HybridHOT: Learning Human-Centric Priors for Fine-Grained Human–Object Contact Segmentation

Qihui Chen[1], Junwen Chen[1], and Keiji Yanai[1]

The University of Electro-Communications, Tokyo, Japan
{chen-q,chen-j}@mm.inf.uec.ac.jp, yanai@cs.uec.ac.jp

Abstract. Human–object conTact (HOT) segmentation is the dense prediction task of assigning each image pixel to background or to the contacting human body part. Contact regions are often small and visually ambiguous, and the plausible contacting part depends on body configuration and scene geometry. We hypothesize that a key limitation is the form of human-centric guidance available to the model, and that fine-grained contact should be steered by priors that are both learned and spatially resolved. We present HybridHOT, a human-centric prior learning framework for HOT segmentation. Specifically, HybridHOT uses a human-specialized foundation encoder to produce two learned priors—a global part-presence prior and a dense per-part contact prior given by contact logits supervised with binary contact masks—which together guide a mask-transformer decoder. The decoder initializes its part queries from the global prior, fuses the dense contact prior into the mask logits, and conditions its pixel features on depth and person masks precomputed from the input image. On HOT-Annotated and HOT-Generated, HybridHOT achieves validation mIoU / AD-Acc of 0.396/0.682 and 0.406/0.607, respectively, substantially higher than the published results of the CLIP-based P3HOT on both datasets. The source code will be available at https://github.com/KeikiChen/HybridHOT.

Keywords: Human–object contact detection · Fine-grained contact localization · Human-centric priors · Body-part segmentation

1 Introduction

Understanding where people physically contact their surroundings is an important cue for human-centric scene understanding. Beyond recognizing that a person and an object co-occur, many downstream systems need to know where the body touches the world. A service robot receiving a handed-over cup must identify which hand is in contact with the object; an activity-analysis system separates "carrying" from "standing beside" by the contact configuration; and AR/VR and safety-monitoring applications rely on pixel-level contact to anchor virtual content or to flag unsafe poses. Human–object contact (HOT) segmentation formalizes this need as a dense prediction problem: given a single RGB

image, assign each pixel to background or to one of a fixed set of contacting body parts [5]. The pixel-level formulation provides the precise contact regions required by these downstream applications. Localizing contact at the pixel level is challenging for three coupled reasons. First, contact regions are small and have weak appearance cues: a hand on a table or a foot on a pedal occupies few pixels and is easily confused with the surrounding body or object. Second, contact depends on body configuration — the same object may be touched by a hand, shoulder, or hip — so appearance alone is often insufficient to identify the contacting part. Third, contact is a 3D geometric event observed in 2D, making cues such as depth and person extent important but noisy. Existing dense predictors often rely on generic features or global part cues, leaving the decoder with limited spatial guidance for contact localization.

Our key insight is to represent human-centric contact guidance with both global and dense learned priors. A global signal indicates which body parts are likely to participate in contact, while a dense per-part signal provides spatial evidence of where contact occurs. These signals are complementary: the former provides image-level semantic guidance, whereas the latter supports pixel-level localization. We therefore inject them at different decoding stages rather than using them only to rescale the final predictions. We further incorporate depth and person-mask cues through learned feature modulation and attention bias, allowing geometric information to guide contact prediction without fixed handcrafted rules.

Building on this insight, we present HybridHOT, a human-centric prior learning framework for human–object contact segmentation. A human-centric encoder produces multi-scale features, from which two lightweight heads predict a global contact-presence prior S_{cls} and a dense per-part contact prior S_{map}, supervised with binary contact masks. These priors guide a mask-transformer decoder: S_{cls} initializes the part queries, S_{map} is fused into the mask logits through a learnable gate, and the decoder conditions its pixel features on depth and person-mask geometry. Encoder and decoder comparisons indicate that the human-centric representation accounts for most of the improvement over the CLIP-based baseline, while prior-guided decoding provides additional gains in SC-Acc, C-Acc, and mIoU. Our contributions are:

- **Human-centric representation.** Our framework uses a human-centric encoder whose body-aware representations support fine-grained contact prediction. Comparisons under matched resolution and decoder indicate that this representation accounts for most of the improvement over the CLIP-based baseline.
- **Dual contact priors.** We learn a global contact-presence prior indicating which body parts participate in contact and a dense per-part prior estimating where contact occurs. The dense prior is supervised with binary contact masks and reused as decoder guidance rather than emitted as a standalone segmentation output.
- **Prior-guided decoding.** We inject the two priors at complementary stages of a mask-transformer decoder through query initialization and output-logit

fusion, while depth and person-mask cues condition pixel features and attention. Encoder×decoder and component-wise ablations analyze the contributions of representation, learned priors, and geometry-aware decoding.

