EFFICIENT DENSE FEATURES WITH BRIXEL

Alexander Lappe[1,2] **Martin A. Giese**[1]
[1]Hertie Institute, University of Tübingen [2]IMPRS-IS
alexander.lappe@uni-tuebingen.de

ABSTRACT

Pretrained on large image datasets, recent dense-feature extractors can produce very fine-grained spatial feature maps, enabling state-of-the-art performance on spatial reasoning tasks. However, computing these feature maps requires the input image to be available at very high resolution, as well as large amounts of compute due to the squared complexity of the transformer architecture. To address these issues, we propose BRIXEL, a simple knowledge distillation approach that has the student learn to reproduce its own feature maps at higher resolution. Despite the simplistic approach, BRIXEL outperforms baseline models by large margins on downstream tasks when the resolution is kept fixed, allowing for more efficient spatial reasoning.

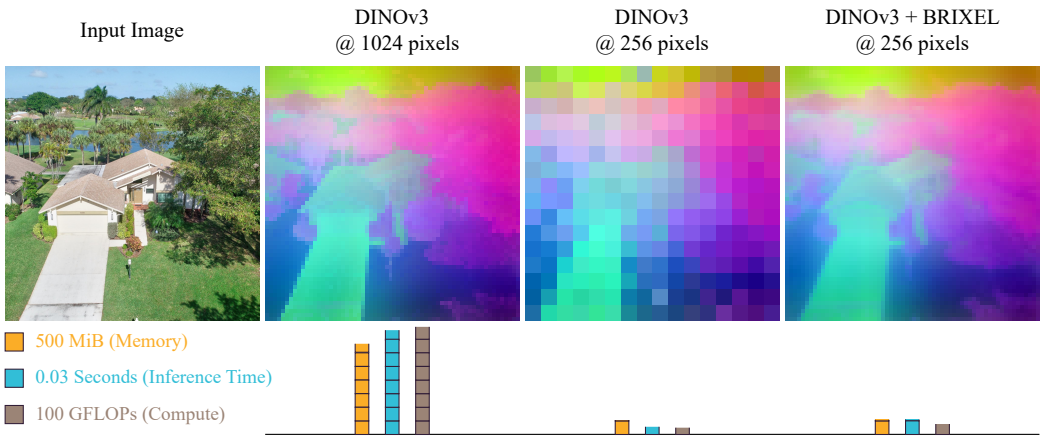

Figure 1: Recent dense feature extractors are able to operate at very high resolution, albeit at great computational cost. We propose BRIXEL, a simple self-distillation approach that produces dense feature maps while circumventing the Vision Transformer's quadratic scaling.

1 INTRODUCTION

In the late 1970s, Pink Floyds's *Another Brick in the Wall* raised the question of whether teachers actually make their students smarter. Here, we revisit the idea of a teacher that makes its students *more* dense in the context of self-distillation of dense image features. Great strides have been made in unsupervised pretraining of vision foundation models in recent years. The highly flexible Vision Transformer (ViT) Dosovitskiy et al. (2021) architecture has enabled a variety of very powerful, general-purpose feature extractors Radford et al. (2021); Zhai et al. (2023); Zhou et al. (2021); Touvron et al. (2022); Caron et al. (2021); Oquab et al. (2023); Siméoni et al. (2025); Bolya et al. (2025). One of the most intriguing properties of these ViT-based models is that they learn not only global image representations, but also dense descriptors for local image regions. While the quality of these dense features varies across models Siméoni et al. (2025); Banani et al. (2024), the state-of-the-art models for fine-grained spatial tasks such as depth estimation Yang et al. (2024) or segmentation Ravi et al. (2024); Siméoni et al. (2025) rely on this mechanism.

