arXiv:2303.04664v2 [cs.CV] 1 Aug 2024

面向高效视觉Transformer预训练的质心中心化建模

Xin Yan、Zuchao Li* 和 Lefei Zhang

中国武汉市武汉大学
zcli.charlie@gmail.com

摘要。 掩码图像建模（MIM）是一种基于视觉Transformer（ViT）的新型自监督视觉预训练范式。以往的方法可分为基于像素和基于词元两类，分别使用原始像素或参数化分词器模型生成的离散视觉词元。我们提出的基于质心的方法 **CCViT** 利用 k 均值算法进行聚类，以获取用于图像建模的质心，无需对分词器模型进行监督训练，且仅需数秒即可生成。这个非参数化质心分词器只需数秒即可创建，并且能够更快地进行词元推理。质心既能表示图像块像素，也能表示索引令牌，同时具备局部不变性。具体而言，我们采用图像块掩码和质心替换策略来构造损坏输入，并通过两个堆叠的编码器块预测损坏的图像块词元并重建原始图像块像素。实验表明，我们的**CCViT** 在 ImageNet-1K 分类任务上，搭配 ViT-B 时达到 84.4% 的 Top-1准确率，搭配 ViT-L 时达到 86.0%。我们还将预训练模型迁移至其他下游任务。无需外部监督或来自其他模型的蒸馏训练，我们的方法便取得了与近期基线相当的结果。[1]

关键词： 大规模视觉预训练· 自监督学习 · 掩码图像建模。

1 引言

过去几年中，视觉Transformer（ViT）[15] 已被证实在图像建模及其他任务中表现出色。然而，实证研究表明，与以往基于 CNN 的模型相比，视觉Transformer需要更多的数据。

这一问题已通过自监督学习成功解决。尤其是在自然语言处理领域，BERT[13] 提出了掩码语言建模（MLM）来解决这一问题。这种先掩码再预测的任务会遮盖输入数据中一定比例的

* 本研究得到了国家自然科学基金委员会（编号：62306216）和湖北省自然科学基金（编号：2023AFB816）的资助。[1] 代码可在 https://github.com/Cakeyan/CCViT 获取 Public

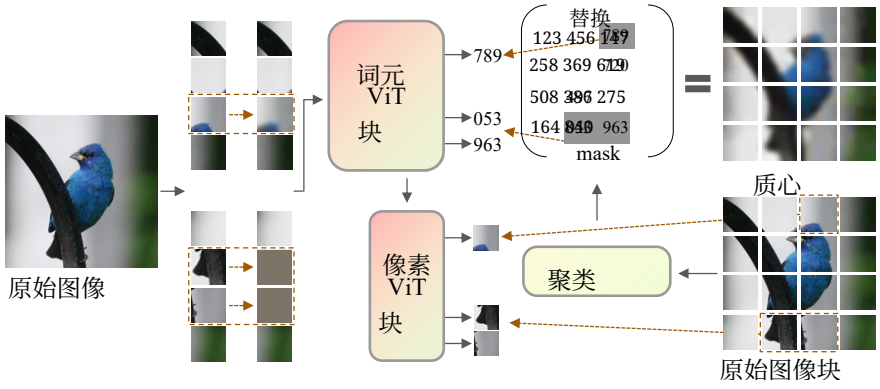


图 1. 所提出的 CCViT 架构。 我们将质心视为两个方面：词元 ID 和图像块像素。我们的以质心为中心的预训练旨在同时预测质心索引并重建图像块像素。我们对部分图像块应用分块掩码（*e.g.*，40%），并用相应的质心替换部分剩余图像块（*e.g.*，10%）。所有受损图像块都会输入 ViT 模块。

内容，然后学习预测被掩码的目标，展现了其利用大规模无标注数据的优势。受掩码语言建模启发，多项研究尝试将这一任务引入计算机视觉领域。BEiT[2] 将离散视觉词元视为目标，提出了掩码图像建模（MIM）：使用分词器将图像转换为词元，而该分词器通过基于离散变分自编码器（dVAE）[32]方法、配备解码器的额外训练阶段获得。在预训练期间，BEiT以受损图像为输入，并学习恢复被掩码的词元。与此相反，MAE[21] 通过额外的解码器重建原始图像像素。