2 Related Work

2.1 Human–Object Contact Detection

Human–object interaction detection localizes human–object pairs and recognizes interaction categories [8, 9, 15], but it usually does not identify the contacting body part or the pixel-level contact region. HOT detection instead targets dense, part-level contact localization by identifying the body regions in contact with objects. Chen et al. [5] introduced the task and the HOT dataset, casting it as part-level segmentation and using a part-attention branch to exploit surrounding body-part and scene context. Subsequent methods incorporate additional geometric and semantic cues, including pseudo-3D scene reconstruction [28], depth-aware perspective interaction with object texture restoration [29], and prompt guidance with human proximal perception and a regional joint loss [27]. These methods supplement contact estimation with geometric or semantic cues; we instead learn part-resolved spatial priors and inject them directly into the decoding process. Beyond whole-body HOT, a line of work addresses *specific* contacts, e.g., hand–object contact datasets [2] and foot–ground contact estimation from monocular video [23], typically over a limited set of part–object pairs. A parallel line estimates contact in *3D*: dense full-body human–scene contact [10, 12], human–scene interaction priors [11], in-the-wild 3D contact [25], and joint human–object reconstruction [32]; these require 3D supervision or scene geometry, whereas we target single-image, pixel-level, whole-body 2D contact. Most related to our work is the recent PaIR-Net [26], which jointly learns action recognition and contact localization and also exploits a body-part prior. PaIR-Net is evaluated on its own PaIR-1/PaIR-2 protocol rather than the HOT-Annotated/HOT-Generated benchmark, so direct numerical comparison is currently unavailable. PaIR-Net predicts a global descriptor of contact-relevant body parts and uses it to gate channels and rescale the human–object box; its dense contact prediction is subsequently reused by the action branch. HybridHOT uses a different prior formulation and decoding strategy. We learn both a global contact-presence prior and a part-resolved dense contact prior: the former initializes part queries, while the latter is explicitly mask-supervised and fused back into the segmentation logits. Thus, the dense signal in HybridHOT acts as an intermediate decoder prior rather than as a standalone prediction reused by another task branch.

2.2 Human-Centric Vision Models

A complementary line builds representations specialized to humans. General-purpose encoders transfer broadly but are not specialized for body parts, while

human-centric pretraining and architectures excel on dense human tasks, e.g. ViT pose baselines [33] and part-attention body regressors [17]. Sapiens [13] and Sapiens2 [14] are pretrained on large-scale human image collections. Sapiens supports body-part segmentation, depth, pose, and normal estimation; Sapiens2 extends the human-centric model family with strong body-part segmentation and other dense human-vision tasks. We adopt Sapiens2 as the encoder because its body-aware features provide part-level representations suitable for predicting our contact priors; a strong encoder alone, however, does not explicitly provide spatially resolved contact guidance, which motivates our priors and prior-guided decoder.

2.3 Query-Based and Mask-Transformer Segmentation

Query-based decoders are standard for segmentation: DETR [4] introduced object queries decoded by cross-attention, and MaskFormer and Mask2Former [6,7] predict masks as the dot product between learned queries and a per-pixel embedding, with masked attention focusing each query on its region. In standard mask transformers, the queries are generic and the pixel embedding is not explicitly conditioned on scene geometry. We build on this family but initialize part queries from a learned global prior and condition the pixel embedding on depth and person-mask geometry before decoding. Related mechanisms we use include feature-wise affine modulation [21], promptable person masks from SAM-style segmenters [3, 16], and monocular depth [1, 19].

3 Method

3.1 Problem Formulation

Let $I \in \mathbb{R}^{3\times H\times W}$ be an image. HOT segmentation assigns every pixel of a downsampled grid $h \times w = \frac{H}{4} \times \frac{W}{4}$ to one of $C = K + 1$ labels — background or one of $K = 17$ contacting body parts. We predict a logit volume $\hat{Y} \in \mathbb{R}^{C\times h\times w}$ and take the per-pixel prediction as $\arg\max_c \hat{Y}$. We additionally use two auxiliary geometric inputs, precomputed once from the same RGB image by off-the-shelf models (ZoeDepth [1] and text-prompted SAM [16]; supplementary Sec. A) and resampled to the output grid: a monocular depth map $D \in [0,1]^{h\times w}$, min–max normalized per image, and R person masks $M \in [0,1]^{R\times h\times w}$ (the ten largest during training and all masks at evaluation), from which we form a single foreground map by union, $\bar{m} = \mathrm{clip}\big(\sum_{r=1}^{R} M_r, 0, 1\big) \in [0,1]^{h\times w}$. We use learned human-centric priors to guide the prediction of $\hat{Y}$, rather than relying only on generic features and a single global zero-shot prior.

3.2 Overview

HybridHOT (Fig. 1) has two parts. A human-centric encoder $\mathcal{E}$ maps I to a feature pyramid $\{x_l\}_{l=1}^{4}$ and, from its deepest feature, two priors: a *global contact-presence prior* $S_{\mathrm{cls}} \in [0,1]^K$ (which body parts are in contact) and a *dense*

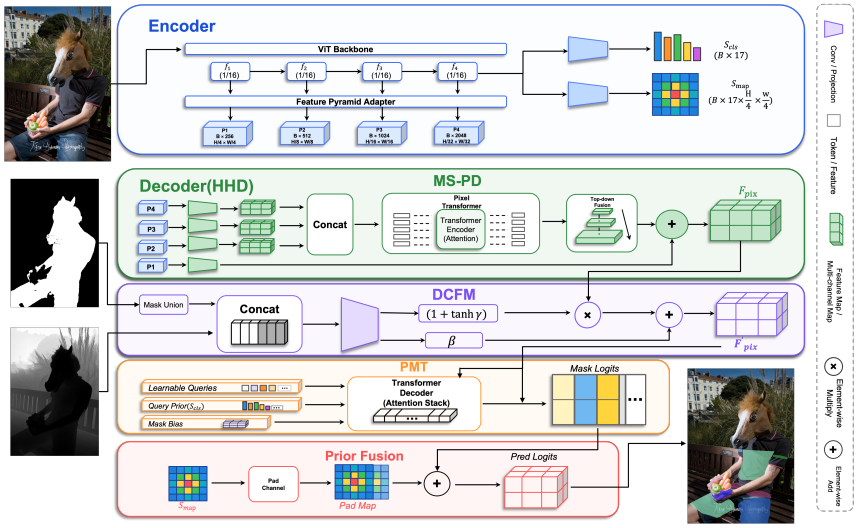


Fig. 1: Overview of HybridHOT. A human-centric encoder produces multi-scale features and two priors: a global contact-presence prior S_{cls} and a dense per-part contact prior S_{map}. The HHD decoder constructs a multi-scale pixel embedding (MS-PD), conditions it on depth and person-mask geometry (DCFM), decodes S_{cls}-conditioned person-aware queries (PMT), and fuses S_{map} into the final segmentation logits.

