ICON: Incremental CONfidence for Joint Pose and Radiance Field Optimization Supplementary Material

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/

In this supplementary material, we include:

1. Video visualizations of (a) the baseline failure modes identified in the main paper (section 1) and (b) ICON predictions on CO3D and self-recorded videos.
2. Per-scene performance breakdown of ICON in the main paper (section 2)
3. Discussions on evaluation differences with NoPE-NeRF [2] and LocalRF [6].
4. Additional benchmarking on remaining categories on CO3D [8] (section 4).
5. A proof of concept experiment ICON for scene-level joint pose + NeRF learning on ScanNet [4] (section 5).
6. Our discussion on the limitations and future directions of the proposed method (section 6).

1. Visualizing failure modes

In the supplementary materials, we provide several video visualizations of the failure modes discussed in section 3 of the main paper and visualized in 2D in Fig.3 in the main paper. Note that the failure modes come from baseline methods that jointly optimize pose and NeRF, where ICON is designed to avoid these failure modes. We provide a catalog here:

- **ToyTruck_Fragmentation.mp4**: we lift the point cloud from 7 frames using the predicted color and depht value. It is the same toytruck used in Fig.3 in the main paper. We observe that each pose produces its own mini radiance field, mutually invisible from each other's viewpoint. Poses fly through this tube flipbook-style, seeing a single toytruck.
- **Apple_GBR.mp4**: we lift one frame to point-cloud using the predicted color and depth value. It is the same scene used in Fig.3 in the main paper. We can observe the Generalized Bas Relief [1] effect, where the apple appears to be rotating to the opposite direction as the camera. At a breaking point, we can observe that the apple is indeed concave in the table.
- **Bench_GBR.mp4**: similar to the apple example, we provide an additional one to show GBR effect with a bench.
- **Toaster_Overlapping_Registration.mp4**: we use marching cube to collect a set of 3D points and visualize the point cloud. It is the same toaster used in Fig.3 in the main paper. We can observe the effect of overlapping registration, where one side is empty due to no coverage from the cameras and the other side is blurry due to overlapped camera coverage.

In addition, we include several video visualization of ICON on CO3D and self-recorded videos. The visualizations are composed of two parts. It begin with a clip showing the original video and shape evolution over training. The shape by converting radiance field from ICON through Marching Cube, which is gradually refined as training continues. The viewpoint to render the shape is decided by the camera pose prediction from ICON. This show the reconstruction quality of ICON as well as the pose predictions, in particular for dynamic objects. These videos are titled with "{object_name}_Evo.mp4".

2. Per-scene performance breakdown

We expand ICON results presented in main paper in section3 on CO3D full scene, CO3D object-only and HO3D [5] to document per-scene performance. Results are summarized in Tab. 1, Tab. 2 and Tab. 3.

3. Evaluation differences

For novel-view synthesis, NoPe and LocalRF use a different protocol: instead of directly aligning the predicted and GT camera (as in BARF, NeRF--, and ICON), test image poses are initiated from the closest training pose (using GT), followed by test-time optimization to refine. This increases view-synthesis quality significantly when pose predictions are poor, boosting NoPe, L2G, and LocalRF PSNR to 16.6, 17.4, and 15.4 respectively. For data, we also observe prior works using short trajectories with $\sim 40^\circ$ max rotation (NoPe), coarse pose initialization (L2G), and long

*Equal contribution.

Category	Scene	ATE	ATE_{rot}	PSNR	SSIM	LPIPS
apple	189_20393_38136	0.027	0.09	24.83	0.74	0.32
ball	123_14363_28981	0.454	2.31	16.43	0.43	0.74
bench	415_57121_110109	0.002	0.12	26.03	0.69	0.33
book	247_26469_51778	0.219	1.41	26.79	0.76	0.30
bowl	69_5376_12833	0.338	2.02	15.33	0.35	0.68
broccoli	372_41112_81867	0.022	0.14	26.40	0.79	0.35
cake	374_42274_84517	0.040	0.31	23.85	0.76	0.26
hydrant	167_18184_34441	0.092	0.69	19.05	0.54	0.49
mouse	377_43416_86289	0.240	1.33	22.33	0.71	0.36
orange	374_42196_84367	0.200	3.86	24.71	0.80	0.35
plant	247_26441_50907	0.190	1.95	16.30	0.43	0.59
remote	350_36761_68623	0.043	0.28	27.08	0.66	0.42
skateboard	245_26182_52130	0.061	0.34	21.37	0.67	0.58
suitcase	109_12965_23647	0.110	1.37	17.77	0.61	0.48
teddybear	34_1479_4753	0.050	0.55	24.08	0.76	0.32
toaster	372_41229_82130	0.240	2.57	20.11	0.53	0.50
toytrain	240_25394_51994	0.170	1.92	19.08	0.66	0.49
toytruck	190_20494_39385	0.010	0.17	27.39	0.87	0.15
Avg		0.138	1.16	22.24	0.65	0.43

Table 1. Per-scene performance of ICON on CO3D full scene evaluation.

