

Orient Anything V2: Unifying Orientation and Rotation Understanding

**Zehan Wang^{1,2*}, Ziang Zhang^{1*}, Jiayang Xu¹, Jialei Wang¹,
Tianyu Pang^{3†}, Chao Du³, Hengshuang Zhao⁴, Zhou Zhao^{1,2‡}**

¹Zhejiang University; ²Shanghai AI Lab; ³Sea AI Lab; ⁴The University of Hong Kong

<https://orient-anythingv2.github.io/>

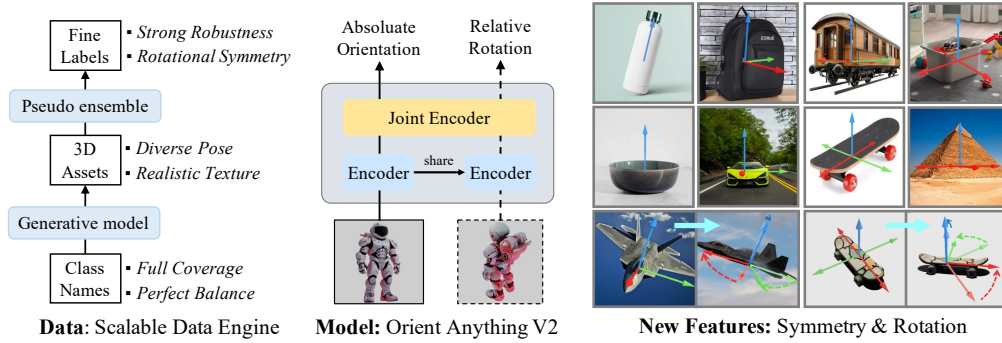


Figure 1: Overview of Orient Anything V2. We upgrade the foundation orientation estimation model from both **Data** and **Model** perspectives. It unifies the understanding of object orientation and rotation, achieving better estimation accuracy and gaining the **New Features** to handle rotational symmetry and relative rotation. Zoom in for the best view.

Abstract

This work presents **Orient Anything V2**, an enhanced foundation model for unified understanding of **object 3D orientation and rotation** from **single or paired images**. Building upon Orient Anything V1, which defines orientation via a single unique front face, V2 extends this capability to handle objects with diverse rotational symmetries and directly estimate relative rotations. These improvements are enabled by four key innovations: **1)** Scalable 3D assets synthesized by generative models, ensuring broad category coverage and balanced data distribution; **2)** An efficient, model-in-the-loop annotation system that robustly identifies 0 to N valid front faces for each object; **3)** A symmetry-aware, periodic distribution fitting objective that captures all plausible front-facing orientations, effectively modeling object rotational symmetry; **4)** A multi-frame architecture that directly predicts relative object rotations. Extensive experiments show that Orient Anything V2 achieves state-of-the-art zero-shot performance on *orientation estimation*, *6DoF pose estimation*, and *object symmetry recognition* across 11 widely used benchmarks. The model demonstrates strong generalization, significantly broadening the applicability of orientation estimation in diverse downstream tasks.

*Equal Contribution.

†Project Leader.

‡Corresponding Author.

1 Introduction

Estimating object orientation from images is a fundamental task of computer vision. 3D object orientation information plays crucial roles in robot manipulation [49, 37, 22], autonomous driving [25], AR/VR [11, 30, 4, 57], and spatial-aware image understanding [21, 28, 60] and generation [56, 50, 35].

Orient Anything V1 [48] is a foundation model for estimating the object orientation aligned with an object’s unique front face. While it exhibits strong robustness and accuracy in absolute orientation estimation, it lacks an understanding of rotation (despite its intrinsic link to orientation). This deficiency results in difficulties handling numerous rotationally symmetric objects (simply classifying them as having no front face) and understanding object rotation relative to a specified reference frame. These limitations around rotation understanding restrict its utility in many downstream tasks.

In this work, we aim to develop an enhanced orientation estimation model, Orient Anything V2, with stronger generalization and deeper understanding of both object orientation and rotation. Our contributions include a scalable data engine and a more elegant model framework.

From the *data* perspective, Orient Anything V1 uses advanced VLM [12, 31] to annotate real 3D assets from Objaverse [8, 7]. Building on this data-driven motivation, we leverage advanced 3D generation models [51, 61] to further speed up data scaling-up and improve the data coverage and balance. Additionally, we assemble pseudo labels predicted by the V1 model across multi-view renderings and refine them through model-in-the-loop calibration. The proposed data engine enables highly cost-effective and flexible data scaling up, delivers robust annotation performance, and shows a strong understanding of rotationally symmetric objects. Our final dataset includes 600K assets, $12\times$ larger than the existing orientation dataset, with significantly higher annotation quality, accurately identifying 0 to N valid front faces.

From the *model* perspective, we first propose symmetry-aware orientation distribution, explicitly teaching the model to capture and predict rotational symmetry. Moreover, our model supports multi-frame input to directly predict relative rotations between frames. This design effectively bridges the knowledge transfer between absolute orientation and relative rotation, showing strong potential in reference-known scenarios.

Our experiments demonstrate the enhanced and novel capabilities of our model. It achieves superior performance on zero-shot orientation estimation and sets new records on zero-shot rotation estimation (i.e., 6DoF pose estimation [49, 26]), while also accurately handling and predicting different rotational symmetries.

To summarize, we propose Orient Anything V2, which improves Orient Anything V1 as follows:

- We propose a data engine that cost-efficiently scales up 3D asset collection and robustly annotates the 0 to N valid front faces to capture different object rotational symmetries.
- We introduce symmetry-aware distribution fitting as a learning objective, allowing the model to directly predict all plausible object orientations.
- We extend the model architecture to support multi-frame input, enabling it to directly estimate relative object rotations over the reference frame.
- Our model demonstrates strong zero-shot generalization across absolute orientation estimation, relative rotation estimation, and object symmetry recognition.

2 Related Work

2.1 Object Rotational Symmetry

Rotational symmetry [36, 34] indicates that an object may retain its original shape after being rotated by certain angles. This property is commonly found across various objects. Understanding an object’s rotational symmetry is critical for 3D object recognition and generation [24, 59], pose estimation [17, 6], and robotic manipulation [39]. While some existing works [40] attempt to detect 3D rotational symmetry from single-view 2D images, they are constrained by limited training data and lack zero-shot generalization to open-world scenarios.



Figure 2: Real assets from Objaverse suffer from (a) low-quality texture and (b) limited realism.

