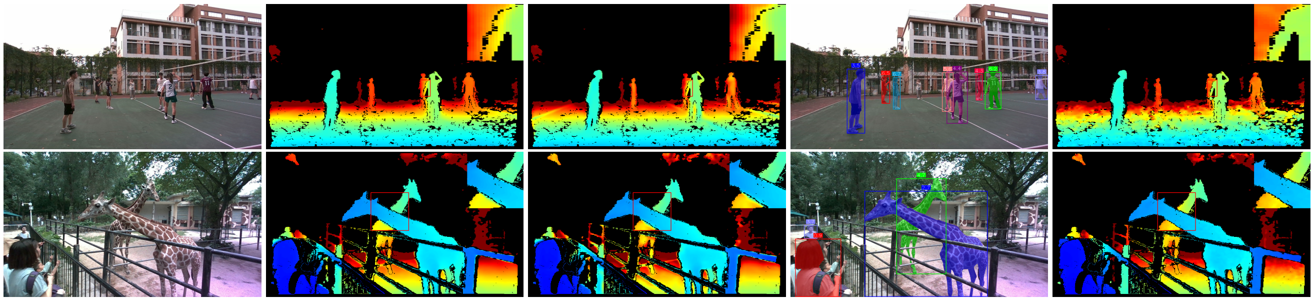
Instance-Level Video Depth in Groups Beyond Occlusions

Yuan Liang[1,2], Yang Zhou[1,2], Ziming Sun[1,2], Tianyi Xiang[1,2], Guiqing Li[1], and Shengfeng He[2*]
[1]South China University of Technology [2]Singapore Management University



(a) Objects in a Group (b) Depth Anything V2 [53] (c) Zoedepth [4] (d) Instance Mask (e) Ours

Figure 1. A typical scenario involving a complex dynamic group that cannot be tackled by existing monocular depth estimation methods. We contribute the first Group Instance Depth Dataset and a novel instance-aware depth estimation framework. Our method effectively captures instance-wise depth relationships, mitigating occlusion ambiguities and improving depth accuracy in dynamic multi-object scenes.

Abstract

Depth estimation in dynamic, multi-object scenes remains a major challenge, especially under severe occlusions. Existing monocular models, including foundation models, struggle with instance-wise depth consistency due to their reliance on global regression. We tackle this problem from two key aspects: data and methodology. First, we introduce the Group Instance Depth (GID) dataset, the first large-scale video depth dataset with instance-level annotations, featuring 101,500 frames from real-world activity scenes. GID bridges the gap between synthetic and real-world depth data by providing high-fidelity depth supervision for multi-object interactions. Second, we propose InstanceDepth, the first occlusion-aware depth estimation framework for multi-object environments. Our two-stage pipeline consists of (1) Holistic Depth Initialization, which assigns a coarse scene-level depth structure, and (2) Instance-Aware Depth Rectification, which refines instance-wise depth using object masks, shape priors, and spatial relationships. By enforcing geometric consistency across occlusions, our method sets a new state-of-the-art on the GID dataset and multiple benchmarks. Our code and dataset can be found at `https://github.com/ViktorLiang/GID`.

1. Introduction

The success of monocular depth estimation has been largely driven by the availability of large-scale datasets, which enable models to learn depth relationships from diverse visual cues. Recent advancements in foundation models, such as Depth Anything V1 [52] and Depth Anything V2 [53], have demonstrated that large-scale training improves generalization across various domains. These models leverage extensive data and powerful neural architectures to estimate depth from a single image without scene-specific supervision. However, despite their large-scale training, they still struggle in complex real-world scenarios, particularly those involving occlusions.

One key limitation stems from the nature of existing datasets, which predominantly focus on structured environments such as indoor scenes [14, 44] and autonomous driving [19], where objects are mostly static and occlusions are minimal. In contrast, real-world activity scenes, such as sports events, performances, and crowded environments, feature multiple interacting objects that frequently occlude one another. These occlusions introduce depth ambiguities that pose significant challenges for monocular depth estimation models.

Beyond dataset limitations, most monocular depth estimation methods, even those trained on large-scale datasets, formulate depth estimation as a global regression problem, optimizing per-pixel depth errors [16, 41, 52, 53, 55]. While this approach allows models to generalize across diverse scenes, it fails in cluttered environments where depth dis-

[*]Corresponding author: Shengfeng He (shengfenghe@smu.edu.sg).

continuities arise due to overlapping objects. Since these methods do not explicitly model object-level depth relationships, they struggle to infer occluded regions correctly, often producing depth inconsistencies in multi-object interactions (see Figure 1b). Some approaches [2, 4, 18] attempt to address this issue using relative depth decoders, but their predictions are constrained by a global depth range and do not explicitly account for occluded object structures. As a result, they fail to maintain depth consistency in occluded regions, particularly in complex, multi-object environments (see Figure 1c).

To overcome these challenges, we argue that depth estimation must incorporate instance-level semantics, explicitly modeling object interactions, occlusion boundaries, and relative depth hierarchies to handle occlusions and depth discontinuities. A naive approach might be to use off-the-shelf instance object detectors, but these models also suffer from occlusions. When objects overlap, they often fail to segment occluded instances, miss detections, or lose object identities across frames. Since these detectors struggle with the same occlusion challenges as depth estimation, addressing these issues requires advancements in both training data and methodology.

From a dataset perspective, we introduce the *Group Instance Depth (GID) dataset*, the first large-scale video depth dataset explicitly designed for dynamic, multi-object scenes. GID comprises over 101,500 frames covering diverse activities, including sports, dance, and animal interactions. Each frame is annotated with instance-wise depth, bounding boxes, segmentation masks, and consistent object identities. Unlike existing datasets that rely on optical flow or stereo matching for depth annotation [5, 47], which often introduce inconsistencies in dynamic settings, GID provides high-fidelity depth annotations captured using depth sensors, ensuring greater accuracy in real-world scenarios.

