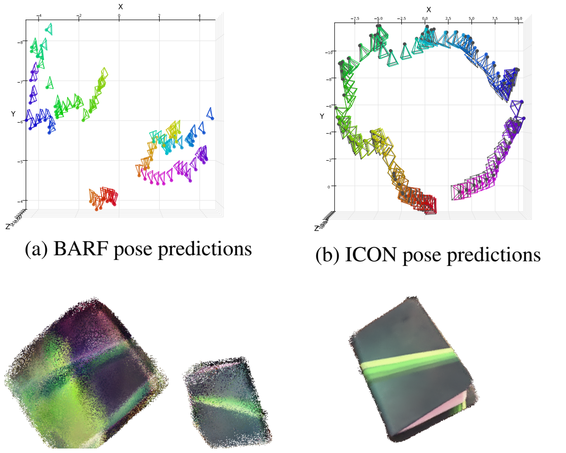
# ICON: Incremental CONfidence for Joint Pose and Radiance Field Optimization

Weiyao Wang, Pierre Gleize[*], Hao Tang[*], Xingyu Chen, Kevin J Liang, Matt Feiszli
FAIR, Meta
{weiyaowang,gleize,haotang,xingyuchen,kevinjliang,mdf}@meta.com
weiyaowang.github.io/icon/

## Abstract

*Neural Radiance Fields (NeRF) exhibit remarkable performance for Novel View Synthesis (NVS) given a set of 2D images. However, NeRF training requires accurate camera pose for each input view, typically obtained by Structure-from-Motion (SfM) pipelines. Recent works have attempted to relax this constraint, but they still often rely on decent initial poses which they can refine. Here we aim at removing the requirement for pose initialization. We present Incremental CONfidence (ICON), an optimization procedure for training NeRFs from 2D video frames. ICON only assumes smooth camera motion to estimate initial guess for poses. Further, ICON introduces "confidence": an adaptive measure of model quality used to dynamically reweight gradients. ICON relies on high-confidence poses to learn NeRF, and high-confidence 3D structure (as encoded by NeRF) to learn poses. We show that ICON, without prior pose initialization, achieves superior performance in both CO3D and HO3D versus methods which use SfM pose.*



(a) BARF pose predictions (b) ICON pose predictions

(c) BARF novel-view synthesis (d) ICON novel-view synthesis

Figure 1. **Novel view and pose visualizations of ICON and BARF when no initial pose is available.** We train on a flyaround video of a book from CO3D [43]. BARF [23] trajectories exhibit fragmentation: camera poses split into two forward-facing clusters and create two books. ICON provides high-quality view synthesis and precisely recovers poses. The colored triangle meshes represent ICON predicted poses, and grey ones represent groundtruth.

## 1. Introduction

Robustly lifting objects into 3D from 2D videos is a challenging problem with wide-ranging applications. For example, advances in virtual, mixed, and augmented reality [28] are unlocking new interactions with virtual 3D objects; 3D object understanding is important for robotics as well (*e.g.* manipulation [18, 42, 64] and learning-by-doing [7, 65]).

Bringing objects to 3D requires both extracting 3D structure and tracking 6DoF pose, but existing approaches have limitations. Many [1, 63, 66] rely on depth, which is a powerful signal for 3D reasoning. However, accurate depth typically requires additional sensors (*e.g.* stereo, LiDAR), which add cost, weight, and power consumption to a device, and is thus often not widely available. Without this depth signal, these methods often fail. Solving only half the problem is also common: 3D object reconstruction methods often assume pose [34, 36, 39, 43, 53, 61, 71], and object pose estimation methods often assume a 3D model (*e.g.* CAD) [21, 41, 68]. This chicken-and-egg problem often limits the applicability of these approaches.

Here we aim to tackle both problems jointly, learning both an implicit 3D representation and per-frame camera poses from a single monocular RGB video. We supervise both 6DoF poses and reconstruction with a dense photometric loss, projecting the 3D representation onto the 2D input frames. Specifically, we represent objects/scenes as a Neural Radiance Field (NeRF) [34] to obtain 2D rendering.

While recent works [17, 23, 24, 57, 62, 72] have shown that poses can to some extent be (jointly) learned in this setting, they are most effective when used to refine initial poses with moderate noise. For example, [62] shows they begin to fail when pose noise exceeds approximately 20 degrees of rotation error; more complex trajectories are unrecoverable. Indeed, these methods also fail on even moderately-

[*] Equal contribution.

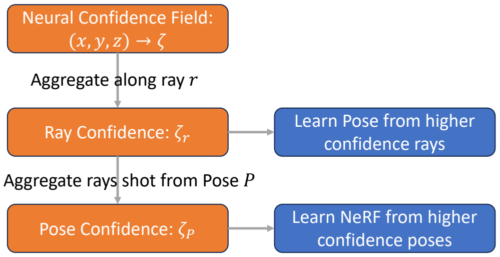

Neural Confidence Field: $(x, y, z) \rightarrow \zeta$
Aggregate along ray $r$
Ray Confidence: $\zeta_r$
Learn Pose from higher confidence rays
Aggregate rays shot from Pose $P$
Pose Confidence: $\zeta_P$
Learn NeRF from higher confidence poses

Figure 2. **ICON overview**. ICON constructs a Neural Confidence field on top of NeRF to encode confidence $\zeta$ for each 3D location. The confidence is then used to guide the optimization process.

complex trajectories, for example a full 360-degree fly-around of an object (Sec. 4). This means SfM preprocessing remains a prerequisite for constructing a radiance field.

One approach may be to focus on the large-noise setting, aiming to resolve larger pose changes. This can be promising [30], but we choose to focus on the incremental case. This arises naturally in real-world settings where video is input, *e.g.* embodied AI. We take inspiration from incremental SfM [49] and SLAM [9], training pose and NeRF jointly in an incremental setting. In this setup, the model takes a stream of video frames, one at a time. Leveraging a motion-smoothness prior, we initialize an incoming frame with the previous frame's pose. Information between frames is exchanged through view synthesis from NeRF.

However, the interdependence between 3D structure and pose presents a major challenge: high photometric error may be attributable to a poor 3D model or a large error in pose. We observe and analyze several interesting failure modes, including fragmentation, a generalization of the classical Bas-Relief ambiguity [2], and overlapping registration (see Fig. 3). To address these difficulties, we propose **ICON** (Incremental CONfidence). The intuition is simple (Fig. 2): "When pose is good, learn the NeRF; when the NeRF is good, learn pose." ICON interpolates between these two regimes, using a measure of confidence obtained from photometric error, and maintaining a NeRF-style "Neural Confidence Field" to store confidence in 3-space. Confidence is also used as a signal to guide optimization, helping to identify (and escape from) local minima.