与自然语言处理中主要以语言词元为目标不同，由于视觉与语言之间的信息密度不同，计算机视觉领域出现了各种重建目标，包括视觉词元[26,14],高层特征[11], 原始像素[21] 以及原始图像特征[35], 。我们可以将这些目标大致分为两类模型：基于词元的掩码图像建模，以词元或高层特征作为预训练目标，通常需要额外训练一个参数化分词器模型，将图像从连续表示抽象为离散表示；以及基于像素的掩码图像建模，它鼓励模型重建原始像素或方向梯度直方图等原始特征，而无需使用分词器。

然而，这两种方法都存在缺陷。首先，它们都会引入冗余的参数化模块，将潜在表示转换为原始像素或视觉词元，而在微调时又会丢弃这些模块。这些模块需要额外的训练阶段，或增加预训练的计算开销。更重要的是，它们可能需要额外的数据集或强监督。此外，参数化分词器生成的每个视觉词元都无法表示对应的图像块，而实际图像与先掩码后预测

框架相矛盾，因为该框架依赖词元之间的相关性进行掩码推理。这也解释了为什么在基于标记的掩码图像建模中，预训练阶段通常难以实现较高的词元预测准确率。我们将在实验部分和表4中提供全面的证明。

在这项工作中，我们提出了**CCViT**，即**C**entroid-**c**entered**V**ision **T**ransformer（质心中心化视觉 Transformer），如图 1 所示。具体而言，对于基于质心的分词器，我们在普通图像块像素上采用高效的聚类方法（例如 K-means 算法），以确定每个图像块的令牌编号，即最近质心的索引。与需要在预训练前耗费大量训练资源的参数化分词器不同，只需使用训练集中的极小部分数据，花费几秒钟即可轻松获得质心。此外，与其他参数化分词器相比，获取视觉词元的速度更快。更重要的是，我们仅在 ImageNet-1K[12] 的训练集上进行聚类以获得质心，不会引入 DALL-E[2,32],等大型或私有数据集，也不会将 CLIP[30,31] 等大型模型蒸馏到分词器中。预训练期间，我们将图像拆分为若干图像块，并掩码掉其中一部分。随后，我们将所有受损图像块输入骨干 ViT 编码器，以重建这些图像块的质心索引。

我们还提出了一种全新的掩码图像建模视角，即基于质心的掩码图像建模，它简化了流程，无需额外增加训练成本。在基于质心的掩码图像建模中，我们聚焦于聚类质心，它们既可视为质心图块像素，也可视为质心索引令牌。为充分利用这一双重属性，我们采用随机替换掩码策略，将部分图像块替换为最近质心的像素，以促使模型将质心的像素表征与其对应的词元 ID 对齐。我们还堆叠了两个具有不同学习目标的视觉 Transformer 模块，包括令牌ViT模块和像素 ViT 块。进一步的实验与分析表明，这两种策略能有效利用这一双重属性并提升性能。

我们在 ImageNet-1K 上进行自监督预训练，并在下游任务（即图像分类和语义分割）上微调预训练模型。实验结果表明，我们的方法能够获得出色的图像表征。具体而言，使用 ViT-B，我们在 ImageNet-1K 分类上达到 84.3%，在 ADE20K 上达到 48.4 的平均交并比，采用相同的 300 个预训练轮数时，性能超过 BEiT $+1.4\%$ 和 $+3.7$。为展示模型的扩展能力，我们采用更大的骨干网络和更长的预训练轮数，并取得显著提升。消融研究表明，同时学习质心的两种形式——像素和词元，优于单独学习任一形式。此外，我们分析了基于质心的掩码图像建模为何更适用于先掩码后预测框架，并具有最高的词元预测准确率。本文的贡献总结如下：

– 我们提出了一种基于 k 均值算法聚类的新型非参数视觉分词器。该分词器创建速度快，且仅需使用一小部分训练集，无需外部数据。我们还提出了一种新颖的随机替换策略和一种堆叠模型架构。

– 我们根据预训练目标，将现有的所有掩码图像建模方法分为基于词元的掩码图像建模和基于像素的掩码图像建模，并提出了基于质心的掩码图像建模，通过将质心同时视为图像块像素和索引令牌，提供了双重视角。与基于词元的掩码图像建模相比，该方法所需的分词器训练成本更低；与基于像素的掩码图像建模相比，该方法省去了编码器，因此计算效率更高。
– 我们在下游任务和迁移任务上进行了大量实验，证明了我们在无外部监督和蒸馏的情况下达到了最先进的性能。此外，我们还分析了效率，包括浮点运算次数、模型参数等。
– 我们全面研究了不同的视觉分词器，发现与我们具有双重性质的质心相比，以往的视觉词元无法表示相应的图像块。实验表明，我们的质心分词器表现稳健，也可扩展到包括像素重建在内的其他任务。

