CSCI566 Project Proposal

TTop: Reliable Depth for Small Tabletop Objects

Yumeng He
University of Southern California
heyumeng@usc.edu

Xiaoying Wang
University of Southern California
xwang648@usc.edu

Jingkai Shi
University of Southern California
shikings@usc.edu

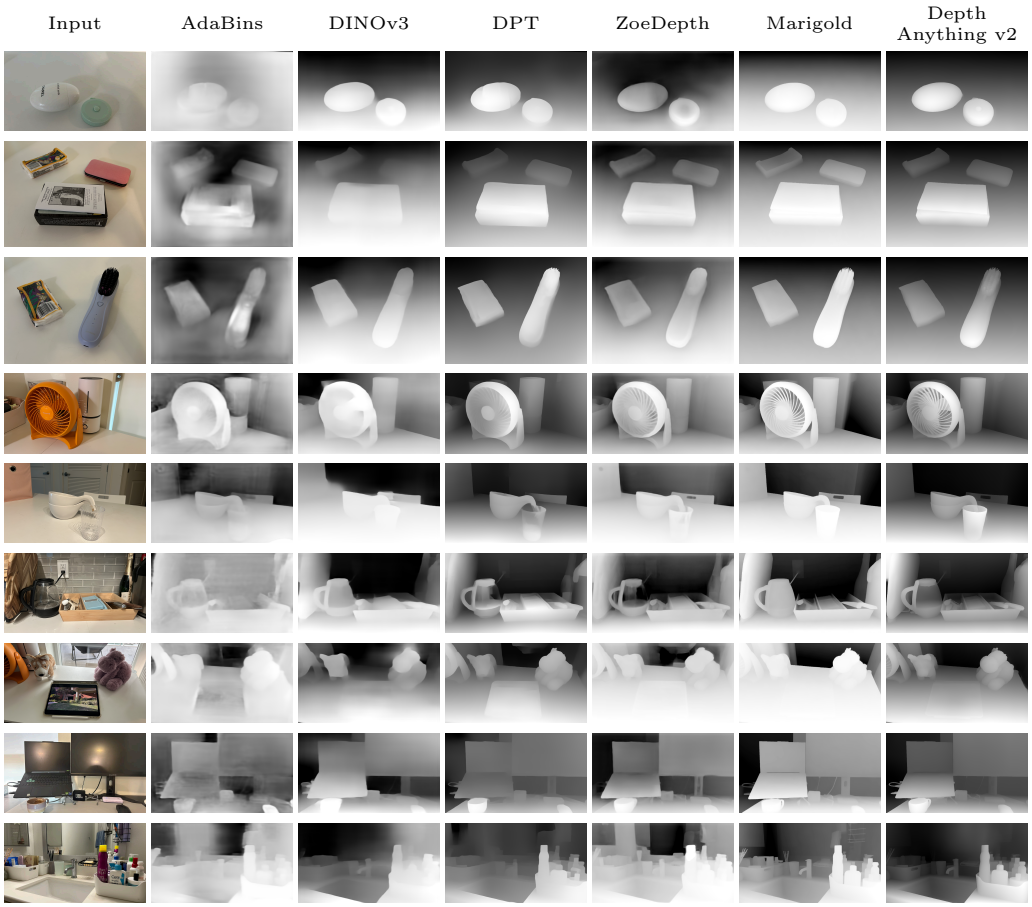

Figure 1: Qualitative comparison of monocular depth on tabletop scenes. Each row shows one RGB scene (left) and the corresponding depth predictions from six off-the-shelf baselines: AdaBins, DINOv3, DPT, ZoeDepth, Marigold, and Depth Anything v2. The rows are curated to stress common tabletop failure modes such as fine-scale geometry, transparent materials, pictorial 3D, and strong cast shadows.

Preprint. Under review.