Our focus is on object orientation relative to a semantic "front" face. The number of possible valid front-facing orientations an object possesses is determined by its rotational symmetry around its vertical axis. For example, 180-degree symmetry means there are two distinct valid front faces. Objects with continuous rotational symmetry (symmetric at any angle), like balls, are considered to have no meaningful direction. In this work, we broaden the applicability of orientation estimation models by enabling the prediction of an object’s azimuthal symmetry from a single 2D image. Our model demonstrates impressive zero-shot rotational symmetry recognition performance.

2.2 Relative Rotation Estimation

Predicting an object’s rotation in the query frame relative to the reference frame is a fundamental capability in 6DoF pose estimation [26, 14, 55] and is crucial for robotics applications. Early methods [23, 18, 9] focused on specific instances or object categories. More recent approaches like OnePose [44] and OnePose++ [15] estimate object rotation by solving 2D-3D correspondences across views. POPE [10] follows a similar idea and achieves zero-shot rotation estimation with a single reference frame with the help of SAM [19] and DINOv2 [33]. However, the reliance on pixel matching makes these methods prone to failure under large viewpoint changes.

In contrast, we propose a purely implicit learning approach. Leveraging the inherent coupling between rotation and orientation, we extend the orientation estimation model to support multi-frame inputs, enabling direct zero-shot relative rotation prediction between arbitrary views.

2.3 Single-view Orientation Estimation

Estimating an object’s 3D front-facing orientation (interpreted as its rotation relative to the canonical front view) from a single view, requires the model to have an inherent understanding of different objects’ standard poses and front-facing appearances. Earlier works [53, 46, 42] are mainly limited to a small number of categories or specific domains. More recently, ImageNet3D [29] introduce a large-scale dataset with manually annotated 3D orientations. Orient Anything [48] achieves robust orientation estimation for any objects in any scenes by leveraging an advanced automated annotation pipeline, improved learning objectives, and real-world knowledge from the pre-trained vision model.

In this work, we further address several limitations of Orient Anything and upgrade the orientation estimation model from both data-driven (novel and scalable data engine) and model-driven (direct symmetry and rotation prediction) perspectives, resulting in Orient Anything V2.

3 Revisiting Orient Anything V1

Orient Anything V1 pioneers zero-shot object orientation estimation from single images. It introduces a VLM-based pipeline to annotate front faces of Objaverse 3D assets [8, 7], learns orientation estimation from their renderings via distribution fitting. It also provides a confidence score to indicate whether an object has a unique front face. To further advance the orientation prediction foundation model, we first dig into the potential limitations in its training data and framework.

Disadvantages of Real 3D Assets 1) *Imbalanced Category Distribution*: Stemming from the human biases in asset creation, real 3D datasets, such as Objaverse [8, 7], suffer from significant class imbalance. Common categories like buildings and characters make up a large proportion, while

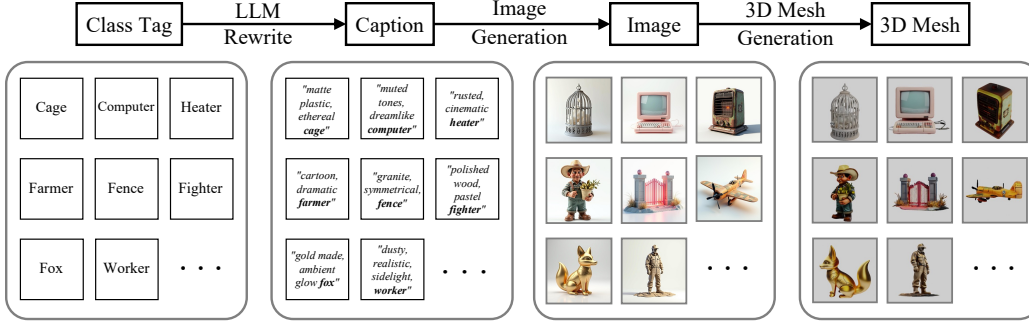


Figure 3: Overview of 3D Asset Synthesis Pipeline. We begin with class tags and use a series of advanced generative models to progressively generate high-quality 3D assets.

others, like uncommon animals, are severely underrepresented. 2) *Inconsistent Data Quality*: Current large-scale 3D datasets often lack high-quality assets with complete geometry and rich surface details. (Fig. 2 a) Moreover, many human-created meshes exhibit fixed poses, leading to a substantial domain gap from real-world object variations (Fig. 2 b).

Limitations of Object Rotation Understanding 1) *Ignored Rotational Symmetries*: Orient Anything V1 defines orientation based on the single, unique front face, overlooking the different rotational symmetries (i.e., multiple valid "front" faces). For the many symmetric objects in real world, the model cannot effectively distinguish or identify their potential orientations. 2) *Unsupported Relative Rotations*: The relative rotation between two views and the front-facing orientation (essentially the rotation relative to the front view) are inherently coupled. However, estimating relative rotation through independent absolute orientation predictions suffers from significant error accumulation, causing Orient Anything V1 to often fail in relative rotation estimation.

4 Scalable Data Engine

4.1 3D Asset Synthesis

Motivated by the recent remarkable progress in generative models and the successful application of synthetic data in downstream tasks [45, 56], we explore whether *synthetic 3D assets* can serve as scalable, high-quality data sources for orientation learning. To fully harness modern generative models, we construct our asset synthesis pipeline as a structured process: *Class Tag* $\rightarrow$ *Caption* $\rightarrow$ *Image* $\rightarrow$ *3D Mesh*, as detailed below:

Step 1: Class Tag $\rightarrow$ Caption. To ensure broad category coverage and diversity, we follow SynCLR’s approach [45], starting from ImageNet-21K [38] category tags, and use Qwen-2.5 [54] to generate rich captions that describe detailed object attributes and diverse poses. *Step 2: Caption $\rightarrow$ Image*. We use the state-of-the-art text-to-image model, FLUX.1-Dev [20], to generate images following the captions. Besides, we enhance captions with positional descriptors to promote explicit 3D structure and upright pose. *Step 3: Image $\rightarrow$ 3D mesh*. We employ the leading open-source image-to-3D model, Hunyuan-3D-2.0 [61], to produce high-quality 3D meshes from the synthesized images.

Finally, we generate 600k 3D assets in total, with approximately 30 items for each class tag in ImageNet-21K. These assets feature complete geometry, detailed textures, and balanced category coverage. In terms of scale, the new synthetic dataset is $12\times$ larger than the filtered real dataset used in Orient Anything V1.

4.2 Robust Annotation

Orient Anything V1 employs VLM to annotate the unique canonical front view of 3D assets. However, this approach is limited by the VLM’s underdeveloped spatial perception ability and struggles to handle diverse rotational symmetries. To address these challenges, we introduce a more effective and robust system for annotating 3D asset orientations.

