotations do not contain segmentation masks. Even though OCHuman 1.0 masks could be mapped to a subset of OCHuman-Pose annotations, we removed them from the whole dataset to prevent confusion. Therefore, OCHuman-Pose is not suitable for mask-related tasks such as pose-to-segmentation or mask detection.

Third, as mentioned above, not all instances in all images are annotated and there are no "crowd" annotations. The dataset assumes that such predictions will have lower confidence than all predictions corresponding to annotated instances and the error will not show in COCO-style evaluation.

The same assumption applies to human pose evaluation. We believe that the assumption is reasonable for the in-the-wild human pose estimation task, which is the primary target of this dataset.

Appendix B Selecting Keypoints for SAM Prompting

Sec. 5.3.2 shows in the main paper that visibility is the best single variable for selecting SAM-pose2seg prompts. This supplementary section expands that ablation and addresses the follow-up question of why visibility works best, and whether combining it with another confidence-like variable can improve the selection further.

We analyze prompt selection through *in-mask* and *out-of-mask* keypoints. An *in-mask* keypoint is a keypoint which lies inside the ground-truth mask of the prompted person. Such a keypoint is therefore a correct positive prompt for SAM-like models such as SAM2 [8] and SAM-pose2seg. This notion is related to visibility, but it is not identical to it: a self-occluded keypoint can still lie inside the target mask and remain a correct positive prompt. Conversely, out-of-mask keypoints

[3] https://github.com/MiraPurkrabek/PoseAnnotator

are undesirable prompts because they typically correspond to another person, object or background.

Our analysis suggests that visibility is not merely a proxy for generic confidence. If that were the case, expected OKS or MaskPose confidence would be expected to behave similarly, but they do not. Instead, visibility is the variable most directly aligned with the actual prompting objective: selecting keypoints that fall inside the target mask. We also study multi-variable combinations and find that visibility can be improved slightly by adding presence probability, but the gain is negligible in the top-3 prompting regime used by SAM-pose2seg.

All experiments in this section are conducted on OCHuman validation set [6] unless stated otherwise. At the end of the section, we also discuss results on COCO [10] and CIHP [5].

B.1 Single-Variable Analysis

Fig. B2 compares four single-variable predictors: visibility, expected OKS, presence probability, and confidence. The figure overlays the densities of in-mask and out-of-mask keypoints for each score. The in-panel AUC is the 3-fold cross-validated ROC-AUC of the corresponding single-variable classifier, and p@3 measures the precision of the top-3 selected keypoints. The confidence plot is highlighted because it comes from MaskPose-B rather than PMPose.

Among the single-variable options, visibility is by far the best discriminator. Its in-mask distribution is strongly concentrated near 1, while out-of-mask keypoints accumulate near 0, which yields both the highest AUC and the best p@3. This directly explains the main-paper ablation in Tab. 7: visibility is the best standalone score for selecting positive prompts for SAM-pose2seg. While expected OKS and confidence also contain useful information, they do not separate in-mask from out-of-mask keypoints nearly as well as visibility. Presence probability is the weakest single variable. It is designed to capture whether the keypoint is in the image crop at all, not whether the keypoint lies inside the correct instance mask.

B.2 Multi-Variable Analysis

Fig. B3 studies whether combining PMPose variables further improves the prediction of in-mask status. The figure shows logistic models for all pairwise combinations and for the full three-variable model. The horizontal axis is therefore no longer a raw PMPose output, but the logistic predicted (σ) probability that the keypoint is in-mask.

The best pairwise combination is visibility with presence probability. It slightly improves over visibility alone, which suggests that presence probability carries a small amount of complementary information. Intuitively, presence probability is helpful in cases where the keypoint may be visually plausible but is near the crop boundary, outside the image, or otherwise uncertain in a way that visibility alone does not fully capture. However, the gain is extremely small. The visibility + expected OKS model is slightly worse, and expected OKS + presence probability is much worse, which again shows that visibility is the dominant variable. Using all three variables does not improve further; in fact, it is essentially tied with visibility + presence probability.

