NAVI: Category-Agnostic Image Collections with High-Quality 3D Shape and Pose Annotations

Varun Jampani* **Kevis-Kokitsi Maninis*** **Andreas Engelhardt** **Arjun Karpur**
Karen Truong **Kyle Sargent** **Stefan Popov** **André Araujo**
Ricardo Martin-Brualla **Kaushal Patel** **Daniel Vlasic** **Vittorio Ferrari**
Ameesh Makadia **Ce Liu†** **Yuanzhen Li** **Howard Zhou**

Google

Abstract

Recent advances in neural reconstruction enable high-quality 3D object reconstruction from casually captured image collections. Current techniques mostly analyze their progress on relatively simple image collections where Structure-from-Motion (SfM) techniques can provide ground-truth (GT) camera poses. We note that SfM techniques tend to fail on in-the-wild image collections such as image search results with varying backgrounds and illuminations. To enable systematic research progress on 3D reconstruction from casual image captures, we propose 'NAVI': a new dataset of category-agnostic image collections of objects with high-quality 3D scans along with per-image 2D-3D alignments providing near-perfect GT camera parameters. These 2D-3D alignments allow us to extract accurate derivative annotations such as dense pixel correspondences, depth and segmentation maps. We demonstrate the use of NAVI image collections on different problem settings and show that NAVI enables more thorough evaluations that were not possible with existing datasets. We believe NAVI is beneficial for systematic research progress on 3D reconstruction and correspondence estimation. Project page: https://navidataset.github.io

1 Introduction

The field of 3D object reconstruction from images or videos has been dramatically transformed in the recent years with the advent of techniques such as Neural Radiance Fields (NeRF) [26]. With recent techniques, we can reconstruct highly detailed and realistic 3D object models from multiview image captures, which can be used in several downstream applications such as gaming, AR/VR, movies, etc.

Despite such tremendous progress, current object reconstruction techniques make several inherent assumptions to obtain high-quality 3D models. A key assumption is that the near-perfect camera poses and intrinsics are given or readily available via traditional Structure-from-Motion (SfM) pipelines such as COLMAP [31]. This assumption imposes several restrictions on the input image collections. The input images have to be of sufficiently high-quality (e.g. non-blurry) and the number of input images should also be high (typically $>$ 30-50) for SfM to estimate sufficient correspondences across images. In addition, SfM techniques typically fail on internet image sets that are captured with varying backgrounds, illuminations, and cameras. Such internet image collections do not require active capturing and are widely and readily available, such as product review photos or image search results (*e.g.*, internet images of Statue-of-Liberty, Tesla Model-3 car, etc.). It is highly beneficial to develop 3D object reconstruction techniques that can automatically produce high-quality 3D models from such image collections in the wild.

*Equal contribution
†C. Liu's current affiliation is Microsoft

Preprint. Under review.

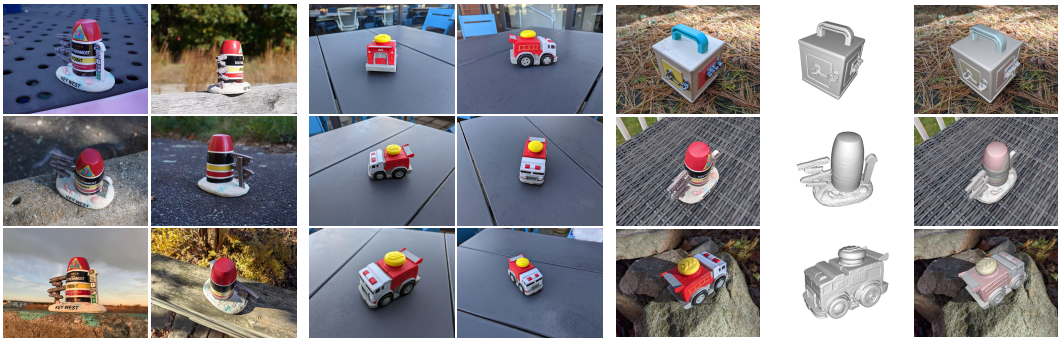

Figure 1: **NAVI dataset overview**. NAVI dataset consists of both multiview and in-the-wild image collections, where each image is aligned with the corresponding 3D scanned model resulting in high-quality 3D shape and pose annotations.

In this work, we propose a new dataset of image collections which we refer to as 'NAVI' (Not AVerage Image dataset). Specifically, our dataset contains two types of image collections with near-perfect camera poses and 3D shapes: 1. Standard multiview object captures and 2. In-the-wild object captures with varying backgrounds, illuminations and cameras. Fig. 1 shows examples of the in-the-wild and multiview images in NAVI along with the 2D aligned 3D scans. Next, we describe the key distinguishing properties of the NAVI dataset in relation to existing datasets.

Casual captures. Several existing multiview datasets are either synthetic or captured in lab settings [26]. We capture NAVI images in casual real settings using hand-held cameras.

In-the-wild image collections. In addition to typical multiview images, NAVI also provides in-the-wild image collections where objects are captured under varying backgrounds, illuminations, and cameras. SfM techniques usually fail on such image sets and NAVI provides a unique opportunity to systematically research joint shape and camera estimation from in-the-wild image collections.

