InclusiveHuman-10K: Towards Inclusive Human Parsing Beyond the Intact-Limb Assumption

Heming Du[1]*, Jiaying Ying[1]*, Xiaofeng Cao[2], Zhu Li[3], Yuanyuan Liu[3], Kaihao Zhang[4], Xin Chen[3], and Xin Yu[5]†

[1] The University of Queensland [2] Tongji University [3] Hangzhou Dianzi University
[4] Australian National University [5] The University of Adelaide
Project Page *Equal Contribution †Corresponding Author

Fig. 1: Demonstration of InclusiveHuman-10K (IH-10K). We present an inclusive body parsing dataset for individuals with limb deficiencies. Different colors denote different body part labels. IH-10K provides pixel-wise annotations for residual limbs, prosthetic components, and mobility assistive devices. We fill a gap in human parsing benchmarks and thus enable inclusive evaluation on the underrepresented population.

Abstract. Human body parsing datasets and benchmarks define part labels for intact anatomies and focus on body parts and clothing. Widely used datasets implicitly assume intact anatomies and therefore do not include residual-limb or prosthetic classes. This leaves people with limb deficiencies excluded from body parsing model training and standard evaluation, creating a fairness gap. To the best of our knowledge, InclusiveHuman-10K (IH-10K) is the first human parsing benchmark centered on individuals with limb deficiencies, with explicit labels for residual limbs, prostheses, and mobility assistive devices. Specifically, IH-10K contains more than 10k images and around 15k person instances from everyday, sports, and clinical scenes, each with pixel-wise body parsing masks. This benchmark can support e-commerce, human-machine interaction, rehabilitation and assistive technology, human-centric image understanding, and editing. Moreover, we extend conventional body parsing categories with explicit classes for residual limbs, prostheses, and mobility assistive devices. On our IH-10K, body parsing models can be trained and evaluated on disability-related regions for this population with limb deficiencies, instead of treating them as background or noise. To assess

whether disability-related regions are systematically underserved, we report group-wise evaluation metrics across disability-related and common classes. Our experiments show that supervised models struggle on these regions due to large appearance variation and broken limb structure. Meanwhile, zero-shot methods perform poorly, suggesting that these concepts are absent from previous benchmarks and under-covered in open-vocabulary pretraining. This further highlights the representational and measurement gaps and signifies the necessity of our IH-10K benchmark. We hope this work helps build vision systems that serve people with limb deficiencies more reliably and fairly, increasing the likelihood of meaningful social benefit.

1 Introduction

Human body parsing is a core building block for human-centric vision systems, so it should work for everyone. It defines pixel-level part labels and underpins a wide range of applications, including augmented and virtual reality interaction, virtual try-on, animation, and activity analysis in sports and rehabilitation. Existing datasets such as LIP [28], CIHP [27], and MHP v2 [56] implicitly assume intact anatomies, including parts such as hair, face, arms, legs, and clothing, but lack residual-limb or prosthetic classes. Without explicit labels, disability-related regions are often absorbed into "background" during training and evaluation, which hides failures and prevents targeted improvements. As shown in Figure 2, UniParser trained separately on MHP v2 and CIHP often misses residual-limb regions and mislabels prosthetic parts. This omission leaves individuals with limb deficiencies unseen during training and unmeasured during evaluation, creating a fairness gap in current human parsing practice.

However, people with limb deficiencies form a large and growing group[1] that is largely overlooked by current human parsing benchmarks. Guidance from the World Health Organization[2] estimates that about 0.5% of the population needs prosthetics or orthotics at any time, which implies tens of millions worldwide. This gap is not only about fairness. It also affects reliability in assistive, rehabilitation, and clinical workflows that depend on part-level body understanding. When these individuals participate in daily life, sports, and clinical activities, parsing models should explicitly identify residual limbs and prosthetic regions rather than classifying them into the background, which is crucial for downstream systems that rely on part-level body understanding and editing. Failing to identify disability-related anatomy would reduce the reliability and equity of downstream applications.

To close this gap in existing research, we introduce InclusiveHuman-10K (IH-10K), a 10K-image human parsing benchmark centered on individuals with limb deficiencies. Our goal is to increase the likelihood of meaningful social benefit by making disability-related regions first-class labels with clear, reproducible evaluation. Figure 1 gives an overview of IH-10K and the added disability-related

[1] WHO Assistive technology factsheet.
[2] Service Provision Guidance for Prosthetic & Orthotic Services

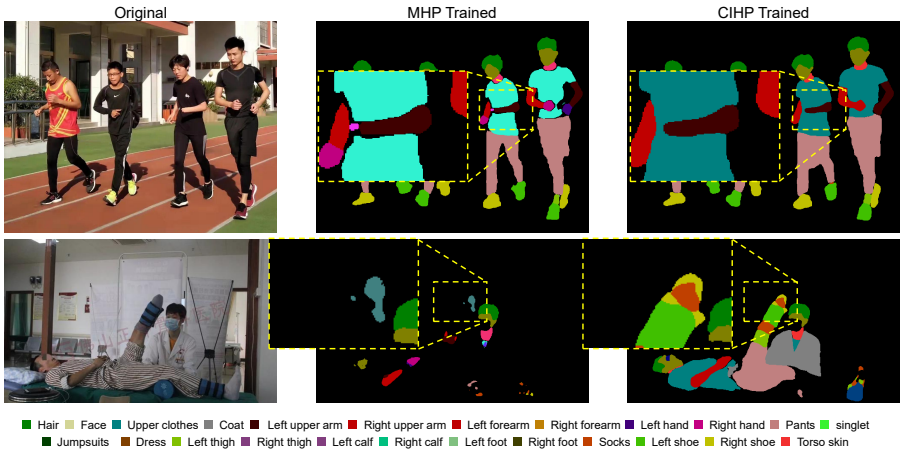

Fig. 2: Failure cases of UniParser trained on MHP v2 and CIHP when applied to individuals with limb deficiencies. Left: input images. Middle/right: predictions of UniParser trained on MHP v2 / CIHP. Because MHP v2 and CIHP do not include labels for residual limbs or prosthetic parts, UniParser often assigns these regions to background or nearby intact-part classes (yellow dashed boxes).

labels. In particular, IH-10K introduces new classes for residual limbs, prostheses, and mobility assistive devices in addition to conventional body parsing categories. These labels are not merely extra category names. They make models learn varied residual-limb shapes, prosthetic configurations, mobility aids, stump-prosthesis boundaries, and body-device contact. Limbs may terminate early, continue through prosthetic components, or connect to mobility aids. As a result, models must parse body layouts beyond a fixed intact-limb assumption.

Specifically, IH-10K contains 10,555 images, 14,939 person instances, and 160,133 segmentation masks across everyday, sports, and clinical scenes. Its label set follows common body parsing practice, while adding explicit classes for residual limbs, prosthetic components, and mobility assistive devices. Each person is annotated with a pixel-wise parsing mask under an annotation protocol, followed by independent full-pass review, annotator revision, and lead-reviewer quality control. IH-10K covers a wide age range and diverse real-world contexts, enabling subgroup-resolved evaluation over common and disability-related classes. Data is released through the project page under the CC BY-NC-SA 4.0 license, with approval from the institutional Human Research Ethics Committee.

