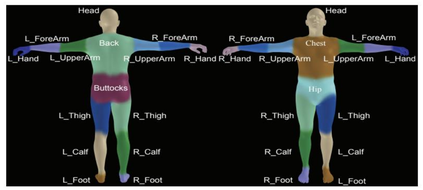
HybridHOT: Learning Human-Centric Priors for Fine-Grained Human-Object Contact Segmentation

Qihui Chen, Juwen Chen, Ynai Keiji ; The University of Electro-Communications, Tokyo

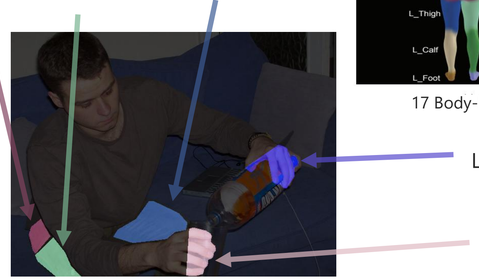
ACCV DEC 14-18, 2026 OSAKA JAPAN

Human-Object Contact Segmentation(HOT)

Left Thigh, Right Thigh and Buttocks rest on the sofa.



17 Body-Part labels(Rear and Font views)[1]



Left Hand holds the bottle.

Right Hand holds the glass.

The Human-Object conTact(HOT) is a pixel-level prediction task. Given as single RGB image, it labels each pixel as background or as one of 17 contacting body parts.

HOT datasets. The HOT benchmark contains 35,287 images and 162,267 labeled contact regions. HOT-Annotated uses manually drawn contact polygons, and HOT-Generated uses 3D human-scene contacts projected into 2D-images.

Introduction

- **Motivations**

Recent advances in embodied AI highlight the need to understand the physical world and how people interact with it. HOT task is gaining attention for its potential to support robotic manipulation and human-robot interaction by revealing where people touch objects.

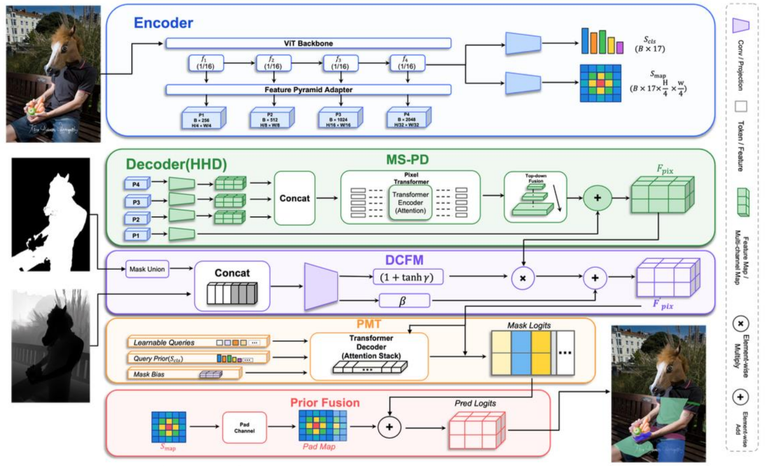
Prior work on HOT uses general-purpose pretrained encoders. P3HOT, for example, uses CLIP RN50, pretrained on web image-text pairs, to match images with body-part contact prompts. The resulting image-level scores indicate with parts may be in contact and provide limited spatial guidance for fine-grained body-part segmentation and pixel-level contact localization.

We aim to learn complementary contact priors using a human-centric visual backbone to better delineate contact regions on individual body parts and improve pixel-level segmentation accuracy.

- **Challenges**
  - Small contact region with weak visual cues.
  - Pose-dependent ambiguity in which body part touches an object.
  - Depth ambiguity in a single RGB image.

- **Contributions**
  - Use a human-centric backbone for fine-grained body-part feature.
  - Global and dense contact priors that identify which parts touch and where.
  - Prior-guided decoding through query initialization and logit fusion, with depth and person-mask cues.

Method



- **Human-Centric Encoder**

We initialize a human-centric backbone with body-part segmentation weights and extract features after ViT blocks 6, 12, 18, and 24. A lightweight adapter builds a feature pyramid with strides 4, 8, 16, and 32.