Abstract

Depth estimation in tabletop scenes often fails for small objects that provide weak visual cues. These objects yield fat boundaries, missing parts, and unstable scale, which block downstream steps such as creating tight three-dimensional bounding boxes and accurately determining orientation, scale, and position in real-to-sim pipelines. While large structures are handled reasonably well by current monocular methods, small items remain brittle. We work in the single-image setting and target this gap and show that our design also benefits larger objects. We cast refinement as a depth to depth denoising problem and propose a two stage pipeline. Stage one builds pseudo ground truth by segmenting each object with a large language model guided semantics to ignore mirrors, paintings, and transparent or reflective surfaces, and to produce text prompts for grounded segmentation, and applying instance wise optimization that combines local geometric priors such as piecewise planar and quadric surfaces, edge-aware boundary regularization, and scene-level scale and shift calibration. This stage also produces a per pixel confidence map. Stage two trains a noise-conditioned model that takes a noisy depth map as input and returns a refined depth map in a single forward pass. Experiments on our newly collected tabletop dataset of about 3000 images with ground truth depth, with 200 held out photos for validation, show consistent gains over Depth Anything V2, Marigold, MiDaS, and ZoeDepth in delta accuracy and SILog, with sharper boundaries and more accurate three-dimensional boxes and six degrees-of-freedom (DoF) pose estimation. Used as a drop-in depth enhancer from a single RGB image, our outputs also improve downstream semantic segmentation, depth estimation, object detection, and object discovery, and the dataset and evaluation protocol constitute an additional contribution.

1 Introduction

Monocular depth is central to a wide range of downstream tasks, from three-dimensional (3D) reconstruction, navigation, and scene understanding to modern applications in content creation and interactive simulation [1, 2]. In tabletop scenes, depth quality directly governs grasp success, collision avoidance, and precise placement for robot manipulators, while also stabilizing real-to-sim pipelines and automated hyperparameter optimization [3–5]. However, current monocular estimators struggle with small, texture-poor, and self-occluded objects, producing thick/over-smoothed boundaries, missing parts, and unstable scale that degrade 2D boxes, 6D pose, and grasp scoring [6–10]. We revisit tabletop depth as a single-image depth-to-depth refinement with instance-aware geometry and semantic guidance [11–13], yielding sharper boundaries, stable scale, and manipulation-ready depth for robotics and perception.

Despite the rapid progress of monocular foundation models, tabletop depth remains brittle due to four recurring factors. (i) Small parts and high-frequency details are easily smoothed out, yielding thick boundaries and missing structures [14, 15]. (ii) Transparent and specular surfaces (e.g., glass, mirrors, glossy plastics) entangle physical geometry with reflections and refractions [16, 17]. (iii) Screens and photographs introduce rich but misleading internal depth cues on a planar surface [11, 13, 17]. (iv) Strong cast shadows obscure contours or create spurious edges [15, 18]. These factors amplify the ill-posedness of monocular depth and highlight a shortage of instance-aware priors in current estimators.

Our key insight is to treat tabletop scenes as sets of manipulable instances and to refine depth by coupling geometry with semantics at the instance level. Concretely, we construct pseudo ground-truth depth and a confidence map via instance-aware optimization guided by vision-language semantics and material/support priors (e.g., support plane, part identity, reflectivity). A lightweight transformer then performs depth-to-depth refinement in a single forward pass, trained with confidence-weighted, boundary-preserving losses and scale-stabilizing objectives that anchor depth across objects and the support surface [14, 19].

Across diverse real tabletop photos, our approach yields sharper boundaries, recovered small parts, and a more stable relative scale, facilitating downstream segmentation and manipulation.

The main contributions of this paper are summarized below:

- We introduce an object-centric refiner with boundary-preserving objectives that explicitly protect thin parts and high-frequency details, turning over-smoothed edges and missing micro-structures into sharp, contiguous geometry.
- We incorporate semantics-driven priors and confidence modeling for transparent/specular regions (glass, mirrors, glossy plastics), suppressing spurious depths from reflections/refractions and preventing "hallucinated" virtual content from contaminating physical geometry.
- We detect and regularize planar pictorial surfaces (screens, photographs) so that depicted 3D scenes are treated as flat carriers while nearby real objects retain correct relief, reducing depth leakage from 2D imagery.
- We curate a high-precision dataset of 3,000 pixel-aligned triplets: RGB image, semantic segmentation mask, and depth map, covering small parts, transparent/specular materials, pictorial surfaces, and strong shadows. The dataset is built to stress the above failure modes and to support future training and benchmarking of instance-aware depth refinement.