We benchmark inclusive body parsing on IH-10K by reporting subgroup-resolved results over disability-related and common classes. This makes performance gaps visible and comparable, instead of being masked by overall scores dominated by common classes. We evaluate strong supervised and zero-shot parsing methods, including AIParsing [81], UniParser [12], SegFormer [70], Mask2Former [10], and SAM2 [54]. Across methods, we observe large and consistent errors on

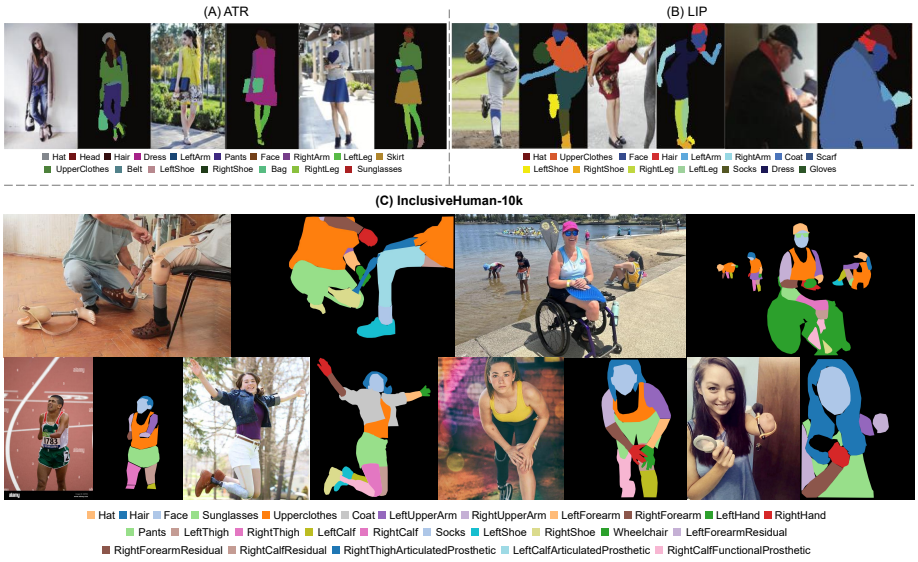

Fig. 3: Demonstration of differences in annotated part categories across body parsing datasets. The figure shows a subset of categories for clarity. (A) ATR and (B) LIP focus on clothing and intact limbs. (C) InclusiveHuman-10K adds residual limbs, prosthetic components, and mobility assistive devices, which supports inclusive analysis and fair evaluation.

disability-related regions. This gap underscores the need for IH-10K and inclusive body parsing models that work beyond the intact-limb assumption.

IH-10K turns residual limbs, prosthetic components, and mobility aids into explicit part labels with pixel-level masks and clear evaluation. This converts a previously hidden blind spot in human parsing into a measurable and repeatable benchmark setting. It also makes performance gaps visible through subgroup-resolved reporting on common vs disability-related classes. As a result, future methods can be compared on whether they truly generalize to diverse limb structures, instead of only fitting intact anatomies. By making these gaps measurable, IH-10K can steer future work toward models that serve people with limb deficiencies more reliably and fairly. We expect IH-10K to serve as a standard testbed for inclusive human parsing and to accelerate progress toward fair and reliable human-centric vision. Our contributions are summarized below:

- We introduce InclusiveHuman-10K (IH-10K), a 10K-image human parsing benchmark centered on individuals with limb deficiencies, with pixel-wise labels for residual limbs, prosthetic components, and mobility assistive devices.
- We establish an inclusive benchmark protocol with subgroup-resolved evaluation over common and disability-related classes, making disparities explicit and comparable.

Table 1: Comparison of human parsing benchmarks. #Images/Frames reports the number of images for static datasets and the number of frames for VIP. People/sample reports the average number of annotated person instances per image when available. IH-10K is the first human parsing benchmark that labels residual limbs, prosthesis components, and mobility aids, and it supports subgroup evaluation over common vs disability-related classes.

Dataset	#Images /#Frames	People / sample	Residual limb	Prosthesis parts	Mobility aids	Subgroup eval
PASCAL-Person-Part [8]	3,533	2.2	✗	✗	✗	✗
ATR [47]	17,700	1.0	✗	✗	✗	✗
LIP [28]	50,462	1.0	✗	✗	✗	✗
CIHP [27]	38,280	3.4	✗	✗	✗	✗
MHP v2 [56]	25,403	3.0	✗	✗	✗	✗
VIP [85]	21,247	2.9	✗	✗	✗	✗
InclusiveHuman-10K (Ours)	10,555	1.4	✓	✓	✓	✓

– We benchmark strong supervised and zero-shot methods on IH-10K and provide analysis that reveals large and consistent errors on disability-related regions, setting clear targets for future inclusive body parsing research.

2 Related Work

2.1 Human Body Parsing

Benefiting from advances in deep learning algorithms and large-scale annotated data [3, 6, 14, 15, 19, 20, 26, 39, 42–44, 63], human-centric vision has made steady progress in recent years.

Datasets. Human parsing datasets evolve from coarse part labels to in-the-wild and crowded scenes. PASCAL-Person-Part and ATR focus on constrained images or fashion-centric settings [8,47]. LIP adds diverse poses, scales, and lighting [28]. CIHP and MHP v2 support multi-person parsing under heavy occlusion, while VIP extends to videos with cross-frame instance links [27, 56, 85]. However, these benchmarks do not annotate residual limbs, prosthetic components, or mobility assistive devices. Thus, standard pretraining on ADE20K or COCO-Stuff and fine-tuning on LIP, CIHP, MHP v2, or VIP do not evaluate parsing behavior on individuals with limb deficiencies [5, 27, 28, 56, 84, 85]. Figure 3 compares the label spaces and highlights the missing disability-related parts in prior datasets.

Methods. Human parsing inherits architectures from semantic segmentation, including FCN, DeepLab, PSPNet, UPerNet, HRNet, and transformer-based models such as SETR and SegFormer [7, 48, 65, 69, 70, 82, 83]. Mask classification frameworks such as MaskFormer and Mask2Former and instance pipelines such as Mask R-CNN remain strong baselines [10, 11, 32]. Parsing-specific work further improves part structure and boundaries, such as CE2P and SCHP [45, 56]. AIParsing combines an anchor-free detector with an edge-guided parsing head [81]. UniParser unifies instance and part correlation learning with pixel-level outputs [12]. CID proposes contextual instance decoupling for crowded multi-person

Fig. 4: Demonstration of the InclusiveHuman-10K (IH-10K) dataset. Representative samples show pixel-wise body parsing for residual limbs, prosthetic components, and mobility assistive devices across varied scenes, poses, and viewpoints. Segmentation masks are overlaid to visualize part boundaries.

scenes [64]. Yet most methods assume intact-limb label spaces, so disability-related regions stay unseen in training and unmeasured in reporting.

What IH-10K adds and what it enables. As summarized in Table 1, standard human parsing benchmarks omit residual limbs, prosthesis parts, and mobility aids, and they do not support subgroup evaluation over common vs disability-related classes. IH-10K makes these regions explicit part labels with pixel-wise masks, introducing non-intact limb structures, sharp device boundaries, and long-tailed occluded parts. It also enables direct measurement on disability-related regions via part-level metrics (e.g., PCP and AP^p), subgroup-resolved reporting, and boundary-sensitive evaluation.

2.2 Individuals with Limb Deficiencies

Recent progress in human-centric vision has been driven by stronger models and richer annotated data [13, 24, 34–37, 50, 51, 68, 73–77]. However, a small set of works directly includes people with limb loss or prostheses [16, 18, 80]. LDPose defines residual-limb endpoints and asks models to predict missing joints; it reports large errors from common pose models on residual regions and provides a dataset focused on limb-deficient pose [79]. InclusiveVidPose scales this direction to videos and provides a large video-based pose dataset for individuals with limb deficiencies, with residual-limb keypoints and rich labels such as segmentation masks, tracking IDs, and per-limb prosthesis status [17]. ProGait targets transfemoral prosthesis users in video and supports segmentation, 2D pose, and gait analysis; it shows that off-the-shelf models degrade when prostheses appear [78]. Clinical biomechanics releases motion capture and force datasets for amputees, which supports gait research without visual parsing [33]. Human–computer interaction work studies wheelchair use and access in daily life, but it does not define residual-limb primitives for vision tasks [38, 46]. In summary, these works

motivate inclusion, but the field still lacks a large pixel-level parsing dataset with explicit residual-limb and prosthesis labels and subgroup-resolved evaluation, which is the gap that IH-10K fills. By making these failure modes visible and measurable, IH-10K encourages new parsing models and training objectives that handle non-intact body topology and assistive-device boundaries, rather than optimizing only for intact-part accuracy on existing benchmarks.

3 InclusiveHuman-10K

3.1 Data Sourcing

We source public images from LDPose [79] that show residual limbs, prostheses, or mobility aids in everyday, sports, and clinical scenes (Figure 4). We use carefully chosen keywords and their synonyms to retrieve candidates. We keep only images that allow research use under their licenses, and we record the source URL, access date, and license. We require at least one person with enough pixel detail for part parsing, then we remove near-duplicates and balance domains and scene types during collection. This follows prior in-the-wild human parsing datasets such as LIP [28] and CIHP [27]. Such in-the-wild data is important because modern parsers rely on varied training images. Broader data also helps reveal cases that common benchmarks do not cover [21–23, 25, 58–62, 66, 67, 72].

