Neural Feature Fusion Fields

3D Distillation of Self-Supervised 2D Image Representations

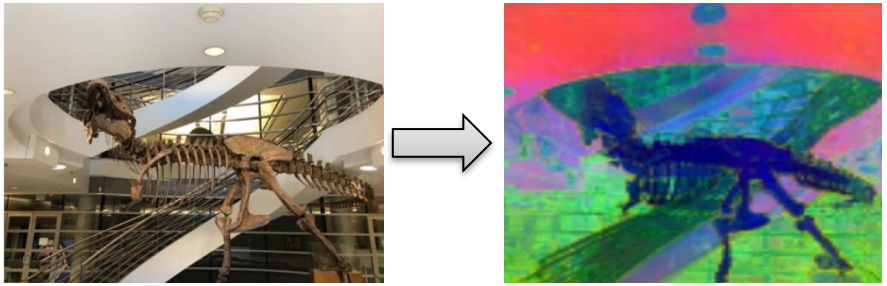
Vadim Tschernezki, Iro Laina, Diane Larlus, Andrea Vedaldi

Outline

- Motivation
- Background
  - 2D Feature Extraction
  - Neural Rendering
  - Semantic NeRF
- Neural Feature Fusion Fields (N3F)
- Results
  - Qualitative & quantitative
- Limitations & future work

Motivation

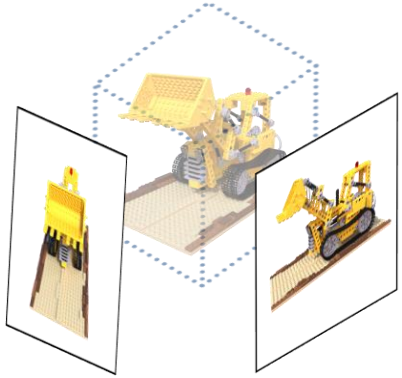
Semantic scene understanding



[2]

Feature extraction

3D scene understanding



[15]

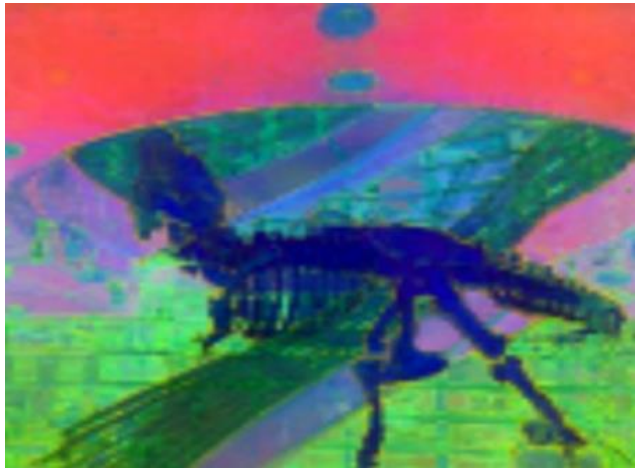
Neural rendering

- Each idea helps improving the other one
- Same input domain

→ combine both methods

Feature Extraction - DINO

- Operates in 2D
- Architecture based on Vision Transformer (ViT)
- Self-supervision
- Self-distillation