2 相关工作

自监督视觉学习多年来一直受到研究，旨在将这种学习范式引入视觉预训练。各种方法采用不同的前置任务进行预训练，包括拼图任务[27],着色[24] 和旋转角度预测[19]。对比学习也是视觉表征学习[7,9,20,22,36]中的一种趋势。这些方法通常依赖数据增强方法。此外，早期研究采用聚类方法来学习视觉表征[3,4,25,39]。最近，iGPT[6] 通过使用 k 均值算法对（R、G、B）像素取值进行聚类，创建了一个 9 位调色板（ $K = 512$ ），并将聚类令牌序列作为自回归和 BERT 目标的直接输入。相比之下，我们的方法以图像块作为输入，并以质心索引作为预训练目标。

掩码图像建模已广泛应用于视觉预训练领域，作为自然语言处理中掩码语言建模方法的对应方法，例如BERT[13]。自视觉Transformer[15] 首次突破架构障碍以来，掩码图像建模（MIM）迅速取得了显著成功[5,8,16]。MIM 在预训练阶段随机掩盖图像的一部分，并对其进行重建。由于预训练阶段的重建目标有所不同，而自然语言处理主要以词元为目标，因此 MIM 方法主要分为两种范式，即也就是基于词元的掩码图像建模和基于像素的掩码图像建模。基于词元的掩码图像建模源自已有方法 BEiT[2], ，需要参数化分词器来生成词元或高层特征，作为预训练目标。具体而言，BEiT 和 VIMPAC[34] 使用来自 DALL-E[32,33]的离线离散变分自编码器分词器。PeCo[14] 在 VQGAN[17] 训练中将 MoCov 3[10] 用作感知模型。mc-BEiT[26] 也着眼于感知相似性，并构建码本用于预训练。iBOT[41], 、SdAE[11]和 data2vec[1] 采用自蒸馏方法。BEiTv2[30] 也使用蒸馏方法，并引入向量量化知识蒸馏[28] 来训练分词器，使用

CLIP[31]。采用非参数化分词器的基于像素的掩码图像建模方法（如 MAE[21]和 SplitMask[16],）将原始像素或图像块视为预训练目标，而非词元，并且需要冗余的解码器。MaskFeat[35] 进一步引入手工设计的方向梯度直方图特征作为目标。与这些方法不同，我们的基于质心的掩码图像建模使用非参数分词器，在预训练中同时建模词元和像素。我们使用 K-means 聚类算法生成质心，从而避免分词器训练的成本。

3 掩码图像建模

掩码图像建模（MIM）最早由 BEiT 提出，它将 BERT 式预训练引入计算机视觉，并成功复现了自然语言处理领域的成功。具体而言，它首先将输入的二维图像拆分为一系列图像块，然后掩盖其中一定比例的图像块。与像素相比，基于词元的掩码图像建模的预训练目标是利用更高语义层级的视觉上下文（即离散视觉词元）重建受损图像。形式上，给定输入图像 $x \in \mathbb{R}^{C\times H\times W}$，首先将其拆分为 $n = HW/P^2$ 个图像块，并展平为 $\{x_i^p\}_{i=1}^n$，其中 (P, P) 是图像块的分辨率，且 $x^p \in \mathbb{R}^{n\times C\times P^2}$。它通过掩码比例 r_m 对输入图像 x 采样生成掩码，并根据被掩盖的位置 $\mathcal{M}$ 生成受损图像 $\hat{x}$：

$$\hat{x} = \{x_i^p \boldsymbol{E}_p \mid i \notin \mathcal{M}\}_{i=1}^n \bigcup \{e_m \mid i \in \mathcal{M}\}_{i=1}^n \tag{1}$$

