SAID-NeRF: Segmentation-AIDed NeRF
for Depth Completion of Transparent Objects

Avinash Ummadisingu[1], Jongkeum Choi[2], Koki Yamane[3],
Shimpei Masuda[1], Naoki Fukaya[1], Kuniyuki Takahashi[1]

Abstract— Acquiring accurate depth information of transparent objects using off-the-shelf RGB-D cameras is a well-known challenge in Computer Vision and Robotics. Depth estimation/completion methods are typically employed and trained on datasets with quality depth labels acquired from either simulation, additional sensors or specialized data collection setups and known 3d models. However, acquiring reliable depth information for datasets at scale is not straightforward, limiting training scalability and generalization. Neural Radiance Fields (NeRFs) are learning-free approaches and have demonstrated wide success in novel view synthesis and shape recovery. However, heuristics and controlled environments (lights, backgrounds, etc) are often required to accurately capture specular surfaces. In this paper, we propose using Visual Foundation Models (VFMs) for segmentation in a zero-shot, label-free way to guide the NeRF reconstruction process for these objects via the simultaneous reconstruction of semantic fields and extensions to increase robustness. Our proposed method *Segmentation-AIDed NeRF* (SAID-NeRF) shows significant performance on depth completion datasets for transparent objects and robotic grasping.

I. INTRODUCTION

Acquiring accurate 3d information of transparent objects is a crucial but difficult challenge in Robotics and Computer Vision. These objects demonstrate several non-Lambertian properties leading to highly view-dependent appearance, if at all visible due to transparency, and specular effects like reflection, refraction, scattering, and dispersion complicating both visual perception and depth capture via RGB-D cameras. Alternative sensor modalities have been studied but have limitations due to cost, calibration and sensitivity [1].

Machine learning approaches can be utilized to estimate or complete noisy or missing depth values [2]–[5]. Deep Learning methods typically require large amounts of data, but the acquisition of accurate labels for depth is challenging due to the difficulties in accurately simulating them or complex labeling setups needed [2], [5], [6]. Furthermore, enhancing the out-of-distribution generalization capabilities of these methods remains a considerable hurdle.

A contrasting approach is the use of NeRF-based methods for 3d scene reconstruction from multiple views. NeRFs allow for a learning-free, per-scene optimization approach,

sidestepping issues related to training and generalization [7]–[9]. However, they are often limited by the quality and quantity of views supplied. This is particularly important for the sufficient capture of transparent object surfaces from different views and involves potentially inducing specular effects like reflections, by adding additional light sources [10]. While NeRFs demonstrate impressive visual reconstructions of transparent objects, the recovered depth is often too inaccurate/noisy to use in applications like robotic grasping.

In this work, we propose a method **S**egmentation-**AID**ed **NeRF** (SAID-NeRF) that exploits instance masks generated by Visual Foundation Models (VFMs) [11], [12] to enhance NeRF's estimation of surfaces of transparent objects that are otherwise ambiguous when relying purely on RGB (Fig. 1 (b)). Having been trained on related visual tasks on internet-scale datasets for generalization, VFMs encounter a large variety of transparent objects under numerous conditions that can be leveraged. In addition, for scenarios where objects in the scene are difficult to separate such as overlapping or cluttered objects, we present a simple heuristic to generate a hierarchy of non-overlapping masks for label-free use (Fig. 1 (c)). This allows the reconstruction of the objects in reasonable time (seconds) using a handful of views. Then, the depth information acquired using the proposed method is used by the robot for object grasping.

The novelty and contributions of this work are
- Utilization of instance masks from VFMs to guide surface density acquisition of transparent objects.
- Extensions of NeRF that enable robust and reliable estimation under difficult conditions in seconds.
- A method for the grouping of label-less object segments into a hierarchical set of non-overlapping masks for the label-free generation.
- Evaluation of the proposed method on a large-scale transparent object depth completion dataset.
- Evaluation of the proposed method through robotic grasping of several transparent objects under challenging settings.

II. RELATED WORK

A. Transparent Object Segmentation

Deep Learning models typically require large amounts of annotated data to learn, which is made challenging by the difficulty in acquiring ground truth masks and the diversity of appearance of transparent objects. The generation and use of synthetic data as an alternative is challenging due to the difficulty in accurate simulation of specular effects potentially