Building on this dataset, we propose a novel depth estimation framework, *InstanceDepth*, that explicitly incorporates instance-wise depth information to improve depth accuracy in occluded and multi-object scenes. Our method follows a two-stage training paradigm: (1) a *Holistic Depth Initialization*, which estimates coarse depth layers to capture the overall scene structure, and (2) an *Instance-Aware Depth Rectification*, which corrects depth inconsistencies by incorporating instance masks, object shapes, and relative spatial relationships between overlapping instances. Given that depth discontinuities often arise from occlusions and missing contextual information, we refine the initial depth predictions by ensuring that each instance's depth better aligns with its intrinsic shape characteristics, category constraints, and spatial positioning relative to other occluding or occluded objects. This refinement enforces geometric consistency across instances, mitigating depth ambiguities in cluttered scenes. By explicitly integrating instance-level depth cues, our approach not only preserves the global scene structure but also enhances fine-grained depth estimation at the object level (see Figure 1d & 1e).

In summary, our main contributions are as follows:

- We introduce the Group Instance Depth (GID) dataset, the first large-scale video depth dataset for multi-object dynamic scenes, comprising 101,500 frames with instance-wise depth, segmentation, and tracking annotations.
- We propose *InstanceDepth*, the first occlusion-aware depth estimation framework designed for complex multi-object environments.
- We develop a two-stage training paradigm: (1) *Holistic Depth Initialization*, which establishes a coarse depth structure for the scene, and (2) *Instance-Aware Depth Rectification*, which refines instance-wise depth using object masks, shape priors, and spatial relationships between overlapping instances.
- Our method achieves state-of-the-art performance on our GID dataset and multiple depth estimation benchmarks, demonstrating superior occlusion handling and robust depth consistency in dynamic scenes.

2. Related Work

Image Monocular Depth Estimation. Monocular depth estimation infers depth from a single image and is primarily approached through supervised or self-supervised learning. Supervised methods [10, 15, 18, 32, 33, 41] rely on ground truth depth, but acquiring such data is costly and labor-intensive. DPT [42] introduced transformer-based architectures to enhance feature extraction, while synthetic datasets [36] and structure-from-motion techniques [30, 31] have been explored to mitigate data limitations.

Recent foundation models, such as Depth Anything [52] and Depth Anything V2 [53], leverage large-scale pseudo-labeling pipelines to improve generalization. Depth Anything V2 further replaces real labels with synthetic data and scales up model capacity for finer depth predictions. Similarly, self-supervised methods [20–22, 29] estimate depth without ground truth, relying on stereo pairs [20] or occlusion-aware losses [21]. RA-Depth [24] dynamically adapts resolution, while Shi *et al*. [43] address reflective surfaces by distilling multi-view 3D information.

While foundation models [52, 53] demonstrate strong generalization, they struggle with occlusions and instance-wise depth consistency. Our work follows the supervised paradigm and introduces a video depth dataset tailored for dynamic activity scenes, complementing these models by improving depth estimation in complex environments with object interactions and occlusions.

Video Monocular Depth Estimation. Video depth estimation extends monocular depth learning by enforcing temporal consistency across frames. One approach refines single-frame predictions at inference using geometric constraints [35], pose optimization [28, 59], motion adaptation [45, 57], or embedded correlation [8, 58], but these methods require test-time training, making them computationally expensive. Alternatively, training-based approaches integrate spatial-temporal information to enable inference without finetuning. ST-CLSTM [56] employs an LSTM with GAN supervision, while Feng *et al.* [17] introduce an occlusion-aware cost volume by separating object motion. NVDS [47] ensures temporal consistency via cross-attention, with NVDS+ [48] stabilizing predictions through a plug-and-play module. DepthCrafter [25] generates long-term coherent depth sequences via stitching-based inference.

To enhance scalability, Depth Any Video [51] introduces synthetic training pipelines and generative diffusion priors, while Video Depth Anything [9] extends Depth Anything V2 for long-video inference. Our work diverges by addressing long-term video depth estimation in dynamic activity scenes with multiple objects, ensuring robustness against occlusions and prolonged interactions.

Video Depth Dataset. The progress of video depth estimation is constrained by the lack of diverse datasets. KITTI [19] provides high-quality ground truth for driving scenes but lacks variety. NYU Depth [44] and ScanNet [14] focus on indoor environments, while synthetic datasets like Sintel [5] and Tartanair [46] expand dataset diversity. More recently, Wang *et al.* [47] introduced a large-scale dataset for in-the-wild depth estimation.

Despite these advancements, most datasets emphasize static scenes with minimal occlusions and lack instance semantics. To bridge this gap, we introduce a new video depth dataset enriched with dynamic human activities across indoor and outdoor settings, incorporating instance labels to better model real-world depth estimation challenges.

3. GID Dataset

Existing video depth datasets primarily rely on synthetic depth for multi-object scenes [14, 26] or animated content [5, 54], lacking real-world fast-moving and deformable objects. To bridge this gap, we introduce the Group Instance Depth (GID) dataset, capturing real-world activity scenes with commercial depth cameras. GID is recorded using two high-quality depth sensors, the Intel RealSense D455 and Microsoft Azure Kinect DK, which provide synchronized RGB and depth images, ensuring precise depth representation in dynamic human and animal activities.

Image and Depth Capturing. To capture diverse real-world activities, we recorded scenes in sports courts (basketball, badminton, volleyball, dance, table tennis) and a zoo environment. Data collection occurred both day and night to include varied lighting conditions. Cameras were intentionally moved during recording to create dynamic scenes featuring primary subjects such as humans, basketballs, and rackets.

