Improving 2D Feature Representations by 3D-Aware Fine-Tuning

Yuanwen Yue[1], Anurag Das[2], Francis Engelmann[1,3],
Siyu Tang[1], and Jan Eric Lenssen[2]

[1] ETH Zurich [3] Google
[2] Max Planck Institute for Informatics, Saarland Informatics Campus

In the appendix, we provide (1) experiments with other DINOv2 ViT variants (Appendix A) (2) experiments on more tasks and heads (Appendix B) (3) experiments on impact of feature dimensions for linear probing (Appendix C) (4) more visualization and K-Means clustering of features (Appendix D).

A Experiments With More DINOv2 ViT Variants

To demonstrate that the effectiveness of our 3D-aware fine-tuning is agnostic to DINOv2 architecture variants, we conduct additional experiments using the ViT-Base architecture with a feature dimension of 768. We show the results of semantic segmentation and depth estimation across multiple in-domain and out-of-domain datasets in Tab. 1 and Tab. 2, respectively. We observe a similar trend of improvement with the ViT-B architecture. For example, on the fine-tuning dataset ScanNet++, incorporating our fine-tuned features brings an improvement of 3.47% mIoU on semantic segmentation and reduces 0.03 RMSE on depth estimation. On other indoor datasets NYUv2 and out-of-domain dataset ADE20k, our 3D-aware fine-tuning consistently improves the original DINOv2. This experiment indicates that our 3D-aware fine-tuning is applicable to different ViT architectures and readily benefits downstream tasks.

Table 1: Results of ViT variants on semantic segmentation. Our 3D-aware fine-tuning yields consistent improvements on semantic segmentation for both ViT-S and ViT-B architectures.

		ScanNet++ [7]			NYUv2 [6]			ADE20k [8]		
Method	Arch.	mAcc (↑)	mIoU (↑)	aAcc (↑)	mAcc (↑)	mIoU (↑)	aAcc (↑)	mAcc (↑)	mIoU (↑)	aAcc (↑)
DINOv2 [3]	ViT-S	40.84	30.19	80.25	76.88	65.55	82.43	56.74	44.28	79.73
+ Ours	ViT-S	**43.4**	**32.76**	**83.54**	**80.52**	**67.5**	**83.37**	**58.71**	**45.93**	**81.05**
DINOv2 [3]	ViT-B	42.99	32.72	82.05	80.56	68.45	84.03	59.11	47.16	80.79
+ Ours	ViT-B	**46.35**	**36.19**	**85.5**	**80.58**	**70.56**	**85.72**	**62.18**	**49.5**	**82.52**

Table 2: Results of ViT variants on depth estimation. Our 3D-aware fine-tuning yields consistent improvements on depth segmentation for both ViT-S and ViT-B architectures.

		ScanNet++ [7]		NYUv2 [6]		KITTI [2]	
Method	Arch.	RMSE (↓)	Rel (↓)	RMSE (↓)	Rel (↓)	RMSE (↓)	Rel (↓)
DINOv2 [3]	ViT-S	0.3742	0.2836	0.4423	0.1392	3.0322	0.0965
+ Ours	ViT-S	**0.3361**	**0.2401**	**0.4198**	**0.1300**	**2.9125**	**0.0891**
DINOv2 [3]	ViT-B	0.3439	0.2576	0.3986	0.1218	2.9071	0.095
+ Ours	ViT-B	**0.3174**	**0.2324**	**0.3802**	**0.1171**	**2.7923**	**0.0897**

Table 3: Results on image classification. Our features do not improve image classification results.

Method	Acc. (↑)
DINOv2 [3]	**80.02**
+ Ours	80.00

Table 4: Results with DPT head on depth estimation. Beyond linear probing, we evaluate with DPT head for depth estimation and observe consistent improvement.

Method	RMSE (↓)	Rel (↓)
DINOv2 [3]	0.3027	0.2149
+ Ours	**0.2830**	**0.1936**

B Experiments on More Tasks and Heads

Image classification. We additionally evaluate our approach with DINOv2 small on image classification. We train a linear probing on ImageNet-1K [5] for 12500 iterations on a single GPU. As shown in tab. 3, our features do not improve image classification results. This is expected as classification mainly relies on `CLS` token of ViT while our method aims to improve image patch features.

DPT head. Beyond linear probing, we evaluate DINOv2 small with the DPT head [4] for depth estimation on ScanNet++. In comparison with the linear probing results (Tab. 2 in the main paper), the DPT head improves both results and our features are still helpful in this setup (see Tab. 4). This demonstrates that improvement brought the 3D-aware features is not limited to linear probing but also applicable to more complex heads.

