

Figure 7: Human refinement interface.

A Human Refinement Interface

Figure 7 illustrates our annotation interface for human refinement of LabelAny3D’s automatically generated pseudo labels on COCO3D. The top panel displays the point cloud and corresponding 3D bounding boxes, which annotators can rotate and translate for an optimal viewing angle. The lower-left panel shows the attributes of the selected bounding box. Annotators can modify specific parameters (e.g., width) using keyboard shortcuts. They also have the option to delete a bounding box if it corresponds to an invalid 3D object (e.g., a person on a poster). The lower-right panel visualizes the 2D projections of the current 3D boxes to provide additional context. Annotators may also choose to discard the entire image if the relative geometry is incorrect or no valid 3D objects are present.

B Annotation Efficiency

Table 4 presents a category-wise overview of AP_{3D} , AR_{3D} , and average IoU_{3D} for the pseudo annotations generated by LabelAny3D, evaluated against our refined COCO3D benchmark. The results demonstrate that our pipeline produces high-quality annotations with human-like accuracy. In addition to achieving high precision, the average IoU exceeds 0.40 for the majority of categories, indicating strong alignment of spatial layout.

Table 2 compares the overall AP_{3D} of our method against the baseline OVM3D-DET [27], with our approach achieving a higher AP of 64.17. Among the 5,373 annotations in the COCO3D benchmark, 3,146 were accepted by human annotators without modification, while 2,227 required only minor refinement. Just 466 were rejected due to issues such as reflections, 2D object representations, or insufficient point cloud quality.

These results demonstrate that LabelAny3D produces high-quality pseudo annotations that can serve as effective initialization for human annotators. With minimal effort required for refinement, our pipeline significantly reduces manual workload and accelerates the overall annotation process.

Table 4: **Per-category 3D annotation quality for the top 50 categories.** Comparison between pseudo annotations from LabelAny3D and human-refined annotations. Ranking is based on IoU_{3D} .

Category	$\text{AP}_{3D}\uparrow$	$\text{AR}_{3D}\uparrow$	$\text{IoU}_{3D}\uparrow$	Category	$\text{AP}_{3D}\uparrow$	$\text{AR}_{3D}\uparrow$	$\text{IoU}_{3D}\uparrow$
sports ball	81.58	88.26	80.94	bed	88.84	93.33	76.32
fire hydrant	91.14	94.80	75.16	airplane	90.61	93.53	68.83
couch	70.05	88.57	67.89	snowboard	63.50	72.14	68.27
parking meter	92.56	94.29	63.94	mouse	68.98	78.70	63.57
vase	89.89	92.73	61.38	skateboard	69.87	80.73	60.70
fork	64.76	75.65	58.93	cat	72.94	80.49	57.66
bicycle	83.45	94.65	56.60	microwave	82.65	90.53	56.00
dog	79.54	90.36	54.73	tie	48.71	67.83	54.60
bus	84.28	93.61	52.20	dining table	86.38	92.22	51.59
bench	84.80	89.81	49.43	surfboard	50.10	65.00	48.24
sink	60.22	71.95	46.97	sandwich	65.13	77.94	47.38
refrigerator	72.87	86.52	45.04	kite	70.54	80.71	44.99
laptop	69.56	84.69	43.92	cake	80.60	90.40	43.38
oven	66.77	79.05	43.29	umbrella	92.28	95.21	41.30
tv	43.71	62.54	41.10	bowl	79.32	88.40	40.70
horse	76.75	89.84	40.38	bear	82.62	89.71	40.28
backpack	69.39	82.11	38.90	keyboard	47.08	61.82	38.36
pizza	70.90	81.82	37.22	skis	40.44	64.25	36.93
traffic light	62.11	77.12	36.80	cup	89.94	93.44	36.35
remote	59.63	71.60	35.49	tennis racket	38.10	59.31	31.62
clock	39.06	57.36	33.25	elephant	87.71	98.82	33.19
handbag	68.75	82.28	32.23	wine glass	97.11	98.51	31.76
chair	61.75	88.63	32.60	horse	76.75	89.84	40.38
boat	80.06	90.29	12.15	bird	70.48	85.53	16.68
motorcycle	92.65	96.77	13.99	book	65.94	76.75	13.57

C Implementation details

During annotation, we exclude objects whose masks contain fewer than 400 pixels. Following [30], we also discard objects whose masks overlap the image boundary by more than 10 pixels, treating them as truncated.

For 2D–3D matching, we first estimate the camera elevation angle using the elevation module from One-2-3-45 [43], based on the amodal-completed object crop. We then render 8 views of the object at the estimated elevation, with azimuths spaced at 45-degree intervals. After computing correspondences between the amodal crop and the 8 rendered views, we obtain an initial camera pose. Using this pose, we render the mesh again and perform an additional 2D–3D matching step to refine the camera pose. Generating pseudo annotations for a single object takes approximately one minute.

