Auto3R: Automated 3D Reconstruction and Scanning via Data-driven Uncertainty Quantification

Chentao Shen[1,2], Sizhe Zheng[2], Bingqian Wu[2], Yaohua Feng[1], YuanChen Fei[3], Mingyu Mei[1], Hanwen Jiang[4], and Xiangru Huang[2]

[1] Zhejiang University, Hangzhou, China
[2] Westlake University, Hangzhou, China
[3] Hunan University, Changsha, China
[4] Adobe Research, California, USA
Corresponding author: huangxiangru@westlake.edu.cn

Fig. 1: Left: The scanning viewpoints selected by Auto3R, exhibiting a tendency to converge toward areas with occlusion. Right: Auto3R achieves accurate reconstruction quality and significantly outperforms the state-of-the-art methods.

Abstract. Traditional high-quality 3D scanning and reconstruction typically relies on human labor to plan the scanning procedure. With the rapid development of embodied systems such as drones and robots, there is a growing demand of performing accurate 3D scanning and reconstruction in an fully automated manner. We introduce Auto3R, a data-driven uncertainty quantification model that is designed to automate the 3D scanning and reconstruction of scenes and objects, including objects with non-lambertian and specular materials. Specifically, in a process of iterative 3D reconstruction and scanning, Auto3R can make efficient and accurate prediction of uncertainty distribution over potential scanning viewpoints, without knowing the ground truth geometry and appearance. Through extensive experiments, Auto3R achieves superior performance that outperforms the state-of-the-art methods, particularly on challenging viewpoints. We also deploy Auto3R on a robot arm equipped with a camera and demonstrate that Auto3R can be used to effectively digitize real-world 3D objects and delivers ready-to-use and photorealistic digital assets. Our code is available at https://tomatoma00.github.io/auto3r.github.io/.

Keywords: Active reconstruction · Uncertainty quantification · Gaussian splatting

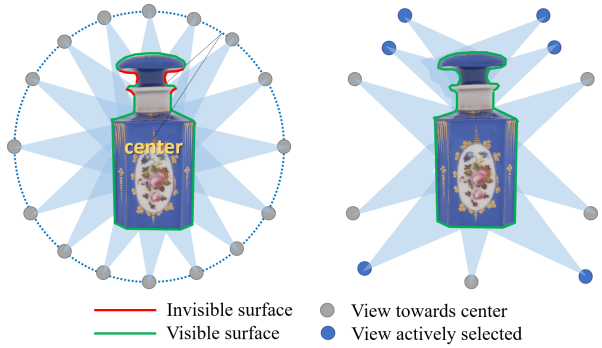

Fig. 2: Left: Reconstruction from randomly sampled viewpoints can lead to incomplete observation (highlighted with red surface) and thus low-quality results. Right: Active planning formulated with uncertainty prediction solves the problem by selectively chose observation viewpoints for scanning.

1 Introduction

Acquiring high-fidelity 3D assets is essential for applications in gaming, film, and virtual reality. While recent advances in 3D reconstruction—particularly approaches based on 3D Gaussian Splatting (3DGS) [21] and neural rendering [30], have greatly improved visual quality, the reconstruction pipeline itself remains costly and labor-intensive. In practice, generating a high-quality 3D model still needs capturing numerous informative views, as shown in Fig. 2, involving manual trajectory planning and repeated quality inspection [9, 24, 31]. To address this challenge, we study the problem of Active Reconstruction, which aims to automate and accelerate the data acquisition process. It is commonly formulated as an iterative loop: from the currently captured images, the system performs reconstruction, selects the next viewpoints expected to yield the greatest improvement, then captures these views. This procedure reduces human involvement, improves efficiency, and is evaluated by the final reconstruction quality as well as the resources consumed (e.g., number of views or total runtime).

The key challenge in active reconstruction lies in developing a reliable and efficient **Uncertainty Quantification (UQ)** model to guide viewpoint selection. This is non-trivial, as the UQ model must determine where to scan next based solely on the current reconstruction results. Ideally, it should accurately identify regions of low reconstruction quality while avoiding redundant views, thereby improving reconstruction efficiency and overall quality.

Traditional UQ models [17,45] rely on analytical metrics derived from statistics and information theory, such as Fisher information or mutual information. However, these metrics serve only as *indirect proxies* of the true reconstruction error. While they approximate the expected information gain from adding a new view, they fail to capture complex photometric and geometric inconsistencies arising from material properties or occlusions. In addition, their reliance on

analytical approximations (e.g., local linearization or Hessian estimation) leads to significant computational overhead, hindering real-time performance.

Recent works such as PUN [49] and Active-View-Selector (AVS) [42] adopt data-driven approaches that learn uncertainty directly from rendered or captured images. However, these methods remain limited. PUN predicts a global uncertainty field from a single input image, disregarding the intermediate reconstruction states and thus lacking scene-specific adaptivity. AVS improves upon this by comparing rendered candidate views with existing captures, yet it only estimates 2D image-level reconstruction errors (e.g., SSIM) without explicitly reasoning about 3D geometry or depth reliability.

In contrast, our method introduces a **joint 2D–3D uncertainty formulation** that integrates rendered color, depth, and their uncertainties within a unified data-driven framework. Unlike PUN and AVS, Auto3R explicitly couples photometric and geometric uncertainty, enabling more precise identification of ambiguous or incomplete regions. This design allows the system to select viewpoints that maximize the expected reduction in reconstruction uncertainty while remaining computationally efficient.

To summarize, our contributions are:

- **A fully automated 3D reconstruction framework.** We introduce **Auto3R**, the depth aware data-driven uncertainty quantification (UQ) model tailored for active 3D reconstruction and scanning.
- **Data-driven uncertainty quantification.** We design a dual-branch UQ model that fuses color and depth cues via depth-aware blending and reweighting for precise and efficient viewpoint selection.
- **Real-world robotic deployment.** We extend Auto3R with a video-based UQ module for path-level uncertainty estimation, enabling efficient and adaptive robotic scanning in real-world environments.