C Experiments on Feature Dimensions

We concatenate original 2D features with our fine-tuned features, which will introduce increased feature dimension. In this experiment, we compare with DINOv2 small with duplicate features for linear probing of semantic segmentation and depth estimation on ScanNet++. As shown in Tab. 5, simply duplication ② only leads to little improvement compared with incorporating our fine-tuned features ③. This verifies that it is not the number of feature dimensions that leads to improvement.

Table 5: Results of duplicating DINOv2 features for linear probing. We verify that it is not the number of feature dimensions that leads to improvement by showing that simple duplication of original features does not help.

	F_{dim}	mIoU (↑)	RMSE (↓)
① DINOv2	384	30.19	0.3742
② DINOv2 × 2	768	30.31	0.3676
③ DINOv2 + Ours	768	**32.76**	**0.3361**

D Visual Analysis of Features

We train feature Gaussians and conduct 3D-aware fine-tuning on ScanNet++. In Fig 1, we visualize the features rendered by pre-trained feature Gaussians (4^{th} column), features of DINOv2 (2^{nd} column) and our fine-tuned features (3^{rd} column). The colors of features in all visualizations are produced using principle component analysis (PCA). The standard DINOv2 features suffer from noise and rough object boundaries. After lifting those features to 3D by training feature Gaussians, we observe the rendered features enjoy cleaner and sharper object boundaries. We then fine-tune DINOv2 using those rendered features, which results in compact and clean feature representations.

Although the fine-tuning is only conducted on ScanNet++, we observe the resulting fine-tuned DINOv2 can generalize to other indoor datasets (*e.g.* NYUv2 and ScanNet) and produces cleaner feature maps and more pronounced structure details (Fig. 2). Similar patterns can also be found in out-of-domain datasets (*e.g.* Pascal VOC, ADE20k and KITTI), as shown in Fig. 3. Visualizations of these feature representations indicate that 3D-aware fine-tuning is helpful and transferable. We observe the improvements are mainly reflected in two aspects: (1) cleaner and more compact feature maps. (2) clearer object boundaries and structured details emerge.

Feature clustering. We also use a simple K-Means clustering to directly examine the semantic concepts encoded in the feature representations. We show the K-means clustering results in Fig. 4. The improvements in our features are directly reflected in those simple clustering results. As shown in Fig. 4, the K-Means results of DINOv2 (3^{rd} column) are strongly affected by artifacts and noise. By contrast, our clustering results (5^{th} column) are much cleaner and more compact. In addition, we observe the PCA features and K-Means clustering of our 3D-aware fine-tuned features exhibit higher temporal consistency than the standard DINOv2 features. Please check our demos on our project page to see the full visualizations of video sequences.