Near-perfect 3D geometry and camera poses. We use high-quality 3D scanners to get 3D shape ground-truth and also obtain high-quality 3D camera pose annotations with manual 2D-3D alignment along with rigorous verification. This is in contrast to several recent datasets such as [40] that rely on SfM to provide GT, thereby limiting the image capture setups.

Accurate dense correspondences. We provide accurate per-pixel correspondences using the 3D shape alignments. While most real-world datasets for correspondence evaluation rely on known homographies [3] or sparse keypoint annotations recovered from estimated geometry [8, 18], NAVI's precise 2D-3D alignments lead to accurate and dense object correspondences.

Derivative annotations such as pixel-accurate object segmentation and monocular depth can be easily derived from high-quality 2D-3D alignments in NAVI.

Category-agnostic. Objects in the NAVI dataset are category-agnostic with image collections that do not have any category-specific shapes, which is in contrast to widely-used 2D-3D datasets [37, 41].

To demonstrate the utility of NAVI, we benchmark and analyze some representative techniques on three different problem settings: multiview object reconstruction, 3D shape and pose estimation from in-the-wild image collections, and dense pixel correspondence estimation from image pairs. In addition to these problem settings, one could also use NAVI images for other single-image vision problems such as single image 3D reconstruction, depth estimation, object segmentation, etc.

2 NAVI dataset construction

Challenges. It is worth emphasizing the challenges in our data construction by taking a look at some existing 2D-3D aligned datasets. Several works [6, 7, 20, 9, 12] propose synthetic 3D assets which are used to render 3D-aligned images. Real-world datasets such as Scan2CAD [2] and Pascal3D+ [41] use nearest intra-category CAD models for alignment w.r.t 2D images, resulting in only coarse annotations. Similarly, IKEA Objects [23] and Pix3D [37] annotate retrieved images by aligning one 3D CAD model to images using point correspondences. Even for datasets with mostly exactly-matching products [37], slight deformations and moving parts that appear different on images with

respect to their 3D scan can lead to inaccurate alignments. Different instances of the same object can also have different shapes due to other factors (e.g. shoes of different sizes are not uniformly scaled versions etc.) Fig. 2 shows the sample alignments from the existing datasets showcasing the challenges in obtaining near-perfect 3D shapes and in the 2D-3D alignment.

Rationale. To avoid such issues in NAVI, we selected rigid objects without moving parts, manually scanned the object shape and took image captures of *the same* objects in diverse real-world settings. We then use our interactive alignment tool to obtain near-perfect 2D-3D alignments with precise pose control during the annotation. Most datasets, including our earlier attempts, use a multi-stage alignment process that involves annotating point correspondences and then optimizing the object pose. Even though this is a more scalable approach for dataset creation, the alignments are not as accurate as we want. The NAVI dataset construction consists of 4 steps: (1) Scanning the 3D objects, (2) Capturing image collections, (3) 2D-3D alignment, and (4) Alignment verification.

1. Scanning the 3D objects. We collect 36 rigid objects and use two professional 3D scanners, EinScan-SP [14] and EinScan Pro HD [13], to obtain high-quality 3D object scans. We center the scans at origin, but do not normalize the shapes to preserve their metric dimensions (in mm). Fig. 3 displays some NAVI images and their aligned 3D scans. Notice the diverse and category-agnostic nature of the objects.

Figure 2: **2D-3D alignments from existing datasets** have issues as 3D models do not exactly match the corresponding 2D image due to model or configuration discrepencies.

2. Capturing image collections. For each object, we captured two types of image collections: in-the-wild, and multiview. In-the-wild captures contain images with different backgrounds, illumination, and cameras. multiview captures offer the standard multiview setup: same camera, object pose, and environment, but with different camera poses. For practical utility, we captured the images in casual settings with hand-held cameras ranging from mobile phones to DSLRs and mirrorless cameras. In total, we use 12 different cameras to capture around 10.5K images with 2.3K in-the-wild images and 8.2K multiview images. More dataset details are present in the supplementary material.

3. 2D-3D alignment. The goal is to obtain near-perfect 2D-3D alignments; *i.e.*, accurate 6DoF rigid object transformations along with accurate camera intrinsics. We developed an interactive tool on which the user can progressively align the 3D object by rotating and translating it in 3D, using the mouse. Since we know the cameras used to capture the images, we initialize the camera focal length, which can be further refined during the alignment process. Our interactive tool gives the user full control over the alignments, and we observe that this leads to higher-quality poses than alternative implicit alignment tools that optimize the pose from 3D↔2D point correspondences [37]. We trained 10 dedicated annotators for our alignment task allowing us to obtain higher quality annotations than several existing datasets that rely on generic non-expert annotators. Refer to the supplementary material for more details on the alignment tool and the process.

4. Alignment verification. To ensure high-quality annotations, we further manually verify each 2D-3D alignment with 2 expert annotators. Specifically, we overlay the 3D shape onto the 2D image and ask trained annotators to label them as 'incorrect' if the alignments look even slightly wrong. For images labeled 'incorrect', we repeat the 2D-3D alignment and verification steps. After two stages of alignment and verification, we discard around 7% of the original captured images. We further annotate images with a binary occlusion label to indicate if the object is occluded by other objects. We exclude occluded object images from our validation sets for different tasks to avoid introducing artifacts in the metrics.