We perform quantitative evaluation of ICON on CO3D [43], HO3D [15], and LLFF [33]. While joint pose-and-3D baselines often fail catastrophically, ICON achieves strong performance on CO3D, comparable to NeRFs trained on COLMAP [49] pose and surpassing a wide selection of baselines, such as DROID-SLAM [56] and PoseDiffusion [60]. In addition, we evaluate on CO3D videos with background removed; this significantly increases the difficulty since background texture makes camera pose extraction easier. We note that this case (a single masked object in isolation) is quite valuable: success here means a method will work whether the camera is moving, the object is moving, or both. ICON achieves superior performance to NeRF+COLMAP pose and a wide selection of baselines. Finally, ICON outperforms RGB baselines and is comparable to SOTA RGB-D method BundleSDF [66] on dynamic hand-held objects in HO3D.

To summarize, we make the following contributions:

1. We propose an incremental registration for joint pose and NeRF optimization. This setup removes the requirement for pose initialization in common video settings.
2. We systematically study this incremental setup and discover several challenges. Based on the observations, we propose ICON, an optimization protocol based on confidence in spatial locations and poses.
3. We evaluate ICON with a focus on object-centric datasets. ICON is SOTA among RGB-only methods, and is even competitive with SOTA RGB-D methods.

## 2. Related Work

**Neural Radiance Field** (NeRF) [34] is a powerful technique to represent 3D from posed 2D images for novel view synthesis. One major limitation of NeRF resides in its requirement for accurate camera poses. Recent works, including NeRF-- [62], BARF [23], SCNeRF [17], SiNeRF [67], NeuROIC [20], IDR [70], GARF [8] and SPARF [57] have attempted to relax this requirement by jointly optimizing poses and NeRF. Despite the promising direction, they work the best when refining noisy initial poses and are limited by the robustness of initial pose estimation methods. One direction the community takes to further reduce the dependency on pose is by adding additional components or signals for initial pose estimations, such as GANs [30], SLAM [44], shape priors [73], depth [4, 32], correspondence [6, 32], and coarse annotations [5]. We tackle this problem from a different angle, where we propose an incremental setup of joint NeRF and pose optimization. Our proposed method ICON does not use additional signals and achieve strong performance on challenging scenarios when camera poses are difficult to obtain.

**Pose estimation (Object)** aims to infer the 6 Degrees-of-Freedom (DoF) pose of an object from image frames. The line of work can be classified into two main categories: image pose estimation [21, 68, 74] and video pose tracking [35, 52, 55], where the former mostly focuses on inferring pose from sparse frames and the latter takes the temporal information into consideration. However, many methods in video or image pose estimation assume known instance- or category-level object representations, including object CAD models [21, 22, 35, 52, 54, 59, 68] or pre-captured reference views with known poses [25, 40]. Recently, BundleTracks [63] removes the need for such object priors, thus generalizing to pose tracking for unseen novel objects, and BundleSDF [66] improves pose tracking by constructing a neural representation for the object. However, both require depth information, limiting their applications.

**SLAM (Simultaneous Localization and Mapping)** builds a map of the environment while simultaneously determining its location within that map [10, 12, 13, 19, 37, 38, 77]. While most SLAM methods focus on understanding camera pose movement in a static environment, object-centric SLAM [29, 31, 45, 46, 50] focus on learning object pose in a dynamic environment. However, most of those methods require depth signal [29, 31, 45] and struggle with large occlusion or abrupt motion [66].

## 3. Method

ICON takes streaming RGB video frames as input and produces 3D reconstructions and camera pose estimates. ICON incrementally registers each input frame to optimize 3D reconstruction guided by confidence: the 3D reconstruction is learned more from frames with high confidence pose, and pose relies on 3D-2D reprojection from higher confidence areas of the 3D reconstruction.

### 3.1. Preliminaries: Neural Radiance Fields

ICON relies on Neural Radiance Fields (NeRF) to represent a 3D reconstruction: NeRF encodes a 3D scene as a continuous 3D function through a multilayer perceptron (MLP) $f$ parameterized by $\Theta$: 3D point $x$ and viewing direction $d$ form the input $(\boldsymbol{x}, \boldsymbol{d}) \in \mathbb{R}^{\mathbf{5}} \rightarrow (\mathbf{c}, \sigma) \in \mathbb{R}^{\mathbf{4}}$, where $\mathbf{c} \in \mathbb{R}^{\mathbf{3}}$ is the color and $\sigma$ is the opacity. To generate a 2D rendering of a scene at each pixel $p = (u, v)$ in image $\hat{I}_i$ from camera pose $P_i$, NeRF uses a rendering function $\mathcal{R}$ to aggregate the radiance along a ray shooting from the camera center $o_i$ position through the pixel $p$ into the volume:

$$\hat{I}_i(p) = \mathcal{R}(p, P_i|\Theta) = \int_{z_{\text{near}}}^{z_{\text{far}}} T(z)\sigma(\mathbf{r}(z))\mathbf{c}(\mathbf{r}(z), d)dz \tag{1}$$

where $T(z) = \exp(-\int_{z_{\text{near}}}^{z} \sigma(\mathbf{r}(z))dz)$ is the accumulated transmittance along the ray, and $\mathbf{r}(z) = o_i + zd$ is the camera ray from origin $o_i$ through $p$, as determined by camera pose $P_i$. NeRF implements $\mathcal{R}$ by approximating the integral via sampled points along the ray, and is trained through a photometric loss between the groundtruth views $I_i$ and the rendered view $\hat{I}_i$ for all images $i = 1, ..., N$:

$$\Theta^* = \arg\min_\Theta \mathcal{L}_p(\hat{I}|I, P), \text{where } \mathcal{L}_p(I, \hat{I}) = \sum \|I_i - \hat{I}_i\|^2 \tag{2}$$

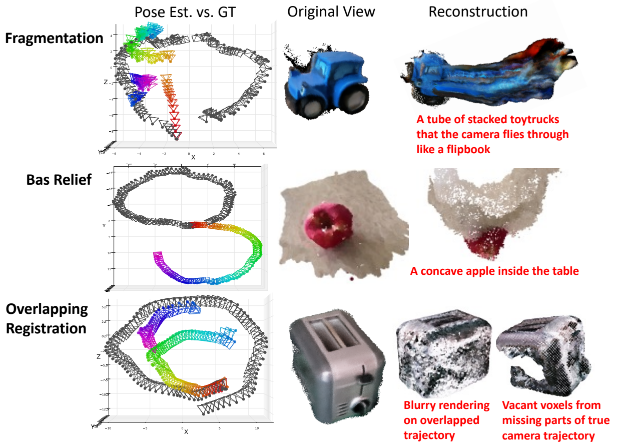


Figure 3. **Three major failure modes of joint pose and NeRF optimization: fragmentation, Bas Relief, and overlapping registration**. Predicted (colored) and GT (grey) poses are shown. **Fragmentation**: Pose and NeRF break apart, producing separate, mutually invisible radiance fields. Here a tube of toy trucks is created, each occluding the next, which the poses fly through flipbook-style, each seeing a single toy truck. See also Fig. 1, where independent reconstructions occur in different regions of 3-space. **Bas Relief**: Due to an inherent ambiguity in RGB reconstruction, the model constructs a "relief" by creating a concave apple inside the table, resulting in camera trajectories inverted by 180 degrees. **Overlapping Registration**: Two subsets of the pose trajectory are trapped in a local minimum, incorrectly observing the same part of the radiance field, leading to blurry rendering and empty voxels. Here, one side of the toaster is blurry due to overlapping views, while the other has no views and is vacant.