2 Related Works

Traditional reconstruction methods often depend on a fixed, manually captured set of images (e.g., spherical or grid-based sampling). Such approaches typically require a large number of input views to achieve satisfactory reconstruction quality, leading to inefficiencies in both data acquisition and computational cost. To address this limitation, automated reconstruction methods have been proposed to actively predict the next viewpoint that is expected to maximize the improvement in reconstruction quality.

Automated 3D Reconstruction has been extensively studied, evolving from explicit geometric representations (e.g., point clouds, meshes, voxels) to neural implicit and hybrid representations (e.g., NeRF, SDF, 3DGS). For *explicit representations*, point cloud–based methods [5–7] identify under-observed regions via point density and visibility analysis, while mesh-based approaches [23, 25, 43] detect incomplete surfaces from curvature and normal consistency. Voxel-based methods [10, 12, 22, 32, 33, 40, 41] estimate view utility through occupancy and

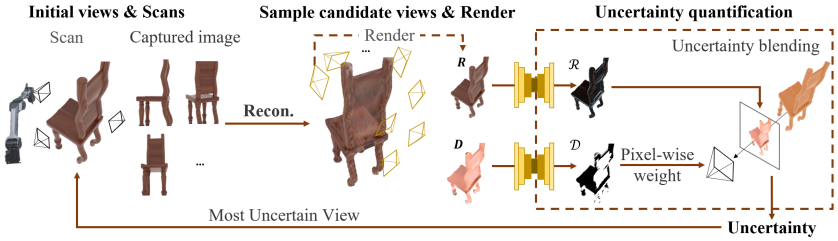

Fig. 3: An illustration of our automated scanning and reconstruction methods. Given a set of scanned images, we reconstruct the scene and render the corresponding RGB images and depth maps from candidate scanning viewpoints. We proposed an uncertainty quantification on them, obtaining the uncertainty of each candidate viewpoint. Finally scan on the most uncertain viewpoints, and repeat the process.

uncertainty measures. However, these geometry-driven techniques rely on handcrafted priors, making them sensitive to noise and less generalizable to complex or reflective surfaces. With *NeRF-based active reconstruction*, models such as ActiveNeRF [34], NeRFDirector [44], and NVF [46] estimate information gain or visibility to guide view selection, but remain computationally intensive and limited to synthetic settings. Recently, *3DGS-based methods* show promise for efficient active reconstruction. ActiveGS [19], ActiveSplat [27], FisherRF [17], and GauSS-MI [45] adopt uncertainty or visibility-driven selection, yet still rely on analytic approximations (e.g., Hessian or mutual information) that are costly and weakly correlated with perceptual quality. In contrast, our *Auto3R* introduces a data-driven uncertainty quantification model that learns directly from rendered color and depth, achieving accurate, material-robust view planning.

Uncertainty Quantification. The core of automated reconstruction lies in accurately estimating the *uncertainty* of the current reconstruction. Uncertainty quantification (UQ) is central to active reconstruction, as it guides the system to prioritize viewpoints that most effectively reduce geometric or photometric ambiguity while avoiding redundant captures. Existing methods can be categorized by the *space in which uncertainty is defined*. Methods that operate in the *3D domain* (hereafter *3D uncertainty*) typically attach uncertainty parameters to geometric primitives such as voxels or Gaussians. Early works estimate uncertainty from visibility and occupancy [22, 23], while later studies measure geometric confidence through distribution variance [1, 26, 35–37, 47]. Information-theoretic models [15, 17, 19, 45] evaluate potential information gain using Fisher or Shannon information, and recent works [28] model uncertainty in latent feature space. Although effective, these 3D-based approaches rely on analytical approximations or handcrafted priors, often missing photometric inconsistencies or reflectance-induced ambiguities. Other methods define uncertainty in the *2D rendering space*. Rendering-based methods [18, 39] estimate pixel-wise variance, whereas noise-based approaches [3, 34, 42] infer uncertainty from observation noise using fixed priors. NeRF-based UQ models [14, 29, 44] extend this idea to per-ray estimation but remain computationally expensive. In contrast,

our approach jointly learns 2D and 3D uncertainty representations directly from rendered color and depth, bridging appearance and geometry. This data-driven formulation captures both photometric artifacts and structural incompleteness, enabling accurate, efficient, and generalizable UQ for active reconstruction.

3 Overview

In this section, we briefly introduce the key components of our algorithm.

Framework. As shown in Fig. 3, we begin reconstruction from a small set of initial viewpoints. After several iterations, a set of candidate viewpoints is sampled and evaluated based on the uncertainty of the current reconstruction under each candidate view. The candidate viewpoint with the highest uncertainty is then selected for scanning, and the newly captured image is added to the training set for iterative refinement

3DGS-based Reconstruction. We adopt 3D Gaussian Splatting (3DGS) as our scene representation due to its fast convergence and suitability. For viewpoint selection, we render both color and depth map from the 3D Gaussians for each candidate viewpoint, enabling uncertainty evaluation for view planning.

Data-driven Uncertainty Prior. We propose a module to quantify the uncertainty of rendered color and depth, which typically manifests as Gaussian-induced blur, aliasing artifacts, or structural distortions. A lightweight network takes the rendered images and depth maps as input and predicts per-view *render uncertainty* $\mathcal{R}$ and *depth uncertainty* $\mathcal{D}$ for each candidate viewpoint.

Uncertainty Quantification (UQ). Our method jointly models uncertainty inferred from both 2D appearance and 3D geometry. We introduce a depth-aware uncertainty blending scheme that combines the render uncertainty $\mathcal{R}$ with depth information using depth values as blending weights. Since depth itself carries uncertainty, each depth layer is further re-weighted by its corresponding depth uncertainty before the final aggregation.

Automated Scanning. We further extend our UQ model to plan a continuous scanning path consisting of multiple viewpoints, rather than selecting a single view at a time. A lightweight video-based network takes as input a sequence of rendered images along the candidate path and predicts an overall uncertainty score, which guides the robot in selecting the optimal scanning trajectory.

4 Method

In this section, we introduce the detail of data-driven image uncertainty prior (Sec. 4.1), uncertainty quantification (Sec. 4.2), and the extension to real-world scanning (Sec. 4.3).

