ConMax3D: Frame Selection for 3D Reconstruction Through Concept Maximization

Akash Malhotra[1,2], Nacéra Seghouani[2], Gilbert Badaro[1] and Christophe Blaya[1]
[1]*Amadeus, Sophia Antipolis, France*
[2]*Université Paris-Saclay, LISN, Paris, France*
{*christophe.blaya, gilbert.badaro*}*@amadeus.com,* {*akash.malhotra, nacera.seghouani*}*@lisn.fr*

Keywords: Frame Selection, 3D Reconstruction, Semantic Segmentation, Multi-View Synthesis, NeRF, Gaussian Splatting.

Abstract: This paper proposes a novel best frames selection algorithm, ConMax3D, for multiview 3D reconstruction that utilizes image segmentation and clustering to identify and maximize concept diversity. This method aims to improve the accuracy and interpretability of selecting frames for a photorealistic 3D model generation with NeRF or 3D Gaussian Splatting without relying on camera pose information. We evaluate ConMax3D on the LLFF dataset and show that it outperforms current state-of-the-art baselines, with improvements in PSNR of up to 43.65%, while retaining computational efficiency.

1 INTRODUCTION

Creating a 3D model of an object or a scene using multiple images from different viewpoints has been a long-standing problem in computer vision. Prior to the advent of deep learning techniques for 3D reconstruction[(Mildenhall et al., 2019), (Lombardi et al., 2019), (Fridovich-Keil et al., 2023)], traditional methods such as structure from motion (Schonberger and Frahm, 2016) and multiview stereo (Seitz et al., 2006) were widely used. The introduction of Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) revolutionized novel view rendering by leveraging neural networks to create photorealistic 3D models where color is a function of camera pose. This innovation has led to a surge in research on radiance fields, including enhanced NeRF models such as Instant-NGP (Müller et al., 2022), MipNeRF (Barron et al., 2021), and ZipNeRF (Barron et al., 2023), as well as alternative techniques such as Gaussian Splatting (3DGS) (Kerbl et al., 2023) and related works[(Gao et al., 2022), (Wu et al., 2024b)].

However, radiance field-based methods often require numerous frames to train high-quality 3D representations. This challenge stems from the absence of a systematic approach for capturing optimal frames, as the requirements vary significantly based on object geometry and color distribution (Pan et al., 2024).

Techniques such as ActiveNeRF (Pan et al., 2022) and related works[(Goli et al., 2024), (Jin et al., 2023)] have proposed uncertainty estimation as a strategy to address this issue. While uncertainty-based techniques outperform random sampling, they necessitate modifications to the architecture and training regime of 3D reconstruction models such as NeRF, which could increase the training cost and complexity.

Other methods such as the one presented by (Pan et al., 2024) employ the Tammes Problem (Lai et al., 2023) to predict frame positions based solely on camera poses. Although effective for synthetic data and spherical camera poses, this approach is less applicable to real-world data where not only the frame selection is influenced by scene geometry and color distribution, but also the camera pose distribution may not be spherical. Moreover, previous approaches do not typically incorporate high-level concepts such as parts of objects within the 3D scene, which could enhance the interpretability of the frame selection process.

Given the constraints on computational resources, it is often impractical to use all available frames for 3D reconstruction. For example, a typical video captured by a smartphone is 60 frames/second and may contain over thousands of frames for a few minutes capture. Also, as observed by (Orsingher et al., 2023), the quality of reconstruction by NeRF has diminishing returns as the number of frames increases for a scene, particularly if there is a significant overlap between the frames. Consequently, the challenge becomes selecting the best subset of frames (or views)

Malhotra, A., Seghouani, N., Badaro, G. and Blaya, C.
ConMax3D: Frame Selection for 3D Reconstruction Through Concept Maximization.
DOI: 10.5220/0013258800003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 598-609
ISBN: 978-989-758-728-3; ISSN: 2184-4321

within a specified budget k that maximizes the quality of the 3D reconstruction. This constraint necessitates a selection strategy that is both accurate and fast to ensure that the chosen frames capture the essential features and variations of the scene. This paper introduces a novel algorithm named **ConMax3D**, which employs image segmentation followed by clustering to identify key concepts within a set of multiview images. A "concept" is defined as a recurring pattern in pixel color distribution across multiple images, as presented by (Asano et al., 2019). In our approach, after generating concepts, the frames are selected with the objective to maximize the inclusion and coverage of diverse concepts.

In addition to ConMax3D, Inspired by PC-NBV (Zeng et al., 2020), we propose an enhanced baseline called **Point Cloud Maximization for Frame Selection**, which first constructs a point cloud representation of the scene and then uses a heuristic based greedy frame selection strategy. Unlike PC-NBV, it does not use a neural network, which has a training overhead and potentially generalization problems in out of distribution scenes.

We compare ConMax3D with Random Sampling, Furthest View Sampling (FVS) which are also used as baselines. Point Cloud Maximization is used as an advanced baseline, and ActiveNeRF (Pan et al., 2022), is used as an state of the art baseline.

The main contributions of this paper is to propose a best frames selection algorithm that has the following characteristics:

- **Camera Pose Independence:** ConMax3D does not require camera pose information, making it applicable in varied and realistic environments using only RGB images as input.
- **Model Independence:** This method is decoupled from specific 3D reconstruction models, enhancing its utility across different radiance field-based reconstruction techniques such as NeRF (Mildenhall et al., 2021), 3D Gaussian Splatting (Kerbl et al., 2023), and others that use view dependency for color prediction.
- **Concept-Based Selection:** High-level concepts are used for frame selection, improving the interpretability of the process.

We demonstrate the effectiveness of our approach through extensive experiments on both spherical camera configurations, in which all the cameras are facing towards and are equidistant from the object centroid, and non-spherical configurations, where the cameras can be placed in arbitrary positions and orientations. Non-spherical configurations are more challenging both for frame selection and 3D reconstruction algorithms and are closer to real-world captures.

Our results show significant improvements, with gains up to 43.65% in PSNR, showing the potential of this approach in reducing the number of required frames while maintaining high-quality reconstructions.