- **Human-Centric Priors**

From the deepest feature map, and MLP on globally pooled features predicts which body parts are in contact(S_{cls}). A convolutional head predicts where contact occurs on each part(S_{map}).

- **Hybrid Human Decoder(HHD)**

We build a Mask2Former-style decoder that uses monocular depth, person masks, and learned contact priors to guide pixel-level contact segmentation.

  - **MS-PD.** Cross-scale self-attention and top-down fusion combine multi-scale features into a stride-4 pixel embedding.
  - **DCFM.** Monocular depth and the union of person masks guide spatially varying scaling and shifting of pixel features.
  - **PMT.** Each class is assigned one query. The global prior S_{cls} offsets the initial foreground queries, and a person-mask bias guides their masked cross-attention.
  - **Prior Fusion.** The decoder adds a learned multiple of the dense prior S_{map} to the final foreground logits.

Results

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**

Results on HOT-Annotated and HOT-Generated datasets.

- **Comparison with Prior Work**

Our HybridHOT achieves state-of-the-art performance across all five reported metrics on HOT benchmark datasets, HOT-Annotated and HOT-Generated.

- **Prior and Encoder Ablation**

  - Either contact prior improves all five metrics over the no-prior baseline.

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_{map} (S_{cls} only)*	58.59	81.75	**0.387**	**0.415**	0.701
– S_{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

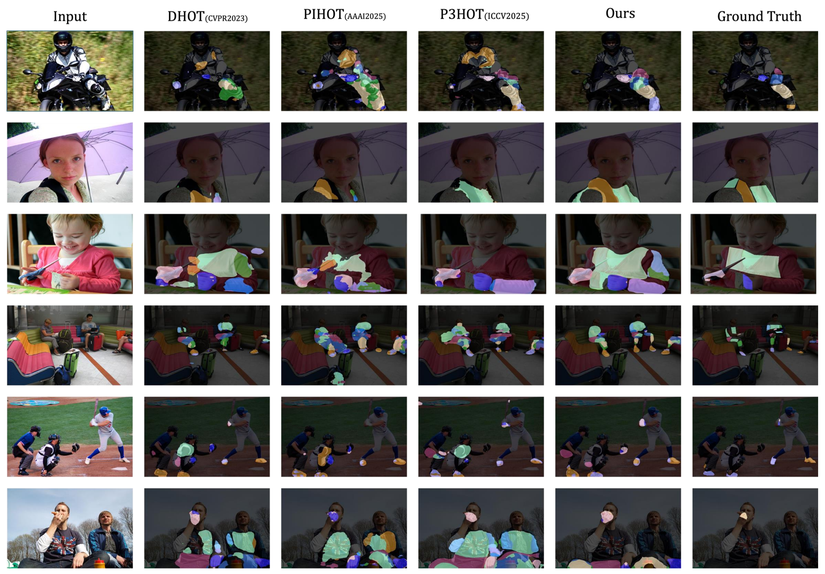
Human-Centric Prior Ablation

  - With the decoder held fixed, human-centric Sapiens2 outperforms general-purpose ViTs of similar scale across all five metrics.

Encoder	Encoder Params (M)	SC-Acc (%)	C-Acc (%)	mIoU	wIoU	AD-Acc
CLIP ViT-L/14 [7]*	303	42.42	79.03	0.188	0.235	0.347
DINOv3-L/16 [8]*	303	57.38	88.50	0.339	0.346	0.635
Sapiens2-0.4b*	395	**60.37**	**89.38**	**0.373**	**0.386**	**0.688**

Encoder Ablation

- **Qualitative Results**



Conclusion

HybridHOT achieves 0.396 mIoU and 0.682 AD-Acc on HOT-Annotated and 0.406 mIoU and 0.607 AD-Acc on HOT-Generated. Ablations on HOT-Annotated show that the human-centric encoder accounts for most of the mIoU gain, with prior-guided decoding providing a further improvement.

Reference & Acknowledgements

[1] Y. Wang et al. "Prompt Guidance and Human Proximal Perception for HOT Prediction with Regional Joint Loss." ICCV, 2025.

Supported by JST CRONOS JPMJCS24K4 and JSPS KAKENHI 26K23875.