contact prior $S_{\text{map}} \in \mathbb{R}^{K\times h\times w}$, given by per-part contact logits that indicate where each body part is in contact. A prior-guided mask decoder $\mathcal{D}$ then predicts the output logits $\hat{Y}$ from the pyramid features $\{x_l\}$, the two priors S_{cls} and S_{map}, and the auxiliary geometric inputs D and $\bar{m}$. The two priors enter decoding at complementary stages: S_{cls} conditions the part queries and S_{map} is fused into the output.

3.3 Human-Centric Encoding and Prior Extraction

We instantiate $\mathcal{E}$ with a Sapiens2 ViT [14] and tap its features after blocks 6, 12, 18, and 24; all four taps lie on the same $\frac{1}{16}$-resolution patch grid. A lightweight pyramid adapter projects them with 1×1 convolutions and resamples them by factors of 4, 2, 1, and $\frac{1}{2}$ into the strided multi-scale layout $\{x_l\}$ (strides 4 to 32) that the decoder consumes. Writing f for the deepest tapped feature, we predict the two priors as

$$S_{\text{cls}} = \sigma\big(\text{MLP}(\text{GAP}(f))\big) \in [0,1]^K, \qquad S_{\text{map}} = \text{Up}_4\big(\text{Conv}(f)\big) \in \mathbb{R}^{K\times h\times w}, \quad (1)$$

where GAP is global average pooling and Up_4 is bilinear $\times 4$ upsampling. For body-part class k, $(S_{\text{cls}})_k$ is the image-level probability that at least one pixel of that class is in contact, whereas $(S_{\text{map}})_{k,u,v}$ is an unnormalized contact logit at location (u,v). The former offsets the corresponding foreground query before

decoding (Eq. (4)); the latter is directly supervised and added to the corresponding output logit through the learned scalar α (Eq. (7)). Since α starts at 0, early predictions rely on the mask decoder while the S_{map} head still learns from its own BCE loss. S_{map} is used only as an intermediate dense prior and is not emitted as a standalone segmentation output; its direct loss is not decoder-layer deep supervision.

3.4 Prior-Guided Mask Decoding

We instantiate the priors in a mask-transformer decoder (HHD). We use $N = C$ class-indexed queries, each predicting one class logit map, and decode in three steps: build a pixel embedding, condition it on scene geometry, and refine prior-conditioned queries against it. Its components are evaluated in Sec. 4.6.

Multi-scale pixel embedding. Lateral projections $p_l = \phi_l(x_l) \in \mathbb{R}^{d \times H_l \times W_l}$ ($d = 256$) map the pyramid to a common width. We contextualize the three coarsest levels ($l \geq 2$) with L_p pre-norm self-attention layers applied to their flattened union, using a level embedding e_l in each token and a 2D sinusoidal position embedding π_l in the attention queries and keys, and then fuse them top-down with the finest lateral p_1 through FPN-style aggregation [20] into a single high-resolution embedding

$$F = \mathrm{FPN}\Big(p_1,\ \mathrm{SA}\big(\{\,\mathrm{flat}(p_l) + e_l\,\}_{l\geq 2};\ \{\pi_l\}_{l\geq 2}\big)\Big) \in \mathbb{R}^{d\times h\times w}, \tag{2}$$

which lets distant pixels exchange context — useful for small contact regions whose disambiguation depends on the surrounding body and object.

Geometry-conditioned modulation. Contact localization benefits from cues about relative depth and human spatial extent. We therefore condition F on $(D, \bar{m})$ through feature-wise modulation [21]:

$$(\gamma, \beta) = \psi\big([\,D\,;\,\bar{m}\,]\big), \qquad \tilde{F} = (1 + \tanh\gamma) \odot F + \beta, \tag{3}$$

where ψ is a small convolutional network and $\odot$ is the element-wise product with spatially varying $\gamma, \beta \in \mathbb{R}^{d\times h\times w}$. The residual parameterization constrains the multiplicative modulation around an identity transform and avoids unconstrained feature scaling.

Prior-conditioned queries. We maintain N learnable base query embeddings $\bar{Q} = \{\bar{q}_n\}_{n=0}^{N-1} \subset \mathbb{R}^d$. The initialized queries are obtained by conditioning each foreground base query on its global contact-presence prior:

$$q_n^0 = \begin{cases} \bar{q}_0, & n = 0, \\ \bar{q}_n + W_s\big((S_{\text{cls}})_n\big), & n = 1, \ldots, K, \end{cases} \tag{4}$$

where $(S_{\text{cls}})_n$ is the n-th entry of the global prior and $W_s(\cdot)$ is a learned affine map $\mathbb{R} \to \mathbb{R}^d$, so each foreground query is conditioned on the predicted contact presence of its corresponding body part. Pixel tokens $V = \mathrm{flat}(\tilde{F})$ serve as values; $V + \pi$ serves as keys, where π is a 2D position embedding.