This paper is organized as follows: we present related work in Section 2 to provide a comprehensive review of recent advancements in frame selection techniques for 3D reconstruction. We then outline our proposed framework in detail in Section 3, highlighting each component of the system, from image segmentation to concept maximization. We also describe the Point Cloud Maximization. This is followed in Section 4 by the description of our experimental setup, including the reconstruction models used, dataset, evaluation metrics, and the comparative performance of our method against existing approaches. Finally, in Section 5, we discuss the implications of our findings, address potential limitations, and suggest directions for future research in this domain.

2 RELATED WORK

Recent advancements in Neural Radiance Fields (NeRF) have focused on improving efficiency through reduced frame requirements and enhanced computational strategies. We review key contributions that align closely with our work but differ significantly in approach and methodology.

Semantic Consistency. Several works address the issue of semantic consistency and overfitting in NeRF implementations: DietNeRF (Jain et al., 2021) introduces a semantic consistency loss using a pretrained image classifier to ensure that rendered images are photorealistic and semantically consistent. PixelNeRF (Yu et al., 2021) proposes a NeRF variant conditioned on pixel-aligned features from a pretrained CNN, improving reconstruction robustness and generalization. While these methods focus on pixel-specific features or semantic consistency, our concept maximization approach selects diverse frames based on overall conceptual coverage, offering a different perspective on improving NeRF performance.

Uncertainty Quantification. Uncertainty estimation has emerged as a key strategy for optimizing view selection: ActiveNeRF (Pan et al., 2022) integrates active learning to enhance NeRF training by modeling radiance field values as a Gaussian distribution and using variance as the measure of un-

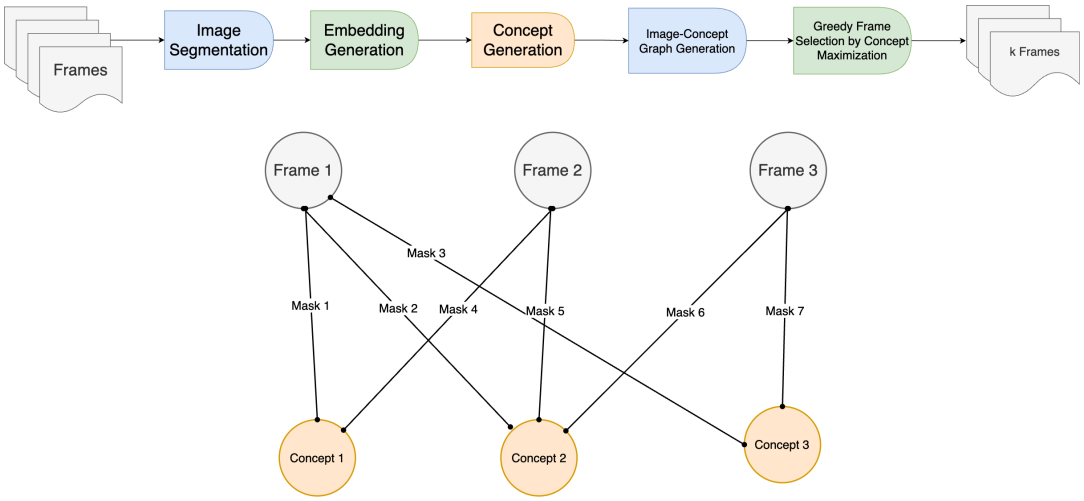


Figure 1: We propose the ConMax3D framework, which first segments the images using SAM, then clusters the obtained masks into "concepts," then creates an Image-Concept graph based on the Image-mask-concept relations, and finally selects the best k frames maximizing the conceptual diversity and coverage in a greedy manner.

certainty. BayesRays (Goli et al., 2024) introduces a post-hoc framework for uncertainty quantification in pre-trained NeRF models. NeU-NBV (Jin et al., 2023), NeurAR (Ran et al., 2023), and Smith et al. (Smith et al., 2022) propose methods for next-best-view (NBV) planning using uncertainty maps and occupancy-based models.

These techniques use only low level information (pixel-level or ray level) and modify the model architecture and/or training. Our framework, ConMax3D, does not interact with the model used and is used in the pre-processing step.

Efficient Reconstruction with Fewer Frames. Some approaches aim to improve NeRF and 3D Gaussian Splatting reconstructions with a limited number of frames. RegNeRF (Niemeyer et al., 2022), InstantSplat (Fan et al., 2024), and MVSplat (Chen et al., 2025) demonstrate efficient reconstructions by optimizing the frame rendering process. While these approaches provide better reconstruction with fewer frames, their goal is fundamentally different. These methods excel in scenarios with fewer images, optimizing for making most of what is available. In contrast, ConMax3D addresses a different challenge: selecting the optimal subset of frames from a large pool of Images (e.g., video sequences) under specific constraints such as GPU memory and training time. This selection process enhances the applicability of any subsequent reconstruction, including those performed by sparse and efficient methods.

Autonomous Data Collection. Frameworks for optimizing the NeRF training process through autonomous data collection have been proposed: AutoNeRF (Marza et al., 2024) develops an autonomous data collection framework through exploration. (Kopanas and Drettakis, 2023) suggest metrics to guide camera placement for better reconstruction quality. ActiveRMAP (Zhan et al., 2022) integrates NeRF with active vision tasks using RGB-only data in a dual-stage optimization alternating NeRF reconstruction and planning.

These methods, although maybe confused as a competitor to our framework, differ in the problem setting. We solve for the scenario when there are already pre-captured frames available.

Frame Selection Optimization. Various strategies have been explored for optimizing frame selection: Cerkezi et al. (Cerkezi and Favaro, 2024) and PC-NBV (Zeng et al., 2020) use object-centric sampling and point clouds for efficient NBV selection. Isler et al. (Isler et al., 2016) use information gain for NBV selection. Zaenker et al. (Zaenker et al., 2021) maximize the Region of Interest (ROI) using an Octree structure.

While our enhanced baseline uses point clouds, which is inspired by PC-NBV, our main approach ConMax3D, makes use of high level concepts in images which is not used in these techniques.