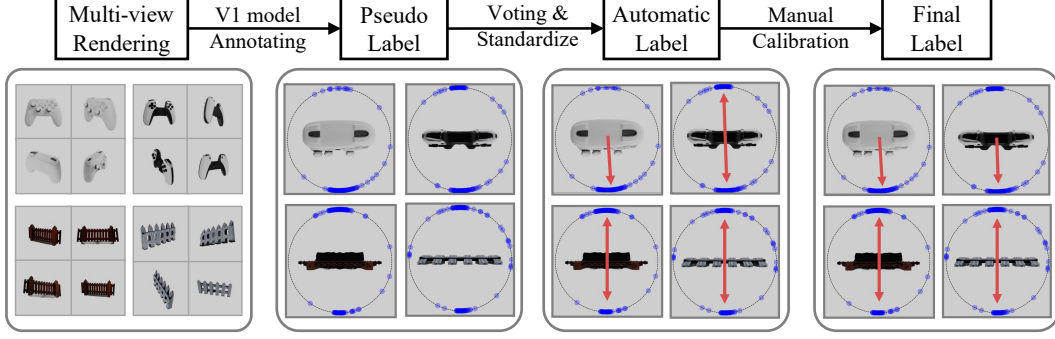


Figure 4: Overview of Robust Annotation Pipeline. "Pseudo Label" visualizes the azimuth direction of pseudo labels and objects in the horizontal plane. By fitting the pseudo labels to standard periodic distribution, we can robustly derive the orientation and symmetry label. Human calibration is only required for categories with symmetry inconsistencies.

Intra-asset Ensemble Annotation We first train an improved orientation estimation model as our automatic annotator, based on the Orient Anything V1 paradigm and incorporating the additional real-world orientation dataset, ImageNet3D. Next, for each 3D asset, we employ this model to produce pseudo-labels for various renderings. Finally, we project these pseudo-labels, obtained from different viewpoints, back into a canonical 3D world coordinate system.

As shown in Fig. 4, the overall distribution of pseudo-labels on the horizon plane clearly indicates the object’s possible orientations. To capture the main direction and rotational symmetries, we first arrange the discrete predicted azimuth angles over $[0^\circ, 360^\circ)$ into a probability distribution $\mathbf{P}_{\text{pseudo}} \in \mathbb{R}^{360}$. This distribution is then fitted to a periodic Gaussian distribution using the least squares method:

$$(\bar{\varphi}, \bar{\alpha}, \bar{\sigma}) = \arg \min_{\varphi, \alpha, \sigma} \sum_{i=0}^{359} \left(\mathbf{P}_{\text{pseudo}}(i) - \frac{\exp\left(\frac{\cos(\alpha(i-\varphi))}{\sigma^2}\right)}{2\pi I_0\left(\frac{1}{\sigma^2}\right)} \right)^2 \quad (1)$$

where $\bar{\sigma}$ is the fitted variance. The phase $\bar{\varphi} \in [0^\circ, 360^\circ)$ represents the main azimuth direction. The periodicity $\bar{\alpha} \in \{1, 2, \dots, N\}$ signifies $360/\bar{\alpha}$ -degree rotational symmetry, possessing $\bar{\alpha}$ valid front faces, while $\bar{\alpha} = 0$ indicates no dominant orientation.

Ensembling multiple pseudo labels in the 3D world effectively suppresses outlier errors from single-view predictions, resulting in significantly more reliable annotations.

Inter-assets Consistency Calibration Building on the rotational symmetry and orientation annotations for individual assets, we further perform human-in-the-loop consistency calibration across assets. Specifically, since our 3D assets are generated based on object category tags, they are naturally grouped by category. We assume that objects of the same category should share the same type of rotational symmetry. Based on this assumption, we analyze the annotated rotational symmetries within each category. If all assets within the same category demonstrate the same symmetries, we directly consider the annotations to be correct. If inconsistencies are found, we manually review all assets in that category to re-annotate or filter out incorrect annotations.

As each asset is annotated independently, the cross-asset consistency check and manual calibration offer an orthogonal perspective that efficiently and effectively enhances annotation reliability. Statistically, across 21k source category tags, we observe only minor inconsistencies in around 15% of categories, each involving a small number of assets. The finding further validates the accuracy and robustness of our ensemble annotation strategy.

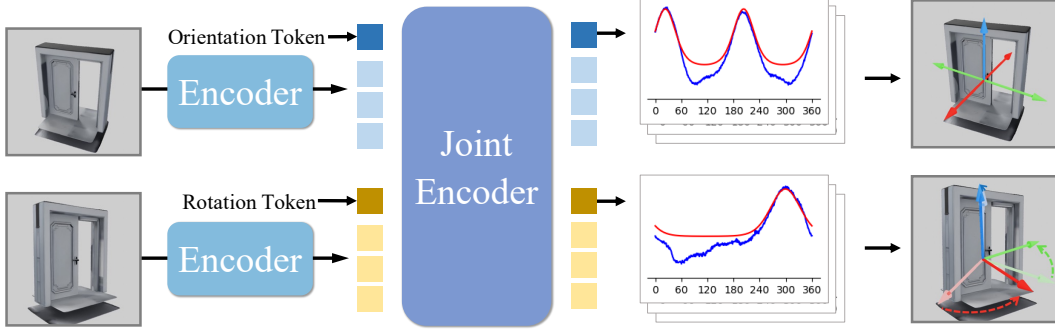


Figure 5: Framework of Orient Anything V2. One or two input frames are tokenized by DINOv2 and then jointly encoded using transformer blocks. We finally employ MLP heads to predict the orientation or rotation distributions from the encoded learnable tokens of each frame.

5 Framework

5.1 Symmetry-aware Distribution

Orient Anything V1 proposes an orientation distribution fitting task that guides the model to learn circular Gaussian distributions over azimuth, polar, and in-plane rotation angles, that preserve the similarity between neighboring angles. Each angle is modeled with a unimodal target distribution centered on a unique front-facing orientation. For symmetric objects with multiple or no semantic front faces, the model additionally predicts a low orientation confidence to filter them out.

To recognize different types of rotational symmetry and enable general orientation prediction for objects with multiple front faces, we further introduce the symmetry-aware periodic distribution as the training target. As discussed in Sec. 4.2, our ensemble annotation and consistency calibration approach enables accurate and robust labeling of 0 to N valid front-facing directions over the horizontal plane. To incorporate these annotations into prediction, we directly model 0 to N valid front faces within the azimuth angle distribution. This design naturally replaces V1’s extra orientation confidence design. Instead, different kinds of rotational symmetries are captured directly from the predicted probability distribution. This more elegant framework enables the model to inherently share knowledge across all object categories.