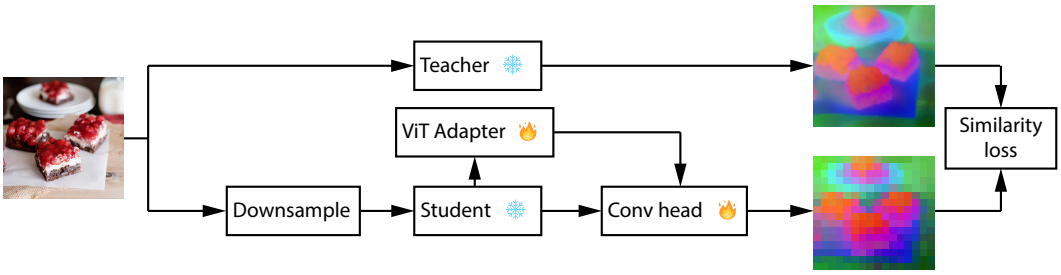

Figure 2: An overview of BRIXEL. The teacher and student network share both architecture and weights, which are all frozen. During training, the student receives a downsampled input image and has to reconstruct the dense features computed by the high-resolution teacher model. To do so, the student is connected to a standard ViT adapter and feeds into a convolutional readout head which fuses the output of the frozen student backbone and the trainable ViT adapter.

The main disadvantage of using Vision Transformers for dense tasks is that the spatial resolution of the features is inherently limited. Since these models divide the input into local patches and, unlike convolutional networks, operate at the same spatial resolution throughout the entire architecture, the final feature resolution is equal to the resolution of input image patches. The most commonly used patch size is 16, meaning that the feature maps will be downsampled by that factor relative to the input image. One way to circumvent this problem is to pretrain the model on very-high resolution images, which results in extremely high-resolved feature maps at test time Siméoni et al. (2025). However, this strategy comes with two caveats: First, the test-time or downstream task images need to be available in much higher resolution than the desired feature resolution, which in many applications will not be the case. Second, the quadratic scaling of the computational complexity of the transformer architecture with respect to the number of input tokens makes this approach very expensive.

This disadvantage is usually addressed by decoupling the heavy semantic and geometric lifting of the ViT from the fine-grained spatial computations. The high-performing dense ViTs mentioned in the previous paragraph rely on a pretrained backbone, and a supervised, task-specific spatial refiner Yang et al. (2024); Ravi et al. (2024); Siméoni et al. (2025) like the ViT adapter Chen et al. (2022) or MaskFormer Cheng et al. (2021; 2022). While leading to great performance, this approach requires substantial further supervised training after the unsupervised pretraining stage, meaning that it is not applicable when little supervised data is available. Therefore, it does not fully align with the desired off-the-shelf downstream task capabilities of these models. To remedy these issues, the goal of this work is to produce high-resolution, *task-agnostic* dense features without any supervision.

2 A TEACHER THAT MAKES THE STUDENT MORE DENSE

As discussed in the previous section, prior work combined pretrained Vision Transformers with a fine-grained refiner module for dense tasks. While training the refiner, the transformer weights may be fine-tuned Ravi et al. (2024) as well or kept frozen Siméoni et al. (2025). In any case, optimizing the weights of the refiner network requires an additional learning signal beyond the (usually unsupervised) pretraining strategy, which is supplied in the form of task-specific label supervision Ravi et al. (2024); Yang et al. (2024); Siméoni et al. (2025).

Since we aim to produce task-agnostic high-resolution features, producing a strong learning signal for the refiner is not as straightforward. However, the recently published DINOv3 model family offers an interesting avenue in that direction. These models are trained at a variety of input resolutions, and are regularized to produce highly consistent dense features across image sizes Siméoni et al. (2025). As prior work has shown that the features of foundation models, coupled with a refiner network, perform well on a variety of tasks, we wonder whether the reconstruction of higher-resolution foundation model features is among those tasks.

Main idea. We propose a simple method for generating high-resolution dense features by distilling fine-grained spatial information captured by a model at high resolution into a refiner network. The

goal of the refiner is to operate in conjunction with the model at low resolution to output the same dense feature map that the model would produce for higher-resolution input. We sketch the training procedure in Fig. 2. It consists of a simple teacher-student setup where the teacher and the student are identical networks with shared, frozen weights. The student is connected to a refiner network and both feed into a convolutional head that outputs the final dense features. During training, the teacher receives input images at high resolution, and the student receives the same image downsampled by a factor of 4 per side. The weights of the refiner network and the head are optimized for the output to mimic that of the teacher.

