kakaoenterprise

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

A Solution for the Common Object in 3D Challenge

Many-View Reconstruction Track
NGR-CO3D@ECCV2022

Taewan Ethan Kim

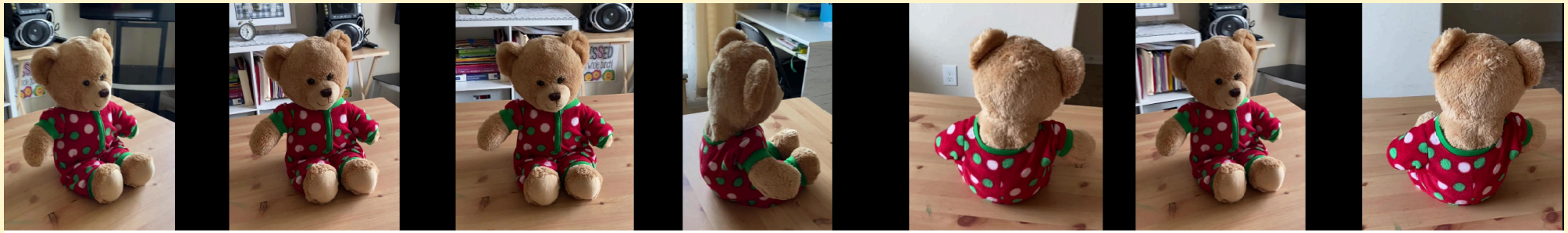
Vision Intelligence,
AI Lab & Service,
Kakao Enterprise

kakaoenterprise

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

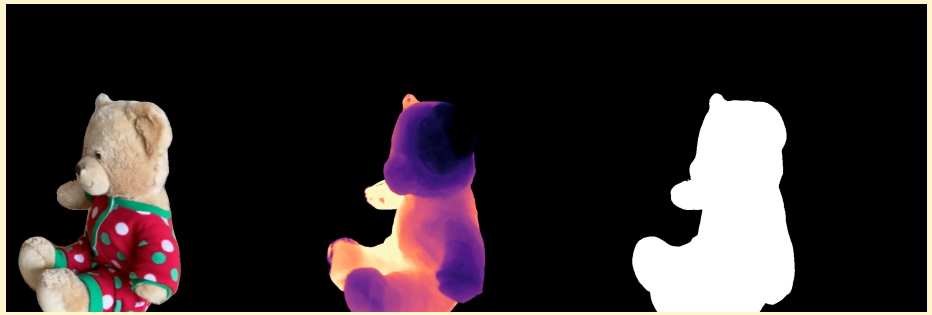
Task Definition

100+ 2D RGBs with Camera poses and ROI masks, provided [1]



An example of Teddybear-dev-0 sequence

Outputs Appearance (RGB), Depth-map (D), and ROI mask (A) for a given test camera poses for 88 (40 and 48 for DEV and TEST) video sequences in total.



An example of our submission output for teedybear-manyview-dev-0

An example of our submission output for bicycle-manyview-test-0

[1] J. Reizenstein et. al, Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction, ICCV, 2021

Our Approach

(At a glance)

We employed the
- Neural Radiance Field (NeRF) [1]
- Tensorial Radiance Field (TensoRF) [2]

We combine both outputs by averaging (late fusion)