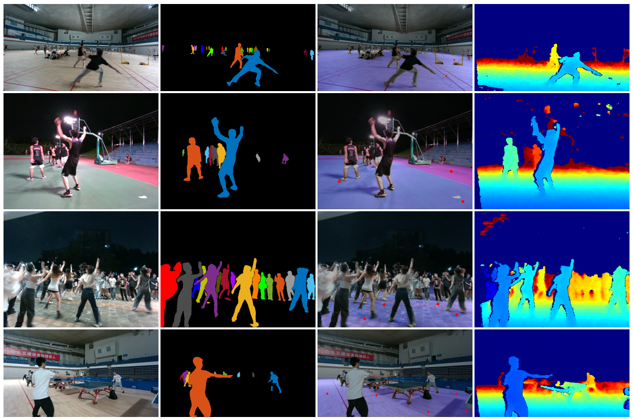


Figure 2. Illustrative mask and depth annotations for various activities, including badminton, basketball, dance, and table tennis. The “Object Mask” images distinguish objects with unique colors, while the “Ground Mask” images highlight prompt dots with red circles. The “Depth” images visualize depth values within a range of 0.01 to 10.0 meters.

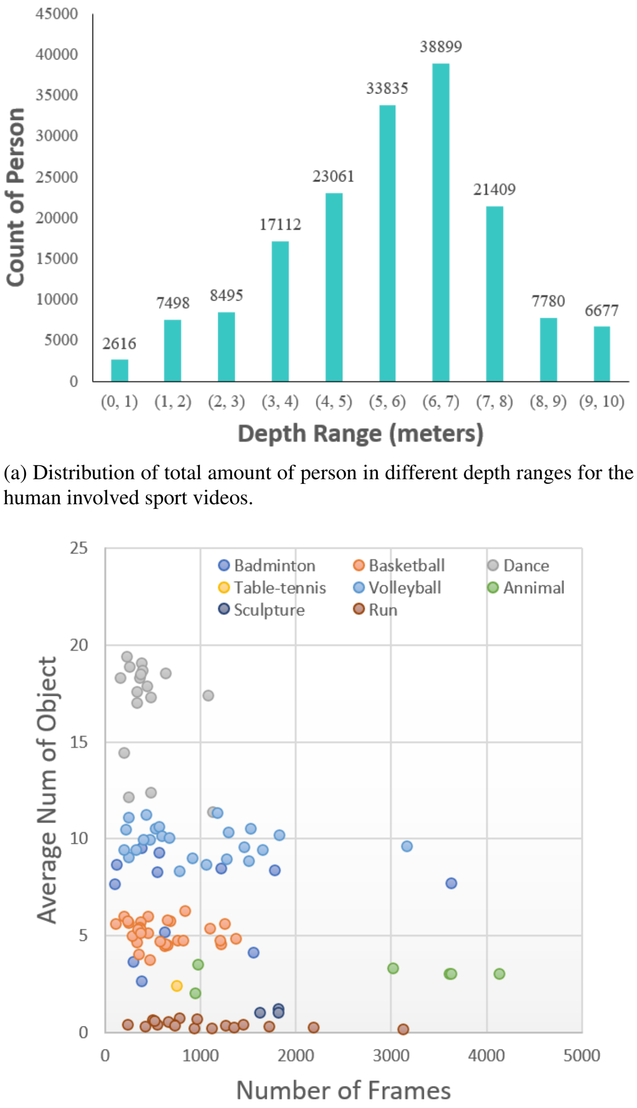
Table 1. Comparison of video depth datasets by video length and object count. Datasets using optical flow for depth annotation are listed in the second section, while those using depth cameras are in the last. For KITTI, only deformable moving objects (e.g., persons) are counted.

Dataset	# Videos	Avg Video Len	Avg # Objects
Sintel [5]	20	50	2.3
KITTI [19]	151	311.2	1.2
NYUDV2 [44]	284	128.2	0
GID (Ours)	112	625.9	8.7

Mask Generation. Mask annotations for GID were generated in two stages: prompt creation and mask extraction using SAM [27]. Bounding boxes outlined objects, while dots marked ground areas to enhance annotation accuracy. These prompts were processed through SAM to produce distinct object and ground masks. Figure 2 showcases sample annotations, capturing intricate details like human limbs, which are often underrepresented in optical flow-based datasets.

We compared GID with existing datasets, including Sintel, KITTI, and NYU Depth [5, 19, 44], in terms of video length and moving object diversity (see Table 1). GID offers longer sequences and a broader range of deformable objects, highlighting its unique contribution to real-world depth estimation research.

Tracking Identity Generation. Tracking identity is essential for consistent video depth estimation. While detection-based tracking methods [7, 49] are viable, we adopt segmentation-based tracking [13] due to its pre-training on large datasets like SAM [27] and open-world datasets [1], ensuring more stable tracking over long sequences. We use



(a) Distribution of total amount of person in different depth ranges for the human involved sport videos.

(b) Average number of objects and the total number of frames in each video. Each point is a video.

Figure 3. Statistical distributions of our proposed GID dataset.

DEVA [13], an off-the-shelf tracker, to generate initial object masks with unique identities. These masks are then matched to prior masks using the highest Intersection over Union (IoU) to maintain identity consistency. Figure 2 illustrates different identities in varying colors.

Statistical Distribution Analysis. We analyze the dataset's statistical distribution, focusing on object counts across depth ranges (Figure 3a). Most objects fall within 4.0–8.0 meters, with fewer at shallow ($< 2m$) or deep ($> 8m$) ranges. Additionally, we examine object counts and frame distributions across video categories (Figure 3b). The Dance category contains the most objects, often exceeding 10 per video, while other categories generally have fewer. Most videos range between 0–1,000 frames, though some exceed this limit. The Table Tennis category has fewer samples due to the difficulty of capturing complete scenes with depth cameras. For the dataset split, we randomly selected 21 videos as the test set, comprising 20,223 frames that encompass motion scenarios and human-animal interaction scenarios. Unlike the train/test split ratios in other datasets [19, 44], we assign a larger proportion (20%) to the test set, which poses an even greater challenge for depth estimation in group scenarios.

