PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers

Xiaoyi Dong[1*†], Jianmin Bao[2*], Ting Zhang[2], Dongdong Chen[3,†], Weiming Zhang[1]
Lu Yuan[3], Dong Chen[2], Fang Wen[2], Nenghai Yu[1], Baining Guo[2]
[1]University of Science and Technology of China
[2]Microsoft Research Asia
[3]Microsoft Cloud + AI
{dlight@mail., zhangwm@, ynh@}.ustc.edu.cn cddlyf@gmail.com
{jianbao, ting.zhang, luyuan, doch, fangwen, bainguo }@microsoft.com

Abstract

This paper explores a better prediction target for BERT pre-training of vision transformers. We observe that current prediction targets disagree with human perception judgment. This contradiction motivates us to learn a perceptual prediction target. We argue that perceptually similar images should stay close to each other in the prediction target space. We surprisingly find one simple yet effective idea: enforcing perceptual similarity during the dVAE training. Moreover, we adopt a self-supervised transformer model for deep feature extraction and show that it works well for calculating perceptual similarity. We demonstrate that such learned visual tokens indeed exhibit better semantic meanings, and help pre-training achieve superior transfer performance in various downstream tasks. For example, we achieve **84.5%** Top-1 accuracy on ImageNet-1K with ViT-B backbone, outperforming the competitive method BEiT by **+1.3%** under the same pre-training epochs. Our approach also gets significant improvement on object detection and segmentation on COCO and semantic segmentation on ADE20K. Equipped with a larger backbone ViT-H, we achieve the state-of-the-art ImageNet accuracy (**88.3%**) among methods using only ImageNet-1K data.

Introduction

Current state-of-the-art self-supervised pre-training methods (Dosovitskiy et al. 2020; Bao, Dong, and Wei 2021; He et al. 2021; Xie et al. 2021; Chen et al. 2022; Wei et al. 2021) for vision transformers focus on masked image modeling (MIM), a task of making predictions for masked patches from the visible patches. The input is usually an image consisting of visible patches and randomly masked patches and each patch is associated with corresponding positional embedding. The prediction target for masked patches varies for different methods, ranging from pixel-level prediction (Dosovitskiy et al. 2020; He et al. 2021; Xie et al. 2021) to feature-level prediction (Bao, Dong, and Wei 2021; Chen et al. 2022; Wei et al. 2021). In this paper, we study the prediction targets and introduce a better prediction target for MIM.

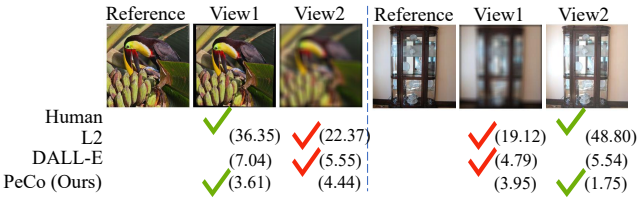

Figure 1: Several examples illustrating the results of different prediction targets on the question that which image (View1 or View2) is "closer" to the Reference image. The number denotes the distance between View1 or View2 and the Reference image. The images with smaller distances are considered more similar. We observe that the proposed PeCo agrees with human judgments while L2 or DALL-E disagree.

We point out that current prediction targets disagree with human judgment when evaluating the similarity between two different images. There are two representative prediction targets in current MIM methods: per-pixel regression and discrete token prediction. Figure 1 illustrates the results of different prediction targets on the question that which image (View1 or View2) is "closer" to the "Reference" for these examples. The reason for such disagreement of current prediction targets may come from the per-pixel loss. Note that the discrete tokens are obtained by a VQ-VAE trained under the objective of reconstruction loss, *i.e.* , per-pixel loss. The per-pixel measure assuming pixel-wise independence is insufficient for assessing structured outputs. For example, blurring causes large perceptual change but small pixel error, while shifting incurs small perceptual change but large pixel error (Zhang et al. 2018). Such disagreement with human visual perception indicates that perceptually similar patches may have divergent prediction targets. This undermines the capability of MIM as it, in principle, is based on context prediction.

We propose that a good prediction target for MIM should coincide with human judgment. In other words, perceptually similar images should be close to each other in the prediction target space. Inspired from the observation in (Zhang et al. 2018) that deep features model low-level perceptual similarity surprisingly well, we introduce this so-called percep-

*Equal contribution, † Corresponding Author
†Work done during an internship at Microsoft Research Asia

tual loss in VQ-VAE for discrete token learning. This loss can be viewed as per-feature loss as it aims to minimize the feature-wise distance between the original image and the reconstructed image. Specifically, we adopt multi-scale deep features from multiple layers at different depth of a self-supervised Transformer. As shown in Figure 1, our proposed new prediction target indeed aligns with human perception judgment. We also show that the proposed visual tokens get much higher linear accuracy than the one without the perceptual loss. It indicates that our new visual tokens exhibit more semantic meanings, which is analogous to texts whose discrete tokens often contain highly semantic information.