[1]A. Ummadisingu, S. Masuda, N. Fukaya, and K. Takahashi are with Preferred Networks, Inc. {ummavi, masuda, fukaya, takahashi}@preferred.jp [2]J. Choi is with the University of Tokyo. This work is done in a part-time job at Preferred Networks, Inc. choijongkeum@g.ecc.u-tokyo.ac.jp [3]K. Yamane is with the University of Tsukuba. This work is done in a part-time job at Preferred Networks, Inc. yamane.koki.td@alumni.tsukuba.ac.jp
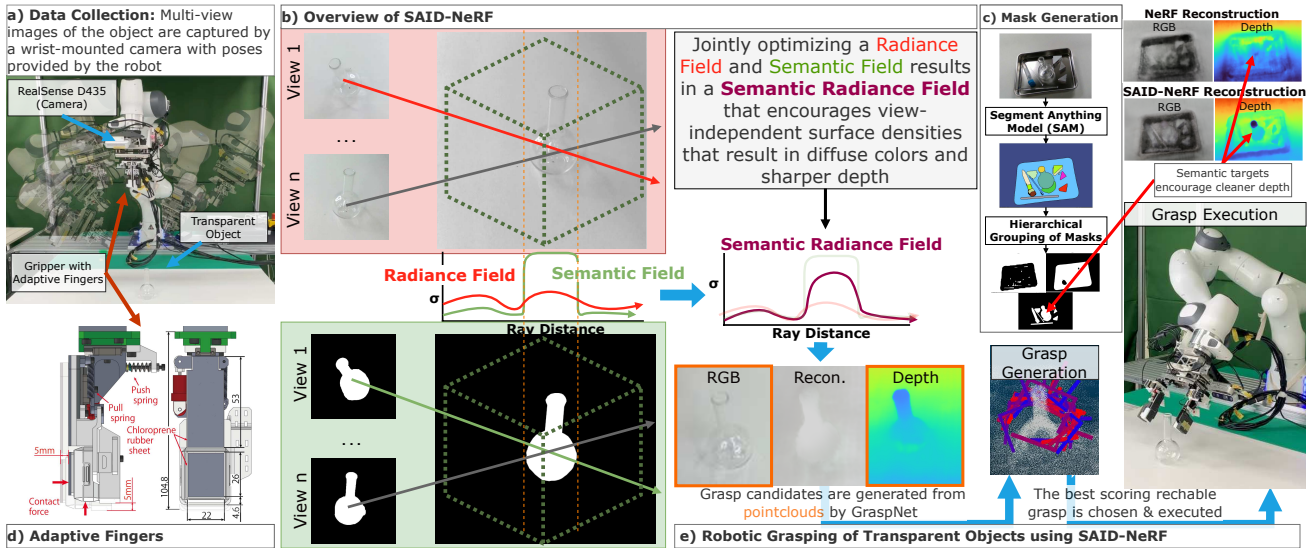
Fig. 1: Overview of the proposed system for robotic grasping using **S**egmentation-**AID**ed **NeRF** (SAID-NeRF) for depth estimation of transparent objects. SAID-NeRF uses Visual Foundation Models (VFMs) for zero-shot segmentation to guide the reconstruction process

inducing a sim2real gap. Several datasets for transparent objects have been proposed for segmentation [13], depth completion and grasping [2], [6]. However, they remain limited in the diversity of views, settings, and objects.

An alternative to methods trained on annotated data is the use of Visual Foundation Models (VFMs). These models are tried on vast amounts of data leveraging self-supervision, other pre-trained models and combine modalities like text to aid scaling and generalization. A VFM of particular interest to this work is the Segment Anything Model (SAM) [12], notable for its excellent zero-shot segmentation results. It is often combined with other methods to create powerful segmentation methods such as Grounded-SAM, Semantic-SAM [14], and SEEM [15].

In this study, we leverage SAM's zero-shot ability to segment transparent objects in wide contexts and conditions.

B. Learning-based Depth Completion

A class of prior works aims for the supervised completion of depth values trained entirely on synthetic data [4], [5], [10] or real data [2], [6] with depth ground truths often found by using 3d models or object markers [2]. A common issue with supervised learning approaches is their generalization gap or the struggles of these models to work on data outside of their training set. This is especially important for transparent objects since they demonstrate a wider range of environment-dependent appearance owing to their specular properties. For a more comprehensive review of the topic, we refer the reader to [1]. There also exists a wider range of monocular depth estimation methods that learn either relative or metric depth estimation [16] and are trained on a large number of scenes for generalization but do not typically offer the millimeter precision or high performance on transparent objects required for successful grasping.

C. Neural Radiance Fields (NeRFs)

NeRFs [7] have emerged as a promising 3d representation method to capture view-dependent appearance. Several stud-

ies have looked at better representations such as Multiresolution Hash Encoding [8] and to improve reconstruction quality and time [17]–[22]. Prior work has also explored using additional sources of supervision such as depth to reduce reconstruction time and views [23] including using depth predicted from trained models [24]–[27]. Others have considered using NeRFs for semantic labeling [28]–[32] and incorporating semantic information into NeRFs. Methods that include semantic labels to objects in the scene typically assume we have a pre-trained NeRF, which is then used to convert 2d semantic masks into 3d [28], [29] or have a focus on label propagation [30]. This can be used for downstream tasks such as object selection, editing, in-painting etc.

By construction, *radiance fields* aim to capture the appearance of the object without consideration of its true shape or depth. This problem is further exacerbated for transparent objects where surface specular effects may be perfectly captured without recovering the object's underlying geometry, rendering them unreliable for robotic tasks. Implicit Surface estimation methods like NeUS [9], [33] are better suited to this task but have longer training times, which is a bottleneck to tasks like robotic grasping. Our proposed method seeks to use the NeRF formulation with semantic information (a Semantic-NeRF) for better surface extraction aimed at robotic grasping, ignoring effects on visual quality.

D. Robotic Grasping with NeRF

Several works have explored using NeRFs for grasping transparent objects. Dex-NeRF [10] uses NeRFs to grasp transparent objects in the scene but required significant computation time in the order of hours. The introduction of Multiresolution Hash Encoding [8] significantly shortened reconstruction time and has driven follow-up work like Evo-NeRF [34]. GraspNeRF [35] uses a pre-trained feature extractor that is optimized jointly with a volumetric rendering loss (for RGB) and a trained volumetric grasp detection network VGN [36] to use grasp quality to guide detection of geometry and is trained end-to-end. Our approach, instead,

uses segmentation masks to overcome the view-dependent appearance of these objects enabling robotic grasping in a reasonable time using a handful of views. We provide a comparison to these methods and show the effectiveness of our proposed method in Section V-B.

III. METHOD

This section describes the proposed method SAID-NeRF for depth completion of transparent objects as well as its proposed use in a robotic grasping system. First, we describe the basics of NeRF and it's variants in the preliminaries (Section III-A) with a discussion of SAID-NeRF's purported working mechanism in the reconstruction of transparent objects (Section III-B) as well as proposed extensions to the architecture (Section III-C) and the methodology to generate hierarchical layers of semantic masks of objects in the scene (Section III-D). The use of SAID-NeRF in a robotic grasping system for the grasping of transparent objects is discussed in Section III-E.