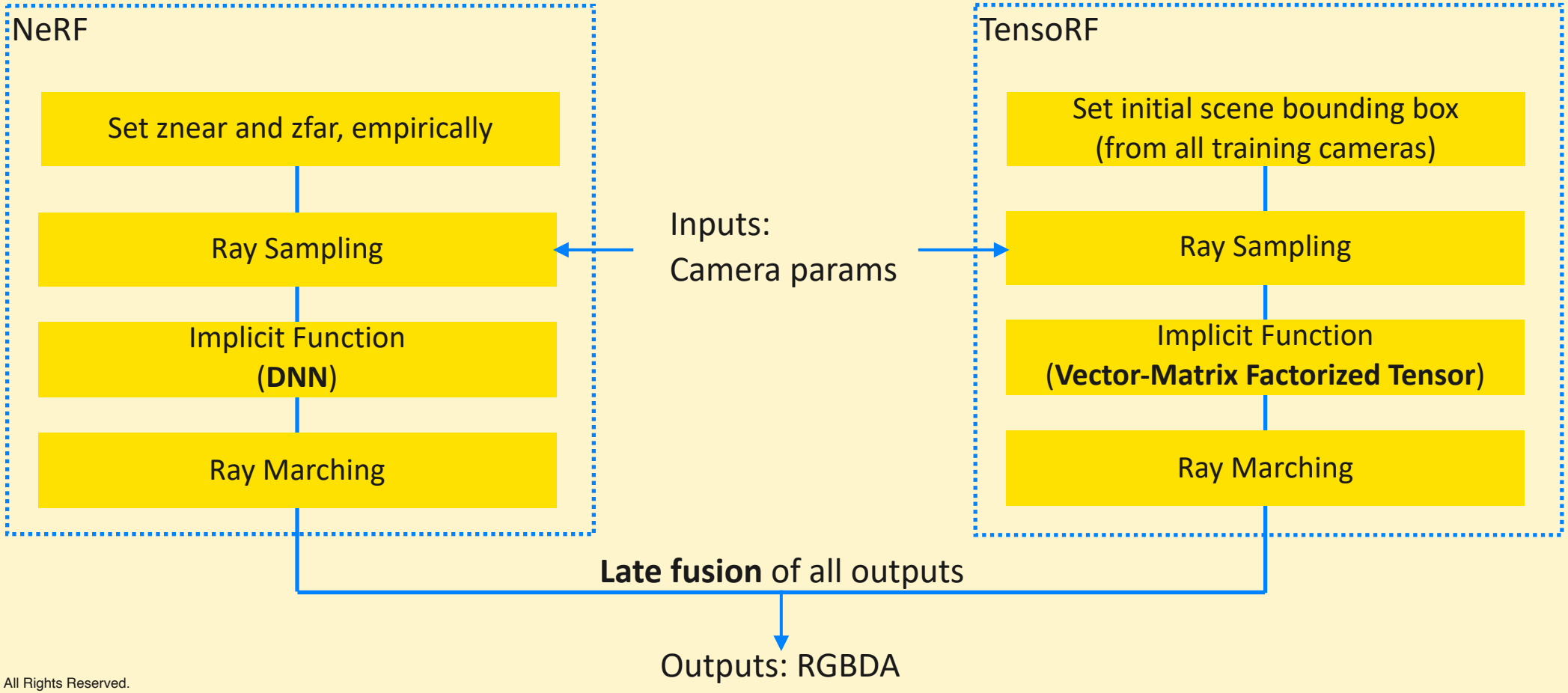
For a better generalization, we applied
- “Sample” Entropy Minimization [3] on densities for NeRF
- L1 sparsity and Total Variation (TV) loss on densities and features for TensoRF

To improve stability of convergence,
- Moving Average of the implicit function (inspired by StyleGAN2 [5])
- Learning-Rate Warmup [4]

[1] B. Mildenhall et. al, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV, 2020
[2] A. Chen et. al, Tensorial Radiance Field, ECCV, 2022
[3] M. Kim et. al, InfoNeRF: Ray Entropy Minimization for Few-Shot Neural Volume Rendering, CVPR, 2022
[4] T. He et. al, Bag of Tricks for Image Classification with Convolutional Neural Networks, CVPR, 2019
[[5] T. Karras et. al, Analyzing and Improving the Image Quality of StyleGAN, CVPR, 2020

Our Approach

(A simple ensemble representation of an object with NeRF and TensoRF)



Our Approach

(Objectives)

RGB loss:

$$\mathcal{L}_{\text{RGB}} = \frac{1}{|\mathcal{M}_{\mathbb{1}}|} \sum_{\mathbf{r}\in\mathcal{R}} \left[\left\| \hat{C}_c(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 + \left\| \hat{C}_f(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 \right] \mathcal{M}(\mathbf{r}) \quad (1)$$

,where $\mathcal{M}(\mathbf{r}) = \begin{cases} 1, & \mathbf{r} \text{ is in the ROI} \\ 0, & \text{otherwise} \end{cases}$ (2)