We denote MIM using the introduced perceptual visual tokens for targets as "PeCo", *i.e.* Perceptual Codebook for BERT pre-training of vision transformers. In the experiments, we demonstrate that equipped with such perceptual visual tokens, PeCo achieves better performance compared with the strong competitor BEiT (Bao, Dong, and Wei 2021) using DALL-E (Ramesh et al. 2021) codebook trained over $250M$ images without the perceptual loss. We fine-tune the pre-trained model on various downstream tasks: image classification, object detection, and semantic segmentation. Experimental results show that our pre-trained model transfers better than BEiT with only the prediction target changed. Concretely, we achieve **84.5**% Top-1 accuracy on ImageNet-1K with ViT-B model, outperforming BEiT by **+1.3%** with the same 800 pre-training epochs. Our approach also gets significant improvement on COCO object detection and semantic segmentation as well as on ADE20K semantic segmentation. Our PeCo also shows strong scalability that when equipped with a larger backbone ViT-H, we achieve the state-of-the-art ImageNet accuracy (**88.3%**) among methods using only ImageNet-1K data.

Related Works

Self-Supervised Learning Self-supervised learning has attracted increasing attention over the past few years, as deep learning networks become more and more data-hungry and it's impossible to label everything in the world. There are two main categories along this path, contrastive and generative (Liu et al. 2021a). One emerging field is self-supervised contrastive learning, training an encoder to the representation measured by contrastive loss (Hadsell, Chopra, and LeCun 2006; Dosovitskiy et al. 2014) via comparing similar and dissimilar samples. The representative methods include MOCO (He et al. 2020; Chen et al. 2020d), SimCLR (Chen et al. 2020b,c), BYOL (Grill et al. 2020), SwAV (Caron et al. 2020). However, contrastive-based methods heavily depend on the strong data augmentation and effective negative sampling.

The other recent resurgent field is generative self-supervised learning, training an encoder and a decoder under the objective of reconstruction loss. The typical objectives, autoregressive and denoising autoencoder, aiming at recovering the corrupted or masked input, has yielded the most successful frameworks (Devlin et al. 2018) in NLP. Thanks to the pre-existing vocabulary in language, recovering the missing word can be transformed into predicting all the possible words with the probability estimation, converting the prediction problem to an easier classification problem. While in CV, on the other hand, most attempts (Van Oord, Kalchbrenner, and Kavukcuoglu 2016; Oord et al. 2016; Chen et al. 2020a) still resort to regression for generative methods due to the lack of visual vocabulary, *e.g.* iGPT (Chen et al. 2020a). Recently, BEiT (Bao, Dong, and Wei 2021) successfully adopts a classifier for prediction by directly adopting a VQ-VAE as the visual tokenizer. Yet there exists a major difference between the language vocabulary and the visual vocabulary. That is, the words of language are highly semantic, while the visual words of images are mostly not. Most recently, numerous works (Bao, Dong, and Wei 2021; He et al. 2021; Xie et al. 2021; Wang et al. 2022b; Dong et al. 2022) based on MIM have been concurrently developed, yet few studied the perceptual level of the prediction targets. In this work, we attempt to learn a perceptual visual vocabulary for BERT pre-training, showing superior transfer performance than BEiT (Bao, Dong, and Wei 2021) and MAE (He et al. 2021).

Discrete Visual Supervision Exploring masked image modeling or image inpainting task for self-supervised pre-trained tasks has never been stopped in vision community, especially when BERT (Devlin et al. 2018) achieves great success in various tasks of NLP. To apply the cross-entropy loss function for vision tasks, iGPT (Chen et al. 2020a) clusters the pixel values to simulate the process of BPE (Sennrich, Haddow, and Birch 2015) process for different words in language. ViT (Dosovitskiy et al. 2020) attempts to directly divide the raw pixel values into multiple groups and assign a discrete label for each group GRB value. Recent work VQ-VAE (Oord, Vinyals, and Kavukcuoglu 2017) proposes to adopt encoder and decoder to quantize the visual contents to a learnable codebook with fixed size.

Perceptual Similarity The term "perceptual similarity" refers to imitating human perception when assessing image similarity. It has been shown in (Zhang et al. 2018) that the internal activations of network trained for classification task surprisingly coincide with human judgment. Such deep features have been widely used in image generation with the goal of synthesizing realistic images. The loss is called perceptual loss or VGG loss as the network used is often VGG architecture. In this paper, we surprisingly discover that this simple loss is super effective in building a better prediction target and significantly improves vision BERT pretraining. Moreover, to enable self-supervised learning, we adopt a self-supervised trained network rather than ImageNet-trained networks and show it also works comparably well. Both these two discoveries are conceptually simple yet super-effective and valuable.

Method

In the natural language processing field, the words are naturally discrete tokens which contain high semantic information. By contrast, vision signals are continuous with redundant low-level information. While there are various ways to discretize the image in prior works, the semantic level of the resulting visual tokens has been largely ignored. In this section, we start by briefly describing the discrete representa-

tion learning from VQ-VAE, and then introduce the process of how to learn a perceptual codebook, followed by BERT pre-training over the learned perceptual visual tokens.