2 Related Work

Monocular depth estimation has been widely studied, from early CNN-based models [20] to more recent transformer approaches such as DPT [21] and DINOv2 [22]. These global methods achieve strong performance on benchmarks but often produce coarse depth in cluttered tabletop settings, where objects are small, textureless, and self-occluded. At the same time, foundation models for segmentation, such as the Segment Anything Model (SAM) [23] have demonstrated the ability to generalize mask generation across domains. Together, these advances highlight the potential of combining segmentation and depth estimation, but also reveal a gap: few methods directly exploit segmentation for object-level depth refinement.

Monocular depth estimation Building on this progression, several representative depth models demonstrate the field's trajectory and limitations. AdaBins [24] extended CNNs by learning adaptive depth bins per-image for metric estimation. Later, transformer-based approaches emerged, such as DPT [21] and ZoeDepth [25]. DPT reassembles patch embeddings from a ViT backbone for dense regression, while ZoeDepth combines a relative depth network with a specialized metric scaling head. Foundation models like DINOv3 [22] showed that powerful self-supervised features require only a simple linear probe for depth prediction. Generative approaches such as Marigold [26] repurpose latent diffusion models to generate highly detailed results. The current state-of-the-art Depth Anything V2 (DAv2) [27], uses a DINOv2 backbone with large-scale pseudo-labeling to achieve strong performance, setting a new benchmark on the NYU-D dataset where its fine-tuned ViT-L model achieves an AbsRel of 0.056 and a δ_1 accuracy of 98.4%.

To ground our work, we qualitatively evaluate existing methods on challenging tabletop scenes, revealing a clear performance hierarchy and common failure modes. The CNN-based AdaBins produces the weakest results, with severely blurred contours. Foundation models like DINOv3 yield smooth but overly flat reconstructions. Supervised methods like DPT and ZoeDepth are more coherent, with DPT excelling at local detail while ZoeDepth offers better global consistency. The generative Marigold provides highly detailed maps but can hallucinate geometry. Depth Anything v2 consistently delivers state-of-the-art detail and sharpness. Despite these advancements, our evaluation reveals that all models still struggle with fine-scale geometry, transparent objects, and mirrors, motivating our work.

Denoising Vision Transformers. The Denoising Vision Transformer (DVT) [28] shows that pre-trained ViTs contain noisy feature maps that degrade dense predictions like depth. By introducing a denoiser module, DVT stabilizes features and improves performance across tasks. Crucially, its two-stage design—first denoising features, then evaluating on downstream tasks—inspires our own methodology, where we refine depth in Stage 1 and train a network in Stage 2. Unlike DVT, which denoises globally across the entire scene, we extend this idea locally: segmenting objects and applying per-object priors to recover sharper boundaries and accurate 3D bounding boxes for tabletop scenes.

Segment Anything. The Segment Anything Model (SAM) [23] introduced a universal framework for object segmentation, later extended by Grounded-SAM [29] for text-prompted detection and masks.

These models are powerful for mask generation but are rarely leveraged for depth refinement. In our pipeline, segmentation serves as a critical intermediate step, enabling object-wise depth refinement and tighter 3D bounding boxes. SAM therefore provides the building block that allows us to address the limitations of global monocular estimators.

3 Pipeline

3.1 Proposed Method

We cast tabletop depth refinement as single-image, depth-to-depth learning. From one RGB image I without intrinsics, we synthesize high-quality pseudo ground truth depth $D^\star$ and a per-pixel confidence map C via an instance-aware optimization pipeline steered by semantic cues from a vision–language model (VLM). These pseudo labels supervise a lightweight transformer that takes a noisy monocular depth map D and predicts a refined map $\hat{D}$ in one forward pass. Training is confidence-weighted and boundary-aware, which sharpens discontinuities, stabilizes global scale, and preserves small objects that are otherwise smoothed away.