Note: Our RGB feature branch only focuses on the ROI by introducing $\mathcal{M}(\cdot)$

ROI loss:

$$\mathcal{L}_{\text{ROI}} = \text{BCE}(\hat{\alpha}(\mathbf{r}), \mathcal{M}(\mathbf{r})) \quad (3)$$

,where

$$\hat{\alpha}(\mathbf{r}) = 1 - \prod_{i\in\mathbf{r}} \exp(-\sigma_i \delta_i) \quad (4)$$

Note: $\mathbf{r}$ is sampled at random location of an image so that both positive and negative samples properly produce gradients for rendered alpha mask $\hat{\alpha}(\mathbf{r})$

kakaoenterprise

EUROPEAN CONFERENCE ON COMPUTER VISION
ECCV TEL AVIV 2022

Our Approach

(TensoRF)

In TensoRF, we adapted
- shrinking the scene bounding box, and
- increasing resolution of tensor-grid in early training (as described in the paper [1])

instead of using separate coarse and fine implicit functions with the importance-ray-sampling.

Actually,

We expected the TensoRF could replace our primary implementation of NeRF.
However, in our case, we didn't observe performance improvement compared with our re-implementation of NeRF.

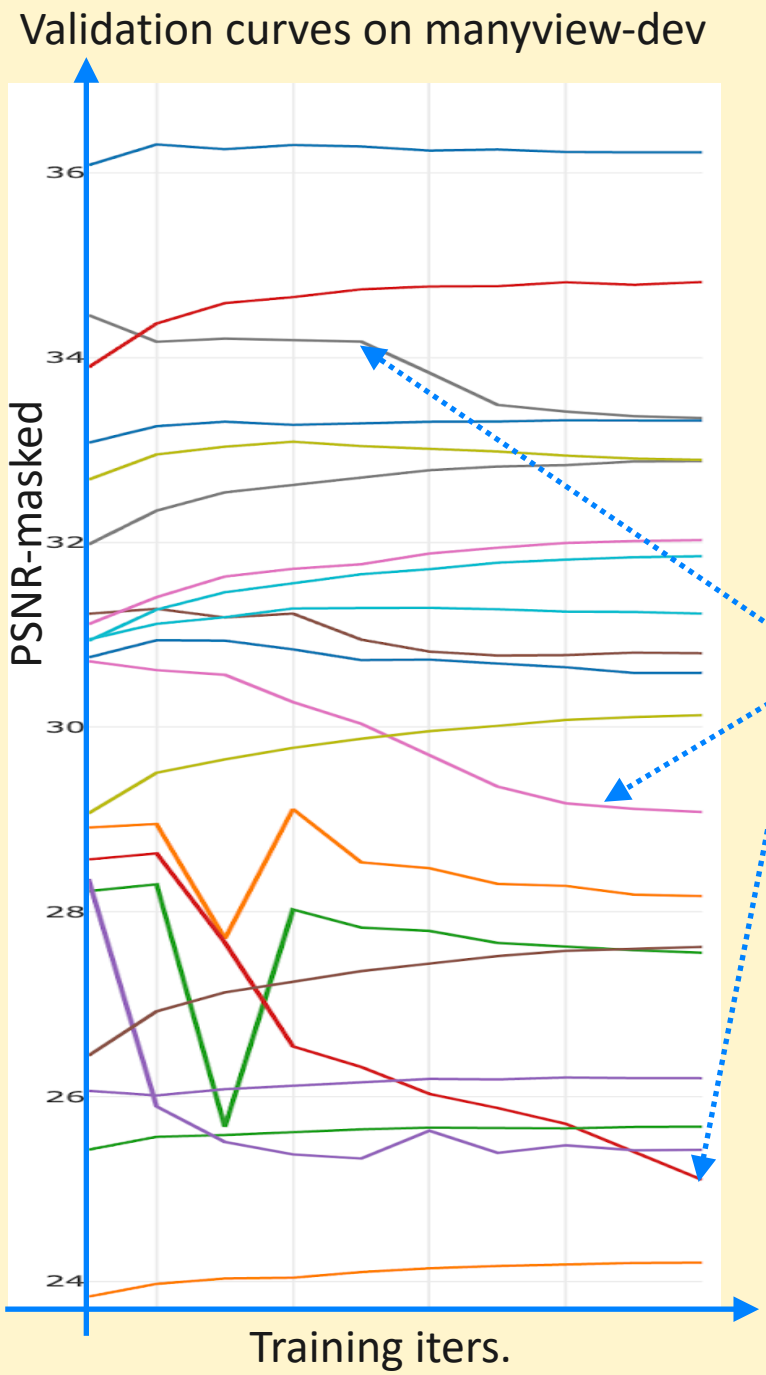
So, we alternatively employed an ensemble approach rather than replacing the NeRF.