Ensemble and Surrogate Objectives. Some methods employ ensemble techniques or surrogate objectives: Density-aware NeRF Ensembles (Sünderhauf et al., 2023) uses NeRF ensembles to quantify uncertainty in reconstruction using ray termination probabilities. SO-NeRF (Lee et al., 2023) employs surro-

gate objectives such as surface coverage and geometric complexity to measure view quality.

While these methods provide valuable insights into improving NeRF quality and efficiency, they differ significantly from our concept-based frame selection strategy. In summary, while each of these approaches contributes uniquely to the field of 3D reconstruction, our method specifically targets the problem of frame selection by leveraging image segmentation and clustering to maximize conceptual diversity. This not only improves the interpretability and relevance of selected frames but also remains independent of camera poses and the reconstruction model used, making it highly adaptable to various 3D reconstruction models available.

3 FRAMEWORK OVERVIEW

In this section, we introduce our primary contribution: **ConMax3D** (Frame selection through Concept Maximization for 3D Reconstruction), an innovative framework for frame selection in multiview 3D reconstruction (see Figure 1). Additionally, we present an enhanced baseline approach: **Point Cloud Maximization for Frame Selection** (see Figure 2). While ConMax3D leverages segmentation masks and clustering to identify concepts for optimal frame selection, the Point Cloud Maximization approach utilizes dense reconstruction techniques.

3.1 ConMax3D Framework

As illustrated in Figure 1, our ConMax3D framework operates through a series of carefully designed steps. Initially, it segments the input images and embeds the resulting sub-images. These embeddings are then clustered to identify high-level concepts within the scene. Subsequently, a frame-concept graph is constructed, enabling the selection of the optimal k frames through an influence maximization approach, guided by the Utility function defined in Equation 1.

3.1.1 Image Segmentation

A critical step in optimizing frame selection for 3D reconstruction is the identification and prioritization of the most informative image regions. We achieve this through a image segmentation that delineates distinct objects and regions within each image. This divides images into semantically meaningful segments, each represented by a mask - a binary or multi-class image that precisely delineates regions of interest.

Our approach employs state-of-the-art segmentation techniques to process a diverse set of RGB images captured from multiple viewpoints. While we primarily utilize the Segment-Anything Model (SAM) (Kirillov et al., 2023), our framework is flexible and can accommodate other advanced models such as[(Yang et al., 2024), (Wu et al., 2024a)]. SAM has many parameters, such as predicted IOU and number of points in grid, that can be set at inference time. By filtering the masks through a threshold predicted IOU, which tells us the confidence score of the predicted mask, we can adjust the conservativeness of the segmentation process, allowing for optimal adaptation to variety of different images.

The segmentation process yields a rich set of sub-images, each corresponding to a unique object or region within the original image. Examples of such segmentations are shown later in Section 4, Figure 3. The segmentation masks are used subsequently in clustering and frame selection steps, ensuring that the most salient and informative image components are leveraged for 3D reconstruction.

3.1.2 Embedding Generation

To enhance computational efficiency, we crop and downscale the segments derived from the previous step. After that, we embed these sub-images using a CNN model. The generation of embeddings for these segments is a crucial process, as it transforms the rich visual information of segmented regions into compact, numerically represented feature vectors. For this task, we leverage the pre-trained EfficientNet architecture (Tan and Le, 2019), chosen for its optimal balance of accuracy and efficiency. However, our framework's flexibility allows for the integration of other state-of-the-art vision models, such as ResNet (He et al., 2016) or CLIP (Radford et al., 2021). We compared the pairwise distances of embeddings generated by Resnet18 and EfficientNet and plotted them as histogram as shown in Figure 4. Since EfficientNet embeddings are better separated (i.e., distribution of pairwise distances have higher variance), they yield better results in clustering, and in general for our framework. We extract the segments through element-wise multiplication of the RGB image with corresponding binary masks. This process ensures uniform segment sizes, enabling efficient batch processing for EfficientNet on GPU hardware for rapid inference. The resulting embeddings encapsulate the essential features of each segment, which is important for clustering. In the next phase, these embeddings facilitate the grouping of segments into semantically coherent clusters based on their distinct visual attributes.

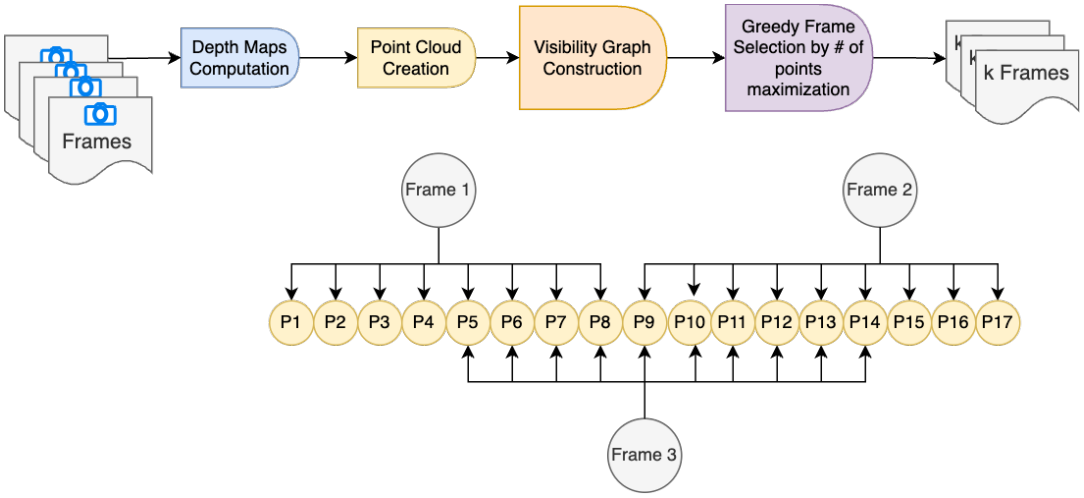


Figure 2: We propose an advanced baseline, Point Cloud Maximization, which computes the visibility graph using stereo matching and then uses a greedy approach to select the best k frames that maximize the number of unique points in the visibility graph.