Category	Scene	ATE	ATE_{rot}	PSNR	SSIM	LPIPS
apple	189_20393_38136	0.255	1.70	26.59	0.95	0.06
ball	123_14363_28981	0.450	2.54	20.27	0.93	0.09
bench	415_57121_110109	0.183	1.22	24.26	0.80	0.19
book	247_26469_51778	0.174	1.36	24.24	0.89	0.13
bowl	69_5376_12833	0.637	4.66	16.91	0.94	0.09
broccoli	372_41112_81867	0.201	1.65	24.63	0.93	0.09
cake	374_42274_84517	0.058	0.46	21.53	0.91	0.12
hydrant	167_18184_34441	0.150	1.05	23.86	0.92	0.12
mouse	377_43416_86289	0.420	7.09	15.93	0.80	0.31
orange	374_42196_84367	0.387	3.84	29.34	0.98	0.02
plant	247_26441_50907	0.075	0.62	18.28	0.75	0.27
remote	350_36761_68623	0.109	0.71	25.38	0.94	0.09
skateboard	245_26182_52130	0.194	1.50	19.51	0.81	0.18
suitcase	109_12965_23647	0.082	0.78	21.17	0.89	0.18
teddybear	34_1479_4753	0.053	0.42	24.56	0.91	0.10
toaster	372_41229_82130	0.225	1.01	20.79	0.94	0.10
toytrain	240_25394_51994	0.159	1.19	20.35	0.83	0.18
toytruck	190_20494_39385	0.066	0.68	26.46	0.95	0.05
Avg		0.215	1.80	22.45	0.89	0.13

Table 2. Per-scene performance of ICON on CO3D object-only evaluation.

trajectories but little rotation (walking forward) (LocalRF); pairwise pose changes are also small ($< 0.5°$). In contrast, CO3D has 360° rotation with an average 2-4° relative rotation between frames. For evaluation, LocalRF and NoPE sample held-out interpolating frames to test, which are close to adjacent training frames on either side. We follow BARF in holding out the last 10% of frames for test; these extrapolating frames have larger viewpoint differences and are more challenging.

	ATE	ATE_{rot}	Trans	PSNR	CD(cm)
SiS1	0.028	3.80	0.017	19.13	0.23
MC1	0.019	5.90	0.049	14.24	0.41
ABF13	0.064	10.67	0.094	11.79	1.72
GPMF12	0.029	11.23	0.056	16.27	0.38
ND2	0.027	7.18	0.015	20.06	0.50
SM2	0.026	5.56	0.032	13.51	0.85
SMu1	0.017	13.19	0.081	14.46	1.02
AP13	0.058	7.06	0.046	20.42	0.50
Avg	0.033	8.07	0.049	16.24	0.70

Table 3. Per-scene performance of ICON on HO3D evaluation. CD stands for Chamfer Distance, measuring mesh quality.

4. Evaluating ICON on other CO3D categories

In this section, we supplement the results reported in the main paper on CO3D [8]. We add a study using all the remaining 33 categories from CO3D and evaluate on the full scene. This makes it possible for us to include symmetric objects such as vase whose poses are indistinguishable in the object-only evaluation. Since no official subset is specified for these categories, we take top-4 instances from each category with highest camera pose confidence and randomly sample one instance for each category. It is worth noting that the "ground-truth" camera poses are estimated by COLMAP, and may not be 100% accurate, especially these categories are not part of the official benchmarking sets. We use the same (hyper-)parameters as the main paper benchmarking on the 18 categories.

We report the results in Tab 4. We observe that most objects achieve similar results as Tab 1. However, there are a few objects where ICON yields imprecise poses, dragging down the average metrics. We believe there are two causes. First, ICON relies on photometric loss and may suffer from changes in the scenes. Many of the scenes where ICON has ≥ 3 degree rotation error have moving shadows (either object or human), strong lighting change (from the builtin flash of the camera) or reflective surfaces. We show a few examples here in Fig. 1. Second, the groundtruth poses used to evaluate the trajectory are generated by COLMAP, which may not be accurate, especially the categories not included in the official benchmarking sets.

