# Novelty Assessment Report

**Paper**: Depth Anything 3: Recovering the Visual Space from Any Views
**PDF URL**: https://openreview.net/pdf?id=yirunib8l8
**Venue**: ICLR 2026 Conference Submission
**Year**: 2026
**Report Generated**: 2025-12-29

## Abstract

We present Depth Anything 3 (DA3), a model that predicts spatially consistent geometry from an arbitrary number of visual inputs, with or without known camera poses. In pursuit of minimal modeling, DA3 yields two key insights: a single plain transformer (e.g., vanilla DINOv2 encoder) is sufficient as a backbone without architectural specialization, and a singular depth-ray prediction target obviates the need for complex multi-task learning. Through our teacher-student training paradigm, the model achieves a level of detail and generalization on par with Depth Anything 2 (DA2). We establish a new visual geometry benchmark covering camera pose estimation, any-view geometry and visual rendering. On this benchmark, DA3 sets a new state-of-the-art across all tasks, surpassing prior SOTA VGGT by an average of 35.7\% in camera pose accuracy and 23.6\% in geometric accuracy. Moreover, it outperforms DA2 in monocular depth estimation. All models are trained exclusively on public academic datasets.



## Core Task Landscape

This paper addresses: **Predicting Spatially Consistent Geometry from Arbitrary Number of Visual Inputs**

A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

### Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Multi-View Stereo and Depth Estimation**
- **Generative and Diffusion-Based 3D Reconstruction**
- **Novel View Synthesis from Sparse Inputs**
- **Unified and Multi-Task 3D Prediction**
- **Domain-Specific Geometric Reconstruction**
- **Shape Completion and Reconstruction from Partial Observations**
- **3D Segmentation and Part-Level Understanding**
- **Spatial Reasoning and Scene Understanding**
- **Specialized Geometric Inference Tasks**

### Complete Taxonomy Tree

- Predicting Spatially Consistent Geometry from Arbitrary Number of Visual Inputs Survey Taxonomy
- Multi-View Stereo and Depth Estimation
  - Cost Volume Aggregation with Geometric Consistency (3 papers)
  - [1] Gomvs: Geometrically consistent cost aggregation for multi-view stereo (Jiang Wu, 2024) View paper
  - [12] Multi-scale geometric consistency guided multi-view stereo (Qingshan Xu, 2019) View paper
  - [21] GC-MVSNet: Multi-view, multi-scale, geometrically-consistent multi-view stereo (Sripad Joshi, 2024) View paper
  - Geometry-Aware Priors and Feature Enhancement (3 papers)
  - [3] Learning multi-view stereo with geometry-aware prior (Kehua Chen, 2025) View paper
  - [14] Deep 3d capture: Geometry and reflectance from sparse multi-view images (Bi, 2020) View paper
  - [18] Multi-View Stereo with Geometric Encoding for Dense Scene Reconstruction (Guidong Yang, 2025) View paper
  - Self-Supervised and Unsupervised Multi-Camera Depth (2 papers)
  - [13] STViT+: improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization (Zhuo Chen, 2025) View paper
  - [46] Self-Supervised Depth Completion Guided by 3D Perception and Geometry Consistency (Cai Yu, 2023) View paper
- Generative and Diffusion-Based 3D Reconstruction
  - Multi-View Diffusion with Geometric Constraints (3 papers)
  - [2] Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models (Liang Hanwen, 2024) View paper
  - [4] SPAD: Spatially Aware Multi-View Diffusers (Yash Kant, 2024) View paper
  - [16] 3D-Adapter: Geometry-Consistent Multi-View Diffusion for High-Quality 3D Generation (Chen Hansheng, 2024) View paper
  - Unified Video and 3D Generation (3 papers)
  - [5] FantasyWorld: Geometry-Consistent World Modeling via Unified Video and 3D Prediction (Jiang Fan, 2025) View paper
  - [6] 4D Driving Scene Generation With Stereo Forcing (Lu Hao, 2025) View paper
  - [25] CoGen: 3D Consistent Video Generation via Adaptive Conditioning for Autonomous Driving (Zhu, 2025) View paper
  - Amodal and Occlusion-Aware 3D Reconstruction (2 papers)
  - [8] AmodalGen3D: Generative Amodal 3D Object Reconstruction from Sparse Unposed Views (Junwei Zhou, 2025) View paper
  - [22] Multi-view 3D object reconstruction and uncertainty modelling with neural shape prior (Ziwei Liao, 2024) View paper