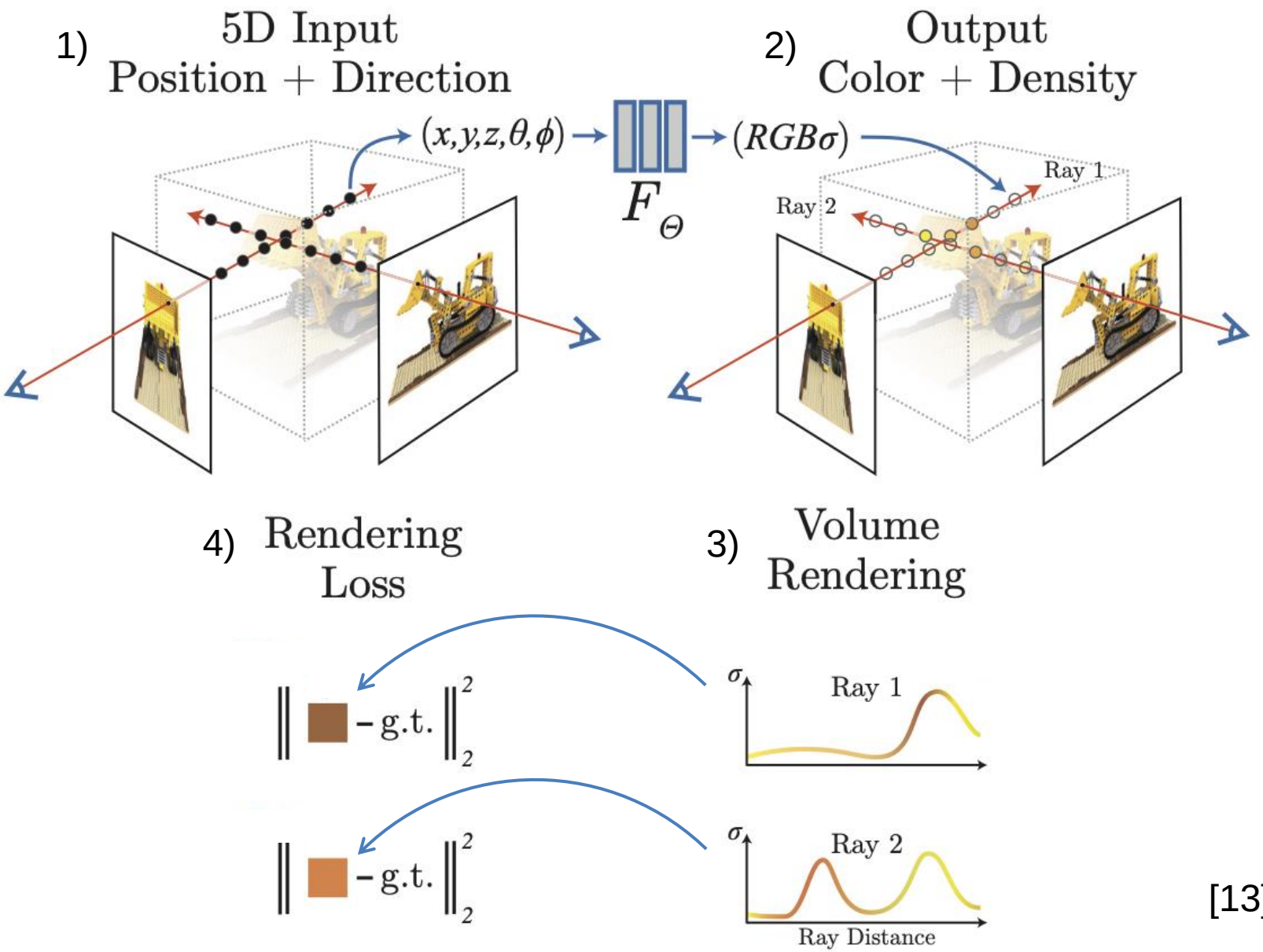
[2]

DINO – results

[11]

- Other 2D feature extractors:
  - MoCo-v3, DeiT

Neural Rendering - NeRF



[13]

NeRF - results

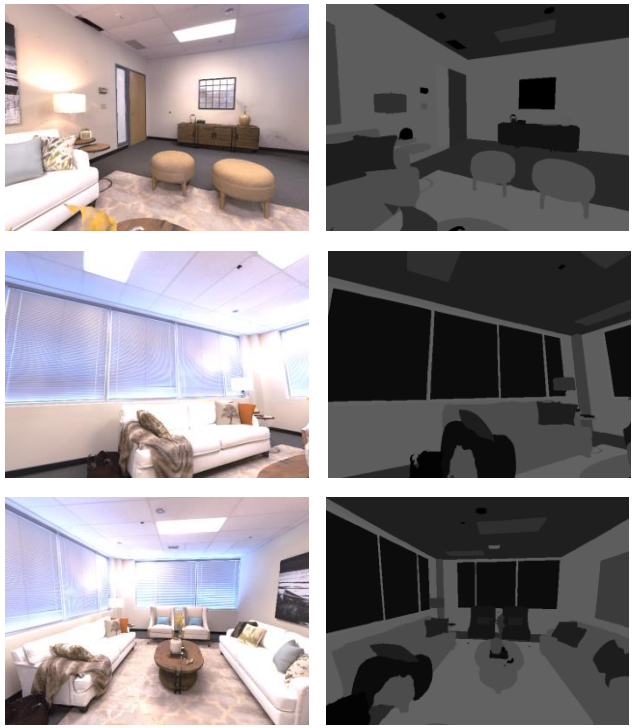
[15]