Derivative annotations. In addition to the full 3D alignments of scans to images, there are several derivative annotations that result from the accurate 2D-3D alignments: Relative camera poses, dense correspondences, metric depth maps, and binary masks. Relative camera poses are an implicit output of alignment, as all objects were posed with respect to their canonical pose. Since we have annotated multiple images of the same object, we obtain dense correspondences on the images by sampling the pixels in mutually visible parts of the 3D shape in image pairs. This enables dense correspondence evaluation both for the standard multiview setup, and for in-the-wild images captured in different environments. Fig. 4 visualizes sample GT pixel correspondences on NAVI image pairs. Furthermore,

Figure 3: **NAVI samples**. Sample object images and the corresponding 2D-3D alignments. NAVI consists of casually-captured and category-agnostic image collections with precise 2D-3D alignments.

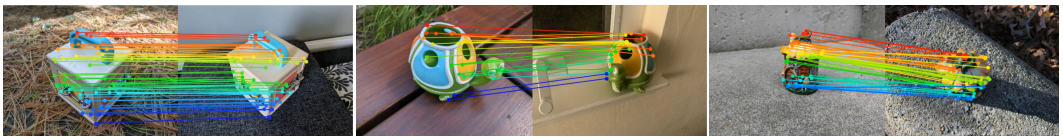

Figure 4: **Pixel correspondences**. Sample image pairs and their corresponding GT pixel correspondences. For visualization purposes, we show sparsely sampled points and color-code the correspondences based on their 2D location from top to bottom.

metric depth maps are obtained by computing the depth of the 3D alignments from the camera viewpoint. The binary object masks are trivially obtained by binarizing the depth maps. Fig. 5 shows sample object depth and mask annotations in NAVI. For simplicity, we refer to our annotations as GT.

3 3D from multiview image collections

Problem setting. Given a set of images taken from different viewpoints, the task is to reconstruct the 3D shape and appearance of an object. The 3D representation can then be used for downstream tasks like scene editing, relighting, and rendering of novel views. Traditional multiview reconstruction pipelines such as Structure-from-Motion (SfM) first reconstruct camera poses together with a sparse object representation followed by a dense reconstruction and potential mesh generation step. After adding materials and textures, the resulting 3D asset can then be used to render new views. More recent techniques such as NeRF [26] optimize neural representations of objects directly on the RGB images with the camera poses obtained from an SfM reconstruction as a pre-processing.

Related datasets. Synthetic multiview scenes [35, 26, 16] are widely adopted for evaluations. In contrast to synthetic scenes that come with precise 3D scene and camera poses but only translate to real-world photography to a limited degree, real scenes usually require off-the-shelf SfM methods [31] for pose estimation. BlendedMVS [43], one of the first multi purpose datasets for stereo reconstruction comes with re-rendered images based on geometry and poses reconstructed via a SfM pipeline. CO3D [29] and Objectron [1] are large-scale datasets with object-centric videos, and provide either a rough point cloud reconstruction of the object [29] or a 3D bounding box [1]. The dataset of [19] offers a handful of 3D laser scans along with the corresponding real-world image collections. Recently, works of [34] and OmniObject3D [40] provide 3D object scans along with multiview image captures in constrained lab settings. These works rely on SfM for semi-automatic 2D-3D alignment. In summary, existing multiview datasets are synthetic [35, 26, 16] or based on reconstructed 3D

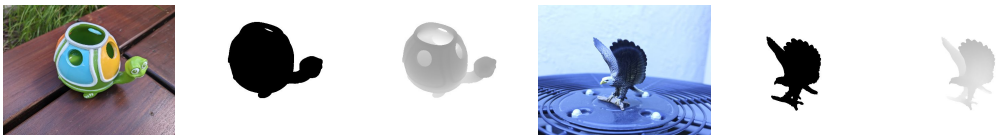

Figure 5: **Sample depths and masks**. 2D-3D alignments on NAVI images allows to readily obtain high quality object depths and mask annotations.

models [43], with rough 3D shapes [29], provide only a limited number of scenes [19] or they consist of image captures in constrained settings [34, 40].

The distinctiveness of NAVI. In contrast, NAVI satisfies multiple requirements by offering highly-accurate 3D shapes and alignments for multiple objects from different categories in different real-world environments and illumination. This allows for more precise evaluation of 3D reconstruction techniques on real-world object image collections.

NAVI dataset and metrics. We split each of the multiview image sets into 80%/20% train/validation sets. The multiview sets are object-centric with an average of 25 images per set (minimum 3 to maximum 180). For even evaluation across the objects, we randomly sample 5 multiview scenes for each object from the subsets that include more than 6 images, resulting in 180 multiview sets for our experiments. We use the standard novel view synthesis metrics, PSNR, SSIM and, LPIPS [46], on validation images and report average metrics across all sets.

Experiment. A key assumption in most existing works is that SfM provided camera poses are good enough for 3D reconstruction. We want to test this hypothesis by evaluating how our annotated camera poses compare against COLMAP [31] poses for off-the-shelf 3D reconstruction techniques. For this, we use the generic and widely-used InstantNGP [28] to reconstruct Radiance Fields from the multiview image sets. For the optimization we use the GT masks to limit the reconstruction to the object area.