A. Preliminary: Neural Radiance Fields (NeRFs)

NeRFs [7] use multilayer perceptions (MLPs) to represent a 3d volumetric scene with a 5d function of position $\mathbf{x} = (x, y, z)$ and viewing direction $\mathbf{d} = (\theta, \phi)$. Specifically, for a given 3d point, NeRF learns an implicit function $\mathbf{f}$ that estimates the density σ and RGB color $\mathbf{c}$ like $\mathbf{f}(\mathbf{x}, \mathbf{d}) = (\sigma(\mathbf{x}), \mathbf{c}(\mathbf{x}, \mathbf{d}))$. To compute the color of a single pixel, we generate a ray $\mathbf{r}$ originating at the center of its projection direction. The expected color $C(\mathbf{r})$ is calculated by integrating the implicit radiance field along this ray to compute the implicit radiance of any object within its path.

$$C(\mathbf{r}) = \int_{t_{near}}^{t_{far}} T(t)\sigma(t)\mathbf{c}(t, d)dt, \qquad (1)$$

where $T(t) = \exp\left(-\int_{t_{near}}^{t} \sigma(s)ds\right)$ and t_{near} and t_{far} are the near and far planes of the camera. NeRF approximates this integral using numerical quadrature along with techniques for efficient rendering [8].

Following [30], we augment the formulation by adding a segmentation rendering MLP before the viewing direction is appended for the RGB(Specular) component as shown in Fig. 2. The viewing direction itself is first projected using a Spherical Harmonics (SH) basis function before it is used. Concretely, we introduce a view-invariant function $\mathbf{s}$ that maps position $\mathbf{x}$ to a distribution over $\mathbf{C}$ semantic labels leading to

$$S(\mathbf{r}) = \int_{t_{near}}^{t_{far}} T(t)\sigma(t)\mathbf{s}(t)dt. \qquad (2)$$

Several extensions to NeRF aim to better capture specular effects by the decomposition of the color component $\mathbf{c}$ [21], [22]. To better encourage the capture of the underlying diffuse color of transparent objects, we follow [37] in decomposing appearance into view-independent diffuse color $c_d(\mathbf{x})$ and a view-dependent specular color $c_s(\mathbf{x}, \mathbf{d})$ with shallow MLPs with the final color obtained as a sum $c = c_s + c_d$ as shown in Fig. 2, with a small L2 penalty on the specular term to encourage diffuse color components.
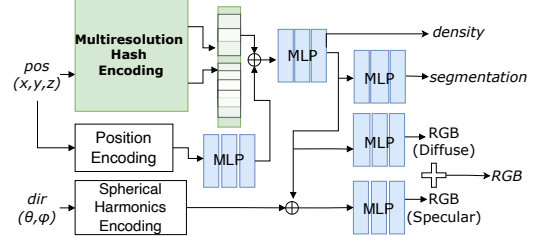

Fig. 2: Architecture diagram of the proposed SAID-NeRF

B. Semantic Fields as NeRF Density Grounding

This section describes how joint optimization of semantic components in Semantic NeRFs can help recover better surface densities for transparent objects. NeRF approximates eq. (1), and the expected color is given by

$$\hat{C} = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i \delta_i))\mathbf{c}_i, \qquad (3)$$

where $T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)$ and $\delta_i = t_{i+1} - t_i$ is the distance between two samples along the ray. In a similar vein to the color estimation, the depth map in NeRF can be approximated using the transmittance T_i and volume density σ_i at each depth sample along the ray:

$$\hat{D} = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i \delta_i))t_i. \qquad (4)$$

When a projected ray encounters an opaque object at some distance t_i, it has a high contribution to the estimated color of the pixel color $\hat{C}$, acquired by a high assignment of volume density σ_i at the point, therefore leading to high depth contribution $\hat{D}$ in eq. (4). However, when it encounters a transparent object, the contribution it has to the color of the pixel $\hat{C}$ is very low (visualized in the radiance field component of Fig. 1 (b)). Therefore, the assigned volume densities are much more diffuse until the ray passes through and perhaps encounters another opaque surface behind it. This means that the values of estimated depth for the ray $\hat{D}$ will be closer to the opaque object behind it, or worse, somewhere in between, as is expected with an expectation of a bimodal distribution.

In this work, we exploit this fact and leverage density assignments from the semantic head of the NeRF as a way to enforce higher surface densities to transparent objects in spite of the negative contribution they have to their color components. Visualized in the Semantic Radiance Field component of Fig. 1 (b), we see the visual effect of these density assignments in the diffuse color acquired by the transparent objects that were provided as masks, as a result of focus on view-independent, diffuse appearance. This phenomenon is also what allows the reduction in the number of views and environment modifications like light sources required by existing methods, such as Dex-NeRF [10].

C. Position Encoding

This section describes the addition of Frequency Embedding back into the architecture of a Multiresolution Hash Encoding-based NeRF for avoiding floaters and warped geometry for patternless surfaces.

Our proposed method is based on Instant-NGP [8]. Inspired by the work of [38], we find the addition of the frequency embedding, by concatenation with the features of the multiresolution hash encoding, significantly improves the quality of reconstructions. We hypothesize this to be an inductive bias towards preferring plane-like surface reconstructions due to positional encoding embedding regions with element-wise coordinates, loosely based on [33]. This is particularly stark when objects are placed on unremarkable backgrounds that are non-patterned or single-colored leading to inaccurate densities and so-called "floaters". These floaters have a particularly destructive effect on robotic grasping by confusing grasp generation and motion planning algorithms.

D. Detection & Semantic Mask Generation

This section describes the detection of all objects in a scene using VFMs in a zero-shot manner and ways to generate roughly consistent semantic masks without the need for class labels. To provide semantic targets to SAID-NeRF, an ordered set of **S** semantic classes that are consistent across views are required. Additionally, the segmentation method must be capable of separating transparent objects, both from their backgrounds and each other. We propose a way to use SAM to generate unlabelled instance segments of an image and then use a straightforward heuristic to group them into a hierarchical set of non-overlapping scene masks that are roughly consistent as visualized in Fig. 1 (c).