Learning Discrete Codebook for Visual Content

We utilize VQ-VAE (Oord, Vinyals, and Kavukcuoglu 2017) to convert the continuous image content into the form of discrete tokens. Consider an image $x \in \mathbb{R}^{H\times W\times 3}$, VQ-VAE is able to represent it with discrete visual codewords $\{z_q^1, z_q^2, \cdots, z_q^N\} \in \mathcal{V}^1 \times \mathcal{V}^2 \times \cdots \times \mathcal{V}^N$, where z_q^i comes from a visual codebook (vocabulary) $\mathcal{V}^i = \{e_k^i \in \mathbb{R}^D\}_{k=1}^{K_i}$ consisting of K_i D-dimensional codewords. Usually we have $K_1 = K_2 = \cdots = K_N = K$ for simplicity, and $N = h \times w$ with $h \times w$ being the spatial resolution of the latent space.

Specifically, to learn such latent codeooks, VQ-VAE contains three major parts: an encoder, a quantizer and a decoder. The encoder maps the input image to intermediate latent vectors $z = \text{Enc}(x)$, where $z \in \mathbb{R}^{h\times w\times D}$.The quantizer is in charge of quantizing each vector at position (i, j) to be codewords coming from the corresponding codebook $\mathcal{V}^{i,j} = \{e_k^{i,j}\}_{k=1}^K \subset \mathbb{R}^D$ according to nearest neighbor assignment. That is,

$$k^* = q(z^{i,j}) = \underset{k\in\{1,2,\cdots,K\}}{\arg\min} \|z^{i,j} - e_k^{i,j}\|. \tag{1}$$

$$z_q^{i,j} = r(k^*) = e_{k^*}^{i,j}, \tag{2}$$

where q is the quantization encoder that maps the vector to an index of the codebook, and r is the quantization decoder that reconstructs the vector from the index. Based on the quantized codewords $z_{\mathbf{q}}$, the decoder aims to reconstruct the input image x. Suppose the reconstruct result is $\hat{x} = \text{Dec}(z_{\mathbf{q}})$. Since the quantizer is non-differentiable, to back-propagate gradient into encoder, the gradient is approximated like the straight-through estimator (Bengio, Léonard, and Courville 2013) and just copied from decoder to encoder (Oord, Vinyals, and Kavukcuoglu 2017). The training objective of VQ-VAE is defined as,

$$\begin{aligned}\mathcal{L}_{\text{VQ-VAE}}(\text{Enc}, \text{Dec}, \{\mathcal{V}\}) = \mathcal{L}_{pixel} + \|\text{sg}[\text{Enc}(x)] - z_{\mathbf{q}}\|_2^2 \\ + \beta\|\text{sg}[z_{\mathbf{q}}] - \text{Enc}(x)\|_2^2.\end{aligned} \tag{3}$$

Here, $\mathcal{L}_{pixel} = \frac{1}{H\times W\times 3}\|x - \hat{x}\|$ is the per-pixel loss, $\text{sg}[\cdot]$ is the stop-gradient operator, β is a loss weight set to 0.25 in all our experiments.

Learning Perceptual Codebook for Visual Content

In the vanilla VQ-VAE, the codebook is learned by an element-wise pixel loss, *i.e.* $\mathcal{L}_{pixel}$, between the original image and the reconstructed image. However, this per-pixel loss may prevent the network from capturing *perceptual* difference since the loss only accounts for the correctness of individual pixels. Therefore, a small shift and rotation operation on the original image may not cause perceptual change but large ℓ_1/ℓ_2 error.

Therefore, we propose a simple yet effective strategy by enforcing perceptual similarity between the original image and the reconstructed one beyond the pixel loss. The perceptual similarity is not based on pixel differences but instead feature differences where the high-level image features extracted from a pre-trained deep neural network. We hope this feature-wise loss will better capture perceptual difference and offer invariance towards low-level variations. We show the comparison of using different losses in Figure 3 from the perspective of image reconstruction, suggesting that images with lower pixel-wise loss may not appear perceptually similar.

Previous works usually adopt a supervised pretrained VGG (Simonyan and Zisserman 2014) network to calculate perceptual loss, since using supervision is not consistent with our purpose of self-supervised pre-training. We turn to the self-supervised models and replace the ConvNet-based model with Vision Transformer, which have a better modeling capability and efficiency. On the other hand, pre-trained models usually encode different levels of semantic information in different layers, to enable our codebook to have rich perceptual information, we adopt multi-scale features from multiple layers of the model to calculate the perceptual loss. Our experiments show that a vision Transformer (ViT-B model) from self-supervised learning works well for calculating perceptual loss.

Formally, let $f_l(x)$ be the normalized activations of the l-th layer of a network F when processing the image x. The size of the feature map is $H_l \times W_l \times C_l$ with H_l being the height, W_l being the width and C_l being the channel dimension. Usually, multi-scale features, more comprehensive and discriminative, from multiple layers at different depth are extracted to calculate the perceptual similarity for better semantic capture. The perceptual metric for the input image x and the reconstructed image $\hat{x}$ can be formulated as,

$$\mathcal{L}_{\text{percep}} = \sum_{l\in\mathcal{S}} \frac{1}{C_l H_l W_l}\|f_l(x) - f_l(\hat{x})\|_2^2, \tag{4}$$

where $\mathcal{S}$ denotes the number of layers from which the features are extracted.