4. InstanceDepth

As illustrated in Figure 4, we introduce InstanceDepth, a novel depth estimation framework designed to improve depth accuracy in multi-object, occlusion-heavy environments. It follows a two-stage process: Holistic Depth Initialization, which establishes a coarse scene-level depth structure, and Instance-Aware Depth Rectification, which refines depth estimates by incorporating instance segmentation, shape priors, and spatial relationships to enforce geometric consistency.

4.1. Holistic Depth Initialization

Holistic Depth Initialization serves as the first stage of our framework, providing a structured depth estimation that captures the overall scene layout. It establishes depth priors by segmenting depth space and refining relative depth estimates, laying the foundation for precise instance-level depth refinement.

To incorporate object-level information in depth estimation, we first apply depth range segmentation to assign depth references to pixels. Prior methods rely on predefined depth ranges per image or pixel, but complex scenes with occlusions require a more adaptive approach. Objects at different depths within overlapping regions should be treated separately.

As shown in Figure 4, we design a multi-scale depth decoder (detailed in Figure 5) that progressively refines features from coarse to fine depth scales, generating depth range features. Based on these features, we first obtain an initial depth range segmentation S and depth estimation D using a lightweight convolutional network. The depth is then refined iteratively at multiple segmentation levels.

Let F_i represent the depth range features at the i-th level, with S_i and D_i denoting the corresponding depth range segmentation and initial depth estimation. The maximum depth value, MAX_d, is divided into rd depth ranges. We evaluate depth range confidence using a small network Φ_d, which takes the depth range features F_i and depth D_i at level i:

$$C_i = \text{Sigmoid}(\Phi_d(F_i, D_i)) \tag{1}$$

The pixel-wise relative error score R_i is then computed by summing the segmentation S_i and confidence C_i along the

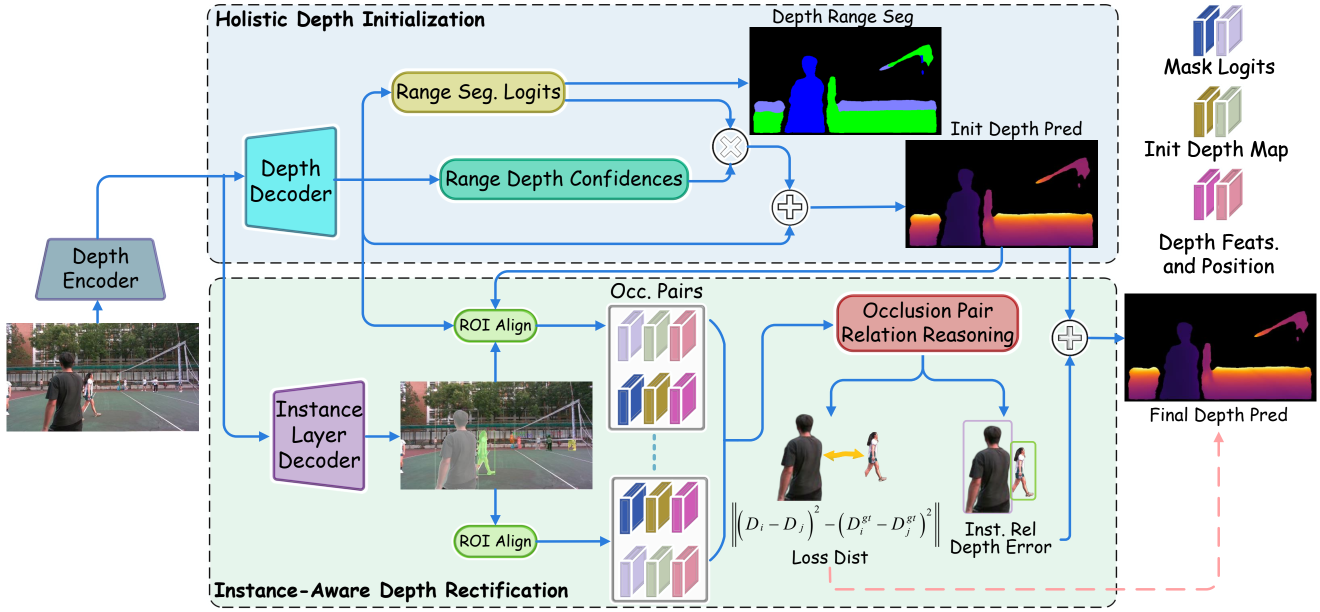


Figure 4. Overview of the proposed method, consisting of three key components. The upper section illustrates the depth estimation training process, incorporating cascaded global depth range blocks and relative depth error prediction. The lower left depicts the layered instance segmentation, which generates instance depth layers and segmentation masks. The lower right presents the instance 3D relation reasoning stage, where instance-wise relative depths are predicted to refine the initial depth estimation.