First, we make the following assumptions about the captured scene: 1) Some surface (e.g. a table) exists on which the object(s) of interest are placed. 2) Images capture this surface prominently (it occupies a significant part of the image) and consistently (present across multiple views). 3) Objects are placed on the surface directly or on top of/within each other, such as tools in a tray or water in a beaker.

A simple algorithm is proposed to group masks into non-overlapping hierarchies, based on checking for overlap between their convex hulls as membership into non-overlapping sets (reminiscent of a union-find algorithm) as described in Algorithm 1. We find that this algorithm, while straightforward, is quite imprecise and fails to handle several settings. However, empirically, we find that it provides reasonably consistent masks as long as the number of sets requested remains reasonably low (≤ 3) and faulty assignments are typically compensated for by other views.

E. Robotic Grasp Planning

This section describes the use of SAID-NeRF in a robotic system for the grasping of transparent objects. A typical vision-based pipeline consists of perception (via RGB-D cameras), completion/estimation of depth, grasp generation, and grasp execution. We use a wrist-mounted RGB-D camera to automate data collection from different angles. By using camera poses acquired from the robot's pose, we overcome the need to use COLMAP [39], and operate in the robot's coordinate system making inter-operation easy and eliminating discarding images for lack of feature matches.

Next, SAID-NeRF processes the collected images and pose information and reconstructs the scene in 3d. Depending


Algorithm 1 Hierarchical semantic mask generation
Input: M masks, c requested number of channels
Output: c masks of size M
1: Sort M in decreasing order of area
2: Compute convex hulls of each mask in M
3: Initialize an empty list for the mask sets, $Sets$
4: **for** each mask i in sorted list of masks **do**
5:     **for** each set s in $Sets$ **do**
6:         check if hull of i overlaps with hulls in set s
7:         **if** there is no conflict **then**
8:             assign mask i to set s
9:         **end if**
10:     **end for**
11:     **if** mask i couldn't be added to existing set **then**
12:         create new set with i and append it to $Sets$
13:     **end if**
14: **end for**
15: **if** $|\,Sets\,| > c$ **then**
16:     merge all masks in $s \in Sets_{i>c}$ into $Sets_c$
17: **end if**
18: **return** $Sets$
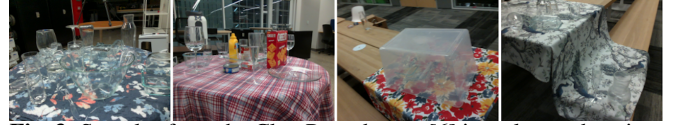
Fig. 3: Samples from the ClearPose dataset [6] in a cluttered setting, with opaque distractors, a translucent cover, and non-planar settings

on the format needed by the grasp generation (RGB-D, Point clouds, Mesh, Truncated Signed Distance Field, etc) the NeRF can flexibly generate the output. To generate point clouds, we accumulate rendered RGB and (estimated) depth values from the NeRF across training views. Point clouds are then used to generate potential grasp candidates using GraspNet [40], the highest-scoring reachable candidate of which is chosen and executed. While several grasp generation methods are applicable, we chose GraspNet [40] motivated by its large-scale training, aim of generalization, as well as its use in several recent works [2], [4], [41].

IV. EXPERIMENT: DEPTH COMPLETION DATASET

A. Experiment Setup

To evaluate our transparent object depth completion capabilities, we chose the ClearPose Transparent Object Dataset [6] that consists of 63 objects arranged in numerous configurations. The choice was motivated by the dataset's convenient categorization, its size, and the availability of pose information. Specifically, we use the provided "downscaled" variant of their dataset, which subsamples images in each set and scene. Some sample images from the dataset are shown in Fig. 3. The various test sets explore a variety of settings, and noticeably, some have as few as 14-23 views taken from a wide range of viewing angles. We show that this does not hinder the proposed method and can perform competitively even in these difficult sparse view conditions.

As baseline methods, we report the results of the depth completion method, TransCG [2], presented in the original work as a baseline. Additionally, we also benchmark the recently proposed depth completion method DREDS [4] trained on a mix of transparent, metallic, rough, and specular

objects. We note that we do not perform any fine-tuning of the methods on the training sets. To evaluate against multi-view NeRF-based approaches, we also benchmark against the *Nerfacto* [42] method which combines several recent advances in NeRF literature. Additionally, to evaluate the use of large-scale monocular depth prediction methods in combination with NeRF, we benchmark against a variant of Nerfacto called the *Depth-Nerfacto* which uses the monocular depth estimation method ZoeDepth [16] and a Ranking-based depth loss from SparseNerf [24] to compensate for the unknown scale and metric of depth values generated which prevent it from being usable directly.

B. Implementation Details

We implement SAID-NeRF using the Nerfacc [43] library for efficient sampling and use their proposal network estimator, first proposed in the Mip-NeRF 360 paper [44]. The base architecture is implemented as per Fig. 2 with the additional position encoding MLP of a single layer of width 16, while the segmentation MLP is two layers of width 32. The loss function for the RGB values (output as a sum of the diffuse and specular heads) is chosen to be the L1 loss. For segmentation targets, we generate three sets of masks (such that the segmentation head outputs a vector of length 3). For their quick generation, we use a "light" variant of HQ-SAM (Called Light-HQ SAM) [45]. The semantic labels are jointly trained along with RGB using a Binary cross-entropy loss. We find that depth supervision imposed on the expected depth values was vital for scenes where backgrounds did not possess significant patterns. We use an L2 loss imposed on the depth values obtained from the RGB-D camera (reminiscent of [23]) despite its inability to capture depth values for transparent objects.