PMPose does not output MaskPose-style confidence, so we cannot form a like-for-like "visibility + confidence" model inside PMPose. However, expected OKS and presence probability are PMPose's confidence-like variables, and combining them with visibility yields at most a marginal improvement. Therefore, the main conclusion remains unchanged: visibility is the key variable, and extra variables provide only negligible gains.

B.3 Top-K Analysis

Binary classification analyzed in Secs. B.1 and B.2 is useful for understanding the overall discriminative power of each variable, but SAM-pose2seg does not operate in a binary classification regime. The target of scoring is to select K keypoints for SAM prompting to maximize IoU with the ground-truth mask. Therefore, analyzing the top-K behavior of each scoring strategy is more directly relevant to the actual prompting performance.

Fig. B4 compares the two most relevant ranking strategies: visibility alone and visibility + presence probability. For each instance, keypoints are sorted by the corresponding score and we measure precision@K and recall@K of the selected set. High precision means that most selected prompts

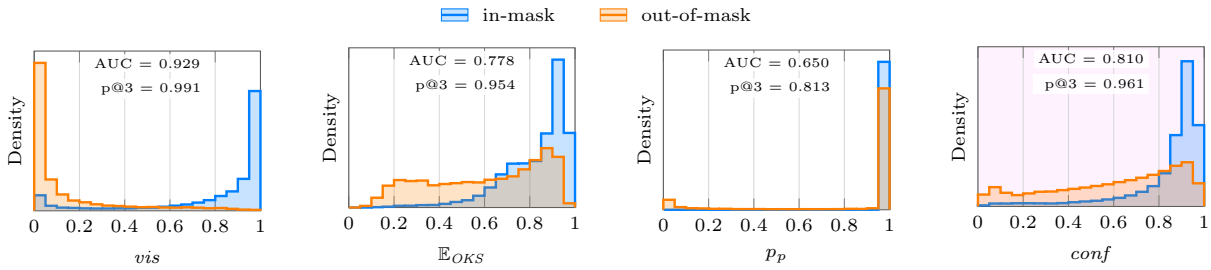


Fig. B2: **Single-variable analysis of prompt-selection scores** on OCHuman validation set [6]. In-mask points are keypoint predictions that lie inside the ground-truth instance mask. Visibility (*vis*), expected OKS ($\mathbb{E}_{OKS}$) and presence probability (p_p) are from PMPose-B while confidence (*conf*) comes from MaskPose-B [7]. The in-graph AUC is 3-fold cross-validated ROC-AUC of predicting in-mask from the variable, and p@3 is the per-instance precision of selecting the top-3 prompts by that score. Visibility yields the cleanest separation between in-mask and out-of-mask keypoints and the best top-3 precision. This makes visibility the strongest standalone criterion for prompt selection for SAM-like models.

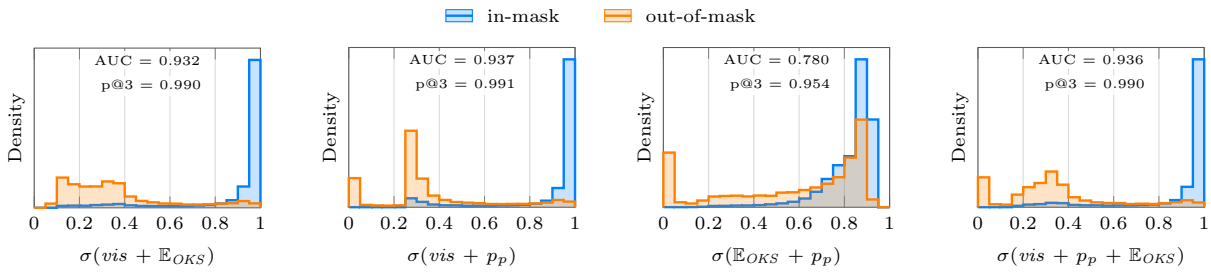


Fig. B3: **Multi-variable analysis of prompt-selection scores** on OCHuman validation set [6]. In-mask points are predictions in ground-truth instance masks. X-axis shows logistic predicted probability (σ) obtained from a model combining PMPose visibility (*vis*), presence probability (p_p), and expected OKS ($\mathbb{E}_{OKS}$). The in-graph AUC is 3-fold cross-validated ROC-AUC, and p@3 is the per-instance precision of selecting the top-3 prompts by that score. *Vis*+p_p is the best model, but only by a very small margin over visibility alone.

are correct positive prompts. High recall means that the ranking brings as many in-mask keypoints as possible to the front.