For training, the target $\mathbf{P}_{\text{azi}} \in \mathbb{R}^{360}$ for the azimuth angle, originally represented as a circular Gaussian distribution, is adapted to be periodic:

$$\mathbf{P}_{\text{azi}}(i|\hat{\varphi}, \bar{\alpha}, \sigma) = \frac{\exp\left(\frac{\cos(\frac{\alpha(i-\hat{\varphi})}{\sigma^2})}{\sigma^2}\right)}{2\pi I_0\left(\frac{1}{\sigma^2}\right)} \quad (2)$$

where $\hat{\varphi}$ and $\bar{\alpha}$ are the phase (azimuth angle) and periodicity (rotation symmetry) fitted from Sec. 4.2, σ is the variance hyper-parameter, and $i = 0^\circ, \dots, 359^\circ$ is the angle index. Target probability distributions for the polar angle $\mathbf{P}_{\text{pol}} \in \mathbb{R}^{180}$ and in-plane rotation angle $\mathbf{P}_{\text{rot}} \in \mathbb{R}^{360}$ are constructed using a similar method, but without the periodicity parameter.

During inference, the predicted angle distributions are fitted to a standard distribution model using the least squares method, similar to Eq. 1. The resulting parameters (azimuth periodicity $\hat{\alpha}$, azimuth angle $\hat{\varphi}$, polar angle $\hat{\sigma}$ and rotation angle $\hat{\delta}$), directly indicate the object’s $\hat{\alpha}$ valid front faces (i.e., symmetric with $360/\hat{\alpha}$ degree rotation) and their corresponding front-facing directions in 3D space.

5.2 Relative Rotation Estimation

To establish a connection between absolute orientation and relative rotation, enabling knowledge sharing and transferring, we modify the network architecture to support dynamic inputs from one or multiple images.

As shown in Fig. 5, we mainly follow VGGT [47], first using a visual encoder, DINOv2 [33], to encode each input image into K tokens, augmented with learnable tokens. The combined set of

Model	SUN-RGBD		ARKitScenes		Pascal3D+		Objectron		ImageNet3D [†]		Ori_COCO
	Med↓	Acc30°↑	Med↓	Acc30°↑	Med↓	Acc30°↑	Med↓	Acc30°↑	Med↓	Acc30°↑	Acc↑
OriAny.V1	33.94	48.5	77.58	35.8	22.90	55.0	30.67	49.6	13.34	71.3	72.4
OriAny.V2	26.00	55.4	36.48	43.2	15.02	72.7	22.62	56.4	15.26	65.2	86.4

Table 1: Zero-shot Absolute Orientation Estimation. [†]: ImageNet3D is used for training Orient Anything V2. To ensure a fair comparison, the compared V1 model is fine-tuned on ImageNet3D. Best results are highlighted in **bold**.

tokens from all frames is then passed into a unified transformer block. The final learnable token corresponding to each frame is used for prediction. Specifically, the learnable token for the first frame is initialized differently and is used to predict the absolute orientation using the symmetry-aware distribution described in Sec. 5.1. Tokens from subsequent frames predict the object rotation relative to the first frame through a similar probability fitting task, but without considering symmetry.

5.3 Training Setting

Our model is initialized from VGGT, a large feed-forward transformer with 1.2 billion parameters pre-trained on 3D geometry tasks. We repurpose its original "camera" token, designed to predict camera extrinsics, to predict object orientation and rotation. This leverages the inherent correlation between camera pose and object rotation. We train the model to fit target orientation (or rotation) distributions using Binary Cross-Entropy (BCE) loss for 20k iterations. A cosine learning rate scheduler is used with an initial rate of 1e-3. Input frames are resized to 518, and random patch masking is used for data augmentation to simulate real-world occlusion. The effective batch size is set to 48, where 1-2 frames are randomly sampled for each training sample. The training dataset comprises the ImageNet3D training set and newly collected 600k synthetic assets. Furthermore, we observe that most objects exhibit only four types of rotational symmetry: $\{0, 1, 2, 4\}$. Therefore, we restrict our training to consider only these four cases. Any fitted periodicity $\bar{\alpha} > 4$ is mapped to 0.

6 Experiment

6.1 Zero-shot Orientation Estimation

Benchmark & Baselines Predicting the 3D orientation of objects from a single image is our core focus. We mainly compare with Orient Anything V1 [48] on ImageNet3D [29] test set and unseen test datasets, SUN-RGBD [41], ARKitScenes [3], Pascal3D+ [52], Objectron [1] and the Ori_COCO [48]. Since current testing datasets often provide only one ground truth orientation, even for symmetric objects, when Orient Anything V2 predicts multiple orientations, we simply select the one closest to facing the camera as the prediction. The main evaluation metrics are the median 3D angle error (Med↓) and accuracy within 30 degrees (Acc30°↑). For Ori_COCO, where 20 samples are collected for each class and annotated within 8 horizontal orientations, recognition accuracy (Acc↑) is used.

Main Results In Tab. 1, we present the comparative results on single view-based orientation estimation. Overall, Orient Anything V2 significantly improves upon V1, benefiting from diverse synthetic data and robust ensemble annotation. On the representative Ori_COCO benchmark, our method achieves 86.4% accuracy and performs well on categories where V1 struggled, such as bicycles. Achieving state-of-the-art results on numerous real-world image datasets highlights our method’s generalization ability.

6.2 Zero-shot Rotation Estimation

Benchmark & Baselines We benchmark zero-shot 6DoF object pose estimation performance under a single reference view. Evaluation is conducted on four widely used datasets: LINEMOD [16], YCB-Video [5], OnePose++ [15], and OnePose [44]. Objects are prepared using the cropping and matching, following [10]. Comparisons are made against three state-of-the-art zero-shot 6DoF object pose estimation methods: Gen6D [27], LoFTR [43], and POPE [10]. Standard metrics for relative object pose estimation are used: median error (Med) and accuracy within 15° and 30° (Acc15 and Acc30), computed for each sample pair.

Model	LINEMOD			YCB-Video			OnePose++			OnePose		
	Med↓	Acc30↑	Acc15↑	Med↓	Acc30↑	Acc15↑	Med↓	Acc30↑	Acc15	Med↓	Acc30↑	Acc15↑
<i>POPE's Sampling (Average rotation angle: 14.85°)</i>												
Gen6D	44.86	36.4	9.6	54.48	23.2	7.7	35.43	41.1	15.8	17.78	89.3	38.9
LoFTR	33.04	56.2	32.4	19.54	68.6	47.8	9.01	89.1	70.3	4.35	96.3	91.8
POPE	15.73	77.0	48.3	13.94	80.1	54.4	6.27	89.6	72.8	2.16	96.2	91.1
OriAny.V2	7.82	98.07	89.7	6.07	91.6	86.4	6.18	99.7	96.6	6.76	99.7	95.7
<i>Random Sample (Average rotation angle: 78.22°)</i>												
POPE	98.03	10.3	4.3	41.88	40.9	27.2	88.21	25.6	19.8	45.73	45.1	37.3
OriAny.V2	28.83	51.6	28.3	15.78	61.2	48.7	12.83	85.5	58.8	11.72	86.7	63.4

Table 2: Zero-shot Relative Rotation Estimation (i.e., pose estimation with one reference view). We evaluate two strategies for sampling query-reference view pairs: (1) query-reference pairs provided by POPE [10], and (2) randomly sampled pairs. The average rotation angles between views for each sampling strategy are 14.85° and 78.22°, respectively.