其中， e_m 是可学习的掩码词元嵌入，$x_i^p \boldsymbol{E}_p$ 表示图像块嵌入的计算过程。

预训练的目标是恢复由分词器 $\{t_i\}_i^n$ 生成的视觉词元，其中 $t \in \mathcal{V}^n$、$\mathcal{V} = \{1, \ldots K\}$ 是包含离散词元 ID 的 K规模码本。受损图像将通过视觉 Transformer 编码器编码为 $\psi(\hat{x}) \in \mathbb{R}^{n\times d}$。该表示将输入线性分类头$lin.$：$\mathbb{R}^d \to \mathbb{R}^K$ 和 Softmax 运算符，以获得每个词元的概率 $p_{\mathrm{MIM}}(t_i \mid \hat{x}) = \mathrm{softmax}_{t_i}(lin. \circ \psi(\hat{x}))$。模型通过交叉熵损失进行训练，其目标是最大化正确词元 t_i 在受损图像$\hat{x}$条件下的对数似然：

$$\max \sum_x \mathbb{E}_{\mathcal{M}} \left[\sum_{i\in\mathcal{M}} \log p_{\mathrm{MIM}}\left(t_i \mid \hat{x}\right) \right] \tag{2}$$

其中， $p_{\mathrm{MIM}}(t_i \mid \hat{x})$是将$\hat{x}$ 输入视觉 Transformer 编码器以预测正确词元 t_i后得到的 Softmax 概率。

4 我们的方法

在我们的 CCViT 中，我们提出了基于质心的掩码图像建模，通过聚类质心对图像进行建模。整体架构如图 2 所示。质心

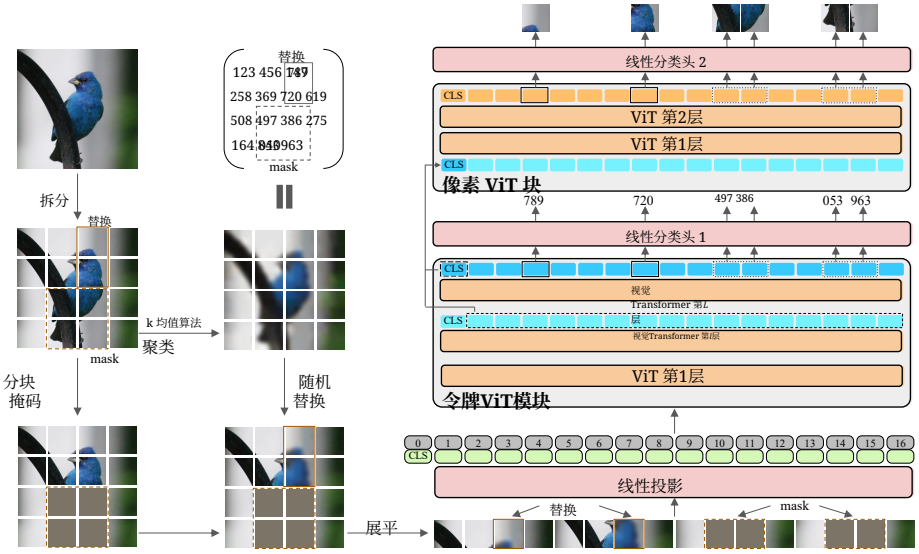


图 2. CCViT 概述。预训练之前，我们在原始像素块上使用 k 均值算法进行聚类，以获得质心。预训练期间，我们会掩码部分图像块，并使用质心随机替换部分未被掩码的图像块。所有图像块都会输入编码器。预训练目标包括质心索引令牌和原始像素。

具有双重视图，即像素视图和标记视图。每个质心都类似于一个平均图像块，因此我们可以直接获取像素。我们也可以轻松获取词元，即质心的索引，因为每个质心块都对应一个索引。正是由于这种双重属性，输出可以是词元或像素，并分别使用不同的损失函数，即交叉熵损失和均方误差。我们将所有图像块输入视觉Transformer编码器骨干网络，包括原始图像块、掩码图像块，以及被质心替换的图像块。

4.1 基于质心的分词器

基于词元的掩码图像建模首先需要将连续图像转换为离散标记，以便使用通过交叉熵损失实现的图像块分类目标来预训练模型。通常，它会采用在大规模数据集上训练的参数化分词器模型。我们的基于质心的分词器无需耗费大量资源来训练额外的参数化分词器模型。我们仅使用聚类方法（例如 K-means 算法）来获取一组质心。通过 K-means 算法得到的质心，其质量并不高度依赖于高标准和大规模的图像数据；只需使用预训练所用训练集的一小部分（例如 ImageNet 的 4%）即可获得，从而避免了基于词元的掩码图像建模中参数化分词器模型常会引入隐式附加数据集的风险。