Camera Poses	PSNR↑	SSIM↑	LPIPS↓
COLMAP	24.04	0.93	0.079
NAVI Poses (GT)	27.54	0.94	0.045

Table 1: **View synthesis metrics using COLMAP and our GT poses.** This shows considerably better performance with our GT poses. This demonstrates that COLMAP [31] can fail on multiview scenes showcasing the use of our pose annotations on multiview scenes.

Results: COLMAP vs. GT poses. Table 1 shows the novel view synthesis metrics on validation images. Results on all the metrics demonstrate considerably better reconstruction with our GT poses compared to using COLMAP poses. COLMAP only registers partial set of views for several cases. This shows that the our GT poses are accurate and are still valuable in the multiview reconstruction setting to analyze reconstruction techniques independent of inaccuracies from the camera registration. While COLMAP poses are arbitrarily rotated and scaled, all NAVI scenes are centered at the origin and in a common coordinate frame. This facilitates evaluation across different objects, especially in the context of grid-based methods like InstantNGP where the scene bounds have some impact on performance.

4 3D from in-the-wild image collections

Problem setting. The aim is to estimate 3D shape and appearance of an object given an *unconstrained* image collection; where the object is captured with different backgrounds, cameras and illuminations. Such image collections are readily available on the internet; *e.g.*, image search results, product review photos, etc. The high variability in the appearance across images makes pose estimation and reconstruction highly challenging compared to the more controlled multiview captures. Techniques need to jointly reason camera poses and illuminations in addition to 3D geometry and appearance. Standard SfM techniques [32, 31] fail to recover camera poses on such in-the-wild image sets.

Existing datasets. Curated object centric image collections from in-the-wild data are scarce. While one could search online image databases for multiple occurrences of the same object or class [42], additional data like camera parameters or object shape as well as the certainty that all images actually depict the same object instance is critical for faithful evaluation. DTU MVS dataset [17] is widely used as a proxy for in-the-wild data [38, 25, 44] as it comes with different lighting conditions for each of the 124 scenes. However, the controlled acquisition environment does not fully reflect in-the-wild conditions. Additionally, 3D scans and depths are of limited quality and coverage since the structured-light scan is only acquired at the given view positions. NeROIC [21] and NeRD [4] provide small collections of scenes for 360° object reconstruction featuring lighting changes and poses reconstructed via SfM. However, no GT object shapes are included. SAMURAI [5] adds eight image sets to the NeRD dataset with different cameras, backgrounds and illuminations; but it only provides RGB images without any GT camera poses or shapes. NAVI dataset subsumes these 8 SAMURAI in-the-wild image collections where we provide near-GT poses and 3D shapes.

The distinctiveness of NAVI. NAVI provides the first real-world in-the-wild image collections with GT 3D shapes and camera poses. For evaluation, existing techniques such as [5] rely on novel view

(a) GT Novel View Image

(b) NeROIC

(c) NeRS

(d) SAMURAI

Figure 6: **Novel view synthesis with in-the-wild 3D reconstruction**. Sample novel view synthesis results of different techniques for 3D reconstruction from in-the-wild image collections.

synthesis metrics on held-out images which entangle the role of estimated camera poses and shapes. It is not possible to assess whether the view synthesis is poor due to a wrongly estimated camera pose or a wrongly estimated 3D object. GT poses and shapes in NAVI wild-sets provide a unique opportunity to systematically analyze different techniques using pose metrics. In addition, NAVI also enables thorough analysis of techniques with controlled noise levels in the camera parameters.

NAVI dataset and metrics. We divide each of the in-the-wild image sets of NAVI into 80% / 20% splits for training and validation respectively, where the techniques optimize a 3D asset using the train images and are evaluated on validation sets. On average there are 65 images in each in-the-wild set with minimum of 46 and maximum of 93 images, respectively. We use 2 different setups for evaluation. First is the standard novel view synthesis metrics that measure PSNR, SSIM and LPIPS [46] scores on held-out validation images. Second is camera pose evaluation where we use Procrustes analysis [15] to compute the mean absolute rotation, translation and scale difference in camera pose estimations for all the images. The camera metrics are a unique feature of NAVI enabled by our near-GT poses, compared to existing real-world datasets with in-the-wild image collections.

Techniques. We analyze four recent reconstruction techniques that can jointly optimize camera poses and can also deal with varying illuminations to some extent: NeRS [45], SAMURAI [5], NeROIC [21], and GNeRF [25]. Different works use different camera initialization and also model the object appearance differently. NeROIC assumes roughly correct COLMAP poses. NeRS and SAMURAI assume rough quadrant pose initialization and GNeRF takes randomly initialized poses. See the supplementary material for a brief introduction of these techniques and refer to their respective papers for more details. While these techniques use either pre-computed or GT objects masks, we use GT object masks in our experiments to ensure fair comparison.