3.2 Category Schema for Inclusive Parsing

We follow common human parsing label sets (LIP, CIHP, and MHP v2) and add disability-related parts, including residual limbs, prosthetic components, and mobility aids, so these regions are named and measured rather than merged into the background.

Residual-limb granularity. To reduce ambiguity at the interface between stump and prosthesis, we label *four segments on each body side*: for the upper limb, we annotate the upper arm and the forearm; for the lower limb, we annotate the thigh and the lower leg. Left and right are annotated separately, which yields eight segment labels in total. This granularity provides clear stump boundaries and consistent labels across amputation levels.

Prosthesis taxonomy. As shown in Figure 5, and following public clinical guidance and patient-education resources, we group prostheses by structural complexity into three levels [1, 2, 52, 53]:

- **Cosmetic prosthesis (appearance).** The goal is appearance restoration with little or no functional intent. In clinical materials, these are often described as cosmetic or passive devices [1, 2, 52, 57]. They have no articulated joints or hinges, and are treated as cosmetic components in our annotations.
- **Functional non-articulated prosthesis (single-piece function).** These devices provide a single primary function through a rigid, non-jointed structure. Examples include running blades and passive hooks [4, 41, 53].
- **Articulated prosthesis (jointed).** These devices contain at least one visible mechanical joint (e.g., knee, ankle, or elbow), enabling a joint-like motion range [2, 52].

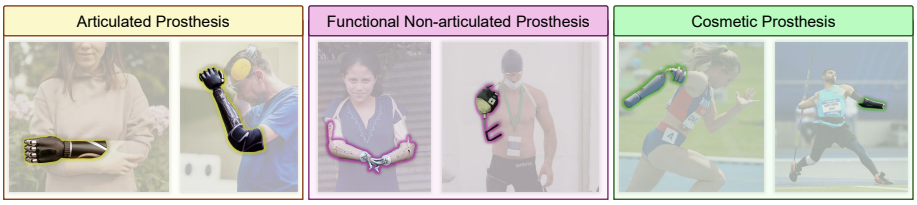

Fig. 5: Demonstration of three prosthesis types. We show cosmetic counterweight prostheses that improve balance without joints or active function, functional non-articulated prostheses that provide a single function (for example running blades or passive hooks), and articulated prostheses with movable joints (for example prosthetic knees or elbows). We annotate these types separately to support part-aware learning and evaluation.

Residual-limb segments and prosthesis types provide supervision for diverse limb structures. Stump length, attachment sites, and joint presence vary across individuals, which encourages models to learn body layouts beyond an intact-limb template.

Disambiguation. When a region boundary is visible, annotators can trace it. When partially occluded, they complete the boundary with the shortest anatomically plausible contour. If the prosthesis subtype cannot be determined or visibility is insufficient, we do not assign a type-specific label and omit that component from per-region scoring.

3.3 Annotation Protocol

Tooling and guidelines. Annotators use X-AnyLabeling[3] with polygon masks for pixel-accurate boundaries. The guideline covers stump boundaries, prosthetic sockets and pylons, mobility aids, and occlusion cases. When a boundary is partly occluded, annotators close it with the shortest anatomically plausible contour.

Annotator training. We recruit 25 trained annotators. We provide written guides and visual examples, a difficult-case manual, and calibration tasks that focus on stump edges, socket–skin transitions, and device components to align style before large-scale labeling.

Review workflow. Each mask is first created by a trained annotator. We split the corpus into batches for quality control. For each batch, a second reviewer performs a full-pass review, records correction requests, and returns the masks for revision. The original annotator revises the masks and resubmits them. A lead reviewer then performs targeted spot checks for cross-batch consistency and resolves unresolved cases. We maintain an audit trail for annotation changes.

3.4 Ethical Considerations and Licensing

Privacy, sensitivity, and takedown. We follow a minimal disclosure principle. We exclude sensitive attributes that are not required for parsing. We provide

[3] X-AnyLabeling

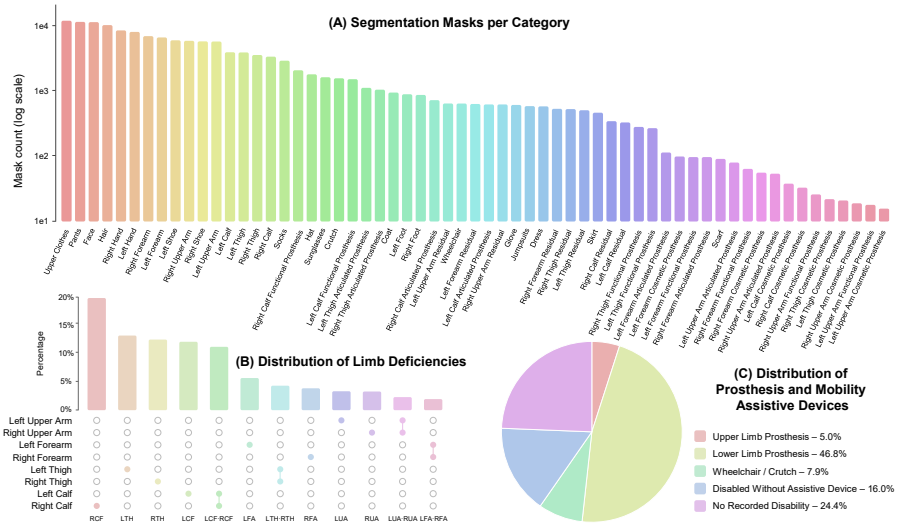

Fig. 6: Demonstration of dataset statistics for InclusiveHuman-10K (IH-10K). (A) Segmentation mask counts per category on a log scale. (B) Distribution of limb deficiencies across upper arm, forearm, thigh, and calf on each side. (C) Distribution of mobility assistive devices. The spread indicates broad diversity across assistive and non-assistive cases.

a documented email channel for removal requests from depicted individuals or rightsholders. Upon verification, we remove or replace the item in the next release and update the audit log and splits.

License and intended use. We release the IH-10K annotations and code under the © CC BY-NC-SA 4.0[4]. Attribution is required. Only noncommercial uses are permitted. Adaptations must use the same license. Users must respect the original third-party licenses attached to source images. IH-10K must not be used for biometric identification, including face recognition or re-identification. IH-10K must not be used for medical diagnosis or clinical decision making.

3.5 Dataset Statistics and Analysis

InclusiveHuman-10K (IH-10K) contains 10,555 images, 14,939 distinct people, and 160,133 segmentation masks.

Composition and distributions. Figure 6 summarizes IH-10K from three views. Figure 6 (A) reports mask counts for five groups, covering prostheses, mobility aids, limb deficiency without assistive devices, and a non-disability group. Counts are shown on a log scale (about 10 to 10^4), revealing a long-tail pattern that is relevant for class imbalance [30]. Figure 6 (B) reports residual-limb

[4] Creative Commons Attribution-NonCommercial-ShareAlike 4.0

Table 2: Fine-tuned body parsing results on IH-10K. We report mean Intersection-over-Union (mIoU), part-based Average Precision (AP^p_{50} and AP^p_{vol}), and Percentage of Correctly Parsed semantic Parts (PCP). Each cell reports the *All* score, with the *Common*/*Disability-Related* breakdown shown in subscript.

Method	Backbone	Metrics (All | Common / Disability-Related)			
		mIoU	APp50	APpvol	PCP
AIParsing [81]	ResNet50	$30.97_{46.34/17.51}$	$31.70_{46.51/6.93}$	$40.06_{46.40/10.72}$	$35.92_{48.06/18.81}$
	ResNet101	$31.49_{50.78/15.60}$	$46.77_{59.73/13.37}$	$47.21_{52.96/15.39}$	$48.81_{58.85/29.62}$
UniParser [12]	ResNet50	$39.06_{56.55/25.18}$	$43.08_{65.07/25.62}$	$24.44_{38.86/13.00}$	$44.66_{65.16/28.38}$
	ResNet101	$41.43_{58.10/28.19}$	$45.67_{67.77/28.11}$	$26.60_{41.33/14.90}$	$47.11_{66.76/31.50}$

frequencies by limb segment and laterality, including bilateral indicators. Figure 6 (C) reports assistive-device settings, which are uneven across categories. We therefore report overall performance and subgroup scores for common and disability-related classes, and include boundary-sensitive evaluation at residual-limb and prosthesis interfaces.