3.2 Stage 1: Pseudo ground truth construction

Given a single tabletop image $I \in \mathbb{R}^{H\times W\times 3}$, our target is a refined depth $\hat{D} \in \mathbb{R}^{H\times W}$. Stage 1 constructs pseudo ground truth $D^\star$ and confidence $C \in [0,1]^{H\times W}$, together with instance masks $\{M_k\}_{k=1}^{K}$. We begin by querying a VLM for a structured scene sketch that lists object categories, materials, coarse locations, and special flags (mirrors, screens, transparent surfaces). This semantic pass serves two purposes: (i) it identifies regions that are inherently unreliable for geometric supervision and therefore should be ignored, and (ii) it proposes geometry priors (planar, piecewise planar, cylindrical, spherical, quadric, thin planar) that regularize subsequent fitting, which is crucial for small parts with weak monocular cues.

Grounded by the VLM's short textual queries, we obtain instance proposals and refine them into clean masks via a modern text-conditioned detector/segmenter. Low-confidence or semantically mismatched proposals are filtered with text–image similarity checks. Each accepted mask is morphologically cleaned, and a signed distance transform provides inner/outer boundary bands that localize later boundary operations without contaminating object interiors.

To initialize depth without camera intrinsics, we aggregate two complementary monocular predictors on I. Following log-depth normalization to $[0,1]$, we form a fused estimate $D^{(0)}$ by a per-pixel selection/weighting that prefers sharper edges where predictors disagree, while retaining a local variance map V as a measure of estimator disagreement. We then correct common monocular artifacts—boundary fattening and occlusion ambiguity—by applying band-limited, edge-aware filtering guided by image gradients; along surface normals crossing mask boundaries, we enforce plausible near–far steps and choose the sharper, lower-variance hypothesis where needed.

Inside each instance M_k, we fit the semantically suggested principal geometry with robust estimators (e.g., RANSAC with robust penalties), treat the current depth as a height field, and model the residual $R_k = D^{(0)} - D_k^{\text{model}}$. Residuals are denoised by instance-restricted total variation or anisotropic diffusion to preserve thin ridges and corners, and very small parts are temporarily upsampled to avoid oversmoothing before being downsampled back. Missing regions are completed by in-instance interpolation for small holes and by Poisson/Laplacian inpainting for larger ones, with strong ridge preservation on thin structures. At the scene level, we fuse instance reconstructions, resolving overlaps by favoring lower model residuals and estimator variance, and, when a table is detected semantically, we fit a planar support and enforce non-penetration at contact bands. A light, edge-aware harmonization outside boundary bands aligns seams while keeping discontinuities crisp.

Finally, we assign per-pixel confidence by combining geometric residuals, boundary proximity, estimator variance, hole-fill ratios, and semantic risk (high for reflective/transparent regions). Concretely,

$$C(x) = \sigma(-\alpha r(x) + \beta d(x) - \gamma v(x) - \delta h(x) - \eta s(x))$$

and we set $C = 0$ in ignored regions. A suite of quality checks (boundary thickness, contact consistency, abnormal residual variance) triggers local fallbacks—e.g., switching to simpler priors—when inconsistencies are detected. The output of Stage 1 is $(D^\star, C, \{M_k\})$ and associated semantics.

3.3 Stage 2: Transformer-based depth refinement

At training time, the model receives a noisy depth map D from any monocular estimator and learns a mapping $f_\theta : \mathbb{R}^{H\times W} \to \mathbb{R}^{H\times W}$ such that $\hat{D} = f_\theta(D)$ approximates $D^\star$. The architecture is a compact encoder–decoder transformer operating on non-overlapping $p \times p$ depth patches. Each patch is flattened in log-depth space, linearly projected to a token, and endowed with 2D sine positional encodings. The backbone comprises LL layers of windowed multi-head self-attention with periodic global tokens to propagate scene-level context needed for consistent scale and long-range symmetry. Optionally, an instance-raster channel and a distance-to-edge map can be concatenated to token embeddings; the base model trains without them. A lightweight convolutional head upsamples the token grid and predicts a residual R in log-depth, yielding $\hat{D} = D + R$. This residual formulation preserves the coarse structure of D and focuses learning capacity on corrections around boundaries, thin parts, and mis-scaled regions.