Our implementation is based on PyTorch3D [56] and Detectron2 [69]. Following [74], we use DINOv2-Base [50] as the image feature encoder and freeze its parameters during training. The model is initialized from the publicly released OVMono3D weights and fine-tuned for 58k steps with a batch size of 64. We train the model using SGD with an initial learning rate of 0.0012, which decays by a factor of 10 at 60% and 80% of training. A linear warm-up is applied for the first 1.8k steps. Training takes approximately 48 hours on 4 NVIDIA A40 GPUs. We apply standard image augmentations during training, including random horizontal flipping and resizing. In addition, a random positional perturbation is applied to the input 2D bounding box, with a maximum offset ratio of 0.2. For evaluation on COCO3D, we use ground-truth 2D boxes as input. For Omni3D [7], we adopt the same 2D detections from Grounding DINO [45] as used in OVMono3D [74].

556 **D COCO3D Benchmark Samples**

557 Figure 8 presents additional samples from our human-refined COCO3D benchmark. The results
558 demonstrate that the relative geometry and spatial layout of the scenes and 3D bounding boxes are
559 highly accurate and align well with human perception.

560 **E More Qualitative Comparisons**

561 We report more qualitative comparisons of OVMono3D [74] with our finetuned variant in Figure 9.
562 Trained with additional pseudo-labeled in-the-wild images, our model produces more accurate
563 predictions for challenging object categories such as animals, athletes, and food.

564 **F Failure Case**

565 Figure 10 illustrates several failure cases of LabelAny3D. In highly occluded scenes, such as the cow
566 in Figure 10(a), the amodal completion model fails to reconstruct the full object, resulting in a 3D
567 bounding box that captures only a partial region. As our method relies on ground-truth 2D instance
568 segmentations, it may incorrectly generate 3D boxes for objects that are not physically present in the
569 3D scene—for example, the person on a television screen in Figure 10(b). Additionally, for crowded
570 scenes, such as in Figure 10(c), the COCO [40, 18] dataset often lacks per-instance segmentation
571 labels. In such cases, instance segmentation models like Grounded SAM [57] also struggle to separate
572 individual objects accurately, causing our method to miss multiple instances.

573 **G Limitations**

574 Our LabelAny3D pipeline relies on depth estimation models to recover scene geometry. While
575 relative depth estimation is mature, it can still fail in challenging scenes. These failure cases can
576 introduce noise into the final 3D bounding box annotations. The pipeline also depends on amodal
577 completion for occluded objects; however, in cases of severe occlusion, the completion model may
578 fail, resulting in misaligned or incomplete pseudo labels. Additionally, our method currently lacks a
579 robust training strategy for learning effectively from noisy pseudo annotations.

580 Since our pipeline integrates depth estimation and amodal completion modules as plug-in APIs,
581 future improvements in these components can be directly incorporated to enhance annotation quality.
582 These limitations highlight the importance of developing more robust auto-labeling frameworks and
583 training strategies for open-vocabulary monocular 3D detection.

584 **H Broader Impact**

585 Our work facilitates efficient 3D annotation of objects from any category in diverse, in-the-wild
586 scenes. By integrating the generated pseudo labels, existing open-vocabulary monocular 3D detectors
587 become more robust to out-of-domain categories (e.g., animals), which can enhance the reliability of
588 autonomous systems such as robots and self-driving vehicles, particularly in safety-critical scenarios.

589 Our dataset is curated from publicly available sources, and therefore does not raise privacy concerns.
590 While our algorithm is category-agnostic and does not introduce explicit bias, the underlying datasets
591 may reflect societal or geographic biases present in the source data. We encourage future work to
592 investigate and mitigate such biases when deploying systems trained on our annotations.

593 **I Licenses**

Table 5: Licenses of assets used.

Asset	License
Cube R-CNN [7]	CC-BY-NC 4.0
Grounding DINO [45]	Apache License 2.0
Segment Anything [34]	Apache License 2.0
Unidepth [53]	CC-BY-NC 4.0
MoGe [65]	MIT License
Depth Pro [6]	Apple License (Link)
InvSR [75]	S-Lab License 1.0 (Link)
One-2-3-45 [43]	Apache License 2.0
TRELLIS [71]	MIT License
Gen3DSR [3]	CC-BY 4.0
MASt3R [37]	CC-BY-NC-SA 4.0
OVMono3D [74]	Apache License 2.0
OVM3D-Det [27]	Apache License 2.0
KITTI [24]	CC-BY-NC-SA 3.0 DEED
nuScenes [8]	CC-BY-NC 4.0
SUN RGB-D [63]	MIT License
ARKitScenes [4]	Apple License (Link)
COCO [41]	CC-BY 4.0
COCONut [18]	Apache License 2.0