Therefore, the overall objective function is,

$$\begin{aligned}\mathcal{L}_{\text{VQ-VAE}_{percep}} = \mathcal{L}_{pixel} + \lambda\mathcal{L}_{percep} \\ + \|\text{sg}[\text{Enc}(x)] - z_{\mathbf{q}}\|_2^2 \\ + \beta\|\text{sg}[z_{\mathbf{q}}] - \text{Enc}(x)\|_2^2,\end{aligned} \tag{5}$$

where λ is the hyper-parameter for the loss weight of $\mathcal{L}_{percep}$, we will study different vaules of loss weight λ in the experiments. The training pipeline of perceptual codebook is illustrated in Figure 2 (a). After training, the encoder and the quantizer are used as tokenizer in the subsequent pre-training process.

BERT Objective over Perceptual Codebook

We adopt the BERT objective to perform the *masked image modeling* task over the discrete visual tokens as in BEiT (Bao, Dong, and Wei 2021), illustrated in Figure 2. For a given image x, the input tokens are image patches which are non-overlappingly split from the whole

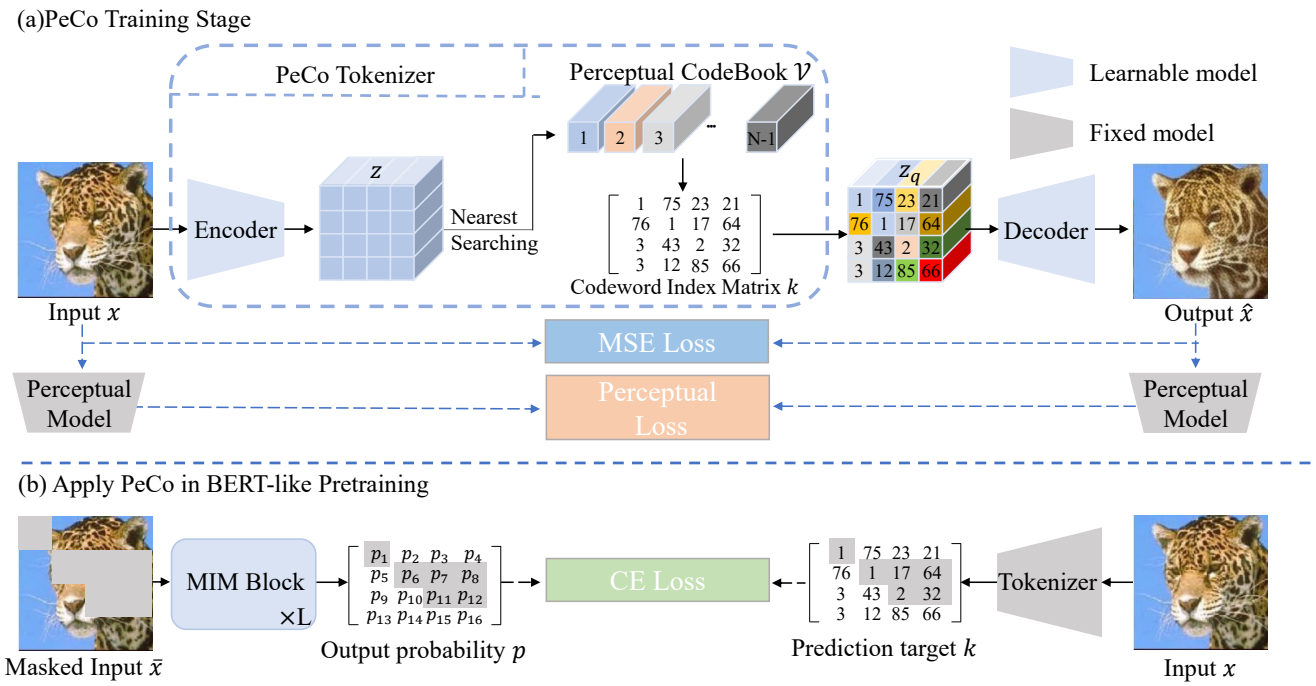

Figure 2: (a) Training pipeline of our Perceptual Coodbook. (b) Apply PeCo in BERT-Like pretraining. Our PeCo provides a more semantic prediction target to the Mask Image Modeling Task.

Figure 3: Image reconstruction with different losses. An example contains three images showing input (left), reconstructed image using pixel-wise loss (middle), and reconstructed image using pixel-wise and feature-wise losses (right). We can see that perceptually the right image appears more similar to the input compared with the middle image, although the middle image gets lower pixel-wise loss.

image, and the output tokens are discrete perceptual visual words obtained through learning Eqn 5. Let the input be $\{x^1, x^2, \cdots, x^N\}$, and the groundtruth output be $\{k^1, k^2, \cdots, k^N\} = q(Enc(x))$. The goal of the masked image modeling is to recover the corresponding visual tokens from the masked input where a portion of input tokens have been masked.