Aligning the feature maps. Let $\mathbf{x} \in \mathbb{R}^{\mathbf{3}\times\mathbf{h}\times\mathbf{w}}$, and $\mathbf{x}_- \in \mathbb{R}^{\mathbf{3}\times\frac{\mathbf{h}}{\mathbf{4}}\times\frac{\mathbf{w}}{\mathbf{4}}}$ denote an input image at high/low resolution respectively. Further, let $T(\cdot)$ denote the teacher and $S_{\boldsymbol{\theta}}(\cdot)$ the student network including the refiner and head, where $\boldsymbol{\theta}$ refers to the trainable parameters. First, we consider the L_1-loss between the outputs, i.e.

$$\mathcal{L}_1(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[||T(\mathbf{x}) - S_{\boldsymbol{\theta}}(\mathbf{x}_-)||_1]. \tag{1}$$

Empirically, we found that this loss function alone resulted in blurry boundaries. To ensure that the refiner produces faithful boundaries, we therefore also encourage it to match the output of Sobel edge detectors in feature space between the student's and teacher's feature maps. As edges at the single feature level are noisy, we compute an SVD on teacher tokens in each batch with gradients detached to find a projection P onto the K highest-variance principal components. Letting ∇_x and ∇_y denote the channel-wise Sobel operators, we define the edge loss

$$\mathcal{L}_{\text{edge}}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[||\nabla_x P(T(\mathbf{x})) - \nabla_x P(S_{\boldsymbol{\theta}}(\mathbf{x}_-))||_1 + ||\nabla_y P(T(\mathbf{x})) - \nabla_y P(S_{\boldsymbol{\theta}}(\mathbf{x}_-))||_1]. \tag{2}$$

Here, the projection P operates token-wise, i.e.

$$P : \mathbb{R}^{C\times\frac{h}{p}\times\frac{w}{p}} \to \mathbb{R}^{C\text{reduced}\times\frac{h}{p}\times\frac{w}{p}}, \tag{3}$$

and p denotes the patch size of the model. Finally, we also include a spectral loss to encourage similar high-frequency components between student and teacher output. To this end, we compute the FFT of both feature maps, convert them to polar coordinates and average amplitudes over concentric circles with fixed radius r to obtain one-dimensional frequency spectra $p_{T(\mathbf{x})}(r)$ and $p_{S(\mathbf{x}_-)}(r)$. We compare high-frequency spectra using the loss

$$\mathcal{L}_{\text{spectral}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}}\left(\log p_{T(\mathbf{x})}(r) - \log p_{S(\mathbf{x}_-)}(r)\right)^2\right], \tag{4}$$

where $\mathcal{R} := \{r | r \geq r_0\}$ contains the high-frequency components. Finally, the overall loss becomes

$$\mathcal{L}_{\text{total}}(\boldsymbol{\theta}) := \mathcal{L}_1(\boldsymbol{\theta}) + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}}(\boldsymbol{\theta}) + \lambda_{\text{spectral}}\mathcal{L}_{\text{spectral}}(\boldsymbol{\theta}). \tag{5}$$

3 Experiments

Models. We perform experiments on the pretrained DINOv3, SIGLIP 2 Tschannen et al. (2025) and Perception Encoder Bolya et al. (2025) models in their ViT-B variants. For the adapter network, we utilize the same architecture that was previously used to apply DINOv3 to supervised dense tasks Siméoni et al. (2025). It is largely based on the initial ViT adapter Chen et al. (2022), with the major difference that it does not feed back into the ViT backbone in which all weights are frozen. The convolutional head operates on the output of the ViT adapter and consists of three shallow residual blocks. All weights except for the frozen backbones are trained from scratch.

Data. As the training is self-supervised, the only necessary criterion for our training data is that its resolution needs to be sufficient for the teacher network to compute high-quality target features. For simplicity, we randomly sample high-resolution images from LAION and the Segment anything database to obtain a training set of 110k images.

Training. We then train the adapter networks at a resolution of 256^2 (so the teacher network operates at a resolution of 1024^2). Each model is trained using Adam on a single NVIDIA A100 for a total of 40k iterations. We set $\lambda_{\text{edge}} = 1$, $\lambda_{\text{spectral}} = 0.1$, $K = 8$ and the learning rate to $1 \cdot 10^{-3}$ with one warmup epoch.

Table 1: Performance of lightweight probes on top of frozen backbone models. We report mean Intersection over Union and Pixel Accuracy for semantic segmentation, Root Mean Square Error for monocular depth estimation and mean Angular Error for surface normals.