Training minimizes a confidence-weighted objective that balances scale-invariant fidelity, edge sharpness, and surface orientation:

$$\ell(x) = \lambda_{\mathrm{si}}\ell_{\mathrm{SILog}}(\hat{D}(x), D^\star(x)) + \lambda_\nabla\|\nabla\hat{D}(x) - \nabla D^\star(x)\|_1 + \lambda_{\mathrm{n}}\big(1 - \langle\hat{n}(x), n^\star(x)\rangle\big)$$

and

$$\mathcal{L} = \sum_x w(x)\,\ell(x), \quad w(x) = C(x)\cdot b(x)\cdot s_{k(x)}$$

Here $b(x)$ boosts pixels within a fixed band around instance boundaries to enforce crisp discontinuities, while s_k emphasizes small objects, for example $s_k = 1/\sqrt{\mathrm{area}(M_k)}$ to counter dataset-level bias against tiny parts. We apply standard photometric/geometric augmentations when forming pseudo labels and mild perturbations in depth space during training to promote robustness across upstream monocular predictors.

At inference, the transformer consumes a single noisy depth map and emits a refined depth in one pass; no semantics, intrinsics, or masks are required. Coupled with Stage 1 supervision, this design yields manipulation-ready depth with sharp edges, stable global scale, and faithful reconstruction of small and thin structures, while remaining a drop-in enhancer for arbitrary monocular estimators.

3.4 Validation Strategy

Inspired by the methodology of DVT [30], which validates upstream improvements by measuring performance on a suite of downstream tasks, our validation strategy is centered on quantifying the performance uplift our refined depth maps provide to three core downstream tasks. For each task, we will employ a state-of-the-art model and evaluate its performance on a standard benchmark, comparing the results obtained with our refined depth against a carefully established baseline.

1. 3D Object Detection. To demonstrate that our depth refinement enables more accurate 3D scene understanding, we will use the 3DETR [31] model on the SUN RGB-D [32] benchmark. We will first reproduce the baseline performance reported in the original 3DETR paper ($59.1AP_{25}$) by running the pre-trained model on the original depth maps to ensure consistency. Subsequently, we will run the same model on our refined depth maps. A significant improvement in mean Average Precision (mAP) over the established baseline will validate the geometric benefits of our approach.

2. 6D Object Pose Estimation. To show that providing our refined depth boosts the accuracy of pose estimation models, we will use the MegaPose [33] framework on the YCB-Video [34] dataset. To ensure a fair comparison, we will first establish our own baseline by submitting pose predictions generated using the original depth maps to the official BOP Challenge evaluation platform. We will then submit a second set of results generated using our refined depth maps. An increase in the official

Average Recall (AR) score over our own baseline will directly demonstrate the positive impact of our depth refinement.

3. Robotic Grasp Planning. To validate that our refined depth allows grasp generation models to produce more reliable grasps, we will use the pre-trained Contact-GraspNet [35] model on the GraspNet-Billion [36] benchmark. We will conduct a controlled comparison using the official evaluation code provided by the GraspNet authors. We will establish a baseline by calculating the Average Precision (AP) on the original depth maps, and then calculate a new AP score using our refined depth maps. An increase in the AP score will directly demonstrate our method's value for robotic manipulation.

3.5 Potential Obstacles and Limitations

While our two-stage approach is designed to rectify key weaknesses in existing monocular depth estimators, its complexity introduces a corresponding set of potential obstacles. The Stage 1 pipeline, in particular, relies on a cascade of heuristic-driven modules, where the success of the entire process hinges on the performance of each component.

Failures in Stage 1 can arise from several sources. First, the entire process is predicated on the semantic interpretation from the Vision-Language Model (VLM). Beyond generating hallucinations or failing to identify challenging content like mirrors, the VLM's choice of geometric prior is critical. An incorrect prior can lead to high residual errors, triggering our fallback mechanism to cycle through alternative priors. However, it is possible that no alternative provides a substantially better fit. Second, the text-to-mask segmentation pipeline is a potential bottleneck. The performance of Grounded-SAM is sensitive to the VLM-generated text prompts. For challenging cases such as mirrored, reflective, or shadow-obscured boundaries, automated segmentation may be unreliable. In such cases, manual inspection and interactive segmentation (e.g., using point or box prompts with SAM) may be required to ensure mask accuracy. These mask imperfections, if not corrected, directly impact all downstream geometry operations.