Precisely, let $\mathcal{M}$ be the set of masked index. Then the masked input $\bar{x}$ is represented as,

$$\bar{x}^i = \begin{cases} x^i, & i \notin \mathcal{M} \\ m, & i \in \mathcal{M} \end{cases}, i = 1, 2, \cdots, N, \tag{6}$$

where m is a learnable mask token as same dimension as non-mask tokens. The masked input tokens are fed into a L-layer vision Transformer with the last layer's hidden output being denoted as $\{h^1, h^2, \cdots, h^N\}$. We aim at recovering the corresponding visual token from the hidden vector at masked positions. To achieve that with the classification loss, a K-way classifier is appended after the hidden vector h^i to get the probability estimation about all possible discrete tokens in the corresponding codebook $\mathcal{V}^i$. Suppose the groundtruth discrete visual tokens corresponding to the masked patches are k^t with $t \in \mathcal{M}$, the pre-training objective can be formulated as,

$$\mathcal{L}_{\text{pre-training}} = -\sum_{t \in \mathcal{M}} \log P(k^t | \bar{x}), \tag{7}$$

where $P(k^t|\bar{x})$ is the estimated target token probability for masked patches of corrupted image $\bar{x}$.

After pre-training the model, we apply the model to various downstream tasks including ImageNet-1K (Deng et al. 2009) classification, COCO object detection (Lin et al. 2014), and ADE20K (Zhou et al. 2017) Segmentation.

Pre-training Details

Vector Quantizer We use the standard k-means algorithm for vector quantization. We set the codebook size K as 8192 for fair comparison. When the size of the discrete latent space K is large, we observe that only a few codewords are selected to represent image and get trained. Many other codewords are wasted. To overcome this issue, we adopt exponential moving averages (Oord, Vinyals, and Kavukcuoglu 2017) to update the codebook which is proved

Methods	pre-train dataset	pre-train epochs	ViT-B	ViT-L	ViT-H	ViT-H_{448}
Training from scratch (i.e., random initialization)						
ViT_{384}	-	-	77.9	76.5	–	–
DeiT	-	-	81.8	–	–	–
ViT	-	-	82.3	82.6	83.1	–
Self-Supervised Pre-Training on ImageNet-1K						
DINO	IN-1K	300	82.8	–	–	–
MoCo v3	IN-1K	300	83.2	84.1	–	–
BEiT	IN-1K	800	83.2	85.2	–	–
BootMAE	IN-1K	800	84.2	85.9	–	–
MAE	IN-1K	**1600**	83.6	85.9	86.9	87.8
PeCo	IN-1K	800	84.5	86.5	87.5	**88.3**

Table 1: Image classification accuracy (%) comparison on ImageNet-1K (IN-1K) of different self-supervised methods using various backbones.

to be useful for increasing utilization of codewords in a codebook.

Perceptual Codebook Learning Setup We train the perceptual codebook using the training set of ImageNet-1K dataset by default. For the encoder and decoder of VQ-VAE, we choose traditional convolutional based backbone. The network contains two residual blocks at each resolution. A self-attention block is applied to the smallest resolution for both encoder and decoder. For perceptual loss, we use the pre-trained 100 epochs ViT-B model from self-supervised method MoCo v3 (Chen, Xie, and He 2021) by default.

BERT Pre-training Setup For computation resource consideration, we use the original ViT-B/16 (Dosovitskiy et al. 2020) as the basic architecture of our backbone to validate the effectiveness of the learned visual codebook, as in BEiT (Bao, Dong, and Wei 2021). The model is pre-trained for 300/800 epochs with the batchsize of 2048. We use a block-wise masking strategy for obtaining the corrupted images with the same setup as BEiT (Bao, Dong, and Wei 2021). We further demonstrate the effectiveness of our approach when scaling to ViT-Large and ViT-Huge backbones.

Experiments

Downstream Tasks

Image Classification aims to classify a given image into its corresponding class category. We use the popular ImageNet-1K dataset. To enable classification, a global average pooling layer is appended after the pre-trained model. We finetune the model with 100 epochs and a cosine decay learning rate that warmups to $4e^{-3}$ with 20 epochs and decays to 0. Following (Bao, Dong, and Wei 2021), the layer-wise learning rate decay is also used and set to 0.65 by default. For more details, please refer to the supplementary materials.

Semantic Segmentation is the task of assigning a label to each pixel of the input image. We compare on the semantic segmentation dataset ADE20K benchmark (Zhou et al. 2017). Here we employ the Upernet (Xiao et al. 2018) as the basic framework. For fair comparison, we follow previous works (Bao, Dong, and Wei 2021) and train Upernet 160k iterations with batch size set as 16, more details are provided in the supplementary material.

Methods	tokenizer dataset	tokenizer #params	BERT pre-train epoch	IN-1K Top-1
BEiT	DALLE(400M)	53.8M	300/800	82.8/83.2
PeCo	IN-1K(1.3M)	37.5M	300/800	84.1/84.5
$PeCo_{lite}$	IN-1K(1.3M)	25.7M	300/800	84.0/84.5

Table 2: Tokenizer comparison with BEiT. Here we report tokenizer training dataset and #parameters. $PeCo_{lite}$ is a lite version of PeCo that reduces the channel number of tokenizer by half.