3.1.3 Concept Generation

The derived embeddings undergo a clustering process utilizing the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBScan) algorithm (McInnes et al., 2017). HDBScan, an evolution of the DBSCAN algorithm (Ester et al., 1996), introduces a hierarchical framework that excels in handling varying cluster densities. The high dimensional image embeddings do not guarantee any specific cluster shape or uniform densities across clusters. So, the HDBScan approach is suitable for our application.

Our implementation of HDBScan is fine-tuned through critical hyperparameters, including minimum cluster size and minimum samples. These parameters allow us to control the granularity of clustering and the algorithm's robustness to noise, ensuring optimal performance across diverse visual scenarios. This clustering process aggregates similar pixel patterns from multiple images into cohesive groups, which we term "concepts." As illustrated in Figure 7, these concepts often correspond to human-interpretable object parts, bridging the gap between low-level visual features and high-level semantic understanding.

3.1.4 Concept Maximization

To identify the most informative frames, we formulate the selection process as an influence maximization problem within a bipartite graph structure (see Figure 1). In this graph, edges connect images to their corresponding concepts, enabling us to maximize concept diversity within the prescribed frame budget k.

Given the combinatorial complexity of selecting the optimal subset, we employ the following strategy: Our frame selection process iteratively identifies and selects frames that maximize the overall concept coverage through a greedy algorithm, detailed in Algorithm 1.

Our algorithm initializes with an empty set of selected frames S. For each candidate image, we establish its concept connections and identify the specific pixels, delineated by segmentation masks, that correspond to each associated concept. We introduce a utility function $U(S,i)$ as given by the equation 1 that measures the contribution of a candidate frame i to the set of already-selected frames S, in terms of the number of new concept-pixels it introduces.

$$U(S,i) = \left| \bigcup_{c \in C(i)} \left(P(i,c) \setminus \bigcup_{s \in S} P(s,c) \right) \right| \quad (1)$$

Here, $C(i)$ represents the set of concepts present in frame i, while $P(i,c)$ denotes the pixels in frame i associated with a specific concept c. Additionally, $\bigcup_{s \in S} P(s,c)$ refers to the set of pixels already covered by the selected frames in S for the concept c.

$U(S,i)$ follows the property of submodularity and monotonicity as shown in the following analysis.

Monotonicity

If $S \subseteq T$, then for any frame i and concept c,

$$\bigcup_{s \in S} P(s,c) \subseteq \bigcup_{t \in T} P(t,c). \quad (2)$$

This implies that removing the pixels already covered $(P(s,c))$ leaves at least as many unique pixels when S

is smaller. Thus,

$$U(S,i) \geq U(T,i), \tag{3}$$

proving monotonicity.

Submodularity

For any $S \subseteq T$ and $i \notin T$, adding frame i to the set S introduces at least as many new pixels as adding i to the larger set T. This is because T already covers all the pixels that S does, and possibly more. Formally, the following inequality is always satisfied:

$$U(S,i) \geq \left| \bigcup_{c \in C(i)} \left(P(i,c) \setminus \bigcup_{t \in T} P(t,c) \right) \right| = U(T,i) \tag{4}$$

This demonstrates the diminishing returns property of the utility function, thereby proving that it is submodular.

For such functions that are both monotonous and submodular, it can be proven that greedy algorithm gives near-optimal results with an approximation ratio of at least $1 - 1/e$ (Nemhauser and Wolsey, 1981).

The first image is also selected based on $U(S,i)$, translating to selecting the image with maximum number and size of semantically recognizable segments. Then the selection process proceeds iteratively in a greedy manner, again maximizing $U(S,i)$ at each step. The algorithm terminates upon reaching the desired frame count k or exhausting the available image pool, thereby constructing a subset of frames that optimally captures the scene's conceptual richness.

Algorithm 1: ConMax3D.

Input:
$C(i)$: Set of concepts connected to image i
$P(i,c)$: Set of pixels for image i under concept c
$U(S,i) = \left| \bigcup_{c \in C(i)} (P(i,c) \setminus \bigcup_{s \in S} P(s,c)) \right|$
Initialize: $S \leftarrow \emptyset$
while $|S| < k$ ***and*** $I \setminus S \neq \emptyset$ **do**
  Choose i from $I \setminus S$ that maximizes $U(S,i)$:
  $i \leftarrow \arg\max_{i' \in I \setminus S} U(S,i')$
  Update $S \leftarrow S \cup \{i\}$
end
return S

3.2 Point Cloud Maximization

We introduce an advanced baseline for Point Cloud Maximization that aims to optimize the capture of unique 3D points. This approach, illustrated in Figure 2, leverages depth maps and camera pose information to achieve superior results.

This framework operates as follows:

1. **Depth Map Computation:** Depth maps are generated for each image in the multiview dataset.
2. **Dense Point Cloud Reconstruction:** Utilizing the depth maps in conjunction with camera pose information, a dense 3D point cloud representation of the scene is reconstructed.
3. **Visibility Graph Construction:** A comprehensive visibility graph is established, linking each point in the reconstructed cloud to its corresponding source images.
4. **Greedy Frame Selection:** Finally, a greedy algorithm is employed, maximizing the number of unique points captured, in a manner analogous to our ConMax3D approach.

The details of this method are shown in Algorithm 2.

Algorithm 2: Point Cloud Maximization.

Input: Total number of frames k, mapping of images to points image2points
Output: Set of selected frames S
Initialize: $S \leftarrow \emptyset$
for *iteration* $= 1$ **to** k **do**
  max_union $\leftarrow 0$
  max_union_idx $\leftarrow -1$
  for j, *points* ***in*** *enumerate(image2points)* **do**
    if $j \notin S$ **then**
      union $\leftarrow$ |set(points) $\cup$ ($\bigcup_{s \in S}$ set(image2points[s]))|
      if *union* $>$ *max_union* **then**
        max_union $\leftarrow$ union
        max_union_idx $\leftarrow j$
      end
    end
  end
  $S \leftarrow S \cup \{$max_union_idx$\}$
end
return S

3.3 Comparative Baseline Approaches

To rigorously evaluate the efficacy of our proposed methods, we implement and assess two additional frame selection algorithms that serve as important baselines. First, we employ a stochastic frame selection process, randomly selecting k frames from a dataset containing N total frames. This method provides a crucial lower bound for performance evaluation.