4.1 Data-driven Image Uncertainty Map

As described in Sec. 3, we represent the object as 3D Gaussians, denoted as $\mathcal{G}$. Each Gaussian $\mathcal{G}^i$ has a mean $\boldsymbol{\mu}^i = (x^i, y^i, z^i)$, covariance $\boldsymbol{\Sigma}^i$, color c^i, and opacity α^i, where the first two define spatial shape and the latter define appearance.

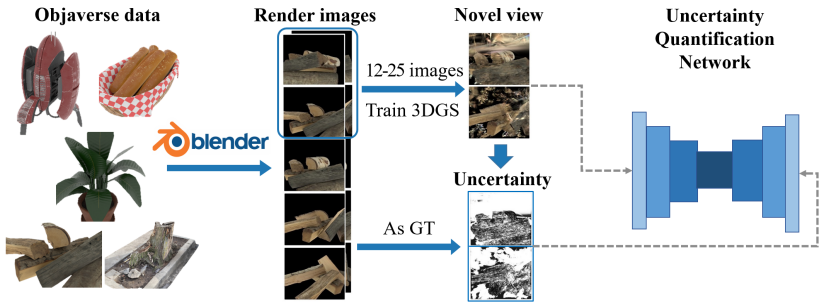

Fig. 4: Training pipeline of our data-driven uncertainty map network. We render 3000 Objaverse objects with random illumination, train 3DGS using 12–25 sparse views, and render novel views for training the model. The uncertainty is then formulated as the SSIM between the rendered and ground-truth novel views.

When projected to a viewpoint, Gaussians are splatted and composited in depth order. The rendered color and depth for pixel (u, v) are computed as:

$$\mathbf{R}(u,v) = \sum_i c^i \alpha^i \prod_{n=1}^{i-1} (1-\alpha^n), \tag{1}$$

$$\mathbf{D}(u,v) = \sum_i z^i \alpha^i \prod_{n=1}^{i-1} (1-\alpha^n). \tag{2}$$

Due to incomplete viewpoint coverage or local overfitting, artifacts such as color distortion, ghosting, and inconsistent geometry often appear in rendered images or depth maps. These visible errors directly indicate regions of high **reconstruction uncertainty**. We therefore model these spatially varying artifacts as a **uncertainty map**, which can be inferred from a single rendered image or depth map. Thus, we employ two lightweight ResNet-50–based image-to-image networks to predict pixel-wise uncertainty maps from rendered color and depth, respectively. Both networks are trained in a self-supervised manner, requiring no additional ground-truth uncertainty labels.

The training process is illustrated in Fig. 4. We select 3,000 objects from the MaterialAnything subset [16] of Objaverse [11] and render them under random HDR environment lighting in Blender 3.2.2. For each object, 60 random camera poses are generated to ensure full object coverage. Among them, 12–25 views are used to train the 3DGS model, while the remaining 35–48 novel views are rendered for uncertainty supervision. For scene-level experiments, we train our model on the map-free-reloc dataset [2] using the same protocol. Because the number of training views varies, the resulting reconstructions naturally exhibit a wide quality range – from coarse to highly accurate – providing diverse supervision for uncertainty learning. We use the pixel-wise SSIM between rendered and ground-truth images as the training signal, enabling the model to learn the mapping from local visual artifacts to uncertainty.

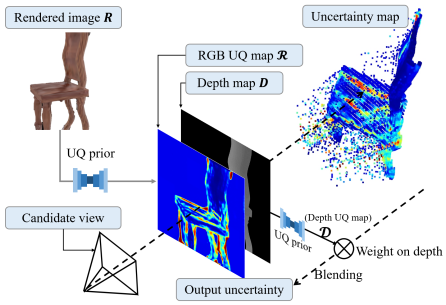

Fig. 5: Illustration of our uncertainty quantification. Rendered-image uncertainty is integrated along depth using perspective projection, where deeper pixels receive higher weights, and is further reweighted by depth-map uncertainty.

4.2 Uncertainty Quantification

After training the UQ network described in Sec. 4.1, given a Gaussian-rendered RGB image $\mathbf{R}$ and depth map $\mathbf{D}$, we obtain two uncertainty maps: the image-based uncertainty $\mathcal{R}$ and the depth-based uncertainty $\mathcal{D}$. We note that relying on either modality alone provides incomplete estimation, as image-based UQ is sensitive to texture and lighting variations, whereas depth-based UQ is less robust to photometric artifacts. Thus, we integrate both cues into a unified uncertainty quantification framework that jointly leverages appearance and geometry.

Depth-aware Blending. For perspective projection, the apparent size of a 3D Gaussian on the image plane is inversely proportional to its distance from the camera: a farther Gaussian projects to a smaller area and influences fewer pixels. As a result, the expected number of contributing Gaussians per pixel grows approximately quadratically with depth. To account for this effect, we propose a **depth-aware blending** strategy that assigns larger weights to deeper pixels during uncertainty aggregation. The blending is formulated as:

$$l_{\text{blend}} = \sum_{(u,v)} \mathbf{D}(u,v) \cdot \mathcal{R}(u,v), \tag{3}$$

where $\mathbf{D}(u,v)$ denotes the depth value and $\mathcal{R}(u,v)$ the corresponding image-based uncertainty at pixel (u,v). This operation effectively emphasizes regions with higher projection overlap and reconstruction ambiguity.