Main Results Tab. 2 includes zero-shot two-view relative rotation estimation results compared with state-of-the-art pose estimation methods. With small relative rotations (using POPE’s sampling), our model achieves the overall best performance across the four datasets. More importantly, our method’s advantage is significantly larger when the relative rotation between the query and reference frame is larger (using random sampling). The significant performance drop of the previous method stems from the reliance on explicit feature matching, which becomes unreliable with large rotations due to less view overlap and scarcer reliable matching points. In contrast, our approach understands images from different viewpoints by considering overall meaning rather than just detailed matching. This makes it more robust to challenging large rotations.

6.3 Zero-shot Symmetry Recognition

Benchmark & Baselines We assess our method’s zero-shot performance in predicting object rotational symmetry in the horizontal plane. This evaluation uses the recent, large-scale 3D object datasets with rotational symmetry annotations: Omni6DPose [58], which contain 149 distinct object classes. To ensure their orientation definition aligns with our front-facing direction, we manually select a subset of 3-5 assets per category and render 2 views per 3D asset for testing. This resulted in 838 testing sample. During inference, models receive a single rendering and predict the four kinds of rotational symmetry predictions. As there are currently no dedicated zero-shot models for predicting object rotational symmetry from a single view, we employ advanced VLMs (Qwen2.5VL-72B [2], GPT-4o [31], GPT-o3 [32], and Gemini-2.5-pro [13]) as baselines. We evaluate their ability to predict horizontal plane rotational symmetry using a multiple-choice format, with recognition accuracy as the metric.

Main Results We present a comparison of our method against various advanced general VLMs for identifying object horizontal rotational symmetry in Tab. 3. Our results indicate that recognizing object rotational symmetry is a challenging problem even for the strongest VLMs, thereby limiting their ability to fully understand the 3D spatial state from 2D images. In contrast, benefiting from high-quality annotations and a unified learning objective, our model achieves 65% accuracy in distinguishing object rotational symmetry. Combining this strong symmetry recognition ability alongside the robust and accurate absolute orientation estimation performance demonstrated in Sec. 6.1, our model can accurately infer multiple potential orientations from a single image in real applications.

	Omni6DPose
	Acc↑
Random	25.0
Qwen2.5VL-72B	55.8
Gemini-2.5-pro	44.4
GPT-4o	62.5
GPT-o3	53.7
OriAny. V2	65.2

Table 3: Zero-shot horizontal rotational symmetry recognition.

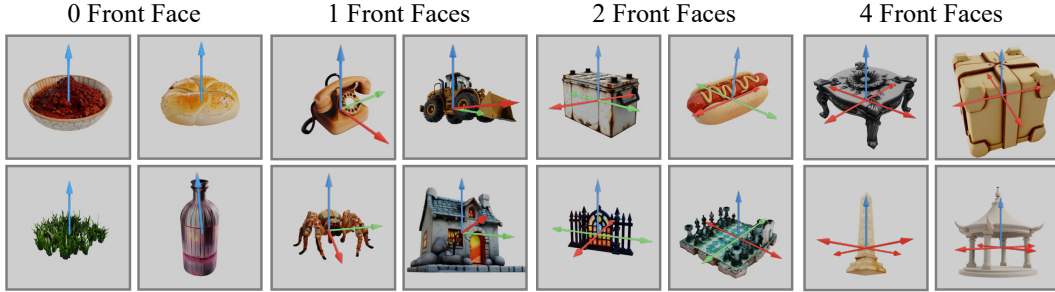


Figure 6: Visualization of synthetic 3D assets and robust annotation.

Row	Assets Type	Assets Number	Initialized Weights	Orientation Estimation			Rotation Estimation			
				Objectron		Ori_COCO	LINEMOD		YCB-Video	
				Med↓	Acc15↑	Acc↑	Med↓	Acc15↑	Med↓	Acc15↑
1	Real	40K	VGGT	25.05	54.3	74.8	10.70	69.8	15.49	72.5
2	Synthetic	40K	VGGT	24.44	54.6	74.6	10.16	74.1	7.28	76.2
3	Synthetic	200K	VGGT	23.82	55.0	75.2	10.22	74.5	7.49	78.6
4	Synthetic	400K	VGGT	25.09	53.9	75.4	9.78	76.3	6.48	80.5
5	Synthetic	600K	VGGT	22.62	54.8	86.4	7.82	89.7	6.07	84.7
6	Synthetic	600K	DINOv2	26.70	52.6	79.0	15.11	49.1	13.78	52.6
7	Synthetic	600K	None	62.08	13.5	25.6	16.54	45.3	13.93	52.2

Table 4: Ablation study. For the rotation estimation, we employ POPE’s sampling pairs.

6.4 Ablation Study

Quality of Synthetic 3D Assets Fig. 6 visualizes our synthetic dataset and the labelled orientation, qualitatively demonstrating the high quality of both the synthetic data and its annotations. Quantitatively, Rows 1 and 2 of Tab. 4 show the comparison of training with an equal amount of annotated real or synthetic 3D assets. We observe that both data sources yield comparable results for absolute orientation estimation. However, for rotation estimation (on LINEMOD and YCB-Video), training with synthetic assets provides a significant advantage. This may be because synthetic assets possess richer, more realistic textures, which are more crucial for understanding rotation.

Effect of Scaling Data In Rows 2, 3, 4, and 5 of Tab. 4, we explore the impact of data scale on the performance of final orientation and rotation estimation. Overall, with the same training step, encountering more diverse data and 3D assets during training leads to better overall performance. Specifically, we find that rotation estimation is more sensitive to data scale than orientation estimation. This may be because orientation relies on overall semantics and structure, while rotation estimation requires understanding diverse textures and fine-grain details to capture cross-view relationships.

Effect of Geometry Pre-training Tab. 4 (Rows 5-7) presents our experiments of different model initialization strategies. Training without any pre-trained initialization yields the worst results. Initializing the separated visual encoder with DINOv2 introduces valuable high-quality semantic and object structure information, leading to substantial performance gains. We observe further improvements in rotation estimation by using VGGT, pre-trained specifically on 3D geometric tasks, which boosts the model’s comprehension of object geometry.