The two rankings are nearly indistinguishable up to roughly 7–8 prompts. This is especially important because SAM-pose2seg uses only the top 3 prompts. The insets in Fig. B4 show that for $K = 2$ and $K = 3$ the curves are effectively identical, and even at $K = 4$ the difference is negligible. Thus, while visibility + presence probability is the best multi-variable model in terms of global AUC, that advantage does not translate into a meaningful practical gain in the operating regime used by SAM-pose2seg.

For this reason we prefer visibility alone in the final method. It is the simplest model, it is already the best single-variable predictor, and the more complex alternatives do not provide a meaningful benefit at top-3 prompting. By Occam's razor, the visibility-only rule is therefore the most justified design choice.

We observed qualitatively similar behavior on COCO [10] and CIHP [5]. On COCO, the best model (by classification AUC) is the combination of all three PMPose variables, but the difference from visibility alone is again very small and does not change the practical top-K behavior. On CIHP, the results follow the same pattern

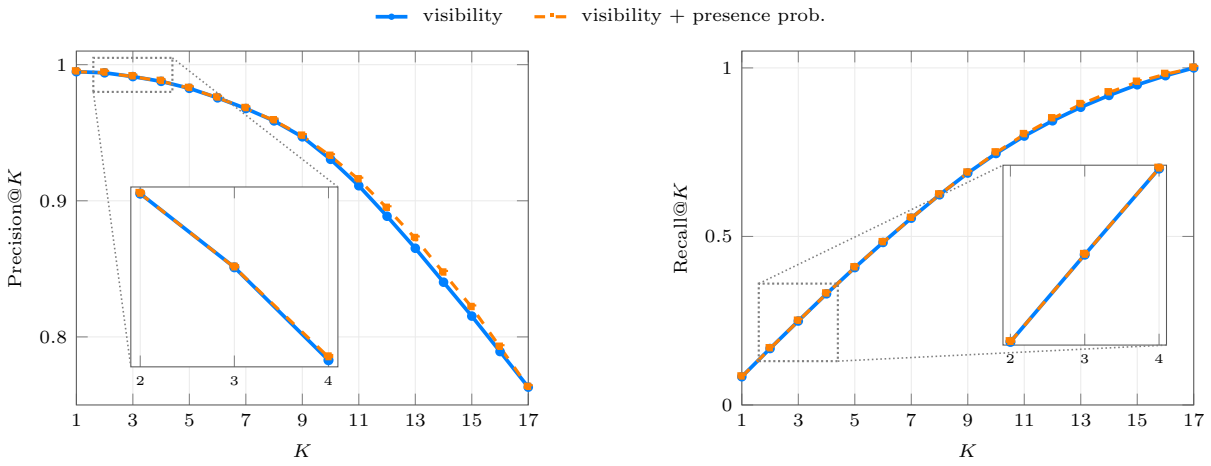


Fig. B4: **Average per-instance top-K precision and recall** of *vis* and *vis*+p_p models on OCHuman validation set [6]. The inset zooms focus on $K \in [2, 4]$, the operating regime most relevant to SAM-pose2seg. The two ranking strategies are virtually identical for $K \leq 5$ and remain very close until roughly 7–8 points. The slight global advantage of *vis*+p_p does not translate into a practical gain in the top-3 regime, so visibility alone remains the preferred rule.

as OCHuman [6], but the in-mask and out-of-mask classes are harder to separate overall, which is consistent with the denser crowds and more ambiguous masks in that dataset.