As a more advanced baseline, we implement the Furthest View Sampling (FVS) algorithm (Eldar et al., 1997), which employs a positionally informed selection strategy. FVS begins by randomly sampling the first frame, then iteratively selects subsequent frames based on their maximal distance from

the currently selected set, using the minmax criterion until the desired number of frames k is reached. FVS aims to maximize the spatial diversity of selected camera positions, capturing a comprehensive range of perspectives of the 3D scene.

By including these baselines and ActiveNeRF, which is the state of the art, in our evaluation, we provide a comprehensive comparison that highlights the advancements and unique strengths of our proposed ConMax3D and Point Cloud Maximization methods.

3.4 Metrics

We use the standard metrics to assess the quality of 3D multiview 3D Reconstructions. Peak Signal to Noise Ratio (PSNR) is used to assess the pixel level accuracy in reconstructions. PSNR quantitatively measures the ratio of maximum signal power to the noise affecting the signal, providing a convenient numerical reference for how closely a reconstructed image matches the ground truth in terms of pixel-level fidelity.

Structured Similarity Index Metric (SSIM) (Wang et al., 2004) goes beyond this pixel-level comparison by modeling the perceived change in structural information, luminance, and contrast, aligning better with human visual perception.

Meanwhile, Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) takes advantage of deep neural network features trained to mimic human judgments of image similarity, offering a more perceptual measure of quality. By jointly reporting PSNR, SSIM, and LPIPS, we capture complementary aspects of reconstruction quality ranging from low-level pixel fidelity to high-level perceptual resemblance.

4 EXPERIMENTS AND RESULTS

For our 3D reconstruction evaluations, we employed the vanilla Neural Radiance Field (NeRF) model and 3D Gaussian Splatting (3DGS), using the LLFF dataset (Mildenhall et al., 2019). This dataset provides eight diverse realistic scenes with two configurations: spherical and non-spherical. Our objective is to select k frames from N available frames, where k is the budget and N is the total number of frames in the scene. We used metrics PSNR, SSIM, and LPIPS (Zhang et al., 2018) to assess reconstruction quality.

Our NeRF model was trained for 50,000 epochs using the images selected by the respective frame selection algorithms. Additionally, we trained 3D Gaussian Splatting for 30,000 epochs using the `gsplat` library (Ye and Kanazawa, 2023) for the same images. The remaining images were used as test images to evaluate the model. For comparison, the Random Sampling and Furthest View Sampling (FVS) methods were also executed. ActiveNeRF (Pan et al., 2022) was included for comparison, with results taken from the literature. For that reason the results of ActiveNeRF are omitted for non-spherical configuration.

As mentioned before, by utilizing both NeRF and 3D Gaussian Splatting in our experiments, we demonstrate that our frame selection methods are model-agnostic. This independence from specific reconstruction models enhances the generalizability and broad applicability of our proposed techniques across various 3D reconstruction paradigms.

4.1 Experimental Setup and Methods

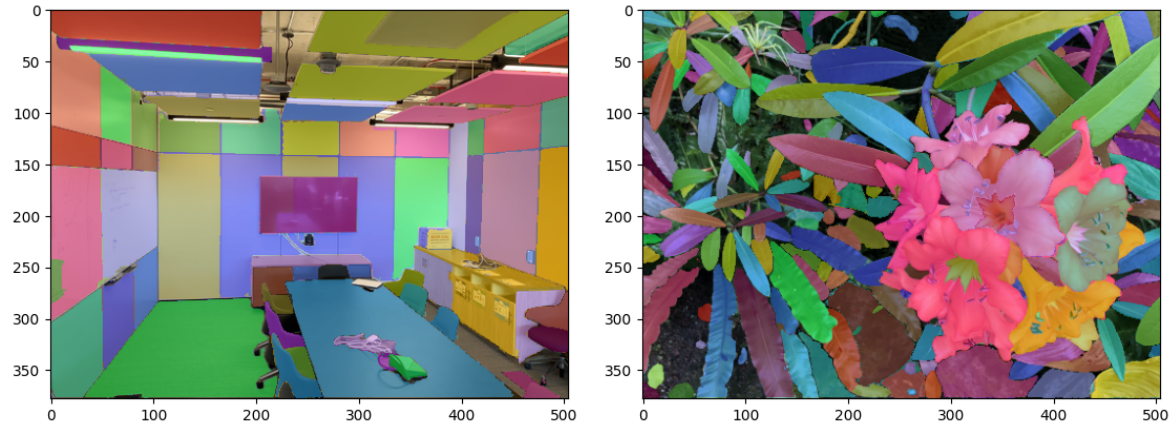
Our experimental protocol was designed to rigorously evaluate the proposed frame selection methods across various conditions. We utilized the LLFF dataset, downsampling the images to a resolution of 378×504 pixels to balance computational efficiency with the preservation of salient features.

4.1.1 ConMax3D Framework

Our ConMax3D Framework incorporated several key steps. On a dataset of around 50 images, the entire pipeline takes approximately 15 minutes to run on a single GPU. We began with image segmentation using the Segment-Anything Model (SAM) (Kirillov et al., 2023), setting the predicted IOU threshold to 0.8 to strike a balance between segmentation granularity and robustness. The resulting segmented regions were then cropped and downscaled by a factor of 4 to enhance computational efficiency.

To mitigate the impact of noisy masks generated by SAM, we implement several strategies:
1. We remove small masks that contain fewer pixels than the square root of the product of the image's height and width $\sqrt{(H * W)}$.
2. We set the prediction IOU to 0.8 for SAM, which is not too low to avoid noisy masks.
3. We exclude outlier clusters identified by HDBScan from the Image-Concept Graph. These outliers typically contain 20-40% of the masks and do not fit well within any established cluster, reflecting their noise-dominated nature.

For embedding generation, we utilized the EfficientNet model (Tan and Le, 2019) to create compact, information-rich representations of the processed segments. These embeddings were then clustered using



(a) Room Scene (b) Flower scene

Figure 3: In this figure we show the effectiveness of the SAM in segmenting different kinds of images. The individual segments of different images in a scene are clustered into concepts, which are then used for frame selection in ConMax3D framework.