Method	Pose Init	PSNR↑		SSIM↑		LPIPS↓		Translation↓		Rotation °↓	
		S_C	$\sim S_C$	S_C	$\sim S_C$	S_C	$\sim S_C$	S_C	$\sim S_C$	S_C	$\sim S_C$
NeROIC [21]	COLMAP	19.77	-	0.88	-	0.1498	-	0.09 ± 0.12	-	42.11 ± 17.19	-
NeRS [45]	Directions	18.67	18.66	0.92	0.93	0.1078	0.1067	0.49 ± 0.21	0.52 ± 0.19	122.41 ± 10.61	123.63 ± 8.80
SAMURAI [5]	Directions	25.34	24.61	0.92	0.91	0.0958	0.1054	0.24 ± 0.17	0.35 ± 0.24	26.16 ± 22.72	36.59 ± 29.98
GNeRF [25]	Random	8.30	6.25	0.64	0.63	0.52	0.57	1.02 ± 0.16	1.04 ± 0.09	93.15 ± 26.54	80.22 ± 27.64
NeROIC [21]	GT	22.75	21.31	0.91	0.90	0.0984	0.0845	0.07 ± 0.24	0.01 ± 0.01	33.17 ± 19.63	31.90 ± 11.11
NeRS [45]	GT	17.92	18.02	0.92	0.93	0.114	0.1098	0.62 ± 0.19	0.65 ± 0.20	86.96 ± 27.63	89.43 ± 22.60
SAMURAI [5]	GT	25.65	25.59	0.92	0.92	0.0949	0.0881	0.16 ± 0.14	0.25 ± 0.26	21.55 ± 21.72	28.25 ± 26.71

Table 2: **Metrics for 3D shape and pose from image collections in the wild.** View synthesis and pose metrics over two subsets from all wild-sets depending on the success of COLMAP (S_C / $\sim S_C$). Rendering quality is evaluated on a holdout set of test views that are aligned as part of the optimization without contributing to the shape recovery. We include GNeRF as a separate baseline although this method is not designed for multi-illumination data. We report metrics with the methods' default camera initialization as well as initializing with the GT poses that come with NAVI.

4.1 Analysis

COLMAP vs. GT poses. Table 2 shows the view synthesis performance and camera pose errors for different techniques and camera initializations. We observe that COLMAP reconstruction only works for a subset of scenes S_C (19 out of 36 scenes) for which the camera pose estimation using COLMAP yields more than 10 cameras. For comparisons with NeROIC that rely on COLMAP initialization, we separately report the metrics on scenes S_C where COLMAP works and those where COLMAP fails ($\sim S_C$). We omit one scene (vitamins bottle) that shows some inconsistencies between views because of a moving cap. Compared to the results from Section 3, the increased complexity of the task is reflected in lower performance. Comparing the performance of NeROIC with COLMAP to the initialization with NAVI GT poses on the S_C subset, it is clear that the NAVI GT poses are

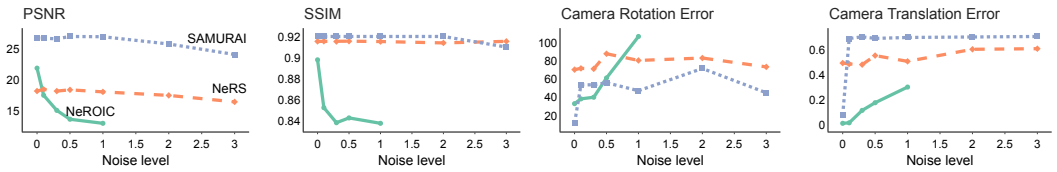

Figure 7: **Analysis with varying camera noise**. For different techniques, we initialize cameras with different levels of noise added to the GT poses for in-the-wild sets. To limit the computation, we report the mean over a subset of four objects of medium complexity.

also superior in this setting. In addition to any COLMAP inaccuracies, the 3D reconstruction task becomes harder as the number of images shrinks due to incomplete COLMAP pose recovery that recovers only a subset of views. Optimizing with GT poses can give insights into the additional challenges of the in-the-wild task independent of any dependency like COLMAP. This enables us to observe the other limitations that have an impact on in-the-wild reconstruction quality like the illumination model in SAMURAI or material model in NeRS.

Comparing different methods. Table 2 shows that SAMURAI performs best although the camera reconstruction quality varies drastically from scene to scene as can be seen in the large uncertainty. This is partly by design as views with large reconstruction errors are discarded over the course of optimization in this approach. It should be noted that data similar to NAVI guided SAMURAI's design. The results indicate that there are aspects covered by this data not available in other datasets (predominantely synthetic) used for evaluations so far. Fig. 6 shows sample novel view synthesis results of different techniques on an example from the "Keywest showpiece" validation set. This is a challenging object with high frequency details (e.g. text), some symmetry, and glossy surface areas. We can observe different artifacts characteristic for the evaluated methods like the rotated view and the high specularity in NeRS, texture smoothness in SAMURAI, and floater artifacts in NeROIC. NAVI includes several challenging objects that are well suited to evaluate the methods' limits. Supplementary material provides further results and analysis of the distribution of the reconstruction errors over different objects in the dataset.

Camera metrics. Thanks to the GT camera pose annotations, both the novel view synthesis and camera evaluations can be done on the same data where multiple datasets, often including synthetic data had to be used in the past. Together with the GT masks from NAVI all the confounding varying assumptions on the input data across different techniques can be made uniform here. For all the techniques, camera errors are relatively high overall, still there is a correlation between pose error and view synthesis quality. NeRS shows a surprisingly large camera pose error. It can be visually confirmed that test views are not that well aligned, still 3D mesh generation based on the training views works relatively well. Camera pose not being a focus in the original work, techniques like NeRS can benefit from explicit pose evaluations for technical improvements.