我们使用 K-means 算法将每个原始图像块划分为 K个簇。给定图像块 $\{x_i^p\}_{i=1}^n$，将其展平为维度为 $D = C \times P^2$的向量。因此，维度为 N D的向量 $\mathcal{X} = \{x_i^v \in \mathbb{R}^D\}_{i=1}^N$ 将找到 k 个质心 $\{\mathcal{C}_k \in \mathbb{R}^D\}_{k=1}^K$，使代价最小：

$$\mathbb{E}(\mathcal{C}_1, \ldots, \mathcal{C}_K) = \frac{1}{N} \sum_{i=1}^{N} \|x_i^v - \mathcal{C}_{a(i)}\|_2 \tag{3}$$

其中，N 是 k 均值算法簇中的训练图像数量，$a(i)$是一个分配函数，定义为 $a(i) = \arg\min_{k \in \{1,\ldots,K\}} \|x_i^v - \mathcal{C}_k\|_2$。我们的基于质心的分词器具有局部不变性，因为我们只从局部视角分别处理每个图像块。

在将连续图像块转换为离散视觉词元的过程中，对于图像块 $x_i^p \in \mathbb{R}^{C \times P^2}$ 及其展平后的向量 $x_i^v \in \mathbb{R}^D$，其索引 $t_i \in \{1, \ldots K\}$ 可通过查找最近质心的索引获得，即：$t_i = \arg\min_{k \in \{1,\ldots,K\}} \|x_i^v - \mathcal{C}_k\|_2$

值得注意的是，由于我们只进行聚类操作，且不改变每个图像块的维度，因此质心本身就是一个易于可视化的图像块。这使得我们的基于质心的非参数分词器具有双重属性，既可视为像素，也可视为词元。换言之，我们可以根据词元 ID 得到输入图像块的近似版本。

4.2 用于建模的图像破坏

自监督预训练是一种破坏图像，然后利用图像内部相关性重建原始图像或其他形式的过程。在我们的 CCViT 中，我们采用了两种破坏策略：分块掩码和质心替换。与随机掩码相比，分块掩码更具挑战性，因为它提供的视觉上下文更少。在质心替换中，我们随机将一定比例的剩余图像块替换为对应的质心。这种技术使模型能够学习质心与索引、质心近似像素和原始像素之间的映射关系。这让我们的模型更抗噪，从而取得更好的效果。

形式化地，给定掩码比例 r_m 和替换比例 r_{re}，掩码位置可表示为 $\mathcal{M} \in \{1, \ldots, n\}^{r_m \times n}$，替换位置为$\mathcal{R} \in \{1, \ldots, n\}^{r_{re} \times n}$。注意，替换位置始终是未掩码的位置，以避免被掩码，即，$\mathcal{M} \bigcap \mathcal{R} = \emptyset$。因此，输入图像 $x \in \mathbb{R}^{C \times H \times W}$ 及其图像块序列 $\{x_i^p\}_{i=1}^n$ 可以被破坏为 $\tilde{x}$:

$$\tilde{x} = \{x_i^p \boldsymbol{E}_p \mid i \notin \mathcal{M} \bigcup \mathcal{R}\}_{i=1}^n \bigcup \{e_m \mid i \in \mathcal{M}\}_{i=1}^n \bigcup \{C_{a(i)} \boldsymbol{E}_p \mid i \in \mathcal{R}\}_{i=1}^n \tag{4}$$

其中，e_m 是可学习的掩码词元嵌入，$\mathcal{C}_i$ 是质心。$a(i)$是分配函数，其定义为 $a(i) = \arg\min_{k \in \{1,\ldots,K\}} \|x_i^v - \mathcal{C}_k\|_2$。

4.3 视觉Transformer骨干网络

如图 2 所示，我们使用视觉Transformer作为骨干网络。输入为受损图像块 $\tilde{x}$ 及其受损位置 $\mathcal{T} = \mathcal{M} \bigcup \mathcal{R}$，这些图像块经过展平并映射为 d维图像块嵌入 $\tilde{x_i}\boldsymbol{E}_p$，其中$\boldsymbol{E}_p \in \mathbb{R}^{d\times(C\times P^2)}$ 是由单个二维卷积层构成的映射投影。

