# Self-supervised pre-raining of Vision Transformers for dense prediction tasks

Jaonary Rabarisoa, Valentin Belissen, Florian Chabot, Quoc-Cuong Pham

# Self-Supervised Pre-training of Vision Transformers for Dense Prediction Tasks

Jaonary Rabarisoa Valentin Belissen Florian Chabot Quoc-Cuong Pham
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{firstname.lastname}@cea.fr

## Abstract

*We present a new self-supervised pre-training of Vision Transformers for dense prediction tasks. It is based on a contrastive loss across views that compares pixel-level representations to global image representations. This strategy produces better local features suitable for dense prediction tasks as opposed to contrastive pre-training based on global image representation only. Furthermore, our approach does not suffer from a reduced batch size since the number of negative examples needed in the contrastive loss is in the order of the number of local features. We demonstrate the effectiveness of our pre-training strategy on two dense prediction tasks: semantic segmentation and monocular depth estimation.*

## 1. Introduction

Recently, Vision Transformers [11] (ViT) have become a powerful alternative to Convolutional Neural Networks (CNN) when solving computer vision tasks. On several benchmarks (image classification, object detection, semantic segmentation, monocular depth prediction, ...) most of the top performing methods use a ViT as a core component. This performance can be explained by a reduced inductive bias in ViT, compared to CNN, which leads to better generalization capability. But this comes at a cost: the need for a large annotated dataset [26]. For a given task, when the amount of training data is limited, pre-training combined with transfer learning is the most successful way to use ViTs.

Several works deal with pre-training ViT models. They can be supervised [26] by solving a classification task on a large dataset [22, 24] or self-supervised [5, 8, 28]. Generally, these approaches seek to learn a global representation at the image level during the pre-training phase, which is then used as initialization to different downstream tasks. Even though they exhibit a good performance when transferred to a pixel-level task such as semantic segmentation [12], we argue that their performance could be improved by taking into account the local aspect of the downstream task.

In this work, we propose to learn discriminative local features using a new self-supervised learning approach. Our pretext task compares representations of every part of a signal to its global representation using contrastive loss [6, 29]. With this strategy, we learn part representations that are informative about the global context allowing a better initialization for future dense downstream tasks. For instance, in semantic segmentation we predict the class of every pixel of an image. The classes to be predicted have high semantic level and define a global concept on the image. The representation of each pixel should then have a sufficient information about the class it belongs to. Several lines of work have demonstrated that contrasting local and global views is a good strategy to learn a representation. In [14] the authors use it to learn nodes and a graph representation. The multi-crop strategy presented in [4] compares several views generated at different scale levels. Another benefit in contrasting local and global views is that we naturally have several local views and then have access to more negatives samples. Our approach is then less sensitive to the batch size. Finally, as shown in [17], using several positives in the contrastive loss [6] helps to learn a better representation.

Our contribution is two-fold: 1) We propose a new pre-training strategy of Vision Transformers [11] designed for dense prediction tasks based on local to global contrastive learning. 2) We present a comprehensive study to demonstrate the effectiveness of our pre-training method on the tasks of semantic segmentation and monocular depth estimation.

## 2. Related Work

**Self-supervised learning** Self-supervised learning is a set of techniques used to pre-train a neural network with unlabelled data before solving a specific downstream task. The most successful approaches use contrastive representation learning [29]. They predict the next representation of a sequence from the current context by solving a classification problem where the target representation is the positive class and any other vectors from the dataset are considered as negative. A standard loss function is the Noise-Contrastive Estimation loss (InfoNCE) [13] which is shown to be a lower bound of the mutual information between the current context