Analysis with varying camera noise. Annotated camera parameters in NAVI allow for a controlled study of how different techniques work with increasing amount of camera noise in their camera initialization. Specifically, we add normal distributed noise with zero mean and varying standard deviation to the annotated poses before feeding it as input to different techniques. The rotational change is limited to +/- 90° and the translation noise scales with the mean distance of the cameras to the object. A noise level of 1.0 translates to a standard deviation of 10% of the mean distance for translation and 18° standard deviation for the rotation noise on a linear scale. Fig. 7 shows the plots with novel view synthesis and camera metrics for SAMURAI, NeRS and NeROIC. While the pose error generally increases as the noise level increases, the camera rotation error is not strictly monotonically increasing, for example. This points to the shape of the loss landscape with local minima. Both SAMURAI and NeRS seems relatively robust with varying camera noises, while NeROIC performance degrades with increasing camera noise. SAMURAI seems to be robust to large noise levels but, except for GT poses, yields a high translation error. This might stem from the camera multiplex initialization and view weighting scheme. Translation can also be approximated by a focal length change to some extend which could also happen in SAMURAI where the global scene bound is part of a regularization that prefers cameras around the mean radius. NeROIC performs very well under small noise levels but cameras rotate too far away from the object bounding box for higher camera noise levels. It seems like small rotation errors can be compensated by the neural network (if conditioned on direction) to some extent here. In summary, different methodologies seem to be needed for different strengths of camera noise. NAVI can help systematically investigate

Method	Correspondence Metrics		Relative Pose Metrics		
	Precision@0.2↑	Dense-Recall@15px↑	AUC@5°↑	AUC@10°↑	AUC@20°↑
SIFT + MNN	2.8 / 1.2	6.3 / 2.2	7.3 / 5.9	13.6 / 11.9	23.5 / 23.0
SIFT + NN-Ratio	10.2 / 4.8	6.4 / 2.2	6.2 / 4.2	11.9 / 8.1	22.7 / 23.1
SuperPoint + MNN	6.1 / 3.3	9.7 / 5.6	10.0 / 8.2	19.2 / 16.0	31.6 / 28.0
SuperPoint + NN-Ratio	24.7 / 18.7	11.4 / 6.8	9.0 / 7.6	17.5 / 15.0	29.0 / 26.5
SuperPoint + SuperGlue	26.8 / 23.8	12.6 / 9.1	12.1 / 10.8	22.2 / 20.1	34.6 / 32.3
LoFTR	19.2 / 13.4	16.3 / 8.5	12.2 / 9.8	22.5 / 18.4	34.2 / 30.0

Table 3: **Correspondence and relative pose estimation.** For each metric, we present multiview (left) / in-the-wild (right) results. To calculate precision, we filter correspondences (0.2 confidence) and use a reprojection error of 3 pixels. For all other metrics, we consider the entire prediction set.

how the camera optimization performs in a technique thereby informing on several useful design choices for technical improvements (e.g. larger vs. smaller pose updates, regularization weights, initialization and fine-tuning). In addition, investigations around the breaking point of a method can lead to valuable insights into the task of joint shape and camera optimization.

5 Correspondence estimation

Problem setting. Given a pair of images of the same object, the goal of correspondence estimation is to match a set of object pixels from one image to the corresponding pixels in the second image. By definition, an image point can have at most one correspondence in the other image as some points may be unmatched due to occlusion. Image pair correspondences are fundamental for the downstream tasks of 3D reconstruction and pose estimation, where a robust estimator is often used to recover the underlying relative camera rotation and translation.

Existing datasets. Finding a suitable dataset for training and evaluating correspondence estimation methods can be a challenge. SPair-71k [27] and CUB [39] provide in-the-wild semantic correspondences, but these correspondences associate parts of different objects and have limited use in instance-level tasks. Manually labeling fine-grained, instance-level correspondences is a time-consuming and error-prone task, so datasets must rely on either known real [3] or synthetic [10] homographies, or complete scene information [8, 18, 22, 33]. However, synthetic homography pairs suffer from unrealistic image distortion, and many of the latter datasets focus only on indoor/outdoor scenes and not object-centric imagery. Alternatively, high-quality 3D models [12] can be used to render object-focused image pairs with known correspondences, but methods may suffer from a wide domain gap when transferring knowledge from synthetic renderings to real world scenes.

The distinctiveness of NAVI. In contrast, the NAVI dataset annotations allow us to generate real-world image pairs with dense per-pixel correspondences, due to the precise 2D-3D alignments. This provides a unique opportunity to have novel *dense* evaluation metrics for correspondence estimation techniques. Additionally, the NAVI *in-the-wild* collections allow correspondences to be annotated across images with different backgrounds, lighting conditions, and camera models. For example, Fig. 4 shows sample pixel correspondences on NAVI *in-the-wild* image pairs.

NAVI dataset and metrics. We sample two types of correspondence datasets in NAVI. The first dataset contains randomly sampled image pairs *within the same multiview set* to represent the scenario of a fixed scene and camera model. The second dataset contains randomly sampled pairs *from the in-the-wild set* to emulate the variety of backgrounds, illuminations, and cameras. For each image pair, we can use the complete camera-object knowledge to label ground truth correspondences between the two images while respecting self-occlusions. We sample up to 707 multiview pairs and 1035 in-the-wild pairs per object resulting in the validation sets with 24745 and 35931 pairs, respectively. We limit GT correspondence labels to object pixels, since the data annotation process limits the available depth information to object points only. Additionally, we resize each image before evaluation such that their largest dimension is 1200 pixels.