We then allow the optimization to run on the dataset and cap the training steps to a specific time limit T, which is 20 seconds for the robotic experiments and 60 seconds for ClearPose. We note that this runs counter to the step-based counting that is usually reported but is more meaningful in an application environment, as the hyperparameters and scene can significantly impact the computation time of a single step. We run all the methods described here on an NVIDIA RTX4000 Ada GPU, except the SAID-NeRF results for the ClearPose Transparent Object Depth Completion dataset, which were run using an NVIDIA Tesla V100 GPU.

C. Ablation Study & Data Requirements

In this section, we ablate and evaluate the performance of each component of the proposed method described in Section III. We systematically remove semantic supervision, position encoding, and depth supervision to study their contribution. Additionally, we report a vanilla variant of Instant-NGP as a baseline without any of the proposed extensions, trained for $10\times$ longer. For evaluation, we choose the *Heavy Occlusion* subset of the ClearPose dataset [6].

Table I shows the results. We find that the components with the largest impact are depth supervision followed by position encoding. Both of these have significant regularizing

TABLE I: Ablation study for the various components proposed and discussed in Section III

Method	RMSE	MAE	REL
SAID-NeRF (Full) (ours)	**0.0843**	**0.0599**	**0.0698**
SAID-NeRF w/o Position Encoding	0.1890	0.1013	0.1202
SAID-NeRF w/o Semantic Info.	0.1308	0.0832	0.0964
SAID-NeRF w/o Diffuse/Specular	0.0921	0.0675	0.0788
SAID-NeRF w/o Depth Supervision	0.6328	0.6100	0.7854
Instant-NGP	0.2051	0.1754	0.2250

TABLE II: SAID-NeRF results on reducing the number of views for the **Heavy Occlusion** dataset

Views	RMSE	MAE	REL
All	0.0837	0.0586	0.0673
60	0.0828	0.0587	0.0669
30	0.0912	0.0655	0.0733
15	0.1092	0.0794	0.0873
8	0.1196	0.0883	0.0971
4	0.1298	0.0981	0.1088

TABLE III: SAID-NeRF results on reducing the NeRF reconstruction time for the **Heavy Occlusion** dataset (all views)

Train Time	RMSE	MAE	REL
10s	0.1025	0.0753	0.0884
15s	0.0993	0.0737	0.0846
30s	0.0877	0.0626	0.0721
60s	0.0828	0.0587	0.0669
90s	0.0842	0.0610	0.0702

effects on the geometry of the scene as a whole and make the difference between a coherent scene geometry. Semantic information and the Diffuse-Specular split, as expected, have a small effect since they are both extensions proposed to aid the capture and reconstruction of transparent objects and have little to do with opaque objects, which make up a big portion of the scene. Additionally, we find that the proposed method SAID-NeRF, with all its proposed extensions, can significantly outperform vanilla Instant-NGP using as few as four images (and evaluating on all the available views) or training for under 10 seconds (Table II and Table III).

D. Experiment Results

The results of the methods, including the proposed SAID-NeRF (ours), are presented in Table IV. The evaluation protocol and metrics closely follow that set by [5]: The Root Mean Squared Error in meters (RMSE), the median error relative to the depth (REL), and percentages of pixels with predicted depths falling within an interval ($[\delta = |predicted - true|/true]$ where $\delta \in \{1.05, 1.10, 1.25\}$. The error is averaged first across the scene and then within the dataset for the fairness of multi-view methods.

We see that the proposed SAID-NeRF outperforms TransCG, DREDS, and Nerfacto in most of the test sets. Both the "New background" and "Heavy occlusion" test sets have between 67-89 images, but the performance differences between them across all methods are remarkable. The "Filled liquid", "Opaque distractor", "Translucent cover" and "Non-planar" datasets only have between 14-23 images. Remarkably, we see that the performance of SAID-NeRF remains relatively consistent despite this drop in size by a factor of 3 and increased scene complexity. Although DREDS outperformed SAID-NeRF on the "Translucent cover" set, the difference is not significant and its remarkable performance across all sets is roughly consistent. This stability in performance may be highly sought after in applications that require a large degree of generalization. We also note that DREDS, which was trained not just for transparent objects but also for rough, specular, and shiny objects, is more stable throughout. A visualization of completed/estimated
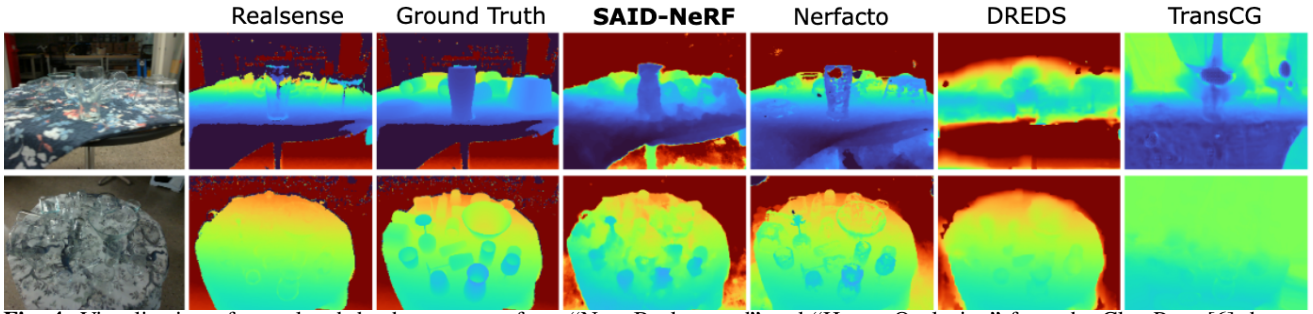
Fig. 4: Visualization of completed depth on scenes from "New Background" and "Heavy Occlusion" from the ClearPose [6] dataset