### 3.2. Incremental frame registrations

A major limitation for these joint pose and NeRF optimization methods is a requirement for good initial poses. If $\{P_i\}$ contain a diverse set of viewpoints and are initialized all from identity, these methods often collapse. For example, a simple but common collapsing solution is fragmentation: each frame creates its own fragmented 3D representation, all mutually invisible to the other views (**Fragmentation** fig. 3). Indeed, BARF [23] collapses on all sequences of the CO3D dataset when the poses $\{P_i\}$ consist of a closed-loop flyaround (see Tab. 1). As discussed in [62], when no pose prior is provided, a breaking point of 20 degree rotation difference for the whole trajectory is observed.

To tackle this problem, we rely on a simple yet effective intuition: camera motions in videos are smooth. Therefore, given a frame $I_i$ in a video, its camera pose $P_i$ is likely to be close to $P_{i-1}$. We leverage this observation and propose to register frames incrementally following the temporal order.

**Implementation.** At the start of training, we jointly optimize NeRF parameters $\Theta$ and poses $\{P_1, P_2\}$ from the first two frames $\{I_1, I_2\}$. After every $k$ iterations, we add a new frame $I_i$ and initialize its pose $P_i$ by $P_{i-1}$. We freeze the learning rate on poses $\{P_i\}_{i=1}^N$ and NeRF $\Theta$ until all frames are registered. A learning rate decay schedule may be applied after all $N$ images are added.

### 3.3. Confidence-Based Optimization

The incremental registration process aims at providing good initialization for the camera poses. However, optimizing poses and NeRF using photometric losses is highly non-convex and contains many local minima [24, 72]. In addition, an incorrectly optimized pose may provide misleading learning signals towards NeRF, increasing the possibility for poses to re-register incorrectly on already registered viewpoints (**Overlapping Registration** fig. 3).

To tackle these, we propose a confidence-guided optimization schema. The intuition is simple: when a pose $P_i$ is confident, it should be trusted more to improve the learned NeRF $f(\Theta)$; when a ray sampled from $P_i$ contains locations that are confident, it should be weighted more to adjust the poses. When pose confidence drops dramatically for a new frame, it is likely that the pose got stuck in a local minima, so we perform a restart to re-register this pose. This is similar to the trial and error strategy of COLMAP [49]. We next describe how we measure confidence for each pose $P_i$ and each point/viewing direction $(\boldsymbol{x}, \boldsymbol{d})$ in 3D.

**Encoding confidence in 3D**. We construct a Neural Confidence Field on top of NeRF: given an input 3D location and direction $(\boldsymbol{x}, \boldsymbol{d})$, NeRF $f$ also predicts confidence $\zeta_{(\boldsymbol{x},\boldsymbol{d})}$. We add one fully-connected layer on top of the features, followed by a sigmoid, similar to the color prediction head.

The confidence for a ray $\boldsymbol{r}$, is then aggregated through volumetric aggregation similar to opacity rendering:

$$\zeta_{\boldsymbol{r}} = (\int_{z_{\text{near}}}^{z_{\text{far}}} \mathcal{P}(z)dz)(\int_{z_{\text{near}}}^{z_{\text{far}}} \mathcal{P}(z)\zeta(\mathbf{r}(z), d)dz) + (1 - \int_{z_{\text{near}}}^{z_{\text{far}}} \mathcal{P}(z)dz)(\int_{z_{\text{near}}}^{z_{\text{far}}} \zeta(\mathbf{r}(z), d)dz) \quad (3)$$

where $\mathcal{P}(z) = T(z)\sigma(\mathbf{r}(z))$. We note that the first term is more prominent when the pixel is opaque whereas the latter is more prominent for transparent pixels.

**Measuring confidence**. We measure confidence by how well a pixel reprojects in 2D through photometric error. Given a ray and its confidence $\zeta_{\boldsymbol{r}}$, we minimize $\mathcal{L}_{\text{conf}} = \|e^{-\mathcal{E}/\tau} - \zeta_{\boldsymbol{r}}\|^2$, where $\mathcal{E}$ is the photometric error used to train NeRF and $\tau$ is a temperature parameter. $\mathcal{L}_{\text{conf}}$ is only used to train the confidence head; the gradient is stopped before NeRF parameters $\Theta$ or poses.

**Pose confidence**. We compute pose confidence $\zeta_{P_i}$ for pose $P_i$ by aggregating confidence over rays sampled from $P_i$. At the start, $P_1$ has confidence 1 and others have confidence 0. During training, we use a momentum schedule to update pose confidence: at training iteration $t$, we sample $B$ rays $\{\boldsymbol{r}_j^i\}_{j=1}^B$ from pose $P_i$, and update confidence $\zeta_{P_i}^t$ as

$$\zeta_{P_i}^t = \beta\zeta_{P_i}^{t-1} + (1-\beta)\frac{1}{B}\sum_{j=1}^{B}\zeta_{\boldsymbol{r}_j^i} \quad (4)$$

The momentum $\beta$ is 0.9 in our experiments.

**Calibrating loss by confidence**. We use confidence to calibrate $\mathcal{L}$. Intuitively:

- When we compute gradients for NeRF parameters $\Theta$, the loss is weighted by $\{\zeta_{P_i}\}$, the pose confidence.
- When we compute gradients for pose $\{P_i\}$, the per-ray loss is weighted by $\{\zeta_{\boldsymbol{r}}\}$, the ray confidence.