Findings and implications for inclusive benchmarking. These statistics highlight three points. First, mask counts follow a long-tail distribution across groups and parts, which motivates imbalance-aware analysis [30]. Second, residual-limb annotations span upper and lower limbs, both sides, and bilateral cases, creating varied boundary and topology conditions for parsing. Third, assistive-device settings are not uniform, so a single overall score can hide subgroup gaps. These properties make subgroup evaluation necessary for inclusive benchmarking. Subgroup-resolved reporting keeps performance gaps visible across disability-related and common classes, in line with fairness-focused evaluation practice [31]. Because disability-related labels often have sharp interfaces, we also motivate boundary-sensitive evaluation to capture errors that region overlap may understate [9].

4 Experiments

4.1 Task and Dataset

We evaluate three tasks on IH-10K: (1) fine-tuned parsing, including (i) fine-tuned body parsing with dedicated human parsing models and (ii) supervised part segmentation with standard semantic segmentors, (2) zero-shot segmentation, and (3) open-vocabulary detection (OVD)[5]. Zero-shot segmentation includes three sub-settings: open-vocabulary instance segmentation (OVIS), open-vocabulary semantic segmentation (OVSeg), and promptable segmentation (PromptSeg) with point or box prompts. We use an image-level random split with a 0.7:0.1:0.2 ratio for training, validation, and testing, and release splits to support reproducibility.

[5] Due to the page limit, all open-vocabulary detection experiments and analysis appear in the supplementary material.

Table 3: Fine-tuned segmentation baselines on IH-10K. We report mIoU, boundary IoU, mAcc, and mDice on *All*, *Common*, and *Disability-Related* classes.

Method	Backbone	Metrics			
		mIoU	BIoU	mAcc	mDice
FCN [48]	ResNet50	**21.36** 35.25/**9.92**	12.00 18.22/6.87	29.63 46.33/15.87	**31.24** **48.93**/**16.67**
	ResNet101	21.14 **35.32**/9.47	**12.82** **19.53**/**7.29**	**30.20** **47.36**/**16.07**	30.75 48.80/15.89
SegFormer [70]	MIT-B0	18.56 31.40/7.99	10.27 16.24/5.34	25.37 41.64/11.97	27.42 44.31/13.52
	MIT-B1	21.17 34.85/9.91	12.02 18.39/6.77	29.10 45.97/15.21	30.84 48.40/16.39
	MIT-B2	22.49 36.84/10.66	13.34 19.95/7.89	30.79 48.33/16.35	32.29 50.56/17.25
	MIT-B3	23.79 39.06/11.22	14.13 21.13/8.37	32.99 51.02/18.15	33.95 53.20/18.09
	MIT-B4	**25.56** **40.66**/**12.76**	14.87 22.06/**8.95**	**34.57** **52.96**/**19.43**	**36.12** **55.01**/**20.56**
	MIT-B5	24.54 39.73/12.04	**14.94** **22.23**/8.94	33.51 52.09/18.21	35.06 54.00/19.46
Mask2Former [10]	ResNet50	23.79 39.92/10.50	15.59 23.88/8.76	34.63 55.06/17.80	33.84 54.23/17.04
	ResNet101	25.37 41.14/12.39	16.27 24.42/9.56	37.03 55.08/22.17	37.03 55.50/19.93
	Swin-T	26.20 43.12/12.26	16.93 25.80/9.63	37.77 58.57/20.64	36.73 57.55/19.57
	Swin-S	24.11 40.09/10.95	15.80 23.68/9.31	36.75 56.49/20.50	34.06 54.12/17.54
	Swin-B	26.63 **45.75**/10.89	17.24 26.56/9.57	39.68 **62.69**/20.73	36.64 **60.16**/17.28
	Swin-L	**28.07** 45.05/**14.09**	**19.41** **28.67**/**11.78**	**40.66** 60.85/**24.04**	**38.81** 59.21/**22.01**
SegNeXt [29]	Mscan-t	23.02 35.92/12.39	13.48 19.77/8.31	32.33 48.05/19.38	33.37 49.59/20.02
	Mscan-s	24.30 39.12/12.10	15.48 22.31/9.86	34.24 51.33/20.17	34.76 53.15/19.61
	Mscan-b	27.54 43.74/14.20	17.21 25.31/10.55	38.44 58.11/**22.25**	38.67 58.23/22.56
	Mscan-l	**28.19** **45.01**/**14.34**	**17.78** **26.12**/**10.92**	**38.62** **58.81**/21.99	**39.33** **59.59**/**22.64**

4.2 Evaluation Protocol

We report results in three groups: *All*, *Common*, and *Disability-Related*. *Common* contains 27 classes aligned with standard human parsing taxonomies. *Disability-Related* contains 34 classes covering residual limbs, prosthetic components, and mobility assistive devices.

For fine-tuned body parsing, we report mean Intersection-over-Union (mIoU), Percentage of Correctly parsed semantic Parts (PCP), and part-based Average Precision (AP^p), including AP^p_{50} and AP^p_{vol} following standard multi-human parsing evaluation. For fine-tuned part segmentation, we report mean Intersection-over-Union (mIoU), Boundary Intersection-over-Union (BIoU), mean accuracy (mAcc), and mean Dice score (mDice). For open-vocabulary instance segmentation, we report COCO Average Precision (AP) and COCO Average Recall with at most one detection per image (AR@1), and we also report mIoU, BIoU, and mAcc for mask quality comparability. For open-vocabulary semantic and promptable segmentation, we report mIoU, BIoU, and mAcc.

4.3 Baselines and Experimental Setup

Fine-tuned body parsing. We evaluate several publicly available human parsing models, including AIParsing [81] and UniParser [12], under the standard multi-human parsing evaluation.

Fine-tuned segmentation. We evaluate FCN [48], SegFormer [70], Mask2Former [10], and SegNeXt [29] as widely used semantic segmenters. We use these models as reproducible baselines for our label space and subgroup metrics.

Zero-shot segmentation. We evaluate OVIS methods including Grounded-SAM [55], Grounded-SAM2 [55], X-Decoder [86], and SEEM [87], and OVSeg

Table 4: Zero-shot segmentation results on IH-10K. We report AP, AR@1, and mask-quality metrics (mIoU, BIoU, mAcc) on *All*, *Common*, and *Disability-Related* classes.

	Method	Backbone	Prompt	Metrics (All | Common / Disability-Related)				
				AP	AR@1	mIoU	BIoU	mAcc
OVIS	Grounded SAM [55]	Base	Text	4.74 10.22/0.38	4.94 10.50/0.53	5.99 11.37/1.71	5.85 11.30/1.52	14.27 27.88/3.47
		Large	Text	5.02 10.68/0.52	5.26 10.96/0.73	6.04 11.41/**1.78**	5.94 11.39/1.62	**14.34** 27.64/**3.77**
		Huge	Text	**5.12** **10.87**/**0.55**	**5.37** **11.15**/0.77	**6.10** **11.56**/1.77	**6.03** **11.58**/**1.63**	**14.34** **27.97**/3.51
	Grounded SAM2 [55]	Tiny	Text	2.46 5.28/0.22	3.04 6.47/0.31	4.62 8.94/1.19	4.30 8.58/0.90	8.55 17.18/1.69
		Small	Text	2.49 5.34/0.22	3.06 6.51/0.33	4.85 9.31/1.31	4.50 8.91/1.00	8.68 17.48/1.69
		Base	Text	2.54 5.40/0.27	3.13 6.60/0.39	4.65 8.91/1.27	4.34 8.54/1.00	8.56 17.11/1.77
		Large	Text	2.60 5.46/0.32	3.20 6.64/0.46	4.69 8.97/1.30	4.38 8.59/1.04	8.66 17.31/1.80
	X-Decoder	Focal-T	Text	0.25 0.20/0.29	1.08 0.65/1.42	0.54 0.42/0.65	0.44 0.36/0.51	3.55 4.50/2.80
		Focal-L	Text	0.26 0.15/0.35	1.50 1.81/1.25	1.04 1.37/0.79	0.94 1.32/0.64	4.68 8.14/1.93
	SEEM [87]	SAM-B	Text	0.12 0.01/0.20	0.86 0.01/1.53	0.38 0.24/0.50	0.30 0.16/0.41	3.48 4.35/2.79
		SAM-L	Text	0.14 0.03/0.22	0.89 0.03/1.57	0.37 0.18/0.52	0.29 0.14/0.41	3.41 4.25/2.73
		Focal-T	Text	0.26 0.14/0.34	1.29 1.40/1.21	1.03 0.95/1.10	0.89 0.97/0.83	3.93 5.29/2.86
		Focal-L	Text	0.16 0.20/0.12	0.99 1.33/**7.13**	0.60 0.50/0.69	0.56 0.50/0.61	3.01 5.42/1.09
OVSeg	GroupViT [71]		Text	-	-	0.53 0.82/**0.29**	0.21 0.34/0.12	5.05 7.18/3.36
	CLIPSeg [49]		Text	-	-	**2.68** **5.81**/0.19	**1.83** **3.95**/**0.15**	**15.40** **28.39**/**5.08**
PromptSeg	SAM [40]	Huge	Box	-	-	75.55 73.38/77.26	71.40 69.62/72.82	82.80 81.77/83.62
		Huge	Point	-	-	18.03 9.53/24.76	18.10 10.29/24.31	32.78 26.07/38.10
	SAM2 [54]	Tiny	Box	-	-	**77.06** **74.67**/78.96	**71.90** **70.04**/73.37	85.42 **84.79**/85.92
		Tiny	Point	-	-	27.09 19.84/32.84	26.19 20.24/30.92	42.53 41.77/43.13
		Small	Box	-	-	73.35 68.27/77.39	68.04 63.82/71.40	81.59 76.89/85.31
		Small	Point	-	-	27.05 21.47/31.48	25.45 21.19/28.84	42.13 41.46/42.66
		Base	Box	-	-	75.84 71.25/**79.48**	70.60 66.60/**73.78**	**85.55** 82.16/**88.24**
		Base	Point	-	-	24.47 20.80/27.39	23.93 20.93/26.32	39.59 40.39/38.95
		Large	Box	-	-	75.20 70.17/79.36	69.98 65.73/73.36	84.58 82.15/86.51
		Large	Point	-	-	23.63 19.66/26.78	23.29 19.74/26.12	39.00 38.98/39.02