Person-aware masked decoding. Each of L_d decoder layers applies masked cross-attention, query self-attention, and an FFN. At layer ℓ the cross-attention uses an additive bias $A^\ell \in \mathbb{R}^{N\times hw}$ that combines region focusing with a person prior,

$$A^\ell_{n,j} = \underbrace{-\tau\,\mathbb{1}\big[m^{\ell-1}_{n,j} < 0\big]\,\big(1-\rho^\ell_n\big)}_{\text{focus on current region}} + \underbrace{\mathbb{1}[n\geq 1]\,\log\max(\bar{m}_j,\varepsilon)}_{\text{attend to people}}, \tag{5}$$

where $\mathbb{1}[\cdot]$ is again the indicator function, $\rho^\ell_n = \mathbb{1}\big[m^{\ell-1}_{n,j} < 0\ \ \forall j\big]$ re-enables a query whose attention row would otherwise be fully masked, $\tau = 10^4$ is a large finite penalty used in place of $-\infty$ for fp16/fp32 stability, $\varepsilon = 10^{-6}$ floors the person-mask union $\bar{m}$ before the log, and the previous-layer logits $m^{\ell-1}$ are detached. Before the first layer, m^0 is computed from the initialized queries q^0. Per-layer masks are read out as a dot product between projected queries and the pixel embedding,

$$m^\ell_n = \varphi_q\big(q^\ell_n\big)^\top \tilde{F} \ \in\ \mathbb{R}^{h\times w}. \tag{6}$$

Equations (4) to (6) give foreground queries a part prior (S_{cls}) and a person-mask bias ($\bar{m}$). The background query receives neither S_{cls} conditioning nor the person-mask attention term.

Output prior fusion. Finally the dense contact prior refines the last-layer logits through a learnable gate,

$$\hat{Y} = m^{L_d} + \alpha \cdot \mathrm{pad}(S_{\text{map}}), \tag{7}$$

where pad inserts a zero background channel and α (initialized to 0) is learned, so the model determines how strongly to incorporate the dense prior into the final segmentation logits. The two priors thus act at complementary points — S_{cls} at the query inputs (Eq. (4)) and S_{map} at the output (Eq. (7)) — which is what we mean by prior-guided decoding.

3.5 Training

The final output is supervised by a cross-entropy loss $\mathcal{L}_{\text{seg}}$ in which the background class is down-weighted by a factor of 0.02 to increase the relative contribution of small contact regions. The global prior uses $\mathcal{L}_{\text{sim}} = \mathrm{BCE}(S_{\text{cls}}, \bar{s})$, where $\bar{s}_k = 1$ iff body part k has at least one contact pixel. The dense prior uses binary cross-entropy with logits, $\mathcal{L}_{\text{smap}}$, against the per-part binary contact masks $\bar{P}$ on the output grid. The training objective is

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda\big(\mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{smap}}\big), \quad \lambda = 0.01. \tag{8}$$

4 Experiments

4.1 Environment Setup

We implement HybridHOT in PyTorch on top of a unified encoder/decoder builder, so that encoder and decoder can be swapped independently for the

ablations. The encoder is a Sapiens2-0.4b ViT initialized from segmentation-pretrained weights [14]; the decoder is a compact mask transformer with S_{map} fusion weight α initialized to 0. Optimization uses two AdamW optimizers, one for the encoder and one for the decoder, with weight decay 10^{-4} and polynomial learning-rate decay with power 0.9. For HybridHOT on HOT-Annotated, the initial learning rate is 5×10^{-5} for both encoder and decoder. The 224×224 control uses a 56×56 output grid; the main configuration trains and evaluates at 384×384 resolution with a 96×96 output grid. In the tables, our results obtained at 384×384 are marked with *. Full training, hardware, and software details, together with the evaluation protocol, are reported in the supplementary material (Sec. A).

4.2 Dataset

We evaluate on the HOT benchmark for pixel-level human–object contact segmentation [5]. The benchmark consists of two datasets, HOT-Annotated and HOT-Generated, constructed from manually annotated 2D contact sources including V-COCO [9], HAKE [18], and Watch-n-Patch [31], together with generated 3D-scene contact sources such as PROX [10]. The task defines 18 classes: background plus 17 body parts, namely head, chest, back, buttocks, hip, and the left/right upper arm, forearm, hand, thigh, calf, and foot. Following the HOT and P3HOT protocols, we train separate models on the two datasets and evaluate the reported HybridHOT results on their corresponding validation splits.

4.3 Metrics

We report five quantities: semantic contact accuracy (SC-Acc), contact accuracy (C-Acc), mean IoU over the 17 parts (mIoU), contact-frequency-weighted IoU (wIoU), and adaptive accuracy (AD-Acc); for all, higher is better. SC-Acc and C-Acc are reported in percent, while mIoU, wIoU, and AD-Acc are reported as decimal values. We use the metric definitions and the public evaluation code of the HOT [5] and P3HOT [27] benchmarks, and compute all metrics at the native output resolution against the correspondingly downsampled ground truth. AD-Acc counts false positives inside the person masks that are used as auxiliary input.