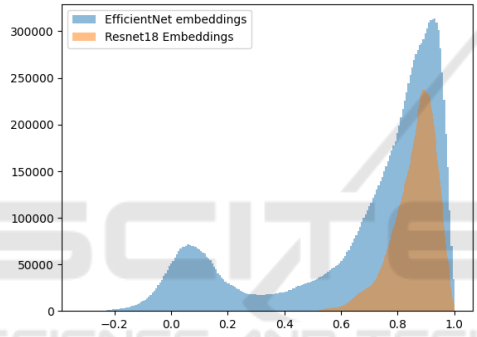


Figure 4: Pairwise distances of embeddings generated by EfficientNet and Resnet are shown as histogram plots respectively. The Efficientnet embeddings have a higher variance, which can be interepreted as the embeddings having better ”separation” in the embedding space, leading to better clustering.

the HDBScan algorithm to group similar pixel patterns across multiple images into “concepts.” We dynamically set the minimum cluster size to $N/4$, where N is the total number of frames, allowing the clustering to adapt to the dataset’s scale. To maintain cluster quality and reduce noise, we discard outlier clusters.

The relationships between images and their associated concepts were then modeled as a graph structure. We applied our greedy frame selection algorithm, as detailed in Algorithm 1, to this graph to maximize concept diversity in the selected subset of frames.

4.1.2 Point Cloud Maximization

For our Point Cloud Maximization approach, we leveraged COLMAP (Schonberger and Frahm, 2016) to compute depth maps and construct dense point clouds. From these point clouds, we derived a visibility graph that established connections between points the images from which they are visible. We then implemented a greedy algorithm, as outlined in Algorithm 2, to maximize the selection of unique points, thereby optimizing for comprehensive scene coverage. On a dataset of around 50 images, the entire pipeline takes approximately 1 hour to run on a single GPU.

4.2 Results

The results, averaged over eight scenes from the LLFF dataset, are summarized in Table 1 for both spherical and non-spherical cases using NeRF and 3D Gaussian Splatting (3DGS) respectively.

The results of ConMax3D are shown in bold. This method demonstrates significant improvements over the baselines in PSNR for a 10-frame budget in the non-spherical case, with gains up to 43.65% across various scenes. The highest gain is seen in the *room* scene, possibly due to well-detected segments and a larger image set (Please refer to the project github for scene-wise comparison statistics). This indicates that ConMax3D may be particularly suitable for selecting best frames for indoor scene reconstruction. The improvement in the spherical case is similar as can be seen in the table.

In non-spherical setting, which is closer to real world captures, ConMax3D consistently outperforms other methods for both NeRF and 3DGS. In the spherical setting, it outperforms all other methods for NeRF and for one setting (25 frames) with 3D Gaussian Splatting. Notably, the performance gains are observed with ConMax3D except in one case

Table 1: Performance Comparison of Frame Selection Methods on Spherical and Non-Spherical Data (LLFF).

Setting	Method	Spherical Data			Non-Spherical Data		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
10 Frames	Random + NeRF	15.55	0.38	0.36	15.83	0.39	0.38
	FVS + NeRF	20.39	0.59	0.22	17.95	0.50	0.31
	ActiveNeRF-BE + NeRF	18.67	0.45	0.37	-	-	-
	ActiveNeRF-CL + NeRF	20.14	0.66	0.33	-	-	-
	Point Cloud Maximization + NeRF	24.35	0.79	0.15	22.49	0.71	0.23
	ConMax3D + NeRF	**25.60**	**0.81**	**0.13**	**24.63**	**0.79**	**0.14**
	Random + 3DGS	22.46	0.82	0.36	22.33	0.76	0.17
	FVS + 3DGS	22.46	0.83	0.35	23.05	0.78	0.16
	ConMax3D + 3DGS	21.21	0.77	0.40	23.44	0.78	0.15
20 Frames	Random + NeRF	13.98	0.30	0.39	16.46	0.41	0.33
	FVS + NeRF	21.67	0.64	0.19	19.21	0.55	0.26
	ActiveNeRF-BE + NeRF	21.86	0.64	0.30	-	-	-
	ActiveNeRF-CL + NeRF	23.12	0.77	0.29	-	-	-
	Point Cloud Maximization + NeRF	25.70	0.82	0.13	25.74	0.82	0.13
	ConMax3D + NeRF	**25.82**	**0.83**	**0.12**	**27.48**	**0.85**	**0.11**
	Random + 3DGS	24.21	0.86	0.37	25.74	0.85	0.11
	FVS + 3DGS	25.85	0.89	0.34	26.31	0.86	0.11
	ConMax3D + 3DGS	22.98	0.80	0.40	26.50	0.86	0.10
25 Frames	Random + NeRF	26.61	0.83	0.12	26.87	0.85	0.11
	FVS + NeRF	27.33	0.86	0.11	27.42	0.86	0.10
	Point Cloud Maximization + NeRF	26.66	0.84	0.12	26.64	0.83	0.12
	ConMax3D + NeRF	**27.43**	**0.86**	**0.10**	**27.52**	**0.86**	**0.10**
	Random + 3DGS	22.17	0.79	0.42	26.30	0.87	0.10
	FVS + 3DGS	23.39	0.80	0.39	26.85	0.87	0.10
	ConMax3D + 3DGS	23.41	0.81	0.40	26.99	0.87	0.09

FVS+3DGS for 20 frames in the spherical setting (shown in red in Table 1) where FVS used with 3DGS gives better reconstruction quality. This maybe due to some variations due to the stochasticity of the methods used.

As the number of frames increases, the performance differences between ConMax3D and other methods becomes less pronounced. These results underscore ConMax3D's superior ability to maximize conceptual diversity and enhance 3D reconstruction quality, especially when the frame budget is limited and especially when used with NeRF.

The Point Cloud Maximization approach also proved to be a robust method, particularly advantageous when depth data is available or dense reconstruction can be easily done. These findings demonstrate the potential of our proposed frameworks to advance the state-of-the-art in frame selection for 3D reconstruction, regardless of the underlying reconstruction method used.