5. Evaluation on ScanNet

ICON focuses our study on object-centric videos such as CO3D and HO3D. However, ICON does not apply specific design tailored towards object that prevents it to work on other types of videos. Here, we include a preliminary study by benchmarking ICON on ScanNet [4]. We randomly sample 10 out of 20 scenes in ScanNet test set and use a clip of 200 frames with a stride of 2. Scenes with NaN value in

Category	Scene	ATE	ATE_{rot}	PSNR	SSIM	LPIPS
backpack	506_72977_141839	0.060	0.42	20.74	0.59	0.42
banana	612_97867_196978	1.691	11.23	13.04	0.15	0.81
baseballbat	375_42661_85494	0.791	7.83	13.92	0.61	0.68
baseballglove	350_36909_69272	0.054	0.72	20.52	0.43	0.62
bicycle	62_4324_10701	0.700	5.94	15.22	0.19	0.69
bottle	589_88280_175252	0.098	1.18	29.59	0.76	0.38
car	439_62880_124254	0.765	4.43	11.40	0.32	0.87
carrot	372_40937_81628	0.873	2.17	20.86	0.63	0.44
cellphone	76_7569_15872	4.725	19.55	13.26	0.30	0.85
chair	455_64283_126636	0.009	0.28	22.77	0.73	0.27
couch	427_59830_115190	0.140	1.64	25.67	0.84	0.29
cup	44_2241_6750	0.453	2.47	23.50	0.60	0.49
donut	403_52964_103416	2.248	11.89	17.60	0.74	0.57
frisbee	339_35238_64092	0.738	3.75	22.34	0.43	0.66
hairdryer	378_44249_88180	0.022	0.16	25.84	0.82	0.33
handbag	406_54390_105616	0.273	2.32	26.51	0.89	0.26
hotdog	618_100797_202003	2.600	7.23	19.78	0.45	0.78
keyboard	375_42606_85350	1.596	7.04	18.54	0.46	0.60
kite	428_60143_116852	0.029	0.36	18.01	0.30	0.74
laptop	378_44295_88252	1.128	7.92	15.04	0.36	0.59
microwave	504_72519_140728	0.023	0.45	21.17	0.61	0.42
motorcycle	367_39692_77422	0.006	0.14	26.52	0.78	0.30
parkingmeter	483_69196_135585	0.136	2.48	17.24	0.56	0.56
pizza	372_41288_82251	0.036	0.26	27.70	0.69	0.42
sandwich	366_39376_76719	0.411	1.67	19.74	0.53	0.51
stopsign	617_99969_199015	3.229	13.81	13.99	0.40	0.72
toilet	605_94579_188112	0.252	5.48	18.53	0.69	0.41
toybus	273_29204_56363	0.057	0.40	23.34	0.65	0.60
toyplane	405_53880_105088	0.020	0.12	22.20	0.53	0.53
tv	48_2742_8095	0.097	0.81	26.32	0.81	0.39
umbrella	191_20630_39388	1.115	5.73	17.35	0.44	0.60
vase	374_41862_83720	0.100	1.27	29.25	0.85	0.28
wineglass	401_51903_101703	1.191	7.80	21.43	0.58	0.53
Avg		0.778	4.21	20.57	0.57	0.53

Table 4. Per-scene performance of ICON on other 33 categories in CO3D full-scene evaluation.