由于基于质心的建模采用局部视角，我们需要 CLS 标记显式聚合全局表征，以弥补将原始图像块下采样为质心所造成的语义损失。我们添加像素 ViT 块，利用 CLS 标记收集全局信息；这在先前的基于词元的掩码图像建模中已被证明有效[18,30]。输入向量$\boldsymbol{H}_0 = [\boldsymbol{E}_{\texttt{CLS}}, \tilde{x_1}\boldsymbol{E}_p, \ldots, \tilde{x_n}\boldsymbol{E}_p]$ 将经过 L 层的令牌ViT模块 $\psi_t^L(\tilde{x})$，以预测被掩码和替换的词元。我们将第 l 层 $[\boldsymbol{h}_1^l, \ldots, \boldsymbol{h}_n^l]$ 的早期表征与第 L 层 $\boldsymbol{h}_{\texttt{CLS}}^L$的 CLS 标记拼接起来。

拼接后的向量 $\boldsymbol{H}_p = [\boldsymbol{h}_{\texttt{CLS}}^L, \boldsymbol{h}_1^l, \ldots, \boldsymbol{h}_n^l]$ 用作像素 ViT 块的输入，以进行原始像素重建。像素块 $\psi_p^2(\boldsymbol{H}_p)$仅由两层组成，因为质心的词元形式与像素形式密切相关。为了获得预测的词元和重建后的图像，令牌块和像素块的表征将分别通过两个不同的线性头 $lin.1(\cdot)$ 和 $lin.2(\cdot)$映射到质心索引空间和原始像素空间。

因此，我们计算两种损失：使用交叉熵损失衡量真实质心索引与预测索引概率 $p_{\mathrm{CIM}}(t_i \mid \tilde{x}) = \mathrm{softmax}_{t_i}(lin.1 \circ \psi_t^L(\tilde{x}))$ 之间的似然度，该概率是令牌块的输出。使用均方误差（MSE）量化像素块生成的图像块 $\psi_p^2(\boldsymbol{H}_p)^{\mathcal{T}}$ 与原始图像块 $x^{p^{\mathcal{T}}}$ 之间的差异。我们的 CCViT 预训练$\mathcal{L}_{\mathrm{CIM}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{MSE}}$ 的总损失可表示为:

$$\mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{MSE}} = -\sum_{\tilde{x}} \sum_{i\in\mathcal{T}} (\log(p_{\mathrm{CIM}}\left(t_i \mid \tilde{x}\right)) - \frac{1}{\lambda(\tilde{x}^{\mathcal{T}})} \|x^{p^{\mathcal{T}}} - \psi_p^2(\boldsymbol{H}_p)^{\mathcal{T}}\|_2) \quad (5)$$

其中， $\lambda(\tilde{x}^{\mathcal{T}})$ 表示元素数量，而 $\psi_p^2(\boldsymbol{H}_p)^{\mathcal{T}}$ 表示输入所有受损图像块（掩码、替换和原始图像块），并且只在受损部分（掩码和替换图像块）上计算损失。

5 实验

我们使用预训练模型在下游任务上进行实验，包括图像分类和语义分割，并遵循先前工作[2]采用的标准化评估协议。我们还对模型的关键组件进行了简要的消融研究。

5.1 预训练设置

质心聚类。 我们对每个原始像素块采用聚类方法，以获得质心及其对应的索引。我们使用 Faiss-GPU[23] 库运行 k 均值算法，并将簇数设为 $K =$

8192，以便与现有工作进行公平比较。我们使用 ImageNet-1K 训练集中的一小部分（5 万张图像）来训练质心，因为该数据集用于预训练，因此不会隐式引入其他数据集的信息。在实验中，每张图像都被调整为 224×224 分辨率，并拆分为 14×14 个图像块。因此，每个图像块的特征大小为 $d = 3 \times 16 \times 16 = 768$。根据使用 Faiss 对样本进行 k 均值聚类的经验，当迭代次数超过 20 次且训练点数达到 $1000 \times K$ 后，k-means 量化器的效果不再有稳定提升。因此，我们随机选取 ImageNet-1K 训练集中每个类别的 50 张图像，以 20 次迭代训练质心，即使用 5 万张图像和 980 万个图像块，仅占其总量的 4%。这一聚类阶段仅需约 150 秒，与耗时数天的参数化分词器模型训练相比，效率大幅提升。