TABLE IV: Results on the test set of the ClearPose dataset. RMSE, REL, and MAE are reported in meters

	Method	RMSE	REL	MAE	$\delta_{1.05}$	$\delta_{1.10}$	$\delta_{1.25}$
New background	**SAID-NeRF(Ours)**	**0.0724**	**0.0609**	**0.0530**	54	**81**	**97**
	Nerfacto	1.3400	0.2200	0.2300	**63**	79	91
	Nerfacto (ZoeDepth)	0.6253	0.1860	0.1705	**63**	79	91
	TransCG	0.1577	0.1206	0.1177	31	53	81
	DREDS	0.1406	0.1236	0.1032	34	59	87
Heavy occlusion	**SAID-NeRF(Ours)**	**0.0843**	**0.0698**	**0.0599**	48	**76**	**96**
	Nerfacto	0.5800	0.2100	0.2200	50	62	72
	Nerfacto(ZoeDepth)	0.7160	0.2073	0.1952	**54**	71	84
	TransCG	0.1769	0.1371	0.1395	30	53	82
	DREDS	0.1770	0.1624	0.1281	26	47	81
Filled liquid	**SAID-NeRF(Ours)**	**0.0754**	**0.0709**	**0.0581**	46	**74**	**96**
	Nerfacto	0.3300	0.2300	0.2700	43	52	58
	Nerfacto(ZoeDepth)	0.4154	0.2797	0.2356	**47**	56	62
	TransCG	0.1096	0.1024	0.0867	34	59	88
	DREDS	0.1486	0.1387	0.1112	24	46	87
Opaque distractor	**SAID-NeRF(Ours)**	**0.1020**	**0.0890**	**0.0749**	**41**	**65**	**91**
	Nerfacto	0.5800	0.2800	0.2900	37	43	48
	Nerfacto(ZoeDepth)	0.6510	0.3214	0.2946	40	46	52
	TransCG	0.1815	0.1456	0.1347	29	47	75
	DREDS	0.1418	0.1211	0.0946	40	58	87
Translucent cover	**SAID-NeRF(Ours)**	0.1173	0.1152	0.0989	25	46	84
	Nerfacto	0.3800	0.3300	0.3700	19	26	37
	Nerfacto(ZoeDepth)	0.3222	0.3120	0.2650	21	29	45
	TransCG	0.1368	0.1278	0.1132	27	48	78
	DREDS	**0.1102**	**0.1007**	**0.0831**	**33**	**59**	**90**
Non planar	**SAID-NeRF (Ours)**	**0.1116**	**0.0936**	**0.0843**	38	**64**	**93**
	Nerfacto	0.3300	0.2600	0.2600	40	52	63
	Nerfacto(ZoeDepth)	0.3269	0.2164	0.2247	**45**	59	71
	TransCG	0.2193	0.1649	0.1805	27	45	74
	DREDS	0.1559	0.1347	0.1182	31	51	85

depth values from the methods on 3 scenes from the "New Background" and "Heavy Occlusion" datasets are presented in Fig. 4. Although DREDS benchmarked well (in terms of RMSE), from the visualization we find it did not accurately estimate object depths. The proposed method is able to accurately estimate the depth of objects, even transparent objects, without significant loss of shape.

V. EXPERIMENT: ROBOTIC GRASPING

To evaluate the applicability of SAID-NeRF as part of a robotic grasping system of transparent objects, we perform robotic experiments. First, we describe the setup of the experiment as well as present and discuss it's results.

A. Experiment Setup

Our robotic system, shown in Fig. 1 (e), consists of a Franka Emika Panda Arm 7-DoF robotic arm equipped with a custom-made parallel gripper operated by a servomotor (Dynamixel XM430-W350-R). A wrist-mounted Intel RealSense D435 is used to capture RGB-D images for data

collection. The parallel gripper is equipped with adaptive fingers for 5 mm adjustment in the horizontal and vertical directions to adapt to errors and for gentle post-grasp placing. The insides of the fingers are lined with a chloroprene rubber sheet to increase the coefficient of friction.
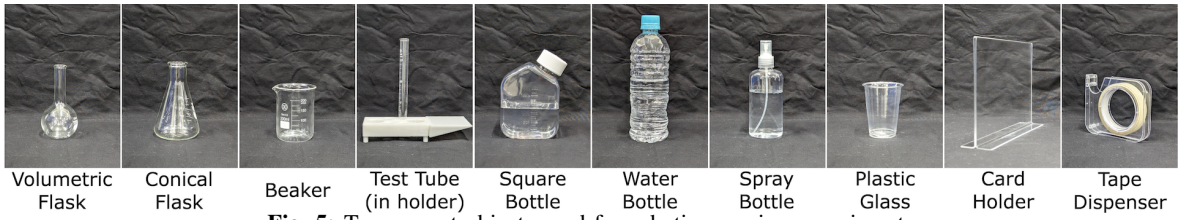
For the grasp targets, we select 10 objects shown in Fig. 5. The objects were chosen amongst laboratory and household equipment to ensure they span a large range of shapes, sizes, and surface gloss such that the difficulties in acquiring depth information would vary. Objects are placed on a white table that is capable of reflecting light (Fig. 1 (a)). Additionally, a single strong directional light was present, which further confused methods that rely on vision.

1) Data Collection: The first step of the process is the collection of datasets. Each object is placed at the center of the workspace, and the robot samples 90 equidistant poses in the hemispheres of radii $[0.4 - 0.6]$ m around the object, visualized in Fig. 1 (a). Using the MoveIt! [46] motion planner, the robotic arm reaches the target pose and records the RGB, depth, and corresponding pose.

2) Grasping Process: After data collection, point clouds are generated using depths completed/estimated by baselines and the proposed method. Using GraspNet [40], we generate and choose the highest scoring, reachable grasp to execute. For each object, we execute 10 grasp attempts and count a success whenever an object is successfully lifted above the table height and is maintained for ≈ 5 seconds. This is done to verify a robust grasp since the objects being grasped are highly slippery and often irregularly shaped.