Depth-Uncertainty Reweighting. During the early stages of reconstruction, depth estimates may be unreliable due to sparse observations. To reduce the impact of inaccurate depth, we reweight each pixel's contribution according to its predicted depth uncertainty. Pixels with higher confidence (i.e., lower $\mathcal{D}$) receive larger weights, leading to the following formulation:

$$l_{\text{blend}^*} = \sum_{(u,v)} \mathbf{D}(u,v) \cdot \mathcal{R}(u,v) \cdot (1 - \mathcal{D}(u,v)). \tag{4}$$

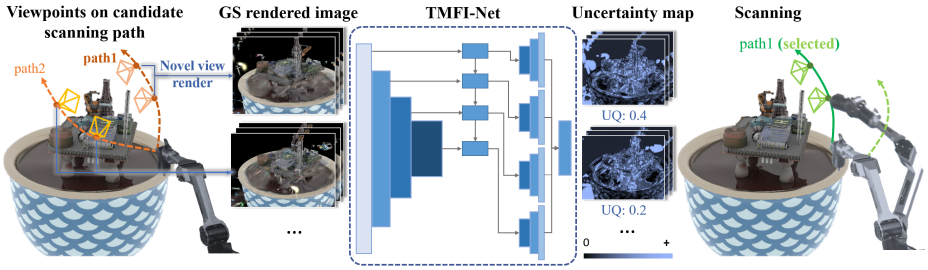

Fig. 6: The extension of uncertainty quantification for real-world scanning. We measure uncertainty on rendered image sampled along potential active trajectories. Then we select the trajectory with larger uncertainty to scan.

Finally, we integrate both blending terms with global regularization over the uncertainty maps:

$$U = l_{\text{blend}} + \lambda_0 l_{\text{blend}^*} + \lambda_1 \sum_{(u,v)} \mathcal{R}(u,v) + \lambda_2 \sum_{(u,v)} \mathcal{D}(u,v), \tag{5}$$

where λ_0, λ_1, and λ_2 control the relative weights of the reweighted blending term and the regularization of image- and depth-based uncertainties, respectively. This formulation jointly models appearance and geometric uncertainty, yielding a scalar uncertainty score U for each candidate viewpoint. The viewpoint with the highest score is selected for reconstruction. We use $\lambda_0 = 0.5$, $\lambda_2 = 0.8$, and $\lambda_1 = 0.8 \times (\text{iteration}/\text{max iteration})$.

4.3 Scanning Path UQ for Robotics Task

In real-world scenarios, the camera is typically mounted on a robotic platform that captures images continuously during motion. Capturing images along the trajectory introduces only minor efficiency overhead compared to pausing at discrete target poses, as the overall acquisition time primarily depends on the number of robot movements.

We extend our image-based UQ model to take a **sequence of images on candidate scanning path** as input, enabling the system to determine the optimal scanning path for sequential image capture. We adapt the encoder of TMFI-Net [50] and design an upsampling module to predict per-frame uncertainty, as illustrated in Fig. 6. Compared with single-image prediction networks, this sequence-based model achieves higher efficiency and a more comprehensive understanding of object surfaces.

During training, each object is rendered with 60 views, and a 3DGS model is trained using a subset of them following Sec. 4.1. We then generate multiple paths connecting nearby viewpoints and interpolate intermediate poses along each path. Continuous image capture is simulated by rendering frames along these trajectories. As before, we use the pixel-wise SSIM between rendered and ground-truth images as the supervision signal for training TMFI-Net.

5 Experiments

We test Auto3R on three different active reconstruction tasks, active normal object reconstruction, specular object reconstruction and scene reconstruction.

5.1 Experimental Setup

Dataset. For objects reconstruction, the normal and specular objects are selected from Objaverse [11], containing objects of various types, ensuring no category overlap between training and testing sets. Notably, we randomly generated 256 candidate viewpoints for view select, and 256 test viewpoints for evaluation, to simulate real-world scanning task, we **did not input initial 3D points before reconstruction**. For scene reconstruction, we use Mip-NeRF360 [4] dataset, which covers complex contexts under different lighting conditions.

Baseline. For fair comparison, all baselines are evaluated under the active view-selection protocol. The baselines are grouped into 2 categories. The first is based on image-based uncertainty, including AVS [42], we also adapt other image uncertainty/quality quantification methods TOPIQ [8], TRES [13], MANIQA [48], MUSIQ [20] to the framework of active reconstruction. On the other hand, we also compare with baselines based on 3D uncertainty, such as FisherRF [17]and Gauss-MI [45]. As for AVS, We re-implemented the CNN version and trained it on our dataset under the same conditions for fair comparison.

Scheduling. For both scene and objects reconstruction, we train the 3DGS model for 30000 iterations. Over these iterations, we add one new view at specific intervals **following the schedule of FisherRF** [17]: [400, 900, 1500, 2200, 3000, 3900, 4900, 6000, 7200, 8500, 9900, 11400, 13000, 14700, 16500, 18400]. In all methods, we provide a final set of 20 views, including the initial views.

Metrics. We evaluate reconstruction quality using standard image-based metrics that measure photometric fidelity and perceptual similarity. We report PSNR, SSIM, and LPIPS for novel view synthesis under test viewpoints. For comprehensive evaluation, each task includes both the average performance across all viewpoints and the worst 5% performance to capture failure cases.

5.2 Evaluation of Uncertainty Quantification

The uncertainty quantification results are shown in Fig. 7. For the objects in the first row, the predicted and GT images exhibit texture differences, which correspond to high uncertainty regions in the GT maps (red/yellow region). The second row compares the predicted and GT uncertainty maps. The heatmaps between predict and GT match well, verifying the model understands reconstruction errors. Quantitative results are provided in **Supplementary Material D**.

5.3 Evaluation of Active Object Reconstruction

In this section, we compare our method and other state-of-the-art methods on the task of active object reconstruction. The corresponding experimental results are illustrated in Tab. 1, Figs. 8 and 9.

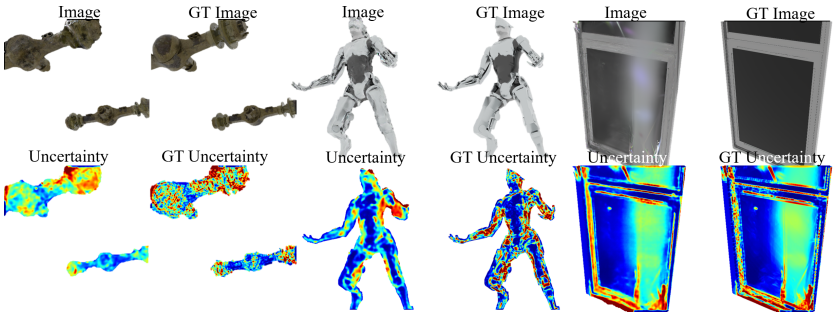

Fig. 7: Visualization of Uncertainty Predictions. Top row: rendered image (input) and ground truth (GT). Bottom row: predicted and GT uncertainty maps. Color encoding from blue (low) to red (high) indicates uncertainty magnitude.