and the next features in the sequence. Based on this loss, SimCLR [6, 7] learns visual representations by contrasting different views of the same image with other images, by using carefully designed random augmentations and a large training batch size. The method outperforms supervised pre-training in image classification tasks. MoCo [15] alleviates the need of a large batch size by sampling the negative examples from a memory bank and uses a momentum network to compute the target feature vectors. AMDIM [1] is a multi-scale extension of contrastive representation learning from multiple views. It predicts local representations of a CNN feature map from global description vectors at different levels of the network. Having a large number of good quality negative examples is a key factor in the convergence of InfoNCE loss optimization. Non-contrastive approaches propose to solve the multi-view representation learning problem without negative samples. BYOL [12] directly matches the outputs of a Siamese network using mean squared error. SwAV [4] learns the representation by predicting a pseudo-label from one view using the other view. Pseudo-labels are computed by clustering representations with Sinkhorn's algorithm. Barlow Twins [33] minimizes an objective function based on the cross-correlation matrix of the two views.

**Pre-training Vision Transformers** The performance of ViT in computer vision tasks highly depends on the pre-training phase, especially when the training data is small. Steiner *et al.* in [26] study the inter-play between regularization, data augmentation and dataset size when training ViTs for image classification. DeiT [28] is a data-efficient training method using distillation through attention. It shows competitive results when training ViT models on ImageNet-1k with no external data. Other works study the self-supervised pre-training of ViT. DINO [5] belongs to the multi-view representation learning methods. It uses self-distillation and pseudo-labels to learn a visual representation. The authors highlight the importance of local-to-global correspondence through the use of the multi-crop strategy [4]. In BeIT [2], inspired by BERT [9], image patches are encoded into discrete visual tokens and the pre-training objective is to recover these visual tokens from corrupted patch embeddings.

**Transformers for Dense Prediction** DETR [3] is one of the first successful applications of Transformers for dense prediction tasks. It uses a transformer encoder-decoder network to directly predict object location and class without any extra components required in standard object detection pipelines. Segmenter [27] uses ViT as its encoder network and a transformer decoder network similar to the one used in DETR in order to predict semantic segmentation masks. Although we follow a similar architecture design as Segmenter, we mainly focus on the pre-training of the encoder.

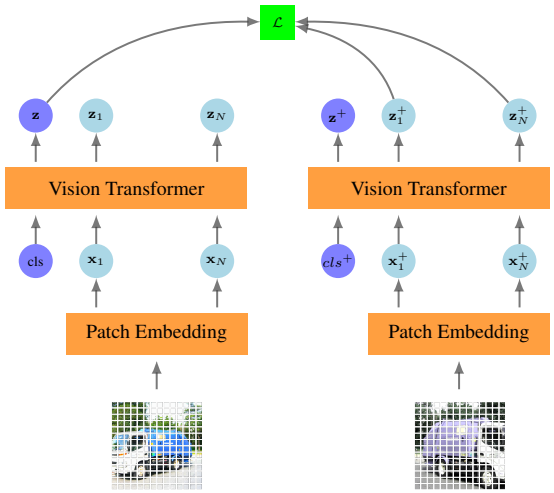


Figure 1. Two augmented views of the same image are generated using random augmentation. Patches of size $16 \times 16$ are extracted from each view and encoded into visual tokens. A learnable class token is used for the global representation. To each token we add its positional encoding and the results are encoded with a ViT. Our loss function contrasts the global representation $\mathbf{z}$ with all patches representation $\mathbf{z}_i^+$.

Applying Transformers to dense vision tasks raises computational complexity issues especially for high-resolution images, as self-attention has quadratic complexity with respect to the number of input tokens. More efficient architectures have been introduced (Swin Tranformer [19], Focal Transformer [32], CSWin Transformer [10], DPT [21]) and are now state-of-the-art for dense prediction tasks. In all of these works, the pre-training phase and the use of large datasets remain critical to reach optimal performance.

In our pre-training approach, we generalize the InfoNCE loss with several positive examples [17] to learn to predict local features from a global image representation. As opposed to AMDIM, our method only works with the external features outputted by the ViT network, which makes it more computationally efficient. Following the same intuition that global-to-local feature prediction [4] allows to learn stronger representations, our approach is not data-driven but only relies on the image patches imposed by the ViT architecture.