Furthermore, several optimization steps introduce their own risks. The initial depth fusion of two models may produce sub-optimal results if both models share a common failure mode—for instance, if both misinterpret a textureless curved surface as planar. In such scenarios, our fusion logic, which selects based on edge sharpness or local variance, lacks a correct source to choose from. The subsequent boundary refinement using guided filtering is susceptible to noise in the guide image (the RGB input); artifacts in the color image can be incorrectly transferred to the depth map, creating plausible but false geometric details. The core geometric fitting and residual smoothing step is also fragile. RANSAC may fail to find a stable fit if the percentage of inliers in the initial depth is too low. Moreover, the residual smoothing process involves an inherent trade-off: aggressive smoothing can erase subtle but meaningful surface details, while gentle smoothing may fail to remove sufficient noise from the residual.

Finally, at the scene level, integration failures can occur. The logic for resolving overlaps between instances might make an incorrect decision, leading to unnatural seams. The final Laplacian harmonization, while edge-aware, could still introduce minor smoothing artifacts at the junctions between different objects. These potential issues underscore the heuristic nature of Stage 1, where the final quality of the pseudo ground truth is a product of many interdependent, and sometimes fragile, automated decisions.

Challenges also exist in the Stage 2 transformer-based refinement. The refinement transformer may learn systematic biases from the pseudo-GT generator. If the Stage 1 pipeline consistently produces overly flat planes or sharp edges, the transformer will learn to replicate these artificial characteristics rather than representing true scene geometry. Furthermore, a key challenge is that a transformer predicting residuals is inherently designed for refinement, not whole-scale reconstruction. For catastrophic errors, such as a hallucinated reflection, the required correction is a large-scale, non-local transformation, which lies outside the typical capabilities of such a model. Our proposed solution addresses this by using the Stage 1 pipeline to generate a large corpus of training pairs that explicitly demonstrate these corrections. By providing the transformer with numerous examples of structurally flawed inputs and their perfectly resolved ground truth counterparts, we hypothesize that the model can learn these complex mappings and overcome the typical limitations of a residual-based refiner.