Methods	pre-train dataset	pre-train epochs	ADE-20K mIoU	COCO AP^{bb}	COCO AP^{mk}
DEiT	IN-1K	300	47.4	44.1	39.8
MoCo	IN-1K	300	47.3	44.9	40.4
BEiT	DALLE+IN-1K	800	47.1	46.3	41.1
MAE	IN-1K	800	47.6	46.8	41.9
MAE	IN-1K	1600	48.1	47.2	42.0
PeCo	IN-1K	800	**48.5**	**47.8**	**42.6**

Table 3: Semantic segmentation mIoU (%) comparison on ADE20K and object detection and instance segmentation comparison in terms of box AP (AP^{bb}) and mask AP (AP^{mk}) on COCO. The backbones for all the methods are the ViT-B.

Object Detection and Segmentation Object detection is to locate objects in a given image and identify each object. We perform fine-tuning on the COCO objection detection and segmentation with the Mask R-CNN (He et al. 2017) framework. Specifically, we add four different scale FPNs to scale the feature map into different size following (Bao, Dong, and Wei 2021). The fine-tuning is conducted with "1x" (12 training epochs) schedule and single-scale input on the COCO training set and test the performance on COCO validation set, following the strategy used in Swin Transformer (Liu et al. 2021b).

Comparison with Previous Works

We first compare our PeCo with previous state-of-the-art works. Here we report ImageNet-1K results with various model sizes. For object detection on CoCo and semantic segmentation on ADE20K, we use ViT-B as the backbone.

Image Classification The Top-1 accuracy on ImageNet-1K classification is reported in Table 1. We compare our method with 1) ViT (Dosovitskiy et al. 2020) and DeiT (Touvron et al. 2021) that are supervisedly trained from scratch with random initialization; and 2) MoCo v3 (Chen, Xie, and He 2021) and DINO (Caron et al. 2021), represent the contrastive learning for self-supervised pre-training; and 3) BEiT (Bao, Dong, and Wei 2021), MAE (He et al. 2021) and

BootMAE (Dong et al. 2022) based on masked image modeling for self-supervised pre-training. It can be seen that our model (PeCo) significantly improves the performance compared with the models trained from scratch, suggesting the effectiveness of pre-training.

Compared with prior self-supervised pre-training models, our model achieves the best performance. For example, our model using ViT-B backbone pre-trained with 800 epochs reaches 84.5% Top-1 accuracy, 1.3% higher than BEiT and 0.9% higher than MAE. Furthermore, we also compare the results on larger backbones, *e.g.* ViT-L and ViT-H. The results are reported in the Table1, showing significantly better performance than previous counterparts. This validates that our perceptual codebook is indeed beneficial for pre-training. Concretely, our model PeCo-H_{448} achieves the best Top-1 accuracy, **88.3**%, on ImageNet-1K without external data, outperforming MAE by 0.5%. This is a new state-of-the-art result using only ImageNet-1K data.

We also report the results pre-trained with 300 epochs in Table 2. Compared with the baseline BEiT (Bao, Dong, and Wei 2021), our model gets $+1.3\%$ improvement for both 300 and 800 pre-training epochs. We further investigate a lite version of tokenizer which reduces the channel number of the original by half. This decreases the extra timecost introduced by the tokenizer by about $2\times$. We can see from Table 2 that with a lite tokenizer, our model still gets competitive performance.

Semantic Segmentation We compare our method with 1) DEiT, which is a supervised pre-training method on ImageNet-1K , 2) MoCo, the contrastive learning based methods, and 3) BEiT (Bao, Dong, and Wei 2021), MAE (He et al. 2021), the state-of-the-art self-supervised learning model. Here we use UperNet (Xiao et al. 2018) framework with 512×512 input and trained for 160K iterations. The evaluation metric is mean Intersection of Union (mIoU) averaged over all semantic categories and we report single-scale results here. The results are given in Table 3. Our method achieve 48.5 mIoU, +1.1 mIoU than supervised based methods. It is also + 1.2 mIoU than MoCo, +1.4 mIoU than BEiT, and +0.9 mIoU than MAE. Our model even achieve better results(+0.4 mIoU) than MAE pre-training with 1600 epochs. This verifies the effectiveness of the perceptual codebook.

Object Detection and Segmentation We further investigate our transfer performance on object detection and segmentation. Here we use Mask-RCNN (He et al. 2017) framework with single-scale input and $1\times$ schedule (12 epochs). We compare with the strong competitor BEiT (Bao, Dong, and Wei 2021) on this dataset. The evaluation metric is box AP for detection and mask AP for segmentation. The comparison is presented in Table 3. Our model with ViT-B as backbone achieve 47.8 box AP and 42.6 mask AP, +3.7 box AP and +2.8 mask AP over supervised methods. Our model also outperform recent work MAE by +1.0 box AP, + 0.7 box AP under the same pre-training epochs. Our model is also higher than MAE pre-training with 1600 epochs.

Methods	LinearProb. on codewords	Classification. on recon.
DALL-E	6.1	18.2
PeCo(w/o $\mathcal{L}_{percep}$)	10.2	17.9
PeCo(ours)	29.7	51.7

Table 4: Evaluation of the semantics of the codewords from linear probling accuracy (%) of codewords on ImageNet-1K and classification accuracy (%) on the reconstructed ImageNet validation images using Deit-T.