基于质心的掩码图像建模。 为进行公平比较，我们采用掩码图像建模方法[2,30,26]所使用的设置。我们使用不含标签的 ImageNet-1K 训练集，通过自监督学习对模型进行预训练。我们将图像分辨率设为224×224，图像块大小设为 16×16，并对基础规模的视觉Transformer（ViT- Base/16）进行预训练。我们采用分块掩码，掩码比例为 40%，并以质心替换 10%，分别对应剩余的 75 个图像块和 20 个图像块。基础规模的 ViT 在令牌块中有 $L = 12$ 层；我们还在其后添加两层像素块，其输入为令牌块第 $l = 9$ 层的输出与最后第 $L = 12$ 层的 CLS 标记拼接后的结果。我们进一步展示了将方法扩展至 ViT-Large 时的有效性。

我们使用 4 块 NVIDIA GeForce RTX 3090 GPU 对模型进行 300 轮预训练。为保持与现有工作的批量大小一致，我们将梯度累积 4 次，使总批量大小达到 2048。更多超参数请参见附录。

5.2 图像分类

在 ImageNet-1K 上进行图像分类是视觉预训练的主要任务，因为该数据集与预训练所用的数据集相同。为确保公平比较，我们在微调时仅使用令牌ViT模块，以保证 ViT 模型参数量相同。我们采用微调 100 轮后的 Top-1准确率作为评估指标。更多超参数请参见附录。

我们在表1中展示了 ViT-B 的图像分类结果。我们的基础尺寸模型仅经过 300 轮预训练，在 ImageNet-1K 分类任务上便达到了 84.3% 的 Top-1准确率，优于掩码图像建模基线方法 BEiT $+1.4\%$、MAE $+0.7\%$。这验证了我们提出的 CCViT 与基于质心的掩码图像建模的有效性，因为我们仅使用 ImageNet-1K 数据集，以及一种易于构建且无需参数的监督信号。

近期研究常将潜在的外部数据集或强监督引入模型，包括 DALL-E 和 CLIP。配备外部信息的基线模型通常能取得显著提升。然而，在不使用外部数据的情况下与其他模型比较，并非公平的评估。我们超越了所有

表1. 我们的方法与先前工作的主要结果对比。下游任务的微调结果主要基于 ViT-Base 模型，包括 ImageNet-1K 上图像分类的 Top-1准确率（%）以及 ADE20K 上语义分割的平均交并比（%）。 †：由 MAE 复现的结果。 ‡：非参数方法，其分词器不含参数。

方法	监督	非参数‡	外部数据	轮数	准确率	平均交并比
监督式ViT[21]	标签	✔	✗	-	77.9	47.4†
Token-based with parametric tokenizer						
BEiT[2]	DALL-E	✗	✔	300	82.9	44.7
mc-BEiT[26]	VQGAN	✗	✗	300	83.9	-
iBOT[41]	自蒸馏	✗	✗	1600	84.0	50.0
PeCo[14]	VQGAN+MoCo	✗	✔	300	84.1	46.7
BEiTv2[30]	CLIP-B	✗	✔	300	85.0	52.7
Pixel-based						
MAE[21]	像素	✔	✗	1600	83.6	48.1
SplitMask[16]	图像块	✔	✗	300	83.6	45.7
MaskFeat[35]	HOG	✔	✗	300	83.6	-
RILS[38]	像素+LAION	✔	✔	1000	83.6	48.1
Centroid-based						
Ours	**质心**	✔	✗	**300**	**84.3**	**48.4**

仅使用 ImageNet-1K 数据的基线模型，甚至超越了 BEiT 和 PeCo 等使用外部数据的基线模型。我们的模型表现不如 BEiTv2，后者采用了非常强的监督，并隐式地融入了来自 CLIP 的数据。与训练轮次更多的模型相比，我们的模型也取得了有竞争力的结果。这表明，以质心为目标的预训练目标，在分类任务上比仅令牌和仅像素更高效。

此外，我们在表 2 中评估了模型在更大型视觉Transformer骨干网络和更长预训练轮数下的扩展能力。具体而言，我们的 ViT-L 模型在 ImageNet-1K 分类任务中达到 86.0%，这表明随着视觉Transformer骨干网络规模增大，模型性能显著提升。我们分别以 $+0.8\%$ 和 $+0.1\%$ 的优势超越了掩码图像建模基线方法 BEiT 和 MAE。通过更长时间、更大规模的预训练，我们的模型取得了显著提升，这证明模型能够从扩展规模中获益。