[1] A. Chen et. al, Tensorial Radiance Field, ECCV, 2022

kakaoenterprise

Our Approach

(Regularization)



Colored lines denote the PSNR-masked metric of the validation set for randomly selected dev-sequences.

Some dev-sequences show overfitting phenomena even in the 100+view setting.

To mitigate this,
We added (unsupervised) entropy minimization loss for predicted densities as well as the early-stopping.

Our Approach

(Regularization)

Per-Sample-Entropy
(for the fine network of NeRF):

$$\mathcal{L}_{\text{ent}} = \frac{1}{\#\text{rays} \times |\mathbf{r}|} \sum_{\#\text{rays}} \sum_{i \in \mathbf{r}} H_i \tag{1}$$

$$H_i = -p_i \log p_i \tag{2}$$

$$p_i := \alpha_i$$

$$\alpha_i = 1 - \exp(-\sigma_i \delta_i) \tag{3}$$

L1 sparsity and TV loss is the same in [1]

[1] A. Chen et. al, Tensorial Radiance Field, ECCV, 2022

kakaoenterprise

EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

Training Parameters

NeRF
Raysampler:
  #rays: 1024
  #samples-per-ray: 96
  znear/zfar: 3/24 (mouse, manyview_test_0: 12/32)
Optimizer:
  Learning-rate: 0.0005
  optimizer/lr-scheduler: Adam, Cosine-scheduler with warmup
  max-iters: 350000~400000
Implicit_fuction:
  noise-for-density-output: 0.05
  moving-average: 0.9999
  implicit_function, positional encoding: same as [1]
Loss:
  weight-rgb: 1.0
  weight-binary-cross-entropy: 1.0
  weight-entropy: 1e-5~5e-5
Jitter-camera:
  angle-std: np.pi * 0.0625
clip_gradient_norm: 1.0

TensoRF
Raysampler:
  #rays: 4096
  #samples-per-ray: 384
  znear/zfar: 3/24 (mouse, manyview_test_0: 12/32)
Optimizer:
  Learning-rate: 0.02, 0.001
  optimizer/lr-scheduler: Adam, Cosine-scheduler with warmup
  max-iters: 350000
Implicit_fuction (Vector-Matrix Factorization):
  Init-resolution: 128
  Final-resolution: 1024
  grid-upsampling-iters: [30000, 50000, 65000, 91300, 116200]
  scene-bbox-shrinking-iters: [20000, 40000]
  moving-average: 0.9999
Loss:
  weight-rgb: 1.0
  weight-binary-cross-entropy: 1.0
  weight-entropy: 5e-5
  weight-TV: 1.0
  L1-sparsity: 5e-5
Jitter-camera:
  angle-std: np.pi * 0.0625
clip_gradient_norm: 0.2

We applied the same hyper-parameters for almost all categories.
Simultaneous training of all categories at once via distributed-package of PyTorch [2].
“Raysamplers" shipped in Pytorch3d [3], “VMFactorizedVoxelGrid” class in Implictron project [4]
A100 servers (up to ~136 gpu cards)
Training time: ~1.8 days (including validating)