- Novel View Synthesis from Sparse Inputs
  - Generalizable Novel View Synthesis without 3D Priors (2 papers)
  - [9] The Less You Depend, The More You Learn: Synthesizing Novel Views from Sparse, Unposed Images without Any 3D Knowledge (Wang Haoru, 2025) View paper
  - [11] Worldmirror: Universal 3d world reconstruction with any-prior prompting (Liu Yifan, 2025) View paper
  - 3D Gaussian Splatting for Sparse Views (3 papers)
  - [10] Coherentgs: Sparse novel view synthesis with coherent 3d gaussians (Avinash Paliwal, 2024) View paper
  - [23] VA-GS: Enhancing the Geometric Representation of Gaussian Splatting via View Alignment (Li Qing, 2025) View paper
  - [49] Breaking the Vicious Cycle: Coherent 3D Gaussian Splatting from Sparse and Motion-Blurred Views (Zhankuo Xu, 2025) View paper
  - Neural Implicit Surface Representations (2 papers)
  - [20] PMVC: Promoting Multi-View Consistency for 3D Scene Reconstruction (Chushan Zhang, 2024) View paper
  - [29] Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild (Jason Zhang, 2021) View paper
- Unified and Multi-Task 3D Prediction
  - All-in-One Geometric Prediction with Flexible Priors ★ (1 papers)
  - [0] Depth Anything 3: Recovering the Visual Space from Any Views (Anon et al., 2026) View paper
  - Feed-Forward Gaussian Splatting with Semantic Fields (2 papers)
  - [17] UniForward: Unified 3D Scene and Semantic Field Reconstruction via Feed-Forward Gaussian Splatting from Only Sparse-View Images (Tian Qi-jian, 2025) View paper
  - [24] ST-GS: Vision-Based 3D Semantic Occupancy Prediction with Spatial-Temporal Gaussian Splatting (Yan, 2025) View paper
- Domain-Specific Geometric Reconstruction
  - Autonomous Driving Scene Reconstruction (2 papers)
  - [28] TheSHY-3D: Texture and Structure HarmonY for Multi-View 3D Object Detection (Wenquan Zhang, 2025) View paper
  - [48] RoboTransfer: Geometry-Consistent Video Diffusion for Robotic Visual Policy Transfer (Liu Liu, 2025) View paper
  - Indoor and Panoramic Scene Reconstruction (2 papers)
  - [27] 360 layout estimation via orthogonal planes disentanglement and multi-view geometric consistency perception (Zhijie Shen, 2024) View paper
  - [39] 360mvsnet: Deep multi-view stereo network with 360deg images for indoor scene reconstruction (CY Chiu, 2023) View paper
  - Facade and Large-Scale Scene Reconstruction (2 papers)
  - [32] Towards scalable multi-view reconstruction of geometry and materials (Carolin Schmitt, 2023) View paper
  - [40] Holistic Inverse Rendering of Complex Facade via Aerial 3D Scanning (Xie Zixuan, 2023) View paper
  - Cross-Daylight and Illumination-Invariant Matching (2 papers)
  - [7] Topology-Aware Multi-View Street Scene Image Matching for Cross-Daylight Conditions Integrating Geometric Constraints and Semantic Consistency (Haiqing He, 2025) View paper
  - [43] SatDreamer360: Geometry Consistent Street-View Video Generation from Satellite Imagery (Zhu Bei-Yi, 2025) View paper
- Shape Completion and Reconstruction from Partial Observations
  - Deep Learning-Based Shape Completion (2 papers)
  - [34] High-resolution shape completion using deep neural networks for global structure and local geometry inference (Xiaoguang Han, 2017) View paper
  - [47] 3D shape completion with multi-view consistent inference (Hu Tao, 2020) View paper
  - Multi-View Consistency for Shape and Pose Learning (2 papers)
  - [37] Multi-view supervision for single-view reconstruction via differentiable ray consistency (Shubham Tulsiani, 2017) View paper
  - [38] Multi-view consistency as supervisory signal for learning shape and pose prediction (Shubham Tulsiani, 2018) View paper
- 3D Segmentation and Part-Level Understanding (3 papers)
  - [15] 3D Part Segmentation via Geometric Aggregation of 2D Visual Features (Marco Garosi, 2024) View paper
  - [30] 3D shape segmentation with projective convolutional networks (Evangelos Kalogerakis, 2017) View paper
  - [35] Decomposing a scene into geometric and semantically consistent regions (Stephen Gould, 2009) View paper
- Spatial Reasoning and Scene Understanding (1 papers)
  - [26] Learning spatial common sense with geometry-aware recurrent networks (Tung, 2019) View paper
- Specialized Geometric Inference Tasks
  - Illumination and Reflectance Estimation (2 papers)
  - [41] Lighthouse: Predicting lighting volumes for spatially-coherent illumination (Pratul P. Srinivasan, 2020) View paper
  - [45] SREGS: Sparse-view Gaussian radiance fields with geometric regularization and region exploration. (Xiaotong Li, 2025) View paper
  - Unsupervised Pose Estimation with Geometric Consistency (2 papers)
  - [31] Geometric Consistency-Guaranteed Spatio-Temporal Transformer for Unsupervised Multi-View 3D Pose Estimation (Kaiwen Dong, 2024) View paper
  - [33] Topologically consistent multi-view face inference using volumetric sampling (Li Tianye, 2021) View paper
  - Statistical and Probabilistic Shape Modeling (2 papers)
  - [19] Unsupervised discovery of the shared and private geometry in multi-view data (Sai Koukuntla, 2024) View paper
  - [36] Inferring 3d structure with a statistical image-based shape model (K. Grauman, 2003) View paper
  - Non-Linear Embedding and Dimensionality Reduction (2 papers)
  - [42] TomoGraphView: 3D Medical Image Classification with Omnidirectional Slice Representations and Graph Neural Networks (Johannes Kiechle, 2025) View paper
  - [50] Generative modeling for continuous non-linearly embedded visual inference (Cristian Sminchisescu, 2004) View paper
  - Object Recognition with Accurate Pose (1 papers)
  - [44] What and where: 3D object recognition with accurate pose (Iryna Gordon, 2006) View paper