From the quantitative results, Auto3R shows the highest PSNR and SSIM with lowest LPIPS, indicating low pixel-level error and faithful geometric structure. AVS and FisherRF also shows the acceptable results, but with some issues in the details, such as the keyboard, truck carriage in Fig. 8. For predict time, Auto3R cost about 3ms per image, with enough efficiency for real-time scanning. Generally, Auto3R achieves excellent performance in this task, outperforming the baselines in most cases.

Table 1: Quantitative results of active object reconstruction on the Objaverse dataset. We report average and worst-5% PSNR, SSIM, and LPIPS over all test views. Auto3R achieves the best overall reconstruction quality across all metrics.

Methods	PSNR↑		SSIM↑		LPIPS↓	
	avg	worst	avg	worst	avg	worst
Random	16.09	8.37	0.5972	0.2192	0.3400	0.5766
FisherRF	23.08	17.85	0.8055	0.6320	0.1063	0.2689
AVS	23.94	18.18	0.8479	0.7232	0.0935	0.2191
GAUSS-MI	16.25	12.71	0.4214	0.2495	0.3269	0.6968
TOPIQ+GS	18.56	11.12	0.7150	0.4075	0.2538	0.5031
TRES+GS	18.47	10.45	0.6803	0.1798	0.2790	0.6469
MANIQA+GS	14.34	8.08	0.5032	0.1686	0.4032	0.6348
MUSIQ+GS	15.82	8.64	0.5503	0.2429	0.3696	0.6419
Auto3R	**27.00**	**21.88**	**0.8882**	**0.7819**	**0.0711**	**0.1813**

5.4 Evaluation of Scene Reconstruction

To further validate the performance of our model, we carried out a larger-scale scene-level reconstruction. There are more objects and more intricate structures in the scene, which makes active reconstruction much more difficult.

We show the quantitative results in Tab. 2, where Auto3R delivers the best overall reconstruction quality. It performs best in PSNR and LPIPS and a close

Table 2: Active scene-level reconstruction results on Mip-NeRF360. Auto3R consistently outperforms previous uncertainty-based and image-based methods in both average and worst-5% case reconstruction quality over all test views.

Methods	PSNR↑		SSIM↑		LPIPS ↓	
	avg	worst	avg	worst	avg	worst
Random	13.34	7.14	0.2341	0.1801	0.5916	0.6241
FisherRF	18.74	10.32	0.5771	0.3299	0.2619	0.4788
AVS	18.95	9.28	**0.6062**	0.3709	0.2679	0.5185
GAUSS-MI	18.22	8.99	0.5609	0.3267	0.2749	0.5142
TOPIQ+GS	13.44	7.30	0.2350	0.1779	0.5981	0.6299
TRES+GS	13.58	6.87	0.2388	0.1890	0.5940	0.6291
MANIQA+GS	13.46	6.64	0.2312	0.1807	0.5918	0.6338
MUSIQ+GS	13.67	6.85	0.2340	0.1800	0.5901	0.6137
Auto3R	**19.13**	**10.75**	0.6055	**0.3862**	**0.2520**	**0.4374**

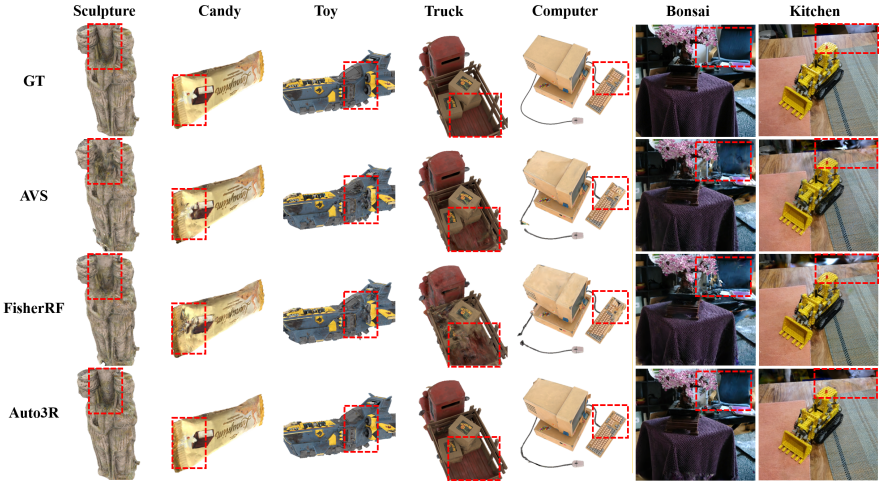

Fig. 8: Qualitative Evaluation of Reconstructed Results: objects (left 5 columns) and scenes (right 2 columns). Rows from top to bottom show ground truth (GT), AVS, FisherRF, and our method.

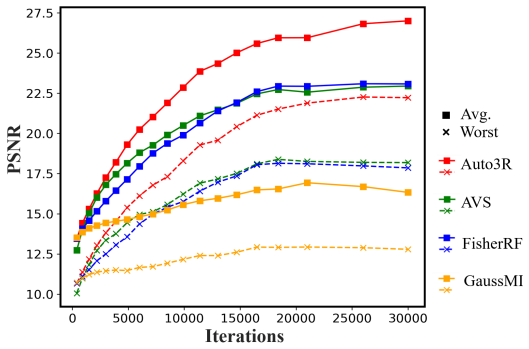

Fig. 9: Per-step curves: average PSNR (solid) and worst-5% PSNR (dashed) with different iterations of next view sampling for each method. Colors differentiate methods.

second in SSIM, preserving fine structures while suppressing artifacts. Moreover, as shown in Fig. 8, we achieve the most accurate results especially on background. These results demonstrate competitive performance for large-scale scenes.