Loss for Tokenizer Training	acc. on IN-1K
L_{pixel}	82.9
L_{pixel} + L_{percep} from SSL ResNet-50	84.0
L_{pixel} + L_{percep} from SSL ViT-B	84.1
L_{pixel} + L_{percep} from Supervised VGG	84.1

Table 5: The performance comparison when using different architectures for calculating the perceptual similarity.

Analysis of Perceptual Codebook

In this section, we ablate our perceptual codebook by using the setting of self-supervised pre-training on ImageNet-1K. The pre-raining epochs is 800.

Semantics of the Codewords The most important question would be: *will the learned perceptual codewords exhibit (more) semantic meanings?* To answer this, we quantitatively evaluate the codewords' semantics from two aspects. (1) We use the codewords of the image as features for classification. An average pooling is conducted over the quantized codewords of the image and we test its linear probing accuracy over ImageNet dataset. (2) We use an ImageNet-1K supervisedly pre-trained DeiT-T (Touvron et al. 2021) (72.2% Top1 accuracy on clean ImageNet val set) to test the classification accuracy over the reconstructed images. We compare with the variant without using the perceptual similarity. The results are given in Table 4. We find that our perceptual codewords get much higher accuracy for both linear evaluation on codewords and classification on the reconstructed images. This indicates that our perceptual codebook exhibits more semantic meanings and benefits the image reconstruction process. We also provide a visualization of the masked region prediction using BEiT (Bao, Dong, and Wei 2021) and our PeCo in Figure 4, showing that our PeCo, with the aid of perceptual codebook, is able to make more semantic predictions for the masked region.

Deep Architectures for Perceptual Similarity Another key question would be: *will the deep architectures for deep perceptual features affect the perceptual codebook learning and thus affect the pre-training performance?* Therefore, we investigate two different deep architectures: convolutional-based backbone ResNet50 (He et al. 2016) and Transformer-based model ViT-B (Dosovitskiy et al. 2020). We study the self-supervised models in order to enable unsupervised pre-training. The results are reported in Table 5. We can see

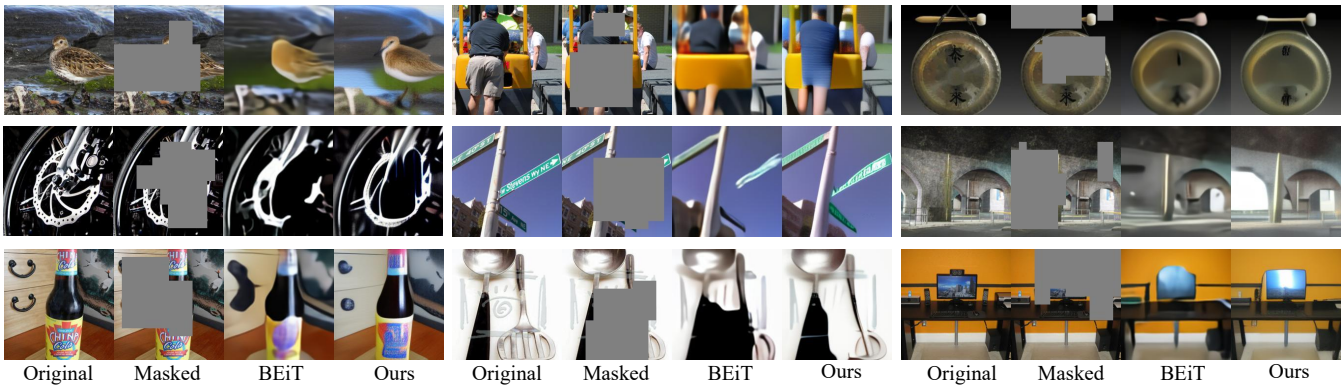

Figure 4: Examples of reconstruction results on ImageNet-1K using BEiT and our PeCo.

Perceptual mechanism	Top-1 acc. on IN-1K
Classification loss on codewords	82.9
Contrastive loss on codewords	82.9
Perceptual loss on images	84.1

Table 6: Performance comparison of different way to inject semantic to the codebook.

Loss functions	Top-1 acc. on IN-1K
$\mathcal{L}_{pixel}$	82.9
$\mathcal{L}_{pixel} + \mathcal{L}_{percep}$	84.1
$\mathcal{L}_{pixel} + \mathcal{L}_{percep} + \mathcal{L}_{adv}$	83.9

Table 7: Performance comparison of different loss functions.

that using convolution-based or Transformer-based network achieves similar performance. In addition, we also report the results using the classical supervised (*i.e.* using label) trained VGG (Simonyan and Zisserman 2014) in Table 5. It can be seen that using supervised model for perceptual metric achieve comparable performance as self-supervised model.

Discussions

We present several in-depth discussions about the proposed model in this section.