## Narrative

Core task: predicting spatially consistent geometry from arbitrary number of visual inputs. The field encompasses a diverse set of approaches organized around how methods handle input flexibility and geometric reasoning. Multi-View Stereo and Depth Estimation focuses on classical correspondence-based techniques that aggregate information across calibrated views, while Generative and Diffusion-Based 3D Reconstruction leverages learned priors to synthesize plausible geometry even from limited observations. Novel View Synthesis from Sparse Inputs emphasizes rendering quality and view interpolation, often trading off geometric accuracy for visual

coherence. Unified and Multi-Task 3D Prediction aims to build flexible architectures that can handle varying numbers of inputs and produce multiple geometric outputs simultaneously, as seen in works like Gomvs[1] and Geometry Aware Prior[3]. Domain-Specific Geometric Reconstruction targets specialized settings such as indoor scenes or human bodies, while Shape Completion and Reconstruction from Partial Observations addresses the challenge of inferring occluded or missing structure. Additional branches cover 3D Segmentation and Part-Level Understanding, Spatial Reasoning and Scene Understanding, and Specialized Geometric Inference Tasks that tackle niche problems like topology-aware matching or amodal completion.

Recent activity highlights tensions between generalization and specialization. Many studies explore how to inject geometric priors into diffusion models (e.g., Diffusion4d[2], FantasyWorld[5]) or how to enforce multi-view consistency in generative pipelines (CoherentGS[10], WorldMirror[11]). Within the Unified and Multi-Task branch, Depth Anything[0] sits alongside methods that emphasize all-in-one geometric prediction with flexible priors, aiming to handle arbitrary input counts without retraining for specific configurations. Compared to Geometry Aware Prior[3], which explicitly incorporates geometric constraints into generative processes, Depth Anything[0] focuses on robust depth estimation that generalizes across diverse visual conditions. Meanwhile, works like SPAD[4] and Stereo Forcing[6] explore how to refine consistency through iterative or adversarial mechanisms. The central open question remains how to balance the expressiveness of learned priors with the reliability of geometric constraints, especially when input views are sparse or uncalibrated.

## Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

### Taxonomy-Level Summary

Both subtopics address unified 3D reconstruction from multiple images, but target different output representations and input assumptions. The original leaf focuses on flexible geometric priors (depth, normals) to produce multiple traditional 3D outputs, while the sibling emphasizes feed-forward prediction of 3D Gaussians with integrated semantic information from uncalibrated images.

**Similarities:** - Both perform unified 3D geometric prediction from visual inputs - Both aim to handle multiple views or arbitrary numbers of images - Both produce spatially consistent 3D representations in a single forward pass

**Differences:** - Original leaf accepts diverse geometric priors (depth, normals, point clouds) as flexible inputs; sibling works with uncalibrated RGB images without requiring geometric priors - Original leaf outputs multiple traditional 3D representations (depth maps, normals, point clouds); sibling outputs 3D Gaussian splats with semantic features - Original leaf excludes fixed input modalities; sibling specifically excludes calibrated inputs and methods without semantic integration - Sibling explicitly integrates semantic fields for joint scene and semantic reconstruction; original leaf focuses on geometric outputs without semantic information

**Suggested Search Directions:** - Hybrid methods that combine flexible geometric priors with semantic feature prediction - Gaussian splatting approaches that can accept optional depth or normal priors - Unified frameworks outputting both traditional geometry (depth/normals) and modern representations (Gaussians) with semantics

### Sibling Subtopics

- **Feed-Forward Gaussian Splatting with Semantic Fields** (leaves: 1, papers: 2)
- Scope: Feed-forward models that predict 3D Gaussians with semantic features from uncalibrated images for unified scene and semantic reconstruction.
- Exclude: Excludes methods without semantic integration or requiring calibrated inputs; see other unified categories.