(a) ADE20k

	mIoU ↑		Pixel Accuracy ↑	
Model	Baseline	Ours	Baseline	Ours
SIGLIP 2	36.6	**40.8**	72.0	**76.0**
DINOv3	46.7	**49.2**	80.5	**82.0**
PE	34.0	**37.8**	73.3	**76.3**

(b) Cityscapes

	mIoU ↑		Pixel Accuracy ↑	
Model	Baseline	Ours	Baseline	Ours
SIGLIP 2	44.2	**50.0**	88.1	**90.9**
DINOv3	61.1	**64.4**	91.6	**93.0**
PE	45.7	**52.0**	90.0	**92.4**

(c) NYU

	RMSE ↓	
Model	Baseline	Ours
SIGLIP 2	0.613	**0.567**
DINOv3	0.354	**0.346**
PE	0.394	**0.381**

(d) NAVI

	RMSE ↓		Angular Error ↓	
Model	Baseline	Ours	Baseline	Ours
SIGLIP 2	0.506	**0.489**	49.4	**48.6**
DINOv3	0.388	**0.380**	41.3	**39.4**
PE	0.469	**0.440**	48.4	**44.9**

Evaluation. We probe whether the increased spatial resolution of the student's feature maps leads to performance gains on dense tasks over the baseline models. In particular, we test the models on semantic scene segmentation using the benchmarks ADE20k Zhou et al. (2017; 2018) and Cityscapes Cordts et al. (2016) at a resolution of 256, as well as monocular depth estimation on NYU Silberman et al. (2012) at 288×384, and monocular depth and surface normal estimation on NAVI Jampani et al. (2023) at 256 pixels. For segmentation, we train a linear probe on top of the frozen backbone model. For depth and surface normal estimation, we train a lightweight non-linear probe adapted from Banani et al. (2024), which again takes only the frozen last-layer features as input. As depth estimation is slightly more involved than classification, we follow the implementation details from Banani et al. (2024) and kindly refer the reader to their work for details. Importantly, the weights of the ViT adapter are frozen during all experiments beyond the self-distillation stage, as we explicitly aim to obtain more dense feature maps without relying on external supervision. As all feature maps have lower resolution than the input images and we are considering pixel-level tasks, we bilinearly interpolate the output of each probe to pixel-level resolution. We report mean Intersection over Union (mIoU) and Pixel Accuracy for semantic segmentation, root mean squared error for depth estimation and mean Angular Error for surface normal estimation in Table 1. Across all models, and tasks, we observe substantial performance increases over the baseline models. We also evaluate BRIXEL at a variety of input

4 Conclusion

We have shown that, using a simple self-distillation strategy, we can faithfully increase the resolution of recent dense feature exptractors. Our findings can be used to both improve performance when the input resolution is fixed, as well as to generate higher-resolution maps when high-quality input images are not available, resulting in more efficient spatial reasoning.

References

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D Awareness of Visual Foundation Models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21795–21806, June 2024. doi: 10.1109/CVPR52733.2024.02059.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception Encoder:

The best visual embeddings are not at the output of the network. In *NeurIPS*. arXiv, April 2025. doi: 10.48550/arXiv.2504.13181.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, October 2021. doi: 10.1109/ICCV48922.2021.00951.

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision Transformer Adapter for Dense Predictions. In *The Eleventh International Conference on Learning Representations*, September 2022.

Bowen Cheng, A. Schwing, and Alexander Kirillov. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In *Neural Information Processing Systems*, July 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289, June 2022. doi: 10.1109/CVPR52688.2022.00135.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, June 2016. doi: 10.1109/CVPR.2016.350.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021.

Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. NAVI: Category-agnostic image collections with high-quality 3D shape and pose annotations. In *NeurIPS*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, July 2023. ISSN 2835-8856.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, July 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos, October 2024.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (eds.), *Computer Vision – ECCV 2012*, volume 7576, pp. 746–760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33714-7 978-3-642-33715-4. doi: 10.1007/978-3-642-33715-4_54.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, August 2025.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, volume 13684, pp. 516–533. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-20052-6 978-3-031-20053-3. doi: 10.1007/978-3-031-20053-3_30.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, February 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. *Advances in Neural Information Processing Systems*, 37:21875–21911, December 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, October 2023. doi: 10.1109/ICCV51070.2023.01100.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130, July 2017. doi: 10.1109/CVPR.2017.544.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic Understanding of Scenes through the ADE20K Dataset, October 2018.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, A. Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer. *ArXiv*, November 2021.