models including CLIPSeg [49] and GroupViT [71]. For PromptSeg, we evaluate SAM [40] and SAM2 [54] with box prompts from the annotated box and point prompts from the mask centroid

4.4 Main Results

Fine-tuned body parsing. Table 2 reports results of AIParsing and UniParser on IH-10K. Across both models, performance on *Disability-Related* classes is much lower than on *Common* classes. UniParser achieves the best mIoU, reaching 41.43 on *All* and 28.19 on *Disability-Related* with ResNet101. However, it remains 29.9 points below its *Common* mIoU of 58.10. The gap is also clear in part-centric metrics. For UniParser with ResNet101, AP^p_{50} drops from 67.77 on *Common* to 28.11 on *Disability-Related*, and PCP drops from 66.76 to 31.50. AIParsing attains higher AP^p_{vol}, yet its *Disability-Related* mIoU and AP^p_{50} remain low (15.60 mIoU and 13.37 AP^p_{50} with ResNet101). Scaling from ResNet50 to ResNet101 improves overall scores, but the *Common* vs *Disability-Related* gap stays large for both models. These results show that even dedicated human parsing models still struggle on residual limbs, prosthesis components, and their boundaries after fine-tuning on IH-10K.

Fine-tuned segmentation. As shown in Table 3, FCN, SegFormer, Mask2Former, and SegNeXt perform much worse on the *Disability-Related* subset than on *Com-*

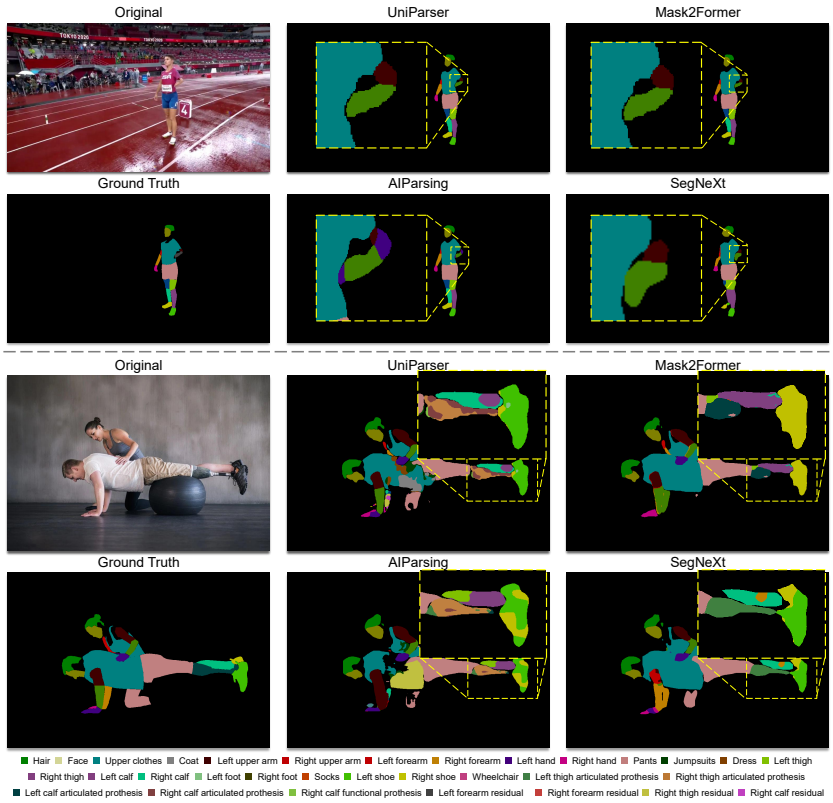

Fig. 7: Failure cases on IH-10K. Left: input images. Middle and right: predicted parsing masks from Mask2Former and UniParser after fine-tuning on IH-10K. Colors indicate the predicted part labels in our label set. Both models may miss or mislabel prosthetic components and residual-limb regions, and they often produce boundary shifts near residual-limb edges and stump–prosthesis seams under small scale, occlusion, and close contact with other people or objects.

mon. Across many model and backbone pairs, Disability-Related mIoU is roughly one third to two fifths of Common mIoU. Boundary Intersection-over-Union (BIoU) drops even more, since it measures overlap in a thin band around boundaries and penalizes boundary shifts more than standard mask IoU. On IH-10K, this points to boundary errors around residual-limb edges and stump-prosthesis interfaces. Scaling up the backbone improves *Common* classes much more than *Disability-Related* classes, so the gap remains across architectures and scales. As a result, a single overall score can hide persistent deficits on disability-related parts, especially for long-tail categories. We therefore report subgroup-resolved results and include boundary-sensitive metrics to keep these failure modes visible and comparable across methods.

Zero-shot segmentation. Table 4 shows a consistent zero-shot gap. Text-driven open-vocabulary methods (Grounded SAM, Grounded SAM2, X-Decoder, SEEM, CLIPSeg, and GroupViT) reach reasonable scores on *Common* classes but collapse on *Disability-Related* classes, and larger backbones do not change this trend. This means that text prompts can often match common body or clothing concepts, but they do not reliably ground residual limbs, prosthetic parts, and mobility aids. The SAM and SAM2 PromptSeg results should be read as ground-truth-prompted mask tracing, not automatic parsing, because their boxes come from annotations and their points come from mask centroids. Additional prompt-model adaptation gives the same message. Tuned Grounded-SAM2 improves All mIoU only from 4.66 to 6.68 and leaves *Disability-Related* mIoU at 1.32. SAM3 fine-tuning gives a larger gain, improving All mIoU from 19.55 to 33.38 and *Disability-Related* mIoU from 5.05 to 14.20. However, the gain is uneven across disability-related subgroups. SAM3 reaches 23.98 mIoU on residual limbs and 54.12 on mobility aids, but only 7.61 on prosthetic parts. It also performs better on upper- and lower-limb cases than on prosthesis-specific subsets. This suggests that IH-10K supervision can teach modern prompt models some residual-limb and device concepts, but prosthetic components remain hard because they are small, rare, and visually close to natural limbs or shoes. SAM2 with ground-truth boxes reaches 75.32 mIoU, and 10% noisy boxes remain close at 73.17. However, using boxes from zero-shot Grounded-SAM2 drops SAM2 to 4.94 mIoU. Thus, high oracle-prompted scores show that the mask decoder can trace these regions once a good prompt is given. The harder problem is to ground, localize, and label disability-related parts without ground-truth prompts.