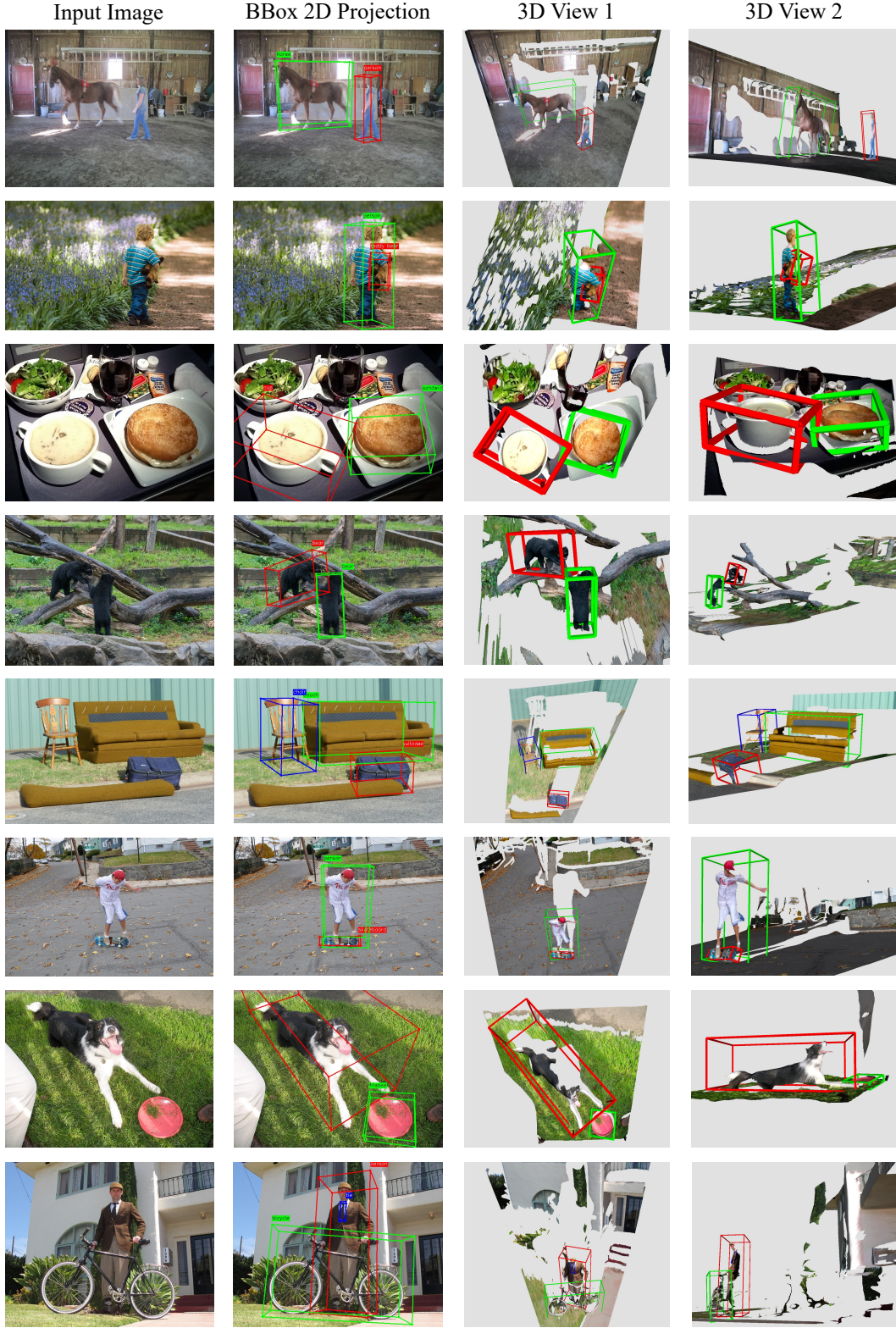


Figure 8: More COCO3D benchmark samples (after human refinement). For each example, we show: (1) the input image, (2) the projected 3D bounding boxes overlaid on the image, and (3–4) two 3D views of the scene point map with the 3D bounding boxes. **Please see our attached files for rendered videos from 3D Scene and BBoxes.**