[1] B. Mildenhall et. al, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV, 202
[2] pytorch: https://pytorch.org/, accessed at 24/10/2022
[3] pytorch3d, https://pytorch3d.org/ , accessed at 24/10/2022
[4] Implictron: https://github.com/facebookresearch/pytorch3d/tree/main/projects/implicitron_trainer

kakaoenterprise

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

Training Parameters

Note that

- We do not use any external datasets or pertained models.
- We do not use any 3D information i.e. depth-map as a supervision

kakaoenterprise

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

Quantitative results

manyview-dev

Employed method	PSNR-Masked (Internal evaluation)
Our Baseline NeRF	31.24 (N / A)
+ Moving Average + Warmup	31.65 (+0.41)
+ Sample-Entropy	31.85 (+0.20)
+ Camera Jittering (Extrinsic Params)	32.00 (+0.15)
+ Ensemble (NeRF + TensoRF)	32.30 (+0.30)

Rank	Participant team	psnr_masked (↑)	psnr_fg (↑)	psnr_full_image (↑)	depth_abs_fg (↓)	iou (↑)
1	**Kakao Enterprise** KE (32.10829284719358)	32.11	27.37	6.16	0.47	0.97

manyview-test

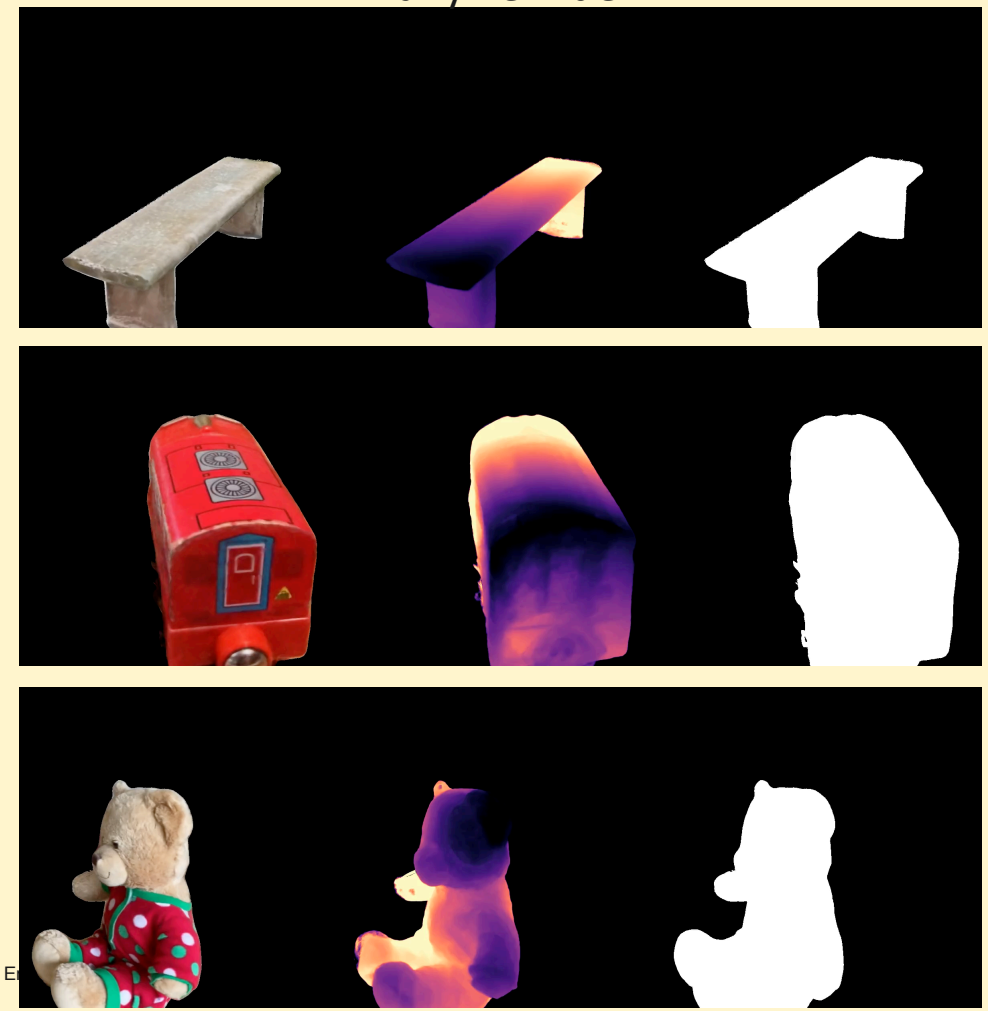
Rank	Participant team	psnr_masked (↑)	psnr_fg (↑)	psnr_full_image (↑)	depth_abs_fg (↓)	iou (↑)
1	**Kakao Enterprise** KE (30.886906286593884)	30.89	24.61	6.29	0.59	0.96