Implicit *vs.* Explicit. The key contribution of our paper is improving the perceptual level of the discrete visual tokens for the subsequent pre-training. We have successfully demonstrated that through a simple strategy, *i.e.* enforcing perceptual similarity over images. One may think that it seems quite implicit for learning perceptual codebook by constraining on images instead of directly exploiting some constraint over the codebook. Indeed, we also experiment in two explicit ways: 1) supervised classification loss over the codewords; 2) constraining a momentum contrastive loss over the quantized codewords through data augmentation in a self-supervised way. We hope that leveraging those forms of high-level classification objective may encode some semantics into the codewords. But empirically we found that such explicit ways are not as effective as the proposed implicit strategy. The results are reported in Table 6. We conjecture that the codebook may learn global semantics from the classification/contrastive loss and thus fail to differentiate different codewords, which is not suitable for pre-training. In contrast, deep features from a pre-trained deep model contain rich and dense semantics.

Perceptual Loss *vs.* GAN Loss. The perceptual loss is widely used in generation tasks with the goal of improving the image quality. We ask the question that *is there a positive relation with the image quality and the perceptual level of the codebook.* In order to explore this, we adopt another technique, adversarial loss in Generative Adversarial Nets(GANs) (Goodfellow et al. 2014), which has been proved to be effective in enhancing the reconstructed image. Specifically, we add a patch-based discriminator D (Li and Wand 2016), aiming to make the original image and the reconstructed one indistinguishable. The adversarial loss is,

$$\min_{Enc,\{\mathcal{V}\},Dec} \max_{D} \mathcal{L}_{\text{adv}} = \log D(x) + \log(1 - D(\hat{x})). \quad (8)$$

We add this loss with a suitable weight 0.4 to Eqn 5 and use the learned codebook for pre-training. The resulting performance is shown in Table 7. We can see that adversarial loss can not bring gain to the transfer performance of pre-training.

Conclusion

In this paper, we argue that a good prediction target for masked image modeling should agree with human perception judgment. Motivated by this observation, we propose a simple yet effective strategy to obtain perceptually discrete tokens, beneficial for BERT pre-training of vision transformers. We present extensive comparisons on various downstream tasks. Our results indeed validate our hypothesis and show superior performance compared with previous state-of-the-art methods. We hope that the deep analysis about the prediction target in our work will lead to a broader exploration of this perspective and even help existing multi-modality foundation model pretraining (Yuan et al. 2021; Wang et al. 2022a).

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant U20B2047, 62072421, 62002334, 62102386 and 62121002.

References

Bao, H.; Dong, L.; and Wei, F. 2021. BEiT: BERT Pre-Training of Image Transformers. *arXiv preprint arXiv:2106.08254*.

Bengio, Y.; Léonard, N.; and Courville, A. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020a. Generative pretraining from pixels. In *ICML*, 1691–1703. PMLR.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. 2020c. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

Chen, X.; Ding, M.; Wang, X.; Xin, Y.; Mo, S.; Wang, Y.; Han, S.; Luo, P.; Zeng, G.; and Wang, J. 2022. Context Autoencoder for Self-Supervised Representation Learning. *arXiv preprint arXiv:2202.03026*.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020d. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, X.; Bao, J.; Zhang, T.; Chen, D.; Zhang, W.; Yuan, L.; Chen, D.; Wen, F.; and Yu, N. 2022. Bootstrapped Masked Autoencoders for Vision BERT Pretraining. In *ECCV*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dosovitskiy, A.; Springenberg, J. T.; Riedmiller, M.; and Brox, T. 2014. Discriminative unsupervised feature learning with convolutional neural networks. *NeurIPs*, 27: 766–774.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *NeurIPs*, 27.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, 1735–1742. IEEE.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked Autoencoders Are Scalable Vision Learners. *arXiv preprint arXiv:2111.06377*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Li, C.; and Wand, M. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, 702–716. Springer.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; and Tang, J. 2021a. Self-supervised learning: Generative or contrastive. *TKDE*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*.

Oord, A. v. d.; Kalchbrenner, N.; Vinyals, O.; Espeholt, L.; Graves, A.; and Kavukcuoglu, K. 2016. Conditional image generation with pixelcnn decoders. *arXiv preprint arXiv:1606.05328*.

Oord, A. v. d.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*, 10347–10357. PMLR.

Van Oord, A.; Kalchbrenner, N.; and Kavukcuoglu, K. 2016. Pixel recurrent neural networks. In *ICML*, 1747–1756. PMLR.

Wang, J.; Chen, D.; Wu, Z.; Luo, C.; Zhou, L.; Zhao, Y.; Xie, Y.; Liu, C.; Jiang, Y.-G.; and Yuan, L. 2022a. OmniVL: One Foundation Model for Image-Language and Video-Language Tasks. In *NeurIPS*.

Wang, R.; Chen, D.; Wu, Z.; Chen, Y.; Dai, X.; Liu, M.; Jiang, Y.-G.; Zhou, L.; and Yuan, L. 2022b. BEVT: BERT Pretraining of Video Transformers. In *CVPR*.

Wei, C.; Fan, H.; Xie, S.; Wu, C.-Y.; Yuille, A.; and Feichtenhofer, C. 2021. Masked Feature Prediction for Self-Supervised Visual Pre-Training. *arXiv preprint arXiv:2112.09133*.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *ECCV*, 418–434.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2021. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *CVPR*, 633–641.