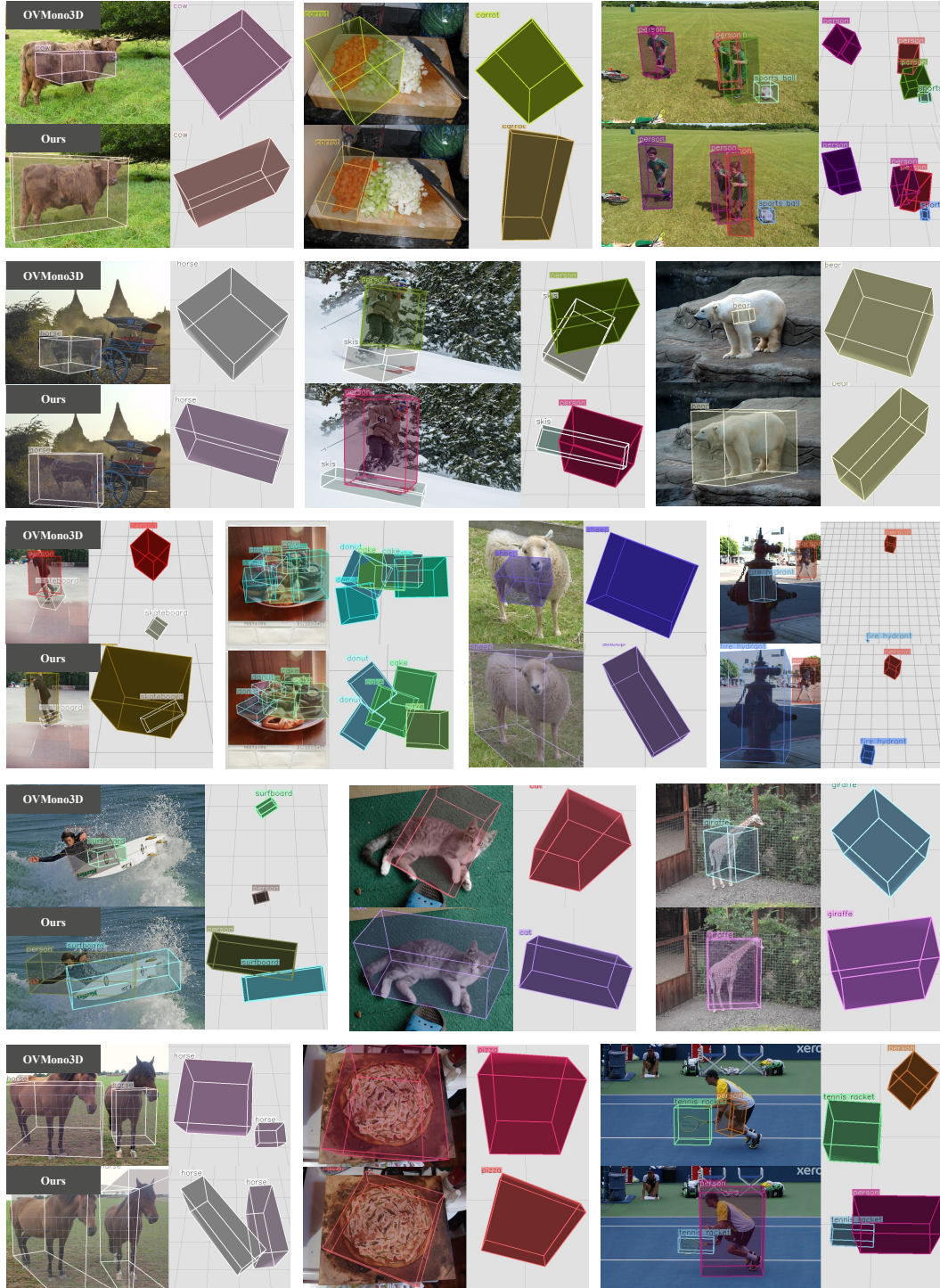
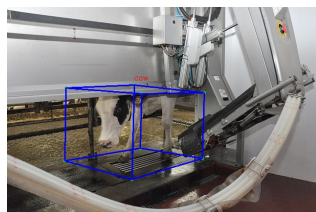
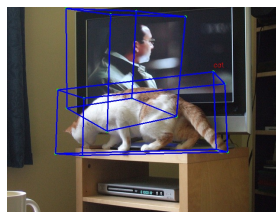


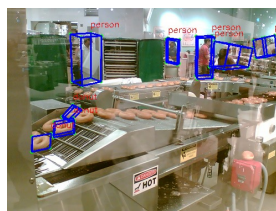
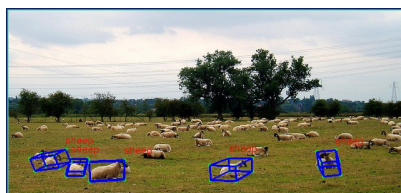
Figure 9: More qualitative open-vocabulary 3D detection results on in-the-wild-images: OVMono3D [74] vs. our finetuned OVMono3D. We display both the 3D predictions overlaid on the image and a top-down view with a base grid of $1\text{ m} \times 1\text{ m}$ tiles.



(a) Highly occluded objects



(b) Objects on 2D plane



(c) Too crowded objects

Figure 10: Failure cases.