At each step, we sample ray $\{\mathrm{r}_j^i\}_{j=1}^B$ from $P_i$. The loss is:

$$\mathcal{L}_{\text{NeRF}}(\Theta|\hat{P}, I) = \sum_i (\sum_j \mathcal{L}(\boldsymbol{r}_j^i))\zeta_{P_i} / (\sum_{i,j}\zeta_{P_i}) \quad (5)$$

$$\mathcal{L}_{\text{Pose}}(\hat{P}|\Theta, I) = \sum_{i,j}\mathcal{L}(\boldsymbol{r}_j^i)\zeta_{\boldsymbol{r}_j^i} / (\sum_{i,j}\zeta_{\boldsymbol{r}_j^i}) \quad (6)$$

$$\mathcal{L}_{\text{all}}(\Theta, \hat{P}|I) = \mathcal{L}_{\text{NeRF}} + \mathcal{L}_{\text{Pose}} + \mathcal{L}_{\text{conf}} \quad (7)$$

**Pose re-init**. Inspired by trial-and-error registration mechanisms in incremental SfM [49], we do a re-initialization from the previous pose if a new image fails to register. We declare failure if we see an abrupt drop in confidence for a newly registered image: after we register $(I_i, P_i)$, we restart if new pose confidence $\zeta_{P_i}$ is less than $\lambda$ standard deviations of the mean of the $K$ previous pose confidences: $\zeta_{P_i} \leq \text{mean}(\{\zeta_{P_j}\}_{j=i-K}^{i-1}) - \lambda \cdot \text{std}(\{\zeta_{P_j}\}_{j=i-K}^{i-1})$. We use $\lambda = 2$ and $K = 10$ throughout our experiments.

### 3.4. Bas-Relief Ambiguity and Confidence-based Restart

Bas-relief ambiguity [2], and the related “hollow-face” optical illusion, are examples of fundamental ambiguity in recovering an object’s 3D structure when objects that differ in shape produce identical images, perhaps under differing photometric conditions like lighting or shadow. For example, a surface with a round convex bump lit from the left may appear identical to the same surface with an concavity lit from the right. We refer generically to such situations as ”Bas-Relief” solutions. Human visual systems are known to employ strong priors (e.g. favoring convexity) to select a particular solution among multiple possibilities.

We observe this phenomenon when jointly optimizing camera poses and NeRF, especially early in optimization when total camera motion is small. The model becomes stuck in a local minimum and cannot escape. For example, a concave version of the scene may be reconstructed when the groundtruth is a convex scene (see **Bas Relief** in Fig. 3). In this example, the camera movement is off by 180 degrees and moves in opposite directions compared to the groundtruth trajectory. We believe that simple priors, using cues like coarse depth, could help produce more human-like interpretations of natural scenes. However, for this study we avoid crafting priors, and remark that our confidence-based calibration of losses helps reduce this issue (16% to 9%).

We also observe that incorrect Bas Relief solutions generally have higher error and lower confidence; Relief solutions tend to be valid for a limited set of viewpoints and wider viewpoints become inconsistent. Hence we propose a generic solution by adopting the restart strategy from incremental SfM. For example, COLMAP restarts to identify different initial pairs if the final reconstruction does not meet certain criteria (e.g. ratio of registered images). For us, we launch $K$ runs independently and measure the confidence after a fixed number of iterations. We pick the one with the highest confidence. In practice, we launch 3 runs and measure the confidence at 10% of the training.

### 3.5. Confidence-based geometric constraint

Following recent works [17, 57], we add a geometric constraint to the optimization. Different from the ray-distance loss [17] and depth consistency loss [57], we adopt sampson distance [16], similar to [60]. We extract correspondence between a frame and its neighbors. We use SIFT [27] features, primarily for fair comparison with COLMAP. At training time, for each pose $P_i$, we sample a pose $P_j$ in its neighborhood, then compute Sampson distance:

$$\mathcal{L}_{\text{Sampson}} = \frac{|x_i F x_j|}{|(x_i F)^1 + (x_i F)^2 + (F x_j)^1 + (F x_j)^2|} \quad (8)$$

where $F$ is the fundamental matrix between $P_i$ and $P_j$ and $(x_i F)^k$ indicates the $k$th element.

**Loss calibration by confidence**. Although geometric cues help constrain the early optimization landscape, the correspondence pairs can be incorrect and/or not pixel-accurate, especially for objects with little texture. This causes the geometric constraint to be detrimental to ICON for obtaining precise poses and reconstructions. We rely on pose confidence $\zeta_{P_i}$ to weight the Sampson distance: for a pair of pose $P_i$ and $P_z$, weight by $1 - \min(\zeta_{P_i}, \zeta_{P_j})$.

## 4. Experiments

**Datasets**. We focus our study on Common Objects in 3D v2 (**CO3D**) dataset [43], a large-scale dataset consisting of turn-table style videos of objects. Ground truth poses are obtained through COLMAP. We train on two versions of the dataset: **full-scene**, which uses the unmodified image frames (both object and background visible), and **object-only**, which removes the background leaving only foreground object pixels. We believe the object-only version is a more challenging yet meaningful evaluation set; in full-scene, objects are often placed on textured backgrounds where COLMAP can successfully extract poses. This implicitly equates object pose and camera pose, and this assumption breaks in dynamic scenes where both object and camera are moving. We use 18 categories specified by the dev set, with "vase" and "donut" removed due to symmetry (indistinguishable in the object-only setting). We select scenes with high COLMAP pose confidence for camera pose evaluation. We clean the masks using TrackAnything [69]; results on original masks are present in the supplementary. To demonstrate performance on dynamic objects, we additionally re-purpose **HO3D** [15] v2 to evaluate the camera pose tracking and view synthesis quality. HO3D consists of static camera RGBD videos capturing dynamic objects manipulated by human hands. We only use the RGB frames for ICON and select 8 clips (each around 200 frames) from 8 videos, each covering a different object. Finally, we show results on **LLFF** [33], a dataset with 8 forward-facing scenes commonly used for scene-level novel view synthesis, especially for NeRFs.

**Architectures and Losses** Our architecture follows NeRF [34] (no hierarchical sampling) and set the image's longer edge to 640. We use the standard MSE loss of NeRF. When using Sampson distance, it is weighted by $10^{-4}$. For the object-only settings in CO3D and HO3D, where object masks are available, we use MSE loss to supervise the opacity. For HO3D, we use hand masks when provided (7 out of 8 clips) to avoid sampling rays from occluded regions.

**Training**. We use BARF [23] settings and train for 200k iterations. For CO3D and HO3D, we skip every other frame to reduce training time, producing sequences around 100 frames. For ICON and its variants, we add a new frame every 1k iterations (CO3D/HO3D) / 500 iterations (LLFF) and freeze the learning rate (100k iterations for HO3D and CO3D, 30k for LLFF). Following BARF, we do not use positional encodings during registration and apply coarse-to-fine positional encoding after registration.