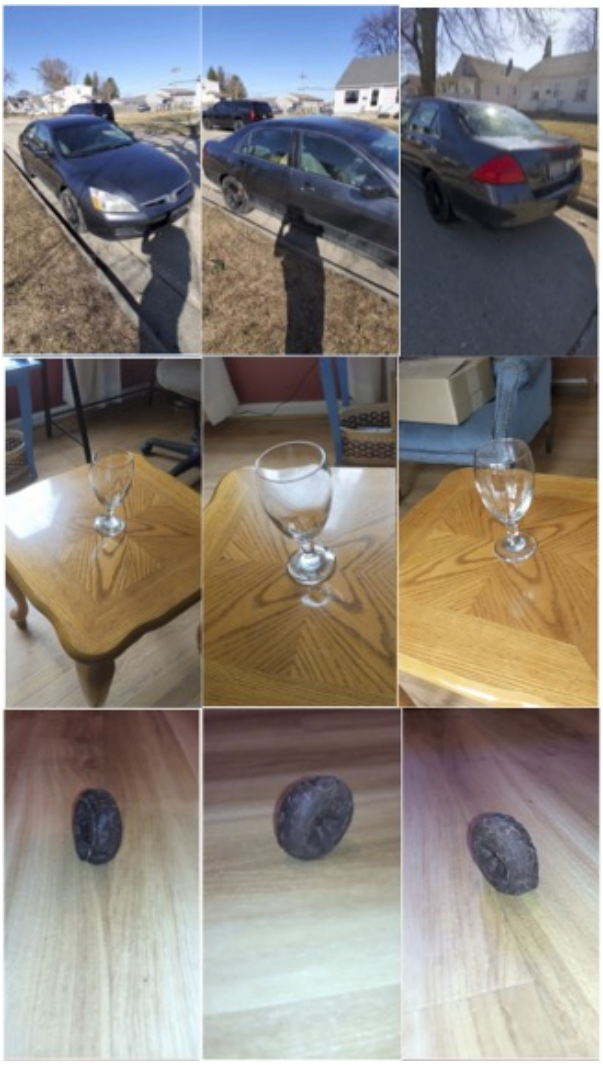

Figure 1. Scenes where ICON produces larger errors. ICON mainly suffer from scenes where photometric loss produces inconsistent supervisions. The **car** example consists of moving human shadow and reflective surface on the car. The **wineglass** example contains transparent surface and light reflections. The **donut** example contains inconsistent lighting, where the flash from the camera generates brighter color in the front and darken the back part. These inconsistencies in different viewpoints cause ICON to produce imprecise camera poses.

camera poses are removed when we sample scenes.

We report camera pose quality following prior works [15] using Relative Pose Error (RPE) on rotation and Absolute Trajectory Error (ATE (m)) for translation. We follow [15] to not use ATE_{rot} because some trajectories in ScanNet has very small translation and aligning the trajectory then evaluate rotation may not be reliable.

We do not change *any* (hyper-)parameters used in CO3D full scene training for ICON to stress test the system on the significantly different scenarios in ScanNet. We include four methods designed to work well on ScanNet for comparison: TartanVO [13], COLMAP [9], DROID-SLAM [11] and current state-of-the-art method ParticleSfM [15]. We note that COLMAP and ParticleSfM may fail to perform well when running only on the short clip, so we run them on the entire video and report the results on the clip. In addition, as noted in [15], since COLMAP often fail on many ScanNet scenes, we use a tuned version following [12].

We report results in Tab 5. Despite having no tuning or change when transferring from CO3D, ICON achieves strong performance on ScanNet compared to the state-of-the-art methods designed to work well on ScanNet style videos. We believe this is a proof-of-concept that ICON can be generalized and adapted to other types of videos.

	TartanVO	DROID	COLMAP	ParticleSfM	ICON
RPE(degree)	1.41	0.56	0.67	0.34	0.47
ATE(m)	0.198	0.066	0.091	0.053	0.092

Table 5. Camera pose evaluation on ScanNet. Despite not optimized for ScanNet scenarios, ICON achieves competitive performance, ranking the second on RPE and third on ATE. The difference between ICON and state-of-the-art method is very small (0.13 degree on rotation and 0.039m on translation)

6. Limitations and future directions

While ICON achieves strong performance to jointly optimize poses and NeRF, it has a few limitations. First, ICON strongly relies on photometric loss as supervision for both NeRF and poses. This relies on the assumption that the color is moderately consistent across different viewpoints. However, this assumption may break in real-world. Although ICON uses confidence to down-weight volumes with inconsistent photometric loss, it will produce imprecise poses (5 to 10 degree rotation error) due to the ambiguity. As shown in Tab 4 and Fig 1, ICON suffers from motion, reflective surfaces, transparency and strong lighting change. We believe leveraging features robust to these changes, such as DINO [3], may help alleviate this issue.

In addition, ICON depends on gradient-based optimization through NeRF [7], which takes hours to train. We believe that combining ICON with more efficient modeling of 3-space will be a promising direction, such as PixelNeRF [14] and FLOW-CAM [10].

References

[1] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *International journal of computer vision*, 1999. 1

[2] Wenjing Bian, Zirui Wang, Kejie Li, Jiawang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. 2023. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 4

[4] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 1, 2

[5] 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. 1

[6] 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. 1

[7] 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. 4

[8] 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

[9] 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. 3

[10] Cameron Smith, Yilun Du, Ayush Tewari, and Vincent Sitzmann. Flowcam: Training generalizable 3d radiance fields without camera poses via pixel-aligned scene flow, 2023. 4

[11] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in neural information processing systems*, 2021. 3

[12] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. NeuralDiff: Segmenting 3D objects that move in egocentric videos. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2021. 3

[13] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. Tartanvo: A generalizable learning-based vo. 2020. 3

[14] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 4

[15] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *European conference on computer vision (ECCV)*, 2022. 3