## Contributions Analysis

**Overall novelty summary.** Depth Anything 3 proposes a unified depth-ray prediction framework using a plain transformer backbone to handle arbitrary numbers of visual inputs with or without known camera poses. The paper resides in the 'All-in-One Geometric Prediction with Flexible Priors' leaf under 'Unified and Multi-Task 3D Prediction'. Notably, this leaf contains only one sibling paper in the taxonomy (the original paper itself), suggesting this specific research direction—combining flexible input handling with minimal architectural specialization—is relatively sparse within the broader field of spatially consistent geometry prediction.

The taxonomy reveals that neighboring research directions are more densely populated. The sibling leaf 'Feed-Forward Gaussian Splatting with Semantic Fields' contains two papers exploring semantic integration with 3D Gaussians. Adjacent branches include 'Multi-View Stereo and Depth Estimation' (nine papers across three sub-categories) and 'Generative and Diffusion-Based 3D Reconstruction' (seven papers). While these areas emphasize explicit multi-view correspondence or generative priors, DA3 diverges by pursuing architectural minimalism and a single prediction target, positioning itself at the intersection of multi-task flexibility and geometric consistency without specialized modules.

Among the 28 candidates examined through semantic search, the teacher-student training paradigm shows one refutable candidate from 10 examined, indicating some prior work in distillation-based approaches for geometric tasks. The core architectural contribution (minimal transformer with depth-ray prediction) found no clear refutations across 10 candidates, suggesting relative novelty in this specific design choice. The visual geometry benchmark contribution also appears distinct, with zero refutations among eight examined candidates. These statistics reflect a limited search scope rather than exhaustive coverage, but suggest the architectural simplification and unified prediction target represent less-explored directions within the examined literature.

Based on the top-28 semantic matches and taxonomy structure, DA3 appears to occupy a sparsely populated niche combining input flexibility with architectural minimalism. The analysis does not cover the full breadth of monocular depth estimation or multi-view stereo literature, focusing instead on methods addressing arbitrary-input geometry prediction. The teacher-student paradigm shows some overlap with existing distillation approaches, while the core architectural choices and benchmark design appear more distinctive within the examined scope.

---

This paper presents **3 main contributions**, each analyzed against relevant prior work:

### Contribution 1: Depth Anything 3 model with minimal architecture and unified depth-ray prediction

**Description**: The authors introduce Depth Anything 3, a unified model that recovers 3D geometry from any number of images using a single plain transformer backbone without architectural modifications. The model employs a minimal depth-ray representation as the sole prediction target, avoiding complex multi-task learning frameworks used in prior work.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

#### 1. MonoDETR: Depth-guided transformer for monocular 3D object detection

**URL**: View paper

**Brief Assessment**

MonoDETR[61] focuses on monocular 3D object detection using depth-guided transformers for detecting objects in driving scenes, not on unified depth prediction from arbitrary numbers of images. The technical approaches and problem domains are fundamentally different.

---

### 2. Transmvsnet: Global context-aware multi-view stereo network with transformers

**URL**: View paper

**Brief Assessment**

TransMVSNet[59] focuses on multi-view stereo depth estimation using transformers for feature matching in cost volume construction, not on unified depth-ray prediction from arbitrary view counts with a single plain transformer backbone.

---

### 3. MVSFormer++: Revealing the Devil in Transformer's Details for Multi-View Stereo

**URL**: View paper

**Brief Assessment**

MVSFormer Plus[67] focuses on multi-view stereo depth estimation using transformer attention mechanisms for feature encoding and cost volume regularization, not on unified depth-ray prediction from arbitrary view counts without architectural specialization.

---

### 4. Is attention all that nerf needs?

**URL**: View paper

**Brief Assessment**

Attention NeRF[66] focuses on novel view synthesis using transformers for neural radiance fields, not on unified depth prediction from multiple images. The candidate addresses view synthesis and rendering, while the original paper presents a depth estimation model with depth-ray representation for 3D geometry reconstruction.