kakaoenterprise

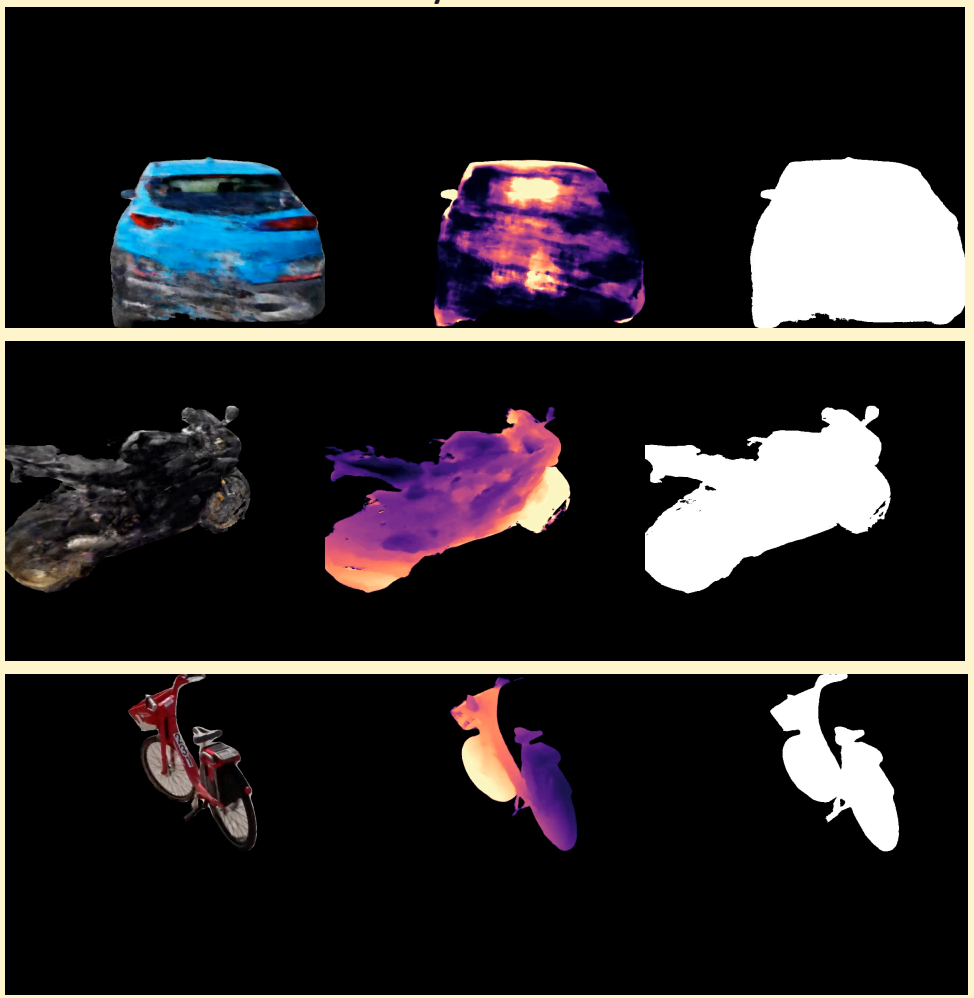
ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

Qualitative results

manyview-dev



manyview-test



kakaoenterprise

EUROPEAN CONFERENCE ON COMPUTER VISION
TEL AVIV 2022

Acknowledgement

We special thanks to our "***deepflow***" team for building and maintaining our machine learning infrastructure.

We special thanks to "***Junha Hyung***" for literature survey, and helpful discussion.

Finally, we special thanks to "***David Novotny***" for his ***endless support in a bunch of troubleshooting***.