[15]

Semantic NeRF

RGB

Semantic labels

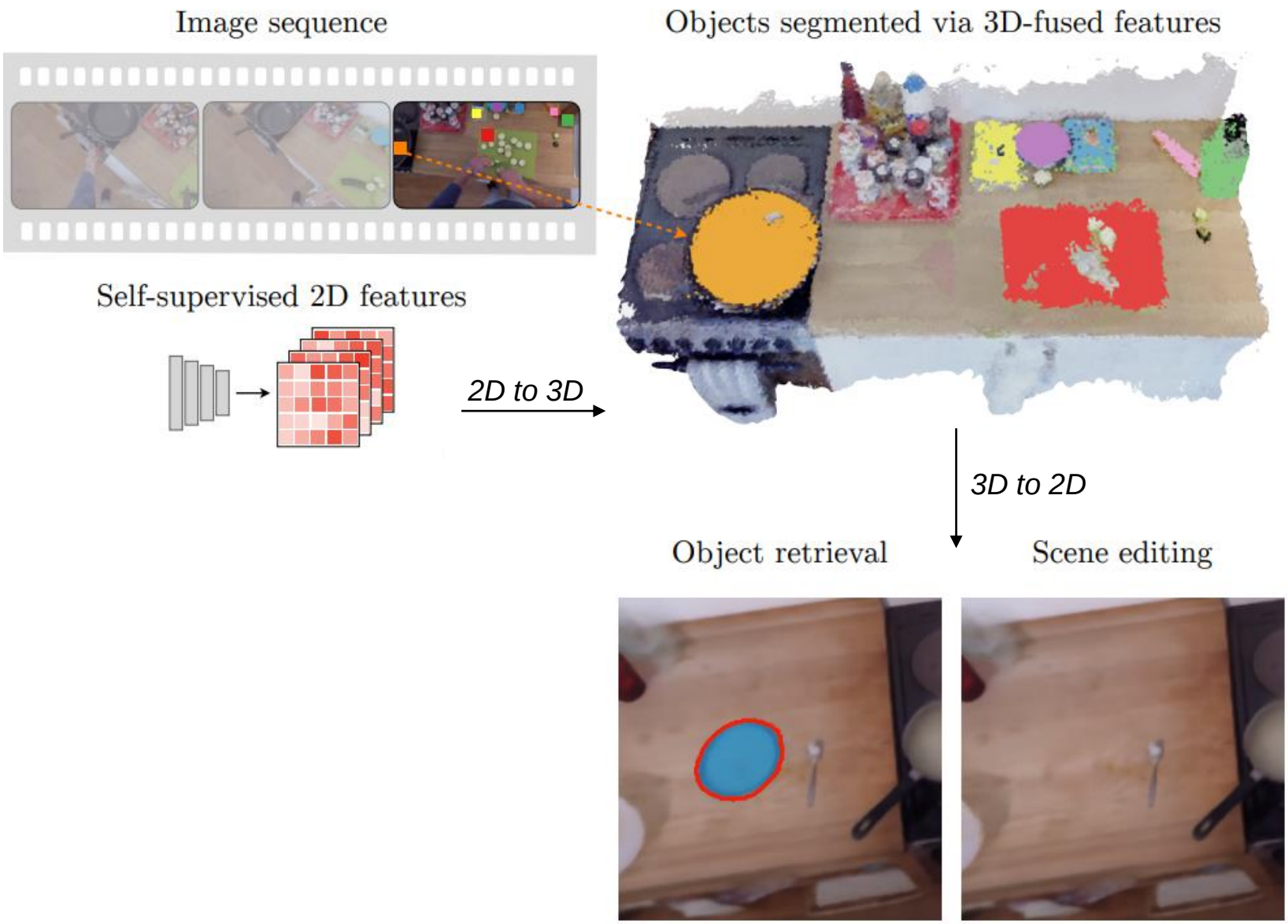


Input [9]