**Evaluation**. Following [23], we evaluate on the last part (typically 10%) of each sequence. We measure camera pose quality with Absolute Trajectory Error (ATE) [76], performing Umeyama alignment [58] of predicted camera centers with ground truth. ATE consists of a translation (ATE) and rotation ($\text{ATE}_{\text{rot}}$) component, evaluating $l2$-distance between camera centers and angular distance between aligned cameras, respectively. For novel view synthesis, we run an additional test-time pose refinement, following standard practices in previous works [23, 57, 62, 72]. We use PSNR, LPIPS [75], and SSIM as metrics.

**Baselines**. We build ICON on top of **BARF** [23], and compare against BARF for joint pose and NeRF optimization. We additionally consider **NoPe**-NeRF [4], which uses *additional* monocular depth estimation, **L2G**-NeRF [6], which applies a local-global alignment module, and **LocalRF** [32], which leverages *additional* monocular depth and optical estimation and progressively registers multiple radiance fields. For novel-view synthesis, we train NeRF with ground truth poses. For pose, we compare against a wide selection of baselines: **PoseDiff** [60] models SfM within a probabilistic pose diffusion framework; concurrent work **FlowCam** [51] solves pose from estimated 3D scene flow; **DROID**-SLAM [56] is a SOTA end-to-end learning-based SLAM system. We also use their predicted poses to initialize and train NeRF. In addition, on object-

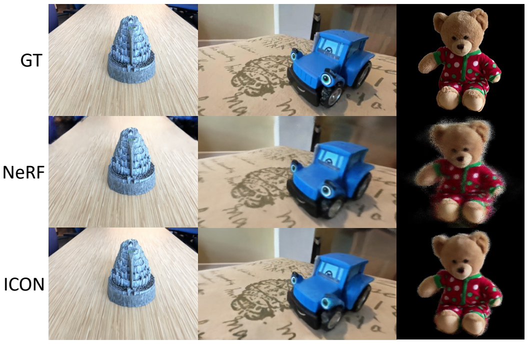


Figure 4. **Novel view synthesis visualization of ICON without poses and NeRF trained with GT poses**. Despite having no pose priors, ICON renders novel views at comparable or higher quality. Results are taken from LLFF and CO3D.

| | ATE | $\text{ATE}_{rot}$ | PSNR | SSIM | LPIPS |
|---|---|---|---|---|---|
| *Pose Source + NeRF* | | | | | |
| DROID | 0.431 | 8.92 | 17.19 | 0.526 | 0.541 |
| FLOW-CAM | 2.681 | 91.28 | 14.40 | 0.441 | 0.689 |
| PoseDiff | 1.973 | 27.25 | 18.82 | 0.563 | 0.520 |
| Groundtruth | - | - | 21.03 | 0.575 | 0.629 |
| *Joint Pose + NeRF optimization* | | | | | |
| BARF | 6.215 | 114.63 | 12.77 | 0.401 | 0.871 |
| GT-Pose+BARF | 0.417 | 3.77 | 19.33 | 0.558 | 0.647 |
| NoPe-NeRF | 5.555 | 115.69 | 10.08 | 0.325 | 0.743 |
| L2G-NeRF | 6.644 | 127.74 | 11.25 | 0.427 | 0.865 |
| LocalRF | 3.715 | 63.42 | 14.60 | 0.467 | 0.693 |
| ICON (Ours) | **0.138** | **1.16** | **22.24** | **0.654** | **0.428** |

Table 1. **Comparison on CO3D [43] full image scenes**. While baseline BARF may fail on CO3D due to larger camera motion overall, ICON can estimate poses very precisely and render novel views at quality similar or better than NeRF trained with GT poses.

only CO3D evaluation, we evaluate poses from state-of-the-art SfM pipeline **COLMAP** [49] and an augment version of COLMAP [47] using learning-based features SuperPoint [11]+SuperGlue [48] (**COLMAP+SPSG**). Though ICON only uses *RGB*, we include popular *RGB-D* methods on HO3D, including DROID with ground truth depth input, **BundleTrack** [63] and state-of-the-art **BundleSDF** [66].

### 4.1. Full scene from CO3D

**ICON is strong on full-scene CO3D.** We compare ICON and baselines on full CO3D scenes in Table 1. Without prior knowledge, BARF must initialize all camera poses as identity. CO3D's flyaround captures of objects result in camera pose variation that significantly exceeds the threshold after which BARF's performance collapses, with an $\text{ATE}_{\text{rot}}$ exceeding 100 degrees. In contrast, ICON's incremental approach recovers significantly more precise camera poses (ATE of 0.137 and $\text{ATE}_{\text{rot}}$ of 1.20), while also achieving better visual fidelity, both qualitatively and quantitatively, as measured by PSNR, SSIM, and LPIPS. Interestingly, ICON still outperforms BARF *even if BARF is provided with the ground truth poses at initialization*. We originally proposed this setting as an upper bound, but we believe this result reflects instability in early iterations of BARF training: CO3D sequences are challenging compared to BARF benchmark scenes (e.g. synthetic dataset from [34]/forward facing LLFF). Camera coverage is sparser, with more drastic lighting changes, and motion blur. Among the 18 scenes, BARF suffers from $\geq$ 10 degree $\text{ATE}_{\text{rot}}$ in 4, dragging down the overall performance. ICON significantly outperforms other pose-free NeRF methods (NoPe, L2G and LocalRF) as well, despite not using additional depth or flow estimations, as visualized in Fig. 5. We note these methods considered easier, unrealistic settings in their experiments; see Supplementary for more details on differences in datasets and evaluation protocols.

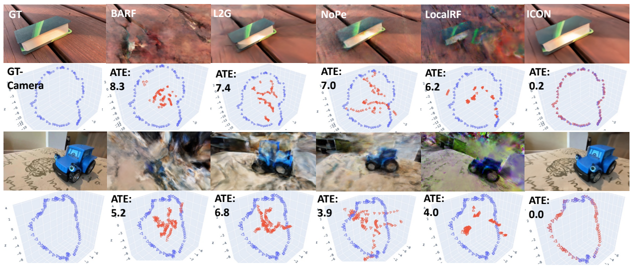


Figure 5. **Novel view synthesis and trajectories of pose-free NeRF methods**. Predicted camera trajectories (red) are aligned with GT (blue). Except ICON, others initialize their test poses by the closest training pose for better quality rendering (see supp).

We also make several comparisons with NeRF [34] and pose prediction methods. We provide NeRF with poses predicted by DROID-SLAM, FLOW-CAM, and PoseDiff, which rely on annotated poses to train or additional signals such as optical flow [55]. However, our joint NeRF and pose training produces better pose estimates (as measured by ATE and $\text{ATE}_{\text{rot}}$), and as a result, NeRF's novel view synthesis suffers in comparison. Even given CO3D's ground truth poses, ICON can outperform NeRF. While this may at first seem surprising, we point out that even the "ground truth" poses in CO3D are not true ground truth; they are generated with COLMAP, which is not perfect. Additionally, in contrast to COLMAP, ICON's joint learning of NeRF and poses means that the estimated poses are specifically optimized to also maximize NeRF quality. We hypothesize that this leads to poses more compatible for learning a NeRF, as reflected by the better performance we observe. Similar observations were presented in prior works [17, 30].