4 Timeline

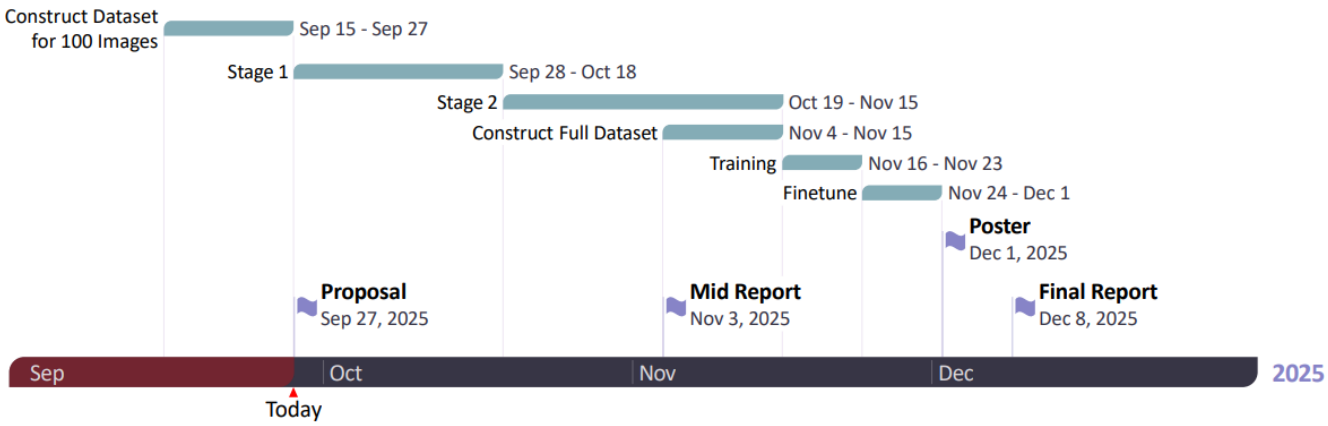

Figure 2: We include a timeline that estimates our progress on this project

References

[1] Richard A. Newcombe, Shahram Izadi, et al. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE Int. Symp. on Mixed and Augmented Reality (ISMAR)*. doi: 10.1109/ISMAR.2011.6092378. URL https://doi.org/10.1109/ISMAR.2011.6092378.

[2] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*. URL https://openaccess.thecvf.com/content_cvpr_2016/papers/Schonberger_Structure-From-Motion_Revisited_CVPR_2016_paper.pdf.

[3] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, et al. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Robotics: Science and Systems (RSS)*. URL https://arxiv.org/abs/1703.09312.

[4] Andreas ten Pas, Marcus Gualtieri, Kate Saenko, and Robert Platt. Grasp pose detection in point clouds. URL https://arxiv.org/abs/1706.09911.

[5] Tomas Hodan et al. Bop challenge 2020 on 6d object localization. In *ECCV Workshops*. URL https://arxiv.org/abs/2009.07378.

[6] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. . URL https://arxiv.org/abs/1907.01341.

[7] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, . URL https://arxiv.org/abs/2103.13413.

[8] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Zoedepth: Zero-shot transfer by combining relative and metric depth. URL https://arxiv.org/abs/2302.12288.

[9] Lintu Yang et al. Depth anything v2. URL https://arxiv.org/abs/2406.09414.

[10] Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Repurposing diffusion-based image generators for monocular depth estimation (marigold). In *CVPR*. URL https://arxiv.org/abs/2312.02145.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. URL https://arxiv.org/abs/2304.02643.

[12] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, et al. Grounded language-image pre-training. In *CVPR*. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Li_Grounded_Language-Image_Pre-Training_CVPR_2022_paper.pdf.

[13] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*. URL https://arxiv.org/abs/2303.05499.

[14] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, . URL https://openaccess.thecvf.com/content_cvpr_2017/papers/Godard_Unsupervised_Monocular_Depth_CVPR_2017_paper.pdf.

[15] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, . URL https://openaccess.thecvf.com/content_ICCV_2019/papers/Godard_Digging_Into_Self-Supervised_Monocular_Depth_Estimation_ICCV_2019_paper.pdf.

[16] Shreeyak S. Sajjan et al. Cleargrasp: 3d shape estimation of transparent objects for manipulation. In *ICRA*. URL https://arxiv.org/abs/1910.02550.

[17] Jie Tan et al. Mirror3d: Depth refinement for mirror surfaces. In *CVPR*. URL https://openaccess.thecvf.com/content/CVPR2021/papers/Tan_Mirror3D_Depth_Refinement_for_Mirror_Surfaces_CVPR_2021_paper.pdf.

[18] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *ECCV*. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123640562.pdf.

[19] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*. URL https://arxiv.org/abs/1406.2283.

[20] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. URL https://arxiv.org/abs/1406.2283.

[21] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction, 2021. URL https://arxiv.org/abs/2103.13413.

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

[24] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4008–4017. IEEE, June 2021. doi: 10.1109/cvpr46437.2021.00400. URL http://dx.doi.org/10.1109/CVPR46437.2021.00400.

[25] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. URL https://arxiv.org/abs/2302.12288.

[26] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation, 2024. URL https://arxiv.org/abs/2312.02145.

[27] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. URL https://arxiv.org/abs/2406.09414.

[28] Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers, 2024. URL https://arxiv.org/abs/2401.02957.

[29] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. URL https://arxiv.org/abs/2401.14159.

[30] Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. In *European Conference on Computer Vision*, pages 453–469. Springer, 2024.

[31] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2906–2917, 2021.

[32] 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.

[33] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022.

[34] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[35] Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13438–13444. IEEE, 2021.

[36] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020.