Semantic 3D reconstruction

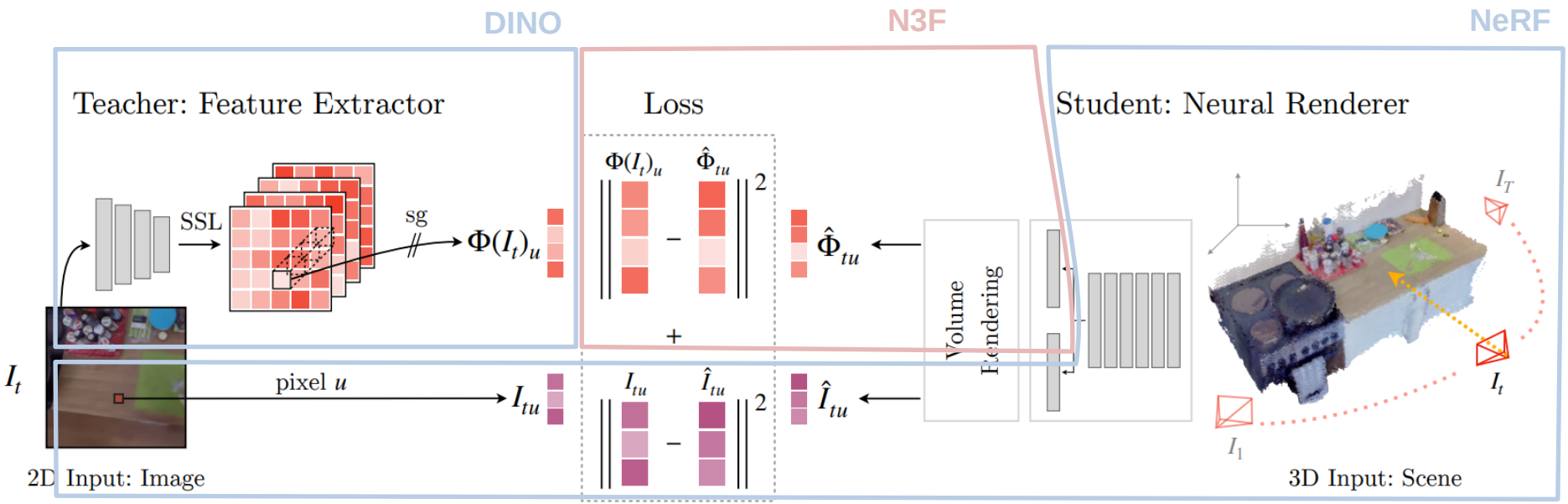
Output [9]

Neural Feature Fusion Fields (N3F) - idea



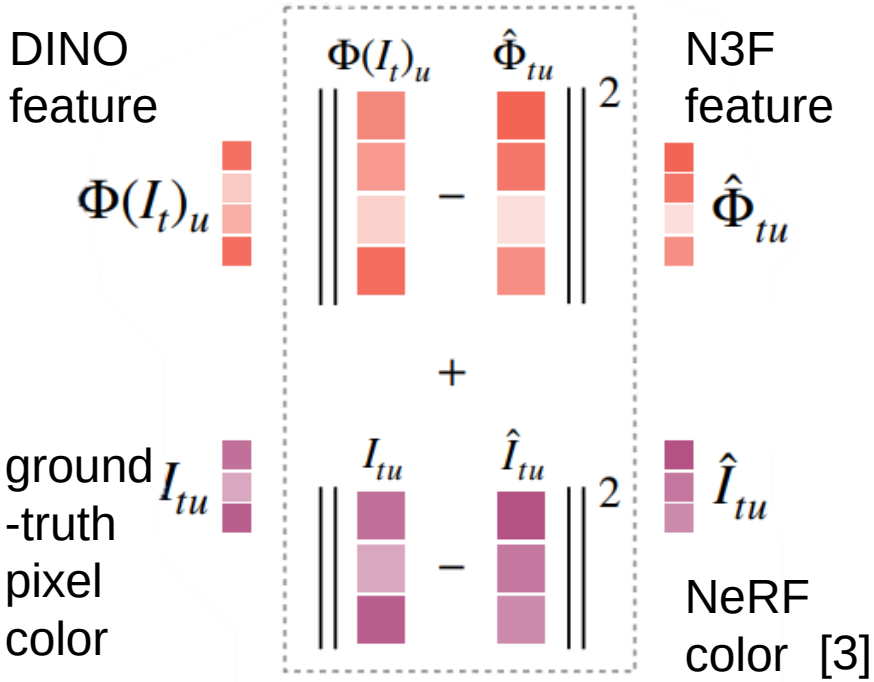
[3]