7 Conclusion

We present Orient Anything V2, an advanced model for unified object orientation and rotation understanding. Through introducing the scalable data engine, a symmetry-aware distribution learning target, and a multi-frame framework, our model enables: 1) Stronger single-view absolute orientation estimation. 2) Advanced two-frame object relative pose rotation estimation. 3) Powerful object horizontal rotational symmetry recognition. In practice, the model can simultaneously and accurately predict multiple valid front faces of objects, making it well-suited for diverse objects and real-world application scenarios.

Limitation While our models exhibit strong generalization to diverse in-the-wild objects in real images, we find that the inherent ambiguity of monocular images leads to less accurate predictions in views with very low information or severe occlusion. Furthermore, the current framework supports a maximum of two input frames. Extending the model to handle more frames will be an important direction for supporting video understanding applications.

Acknowledgements

This work was supported in part by National Key R&D Program of China (No. 2022ZD0162000) and National Natural Science Foundation of China (No. 62222211, U24A20326, 624B2128, 62422606 and 62201484)

References

- [1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7822–7831, 2021.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.
- [4] Lonni Besançon, Anders Ynnerman, Daniel F Keefe, Lingyun Yu, and Tobias Isenberg. The state of the art of spatial interfaces for 3d visualization. In *Computer Graphics Forum*, volume 40, pages 293–326. Wiley Online Library, 2021.
- [5] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015.
- [6] Enric Corona, Kaustav Kundu, and Sanja Fidler. Pose estimation for objects with rotational symmetry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7215–7222. IEEE, 2018.
- [7] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.
- [8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023.
- [9] Ufuk Efe, Kutalmis Gokalp Ince, and Aydin Alatan. Dfm: A performance baseline for deep feature matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4284–4293, 2021.
- [10] Zhiwen Fan, Panwang Pan, Peihao Wang, Yifan Jiang, Dejia Xu, and Zhangyang Wang. Pope: 6-dof promptable pose estimation of any object in any scene with one reference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7771–7781, 2024.
- [11] Aaron L Gardony, Shaina B Martis, Holly A Taylor, and Tad T Brunyé. Interaction strategies for effective augmented reality geo-visualization: Insights from spatial cognition. *Human-Computer Interaction*, 36(2):107–149, 2021.
- [12] Google. Gemini-2.0-flash, 2025. https://aistudio.google.com/prompts/new_chat?model=gemini-2.0-flash-exp.
- [13] Google. Gemini-2.5-pro, 2025. <https://deepmind.google/technologies/gemini/pro/>.
- [14] Qi Guan, Zihao Sheng, and Shibe Xue. Hrpose: Real-time high-resolution 6d pose estimation network using knowledge distillation. *Chinese Journal of Electronics*, 32(1):189–198, 2023.

- [15] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022.
- [16] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012.
- [17] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11703–11712, 2020.
- [18] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6207–6217, 2021.
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.
- [20] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [21] Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. *arXiv preprint arXiv:2504.17207*, 2025.
- [22] Dingzhe Li, Yixiang Jin, Yuhao Sun, Hongze Yu, Jun Shi, Xiaoshuai Hao, Peng Hao, Huaping Liu, Fuchun Sun, Jianwei Zhang, et al. What foundation models can bring for robot learning in manipulation: A survey. *arXiv preprint arXiv:2404.18201*, 2024.
- [23] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022.
- [24] Xiang Li, Zixuan Huang, Anh Thai, and James M Rehg. Symmetry strikes back: From single-image symmetry detection to 3d generation. *arXiv preprint arXiv:2411.17763*, 2024.
- [25] Chuang Lin, Bingbing Zhuang, Shanlin Sun, Ziyu Jiang, Jianfei Cai, and Manmohan Chandraker. Drive-1-to-3: Enriching diffusion priors for novel view synthesis of real vehicles. *arXiv preprint arXiv:2412.14494*, 2024.
- [26] Jian Liu, Wei Sun, Hui Yang, Zhiwen Zeng, Chongpei Liu, Jin Zheng, Xingyu Liu, Hossein Rahmani, Nicu Sebe, and Ajmal Mian. Deep learning-based object pose estimation: A comprehensive survey. *arXiv preprint arXiv:2405.07801*, 2024.
- [27] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *European Conference on Computer Vision*, pages 298–315. Springer, 2022.
- [28] Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024.
- [29] Wufei Ma, Guofeng Zhang, Qihao Liu, Guanning Zeng, Adam Kortylewski, Yaoyao Liu, and Alan Yuille. Imagenet3d: Towards general-purpose object-level 3d understanding. *Advances in Neural Information Processing Systems*, 37:96127–96149, 2024.
- [30] Pedro Monteiro, Guilherme Gonçalves, Hugo Coelho, Miguel Melo, and Maximino Bessa. Hands-free interaction in immersive virtual reality: A systematic review. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2702–2713, 2021.
- [31] OpenAI. Gpt-4o, 2025. <https://openai.com/index/introducing-4o-image-generation/>.
- [32] OpenAI. Gpt-o3, 2025. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- [33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

- [34] Jonathan Palacios and Eugene Zhang. Rotational symmetry field design on surfaces. *ACM Transactions on Graphics (TOG)*, 26(3):55–es, 2007.
- [35] Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J Mitra. Diffusion handles enabling 3d edits for diffusion models by lifting activations to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7695–7704, 2024.
- [36] V Shiv Naga Prasad and Larry S Davis. Detecting rotational symmetries. In *Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*, volume 2, pages 954–961. IEEE, 2005.
- [37] Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, et al. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. *arXiv preprint arXiv:2502.13143*, 2025.
- [38] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.
- [39] Yifei Shi, Zixin Tang, Xiangting Cai, Hongjia Zhang, Dewen Hu, and Xin Xu. Symmetrygrasp: Symmetry-aware antipodal grasp detection from single-view rgb-d images. *IEEE Robotics and Automation Letters*, 7(4):12235–12242, 2022.
- [40] Yifei Shi, Xin Xu, Junhua Xi, Xiaochang Hu, Dewen Hu, and Kai Xu. Learning to detect 3d symmetry from single-view rgb-d images with weak supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4882–4896, 2022.
- [41] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.
- [42] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE international conference on computer vision*, pages 2686–2694, 2015.
- [43] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loft: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.
- [44] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022.
- [45] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15887–15898, 2024.
- [46] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2642–2651, 2019.
- [47] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025.
- [48] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *ICML*, 2025.
- [49] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024.
- [50] Ziyi Wu, Yulia Rubanova, Rishabh Kabra, Drew Hudson, Igor Gilitschenski, Yusuf Aytar, Sjoerd van Steenkiste, Kelsey Allen, and Thomas Kipf. Neural assets: 3d-aware multi-object scene synthesis with image diffusion models. *Advances in Neural Information Processing Systems*, 37:76289–76318, 2024.
- [51] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.
- [52] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014.