Fig. 10: Visualization of Specular Objects Reconstruction.

5.5 Evaluation of Specular Objects Reconstruction

Since 3DGS is insufficient for high-quality reconstruction of specular reflective objects. We therefore compare our framework with other 2D UQ-based methods based on GIR [38] (an inverse rendering 3DGS), we adjust the schedule for adding views and post-augmentation optimization parameters to accommodate its material estimation pipeline. Details of implement active reconstruction for GIR are provided in the **Supplementary Material.A**. And we also conduct the normal 3DGS-based framework for reconstruction with the similar setting.

Tab. 3 and Fig. 10 shows the quantitative results of the specular objects reconstruction, including both 3DGS and GIR-based methods. The results show that our Auto3R outperforms other methods. We can see that combining GIR with our method leads to more accurate reconstruction of regions with reflection.

Table 3: Active reconstruction results for specular objects. We evaluate both 3DGS-based and GIR-based frameworks. Auto3R significantly improves reflection fidelity and quantitative accuracy under challenging non-Lambertian surfaces.

Framework	Methods	PSNR↑		SSIM↑		LPIPS↓	
		avg	worst	avg	worst	avg	worst
3DGS	Random	17.69	12.51	0.6483	0.3563	0.2795	0.5447
	FisherRF	22.84	17.66	0.8248	0.6715	0.0923	0.2689
	AVS	23.82	18.61	0.8652	0.7401	0.0732	0.2419
	TOPIQ+GS	20.92	14.69	0.7719	0.4593	0.1645	0.4536
	Auto3R	**26.41**	**20.94**	**0.8904**	**0.7769**	**0.0686**	**0.1612**
GIR	Random	26.85	20.09	0.8855	0.8278	0.0768	0.0909
	AVS	29.53	20.36	0.9417	0.8319	0.0532	0.0986
	Auto3R	**30.29**	**22.92**	**0.9505**	**0.8791**	**0.0497**	**0.0782**

5.6 Ablation Study

We conduct an ablation study to evaluate the contribution of two key components in uncertainty quantification: (1) the **depth-aware blending**, which aggregates uncertainty along depth to better reflect 3D coverage, and (2) the **depth-uncertainty reweighting**, which adaptively adjusts the blending weights based on confidence in depth estimation. The results are summarized in Tab. 4.

Effect of Depth Blending. Without depth blending (only relies on 2D appearance cues), the model ignores geometric consistency across viewpoints. This leads to unstable uncertainty estimation, particularly in occluded or depth-varying regions. Introducing depth blending improves PSNR from 22.48 to 23.41 (+0.93) and SSIM from 0.7818 to 0.8124 (+0.0306), showing that incorporating geometric cues provides a reliable indicator of poorly reconstructed areas. Among different formulations, the linear depth weighting performs slightly better than quadratic weighting, suggesting that moderate depth dependence balances local and global uncertainty aggregation effectively.

Effect of Depth Uncertainty Reweighting. Further adding depth uncertainty reweighting enhances reconstruction quality across all metrics. By dynamically suppressing unreliable depth regions in early reconstruction stages, this mechanism refines the uncertainty field and prevents overconfident sampling near ambiguous geometry. With this module, PSNR rises from 23.41 to 23.77 and SSIM from 0.8124 to 0.8252, confirming that coupling uncertainty prediction with self-assessed depth confidence improves the robustness of view selection.

Overall Analysis. Combining both depth-aware blending and uncertainty reweighting yields the best performance, accelerating convergence toward high-fidelity reconstructions. The ablation validates that each component contributes complementary benefits—depth blending introduces geometric awareness, while reweighting stabilizes uncertainty estimation throughout reconstruction.

6 Real-world Robotic Deployment

We apply our method on a real-world robot to scanning various objects, a fabric toy, and a circuit board. We use a robot with an RGB camera fixed on the end. Details of the systems can be found in **Supplementary Materials.B**, while the generation of candidate scanning path is also introduced in this material.

Table 4: Ablation on depth blending and depth-uncertainty reweighting. Both contribute to gains in PSNR & SSIM, validating our uncertainty formulation.

Ablation	PSNR↑		SSIM↑		LPIPS↓	
	avg	worst	avg	worst	avg	worst
w/o depth-uq	23.41	17.95	0.8124	0.6494	0.1016	0.2570
w/o depth-blending	22.60	15.94	0.7910	0.5664	0.1085	0.2791
depth^2-blending	23.59	17.60	0.8131	0.6388	0.1004	0.2593
ours	**23.77**	**18.49**	**0.8252**	**0.6703**	**0.0978**	**0.2145**

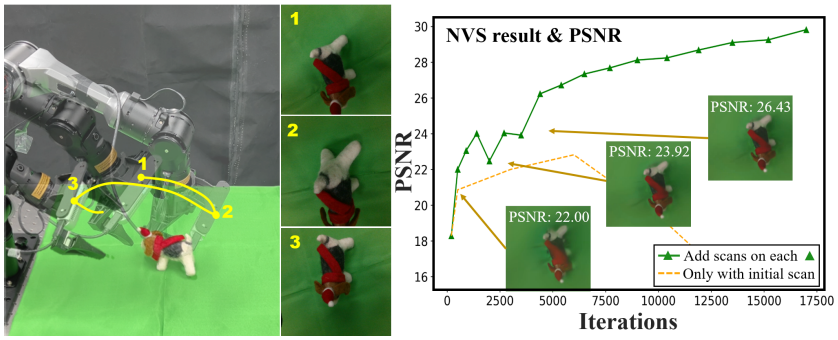

Fig. 11: Top: the first three scanning path of the robot. Bottom: the novel view rendered images and PSNR metrics by Auto3R-extension after each scan.

The layout and the scanning process of the robot is shown in the Fig. 11. And the scanning process with the real-time novel view rendering performance can be find in **Supplementary Material.C** and the **video**.