N3F - method


DINO
N3F
NeRF
Teacher: Feature Extractor
Loss
Student: Neural Renderer
SSL
sg
$\Phi(I_t)_u$
$\hat{\Phi}_{tu}$
I_t
pixel u
I_{tu}
$\hat{I}_{tu}$
Volume Rendering
I_T
I_1
2D Input: Image
3D Input: Scene

[3]

N3F - method



- N3F also handles the feature image $\widehat{\Phi}_t$

Final loss function:

$$\sum_t \left\|\hat{I}_t - I_t\right\|^2 + \lambda \left\|\widehat{\Phi} - \Phi(I_t)\right\|^2$$

Two main contributions:

- Regularization of feature prediction
- Open-world knowledge

Experiments

Experiments

Setup:

1) Simple static scenes (LLFF dataset):
   - N3F with NeRF

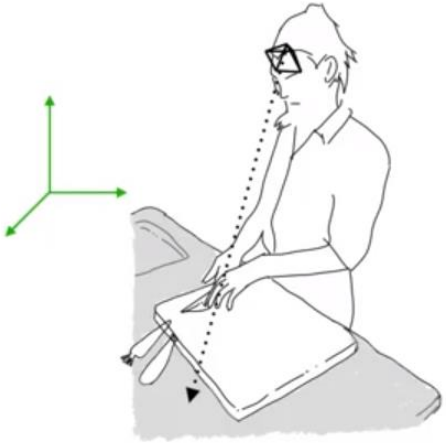
[4]

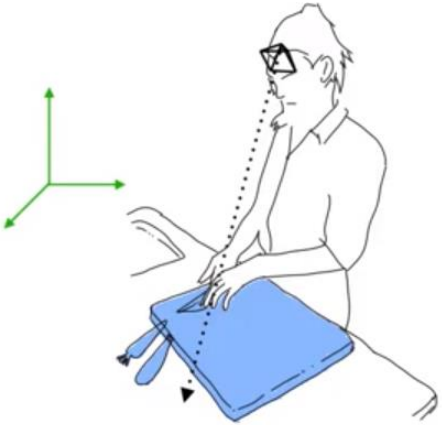
2) Egocentric videos (EPIC-KITCHENS):
   - N3F with NeuralDiff

[7]

NeuralDiff

Neural renderer for egocentric videos





[6]

NeRF

Applications

- Given: collection of images + query image patch

Image: I_t

Patch: R_t

[3]

Feature descriptor:

$$\Phi(I_t)_{R_t}^{avg} = \frac{1}{|R_t|} \sum_{u \epsilon R_t} \Phi(I_t)_u$$

- 2D object retrieval
- 3D object segmentation
- Scene editing
- Amodal segmentation

2D object retrieval - quantitative results

Method	S01	...	S10	Average	(abs gain)
DINO [4] [ViT-B/8]	75.75	...	65.79	64.35	
N3F (DINO)	83.64		82.17	**76.26**	(+11.91)
DINO [4] [ViT-B/16]	77.37	...	62.19	60.15	
N3F (DINO)	88.61		84.93	**78.84**	(+18.69)
MoCo-v3 [8] [ViT-B/16]	70.73	...	60.12	57.51	
N3F (MoCo-v3)	86.67		83.21	**76.15**	(+18.64)
DeiT [58] [ViT-B/16]	55.27	...	52.48	47.51	
N3F (DeiT)	86.02		83.12	**74.53**	(+26.82)