## 3. Approach

Our goal is to learn a pixel-level representation suitable for dense vision prediction tasks in a self-supervised way. In the following, we denote $\mathbf{x} = \{\mathbf{x}_i\}_{i=1\dots N}$ an image with $N$ the number of pixels. Each pixel $\mathbf{x}_i$ belongs to the RGB space. Let $\mathbf{x}^+ = \mathcal{T}(\mathbf{x})$ be an another view of $\mathbf{x}$ obtained by randomly transforming $\mathbf{x}$ with $\mathcal{T}$. The latter is sampled from a set of admissible transformations of $\mathbf{x}$.

### 3.1. Vanilla Constrastive Learning

The vanilla contrastive learning learns a global feature transform $\Phi$ that maps the image $\mathbf{x}$ into a feature vector $\mathbf{z} = \Phi(\mathbf{x}) \in \mathbb{R}^d$ by minimizing the InfoNCE objective:

$$\mathcal{L}_{nce}(\mathbf{z}, \mathbf{z}^+) = -\log \frac{e^{\mathbf{z}^t \mathbf{z}^+ / \tau}}{e^{\mathbf{z}^t \mathbf{z}^+ / \tau} + \sum_{\mathbf{z}^- \in \mathcal{N}(\mathbf{z})} e^{\mathbf{z}^t \mathbf{z}^- / \tau}} \quad (1)$$

where $(\mathbf{z}, \mathbf{z}^+)$ are the global feature vectors of the two views of $\mathbf{x}$ and $\mathcal{N}(\mathbf{z})$ is the set of negative features. That is the set of the features of any other images except $\mathbf{x}$ and its different views. $\tau$ is the temperature. Minimizing (1) aims to group the pair of positive samples $(\mathbf{z}, \mathbf{z}^+)$ together while pushing all negative features away from $\mathcal{N}(\mathbf{z})$.

In practice, $\Phi$ is a deep neural network and the objective function is minimized using stochastic gradient descent. The negative features for each example are sampled from the mini-batch. Having a large number of negative samples is critical while minimizing the InfoNCE loss. Hence, a large batch size is required during the optimization.

### 3.2. Dense Contrastive Learning

We propose to learn a pixel-level feature-transform $\Psi$ that yields features, $\mathbf{z}_i = \Psi(\mathbf{x}_i)$ for each pixel $\mathbf{x}_i$. To learn $\Psi$ we use the InfoNCE loss (1) that compares local and global features in the following way:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{nce}(\mathbf{z}, \mathbf{z}_i^+) \quad (2)$$

In equation (2), $\mathbf{z}$ represents the global representation of the image $\mathbf{x}$ and $\mathbf{z}_i^+$ a local feature from its corresponding view $\mathbf{x}^+$. Minimizing (2) aims to contrast all the patches representation from one view with the global representation of another view. This pretext task is closely related to the dense prediction task of semantic segmentation where we predict the class of each pixel in a given image. We argue that this helps to learn more meaningful local features since it will contain global information.

A similar approach has been proposed to learn nodes and a graph level representation in [14]. While they learn a representation for general graphs we apply it to ViT viewed as a fully connected graph of patches. Also, the multi-crop strategy proposed by [4] brings evidence that contrasting local and global views helps to learn robust features. In our case, we do not generate explicitly the local and global views by data augmentation but rather use the structure imposed by the Vision Transformer.

In practice, the optimization is done by stochastic gradient descent and the negative features are sampled from the mini-batch. With this formulation, the number of negative samples used in the InfoNCE is multiplied by the number of patches. This makes our approach less dependent on the batch size.

## 4. Experiments

In this section we present our experimental setup and main results. We follow the standard evaluation protocols for self-supervised learning. After the pre-training phase, the learned features are evaluated on two downstream tasks: semantic segmentation and monocular depth estimation. At the evaluation time, we add a task related head on top of the ViTs encoder. The whole network is fine-tuned on the target dataset.