- [53] Yang Xiao, Vincent Lepetit, and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3090–3106, 2022.
- [54] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [55] Honghong Yang, Hongxi Liu, Yumei Zhang, and Xiaojun Wu. Fmr-gnet: Forward mix-hop spatial-temporal residual graph network for 3d pose estimation. *Chinese Journal of Electronics*, 33(6):1346–1359, 2024.
- [56] Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. *arXiv preprint arXiv:2503.22236*, 3, 2025.
- [57] Difeng Yu, Xueshi Lu, Rongkai Shi, Hai-Ning Liang, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. Gaze-supported 3d object manipulation in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.
- [58] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216. Springer, 2024.
- [59] Zhaoxuan Zhang, Bo Dong, Tong Li, Felix Heide, Pieter Peers, Baocai Yin, and Xin Yang. Single depth-image 3d reflection symmetry and shape prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8896–8906, 2023.
- [60] Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. *arXiv preprint arXiv:2410.17385*, 2024.
- [61] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: refer to the abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: refer to section 7.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: the paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: refer to section 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: refer to section 6.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: refer to section 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Our results are deterministic.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Refer to section 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: yes

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: refer to section 7.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: the paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: refer to section 6.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: refer to section 4.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: no need.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A More Visualizations of Images in The Wild

In 7 8, 9, 10, 11, 12, 13, we present more visualizations of images from various domains containing different objects. In these images, our model shows strong abilities in single-view absolute orientation estimation, powerful object horizontal rotational symmetry recognition and two-frame object relative pose rotation estimation, further highlighting the impressive zero-shot capability of Orient Anything V2.

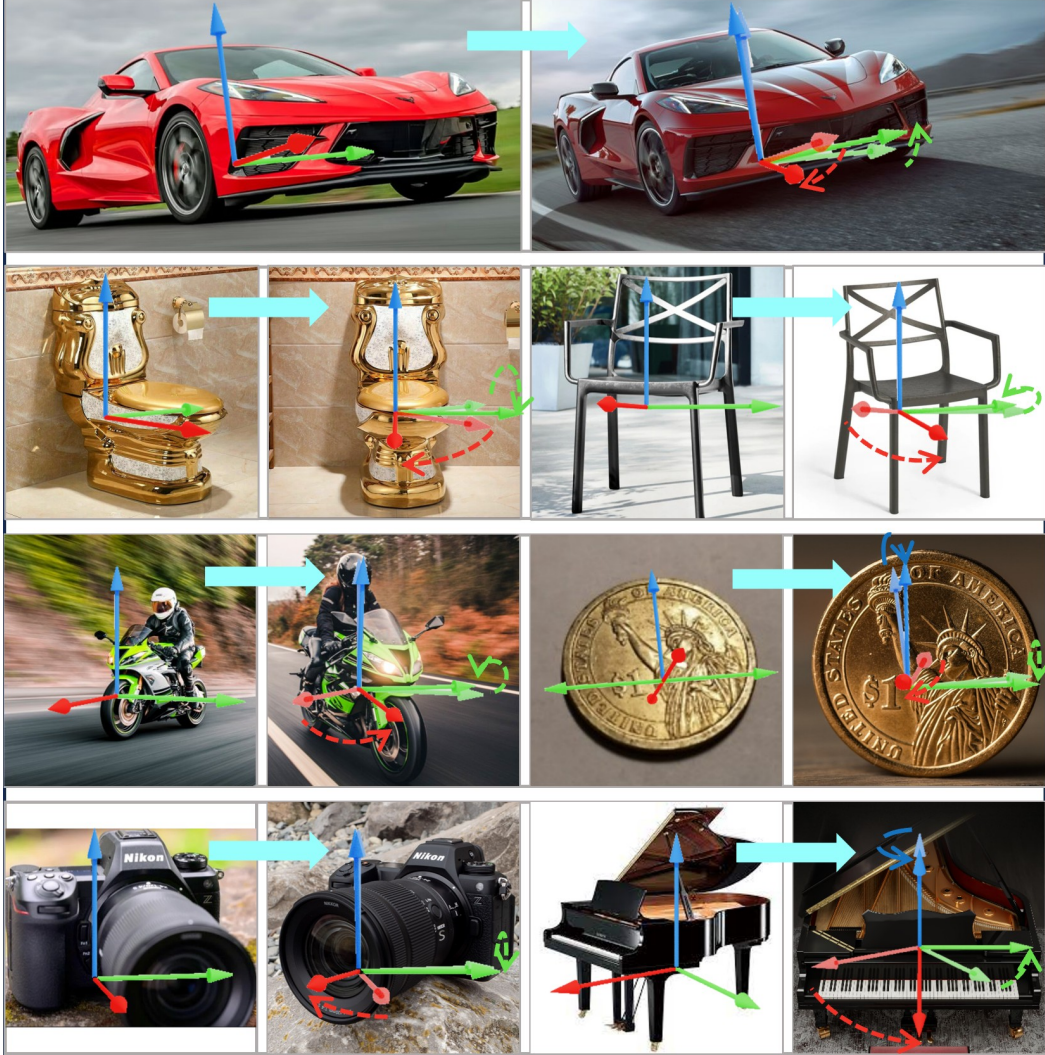


Figure 7: Relative pose rotation estimation for images in the wild.

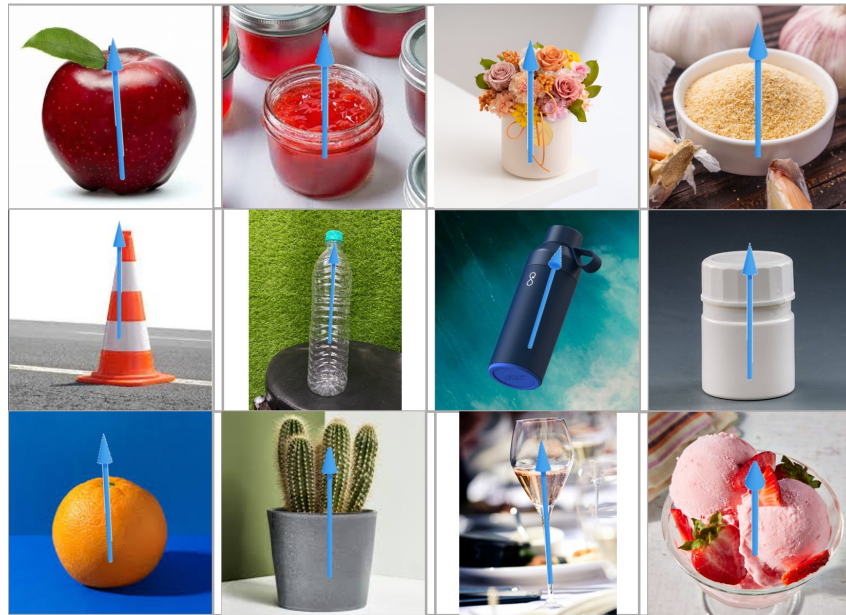


Figure 8: Orientation estimation and Rotational symmetry recognition results on objects has no front direction.