---

### 5. Vision transformers for dense prediction

**URL**: View paper

**Brief Assessment**

Dense Prediction[60] focuses on dense prediction tasks (depth estimation, semantic segmentation) using vision transformers as backbones with convolutional decoders, not on unified multi-view geometry estimation with plain transformers and depth-ray representations.

---

### 6. STViT+: improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization

**URL**: View paper

**Brief Assessment**

STViT Plus[13] focuses on multi-camera depth estimation for autonomous driving using spatial-temporal transformers, not on unified any-view geometry reconstruction with depth-ray representations from arbitrary image counts.

---

### 7. StDepthFormer: Predicting spatio-temporal depth from video with a self-supervised transformer model

**URL**: View paper

**Brief Assessment**

StDepthFormer[64] focuses on spatio-temporal depth forecasting from video sequences using self-supervised learning, while the original paper presents a unified model for any-view geometry reconstruction with depth-ray representation. The candidate addresses temporal depth prediction from videos, not the minimal single-transformer architecture for arbitrary-view geometry that the original claims as novel.

---

### 8. MVSTER: Epipolar transformer for efficient multi-view stereo

**URL**: View paper

**Brief Assessment**

MVSTER[65] focuses on multi-view stereo reconstruction using epipolar transformers for cost volume aggregation, not on unified depth-ray prediction from arbitrary numbers of images with a single plain transformer backbone.

---

### 9. Edge_MVSFormer: Edge-Aware Multi-View Stereo Plant Reconstruction Based on Transformer Networks

**URL**: View paper

**Brief Assessment**

Edge MVSFormer[62] focuses on plant reconstruction using multi-view stereo with edge detection enhancements, not on unified depth-ray prediction from arbitrary views using plain transformers.

---

### 10. Joint depth prediction and semantic segmentation with multi-view sam

**URL**: View paper

**Brief Assessment**

Multi-view SAM[63] focuses on joint depth prediction and semantic segmentation using multi-view stereo with SAM features, not on unified depth-ray prediction from arbitrary view counts with a single plain transformer backbone.

---

## Contribution 2: Teacher-student training paradigm for handling diverse real-world data

**Description**: The authors develop a teacher-student learning approach where a monocular depth teacher model trained on synthetic data generates high-quality pseudo-labels for real-world training data. This strategy addresses noisy and incomplete real-world depth captures while preserving geometric accuracy.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. Distill any depth: Distillation creates a stronger monocular depth estimator

**URL**: View paper

**Brief Assessment**

Distill Any Depth[69] focuses on depth normalization strategies and cross-context distillation for monocular depth estimation, not on the teacher-student paradigm for handling diverse real-world data with synthetic pseudo-labels as described in the original paper's multi-view geometry context.

---

### 2. Unsupervised monocular depth learning using self-teaching and contrast-enhanced SSIM loss

**URL**: View paper

**Brief Assessment**

Self Teaching SSIM[73] uses teacher-student learning for unsupervised monocular depth on augmented vs. non-augmented images at different resolutions, not for generating pseudo-labels from synthetic data to supervise real-world training as in the original paper.

---

### 3. Self-distilled self-supervised depth estimation in monocular videos

**URL**: View paper

**Brief Assessment**

Self Distilled Depth[76] focuses on monocular depth estimation with self-distillation techniques, while the original paper addresses multi-view geometry reconstruction with teacher-student learning for pseudo-labeling real-world data. The candidate's context is too limited to assess overlap with the original's specific teacher-student paradigm for handling noisy real-world depth captures.

---

### 4. Semi-supervised iterative teacher-student learning for monocular depth estimation

**URL**: View paper

**Brief Assessment**

Iterative Teacher Student[68] focuses on monocular depth estimation using teacher-student learning with noise injection for semi-supervised training. The original paper's contribution addresses multi-view geometry reconstruction with synthetic pseudo-labels for real-world data alignment, which is a different application domain and technical approach.

---

### 5. EndoOmni: Zero-shot cross-dataset depth estimation in endoscopy by robust self-learning from noisy labels

**URL**: View paper

**Prior Art Analysis**

EndoOmni[75] demonstrates prior use of a teacher-student training paradigm where a teacher model trained on synthetic/labeled data generates pseudo-labels for real-world training data to address noisy and incomplete annotations. The candidate explicitly describes training a teacher model on labeled datasets to generate pseudo-labels for unlabeled data, and addresses label noise through confidence estimation - a strategy that predates the original paper's claimed novelty. Both papers use teacher models to generate high-quality pseudo-labels for real-world data with noisy annotations, employ confidence-based weighting to handle label quality issues, and align pseudo-labels with original ground truth to preserve geometric accuracy.