Explainability. Using the concept-image graph, we can calculate a variance map, as shown in Figure 5, which is the difference between the selected views and the candidate view according to equation 1. This variance map allows us to pinpoint exactly which concepts and to what extent they are covered in the current selection. Additionally, we can visualize which masks from different images are clustered as concepts, as shown in Figure 7.

5 DISCUSSION AND CONCLUSION

While existing 3D reconstruction models such as NeRF and 3D Gaussian Splatting operate on pixel-level information, human perception is fundamentally concept-based. This work bridges this gap between low-level pixel processing and high-level concept understanding through the ConMax3D approach.

Our method identifies concepts in images and selects optimal frames using a utility function proposed in Equation 1, which evaluates pixel shifts within concepts while prioritizing diverse concepts with larger coverage. This approach not only achieves high reconstruction quality but also provides interpretable insights into why certain images and camera positions result in better or worse rendering outcomes.

While the framework is promising, it has certain limitations in its current form. The SAM-based mask generation serves as a computational bottleneck, particularly for high-resolution images, though faster variants such as (Zhang et al., 2023) could potentially address this. The current HDBScan clustering approach requires pre-computed embeddings and

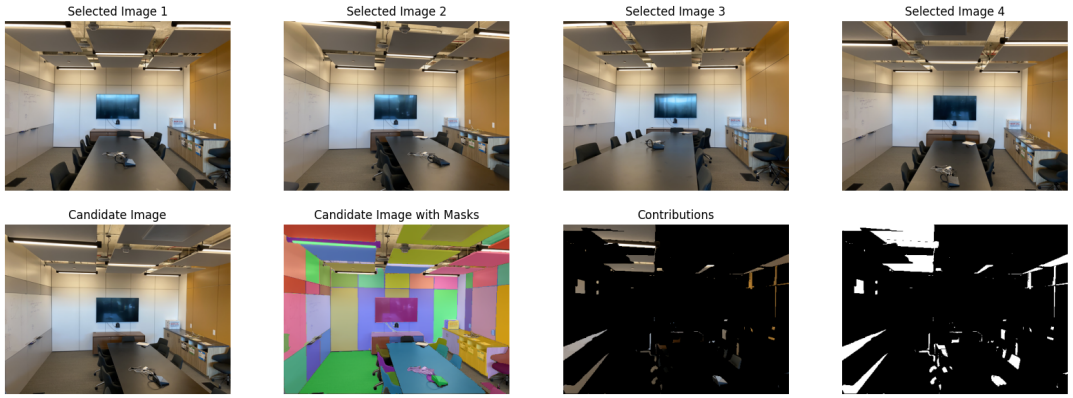


Figure 5: In the top row are shown the frames which are already selected. In the bottom row, the first image from the left is a candidate view, the second is the masks overlay on that candidate view, the third image shows the ”difference” between the selected frames and the candidate view as per the Utility function proposed, and the fourth image is the black and white version of the third to visualize the ”difference”, or the contributions of the candidate view, better.

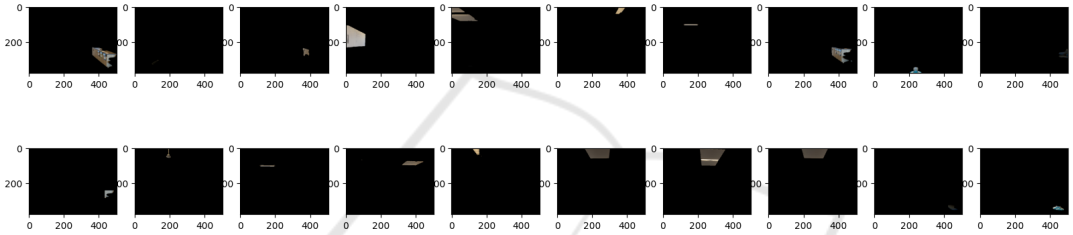


Figure 6: In the clustering step, the HDBScan algorithm classifies some masks (generated by SAM) as outliers. Some examples are shown in this figure. For the Image-Concept graph creation, these outliers are ignored.

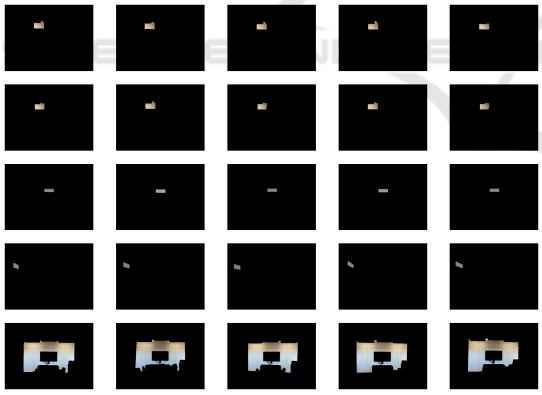


Figure 7: Concept (cluster) examples generated by SAM + HDBScan. In the figure, each row represents a concept and columns represent examples of different masks classified as that concept. Similar masks are grouped into the same clusters (concepts).

generates an outlier cluster of concepts which is ignored in the selection process, potentially problematic when the number of outliers is significant. Deep learning-based clustering algorithms such as (Asano et al., 2019) could potentially overcome these clustering limitations. Furthermore, the greedy selection algorithm provides an approximate solution, which is suboptimal, with a guaranteed approximation ratio of only 0.632 for the maximal coverage problem. While dynamic programming could provide optimal solutions, its prohibitive space complexity makes it impractical for large-scale problems. Graph Neural Networks offer a promising direction for approximating dynamic programming solutions while maintaining scalability (Dudzik and Veličković, 2022). Beyond addressing these limitations, future work would also focus on numerically relating the PSNR and related metrics to the variance map (shown in Figure 5).

Also we compared our frameworks with Furthest View Sampling (FVS), which uses exact camera positions in the benchmarks, and our method does not use camera positions at all. In the real world scenarios, often we have noisy camera poses. It would be interesting to see if ConMax3D benefit from such poses and how will the noise affect FVS.

To conclude, we demonstrate that for high-quality 3D reconstruction in models such as NeRF and 3DGS, considering conceptual diversity and coverage is sufficient for optimal frame selection. This finding not only simplifies the frame selection process but also aligns it with human visual understanding of the scene. Through this conceptual framework, we

provide both better reconstruction quality and interpretable insights into the reconstruction process.