Failure cases. Figure 7 highlights typical errors on disability-related regions after fine-tuning. When people occupy few pixels or are partially occluded, both methods lose fine limb details and small device parts. In cluttered interactions, Mask2Former often maps prosthetic components to the closest intact-part labels, such as calf or shoe, which blurs the stump-prosthesis transition. UniParser can assign prosthesis-related labels in some cases, but it may still mix prosthetic components with adjacent parts when boundaries are weak. We also tag failures for tuned Grounded-SAM2. The dominant failure flags are missed text grounding (85.26% of failed cases), background leakage (42.75%), and missed small parts (26.52%). Disability-related errors also include prostheses parsed as natural limbs and residual limbs being missed. These errors show that IH-10K stresses concept grounding, small parts, non-intact limb layouts, and body-device boundaries.

5 Discussion

Inclusive human parsing should not assume intact anatomies by default. When benchmarks omit residual limbs and prosthetic components, models cannot be trained or evaluated on these regions. IH-10K makes these regions explicit with pixel-level labels and subgroup-resolved reporting, which helps keep performance gaps visible and comparable across methods [31].

Our benchmark also highlights a key limitation of current open-vocabulary systems. In our experiments, text-driven open-vocabulary segmentation and de-

tection perform poorly on disability-related parts, even when they work on common classes. Promptable models produce strong masks only when ground-truth localization is given, so these results should not be read as automatic parsing. Additional prompt-model adaptation shows that fine-tuning can improve modern prompt models, but disability-related grounding remains far below ground-truth-prompted SAM2. This bottleneck matters because zero-shot evaluation is the most accessible setting for the community to reuse models.

IH-10K can benefit downstream systems that rely on part-level body understanding and editing, such as virtual try-on, human image editing, avatar-based interaction, and human-centric content creation. If residual limbs or prosthetic parts are merged into background or clothing, these systems can fail in ways that affect real users. WHO guidance estimates that about 0.5% of a population accesses prosthetic and orthotic services at any one time, which underscores the need for benchmarks that cover these common real-world cases.[6]

We hope IH-10K motivates future work on open-vocabulary recognition of fine-grained body parts beyond an intact-limb template. Promising directions include taxonomy-aware prompting and synonym expansion, long-tail aware training, and boundary-aware objectives and evaluation for stump–prosthesis interfaces [9]. We also emphasize responsible use. IH-10K must not be used for biometric identification or medical diagnosis, and it is released under a non-commercial license (CC BY-NC-SA 4.0).

6 Conclusion

We introduce InclusiveHuman-10K (IH-10K), a benchmark for inclusive body parsing that adds pixel-level labels for residual limbs, prosthetic components, and mobility assistive devices, with expert-reviewed masks across diverse scenes. IH-10K supports subgroup-resolved evaluation over *Common* and *Disability-Related* classes, and reports both parsing-specific and boundary-sensitive metrics. Across fine-tuned body parsing and zero-shot open-vocabulary segmentation, we observe a persistent performance gap on disability-related regions. Promptable models narrow this gap once localization is provided, which suggests a recognition and label-alignment bottleneck for disability-related parts. We hope IH-10K serves as a standard testbed that drives methods and evaluation toward reliable and fair human parsing beyond an intact-limb assumption.

Acknowledgments This research is funded in part by ARC-Discovery grant (DP220100800 to XY) and ARC-DECRA grant (DE230100477 to XY). We thank Advance Queensland Industry Research Projects (AQIRP) for their support and guidance. We thank the area chair and anonymous reviewers for constructive comments.

[6] Service Provision Guidance for Prosthetic & Orthotic Services

References

1. Clinical commissioning policy: Multi-grip upper limb prosthetics. Tech. Rep. 03736, NHS England (2018), https://www.england.nhs.uk/wp-content/uploads/2018/07/Multi-grip-upper-limb-prosthetics.pdf
2. Amputee Coalition: Prosthetic feet (2016), https://amputee-coalition.org/resources/prosthetic-feet/, fact Sheet, Updated 08/2016
3. An, Z., You, J., Li, J., Tang, Y., Li, W., Du, H., Du, S.: Fredn: Spectral disentanglement for time series forecasting via learnable frequency decomposition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 19623–19631 (2026)
4. Beck, O.N., Taboga, P., Grabowski, A.M.: Characterizing the mechanical properties of running-specific prostheses. PLOS ONE **11**(12), e0168298 (2016). https://doi.org/10.1371/journal.pone.0168298, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0168298
5. Caesar, H., Uijlings, J.R.R., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 1209–1218. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00132, http://openaccess.thecvf.com/content_cvpr_2018/html/Caesar_COCO-Stuff_Thing_and_CVPR_2018_paper.html
6. Cao, X., Xu, M., Yu, X., Yao, J., Ye, W., Huang, S., Zhang, M., Tsang, I.W., Ong, Y., Kwok, J.T., Shen, H.T.: Analytical survey of learning with low-resource data: From analysis to investigation. ACM Comput. Surv. **58**(6), 144:1–144:47 (2026). https://doi.org/10.1145/3773075, https://doi.org/10.1145/3773075
7. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018). https://doi.org/10.1109/TPAMI.2017.2699184, https://doi.org/10.1109/TPAMI.2017.2699184
8. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.L.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. pp. 1979–1986. IEEE Computer Society (2014). https://doi.org/10.1109/CVPR.2014.254, https://doi.org/10.1109/CVPR.2014.254
9. Cheng, B., Girshick, R.B., Dollár, P., Berg, A.C., Kirillov, A.: Boundary iou: Improving object-centric image segmentation evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 15334–15342. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.01508, https://openaccess.thecvf.com/content/CVPR2021/html/Cheng_Boundary_IoU_Improving_Object-Centric_Image_Segmentation_Evaluation_CVPR_2021_paper.html
10. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 1280–1289. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.00135, https://doi.org/10.1109/CVPR52688.2022.00135
11. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang,

P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 17864–17875 (2021), https://proceedings.neurips.cc/paper/2021/hash/950a4152c2b4aa3ad78bdd6b366cc179-Abstract.html
12. Chu, J., Jin, L., Teng, Y., Li, J., Wei, Y., Wang, Z., Xing, J., Yan, S., Zhao, J.: Uniparser: multi-human parsing with unified correlation representation learning. IEEE Transactions on Image Processing **33**, 5159–5171 (2024)
13. Deng, W., Campbell, D., Sun, C., Zhang, J., Kanitkar, S., Shaffer, M.E., Gould, S.: Pos3r: 6d pose estimation for unseen objects made easy. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 16818–16828. Computer Vision Foundation / IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.01567, https://openaccess.thecvf.com/content/CVPR2025/html/Deng_Pos3R_6D_Pose_Estimation_for_Unseen_Objects_Made_Easy_CVPR_2025_paper.html
14. Du, H., Du, S., Li, W.: Probabilistic time series forecasting with deep non-linear state space models. CAAI Trans. Intell. Technol. **8**(1), 3–13 (2023). https://doi.org/10.1049/CIT2.12085, https://doi.org/10.1049/cit2.12085
15. Du, H., Li, L., Huang, Z., Yu, X.: Object-goal visual navigation via effective exploration of relations among historical navigation states. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 2563–2573. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00252, https://doi.org/10.1109/CVPR52729.2023.00252
16. Du, H., Ying, J., Chen, X., Wang, S., Li, X., Yu, X.: Same person, different depiction: Counterfactual evaluation of vision-language models on individuals with limb deficiencies. In: European Conference on Computer Vision (ECCV) (2026)
17. Du, H., Ying, J., Wang, S., Li, X., Zhang, K., Yu, X.: Inclusivevidpose: Bridging the pose estimation gap for individuals with limb deficiencies in video-based motion. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=SyQqXAdWUq
18. Du, H., Ying, J., Zhang, K., Zhu, T., Yu, X.: Position: Human-centric vision requires topological generalization beyond fixed skeletal topologies. In: Forty-third International Conference on Machine Learning Position Paper Track (2026), https://openreview.net/forum?id=2N3dVx1soP
19. Du, H., Yu, X., Zheng, L.: Learning object relation graph and tentative policy for visual navigation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII. Lecture Notes in Computer Science, vol. 12352, pp. 19–34. Springer (2020). https://doi.org/10.1007/978-3-030-58571-6_2, https://doi.org/10.1007/978-3-030-58571-6_2
20. Du, H., Yu, X., Zheng, L.: Vtnet: Visual transformer network for object goal navigation. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=DILxQP08O3B
21. Du, X., Sun, H., Lu, M., Zhu, T., Yu, X.: Dreamcar: Leveraging car-specific prior for in-the-wild 3d car reconstruction. IEEE Robotics and Automation Letters **10**(2), 1840–1847 (2024)
22. Du, X., Wang, Y., Sun, H., Wu, Z., Sheng, H., Wang, S., Ying, J., Lu, M., Zhu, T., Zhan, K., et al.: 3drealcar: An in-the-wild rgb-d car dataset with 360-degree views. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26488–26498 (2025)