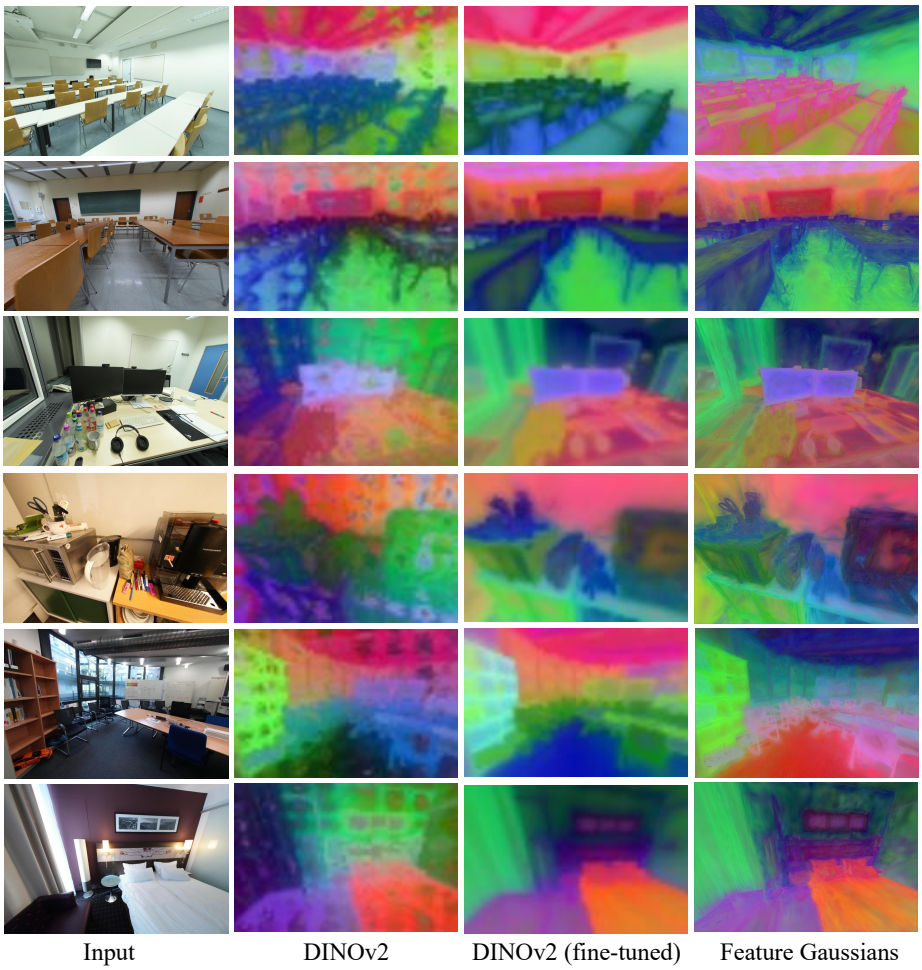

Fig. 1: Feature visualization on ScanNet++ [7]. We visualize the features rendered by pre-trained feature Gaussians (4^{th} column), features of DINOv2 (2^{nd} column) and our fine-tuned features (3^{rd} column). Our 3D-aware fine-tuning helps obtain features and capture detailed structures.

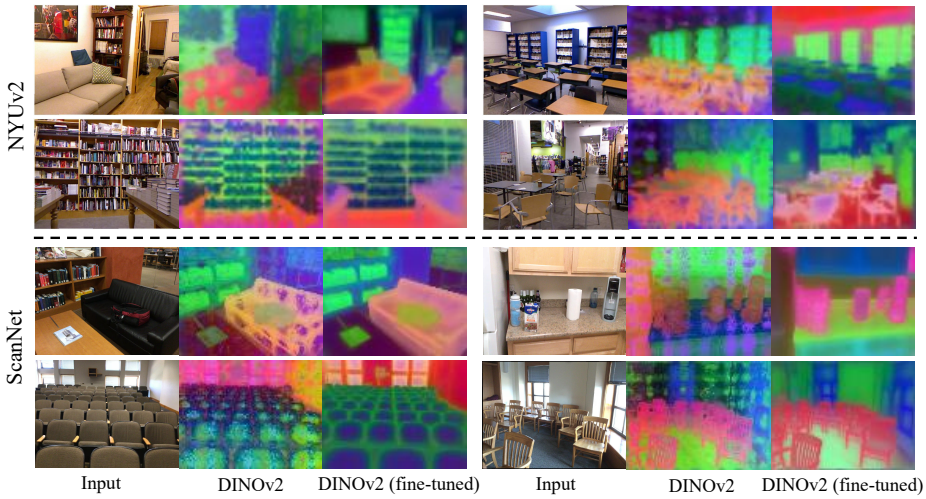

Fig. 2: Feature visualization on indoor datasets NYUv2 [6] and ScanNet [1]. Our 3D-aware fine-tuning helps obtain cleaner features and capture detailed structures.