### 4.1. Pre-Training

We study three different Vision Transformer configurations [11, 28]: ViT-Ti, ViT-S, ViT-B. We set the ViT patch size to $16 \times 16$. The projection network is a 3-layer MLP [6] with a *gelu* activation. The temperature of the InfoNCE loss is fixed at $\tau = 0.1$. The models are pre-trained with an input resolution of $224 \times 224$.

We use AdamW [20] as optimizer, a cosine decay learning rate scheduler and a linear warm-up for $5\%$ of the total epochs. The base learning rate is $10^{-4}$ and is linearly scaled with respect to the batch size. In our current implementation, we do not gather the negative examples across all accelerator devices and only use per device negative to compute the InfoNCE loss (1). The batch size per GPU is set to 128. The pre-training is distributed across 16 Nvidia A100-80G.

By default, we use the same data augmentation policy as SimCLR [6]. First, the image is randomly cropped and resized to $224 \times 224$. Then, the image color is randomly distorted and optionally converted to grayscale. Finally, Gaussian blur is randomly applied.

### 4.2. Semantic segmentation

Semantic segmentation aims to predict the class of each pixel in a given image. We conduct our experiments on the ADE20k dataset [34]. After the pre-training, we append a segmentation head to the pre-trained encoder. The whole network is fine-tuned on the training set. We report results on the validation set. We consider two types of segmentation heads: linear and the UPerNet [2, 30]. The input resolution is set to $512 \times 512$. We use bicubic interpolation to interpolate the positional embedding of the ViT.

The segmentation model is learned using the pixel-wise cross-entropy objective. We use AdamW as optimizer and a base learning rate of $0.0001$ with a polynomial decay policy. During the $64$ epochs of the fine-tuning, we set the weights decay to $0.005$, the stochastic depth to $0.1$ and we apply a layer-wise decay to the learning rate.

Table 1 presents our results for different pre-training methods on the ImageNet-1k and several combinations of Encoder/Head networks. For the Supervised and DINO pre-training we use the encoder weights provided by [26] and [5] and run the fine-tuning task ourselves. It shows that our

pre-training strategy gives the best performance especially for ViT-Ti and ViT-S. We find out that the regularization is critical for the fine-tuning phase when the size of the network increases. For the ViT-B architecture we use a higher value of weights decay and stochastic depth compared to ViT-Ti and ViT-S. Comparatively, a ViT-B pre-trained on ImageNet-21k with BeIT [2] performs better than ours. This suggests that our ViT-B still suffers from overfitting and a the use of a large dataset can help the fine-tuning.

| Encoder | Pre-training | Head | mIoU |
|---|---|---|---|
| ViT-Ti/16 | Supervised* | Linear | 32.4 |
| | Ours | Linear | 36.2 |
| | | UPerNet | **38.0** |
| ViT-S/16 | Supervised* | Linear | 42.1 |
| | DINO* | Linear | 38.8 |
| | Ours | Linear | **42.8** |
| | | UPerNet | **43.2** |
| ViT-B/16 | Supervised* | Linear | 43.1 |
| | DINO* | Linear | 43.2 |
| | Ours | Linear | **45.1** |
| | | UPerNet | **45.1** |
| | *BeIT (i-21k)* | *UPerNet* | ***53.6*** |

Table 1. Performance of the ViT models pre-trained on ImageNet-1k and fine-tuned on ADE20K. We report the mIoU on the validation set. (*) Models are fine-tuned from the weights provided by the authors. (i-21k) indicates pre-training done on ImageNet-21k and result from the paper [2].

### 4.3. Monocular depth estimation

We evaluate the generalization capability of the pre-trained representation by fine-tuning the network on a depth estimation task on the NYU-Depth V2 dataset [25].

Two kinds of regression heads are experimented: on the one hand, a linear series of 4 up-projection convolutions with a stride of 2 until the image resolution is retrieved. On the other hand, we experiment a more sophisticated head inspired from the work of [30] (UPerNet). In both cases, the final activation function is a dilated sigmoid that matches the depth range of the NYU-Depth V2 dataset. The training loss is a berHu loss like in [18] regularized with a depth smoothness term ($\mathbb{L}_2$ gradient loss, inspired from [16]).