### 4.2. Object-only on CO3D

6DoF pose is inherently tricky to annotate, so past datasets often restrict motion to either the object or the camera; in the latter case, visually distinct backgrounds (*e.g.* specially designed patterns, such as QR codes around the object) are often used to make pose trajectory reconstruction easier.

|  | ATE | ATE$_{rot}$ | PSNR | SSIM | LPIPS |
|---|---|---|---|---|---|
| *Pose Source + NeRF* |  |  |  |  |  |
| DROID | 5.903 | 90.25 | 14.54 | 0.181 | 0.818 |
| FLOW-CAM | 6.700 | 120.52 | 13.08 | 0.127 | 0.886 |
| PoseDiff | 4.601 | 64.24 | 15.42 | 0.508 | 0.492 |
| Groundtruth | - | - | 20.77 | 0.718 | 0.301 |
| *COLMAP variants* |  |  |  |  |  |
| COLMAP(11) | 1.177 | 13.62 |  | - |  |
| COLMAP-SPSG(11) | 2.815 | 38.37 |  |  |  |
| COLMAP-SPSG | 3.616 | 43.74 |  |  |  |
| *Joint Pose + NeRF optimization* |  |  |  |  |  |
| GT-Pose+BARF | 2.055 | 17.00 | 15.65 | 0.802 | 0.277 |
| BARF | 6.522 | 114.97 | 8.22 | 0.772 | 0.370 |
| NoPe-NeRF | 6.355 | 116.68 | 5.95 | 0.186 | 0.824 |
| L2G-NeRF | 6.841 | 130.57 | 7.60 | 0.823 | 0.339 |
| ICON (Ours) | **0.215** | **1.80** | **22.45** | **0.893** | **0.132** |

Table 2. **Comparison on CO3D [43] object-only scenes without background**. Despite the challenges with background removal and failure from other methods, ICON can obtain poses at high precision and render novel views at high-quality. Since COLMAP only successfully registered more than 50% of frames on 11 objects, we marked it with “(11)” for comparison. The SPSG version of COLMAP registers for all scenes, and we include a datapoint on the 11 scenes subset that vanilla COLMAP succeeds.

These strategies however do not generalize to more in-the-wild video, especially when both an object and the background (or camera) are moving. For this reason, we also perform evaluations on CO3D with the background masked out; in such a setting, algorithms are forced to only rely on object-based visual signal for estimating pose (Table 4.2).

In this challenging setting, we again observe that BARF fails to estimate accurate poses, as the camera trajectory changes beyond what BARF can correct. Additionally, the difficulty of this setting produces further deterioration of BARF’s novel view synthesis. However, we observe that ICON can still handle such videos, even without signal from the background. This implies ICON is viable for joint pose estimation and 3D object reconstruction on more general videos, when the background cannot be relied on.

As with our full-scene CO3D experiments, we compare with methods for estimating pose, and how well those poses work when fed to a NeRF. We observe that without being able to leverage the background, these methods struggle mightily. Pose prediction ATE and ATE$_{\mathrm{rot}}$ from DROID-SLAM in particular shoot up from 0.431 to 5.903 and 8.92 to 90.25, respectively. With poorer pose, the quality of the learned NeRFs are also correspondingly worse.

For pose in particular, we additionally evaluate COLMAP and its variant COLMAP-SPSG, which replaces SIFT [27] with SuperPoint-SuperGlue [11, 48], on how they predict pose from just the foreground objects of CO3D. We observe that COLMAP performs significantly worse when it cannot rely on background cues, far worse than ICON. We believe this finding to be especially significant, as COLMAP is often considered the gold standard for camera pose alignment, and is often treated as “ground truth” (as in CO3D). This suggests our incrementally learned joint pose and NeRF optimization represents a promising new alternative for posing moving foreground objects, even if the background or camera is also moving.

|  | Input | ATE | ATE$_{\mathrm{rot}}$ | Trans | PSNR |
|---|---|---|---|---|---|
| BARF | RGB | 0.135 | 122.38 | 0.580 | 5.72 |
| ICON |  | 0.033 | 8.07 | 0.049 | **16.24** |
| Baselines |  |  |  |  |  |
| DROID | RGB | 0.187 | 114.71 | 0.548 | - |
| DROID | RGB-D | 0.105 | 51.93 | 0.262 |  |
| BundleTrack |  | 0.046 | 29.45 | 0.158 |  |
| BundleSDF |  | **0.021** | **6.82** | **0.030** |  |

Table 3. **Comparison on HO3D** [15]. ICON works robustly against faster motion (vs CO3D), hand occlusion and lack of background information. In fact, despite only using RGB inputs, ICON can track poses at similar precision as SOTA RGB-D BundleSDF.

## 4.3. Hand-held dynamic objects on HO3D

Understanding handheld objects is of particular importance to many applications, as the very nature of interaction often implies importance, and hands are often the source of object motion. Pose and 3D reconstructions are key components of understanding objects, so the ability to generate them from videos of handheld interactions is of high utility.

We show results on HO3D [15] in Table 3. Again, we primarily compare against BARF for joint object pose estimation and NeRF learning. Similar to the object-only version of CO3D, the background is masked out since it moves differently than object. In addition, HO3D presents challenges with hand-occlusion and faster pose changes than CO3D. As with CO3D, we observe that BARF struggles to properly learn pose, especially with more drastic camera motion across nearby frames. In contrast, ICON handles these challenges well: poses are predicted accurately and textures are rendered properly in novel views (Fig. 6)