We also applied Auto3R and its extension (described in Sec. 4.3), AVS and FisherRF on robotic scanning task. For each task, we start with 4 views in a forward path. For extension of Auto3R, we predict a scanning path and scan an image sequence on the path, for remainning baselines, we keep the same intervals for sampling new view. Tab. 5 shows the results of each method, where the extension of Auto3R achieve the most accurate results, indicating that the effectiveness of extension UQ for scanning. The remaining results indicate that our Auto3R outperforms the other baselines on real-world reconstruction case.

7 Conclusion

Conclusion. In this work, we presented **Auto3R**, an automated framework for active 3D reconstruction and scanning. By introducing a data-driven uncertainty quantification model, Auto3R enables accurate, geometry-aware view planning without manual intervention. Extensive experiments demonstrate that Auto3R achieves superior reconstruction quality compared to existing approaches, and can be deployed on robotic systems for real-world tasks.

Limitations. In real-world tasks, Auto3R sometimes affected by self-occlusion shadows of the robot, which will be a focus of our future investigations.

Table 5: Reconstruction quality under **scanning with Robots**.

Methods	Origin		Path-extension	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Auto3R	22.91	0.8346	**28.56**	**0.8953**
AVS	20.35	0.8159	26.98	0.8894
FisherRF	18.95	0.7415	26.04	0.8835

Acknowledgements

We thank the anonymous reviewers for their valuable feedback.

References

1. Aira, L.S., Valsesia, D., Magli, E.: Modeling uncertainty for gaussian splatting. IEEE Transactions on Neural Networks and Learning Systems **36**(6), 11657–11663 (2025). https://doi.org/10.1109/TNNLS.2025.3553582
2. Arnold, E., Wynn, J., Vicente, S., Garcia-Hernando, G., Monszpart, Á., Prisacariu, V.A., Turmukhambetov, D., Brachmann, E.: Map-free visual relocalization: Metric pose relative to a single image. In: Proc. European Conference on Computer Vision (2022)
3. Bae, G., Budvytis, I., Cipolla, R.: Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In: Proc. IEEE/CVF International Conference on Computer Vision. pp. 13117–13126 (2021). https://doi.org/10.1109/ICCV48922.2021.01289
4. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
5. Border, R., Gammell, J., Newman, P.: Surface edge explorer (see): Planning next best views directly from 3d observations. In: Proc. IEEE International Conference on Robotics and Automation. pp. 1–8 (05 2018)
6. Border, R., Gammell, J.D.: Proactive estimation of occlusions and scene coverage for planning next best views in an unstructured representation. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4219–4226 (2020). https://doi.org/10.1109/IROS45743.2020.9341681
7. Border, R., Gammell, J.D.: The surface edge explorer (see): A measurement-direct approach to next best view planning. International Journal of Robotics Research **43**(10), 1506–1532 (Sep 2024). https://doi.org/10.1177/02783649241230098
8. Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: Topiq: A top-down approach from semantics to distortions for image quality assessment. IEEE Transactions on Image Processing **33**, 2404–2418 (2024)
9. Chen, X., Li, Q., Wang, T., Xue, T., Pang, J.: Gennbv: Generalizable next-best-view policy for active 3d reconstruction. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16436–16445 (2024)
10. Daudelin, J., Campbell, M.: An adaptable, probabilistic, next-best view algorithm for reconstruction of unknown 3-d objects. IEEE Robotics and Automation Letters **2**(3), 1540–1547 (2017). https://doi.org/10.1109/LRA.2017.2660769
11. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–153 (2023)
12. Delmerico, J., Isler, S., Sabzevari, R., Scaramuzza, D.: A comparison of volumetric information gain metrics for active 3d object reconstruction. Autonomous Robots **42** (02 2018). https://doi.org/10.1007/s10514-017-9634-0
13. Golestaneh, S.A., Dadsetan, S., Kitani, K.M.: No-reference image quality assessment via transformers, relative ranking, and self-consistency. In: Proc. IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1220–1230 (2022)

14. Goli, L., Reading, C., Sellán, S., Jacobson, A., Tagliasacchi, A.: Bayes' rays: Uncertainty quantification for neural radiance fields. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20061–20070 (2024). https://doi.org/10.1109/CVPR52733.2024.01896
15. Hanson, A., Tu, A., Singla, V., Jayawardhana, M., Zwicker, M., Goldstein, T.: Pup 3d-gs: Principled uncertainty pruning for 3d gaussian splatting. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5949–5958 (June 2025), https://pup3dgs.github.io/
16. Huang, X., Wang, T., Liu, Z., Wang, Q.: Material anything: Generating materials for any 3d object via diffusion. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26556–26565 (2025)
17. Jiang, W., Lei, B., Daniilidis, K.: Fisherrf: Active view selection and mapping with radiance fields using fisher information. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Proc. European Conference on Computer Vision. pp. 422–440. Springer Nature Switzerland, Cham (2025)
18. Jin, L., Chen, X., Rückin, J., Popović, M.: Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 11305–11312 (2023). https://doi.org/10.1109/IROS55552.2023.10342226
19. Jin, L., Zhong, X., Pan, Y., Behley, J., Stachniss, C., Popović, M.: Activegs: Active scene reconstruction using gaussian splatting. IEEE Robotics and Automation Letters **10**(5), 4866–4873 (2025). https://doi.org/10.1109/LRA.2025.3555149
20. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proc. IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
21. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023)
22. Krainin, M., Curless, B., Fox, D.: Autonomous generation of complete 3d object models using next best view manipulation planning. In: Proc. IEEE International Conference on Robotics and Automation. pp. 5031–5037 (2011). https://doi.org/10.1109/ICRA.2011.5980429
23. Kriegel, S., Bodenmüller, T., Suppa, M., Hirzinger, G.: A surface-based next-best-view approach for automated 3d model completion of unknown objects. In: Proc. IEEE International Conference on Robotics and Automation. pp. 4869–4874 (2011). https://doi.org/10.1109/ICRA.2011.5979947
24. Kriegel, S., Rink, C., Bodenmüller, T., Narr, A., Suppa, M., Hirzinger, G.: Next-best-scan planning for autonomous 3d modeling. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 2850–2856. IEEE (2012)
25. Lee, I.D., Seo, J.H., Kim, Y.M., Choi, J., Han, S., Yoo, B.: Automatic pose generation for robotic 3-d scanning of mechanical parts. IEEE Transactions on Robotics **36**(4), 1219–1238 (2020). https://doi.org/10.1109/TRO.2020.2980161
26. Li, R., Cheung, Y.m.: Variational multi-scale representation for estimating uncertainty in 3d gaussian splatting. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Proc. Conference on Neural Information Processing Systems. vol. 37, pp. 87934–87958. Curran Associates, Inc. (2024)
27. Li, Y., Kuang, Z., Li, T., Hao, Q., Yan, Z., Zhou, G., Zhang, S.: Activesplat: High-fidelity scene reconstruction through active gaussian splatting. IEEE Robotics and Automation Letters (2025)