**Evidence**

Evidence 1 - **Rationale**: Both papers describe using a teacher model trained on high-quality data to generate pseudo-labels for real-world training data with quality issues, demonstrating prior art for this approach. - **Original**: we train a powerful teacher monocular depth model on synthetic data to generate dense, high-quality pseudodepth for all real-world data. crucially, to preserve geometric integrity, we align these dense pseudodepth maps with the original sparse or noisy depth. - **Candidate**: we leverage a self-learning strategy that utilizes a pretrained teacher model to generate pseudo labels. this combined labeled and pseudo-labeled data is then used to train a student model that processes highly perturbed images. this selflearning framework enhances training data variety, promoting g...

---

### 6. Monocular depth estimation via self-supervised self-distillation

**URL**: View paper

**Brief Assessment**

Self Distillation Depth[72] focuses on self-distillation for dynamic scene depth estimation with teacher-student architecture, while the original paper uses teacher-student learning to generate pseudo-labels from synthetic data for real-world training data across diverse multi-view geometry tasks.

---

### 7. Er-depth: Enhancing the robustness of self-supervised monocular depth estimation in challenging scenes

**URL**: View paper

**Brief Assessment**

ER Depth[74] uses a mean teacher paradigm for self-distillation in monocular depth estimation for challenging weather conditions, not for generating pseudo-labels from synthetic data to supervise real-world multi-view geometry training as in the original paper.

---

### 8. Leveraging Near-Field Lighting for Monocular Depth Estimation from Endoscopy Videos

**URL**: View paper

**Brief Assessment**

Near Field Lighting[70] focuses on endoscopy-specific depth estimation using photometric cues and shading representations, not on general multi-view geometry or synthetic-to-real transfer for diverse visual domains.

---

### 9. 3d distillation: Improving self-supervised monocular depth estimation on reflective surfaces

**URL**: View paper

**Brief Assessment**

3D Distillation[77] uses a teacher-student approach for monocular depth estimation on reflective surfaces, but focuses on a different problem domain (handling specular reflections) rather than the general multi-view geometry task addressed in the original paper. The candidate's teacher generates pseudo-labels specifically for reflective surfaces using synthetic data, while the original paper's teacher-student paradigm addresses noisy real-world depth captures across diverse datasets for unified visual geometry estimation.

---

### 10. Exploiting the Potential of Self-Supervised Monocular Depth Estimation via Patch-Based Self-Distillation

**URL**: View paper

**Brief Assessment**

Patch Based Distillation[71] focuses on self-distillation within a single model using patch-based pseudo-labels for fine-grained depth recovery, not on using a separate teacher model trained on synthetic data to generate pseudo-labels for diverse real-world training data.

---

## Contribution 3: Visual geometry benchmark for evaluating pose, geometry, and rendering

**Description**: The authors introduce a comprehensive benchmark spanning 5 datasets with 89 scenes that directly evaluates pose accuracy, depth via reconstruction accuracy, and visual rendering quality. The benchmark includes a novel feed-forward novel view synthesis evaluation across 160 scenes.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. MonoIndoor++:Towards Better Practice of Self-Supervised Monocular Depth Estimation for Indoor Environments

**URL**: View paper

**Brief Assessment**

MonoIndoor Plus[57] focuses on self-supervised monocular depth estimation for indoor environments with specific modules for depth factorization and pose estimation. It does not present a comprehensive benchmark spanning multiple datasets for evaluating pose accuracy, depth reconstruction, and novel view synthesis as described in the original contribution.

---

### 2. Vogtareuth Rehab Depth Datasets: Benchmark for Marker-less Posture Estimation in Rehabilitation.

**URL**: View paper

**Brief Assessment**

Vogtareuth Rehab[58] focuses on marker-less posture estimation in rehabilitation settings using depth datasets, not on comprehensive visual geometry benchmarks spanning camera pose estimation, depth reconstruction, and novel view synthesis across diverse scene types.

---

### 3. E3D-Bench: A Benchmark for End-to-End 3D Geometric Foundation Models

**URL**: View paper

**Brief Assessment**

E3D Bench[51] focuses on benchmarking 3D geometric foundation models across multiple tasks including sparse-view depth, video depth, 3D reconstruction, pose estimation, and novel view synthesis. While both benchmarks evaluate pose estimation, depth/reconstruction, and rendering quality, E3D Bench[51] is designed specifically for evaluating end-to-end 3D geometric foundation models rather than general visual geometry estimators, representing a different evaluation paradigm.