23. Du, X., Wang, Y., Yu, X.: Mvgs: Multi-view regulated gaussian splatting for novel view synthesis (2026), https://arxiv.org/abs/2410.02103
24. Du, X., Wang, Y., Zhan, K., Yu, X.: Mobile-gs: Real-time gaussian splatting for mobile devices. arXiv preprint arXiv:2603.11531 (2026)
25. Du, X., Yu, X., Liu, J., Dai, B., Xu, F.: Ethics-aware face recognition aided by synthetic face images. Neurocomputing **600**, 128129 (2024)
26. Feng, X., Du, H., Fan, H., Duan, Y., Liu, Y.: Seformer: Structure embedding transformer for 3d object detection. In: Williams, B., Chen, Y., Neville, J. (eds.) Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023. pp. 632–640. AAAI Press (2023). https://doi.org/10.1609/AAAI.V37I1.25139, https://doi.org/10.1609/aaai.v37i1.25139
27. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV. Lecture Notes in Computer Science, vol. 11208, pp. 805–822. Springer (2018). https://doi.org/10.1007/978-3-030-01225-0_47, https://doi.org/10.1007/978-3-030-01225-0_47
28. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 6757–6765. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.715, https://doi.org/10.1109/CVPR.2017.715
29. Guo, M., Lu, C., Hou, Q., Liu, Z., Cheng, M., Hu, S.: Segnext: Rethinking convolutional attention design for semantic segmentation. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/08050f40fff41616ccfc3080e60a301a-Abstract-Conference.html
30. Gupta, A., Dollár, P., Girshick, R.B.: LVIS: A dataset for large vocabulary instance segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 5356–5364. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00550, http://openaccess.thecvf.com/content_CVPR_2019/html/Gupta_LVIS_A_Dataset_for_Large_Vocabulary_Instance_Segmentation_CVPR_2019_paper.html
31. Gustafson, L., Rolland, C., Ravi, N., Duval, Q., Adcock, A., Fu, C., Hall, M., Ross, C.: FACET: fairness in computer vision evaluation benchmark. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 20313–20325. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.01863, https://doi.org/10.1109/ICCV51070.2023.01863
32. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2980–2988. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.322, https://doi.org/10.1109/ICCV.2017.322

33. Hood, S., Ishmael, M.K., Gunnell, A., Foreman, K., Lenzi, T.: A kinematic and kinetic dataset of 18 above-knee amputees walking at various speeds. Scientific data **7**(1), 150 (2020)
34. Hou, Y., Gould, S., Zheng, L.: Learning to select views for efficient multi-view understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 20135–20144. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.01903, https://doi.org/10.1109/CVPR52733.2024.01903
35. Hou, Y., Gould, S., Zheng, L.: Scalable deep color quantization: A cluster imitation approach. IEEE Trans. Image Process. **33**, 5273–5283 (2024). https://doi.org/10.1109/TIP.2024.3414132, https://doi.org/10.1109/TIP.2024.3414132
36. Hou, Y., Zheng, L., Torr, P.: Learning camera movement control from real-world drone videos. CoRR **abs/2412.09620** (2024). https://doi.org/10.48550/ARXIV.2412.09620, https://doi.org/10.48550/arXiv.2412.09620
37. Hu, X., Lin, Y., Wang, S., Wu, Z., Lv, K.: Agent-centric relation graph for object visual navigation. IEEE Trans. Circuits Syst. Video Technol. **34**(2), 1295–1309 (2024). https://doi.org/10.1109/TCSVT.2023.3291131, https://doi.org/10.1109/TCSVT.2023.3291131
38. Huang, W., Ghahremani, S., Pei, S., Zhang, Y.: Wheelpose: Data synthesis techniques to improve pose estimation performance on wheelchair users. In: Mueller, F.F., Kyburz, P., Williamson, J.R., Sas, C., Wilson, M.L., Dugas, P.O.T., Shklovski, I. (eds.) Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024. pp. 944:1–944:25. ACM (2024). https://doi.org/10.1145/3613904.3642555, https://doi.org/10.1145/3613904.3642555
39. Khan, M.W., Sheng, H., Zhang, H., Du, H., Wang, S., Coroneo, M.T., Hajati, F., Shariflou, S., Kalloniatis, M., Phu, J., Agar, A., Huang, Z., Golzan, S.M., Yu, X.: RVD: A handheld device-based fundus video dataset for retinal vessel segmentation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/3a71ee306d6991f2f87dd414e0bdf851-Abstract-Datasets_and_Benchmarks.html
40. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 3992–4003. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.00371, https://doi.org/10.1109/ICCV51070.2023.00371
41. Leeds Teaching Hospitals NHS Trust: Multi-grip prosthetic hands & digit (2025), https://www.leedsth.nhs.uk/patients/resources/multi-grip-prosthetic-hands-digit/, leaflet LN005838. Page last reviewed: 04/06/2025
42. Li, J., Du, H., Tang, Y., You, J., Du, S., Yu, W., Li, W.: Lgfnet: Local correlation and global context fusion for multivariate time series forecasting. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2231–2235. IEEE (2026)
43. Li, J., Du, H., Yu, W., Tang, Y., You, J., Du, S., Li, W.: Cadmamba: Clustering adaptive mamba for multivariate time series forecasting. In: ICASSP 2026-

2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3066–3070. IEEE (2026)

44. Li, J., Yu, W., Du, H., Tang, Y., You, J., Du, S., Li, W.: Pamnet: Patch-adaptive mixing network for multivariate time series forecasting. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5836–5840. IEEE (2026)
45. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 3260–3271 (2022). https://doi.org/10.1109/TPAMI.2020.3048039, https://doi.org/10.1109/TPAMI.2020.3048039
46. Li, Y., Mollyn, V., Yuan, K., Carrington, P.: Wheelposer: Sparse-imu based body pose estimation for wheelchair users. In: Flatla, D.R., Hwang, F., Guerreiro, T.J.V., Brewer, R. (eds.) The 26th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2024, St. John's, NL, Canada, October 27-30, 2024. pp. 8:1–8:17. ACM (2024). https://doi.org/10.1145/3663548.3675638, https://doi.org/10.1145/3663548.3675638
47. Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., Yan, S.: Deep human parsing with active template regression. IEEE Trans. Pattern Anal. Mach. Intell. **37**(12), 2402–2414 (2015). https://doi.org/10.1109/TPAMI.2015.2408360, https://doi.org/10.1109/TPAMI.2015.2408360
48. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 3431–3440. IEEE Computer Society (2015). https://doi.org/10.1109/CVPR.2015.7298965, https://doi.org/10.1109/CVPR.2015.7298965
49. Lüddecke, T., Ecker, A.S.: Image segmentation using text and image prompts. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 7076–7086. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.00695, https://doi.org/10.1109/CVPR52688.2022.00695
50. Luo, B., Guo, J., Yao, Y., Ding, X.: Adversarial style optimization: Enhancing vlm jailbreaks by grpo-based stylistic triggers optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11–19 (June 2026)
51. Lv, K., Li, Y., Chen, Z., Wang, S., Han, S., Lin, Y.: Infer the whole from a glimpse of a part: Keypoint-based knowledge graph for vehicle re-identification. In: Walsh, T., Shah, J., Kolter, Z. (eds.) Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025. pp. 5901–5909. AAAI Press (2025). https://doi.org/10.1609/AAAI.V39I6.32630, https://doi.org/10.1609/aaai.v39i6.32630
52. NIH MedlinePlus Magazine: How a prosthetic ankle improves balance control (Mar 2023), https://magazine.medlineplus.gov/article/how-a-prosthetic-ankle-improves-balance-control/
53. Rahnama, L., Soulis, K., Geil, M.D.: A review of evidence on mechanical properties of running specific prostheses and their relationship with running performance. Frontiers in Rehabilitation Sciences **5**, 1402114 (2024). https://doi.org/10.3389/fresc.2024.1402114, https://www.frontiersin.org/journals/rehabilitation-sciences/articles/10.3389/fresc.2024.1402114/full