Several existing works [63, 66] addressing this problem additionally use depth, which provides a powerful signal for 3D object reconstruction and pose. On the other hand, depth requires additional sensors and is not always available, and most visual media on the internet is RGB-only. Interestingly, we find that our results with ICON are competitive with state-of-the-art methods like BundleSDF which do require depth. In addition, although we don’t design or optimize ICON for mesh generation, we include a comparison on mesh by running an off-the-shelf MarchingCube [26] algorithm. We follow the evaluation protocol in [66], use ICP for alignment [3] and report Chamfer distnace. Despite not using depth signals, we found ICON provides competitive mesh quality (0.7 cm) compared to BundleSDF (0.77 cm). We remark that BundleSDF’s reconstruction performed poorly on one scene (2.39 cm); removing the worst scene for both methods, BundleSDF and ICON achieved

|  |  |  |  | CO3D-FullImg |  |  |  |  | CO3D-No Background |  |  |  |  | HO3D |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Incre | Geo. | Calib. | Restart | ATE | $\text{ATE}_{rot}$ | PSNR | SSIM | LPIPS | ATE | $\text{ATE}_{rot}$ | PSNR | SSIM | LPIPS | ATE | $\text{ATE}_{rot}$ | PSNR | SSIM | LPIPS |
| ✓ | ✓ | ✓ | ✓ | **0.138** | **1.16** | **22.24** | **0.654** | **0.428** | **0.215** | **1.80** | 22.45 | **0.893** | **0.132** | **0.033** | **8.07** | **16.24** | **0.863** | **0.164** |
| ✓ | ✓ | ✓ |  | 0.714 | 25.40 | 20.48 | 0.632 | 0.486 | 0.224 | 1.86 | **22.47** | 0.892 | **0.132** | 0.035 | 27.32 | 15.02 | 0.873 | 0.670 |
| ✓ |  | ✓ | ✓ | 1.691 | 28.95 | 18.66 | 0.565 | 0.556 | 0.340 | 3.91 | 21.92 | 0.887 | 0.140 | 0.032 | 19.19 | 14.51 | 0.866 | 0.184 |
| ✓ | ✓ |  |  | 1.283 | 36.82 | 19.05 | 0.567 | 0.562 | 0.972 | 15.94 | 21.03 | 0.875 | 0.163 | 0.046 | 30.50 | 12.86 | 0.863 | 0.290 |
| ✓ |  |  |  | 3.075 | 78.49 | 14.38 | 0.454 | 0.816 | 0.890 | 8.05 | 20.67 | 0.850 | 0.187 | 0.076 | 32.26 | 12.51 | 0.870 | 0.189 |
|  |  |  |  | 6.215 | 114.63 | 12.77 | 0.401 | 0.871 | 6.522 | 114.97 | 8.22 | 0.772 | 0.370 | 0.307 | 131.16 | 7.45 | 0.82 | 0.29 |

Table 4. **Ablation study by removing components when possible**. We remark that all designed component are critical for ICON. In addition, we didn't observe Bas Relief on the CO3D Object-Only (No Background) scenes, so the effect of Restart is minimal.

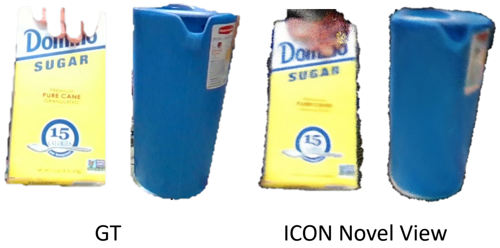


Figure 6. **Visualization of ICON novel view synthesis on HO3D**. ICON can recover shapes and textures accurately.

0.54 cm and 0.56 cm. We believe that this represents the potential of monocular RGB-only methods for object pose estimation and 3D reconstruction.

### 4.4. Ablation studies

**What are the key components in ICON?** We perform ablation studies to gain deeper insight why our proposed methodology leads to such significant improvements in Table 4, examining the impact of incremental frame registration ("Incre."), as well as confidence-based geometric constraint ("Geo."), loss calibration through confidence ("Calib."), and restarts ("Restart"). Note that the top row, with all options enabled, corresponds to our proposed ICON, while the bottom row (with none) is equivalent to BARF. We find all the proposed techniques to be essential.

**ICON works on forward-facing scenes with minor camera motion.** While we primarily focus on the challenging setting of object-centric pose estimation and NeRF representations, ICON does not enforce any object-specific priors. Our approach thus also generalizes to the scene images of LLFF [33], a common benchmark used by the wider NeRF community. Compared to the type of videos in CO3D or HO3D, the images in LLFF tend to be forward-facing, so the camera poses across images have only mild differences. Though easier, being able to recover camera poses in such settings is still important for wider applicability. We find that because the camera poses of LLFF only have limited variation, BARF initialized at identity is able to recover good poses and achieve good PSNR, SSIM, and LPIPS (Table 5). ICON, however, outperforms both BARF and a standard NeRF provided with ground truth poses.

|  | ATE | $\text{ATE}_{\text{rot}}$ | PSNR | SSIM | LPIPS |
|---|---|---|---|---|---|
| GT-Pose+NeRF | - | - | 22.06 | 0.648 | 0.294 |
| BARF | 0.498 | 0.896 | 23.89 | 0.721 | 0.240 |
| ICON | **0.459** | **0.806** | **24.23** | **0.731** | **0.221** |

Table 5. **Comparison on LLFF [33]**. When camera poses have minor or mild motion, BARF works well with identity pose initialization and ICON performs slightly better. ATE is scaled by 100.

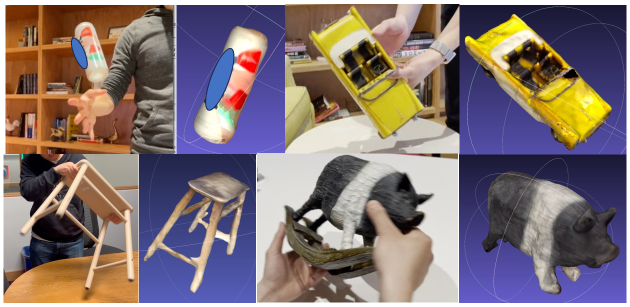


Figure 7. **3D-fy objects in-the-wild.** (left) Videos of hand-manipulated objects; (right) we run ICON on the masked video to get poses+NeRF, and then use Marching Cubes to create meshes.

### 4.5. 3D-fy objects in-the-wild

ICON essentially reduces the need for pose extractions to train NeRFs. We believe it has the potential to unlock key capabilities in real-world applications. As a proof-of-concept, we record several videos of dynamic objects in-the-wild, including object manipulations and object throwing. We ran ICON and convert the NeRF to meshes using off-the-shelf Marching Cubes algorithms (Fig. 7).

## 5. Conclusion

We proposed to study joint pose and NeRF optimization in an incremental setup and highlighted interesting and important challenges in this setting. To tackle them, we have designed ICON, a novel confidence-based optimization procedure. The strong empirical performance across multiple datasets suggests that ICON essentially removes the requirement for pose initialization in common videos. Although our focus is on object-centric scenarios, there are no priors or heuristics that rule out other settings. ICON's LLFF and full-scene CO3D results are strong and show promise for more general types of video input, such as scene reconstruction from moving cameras (*e.g.* egocentric [14]).