For benchmarking, we evaluate both correspondence and pose estimation metrics. We use precision (reprojection error less than 3 pixels) and recall to directly evaluate correspondences, but we define a recall metric that leverages the dense ground truth correspondences made available by the NAVI 2D-3D alignment. For each object pixel visible in the first image, we find the corresponding location in the second image, after filtering out instances of self-occlusion. Given a correspondence prediction set, we calculate the percentage of ground truth matches which have a corresponding prediction whose keypoints are within N pixels of error. We denote this metric *dense recall*, and it provides an understanding of how well-distributed the predicted correspondences are across the co-visible

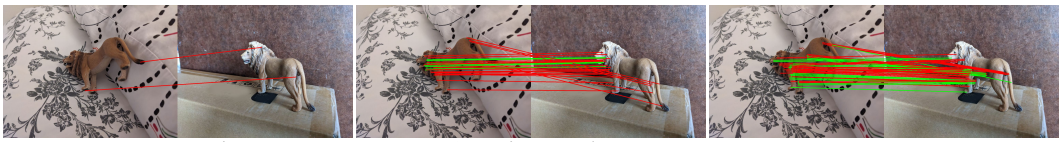

Figure 8: **Sample correspondence results** of different techniques where the correct (within 3 pixels) and incorrect matches are shown in green and red respectively.

regions. In addition, we also estimate relative camera poses from the estimated correspondences and calculate the rotation error between the predicted and ground truth rotation matrices using Rodrigues' formula, and report accuracy within 5°, 10°, and 20° of error following [30].

Techniques. We evaluate the following 4 types of correspondence estimation methods: SIFT + MNN/NN-Ratio [24] that use traditional keypoint detection with heuristic traditional matching; SuperPoint + MNN/NN-Ratio [11] that use learned keypoint detection with traditional matching; SuperPoint + SuperGlue [30] that use both learned keypoint detection and learned matching and; LoFTR [36] that proposes dense learnable matching. We directly evaluate these off-the-shelf models trained on their respective datasets. See the supplementary material for a brief summary of these techniques and refer to the original papers for more details.

5.1 Analysis

Multiview vs. In-the-wild pairs. Table 3 presents the evaluation metrics on the *multiview*/*in-the-wild* image pair datasets in NAVI. Across all metrics, we observe a significant decrease in performance from *multiview* to *in-the-wild* pairs. Traditional methods (i.e. SIFT+MNN/Ratio) are insufficient to handle major changes in lighting conditions, such as ambient lighting and shadows produced by the environment. Learned methods (SuperPoint and SuperGlue) are more robust to changes across *in-the-wild* images with different backgrounds, lighting and cameras. We note that SuperGlue experiences a 3% decrease from *multiview* to *in-the-wild* in Prec@0.2 and a 3.5% decrease in Dense-Recall@15px, compared to 6% and 4.6% for the traditional matcher (SuperPoint + NN-Ratio). We also note that LoFTR proves to be less robust to changes in lighting conditions than the sparse feature-based SuperPoint+SuperGlue method. These results emphasize the importance of exposing learnable features and matchers to sufficient *in-the-wild* image pairs during training.

Dense coverage. Table 3 also shows *dense recall* metric enabled by dense GT correspondences in NAVI. This measures the coverage of pixel correspondences given a wide error tolerance (15 pixels). Local feature techniques are highly dependent on texture-rich regions and suffer from low coverage over smooth/textureless overlapping regions. LoFTR, a dense learnable matcher, performs well on the *multiview* split but is outperformed by SuperPoint+SuperGlue on the *in-the-wild* split. This *dense recall* metric highlights that existing matching techniques recover correspondence sets with low coverage of overlapping object regions, and that the NAVI dataset may serve as a benchmark for this important evaluation metric. Finetuning these methods on object-centric data is likely to yield better performance. Figure 8 shows some sample visual results of correspondences with different techniques.

6 Conclusion and discussion

Use of NAVI in other tasks. In addition to 3D from image collections and correspondence tasks, NAVI can be useful for single-image tasks such as single image 3D reconstruction, monocular depth or normal estimation and object segmentation. There exist several large-scale datasets for these tasks and NAVI can be used as an additional fine-tuning or evaluation dataset. We present some preliminary single image 3D reconstruction experiments in the supplementary material.

Limitations. Scale is the main limitation of the NAVI dataset which consists of only 36 objects and $\approx$10K images. We prioritize annotation quality over quantity; and our current rigorous data capture and annotation pipeline is not easily scalable to collect large datasets. Since the techniques for 3D from image collections usually optimize the 3D models within an image collection, we do not find the small scale of NAVI to be a limiting factor. In the future, we also plan to extend the dataset to videos.

Concluding remarks. In summary, we propose NAVI dataset with multiview and in-the-wild image collections annotated with near-perfect 3D shapes and camera poses. We demonstrated the use of NAVI for better analysis on 3D from multiview image collections, 3D from in-the-wild image

collections and pixel correspondence estimation problems. We believe NAVI is beneficial for a multitude of 3D reconstruction and correspondence tasks.