The analysis above implicitly assumes that all in-mask keypoints are equally good positive prompts once they are selected. This is likely only an approximation: some in-mask keypoints are probably more useful for SAM than others because of local shape ambiguity, occlusion pattern, or proximity to the mask boundary. However, evaluating prompt *usefulness* directly would require running SAM for many keypoint subsets, which quickly becomes combinatorially expensive, so more sophisticated prompt-utility modeling would be needed.

Appendix C Detailed Runtime Analysis

The Fig. C5 complements the component-wise runtime breakdown in Tab. 8 and the main-paper scaling plot in Fig. 8 by comparing BMP latency with other models. All measurements were run on the same machine with a single P100 GPU with 16 GB memory and batch size 1, and we report pure model inference time only, excluding data loading, preprocessing, visualization, and result saving. Per-instance runtime in Tab. 8 is averaged over the full OCHuman validation set [6], whereas Fig. C5 groups images by the number of detected instances and measures total per-image runtime within each bin. For fairness, Sapiens [52] and SAM-3DB [1] are both evaluated together with the same RTMDet-L detector [15]; SAM-3DB is measured in its full published pipeline, including FOV estimation, iterative hand refinement, and mask conditioning. This choice reflects the accuracy/runtime trade-off of the actual model used in the paper, and it also makes clear why SAM-3DB is too expensive to be used inside the iterative BMP loop in place of a 2D pose estimator such as PMPose.

Appendix D Implementation Details

Code, pre-trained weights, released annotation files, and runnable examples are provided on the project website. The linked GitHub repository contains the exact training scripts, configuration files, and inference code, while the appendix below summarizes only the main implementation choices that are useful for understanding the method. Our

goal here is not to duplicate the repository documentation, but to give a concise description of how the released models were obtained.

D.1 Fine-tuning RTMDet

We fine-tune RTMDet-L-ins [15] for 10 epochs on COCO [10] with batch size 16 and a linear learning-rate schedule with base learning rate 0.004. During fine-tuning, a random subset of instances of any class is blacked out to mimic the iterative BMP setting, and their masks are additionally perturbed by erosion or dilation to simulate imperfect masks from previous stages of the loop. The detector is trained on all 80 COCO classes, but only the person class is used in the experiments in this paper.

D.2 PMPose Training

PMPose is trained in multiple stages. We start from ViTPose [42], which itself uses MAE-style pre-training, and first train our MaskPose-style intermediate model on a mixture of COCO [10], MPII [69], and AIC [70]. For COCO we use the ground-truth masks directly, while for MPII and AIC we generate pseudo masks with SAM 2.1 [8] from ground-truth poses and bounding boxes. This stage follows the standard ViTPose schedule: 210 epochs, batch size 128, initial learning rate 5×10^{-4}, and step decays at epochs 170 and 200.

We then replace the MaskPose head with the PMPose head and continue training for 10 epochs with batch size 32 and initial learning rate 5×10^{-4} on the same COCO+MPII+AIC mixture. Throughout both stages, input masks are perturbed to imitate noisy detected masks; we use Voronoi-like mask growth/shrink operations and keep the CropAug strategy of ProbPose [9] with input padding 1.25. The final PMPose model predicts visibility, presence probability, expected OKS, and normalized probabilistic heatmaps in addition to the keypoint locations.

D.3 SAM-pose2seg Fine-tuning

SAM-pose2seg is built on SAM 2.1 Hiera Base Plus [8] and fine-tuned with the official training code on a mixture of COCO [10] and CIHP [5]. As described in Sec. 3.2, we keep the image encoder frozen and optimize only the prompt encoder and mask decoder, which preserves most of the original SAM representation while adapting the model to human pose-guided segmentation. For COCO, both ground-truth masks and ground-truth keypoints are available. CIHP does not provide pose annotations, so we generate pseudo keypoints with PMPose using the ground-truth masks as input.

In training, we use the pose-guided prompting strategy described in Sec. 3.2. We fine-tuned the model for 15 epochs with batch size 1 and initial learning rate 5×10^{-6}.