28. Liao, Z., Waslander, S.L.: Multi-view 3d object reconstruction and uncertainty modelling with neural shape prior. In: Proc. IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3086–3095 (2024). https://doi.org/10.1109/WACV57701.2024.00307
29. Lyu, L., Tewari, A., Habermann, M., Saito, S., Zollhöfer, M., Leimkühler, T., Theobalt, C.: Manifold sampling for differentiable uncertainty in radiance fields. In: Proc. SIGGRAPH Asia Conference Papers. SA '24, Association for Computing Machinery, New York, NY, USA (2024), https://doi.org/10.1145/3680528.3687655
30. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
31. Mostegel, C., Rumpler, M., Fraundorfer, F., Bischof, H.: Uav-based autonomous image acquisition with multi-view stereo quality assurance by confidence prediction. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–10 (2016)
32. Pan, S., Wei, H.: A global max-flow-based multi-resolution next-best-view method for reconstruction of 3d unknown objects. IEEE Robotics and Automation Letters **7**(2), 714–721 (2022). https://doi.org/10.1109/LRA.2021.3132430
33. Pan, S., Wei, H.: A global generalized maximum coverage-based solution to the non-model-based view planning problem for object reconstruction. Computer Vision and Image Understanding **226**, 103585 (2023). https://doi.org/https://doi.org/10.1016/j.cviu.2022.103585
34. Pan, X., Lai, Z., Song, S., Huang, G.: Activenerf: Learning where to see with uncertainty estimation. In: Proc. European Conference on Computer Vision. pp. 230–246. Springer (2022)
35. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: IEEE Conf. Comput. Vis. Pattern Recog.(CVPR) (2020)
36. Ran, Y., Zeng, J., He, S., Chen, J., Li, L., Chen, Y., Lee, G., Ye, Q.: Neurar: Neural uncertainty for autonomous 3d reconstruction with implicit neural representations. IEEE Robotics and Automation Letters **8**(2), 1125–1132 (2023). https://doi.org/10.1109/LRA.2023.3235686
37. Shen, J., Ruiz, A., Agudo, A., Moreno-Noguer, F.: Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations. In: Proc. International Conference on 3D Vision. pp. 972–981 (2021). https://doi.org/10.1109/3DV53792.2021.00105
38. Shi, Y., Wu, Y., Wu, C., Liu, X., Zhao, C., Feng, H., Zhang, J., Zhou, B., Ding, E., Wang, J.: Gir: 3d gaussian inverse rendering for relightable scene factorization. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–12 (2025)
39. Sünderhauf, N., Abou-Chakra, J., Miller, D.: Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields. In: Proc. IEEE International Conference on Robotics and Automation. pp. 9370–9376 (2023). https://doi.org/10.1109/ICRA48891.2023.10161012
40. Vasquez-Gomez, J., Sucar, L., Murrieta-Cid, R.: View/state planning for three-dimensional object reconstruction under uncertainty. Autonomous Robots **41** (01 2017). https://doi.org/10.1007/s10514-015-9531-3
41. Vasquez-Gomez, J., Sucar, L., Murrieta-Cid, R., López-Damian, E.: Volumetric next-best-view planning for 3d object reconstruction with positioning error. International Journal of Advanced Robotic Systems **11**, 1–13 (10 2014). https://doi.org/10.5772/58759

42. Wang, Z., Bhalgat, Y., Li, R., Prisacariu, V.A.: Active view selector: Fast and accurate active view selection with cross reference image quality assessment. arXiv preprint arXiv:2506.19844 (2025)
43. Wu, S., Sun, W., Long, P., Huang, H., Cohen-Or, D., Gong, M., Deussen, O., Chen, B.: Quality-driven poisson-guided autoscanning. ACM Transactions on Graphics **33**(6) (Nov 2014). https://doi.org/10.1145/2661229.2661242
44. Xiao, W., Cruz, R.S., Ahmedt-Aristizabal, D., Salvado, O., Fookes, C., Lebrat, L.: Nerf director: Revisiting view selection in neural volume rendering. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20742–20751 (2024)
45. Xie, Y., Cai, Y., Zhang, Y., Yang, L., Pan, J.: Gauss-mi: Gaussian splatting shannon mutual information for active 3d reconstruction. arXiv preprint arXiv:2504.21067 (2025)
46. Xue, S., Dill, J., Mathur, P., Dellaert, F., Tsiotra, P., Xu, D.: Neural visibility field for uncertainty-driven active mapping. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18122–18132 (2024)
47. Yan, D., Liu, J., Quan, F., Chen, H., Fu, M.: Active implicit object reconstruction using uncertainty-guided next-best-view optimization. IEEE Robotics and Automation Letters **8**(10), 6395–6402 (2023). https://doi.org/10.1109/LRA.2023.3306282
48. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1191–1200 (2022)
49. Zhang, Z., Xu, F., Zhang, M.: Peering into the unknown: Active view selection with neural uncertainty maps for 3d reconstruction. arXiv preprint arXiv:2506.14856 (2025)
50. Zhou, X., Wu, S., Shi, R., Zheng, B., Wang, S., Yin, H., Zhang, J., Yan, C.: Transformer-based multi-scale feature integration network for video saliency prediction. IEEE Transactions on Circuits and Systems for Video Technology **33**(12), 7696–7707 (2023). https://doi.org/10.1109/TCSVT.2023.3278410