4.4 Comparison with State-of-the-Art Methods

Table 1 compares HybridHOT with published methods on both HOT-Annotated and HOT-Generated. The results of DHOT, PIHOT, and P3HOT are taken directly from the respective original papers, whereas the HybridHOT results are obtained by training and evaluating our model independently on the corresponding HOT-Annotated and HOT-Generated splits. We evaluate HybridHOT on the validation splits with the public evaluation code (Sec. 4.3). We do not include PaIR-Net [26], which additionally requires paired human/object/verb targets

Table 1: Comparison with published results on HOT-Annotated and HOT-Generated. Prior results are taken from the respective papers; ours are on the validation splits. *: 384 × 384 input.

Method	HOT-Annotated					HOT-Generated				
	SC-Acc (%)	C-Acc (%)	mIoU	wIoU	AD-Acc	SC-Acc (%)	C-Acc (%)	mIoU	wIoU	AD-Acc
DHOT (ResNet-50, full) [5]	40.70	70.70	0.215	0.260	–	30.40	54.30	0.139	0.167	–
DHOT (ResNet-50, full-OF) [5]	40.10	69.20	0.221	0.250	0.310	24.30	59.20	0.120	0.130	0.249
PIHOT (ResNet-50) [29]	45.30	80.70	0.236	0.286	0.313	34.90	76.30	0.169	0.212	0.254
P3HOT (CLIP RN50) [27]	46.00	74.90	0.256	0.302	0.423	35.20	70.80	0.180	0.231	0.306
HybridHOT (Sapiens2-0.4b)*	**63.11**	**91.95**	**0.396**	**0.402**	**0.682**	**66.72**	**86.63**	**0.406**	**0.557**	**0.607**

and reports results only on its own PaIR-1/PaIR-2 protocol; we contrast the two approaches conceptually in Sec. 2.

HybridHOT obtains higher values than all published results in Tab. 1 on every metric of both datasets. On HOT-Annotated, it raises mIoU from 0.256 to 0.396, AD-Acc from 0.423 to 0.682, and SC-Acc from 46.0 to 63.1 relative to the best published method; on HOT-Generated, it raises mIoU from 0.180 to 0.406, AD-Acc from 0.306 to 0.607, and SC-Acc from 35.2 to 66.7.

However, this comparison is not architecture-controlled. Existing HOT methods primarily use ResNet-50 encoders, including the CLIP RN50 image encoder used by P3HOT [22], whereas HybridHOT adopts the substantially larger Sapiens2-0.4b human-centric encoder and 384×384 input (marked *). We analyze encoder, decoder, resolution, and auxiliary-input effects separately in Sec. 4.6, and provide a fixed-decoder comparison with a large generic ViT, DINOv3-L [24], in the supplementary material (Sec. B).

4.5 Qualitative Analysis

Figures 2 and 3 compare DHOT, PIHOT, P3HOT, and our method in single-person and multi-person scenes, respectively. DHOT and PIHOT are evaluated using their publicly released checkpoints, while P3HOT is re-trained locally by following the official repository, since no public checkpoint is available. In single-person scenes (Fig. 2), HybridHOT produces more spatially coherent and fine-grained contact masks than the compared methods. DHOT and PIHOT often generate fragmented predictions or spread contact responses over large body regions, whereas our method better concentrates the predicted masks around physically plausible contact areas. Compared with P3HOT, HybridHOT better localizes several small and visually ambiguous contact regions, including hands, forearms, and feet. In multi-person scenes (Fig. 3), HybridHOT likewise produces localized part-contact regions in several cases, whereas the compared baselines spread predictions over broader body areas; small or visually ambiguous contacts, however, remain difficult and are occasionally missed. Because HOT provides semantic body-part contact labels rather than person-instance identities, these examples probe whether part-contact regions remain spatially localized in multi-person scenes, not whether each predicted region is assigned to the correct person instance. Finally, a remaining failure mode is left–right confusion between

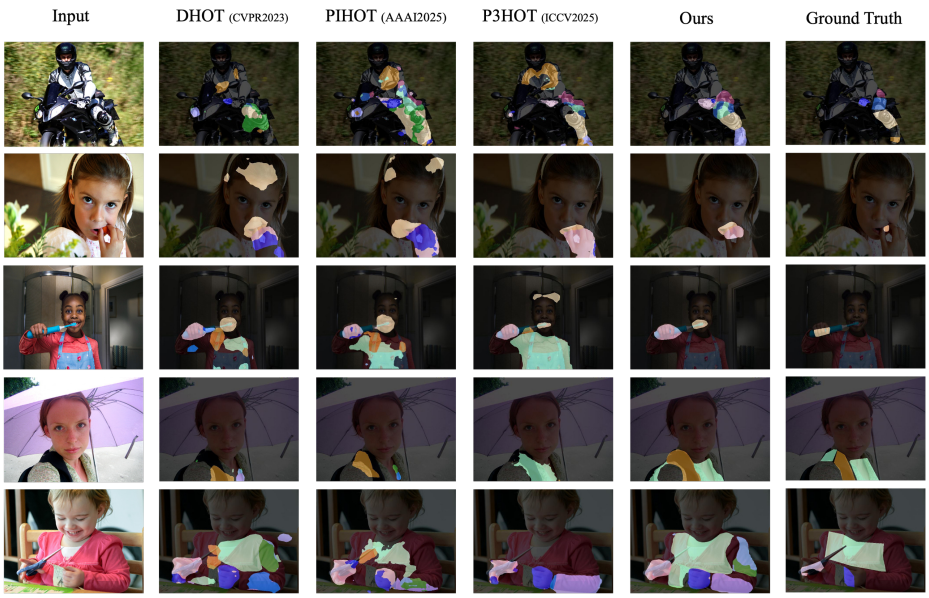


Fig. 2: Qualitative comparison on HOT-Annotated in single-person scenes. DHOT [5] and PIHOT [29]use released checkpoints; P3HOTP3HOT [27] is retrained using its official implementation.

symmetric body parts, where the contact region is localized but assigned to the symmetric counterpart; this issue, together with unannotated-but-plausible predictions, is analyzed in Sec. 5.