---

### 4. Visual SLAM with 3D Gaussian Primitives and Depth Priors Enabling Novel View Synthesis

**URL**: View paper

**Brief Assessment**

Gaussian Primitives SLAM[53] focuses on RGB-D SLAM with 3D Gaussian splatting for dense reconstruction and pose estimation, not on creating comprehensive benchmarks. The candidate evaluates its method on TUM-RGBD dataset but does not introduce a novel benchmark spanning multiple datasets with feed-forward novel view synthesis evaluation.

---

### 5. Revealing scenes by inverting structure from motion reconstructions

**URL**: View paper

**Brief Assessment**

Revealing Scenes[54] focuses on privacy attacks by reconstructing images from SfM point clouds, not on creating benchmarks for evaluating pose accuracy, depth reconstruction, or novel view synthesis quality.

---

### 6. Towards Intelligent Embodied Perception for Indoor Agent

**URL**: View paper

**Brief Assessment**

Intelligent Embodied Perception[55] focuses on indoor embodied perception tasks. The provided context fragments do not contain sufficient detail about benchmark evaluation methodologies to assess overlap with the original paper's comprehensive visual geometry benchmark.

---

### 7. A large-scale, physically-based synthetic dataset for satellite pose estimation

**URL**: View paper

**Brief Assessment**

Satellite Pose Dataset[56] focuses specifically on satellite pose estimation with synthetic data for spacecraft operations, not on general visual geometry benchmarks spanning multiple datasets and tasks like pose, depth reconstruction, and novel view synthesis.

---

### 8. Map-free visual relocalization: Metric pose relative to a single image

**URL**: View paper

**Brief Assessment**

Map Free Relocalization[52] focuses on single-image relocalization using one reference photo, not comprehensive multi-dataset benchmarks for pose, depth reconstruction, and novel view synthesis evaluation.

---

## Appendix: Text Similarity Detection

Textual similarity detection checked 28 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

### 1. Vision transformers for dense prediction

**Detected in**: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

## References


- [0] Depth Anything 3: Recovering the Visual Space from Any Views View paper
- [1] Gomvs: Geometrically consistent cost aggregation for multi-view stereo View paper
- [2] Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models View paper
- [3] Learning multi-view stereo with geometry-aware prior View paper
- [4] SPAD: Spatially Aware Multi-View Diffusers View paper
- [5] FantasyWorld: Geometry-Consistent World Modeling via Unified Video and 3D Prediction View paper
- [6] 4D Driving Scene Generation With Stereo Forcing View paper
- [7] Topology-Aware Multi-View Street Scene Image Matching for Cross-Daylight Conditions Integrating Geometric Constraints and Semantic Consistency View paper
- [8] AmodalGen3D: Generative Amodal 3D Object Reconstruction from Sparse Unposed Views View paper
- [9] The Less You Depend, The More You Learn: Synthesizing Novel Views from Sparse, Unposed Images without Any 3D Knowledge View paper