Supplementary Material. For more plots and scene-wise comparisons, please refer to the following github repository:
https://github.com/akashjorss/Con3DMax.

REFERENCES

Asano, Y. M., Rupprecht, C., and Vedaldi, A. (2019). Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*.

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., and Srinivasan, P. P. (2021). Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864.

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., and Hedman, P. (2023). Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705.

Cerkezi, L. and Favaro, P. (2024). Sparse 3d reconstruction via object-centric ray sampling. In *2024 International Conference on 3D Vision (3DV)*, pages 432–441. IEEE.

Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.-J., and Cai, J. (2025). Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer.

Dudzik, A. J. and Veličković, P. (2022). Graph neural networks are dynamic programmers. *Advances in neural information processing systems*, 35:20635–20647.

Eldar, Y., Lindenbaum, M., Porat, M., and Zeevi, Y. Y. (1997). The farthest point strategy for progressive image sampling. *IEEE transactions on image processing*, 6(9):1305–1315.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Fan, Z., Cong, W., Wen, K., Wang, K., Zhang, J., Ding, X., Xu, D., Ivanovic, B., Pavone, M., Pavlakos, G., et al. (2024). Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*.

Fridovich-Keil, S., Meanti, G., Warburg, F. R., Recht, B., and Kanazawa, A. (2023). K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488.

Gao, K., Gao, Y., He, H., Lu, D., Xu, L., and Li, J. (2022). Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*.

Goli, L., Reading, C., Sellán, S., Jacobson, A., and Tagliasacchi, A. (2024). Bayes' rays: Uncertainty quantification for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20061–20070.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Isler, S., Sabzevari, R., Delmerico, J., and Scaramuzza, D. (2016). An information gain formulation for active volumetric 3d reconstruction. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3477–3484. IEEE.

Jain, A., Tancik, M., and Abbeel, P. (2021). Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894.

Jin, L., Chen, X., Rückin, J., and Popović, M. (2023). Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11305–11312. IEEE.

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Kopanas, G. and Drettakis, G. (2023). Improving nerf quality by progressive camera placement for free-viewpoint navigation.

Lai, X., Yue, D., Hao, J.-K., Glover, F., and Lü, Z. (2023). Iterated dynamic neighborhood search for packing equal circles on a sphere. *Computers & Operations Research*, 151:106121.

Lee, K., Gupta, S., Kim, S., Makwana, B., Chen, C., and Feng, C. (2023). So-nerf: Active view planning for nerf using surrogate objectives. *arXiv preprint arXiv:2312.03266*.

Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., and Sheikh, Y. (2019). Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*.

Marza, P., Matignon, L., Simonin, O., Batra, D., Wolf, C., and Chaplot, D. S. (2024). Autonerf: Training implicit scene representations with autonomous agents. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13442–13449. IEEE.

McInnes, L., Healy, J., Astels, S., et al. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Mildenhall, B., Srinivasan, P. P., Ortiz-Cayon, R., Kalantari, N. K., Ramamoorthi, R., Ng, R., and Kar, A. (2019). Local light field fusion: Practical view synthesis with

prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106.

Müller, T., Evans, A., Schied, C., and Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15.

Nemhauser, G. L. and Wolsey, L. A. (1981). Maximizing submodular set functions: formulations and analysis of algorithms. In *North-Holland Mathematics Studies*, volume 59, pages 279–301. Elsevier.

Niemeyer, M., Barron, J. T., Mildenhall, B., Sajjadi, M. S., Geiger, A., and Radwan, N. (2022). Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490.

Orsingher, M., Dell'Eva, A., Zani, P., Medici, P., and Bertozzi, M. (2023). Informative rays selection for few-shot neural radiance fields. *arXiv preprint arXiv:2312.17561*.

Pan, S., Jin, L., Hu, H., Popović, M., and Bennewitz, M. (2024). How many views are needed to reconstruct an unknown object using nerf? In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12470–12476. IEEE.

Pan, X., Lai, Z., Song, S., and Huang, G. (2022). Activenerf: Learning where to see with uncertainty estimation. In *European Conference on Computer Vision*, pages 230–246. Springer.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ran, Y., Zeng, J., He, S., Chen, J., Li, L., Chen, Y., Lee, G., and Ye, Q. (2023). Neurar: Neural uncertainty for autonomous 3d reconstruction with implicit neural representations. *IEEE Robotics and Automation Letters*, 8(2):1125–1132.

Schonberger, J. L. and Frahm, J.-M. (2016). Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 519–528. IEEE.

Smith, E. J., Drozdzal, M., Nowrouzezahrai, D., Meger, D., and Romero-Soriano, A. (2022). Uncertainty-driven active vision for implicit scene reconstruction. *arXiv preprint arXiv:2210.00978*.

Sünderhauf, N., Abou-Chakra, J., and Miller, D. (2023). Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9370–9376. IEEE.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Wu, J., Jiang, Y., Liu, Q., Yuan, Z., Bai, X., and Bai, S. (2024a). General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3783–3795.

Wu, T., Yuan, Y.-J., Zhang, L.-X., Yang, J., Cao, Y.-P., Yan, L.-Q., and Gao, L. (2024b). Recent advances in 3d gaussian splatting. *Computational Visual Media*, 10(4):613–642.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., and Zhao, H. (2024). Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*.

Ye, V. and Kanazawa, A. (2023). Mathematical supplement for the `gsplat` library.

Yu, A., Ye, V., Tancik, M., and Kanazawa, A. (2021). pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587.

Zaenker, T., Smitt, C., McCool, C., and Bennewitz, M. (2021). Viewpoint planning for fruit size and position estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3271–3277. IEEE.

Zeng, R., Zhao, W., and Liu, Y.-J. (2020). Pc-nbv: A point cloud based deep network for efficient next best view planning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7050–7057. IEEE.

Zhan, H., Zheng, J., Xu, Y., Reid, I., and Rezatofighi, H. (2022). Activermap: Radiance field for active mapping and planning. *arXiv preprint arXiv:2211.12656*.

Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S., and Hong, C. S. (2023). Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.