Funding

This work was supported by the Ministry of the Interior of the Czech Republic under project No. VL01010012 (FACIS2) within the IMPAKT 2 programme, the European Union's Digital Europe Programme under Contract No. 101158609, BMK, BMAW, and the State of Upper Austria within the framework of the SCCH competence center INTEGRATE [Project 1.6 TFI (Transferable Intelligence)], which is part of the FFG COMET Competence Centers for Excellent Technologies Programme, and the Czech Technical University student grant SGS26/074/OHK3/1T/13.

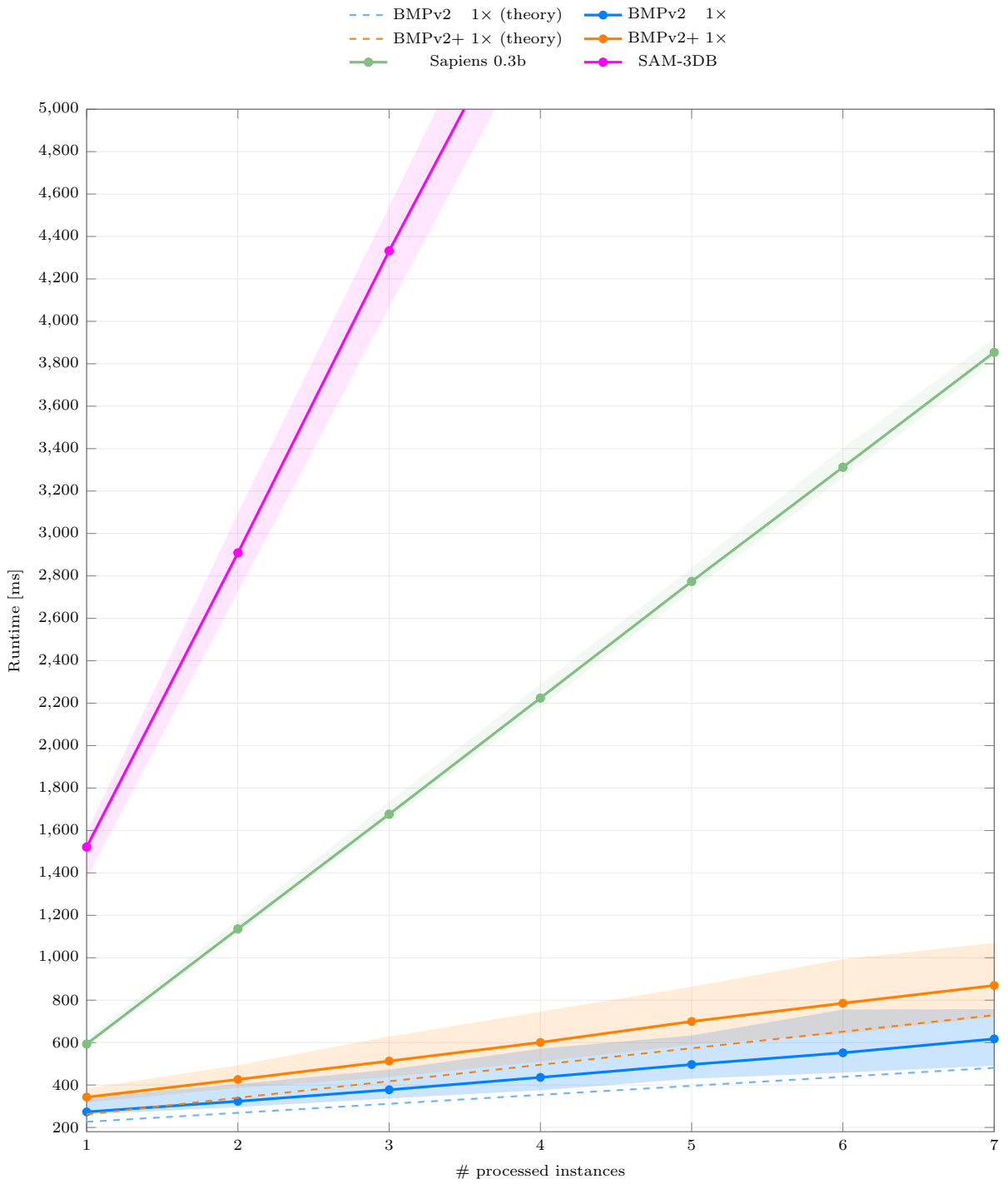


Fig. C5: **BMP, Sapiens [52] and SAM-3DB [1] runtime scaling with the number of processed instances.** Solid lines show median runtime, shaded regions show the range from p05 to p95. Dashed lines show theoretical runtime when naively summing component models. Measured on a P100 GPU with 16GB memory and batch size 1 on OCHuman val. There are too few images with more than 7 detected instances for statistical measurements.

References

[1] Yang, X., Kukreja, D., Pinkus, D., Fan, T., Park, J., Shin, S., Cao, J., Liu, J.-W., Ugrinovic, N., Sagar, A., *et al.*: Sam 3d body: Robust full-body human mesh recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7209–7219 (2026)

[2] Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: Prompthmr: Promptable human mesh recovery. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 1148–1159 (2025)

[3] Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. ArXiv **abs/1805.00123** (2018)

[4] Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.-S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[5] Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 770–785 (2018)

[6] Zhang, S.-H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.-M.: Pose2seg: Detection free human instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 889–898 (2019)

[7] Purkrabek, M., Matas, J.: Detection, pose estimation and segmentation for multiple bodies: Closing the virtuous circle. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9004–9013 (2025)

[8] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.-Y., Girshick, R., Dollar, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: The Thirteenth International Conference on Learning Representations (2025)

[9] Purkrabek, M., Matas, J.: ProbPose: A probabilistic approach to 2d human pose estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 27124–27133 (2025)

[10] Lin, T.-Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (2014)