3) Comparison Methods: For our comparative analysis, we use two-depth completion methods from previous sections, DREDS [4] and TransCG [2], without any additional fine-tuning, and the NeRF method Nerfacto [42], in combination with GraspNet for grasp generation. We also compare against GraspNeRF [35], a high-performance transparent and specular object grasping method with its own grasp generation algorithm VGN [47].

For our proposed SAID-NeRF and Nerfacto, we first create 10 sets of 30 images uniformly sampled from a collected 90 which is used to create the NeRF and generate 10 point clouds. This is done to train and compare the methods under similar environmental conditions. For DREDS and TransCG, we generate point clouds by combining 3-5 sampled views with their completed depth maps. We note that accumulating more views led to significant error accumulation. All point clouds are then cleaned via statistical outlier removal. Grasp-

| Volumetric Flask | Conical Flask | Beaker | Test Tube (in holder) | Square Bottle | Water Bottle | Spray Bottle | Plastic Glass | Card Holder | Tape Dispenser |

Fig. 5: Transparent objects used for robotic grasping experiments

TABLE V: Robotic grasping results for transparent objects

Method	Volumetric Flask	Conical Flask	Beaker	Test Tube (in holder)	Square Bottle	Water Bottle	Spray Bottle	Plastic Glass	Card Holder	Tape Dispenser
SAID-NeRF (Ours)	**9/10**	**9/10**	**9/10**	**8/10**	**7/10**	**10/10**	**10/10**	**10/10**	**9/10**	**9/10**
Nerfacto	0/10	3/10	0/10	7/10	1/10	2/10	7/10	0/10	0/10	0/10
GraspNeRF	7/10	8/10	5/10	1/10	6/10	4/10	**10/10**	1/10	1/10	6/10
TransCG	5/10	3/10	4/10	2/10	2/10	4/10	4/10	0/10	4/10	2/10
DREDS	5/10	2/10	3/10	5/10	1/10	2/10	5/10	0/10	5/10	2/10

NeRF uses 30 input views to generate a Truncated Signed Distance Field (TSDF), and subsequently, grasp candidates.

B. Experiment Results

The success rates (out of 10) of the various methods of grasping the transparent objects are presented in Table V. We see that point clouds generated by our proposed method lead to significant grasp performance across all the objects tested with some representative samples visualized in Fig. 6.

For TransCG, we find that a majority of failure cases involve attempting to grasp noisy pixels that float around in the air despite having run outlier removal. More aggressive processing led to the loss of already challenging point density on the objects. The performance of DREDS, while resulting in similar results to TransCG, is characteristically different. It makes more reasonable grasp attempts but fails when the objects are accidentally pushed away due to the propensity of DREDS to predict smooth depth values along object boundaries in our scenes, especially against the shiny white table, leading to spurious points on its non-camera-facing side. This is exacerbated by our views being sampled above the object, leading to the accumulation of spurious conical structures around the base of the objects, eliminating a significant number of stable side-ward grasps.

Nerfacto, while able to render reasonable RGB shows highly variable behavior in terms of density. It shows reasonable performance with Spray Bottle and Test-tube (with holder), likely due to their opaque cap and base respectively helping with estimated density, while it fails on all other objects. GraspNeRF shows better performance amongst the methods tested and generates grasps different from those from GraspNet. It excels at repeatable grasps for smaller objects such as Spray Bottles and Flasks, while its failure cases appear to be cascading and consistent for objects with low height (Plastic Glass and Beaker) or large volumes (over 30 cm^2) such as the Water Bottle, Card Holder and Test Tube (in holder). We posit this is due to the voxel-based representation used by GraspNeRF not being able to sufficiently capture these objects within its bounds. Using larger values is taxing on GPU memory and causes a loss of details for smaller objects leading to worse grasps overall.

Our proposed method, SAID-NeRF, has the highest success rate on all 10/10 objects and finds that the main failure
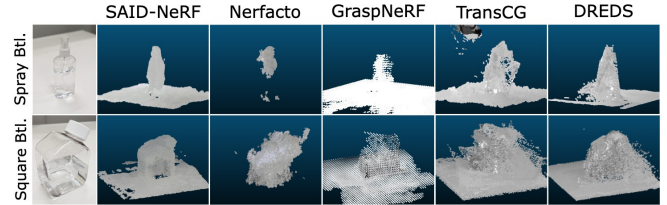
Fig. 6: Point clouds generated by methods tested. GraspNeRF does not use point clouds and its TSDF was converted for visualization

modes are for the Square Bottle which is the heaviest object tested, and causes it to slip due to being picked up by it's cap, or by being accidentally pushed away during approach.

We note that for segmentation targets, two sets of masks are generated since the scenes are relatively simple.

VI. DISCUSSION

Our proposed method, SAID-NeRF, is able to do depth completion of a wide range of challenging transparent objects. Limitations of our approach include the assumptions made by the hierarchical mask generation. While reasonable for typical indoor scenes, they may not transfer to other complex settings. A workaround is using models that provide semantic labels, as briefly mentioned in Section III-D. Another potential issue is the propensity of the semantic components to induce convexity to reconstructed shapes, causing objects with hollow surfaces or fine surface details to be lost. However, objects that can be grasped from the outer edges remain unaffected. While beyond the aim of this work, these details may be recovered by a second stage of fine-tuning without the semantic component.