54. Ravi, N., Gabeur, V., Hu, Y., 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., Girshick, R.B., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025), https://openreview.net/forum?id=Ha6RTeWMd0
55. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024)
56. Ruan, T., Liu, T., Huang, Z., Wei, Y., Wei, S., Zhao, Y.: Devil in the details: Towards accurate single and multiple human parsing. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. pp. 4814–4821. AAAI Press (2019). https://doi.org/10.1609/AAAI.V33I01.33014814, https://doi.org/10.1609/aaai.v33i01.33014814
57. Segura, D., Romero, E., Abarca, V.E., Elias, D.A.: Upper limb prostheses by the level of amputation: A systematic review. Prosthesis **6**(2), 277–300 (2024). https://doi.org/10.3390/prosthesis6020022, https://www.mdpi.com/2673-1592/6/2/22
58. Sun, X., Gazagnadou, N., Sharma, V., Lyu, L., Li, H., Zheng, L.: Privacy assessment on reconstructed images: Are existing evaluation metrics faithful to human perception? In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/2082273791021571c410f41d565d0b45-Abstract-Conference.html
59. Sun, X., Hou, Y., Deng, W., Li, H., Zheng, L.: Ranking models in unlabeled new environments. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 11741–11751. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.01155, https://doi.org/10.1109/ICCV48922.2021.01155
60. Sun, X., Leng, X., Wang, Z., Yang, Y., Huang, Z., Zheng, L.: Cifar-10-warehouse: Broad and more realistic testbeds in model generalization analysis. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=pw2ssoOTpo
61. Sun, X., Li, M., Yuan, K., Sun, M.W., Endo, M., Wu, S., Li, C., Zhang, Y., Wang, Z., Yeung-Levy, S.: Do vlms perceive or recall? probing visual perception vs. memory with classic visual illusions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25861–25870 (June 2026)
62. Sun, X., Yao, Y., Wang, S., Li, H., Zheng, L.: Alice benchmarks: Connecting real world object re-identification with the synthetic. CoRR **abs/2310.04416** (2023). https://doi.org/10.48550/ARXIV.2310.04416, https://doi.org/10.48550/arXiv.2310.04416
63. Tang, Y., Du, H., Du, S., Li, W.: Integrating deep learning into semiparametric network vector autoregressive models. Mathematics **14**(1), 38 (2025)

64. Wang, D., Zhang, S.: Contextual instance decoupling for robust multi-person pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11060–11068 (2022)
65. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3349–3364 (2021). https://doi.org/10.1109/TPAMI.2020.2983686, https://doi.org/10.1109/TPAMI.2020.2983686
66. Wang, J., Lin, Y., Hu, X., Wang, S., Lv, K.: Local motion matters: A deconstruct-recompose paradigm for reinforcement learning pre-training from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9859–9868 (June 2026)
67. Wang, S., Wu, Z., Hu, X., Lin, Y., Lv, K.: Skill-based hierarchical reinforcement learning for target visual navigation. IEEE Trans. Multim. **25**, 8920–8932 (2023). https://doi.org/10.1109/TMM.2023.3243618, https://doi.org/10.1109/TMM.2023.3243618
68. Wu, H., Song, S., Yao, C., Han, S., Wan, H., Lin, Y., Lv, K.: Think how your teammates think: Active inference can benefit decentralized execution. In: Koenig, S., Jenkins, C., Taylor, M.E. (eds.) Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026. pp. 29749–29757. AAAI Press (2026). https://doi.org/10.1609/AAAI.V40I35.40220, https://doi.org/10.1609/aaai.v40i35.40220
69. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V. Lecture Notes in Computer Science, vol. 11209, pp. 432–448. Springer (2018). https://doi.org/10.1007/978-3-030-01228-1_26, https://doi.org/10.1007/978-3-030-01228-1_26
70. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Álvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 12077–12090 (2021), https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html
71. Xu, J., Mello, S.D., Liu, S., Byeon, W., Breuel, T.M., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 18113–18123. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01760, https://doi.org/10.1109/CVPR52688.2022.01760
72. Yao, C., Lin, Y., Song, S., Wu, H., Yang, S., Ma, Y., Lv, K.: Role-level inductive bias for cross-task generalization in multi-agent reinforcement learning. In: Forty-third International Conference on Machine Learning (2026), https://openreview.net/forum?id=oz8kFbXdpj
73. Yao, Y., Gedeon, T., Zheng, L.: Large-scale training data search for object re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 15568–15578.

IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01494, https://doi.org/10.1109/CVPR52729.2023.01494
74. Yao, Y., Wen, Z., Tong, Y., Tian, X., Li, X., Ma, X., Xu, D., Gedeon, T.: Thought graph traversal for test-time scaling in chest x-ray vllms. Pattern Recognit. **179**, 113639 (2026). https://doi.org/10.1016/J.PATCOG.2026.113639, https://doi.org/10.1016/j.patcog.2026.113639
75. Yao, Y., Yang, R., Gedeon, T.: Bipartite mode matching for vision training set search from a hierarchical data server. In: Koenig, S., Jenkins, C., Taylor, M.E. (eds.) Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026. pp. 16128–16136. AAAI Press (2026). https://doi.org/10.1609/AAAI.V40I19.38648, https://doi.org/10.1609/aaai.v40i19.38648
76. Yao, Y., Zheng, L., Yang, X., Naphade, M., Gedeon, T.: Simulating content consistent vehicle datasets with attribute descent. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 12351, pp. 775–791. Springer (2020). https://doi.org/10.1007/978-3-030-58539-6_46, https://doi.org/10.1007/978-3-030-58539-6_46
77. Yao, Y., Zheng, L., Yang, X., Napthade, M., Gedeon, T.: Attribute descent: Simulating object-centric datasets on the content level and beyond. IEEE Trans. Pattern Anal. Mach. Intell. **46**(4), 2489–2505 (2024). https://doi.org/10.1109/TPAMI.2023.3338291, https://doi.org/10.1109/TPAMI.2023.3338291
78. Yin, X., Yang, B., Liu, W., Xue, Q., Alamri, A., Fiedler, G., Gao, W.: Progait: A multi-purpose video dataset and benchmark for transfemoral prosthesis users. CoRR **abs/2507.10223** (2025). https://doi.org/10.48550/ARXIV.2507.10223, https://doi.org/10.48550/arXiv.2507.10223
79. Ying, J., Du, H., Zhang, K., Li, L., Yu, X.: Ldpose: Towards inclusive human pose estimation for limb-deficient individuals in the wild
80. Ying, J., Du, H., Zhang, K., Tweedy, S.M., Yu, X.: Resihmr: Residual-limb aware single-image 3d human mesh recovery for individuals with limb loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13940–13950 (June 2026)
81. Zhang, S., Cao, X., Qi, G.J., Song, Z., Zhou, J.: Aiparsing: Anchor-free instance-level human parsing. IEEE Transactions on Image Processing **31**, 5599–5612 (2022)
82. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 6230–6239. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.660, https://doi.org/10.1109/CVPR.2017.660
83. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 6881–6890. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00681, https://openaccess.thecvf.com/content/CVPR2021/html/Zheng_Rethinking_Semantic_Segmentation_From_a_Sequence-to-Sequence_Perspective_With_Transformers_CVPR_2021_paper.html
84. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: 2017 IEEE Conference on Computer Vision and

Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 5122–5130. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.544, https://doi.org/10.1109/CVPR.2017.544

85. Zhou, Q., Liang, X., Gong, K., Lin, L.: Adaptive temporal encoding network for video instance-level human parsing. In: Boll, S., Lee, K.M., Luo, J., Zhu, W., Byun, H., Chen, C.W., Lienhart, R., Mei, T. (eds.) 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018. pp. 1527–1535. ACM (2018). https://doi.org/10.1145/3240508.3240660, https://doi.org/10.1145/3240508.3240660
86. Zou, X., Dou, Z., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., Peng, N., Wang, L., Lee, Y.J., Gao, J.: Generalized decoding for pixel, image, and language. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 15116–15127. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01451, https://doi.org/10.1109/CVPR52729.2023.01451
87. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/3ef61f7e4afacf9a2c5b71c726172b86-Abstract-Conference.html