## References

[1] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6290–6301, 2022. 1

[2] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *International journal of computer vision*, 1999. 2, 4

[3] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. 7

[4] Wenjing Bian, Zirui Wang, Kejie Li, Jiawang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. 2023. 2, 5

[5] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan T. Barron, Hendrik P.A. Lensch, and Varun Jampani. SAMURAI: Shape And Material from Unconstrained Real-world Arbitrary Image collections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[6] Yue Chen, Xingyu Chen, Xuan Wang, Qi Zhang, Yu Guo, Ying Shan, and Fei Wang. Local-to-global registration for bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8264–8273, 2023. 2, 5

[7] Shuo Cheng, Caelan Garrett, Ajay Mandlekar, and Danfei Xu. NOD-TAMP: Multi-step manipulation planning with neural object descriptors. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition @ CoRL2023*, 2023. 1

[8] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In *The European Conference on Computer Vision: ECCV*, 2022. 2

[9] Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1403–1410. IEEE, 2003. 2

[10] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007. 3

[11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 337–33712, 2017. 6, 7

[12] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 834–849. Springer, 2014. 3

[13] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 3

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Computer Vision and Pattern Recognition*, 2022. 8

[15] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Computer Vision and Pattern Recognition*, 2020. 2, 5, 7

[16] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003. 5

[17] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *International Conference on Computer Vision*, 2021. 1, 2, 5, 6

[18] Daniel Kappler, Franziska Meier, Jan Issac, Jim Mainprice, Cristina Garcia Cifuentes, Manuel Wüthrich, Vincent Berenz, Stefan Schaal, Nathan Ratliff, and Jeannette Bohg. Real-time perception meets reactive motion generation. *IEEE Robotics and Automation Letters*, 3(3):1864–1871, 2018. 1

[19] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, pages 225–234. IEEE, 2007. 3

[20] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. Neroic: Neural rendering of objects from online image collections. *ACM Trans. Graph.*, 41(4), 2022. 2

[21] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European Conference on Computer Vision*, 2020. 1, 2

[22] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d

pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022. 2

[23] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 5

[24] Yunzhi Lin, Thomas Müller, Jonathan Tremblay, Bowen Wen, Stephen Tyree, Alex Evans, Patricio A. Vela, and Stan Birchfield. Parallel inversion of neural radiance fields for robust pose estimation. In *ICRA*, 2023. 1, 4

[25] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *European Conference on Computer Vision*, pages 298–315. Springer, 2022. 2

[26] William E. Lorensen and Harvey E. Cline. Marching cubes: A high-resolution 3d surface construction algorithm. *Computer Graphics*, 21(4):163–169, 1987. 7

[27] David G. Lowe. Object recognition from local scale-invariant features. *International Conference on Computer Vision (ICCV)*, pages 1150–1157, 1999. 5, 7

[28] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: A hands-on survey. *IEEE Transactions on Visualization and Computer Graphics*, 22(12):2633–2651, 2016. 1

[29] John McCormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: Volumetric object-level slam. In *2018 international conference on 3D vision (3DV)*, pages 32–41. IEEE, 2018. 3

[30] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. GNeRF: GAN-based Neural Radiance Field without Posed Camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 6

[31] Nathaniel Merrill, Yuliang Guo, Xingxing Zuo, Xinyu Huang, Stefan Leutenegger, Xi Peng, Liu Ren, and Guoquan Huang. Symmetry and uncertainty-aware object slam for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14901–14910, 2022. 3

[32] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H. Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *CVPR*, 2023. 2, 5

[33] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2, 5, 8

[34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020. 1, 2, 5, 6

[35] Norman Muller, Yu-Shiang Wong, Niloy J Mitra, Angela Dai, and Matthias Nießner. Seeing behind objects for 3d multi-object tracking in rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6071–6080, 2021. 2

[36] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8280–8290, 2022. 1

[37] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 3

[38] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 3

[39] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 1

[40] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10710–10719, 2020. 2

[41] Karl Pauwels and Danica Kragic. Simtrack: A simulation-based framework for scalable real-time object pose detection and tracking. In *International Conference on Intelligent Robots and Systems*, 2015. 1

[42] Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. General in-hand object rotation with vision and touch. In *7th Annual Conference on Robot Learning*, 2023. 1

[43] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 1, 2, 5, 6, 7

[44] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. *arXiv preprint arXiv:2210.13641*, 2022. 2

[45] Martin Runz, Maud Buffier, and Lourdes Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20. IEEE, 2018. 3

[46] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1352–1359, 2013. 3

[47] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 6

[48] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of*

*the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 6, 7

[49] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2, 4, 6

[50] Akash Sharma, Wei Dong, and Michael Kaess. Compositional and scalable object slam. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11626–11632. IEEE, 2021. 3

[51] Cameron Smith, Yilun Du, Ayush Tewari, and Vincent Sitzmann. Flowcam: Training generalizable 3d radiance fields without camera poses via pixel-aligned scene flow, 2023. 5

[52] Manuel Stoiber, Martin Sundermeyer, and Rudolph Triebel. Iterative corresponding geometry: Fusing region and depth for highly efficient 3d tracking of textureless objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6855–6865, 2022. 2

[53] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. *CVPR*, 2021. 1

[54] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the european conference on computer vision (ECCV)*, pages 699–715, 2018. 2

[55] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, page 402–419, Berlin, Heidelberg, 2020. Springer-Verlag. 2, 6

[56] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in neural information processing systems*, 2021. 2, 5

[57] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Computer Vision and Pattern Recognition*, 2023. 1, 2, 5

[58] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991. 5

[59] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 2

[60] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *International Conference on Computer Vision*, 2023. 2, 5

[61] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1

[62] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 1, 2, 3, 5

[63] Bowen Wen and Kostas Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 8067–8074. IEEE Press, 2021. 1, 2, 6, 7

[64] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. Catgrasp: Learning category-level task-relevant grasping in clutter from simulation. *ICRA 2022*, 2022. 1

[65] Bowen Wen, Wenzhao Lian, Kostas E. Bekris, and Stefan Schaal. You only demonstrate once: Category-level manipulation from single visual demonstration. *ArXiv*, abs/2201.12716, 2022. 1

[66] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Muller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. *Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 6, 7

[67] Yitong Xia, Hao Tang, Radu Timofte, and Luc Van Gool. Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 2

[68] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018. 1, 2

[69] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023. 5

[70] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[71] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1

[72] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting neural radiance fields for pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 1, 4, 5

[73] Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. NeRS: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In *Conference on Neural Information Processing Systems*, 2021. 2

[74] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. RelPose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, 2022. 2

[75] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of

deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[76] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7244–7251. IEEE, 2018. 5

[77] Jon Zubizarreta, Iker Aguinaga, and Jose Maria Martinez Montiel. Direct sparse mapping. *IEEE Transactions on Robotics*, 36(4):1363–1370, 2020. 3