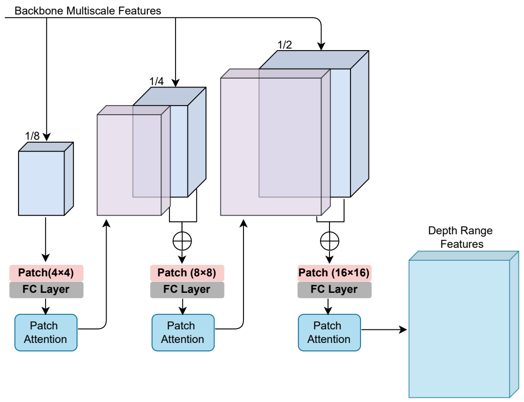


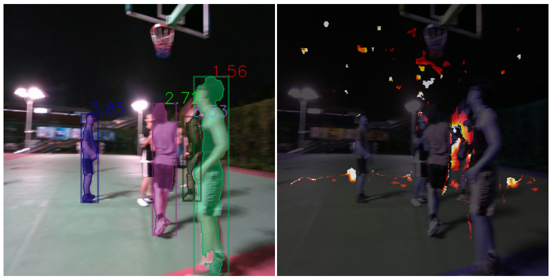
Figure 5. Depth range feature decoder pipeline. It extracts multi-scale features using patch-wise convolution, a linear layer, and patch attention for improved scene geometry understanding.

depth range dimension:

$$R_i = \sum_{i=0}^{rd} (C_i \cdot S_i) \tag{2}$$

Relative error scores are adjusted for each depth range to improve estimation precision. The relative depth error E_i and refined depth at level $(i+1)$, denoted as $D_{(i+1)}$, are computed as:

$$E_i = 2 \cdot (R_i - 1) \cdot \left(\frac{MAX_d}{rd} \right). \tag{3}$$



Instance Depth Layer

Depth Error

Figure 6. Illustrative mask and depth error visualization. Depth errors frequently occur at occlusions between instances, highlighting the need to model instance relationships for accurate depth estimation in dynamic multi-object environments.

$$D_{(i+1)} = D_i + E_i \tag{4}$$

This process is repeated across all depth levels to generate the initial holistic depth, establishing a coarse scene-level depth structure.

4.2. Instance-Aware Depth Rectification

With the initialized holistic depth, we refine instance-wise depth predictions by integrating instance segmentation, depth layer priors, and geometric consistency to address occlusions (see Figure 6). This stage consists of two key components: instance depth layer prediction and occlusion-aware depth refinement, which collectively enhance depth accuracy in multi-object interactions.

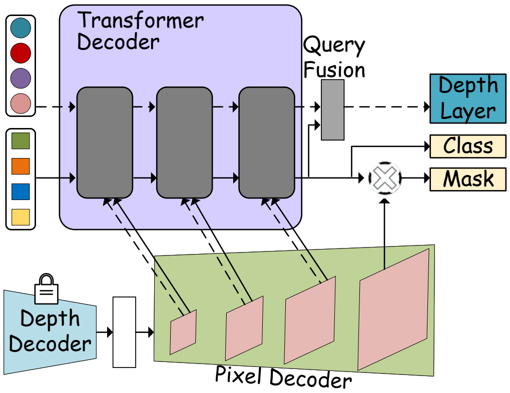


Figure 7. Instance depth layer decoder architecture. Task-specific query embeddings (circles and squares) and pixel decoder parameters (dashed and solid lines) are used for instance depth layer prediction and segmentation. A transformer-decoder-followed query fusion module generates instance-aware depth layers per query.

4.2.1. Instance Depth Layer Prediction

We employ a Mask2Former-based [6, 11, 12] instance segmentation decoder to predict instance masks and their corresponding depth layers, as illustrated in Figure 7. The decoder produces a fixed set of N predictions, each comprising an instance mask Msk^i, category Cls^i, and instance depth layer Dep^i, representing the average depth of the instance. To align predictions with ground truth, we formulate a bipartite matching cost that incorporates depth layer differences, mask IoU, and category classification:

$$\min_{\sigma} \sum_{i=1}^{N} \lambda_1 L_m(\widehat{Msk}^{(i)}, Msk^{\sigma(i)}) + \tag{5}$$

$$\lambda_2 L_c(\widehat{Cls}^{(i)}, Cls^{\sigma(i)}) + \tag{6}$$

$$\lambda_3 L_d(\widehat{Dep}^{(i)}, Dep^{\sigma(i)}), \tag{7}$$

where L_m (mask IoU loss) and L_c (cross-entropy loss) follow Mask2Former [12], while L_d applies a smoothed $L1$ loss for depth layer regression. $\sigma(i)$ denotes the index of the corresponding ground truth object. This formulation jointly optimizes instance segmentation and depth layers while preserving object identities and spatial relationships.

4.2.2. Occlusion-Aware Depth Refinement

To resolve depth ambiguities in occluded regions, we enforce geometric consistency using predicted instance masks and depth layers, as shown in the Occlusion Pair Relation Reasoning module in Figure 4. First, candidate instances are filtered based on category confidence (> 0.9) and mask confidence (> 0.8). For each main instance, overlapping instances ($IoU > 0.1$) are identified, and the nearest in depth (termed the *guest instance*) is retained.

Multi-scale depth features $F^i_{obj} \in R^{2\times C\times H_p\times W_p}$ and geometric priors G_{obj} (including mask logits, normalized coordinates, and global depths) are extracted via ROI alignment [23]. These inputs are processed by an MLP Φ_o to predict relative depth errors E_{obj}:

$$E_{obj} = Sigmoid(\Phi_o([F_{obj}, G_{obj}])) \tag{8}$$

which refines instance depths as:

$$\hat{D}_{obj} = [(E_{obj} \times 2) - 1] \times \overline{D}_{obj} + D_{obj} \tag{9}$$

To supervise the refined depths, we apply a scale-invariant logarithmic loss [16]:

$$L_{obj} = SigLog(\hat{D}_{obj}, DT_{obj}) \tag{10}$$

Additionally, relative depth consistency between occluder-occludee pairs is enforced via L_{dist}:

$$L_{dist} = \sum_{i=1,j\neq i}^{M} ||(\hat{D}_i - \hat{D}_j)^2 - (DT_i - DT_j)^2|| \tag{11}$$

The total refinement loss combines both terms:

$$L_{ref} = \lambda_1 * L_{obj} + \lambda_2 * L_{dist} \tag{12}$$

This dual-loss strategy ensures accurate depth recovery for occluded objects while maintaining coherent spatial hierarchies across instances.