References

[1] A. Ahmadyan, L. Zhang, A. Ablavatski, J. Wei, and M. Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7822–7831, 2021.

[2] A. Avetisyan, M. Dahnert, A. Dai, M. Savva, A. X. Chang, and M. Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2614–2623, 2019.

[3] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] M. Boss, R. Braun, V. Jampani, J. T. Barron, C. Liu, and H. P. Lensch. Nerd: Neural reflectance decomposition from image collections. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[5] M. Boss, A. Engelhardt, A. Kar, Y. Li, D. Sun, J. T. Barron, H. P. Lensch, and V. Jampani. SAMURAI: Shape and material from unconstrained real-world arbitrary image collections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[6] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. *arXiv preprint*, arXiv:1512:03012, 2015.

[7] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Y. Vicente, T. Dideriksen, H. Arora, et al. ABO: Dataset and benchmarks for real-world 3D object understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21126–21136, 2022.

[8] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[9] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv:2212.08051*, 2022.

[10] D. DeTone, T. Malisiewicz, and A. Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016.

[11] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 224–236, 2018.

[12] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022.

[13] EinScan-Pro-HD. https://www.einscan.com/handheld-3d-scanner/einscan-pro-hd/. Accessed: June-03-2023.

[14] EinScan-SP. https://www.einscan.com/einscan-sp/. Accessed: May-26-2023.

[15] J. C. Gower and G. B. Dijksterhuis. *Procrustes problems*, volume 30 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, UK, January 2004.

[16] K. Greff, F. Belletti, L. Beyer, C. Doersch, Y. Du, D. Duckworth, D. J. Fleet, D. Gnanapragasam, F. Golemo, C. Herrmann, T. Kipf, A. Kundu, D. Lagun, I. Laradji, H.-T. D. Liu, H. Meyer, Y. Miao, D. Nowrouzezahrai, C. Oztireli, E. Pot, N. Radwan, D. Rebain, S. Sabour, M. S. M. Sajjadi, M. Sela, V. Sitzmann, A. Stone, D. Sun, S. Vora, Z. Wang, T. Wu, K. M. Yi, F. Zhong, and A. Tagliasacchi. Kubric: a scalable dataset generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 406–413. IEEE, 2014.

[18] Y. Jin, D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi, and E. Trulls. Image matching across wide baselines: From paper to practice. *International Journal on Computer Vision (IJCV)*, 2020.

[19] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (SIGGRAPH)*, 36(4), 2017.

[20] S. Koch, A. Matveev, Z. Jiang, F. Williams, A. Artemov, E. Burnaev, M. Alexa, D. Zorin, and D. Panozzo. Abc: A big cad model dataset for geometric deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9601–9611, 2019.

[21] Z. Kuang, K. Olszewski, M. Chai, Z. Huang, P. Achlioptas, and S. Tulyakov. Neroic: Neural rendering of objects from online image collections. *ACM Transactions on Graphics*, 41(4), jul 2022. ISSN 0730-0301. doi: 10.1145/3528223.3530177. URL https://doi.org/10.1145/3528223.3530177.

[22] Z. Li and N. Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018.

[23] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2992–2999, 2013.

[24] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[25] Q. Meng, A. Chen, H. Luo, M. Wu, H. Su, L. Xu, X. He, and J. Yu. GNeRF: GAN-based neural radiance field without posed camera. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[26] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

[27] J. Min, J. Lee, J. Ponce, and M. Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019.

[28] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (SIGGRAPH)*, 41(4), July 2022. doi: 10.1145/3528223.3530127. URL https://doi.org/10.1145/3528223.3530127.

[29] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 10901–10911, 2021.

[30] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. Superglue: Learning feature matching with graph neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4938–4947, 2020.

[31] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[33] J. Schönberger, H. Hardmeier, T. Sattler, and P. M. Comparative evaluation of hand-crafted and learned local features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[34] R. Shrestha, S. Hu, M. Gou, Z. Liu, and P. Tan. A real world dataset for multi-view 3d reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 56–73. Springer, 2022.

[35] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2437–2446, 2019.

[36] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou. Loftr: Detector-free local feature matching with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8922–8931, 2021.

[37] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman. Pix3D: Dataset and methods for single-image 3D shape modeling. In *CVPR*, 2018.

[38] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[39] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200, 2010.

[40] T. Wu, J. Zhang, X. Fu, Y. Wang, J. Ren, L. Pan, W. Wu, L. Yang, J. Wang, C. Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. *arXiv preprint arXiv:2301.07525*, 2023.

[41] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014.

[42] C.-H. Yao, W.-C. Hung, Y. Li, M. Rubinstein, M.-H. Yang, and V. Jampani. Lassie: Learning articulated shape from sparse image ensemble via 3d part discovery. In *NeurIPS*, 2022.

[43] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[44] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[45] J. Y. Zhang, G. Yang, S. Tulsiani, and D. Ramanan. NeRS: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[46] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See supplementary material.

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Since this is a dataset work, we release the code useful for working with the proposed dataset.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We benchmark existing techniques for experiments

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] We do not propose new techniques in this work, but benchmark existing techniques

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Authors captured the data on their own.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] In the supplementary

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]