[11] Kolomiiets, C., Purkrabek, M., Matas, J.: Sam-pose2seg: Pose-guided human instance segmentation in crowds. arXiv preprint arXiv:2601.08982 (2026)

[12] Zong, Z., Song, G., Liu, Y.: Detrs with collaborative hybrid assignments training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6748–6758 (2023)

[13] Ge, Z.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)

[14] Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)

[15] Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., Chen, K.: Rtmdet: An empirical study of designing real-time object detectors. ArXiv **abs/2212.07784** (2022)

[16] Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European Conference on Computer Vision, pp. 280–296 (2022). Springer

[17] Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019)

[18] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[19] Papandreou, G., Zhu, T.L., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.P.: Towards accurate multi-person pose estimation in the wild. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3711–3719 (2017)

[20] Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)

[21] Cuttano, C., Trivigno, G., Rosi, G., Masone, C., Averta, G.: Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 3395–3405 (2025)

[22] Wei, Z., Chen, P., Yu, X., Li, G., Jiao, J., Han, Z.: Semantic-aware sam for point-prompted instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3585–3594 (2024)

[23] Xiao, S.G., Kabra, R., Li, Y., Lee, D., Carreira, J., Panda, P.: Openworldsam: Extending sam2 for universal image segmentation with language prompts. Advances in Neural Information Processing Systems **38**, 35507–35532 (2026)

[24] Azarian, K., Das, D., Park, H., Porikli, F.: Test-time adaptation vs. training-time generalization: A case study in human instance segmentation using keypoints estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 411–420 (2023)

[25] Azarian, K., Das, D., Park, H., Porikli, F.M.: Test-time adaptation vs. training-time generalization: A case study in human instance segmentation using keypoints estimation. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), 411–420 (2022)

[26] Li, Z., Chen, X., Zhou, W., Zhang, Y., Yu, J.: Pose2body: Pose-guided human parts segmentation. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 640–645 (2019). https://doi.org/10.1109/ICME.2019.00116

[27] Tripathi, S., Collins, M.D., Brown, M.A., Belongie, S.J.: Pose2instance: Harnessing keypoints for person instance segmentation. ArXiv **abs/1704.01152** (2017)