By integrating instance-wise depth information with occlusion-aware geometric reasoning, our approach mitigates depth ambiguities in crowded scenes, achieving robust depth consistency even under severe occlusions.

4.3. Training Details

Our method is implemented using the PyTorch framework [38] and trained on an NVIDIA RTX 4090 GPU. We employ the pretrained DINOv2 [37] as the backbone network for feature extraction.

The training process consists of three progressive phases to ensure effective depth learning at both the scene and instance levels:

- **Global Depth Range Pretraining**: In the first phase, we train the depth range module for 55k iterations with an initial learning rate of $\mathbf{1\times10^{-5}}$. This stage establishes scene-wide depth priors by learning global depth distributions across varying scales.
- **Instance Depth Layer Specialization**: After pretraining the global depth range, we freeze the depth encoder and train the instance decoder for 25k iterations (LR=$\mathbf{1\times10^{-5}}$). This phase focuses on learning object-level depth segmentation, enabling the model to distinguish depth layers for individual instances.

Table 2. Depth estimation results on the GID test set. All models are trained on the GID training set for fair comparisons.

Method	RMS ↓	REL ↓	$\text{RMS}_{\log}$ ↓	Log_{10} ↓	σ_1 ↑
CDMOV [57]	0.657	0.078	0.128	0.182	0.768
NeWCRFs [55]	0.497	0.067	0.098	0.029	0.938
ZoeDepth [4]	0.487	0.068	0.107	0.034	0.949
NVDS [48]	0.511	0.062	0.116	0.037	0.942
DepthAnything V1 [52]	0.433	0.052	0.086	0.024	0.973
DepthAnything V2 [53]	0.435	0.053	0.088	0.026	0.971
PromDep [34]	0.431	0.051	0.084	0.024	0.974
InstanceDepth	**0.397**	**0.045**	**0.077**	**0.019**	**0.983**

Table 3. Depth estimation comparison results on the NYU Depth V2 test set. All methods are trained on NYU Depth V2.

Method	REL ↓	RMS ↓	σ_1 ↑	σ_2 ↑
AdaBins [3]	0.178	0.595	0.698	0.937
TransDepth [50]	0.106	0.365	0.900	0.983
P3Depth [39]	0.104	0.356	0.898	0.981
ZoeDepth [4]	0.077	0.282	0.951	0.994
UniDepth-V [40]	0.058	0.201	0.886	0.984
DepthAnything V1 [52]	0.056	0.206	0.984	**1**
DepthAnything V2 [53]	**0.056**	0.206	0.984	0.998
InstanceDepth	**0.056**	**0.202**	**0.985**	0.999

- **Occlusion-Aware Joint Refinement**: In the final stage, we reverse the freezing strategy—fixing the instance decoder while fine-tuning both the depth encoder and decoder for 25k iterations with a lower learning rate of $\mathbf{1\times10^{-6}}$. This phase enhances depth consistency in occluded regions by refining depth estimates based on spatial relationships between overlapping objects.

5. Experiments

5.1. Datasets and Evaluation Metrics

Since our focus is on depth estimation in dynamic, object-centered activities scenes, we evaluate our method on additional static dataset alongside our GID dataset: NYUDv2 [44], a depth camera-captured dataset featuring indoor environments.

For performance evaluation, we use standard depth estimation metrics, including root mean squared error (RMS), average relative error (REL), and accuracy within thresholds $\sigma_i (i = 1, 2, 3)$, which measure depth estimation precision and robustness across different environments.

5.2. Comparisons with State-of-the-Art Methods

To comprehensively evaluate our approach, we compare InstanceDepth with leading depth estimation methods [4, 34, 48, 52, 53, 55]. Our evaluation covers both dynamic multi-object scenarios and static indoor environments to assess the generalizability of our method.

We first conduct experiments in dynamic multi-object settings using the GID test set. Table 2 presents quantitative comparisons, where InstanceDepth outperforms existing state-of-the-art methods, particularly in occlusion-heavy scenes where depth discontinuities pose significant challenges. Notably, our method surpasses large-scale foundation models, such as Depth Anything V2 [53] and PromDep [34], demonstrating its ability to maintain fine-grained instance-wise depth consistency even in complex object interactions.

Beyond dynamic scenarios, we evaluate InstanceDepth on the NYUDv2 dataset [44] to test its effectiveness in structured indoor environments. For this experiment, we train the Holistic Depth Initialization stage on NYUDv2 using the pretrained encoder from Depth Anything V2 [53]. The results in Table 3 show that our depth range relative error strategy remains effective even in static indoor scenes. However, the performance gains are marginal compared to dynamic environments, likely due to the structured depth layering present in indoor scenes (e.g., walls as persistent background planes), where explicit depth range segmentation is less critical than in cluttered outdoor settings.

Furthermore, we provide qualitative comparisons in Figure 8. As shown, InstanceDepth produces more accurate depth maps than existing methods, particularly in occluded regions where competing models struggle with depth inconsistencies. Our method effectively captures inter-instance relationships and preserves geometric consistency in dynamic, multi-object scenes, reinforcing its robustness in challenging real-world scenarios.

5.3. Ablation Studies

To assess the effectiveness of the proposed two-stage training scheme, we first establish a baseline model using a DepthAnything V2 pretrained encoder combined with a DPT decoder [41]. We then incrementally incorporate the two training stages onto this baseline. Table 4 presents the performance comparisons among these variants. By integrating holistic depth initialization, the model captures the coarse scene structure more effectively, leading to improved performance over the baseline.