VII. CONCLUSION

We introduce Segmentation-Aided NeRF, or SAID-NeRF, an extension of Semantic-NeRFs that uses semantic masks to guide depth estimation for transparent objects. We use Instance Segmentation VFMs in a zero-shot way combined with a simple heuristic to automatically construct semantic masks that the NeRFs use to better recover object surfaces. We find that our method is able to perform well under difficult environmental conditions due to its proposed extensions and outperforms several recent depth completion models trained on large datasets for transparent objects. SAID-NeRF is also able to consistently outperform a strong NeRF method

(Nerfacto) on all these tasks. We further evaluate them for robotic grasping of transparent objects under challenging settings and additionally outperform GraspNeRF, a state-of-the-art NeRF-based grasping method that consists of trained and learnable components.

ACKNOWLEDGMENT

REFERENCES

[1] J. Jiang, *et al.*, "Robotic perception of transparent objects: A review," *IEEE Transactions on Artificial Intelligence*, 2023.

[2] H. Fang, *et al.*, "Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7383–7390, 2022.

[3] Y. R. Wang, *et al.*, "Mvtrans: Multi-view perception of transparent objects," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3771–3778.

[4] Q. Dai, *et al.*, "Domain randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects," in *European Conference on Computer Vision*. Springer, 2022, pp. 374–391.

[5] S. Sajjan, *et al.*, "Clear grasp: 3d shape estimation of transparent objects for manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3634–3642.

[6] X. Chen, *et al.*, "Clearpose: Large-scale transparent object dataset and benchmark," in *European Conference on Computer Vision*, 2022.

[7] B. Mildenhall, *et al.*, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[8] T. Müller, *et al.*, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, July 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530127

[9] P. Wang, *et al.*, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *NeurIPS*, 2021.

[10] J. Ichnowski*, *et al.*, "Dex-NeRF: Using a neural radiance field to grasp transparent objects," in *Conference on Robot Learning (CoRL)*, 2020.

[11] R. Bommasani, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[12] A. Kirillov, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.

[13] E. Xie, *et al.*, "Segmenting transparent objects in the wild," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 696–711.

[14] F. Li, *et al.*, "Semantic-sam: Segment and recognize anything at any granularity," *arXiv preprint arXiv:2307.04767*, 2023.

[15] X. Zou, *et al.*, "Segment everything everywhere all at once," *arXiv preprint arXiv:2304.06718*, 2023.

[16] S. F. Bhat, *et al.*, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.

[17] R. Martin-Brualla, *et al.*, "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections," in *CVPR*, 2021.

[18] D. Rebain, *et al.*, "Derf: Decomposed radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 153–14 161.

[19] M. Boss, *et al.*, "Nerd: Neural reflectance decomposition from image collections," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 684–12 694.

[20] P. P. Srinivasan, *et al.*, "Nerv: Neural reflectance and visibility fields for relighting and view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7495–7504.

[21] X. Zhang, *et al.*, "Nerfactor: Neural factorization of shape and reflectance under an unknown illumination," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 6, pp. 1–18, 2021.

[22] D. Verbin, *et al.*, "Ref-nerf: Structured view-dependent appearance for neural radiance fields," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 5481–5490.

[23] B. Roessle, *et al.*, "Dense depth priors for neural radiance fields from sparse input views," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 892–12 901.

[24] Guangcong, *et al.*, "Sparsenerf: Distilling depth ranking for few-shot novel view synthesis," *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[25] M. A. Uy, *et al.*, "Scade: Nerfs from space carving with ambiguity-aware depth estimates," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[26] M. Prinzler, *et al.*, "Diner: Depth-aware image-based neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 449–12 459.

[27] J. Song, *et al.*, "Därf: Boosting radiance fields from sparse input views with monocular depth adaptation," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[28] S. Vora, *et al.*, "Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes," *Trans. Mach. Learn. Res.*, vol. 2022, 2022. [Online]. Available: https://openreview.net/forum?id=ggPhsYCsm9

[29] A. Kundu, *et al.*, "Panoptic neural fields: A semantic object-aware neural scene representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 871–12 881.

[30] S. Zhi, *et al.*, "In-place scene labelling and understanding with implicit scene representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 838–15 847.

[31] X. Fu, *et al.*, "Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 1–11.

[32] A. Mirzaei, *et al.*, "Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 669–20 679.

[33] Y. Wang, *et al.*, "Pet-neus: Positional encoding triplanes for neural surfaces," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[34] J. Kerr, *et al.*, "Evo-nerf: Evolving nerf for sequential robot grasping of transparent objects," in *6th Annual Conference on Robot Learning*, 2022.

[35] Q. Dai, *et al.*, "Graspnerf: multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1757–1763.

[36] M. Breyer, *et al.*, "Volumetric grasping network: Real-time 6 dof grasp detection in clutter," in *Conference on Robot Learning*. PMLR, 2021, pp. 1602–1611.

[37] J. Tang, *et al.*, "Delicate textured mesh recovery from nerf via adaptive surface refinement," *arXiv preprint arXiv:2303.02091*, 2022.

[38] H. Zhu, *et al.*, "Rhino: Regularizing the hash-based implicit neural representation," *arXiv preprint arXiv:2309.12642*, 2023.

[39] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[40] H.-S. Fang, *et al.*, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[41] X. Liu, *et al.*, "Unseen object few-shot semantic segmentation for robotic grasping," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 320–327, 2022.

[42] M. Tancik, *et al.*, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–12.

[43] R. Li, *et al.*, "Nerfacc: Efficient sampling accelerates nerfs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 18 537–18 546.

[44] J. T. Barron, *et al.*, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," *CVPR*, 2022.

[45] L. Ke, *et al.*, "Segment anything in high quality," in *NeurIPS*, 2023.

[46] D. Coleman, *et al.*, "Reducing the barrier to entry of complex robotic software: a moveit! case study," *Journal of Software Engineering in Robotics (JOSER)*, 2014.

[47] M. Breyer, *et al.*, "Volumetric grasping network: Real-time 6 dof grasp detection in clutter," in *Conference on Robot Learning*, 2020.