Figure 9: Orientation estimation and Rotational symmetry recognition results on objects has one front direction. Part 1.

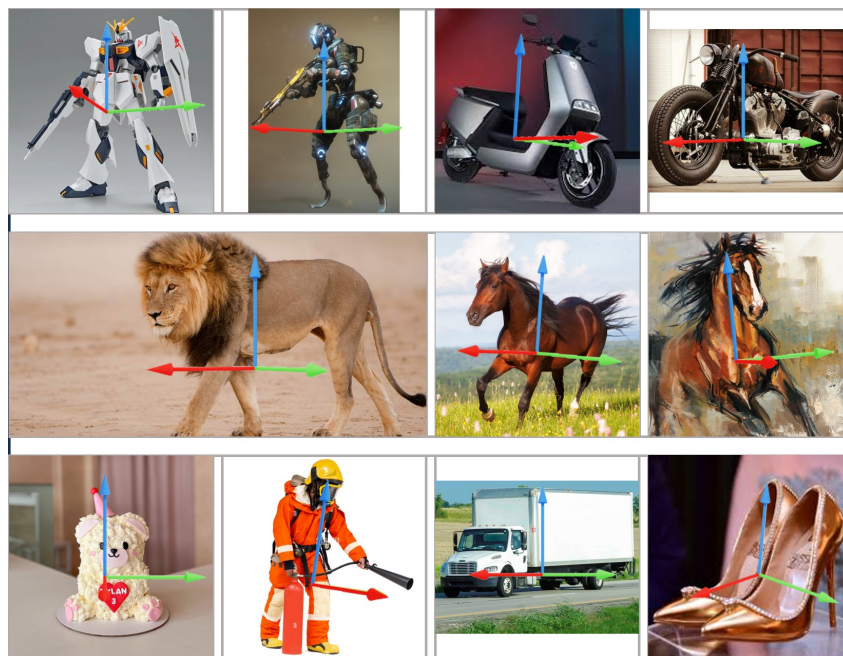


Figure 10: Orientation estimation and Rotational symmetry recognition results on objects has one front direction. Part 2.



Figure 11: Orientation estimation and Rotational symmetry recognition results on objects has two front direction. Part 1.

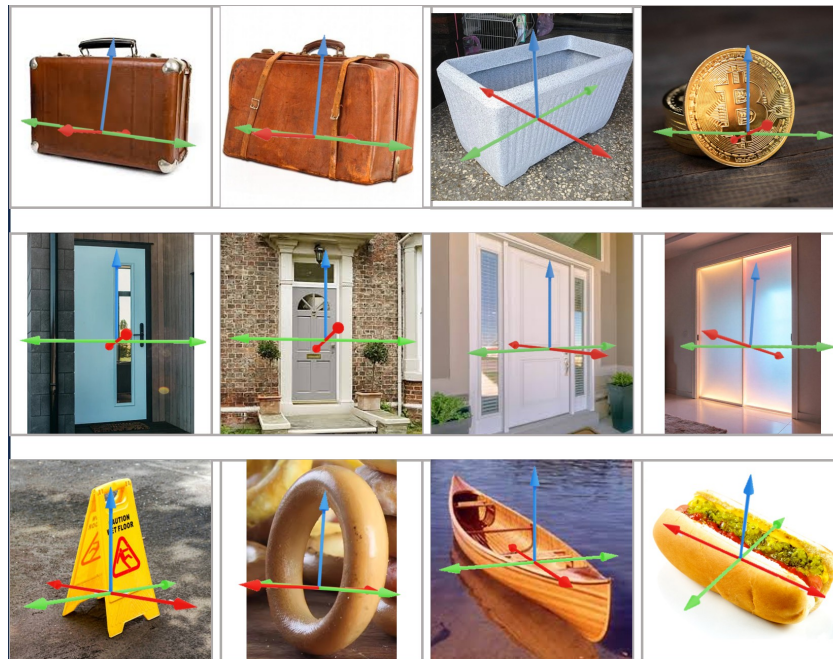


Figure 12: Orientation estimation and Rotational symmetry recognition results on objects has two front direction. Part 2.

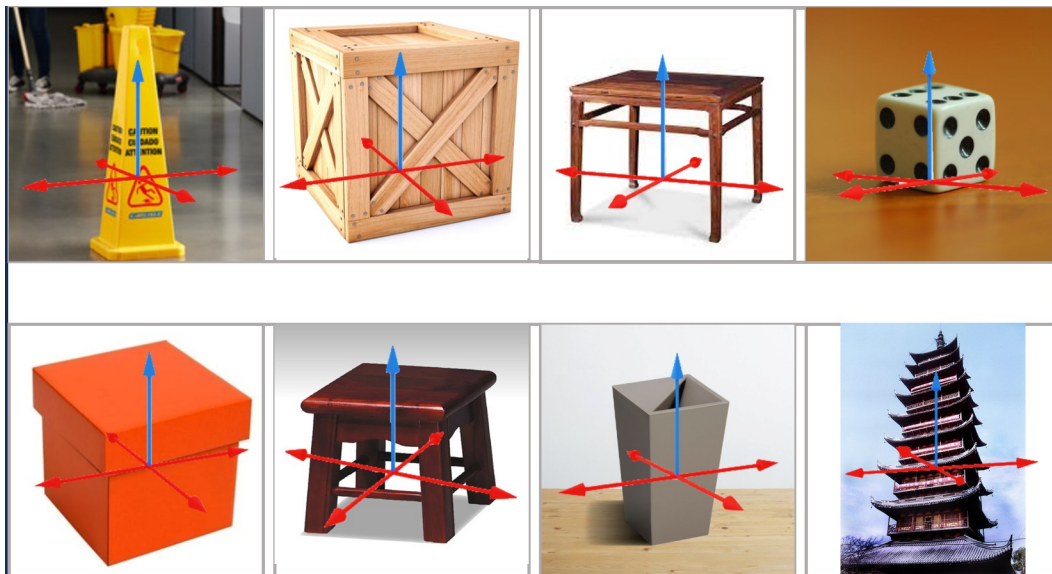


Figure 13: Orientation estimation and Rotational symmetry recognition results on objects has four front direction.