The pre-trained models are fine-tuned on the 47584 image/depth map pairs of the train split, and evaluated on the 654 pairs of the test split. Data augmentation includes random color jitter, crop and horizontal flip. We evaluate our models using standard metrics – absolute relative error (AbsRel), RMSE, threshold accuracy ($\delta_1$). All models are trained for 50 epochs, with the Adam optimizer and a learning rate of $0.0001$.

| Method | | AbsRel | RMSE | $\delta_1$ |
|---|---|---|---|---|
| FCRN-Depth [18] | | 0.127 | 0.573 | 0.811 |
| SimCLR [6] | | 0.134 | 0.557 | 0.833 |
| BYOL [12] | | 0.129 | 0.541 | 0.846 |
| ViT-Ti/16 | Linear | 0.140 | 0.598 | 0.823 |
| | UPerNet | 0.138 | 0.593 | 0.832 |
| ViT-S/16 | Linear | 0.124 | 0.564 | 0.856 |
| | UPerNet | **0.122** | 0.549 | 0.862 |
| ViT-B/16 | Linear | 0.123 | 0.544 | 0.862 |
| | UPerNet | **0.122** | **0.526** | **0.865** |

Table 2. Performance of the pre-trained ViT models when fine-tuned on a downstream depth estimation task (NYU-Depth V2 dataset). For the tiny (ViT-Ti), small (ViT-S) and big (ViT-B) versions, both a linear and UPerNet-style heads are evaluated, with better performance than state-of-the-art reference points.

Our results are presented in Table 2. BYOL [12] and SimCLR [6] are the two main self-supervised reference points. These two models have indeed been pre-trained on the training set of the ImageNet-1k dataset [23] – without using labels – and fine-tuned on the NYU-Depth V2 dataset. As for [12] and for a broader perspective, we also include results from FCRN-Depth [18], which uses supervised pre-training on ImageNet-1k before fine-tuning on NYU-Depth V2. However, since we aim at assessing the performance of the fully self-supervised pre-training and for a fair comparison, we do not report the results of more recent methods that mix ViTs and CNN representations with weights trained in a supervised fashion on large image datasets, like [31]. For the same reason, we also exclude results from models pre-trained on much larger depth datasets, like [21].

Our method outperforms the self-supervised state-of-the-art with respect to all metrics, an observation that holds true for almost all metrics even when using the smaller ViT-S/16 or with the linear up-convolution heads.

## 5. Conclusion

We have presented a new method to pre-train Vison Transformers for dense vision tasks. It is an extension of the contrastive learning framework that compares local to global features. We have shown that this pre-training strategy is competitive on semantic segmentation and monocular depth estimation tasks. For future work we plan to study the effect of more advanced data augmentation on this approach and the use of non contrastive losses as objective function.

## 6. Acknowledgments

This work benefited from the FactoryIA supercomputer financially supported by the Ile-deFrance Regional Council.

## References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views, 2019. 2

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers, 2021. 2, 3, 4

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing. 2

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020. 1, 2, 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 1, 2, 3

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 1, 2, 3, 4

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020. 2

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. 1

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 2

[10] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows, 2021. 2

[11] 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, 2021. 1, 3

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 1, 2, 4

[13] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. 1

[14] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4116–4126. PMLR, 13–18 Jul 2020. 1, 3

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[16] Philipp Heise, Sebastian Klose, Brian Jensen, and Alois Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2360–2367, 2013. 4

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 1, 2

[18] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 4

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 2

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 3

[21] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 2, 4

[22] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021. 1

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 4

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 1

[25] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 4

[26] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers, 2021. 1, 2, 3

[27] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation, 2021. 2

[28] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2021. 1, 2, 3

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 1

[30] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 3, 4

[31] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16269–16279, 2021. 4

[32] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers, 2021. 2

[33] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. 2

[34] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic Understanding of Scenes Through the ADE20K Dataset. *International Journal of Computer Vision*, 127(3):302–321, March 2019. 3