4.6 Ablation Studies

We analyze where the improvement comes from through encoder, decoder, resolution, prior, and input-source ablations on the HOT-Annotated validation split.

(1) Building HybridHOT from the baseline. Table 2 compares encoder and decoder combinations with the CLIP-based baseline; rows marked with * use 384×384 input and the others 224×224, so the Baseline and +Sapiens2 rows also differ in resolution. The matched-resolution controls in Tab. 3 show that replacing the baseline encoder with Sapiens2 raises mIoU from 0.238 to 0.347 at 224×224 and from 0.246 to 0.373 at 384×384, indicating that the larger gain comes from the human-centric encoder.

The HHD decoder provides smaller gains than the encoder replacement, but its effect is consistent on the core contact metrics. With the CLIP-based encoder (both at 224×224), replacing the U-Net-style decoder with HHD improves SC-Acc from 43.3 to 47.3, C-Acc from 73.3 to 83.5, and mIoU from 0.238 to 0.242. With the human-centric encoder (both at 384×384), HHD further improves SC-Acc, C-Acc, mIoU, and wIoU by 2.7, 2.6, 2.3, and 1.6 points, respectively, while

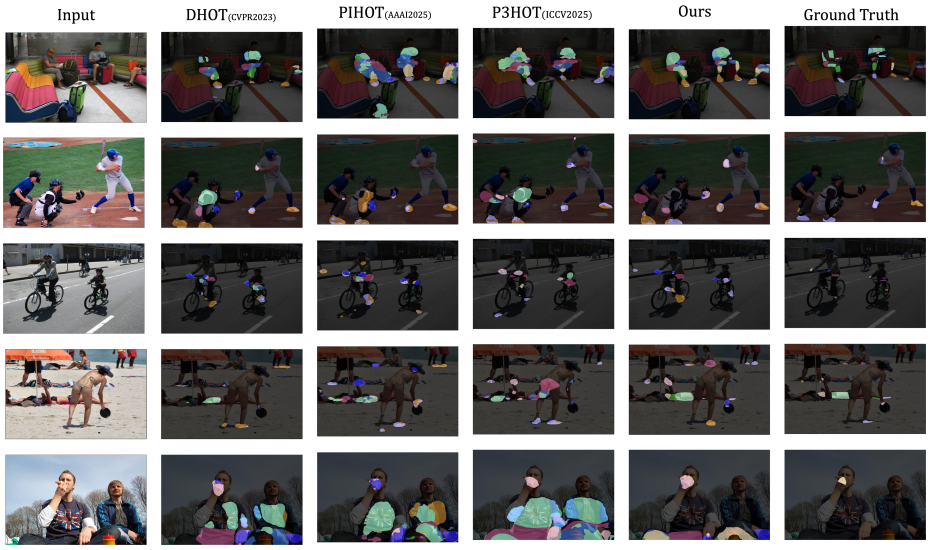


Fig. 3: Qualitative comparison on HOT-Annotated in multi-person scenes.

Table 2: Encoder and decoder comparisons on HOT-Annotated. The highlighted default HybridHOT combines Sapiens2 and HHD; +Data replaces the auxiliary inputs with SAM3 masks and Depth Anything 3 depth computed with Qwen-Image-Edit inpainting. Bold: best per column.*: 384×384 input.

Setting	SC-Acc (%)	C-Acc (%)	mIoU	wIoU	AD-Acc
Baseline (CLIP + U-Net)	43.30	73.33	0.238	0.280	0.441
+Sapiens2*	60.37	89.38	0.373	0.386	0.688
+HHD (CLIP + HHD)	47.27	83.45	0.242	0.280	0.408
+Sapiens2+HHD (full)*	**63.11**	**91.95**	**0.396**	0.402	0.682
+Sapiens2+HHD+Data*	61.95	89.27	0.388	**0.406**	**0.692**

AD-Acc changes slightly from 0.688 to 0.682. Thus, HHD consistently improves SC-Acc, C-Acc, and mIoU across both encoder settings, whereas its effects on the remaining metrics are mixed.

The decoder therefore mainly improves contact assignment rather than all overlap metrics uniformly. Its smaller gain with the CLIP-based encoder is consistent with both priors being predicted from encoder features: less human-specific features provide less informative part-level cues to the prior-guided decoder, so the benefit of HHD depends on the encoder representation rather than being a fixed additive gain.

The "+Data" row further shows that the choice of auxiliary-input sources affects the final metric profile. The alternative pipeline replaces the default person masks with SAM3 [3], the depth estimator with Depth Anything 3 [19], and the LaMa inpainting used for depth estimation with Qwen-Image-Edit [30], applied

Table 3: Resolution×encoder controls with a fixed U-Net-style decoder on HOT-Annotated. *: 384 × 384 input.