[28] Ahmad, N., Khan, J., Kim, J.Y., Lee, Y.: Joint human pose estimation and instance segmentation with poseplusseg. In: AAAI Conference on Artificial Intelligence (2022)

[29] Zhou, D., He, Q.: Poseg: Pose-aware refinement network for human instance segmentation. IEEE Access **8**, 15007–15016 (2020) https://doi.org/10.1109/ACCESS.2020.2967147

[30] Ahmad, N., Khan, J., Kim, J.Y., Lee, Y.: Multiposeseg: Feedback knowledge transfer for multi-person pose estimation and instance segmentation. 2022 26th International Conference on Pattern Recognition (ICPR), 2086–2092 (2022)

[31] Yang, L., Song, Q., Wang, Z., Jiang, M.: Parsing r-cnn for instance-level human analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[32] Cai, Z., Gao, Y., Zheng, Y., Zhou, N., Huang, D.: Crowd-sam: Sam as a smart annotator for object detection in crowded scenes. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)

[33] Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024). Featured Certification

[34] Stoffl, L., Vidal, M., Mathis, A.: End-to-end trainable multi-instance pose estimation with transformers. arXiv preprint arXiv:2103.12115 (2021)

[35] Shi, D., Wei, X., Li, L., Ren, Y., Tan, W.: End-to-end multi-person pose estimation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11069–11078 (2022)

[36] Xiao, Y., Wang, X., Yu, D., Su, K., Jin, L., Song, M., Yan, S., Zhao, J.: Adaptivepose++: A powerful single-stage network for multi-person pose regression. arXiv preprint arXiv:2210.04014 (2022)

[37] Wang, D., Zhang, S.: Contextual instance decoupling for robust multi-person pose estimation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11050–11058 (2022)

[38] Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14676–14686 (2021)

[39] Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)

[40] Papandreou, G., Zhu, T.L., Chen, L.-C., Gidaris, S., Tompson, J., Murphy, K.P.: Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: European Conference on Computer Vision (2018)

[41] Zhou, M., Stoffl, L., Mathis, M.W., Mathis, A.: Rethinking pose estimation in crowds: overcoming the detection information bottleneck and ambiguity. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 14643–14653 (2023)

[42] Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (2022)

[43] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

[44] Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: Hrformer: High-resolution vision transformer for dense predict. Advances in neural information processing systems **34**, 7281–7293 (2021)

[45] Khirodkar, R., Chari, V., Agrawal, A., Tyagi, A.: Multi-instance pose networks: Rethinking top-down pose estimation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 3102–3111 (2021)

[46] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)

[47] Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 1653–1660 (2013)

[48] Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2602–2611 (2017)

[49] Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., Lu, C.: Human pose regression with residual log-likelihood estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11025–11034 (2021)

[50] Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 529–545 (2018)

[51] Luvizon, D.C., Tabia, H., Picard, D.: Human pose regression by combining indirect part detection and contextual information. Computers & Graphics **85**, 15–22 (2019)

[52] Khirodkar, R., Bagautdinov, T., Martinez, J., Su, Z., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: European Conference on Computer Vision (2024)

[53] Wang, Y., Wu, Y., He, W., Guo, X., Zhu, F., Bai, L., Zhao, R., Wu, J., He, T., Ouyang, W., *et al.*: Hulk: A universal knowledge translator for human-centric tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(7), 5672–5689 (2025)

[54] Ci, Y., Wang, Y., Chen, M., Tang, S., Bai, L., Zhu, F., Zhao, R., Yu, F., Qi, D., Ouyang, W.: Unihcp: A unified model for human-centric perceptions. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 17840–17852 (2023)

[55] S.A. Ghasemzadeh, A., Van Zandycke, G., Istasse, M., Sayez, N., Moshtaghpour, A., De Vleeschouwer, C.: Deepsportlab: a unified framework for ball detection, player instance segmentation and pose estimation in team sports scenes. In: 32nd British Machine Vision Conference 2021 (BMVC 2021) (2021)