5.3.1. Ablation on Holistic Depth Initialization

The depth range partitioning strategy plays a crucial role in the holistic depth initialization stage. In this ablation study, we evaluate different partitioning strategies, ranging from fine-grained to coarse-grained depth segmentation. The results in Table 5 indicate that both overly fine-grained and excessively coarse partitioning negatively impact performance. We hypothesize that fine-grained partitioning introduces fragmented depth regions, increasing training complexity and slowing convergence, whereas overly large depth intervals fail to provide sufficient constraints for opti-

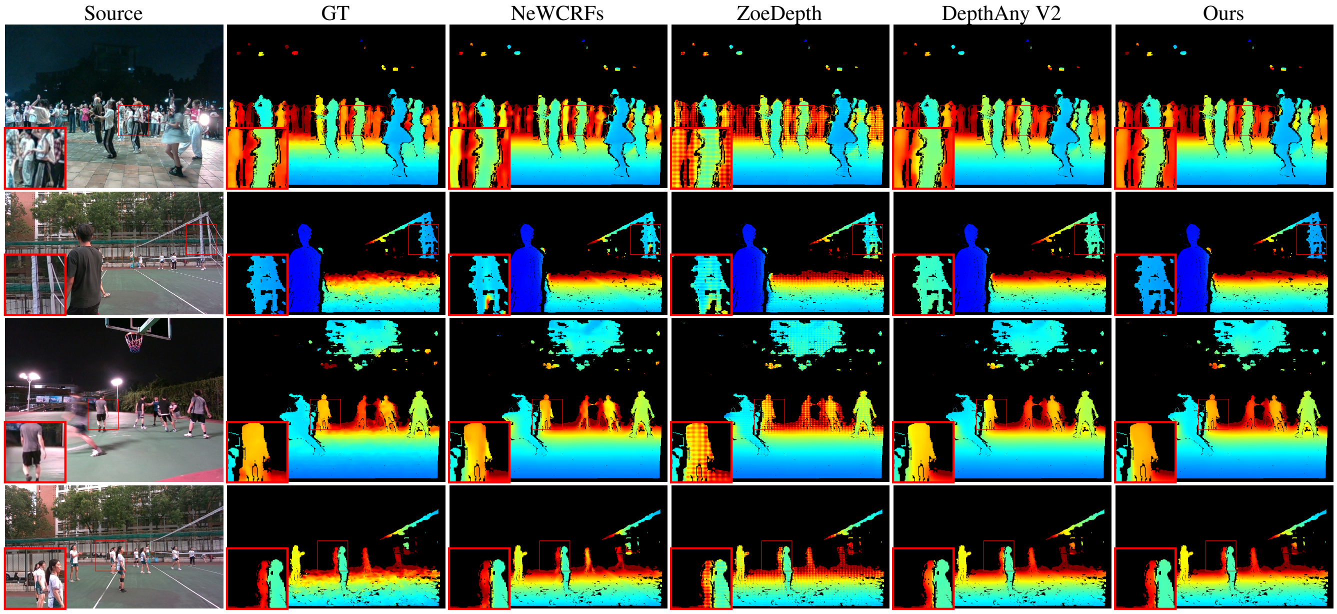


Figure 8. Qualitative comparison of depth estimation results on GID test set. Our method demonstrates superior depth preservation in occluded regions and improved geometric coherence in multi-object scenes.

Table 4. Ablation study on the proposed two-stage framework. The baseline model uses a Depth Anything V2 pretrained encoder with a DPT decoder. "H" and "I" denote the Holistic Depth Initialization and Instance-Aware Depth Rectification stages, respectively.

Method	REL ↓	RMS ↓	$\sigma_1 \uparrow$	$\sigma_2 \uparrow$	$\sigma_3 \uparrow$
Baseline	0.0524	0.4357	0.9757	0.9933	0.9972
Baseline+H	0.0506	0.4188	0.9784	0.9947	0.9975
Baseline+H+I	**0.045**	**0.397**	**0.983**	**0.995**	**0.998**

mizing depth range relative errors.

5.3.2. Ablation on Instance-Aware Depth Rectification

The loss functions in the instance-aware depth rectification stage play distinct roles in optimizing depth accuracy. To assess their individual contributions, we conducted ablation experiments, with results shown in Table 6. The findings indicate that the instance depth refinement loss (L_{obj}) has the most significant impact on performance, whereas the relative distance loss has a more limited effect. We hypothesize that this is because L_{obj} broadly optimizes depth values in occluded regions, whereas the relative distance loss primarily influences cases with large depth discrepancies between the main and guest objects—a scenario that occurs less frequently in typical scenes.

Table 5. Ablation study on the depth range partitioning strategy. The strategy of partitioning ground truth depth into intervals of K meters is denoted as "K meter". This ablation is conducted without the Instance-Aware Depth Rectification stage.

Partitioning By	REL ↓	RMS ↓	$\sigma_1 \uparrow$	$\sigma_2 \uparrow$	$\sigma_3 \uparrow$
1 meter	0.053	0.426	0.977	0.993	0.997
2 meter	**0.051**	**0.419**	**0.978**	**0.995**	**0.998**
3 meter	0.055	0.431	0.972	0.989	0.991

Table 6. Ablation study on the loss functions used in the instance-aware depth rectification stage. All evaluated methods are based on the holistic depth initialization stage.

Method	REL ↓	RMS ↓	$\sigma_1 \uparrow$	$\sigma_2 \uparrow$	$\sigma_3 \uparrow$
L_{obj}	0.0524	0.407	0.974	0.989	0.991
L_{dist}	0.0506	0.412	0.971	0.988	0.991
$L_{obj} + L_{dist}$	**0.045**	**0.397**	**0.983**	**0.995**	**0.998**

6. Conclusion

We introduced GID, a novel dataset for depth estimation in dynamic, real-world group scenes. GID captures diverse human activities with detailed depth annotations, instance masks, and tracking identities, addressing the challenges of multi-object interactions and occlusions. Additionally, we proposed InstanceDepth, an occlusion-aware depth estimation framework that integrates holistic depth initialization and instance-aware depth rectification, significantly improving depth accuracy in complex environments.