Encoder	Decoder	Res.	SC-Acc (%)	C-Acc (%)	mIoU	wIoU	AD-Acc
CLIP-based	U-Net-style	224	43.30	73.33	0.238	0.280	0.441
CLIP-based	U-Net-style	384*	43.01	73.09	0.246	0.293	0.470
Sapiens2	U-Net-style	224	**61.35**	**91.25**	0.347	0.374	0.597
Sapiens2	U-Net-style	384*	60.37	89.38	**0.373**	**0.386**	**0.688**

to all images of the dataset. This pipeline decreases SC-Acc, C-Acc, and mIoU, while slightly improving wIoU and AD-Acc. The alternative pipeline therefore changes different aspects of contact prediction in different directions. Because all auxiliary sources are changed jointly in this experiment, the contribution of each individual source cannot be isolated from this comparison alone.

(2) Isolating resolution from the encoder. Table 3 fixes the U-Net-style decoder and varies the encoder and the input resolution. Increasing the resolution from 224 to 384 changes mIoU only from 0.238 to 0.246 for CLIP and from 0.347 to 0.373 for Sapiens2, both far smaller than the gain from replacing the encoder at either resolution; the main improvement therefore cannot be explained by higher input resolution alone.

(3) Do the learned priors help, and how is each used? Table 4 ablates the S_{cls} and S_{map} heads on the Sapiens2+U-Net-style configuration. In this decoder, S_{cls} scales class outputs, whereas S_{map} provides auxiliary mask supervision during training but is not fused into the output at inference. Among the three ablated variants, removing both heads gives the lowest values on all five metrics (SC-Acc 47.61, C-Acc 65.49, mIoU 0.311), and adding either head alone improves all five (S_{cls} only: 58.59/81.75/0.387; S_{map} only: 54.44/77.49/0.352), showing that both heads contribute to contact prediction in this configuration. The two heads, however, favor different metrics. S_{cls} alone achieves the highest mIoU and wIoU, whereas S_{map} alone achieves the highest AD-Acc. Using both priors produces the highest SC-Acc and C-Acc.

Because the U-Net-style decoder does not consume S_{map} at inference, this table does not measure dense-prior fusion; its training and decoding roles in HHD are evaluated in the supplementary material (Sec. C).

(4) Decoder components. The HHD decoder combines multi-scale pixel embedding (MS-PD, Eq. (2)), query initialization from S_{cls} (Eq. (4)), the person-aware attention bias (Eq. (5)), geometry-conditioned modulation (DCFM, Eq. (3)), and dense-prior fusion at the output (Eq. (7)). Table 2 reports the aggregate effect of replacing the U-Net-style decoder with HHD under both encoder settings, and the supplementary material (Sec. C) ablates each element individually. In the component study, removing either query initialization or person-aware attention reduces mIoU and AD-Acc. The study also separates the two roles of

Table 4: Ablation of the S_{cls} and S_{map} heads with the Sapiens2 encoder and U-Net-style decoder. Here S_{map} supplies auxiliary training supervision and is not fused into the decoder output. Bold marks the best value in each column. *: 384×384 input.

Variant	SC-Acc (%)	C-Acc (%)	mIoU	wIoU	AD-Acc
w/o S_{cls} and S_{map}*	47.61	65.49	0.311	0.371	0.662
$-\,S_{\text{map}}$ (S_{cls} only)*	58.59	81.75	**0.387**	**0.415**	0.701
$-\,S_{\text{cls}}$ (S_{map} only)*	54.44	77.49	0.352	0.396	**0.703**
S_{cls}+S_{map} (both)*	**60.37**	**89.38**	0.373	0.386	0.688

S_{map}: keeping its BCE supervision but removing output fusion reduces mIoU from 0.396 to 0.353, whereas keeping fusion without supervision preserves mIoU but lowers SC-Acc and C-Acc, so supervision shapes the dense prior while fusion determines its effect on the final logits.

Removing MS-PD or DCFM mainly lowers AD-Acc, while the other metrics change in mixed directions; these components therefore trade off AD-Acc against the overlap and contact-accuracy metrics rather than improving all measures uniformly.

(5) Auxiliary-input sensitivity. Beyond the "+Data" row of Tab. 2, the component study in the supplementary material (Sec. C) measures the two geometry-conditioning pathways: relative to the main model, removing DCFM reduces AD-Acc by 0.012 and removing the person-attention bias by 0.032, with mixed effects on the other metrics. These ablations remove conditioning pathways and do not measure the effect of gradually degraded mask or depth quality, which we leave to a systematic perturbation study.

5 Discussion

Although HybridHOT improves both quantitative performance and qualitative contact localization, two limitations remain. First, the model sometimes predicts physically plausible contact regions that are not annotated in the ground truth, particularly around hand–object, arm–object, and body–object interfaces. Such predictions are counted as false positives by the current metrics, suggesting that part of the remaining error may reflect annotation ambiguity rather than model failure alone. A more exhaustively audited evaluation subset could help distinguish genuine false positives from plausible-but-unlabeled contacts.

Second, HybridHOT still exhibits left–right confusion between symmetric body parts, especially hands, arms, thighs, and feet. Because the two sides often have similar local appearance, correct labeling requires global pose, body orientation, and occlusion reasoning. In some cases, the model localizes the correct contact region but assigns it to the symmetric counterpart (Fig. 4). In the top example, the phone-holding hand is split between the left- and right-hand classes; in the bottom example, both feet are localized but their left/right labels



Fig. 4: Failure cases of left–right confusion. Columns: input, HybridHOT, ground truth. Top: the phone-holding hand is split between the left- and right-hand classes. Bottom: both feet are localized but their left/right labels are swapped.

are swapped. These cases indicate that contact localization and anatomical side identification remain partially separable challenges. Future work could incorporate stronger pose-aware supervision, left–right consistency constraints, and symmetry-aware objectives to improve robustness under occlusion and ambiguous body configurations.