- [10] Coherentgs: Sparse novel view synthesis with coherent 3d gaussians View paper
- [11] Worldmirror: Universal 3d world reconstruction with any-prior prompting View paper
- [12] Multi-scale geometric consistency guided multi-view stereo View paper
- [13] STViT+: improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization View paper
- [14] Deep 3d capture: Geometry and reflectance from sparse multi-view images View paper
- [15] 3D Part Segmentation via Geometric Aggregation of 2D Visual Features View paper
- [16] 3D-Adapter: Geometry-Consistent Multi-View Diffusion for High-Quality 3D Generation View paper
- [17] UniForward: Unified 3D Scene and Semantic Field Reconstruction via Feed-Forward Gaussian Splatting from Only Sparse-View Images View paper
- [18] Multi-View Stereo with Geometric Encoding for Dense Scene Reconstruction View paper
- [19] Unsupervised discovery of the shared and private geometry in multi-view data View paper
- [20] PMVC: Promoting Multi-View Consistency for 3D Scene Reconstruction View paper
- [21] GC-MVSNet: Multi-view, multi-scale, geometrically-consistent multi-view stereo View paper
- [22] Multi-view 3D object reconstruction and uncertainty modelling with neural shape prior View paper
- [23] VA-GS: Enhancing the Geometric Representation of Gaussian Splatting via View Alignment View paper
- [24] ST-GS: Vision-Based 3D Semantic Occupancy Prediction with Spatial-Temporal Gaussian Splatting View paper
- [25] CoGen: 3D Consistent Video Generation via Adaptive Conditioning for Autonomous Driving View paper
- [26] Learning spatial common sense with geometry-aware recurrent networks View paper
- [27] 360 layout estimation via orthogonal planes disentanglement and multi-view geometric consistency perception View paper
- [28] TheSHY-3D: Texture and Structure HarmonY for Multi-View 3D Object Detection View paper
- [29] Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild View paper
- [30] 3D shape segmentation with projective convolutional networks View paper
- [31] Geometric Consistency-Guaranteed Spatio-Temporal Transformer for Unsupervised Multi-View 3D Pose Estimation View paper
- [32] Towards scalable multi-view reconstruction of geometry and materials View paper
- [33] Topologically consistent multi-view face inference using volumetric sampling View paper
- [34] High-resolution shape completion using deep neural networks for global structure and local geometry inference View paper
- [35] Decomposing a scene into geometric and semantically consistent regions View paper
- [36] Inferring 3d structure with a statistical image-based shape model View paper
- [37] Multi-view supervision for single-view reconstruction via differentiable ray consistency View paper
- [38] Multi-view consistency as supervisory signal for learning shape and pose prediction View paper
- [39] 360mvsnet: Deep multi-view stereo network with 360deg images for indoor scene reconstruction View paper
- [40] Holistic Inverse Rendering of Complex Facade via Aerial 3D Scanning View paper
- [41] Lighthouse: Predicting lighting volumes for spatially-coherent illumination View paper
- [42] TomoGraphView: 3D Medical Image Classification with Omnidirectional Slice Representations and Graph Neural Networks View paper
- [43] SatDreamer360: Geometry Consistent Street-View Video Generation from Satellite Imagery View paper
- [44] What and where: 3D object recognition with accurate pose View paper
- [45] SREGS: Sparse-view Gaussian radiance fields with geometric regularization and region exploration. View paper
- [46] Self-Supervised Depth Completion Guided by 3D Perception and Geometry Consistency View paper
- [47] 3D shape completion with multi-view consistent inference View paper
- [48] RoboTransfer: Geometry-Consistent Video Diffusion for Robotic Visual Policy Transfer View paper
- [49] Breaking the Vicious Cycle: Coherent 3D Gaussian Splatting from Sparse and Motion-Blurred Views View paper
- [50] Generative modeling for continuous non-linearly embedded visual inference View paper
- [51] E3D-Bench: A Benchmark for End-to-End 3D Geometric Foundation Models View paper
- [52] Map-free visual relocalization: Metric pose relative to a single image View paper
- [53] Visual SLAM with 3D Gaussian Primitives and Depth Priors Enabling Novel View Synthesis View paper
- [54] Revealing scenes by inverting structure from motion reconstructions View paper
- [55] Towards Intelligent Embodied Perception for Indoor Agent View paper
- [56] A large-scale, physically-based synthetic dataset for satellite pose estimation View paper
- [57] MonoIndoor++:Towards Better Practice of Self-Supervised Monocular Depth Estimation for Indoor Environments View paper
- [58] Vogtareuth Rehab Depth Datasets: Benchmark for Marker-less Posture Estimation in Rehabilitation. View paper
- [59] Transmvsnet: Global context-aware multi-view stereo network with transformers View paper
- [60] Vision transformers for dense prediction View paper
- [61] MonoDETR: Depth-guided transformer for monocular 3D object detection View paper
- [62] Edge_MVSFormer: Edge-Aware Multi-View Stereo Plant Reconstruction Based on Transformer Networks View paper
- [63] Joint depth prediction and semantic segmentation with multi-view sam View paper
- [64] StDepthFormer: Predicting spatio-temporal depth from video with a self-supervised transformer model View paper
- [65] MVSTER: Epipolar transformer for efficient multi-view stereo View paper
- [66] Is attention all that nerf needs? View paper
- [67] MVSFormer++: Revealing the Devil in Transformer's Details for Multi-View Stereo View paper
- [68] Semi-supervised iterative teacher-student learning for monocular depth estimation View paper
- [69] Distill any depth: Distillation creates a stronger monocular depth estimator View paper
- [70] Leveraging Near-Field Lighting for Monocular Depth Estimation from Endoscopy Videos View paper
- [71] Exploiting the Potential of Self-Supervised Monocular Depth Estimation via Patch-Based Self-Distillation View paper
- [72] Monocular depth estimation via self-supervised self-distillation View paper
- [73] Unsupervised monocular depth learning using self-teaching and contrast-enhanced SSIM loss View paper
- [74] Er-depth: Enhancing the robustness of self-supervised monocular depth estimation in challenging scenes View paper
- [75] EndoOmni: Zero-shot cross-dataset depth estimation in endoscopy by robust self-learning from noisy labels View paper
- [76] Self-distilled self-supervised depth estimation in monocular videos View paper
- [77] 3d distillation: Improving self-supervised monocular depth estimation on reflective surfaces View paper