EUROPEAN CONFERENCE ON COMPUTER VISION
TEL AVIV 2022

Thank you so much for your listening.

Feel free to contact me, if you have any questions!!!

ethan.y@kakaoenterprise.com

kakaoenterprise

https://kakaoenterprise.github.io/

https://kakaoenterprise.com/

We are looking for someone with passion of this area!!!

kakaoenterprise

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

Quantitative results

Manyview-test

Rank	Participant team	psnr_masked (↑)	psnr_fg (↑)	psnr_full_image (↑)	depth_abs_fg (↓)	iou (↑)
1	KE (30.886906286593884)	30.89	24.61	6.29	0.59	0.96
2	MetaAI-CO3D (NeRF (Implicitron)) B	30.31	23.54	6.28	0.51	0.94
3	MetaAI-CO3D (NeRFormer (Implicitron)) B	26.14	20.32	6.38	0.81	0.77

kakaoenterprise

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

Quantitative results

Manyview-dev

Rank	Participant team	psnr_masked (↑)	psnr_fg (↑)	psnr_full_image (↑)	depth_abs_fg (↓)	iou (↑)
1	KE (32.10829284719358)	32.11	27.37	6.16	0.47	0.97
2	MetaAI-CO3D (NeRF (Implicitron)) B	31.19	23.73	6.16	0.50	0.96
3	MetaAI-CO3D (NeRFormer (Implicitron)) B	30.53	22.98	6.15	0.46	0.97
4	MetaAI-CO3D (NeRF+WCE (Implicitron)) B	30.13	22.60	6.14	0.69	0.93
5	MetaAI-CO3D (SRN+WCE w/o Pos. Embed) B	27.59	20.01	6.13	0.62	0.91
6	MetaAI-CO3D (SRN w/o Pos. Emebed) B	27.24	19.68	6.15	0.73	0.89
7	MetaAI-CO3D (SRN+WCE (Implicitron)) B	24.52	17.09	6.11	0.64	0.80
8	MetaAI-CO3D (SRN (Implicitron)) B	24.33	17.00	6.12	0.64	0.80
9	MetaAI-CO3D (IDR (Implicitron)) B	24.06	16.10	5.92	0.55	0.61

kakaoenterprise

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

Workshop Schedule

SCHEDULE

Google Calendar: link (or view it here)

TIME: 08:45 - 18:15 Israel Time, Monday, Oct 24, 2022

LOCATION: Salon A - David Intercontinental Hotel, Tel Aviv and online via ECCV Platform

The times below are automatically converted to the time zone of your browser.

The workshop starts on Mon Oct 24 2022 14:45:00 GMT+0900 (Korean Standard Time)

TIME	SPEAKER	TITLE
14:45 - 15:00	David Novotny	"Opening and the CO3D Challenge"
15:00 - 15:35	Ben Mildenhall	"Modeling Light for View Synthesis"
15:35 - 16:10	Angjoo Kanazawa	"Towards Dynamic 4D Capture -- Practices and Recommendations"
16:10 - 16:40		Coffee Break
16:40 - 17:15	Niloy Mitra	"Learning a Space of Dense Surface-to-Surface Maps"
17:15 - 17:50	Thomas Müller	"Instant Neural Graphics Primitives"
17:50 - 18:05	Taewan Ethan Kim	CO3D Many-view Winner Presentation
18:05 - 18:20	Zhizhuo Zhou	CO3D Few-view Winner Presentation
18:20 - 20:00		Lunch Break
20:00 - 20:35	Lourdes Agapito	TBA
20:35 - 21:10	Christian Rupprecht	"(De)-Rendering 3D Objects in the Wild"
21:10 - 21:45	Deva Ramanan	"Reconstructing Deformable Objects from Monocular Videos"
21:45 - 22:15		Coffee Break
22:15 - 22:50	Andreas Geiger	"Constraining 3D Fields for Reconstruction and View Synthesis"
22:50 - 23:25	Yaser Sheikh	"Neural Rendering for Photorealistic Telepresence"
23:25 - 24:05		Panel Discussion
24:05 - 24:15		Closing