Acknowledgements. This work is supported by the Guangdong Natural Science Funds for Distinguished Young Scholars (Grant 2023B1515020097), the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No.: AISG3-GV-2023-011), the Singapore Ministry of Education AcRF Tier 1 Grant (Grant No.: MSS25C004), and the Lee Kong Chian Fellowships.

References

[1] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1674–1683, 2023. 3

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, pages 4009–4018, 2021. 2

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, pages 4009–4018, 2021. 7

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1, 2, 7

[5] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 2, 3

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 6

[7] Orcun Cetintas, Guillem Brasó, and Laura Leal-Taixé. Unifying short and long-term tracking with graph hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22877–22887, 2023. 3

[8] Haoxin Chen, Hanjie Wu, Nanxuan Zhao, Sucheng Ren, and Shengfeng He. Delving deep into many-to-many attention for few-shot video object segmentation. In *CVPR*, pages 14040–14049, 2021. 3

[9] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv:2501.12375*, 2025. 3

[10] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *NeurIPS*, 29, 2016. 2

[11] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, pages 17864–17875, 2021. 6

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 6

[13] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, pages 1316–1326, 2023. 3, 4

[14] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 1, 3

[15] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 2

[16] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014. 1, 6

[17] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In *ECCV*, pages 228–244. Springer, 2022. 3

[18] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018. 2

[19] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 3, 4

[20] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017. 2

[21] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *CVPR*, pages 3828–3838, 2019. 2

[22] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, pages 2485–2494, 2020. 2

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 6

[24] Mu He, Le Hui, Yikai Bian, Jian Ren, Jin Xie, and Jian Yang. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *European Conference on Computer Vision*, pages 565–581. Springer, 2022. 2

[25] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *CVPR*, 2025. 3

[26] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *CVPR*, pages 13229–13239, 2023. 3

[27] 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. *arXiv:2304.02643*, 2023. 3

[28] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, pages 1611–1621, 2021. 3

[29] Hanhan Li, Ariel Gordon, Hang Zhao, Vincent Casser, and Anelia Angelova. Unsupervised monocular depth learning in dynamic scenes. In *CoRL*, pages 1908–1917. PMLR, 2021. 2

[30] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. 2

[31] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, pages 4521–4530, 2019. 2

[32] Yuan Liang, Bailin Deng, Wenxi Liu, Jing Qin, and Shengfeng He. Monocular depth estimation for glass walls with context: a new dataset and method. *IEEE TPAMI*, 45 (12):15081–15097, 2023. 2

[33] Yuan Liang, Zitian Zhang, Chuhua Xian, and Shengfeng He. Delving into multi-illumination monocular depth estimation: A new dataset and method. *IEEE TMM*, 2024. 2

[34] Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. *arXiv preprint arXiv:2412.14015*, 2024. 7

[35] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM TOG*, 39(4):71–1, 2020. 3

[36] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. 2

[37] 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. 6

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 6

[39] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *CVPR*, pages 1610–1621, 2022. 7

[40] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, pages 10106–10116, 2024. 7

[41] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3):1623–1637, 2020. 1, 2, 7

[42] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021. 2

[43] Xuepeng Shi, Georgi Dikov, Gerhard Reitmayr, Tae-Kyun Kim, and Mohsen Ghafoorian. 3d distillation: Improving self-supervised monocular depth estimation on reflective surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9133–9143, 2023. 2

[44] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012. 1, 3, 4, 7

[45] Ziming Sun, Yuan Liang, Zejun Ma, Tianle Zhang, Linchao Bao, Guiqing Li, and Shengfeng He. Repose: 3d human pose estimation via spatio-temporal depth relational consistency. In *ECCV*, pages 309–325. Springer, 2024. 3

[46] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 3

[47] Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *ICCV*, pages 9466–9476, 2023. 2, 3

[48] Yiran Wang, Min Shi, Jiaqi Li, Chaoyi Hong, Zihao Huang, Juewen Peng, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Nvds+: Towards efficient and versatile neural stabilizer for video depth estimation. *IEEE TPAMI*, 2024. 3, 7

[49] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *European Conference on Computer Vision*, pages 733–751. Springer, 2022. 3

[50] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *ICCV*, pages 16269–16279, 2021. 7

[51] Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024. 3

[52] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024. 1, 2, 7

[53] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 37:21875–21911, 2024. 1, 2, 7

[54] Ziyu Yang, Sucheng Ren, Zongwei Wu, Nanxuan Zhao, Junle Wang, Jing Qin, and Shengfeng He. Npf-200: A multi-modal eye fixation dataset and method for non-photorealistic videos. In *ACM MM*, pages 2294–2304, 2023. 3

[55] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *CVPR*, pages 3916–3925, 2022. 1, 7

[56] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *ICCV*, pages 1725–1734, 2019. 3

[57] Zhoutong Zhang, Forrester Cole, Richard Tucker, William T Freeman, and Tali Dekel. Consistent depth of moving objects in video. *ACM TOG*, 40(4):1–12, 2021. 3, 7

[58] Weiying Zheng, Cheng Xu, Xuemiao Xu, Wenxi Liu, and Shengfeng He. Ciri: curricular inactivation for residue-aware

one-shot video inpainting. In *ICCV*, pages 13012–13022, 2023. 3

[59] Yulong Zheng, Zicheng Jiang, Shengfeng He, Yandu Sun, Junyu Dong, Huaidong Zhang, and Yong Du. Nexusgs: Sparse view synthesis with epipolar depth priors in 3d gaussian splatting. In *CVPR*, pages 26800–26809, 2025. 3