[56] Rukhovich, D., Sofiiuk, K., Galeev, D., Barinova, O., Konushin, A.: Iterdet: iterative scheme for object detection in crowded environments. In: Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, S+ SSPR 2020, Padua, Italy, January 21–22, 2021, Proceedings, pp. 344–354 (2021). Springer

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

[58] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248–124816 (2015)

[59] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)

[60] Hesse, N., Pujades, S., Romero, J., Black, M.J., Bodensteiner, C., Arens, M., Hofmann, U.G., Tacke, U., Hadders-Algra, M., Weinberger, R., *et al.*: Learning an infant body model from rgb-d data for accurate full body motion analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 792–800 (2018). Springer

[61] Brégier, R., Fiche, G., Bravo-Sánchez, L., Lucas, T., Armando, M., Weinzaepfel, P., Rogez, G., Baradel, F.: Human mesh modeling for anny body. arXiv preprint arXiv:2511.03589 (2025)

[62] Keller, M., Werling, K., Shin, S., Delp, S., Pujades, S., Liu, C.K., Black, M.J.: From skin to skeleton: Towards biomechanically accurate 3d digital humans. ACM Transactions on Graphics (TOG) **42**(6), 1–12 (2023)

[63] Ferguson, A., Osman, A.A., Bescos, B., Stoll, C., Twigg, C., Lassner, C., Otte, D., Vignola, E., Bogo, F., Santesteban, I., et al.: Mhr: Momentum human rig. arXiv preprint arXiv:2511.15586 (2025)

[64] Park, J., Romero, J., Saito, S., Prada, F., Shiratori, T., Xu, Y., Bogo, F., Yu, S.-I., Kitani, K., Khirodkar, R.: Atlas: Decoupling skeletal and shape parameters for expressive parametric human modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6508–6518 (2025)

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

[66] Sárándi, I., Pons-Moll, G.: Neural localizer fields for continuous 3d human pose and shape estimation. Advances in Neural Information Processing Systems **37**, 140032–140065 (2024)

[67] Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1954–1963 (2021)

[68] Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12939–12948 (2021)

[69] Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)

[70] Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., Wang, Y., Wang, Y.: Ai challenger : A large-scale dataset for going deeper in image understanding. ArXiv **abs/1711.06475** (2017)

[71] Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 8726–8737 (2023)

[72] Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: CVPR (2017)

[73] Patel, P., Huang, C.-H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: AGORA: Avatars in geography optimized for regression analysis. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2021)

[74] Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 601–617 (2018)

[75] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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 (2014)

[76] Zhang, T., Huang, B., Wang, Y.: Object-occluded human shape and pose estimation from a single color image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7376–7385 (2020)

[77] Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3334–3342 (2015)

[78] Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: Rtmpose: Real-time multi-person pose estimation based on mmpose. arXiv preprint arXiv:2303.07399 (2023)

[79] Jin, S., Li, S., Li, T., Liu, W., Qian, C., Luo, P.: You only learn one query: learning unified human query for single-stage multi-person multi-task human-centric perception. In: European Conference on Computer Vision, pp. 126–146 (2024). Springer

[80] Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 45–67 (2021)

[81] Ling, E., Huang, D.-K., Hur, M.: Humans need not label more humans: Occlusion copy & paste for occluded human instance segmentation. In: British Machine Vision Conference (2022)

[82] Gama, F., Mísař, M., Navara, L., T. Popescu, S., Hoffmann, M.: Automatic infant 2d pose estimation from videos: comparing seven deep neural network methods. Behavior Research Methods **57**(10), 280 (2025)

[83] Hadera, D., Cech, J., Purkrabek, M., Hoffmann, M.: Blanket: Anonymizing faces in infant video recordings. In: 2025 IEEE International Conference on Development and Learning (ICDL), pp. 1–8 (2025). IEEE

[84] Ruggero Ronchi, M., Perona, P.: Benchmarking and error diagnosis in multi-instance pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 369–378 (2017)