6 Conclusion

We presented HybridHOT for pixel-level human–object contact segmentation, combining a human-centric encoder with a global contact-presence prior for query initialization, a dense per-part contact prior for output-logit fusion, and depth- and person-mask-conditioned decoding.

Encoder and decoder comparisons indicate that the human-centric representation accounts for most of the improvement over the CLIP-based baseline, while prior-guided decoding provides additional gains in SC-Acc, C-Acc, and mIoU across both encoder settings. On HOT-Annotated and HOT-Generated, HybridHOT exceeds all published results we compare against, reaching validation mIoU/AD-Acc of 0.396/0.682 and 0.406/0.607, respectively.

Our analysis also reveals plausible but unlabeled contact predictions and left–right confusion between symmetric parts, motivating stronger pose-aware reasoning and broader cross-scene evaluation.

Acknowledgements This work was supported by JST CRONOS JPMJCS24K4 and JSPS KAKENHI 26K23875.

References

1. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: ZoeDepth: Zero-shot transfer by combining relative and metric depth (2023), arXiv:2302.12288
2. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: ContactPose: A dataset of grasps with object contact and hand pose. In: ECCV. pp. 361–378 (2020). https://doi.org/10.1007/978-3-030-58601-0_22
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris Coll-Vinent, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: SAM 3: Segment anything with concepts. In: ICLR (2026)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020)
5. Chen, Y., Dwivedi, S.K., Black, M.J., Tzionas, D.: Detecting human-object contact in images. In: CVPR. pp. 17100–17110 (2023)
6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1290–1299 (2022)
7. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: NeurIPS (2021)
8. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: CVPR. pp. 8359–8367 (2018)
9. Gupta, S., Malik, J.: Visual semantic role labeling (2015), arXiv:1505.04474
10. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: ICCV. pp. 2282–2292 (2019)
11. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3d scenes by learning human-scene interaction. In: CVPR. pp. 14708–14718 (2021)
12. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: CVPR. pp. 13274–13285 (2022)
13. Khirodkar, R., Bagautdinov, T., Martinez, J., Su, Z., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: ECCV. pp. 206–228 (2024). https://doi.org/10.1007/978-3-031-73235-5_12
14. Khirodkar, R., Wen, H., Martinez, J., Dong, Y., Su, Z., Saito, S.: Sapiens2. In: ICLR (2026)
15. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: HOTR: End-to-end human-object interaction detection with transformers. In: CVPR. pp. 74–83 (2021)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
17. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3d human body estimation. In: ICCV. pp. 11127–11137 (2021)
18. Li, Y.L., Xu, L., Liu, X., Huang, X., Xu, Y., Chen, M., Ma, Z., Wang, S., Fang, H.S., Lu, C.: HAKE: Human activity knowledge engine (2019), arXiv:1904.06539
19. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. In: ICLR (2026)
20. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)

21. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: FiLM: Visual reasoning with a general conditioning layer. In: AAAI. vol. 32, pp. 3942–3951 (2018). https://doi.org/10.1609/aaai.v32i1.11671
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
23. Rempe, D., Guibas, L.J., Hertzmann, A., Russell, B., Villegas, R., Yang, J.: Contact and human dynamics from monocular video. In: ECCV. pp. 71–87 (2020)
24. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., et al.: DINOv3 (2025), arXiv:2508.10104
25. Tripathi, S., Chatterjee, A., Passy, J.C., Yi, H., Tzionas, D., Black, M.J.: DECO: Dense estimation of 3d human-scene contact in the wild. In: ICCV. pp. 8001–8013 (2023)
26. Wang, Y., Lei, Y., Liang, W., Xue, W., Wei, Z., Zhuang, N., Liu, Q.: What-Meets-Where: Unified learning of action and contact localization in images. In: AAAI. vol. 40, pp. 17832–17840 (2026). https://doi.org/10.1609/aaai.v40i21.38841
27. Wang, Y., Lei, Y., Wei, Z., Xue, W., Jiang, X., Zhuang, N., Liu, Q.: Prompt guidance and human proximal perception for HOT prediction with regional joint loss. In: ICCV. pp. 23636–23645 (2025)
28. Wang, Y., Lei, Y., Xiong, Q., Xue, W., Liu, Q., Wei, Z.: DeHOT: Reconstructing pseudo-3d scenes for human-object contact detection. In: Int. Conf. Comput. Big Data Artif. Intell. pp. 497–500 (2024). https://doi.org/10.1109/ICCBD-AI65562.2024.00088
29. Wang, Y., Neng, W., Wei, Z., Lei, Y., Xue, W., Zhuang, N., Xu, Y., Jiang, X., Liu, Q.: Precision-enhanced human-object contact detection via depth-aware perspective interaction and object texture restoration. In: AAAI. vol. 39, pp. 8187–8195 (2025). https://doi.org/10.1609/aaai.v39i8.32883
30. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-Image technical report (2025), arXiv:2508.02324
31. Wu, C., Zhang, J., Savarese, S., Saxena, A.: Watch-n-patch: Unsupervised understanding of actions and relations. In: CVPR. pp. 4362–4370 (2015)
32. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: CHORE: Contact, human and object reconstruction from a single rgb image. In: ECCV. pp. 125–145 (2022). https://doi.org/10.1007/978-3-031-20086-1_8
33. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: NeurIPS (2022)