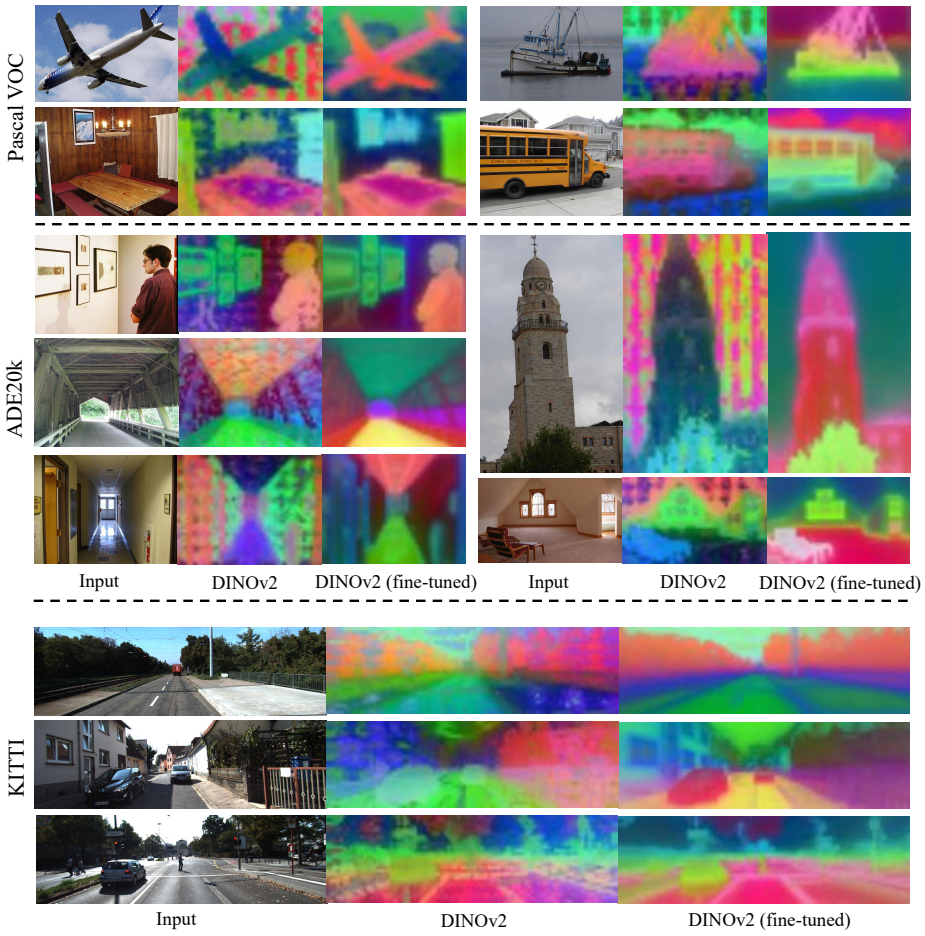

Fig. 3: Feature visualization on out-of-domain datasets. Our 3D-aware fine-tuning is generalizable to out-of-domain datasets and helps obtain cleaner features and capture detailed structures.

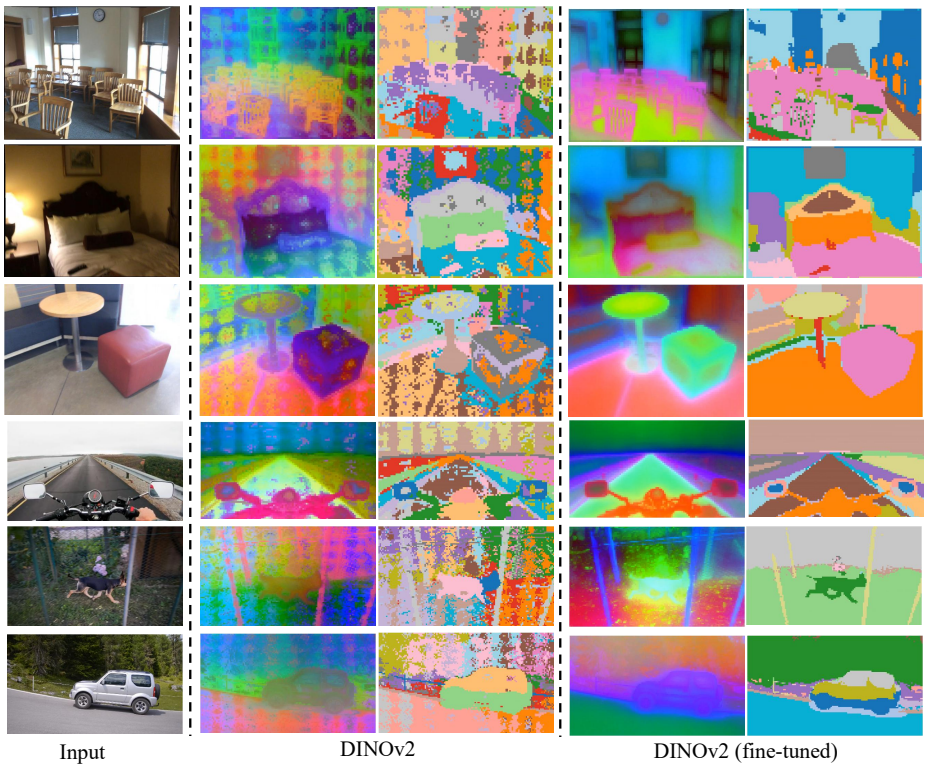

Fig. 4: K-Means clustering of features. We show the PCA features and K-Means clustering results of DINOv2 (2, 3$^{\text{th}}$ columns) and our 3D-aware fine-tuning features (4, 5$^{\text{th}}$ columns). Our K-Means clustering results are much more compact and cleaner than DINOv2.

References

1. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 5
2. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision Meets Robotics: The Kitti Dataset. The International Journal of Robotics Research (2013) 2
3. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning Robust Visual Features Without Supervision. Transactions on Machine Learning Research (2023) 1, 2
4. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision Transformers for Dense Prediction. In: International Conference on Computer Vision (ICCV) (2021) 2
5. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision (2015) 2
6. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor Segmentation and Support Inference From RGBD Images. In: European Conference on Computer Vision (ECCV) (2012) 1, 2, 5
7. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In: International Conference on Computer Vision (ICCV) (2023) 1, 2, 4
8. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene Parsing Through ade20K Dataset. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 1