[3]

2D object retrieval – qualitative results

- Calculate feature values of target images
- Match with query feature descriptor

Query

N3F

Target

DINO

[3]

3D object segmentation

- N3F features can be used “out of the box”
- Segment every point that matches queried mean feature descriptor

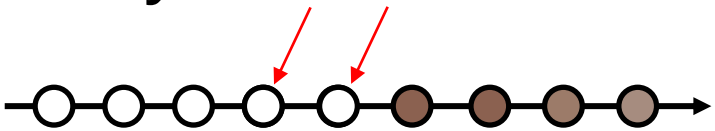
Query

3D Retrieval

[3]

Scene editing

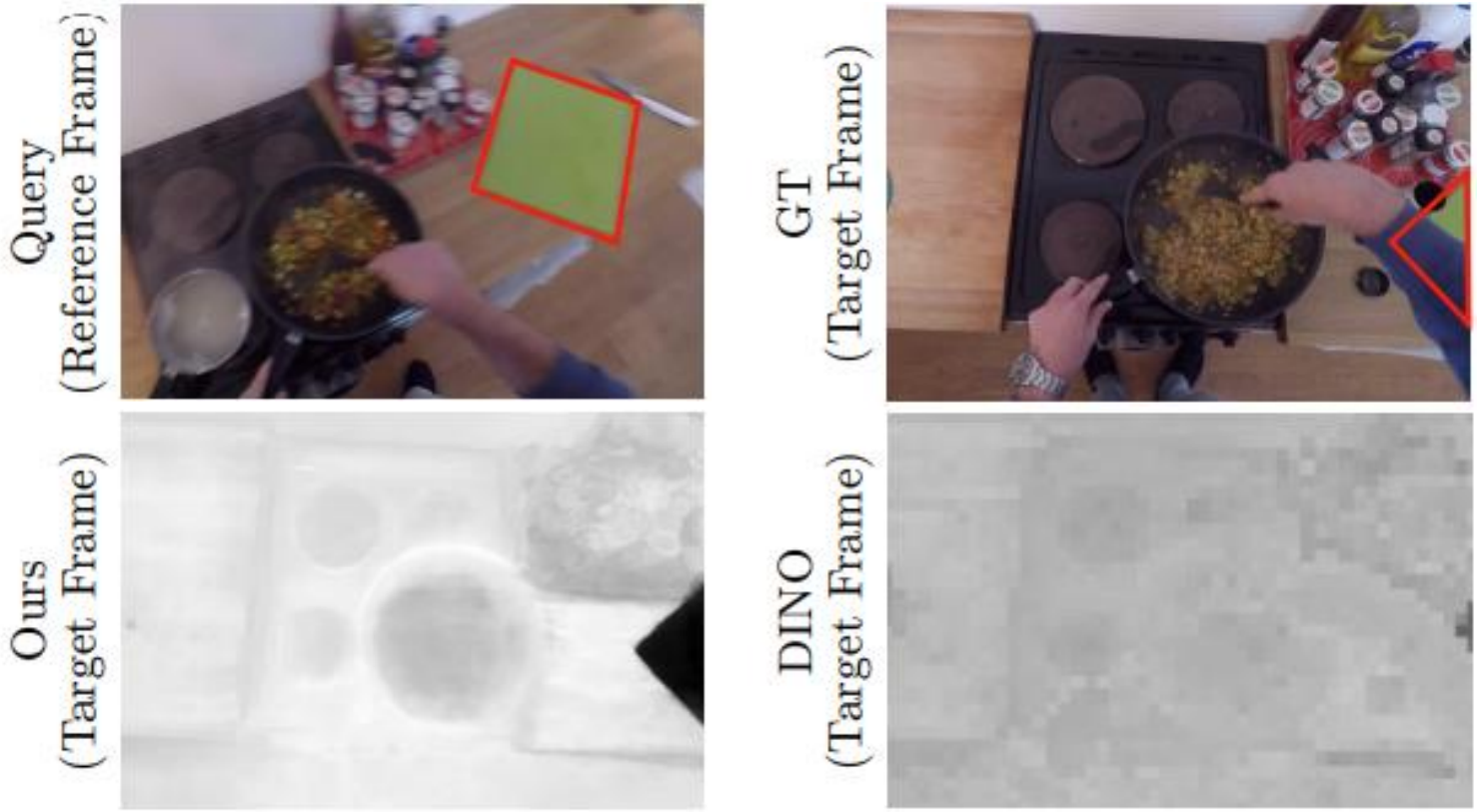
- Remove items from images
- Same as 3D object retrieval
  - ↳ additionally: set volume density to zero



[3]

Amodal segmentation

- Segmenting full extent of an object
- Similar as 3D object retrieval
- Set occupancies to zero for regions that are dissimilar



[3]

Limitations & Future work

Limitations

- Upper bound of performance
- Self-supervised features group related objects
- Reliability on unsupervised machine learning
- Geometry errors by neural renderer

Future work

- Cross-instance correspondence
- Applying N3F in different setups
  - Feature extractor & neural renderer
  - Application domain
- Overcoming time & memory constraints

References

[1] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-place scene labelling and understanding with implicit scene representation. In Proc. ICCV, pages 15838–15847, 2021.

[2] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing NeRF for editing via feature field distillation. arXiv preprint arXiv:2205.15585, 2022.

[3] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi, Neural feature fusion fields: 3D distillation of self-supervised 2D image representations," in Proc. of Joint 3DIM/3DPVT Conference (3DV), 2022.

[4] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng, 2020, NeRF

[5] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In NeurIPS Deep Learning and Representation Learning Workshop, 2015.

[6] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi 2021, NeuralDiff: Segmenting 3D objects that move in egocentric videos, https://www.robots.ox.ac.uk/~vadim/neuraldiff/

[7] Vincent Sitzmann, Michael Zollhofer, and Gordon Wetzstein. Scene representation networks: Continuous 3Dstructure-aware neural scene representations. Proc. NeurIPS, 2019.

[8] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann 2022, Decomposing NeRF for editing via feature field distillation, https://pfnet-research.github.io/distilled-feature-fields/

[9] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison 2021, Semantic-NeRF: Semantic Neural Radiance Fields, https://github.com/Harry-Zhi/semantic_nerf

[10] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In Computer Graphics Forum, volume 41, pages 703–735, 2022.

[11] Piotr Bojanowski, Mike Rabbat, Armand Joulin, Nicolas Ballas, Mathilde Caron, Mahmoud Assran 2021, Advancing the state of the art in computer vision with self-supervised Transformers and 10x more efficient training, https://ai.facebook.com/blog/dino-paws-computer-vision-with-self-supervised-transformers-and-10x-more-efficient-training/

References

[12] Leon Sick 2021, Paper explained: DINO — Emerging Properties in Self-Supervised Vision Transformers, https://towardsdatascience.com/paper-explained-dino-emerging-properties-in-self-supervised-vision-transformers-f9386df266f1

[13] 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 Proc. ECCV, pages 405–421, 2020.

[14] Yannic Kilcher 2021, DINO: Emerging Properties in Self-Supervised Vision Transformers (Facebook AI Research Explained), https://www.youtube.com/watch?v=h3ij3F3cPIk&ab_channel=YannicKilcher

[15] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng 2020, NeRF: Neural Radiance Fields, https://github.com/bmild/nerf

[16] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. Computer Graphics Forum, 2022.

[17] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. arXiv preprint arXiv:2203.15224, 2022.

[18] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. NeuralDiff: Segmenting 3D objects that move in egocentric videos. In Proc. 3DV, pages 910–919, 2021.

[19] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In Yoshua Bengio and Yann LeCun, editors, Proc. ICLR, 2015.

[20] Zhongzheng Ren, Aseem Agarwala, Bryan Russell, Alexander G Schwing, and Oliver Wang. Neural volumetric object selection. In Proc. CVPR, pages 6133–6142, 2022.

[21] Representing Scenes as Neural Radiance Fields for View Synthesis, accessed 22nd November 2022, https://www.matthewtancik.com/nerf

[22] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proc. ICCV, pages 9650–9660, 2021.