5.3 语义分割

语义分割旨在预测每个像素所属的类别，可视为像素级分类任务。我们使用 ADE20K[40]基准，并报告对所有语义类别取平均后的平均交并比（mIoU）。在模型架构方面，我们使用 ViT-Base/16 作为骨干网络，并使用 UPerNet[37] 作为语义分割任务头。为进行公平比

表 2. ImageNet-1K 上扩展能力的微调结果。我们展示了更大的骨干网络（ViT-L）和更长的训练轮次（800 轮）。

方法	监督	ViT-B	+更长	+更大
mc-BEiT	VQGAN	83.9	84.1	85.6
iBOT	蒸馏	83.8	84.0	84.8
PeCo	MoCo	84.1	84.5	86.5
BEiTv2	CLIP-B	85.0	-	86.6
MaskFeat	HOG	83.6	83.9	84.4
BEiT	DALL-E	82.9	83.2	85.2
MAE	像素	-	83.6	85.9
Ours	**质心**	**84.3**	**84.4**	**86.0**

较，我们进行了 160k 步微调，并将批量大小设为 16。更多超参数详情请参阅附录。

根据表1，我们的方法在平均交并比上达到48.4，较BEiT和MAE分别高出 $+3.7$和 $+0.3$。我们还超越了大多数采用相同300个训练轮次的方法，并超过了通过CLIP进行隐式蒸馏的基线方法。根据BEiT的做法，在ImageNet-1K上进行中间微调可进一步提升ADE20K上的性能。我们的平均交并比达到51.6，提升了 $+3.2$。附录中展示了中间结果的详细信息和比较。

5.4 消融研究

表 3. 针对ImageNet-1K分类和ADE20K分割，采用不同预训练设置进行的消融研究。“替换”表示采用随机替换策略。

词元	像素	替换	Top-1准确率	平均交并比
✔	✗	✗	84.12	47.42
✔	✗	✔	84.21	47.89
✔	✔	✔	84.30	48.35

我们使用视觉Transformer-Base模型进行比较，并对关键组件进行了消融研究。评估任务包括ImageNet-1K上的图像分类和ADE20K上的语义分割。预训练设置为300个训练轮次；微调在分类任务中进行100个训练轮次，在分割任务中进行160k步。

结果见表 3。关于预训练目标，我们发现，与仅使用词元或像素作为目标相比，同时使用词元和像素作为目标效果最佳（84.30 vs 84.21，以及 48.35 vs 47.89）。在基于质心的建模中，同时学习词元和像素表征确有裨益。此外，我们对不同的扰乱方法进行了消融实验，发现采用随机替换的模型优于仅采用分块掩码的模型

在分类（84.21 vs 84.12）和分割任务（47.89 vs 47.42）中均是如此。这表明，采用随机替换有助于模型学习像素与词元之间的对应关系。

5.5 基于质心的分词器的进一步分析

表4. 比较掩码前后的未变化词元比例，以及预训练中的词元预测准确率。

方法	Mask			词元预测准确率
	0.1	0.2	0.5	
BEiT	34.34	14.17	1.41	9.49
BEiTv2	59.61	33.56	3.97	4.96
Ours	**90.01**	**80.02**	**50.05**	**40.83**

下游任务的表现与分词器密切相关。使用更大数据集（如 BEiT 中的 DALL-E）训练、并从大型模型（如 BEiTv2 中的 CLIP）蒸馏得到的分词器，往往能取得更好的结果。然而，它们需要在预训练前消耗大量训练资源，而且生成视觉词元的推理时间也更长。

最重要的是，这些参数化分词器生成的每个视觉词元都无法准确表示对应的图像块。单个词元可能与多个图像块相关联，而非专门表示处于相同位置的图像块。即使某个图像块保持不变，如果其他图像块发生改变，该词元也可能随之改变。这会导致词元与图像块之间出现错位。然而，采用先掩码后预测框架的先前基于词元的掩码图像建模方法，其预训练目标与此相矛盾：它旨在根据掩码图像预测视觉词元中被掩码的部分。此外，实际图像块通常具有局部不变性，也就是说，替换一个图像块不会影响其他图像块。现有视觉分词器无法保持这种局部视角，而参数化分词器生成的词元也无法独立表示单个语义信息。

表4中的实验为我们的论点提供了证据。我们首先掩码一定比例的图像块，并将输出词元与原始未掩码图像得到的词元进行比较。我们的质心分词器成功保持了图像块的局部不变性：掩码操作前后令牌编号的变化比例与掩码比例一致。相比之下，BEiT 和 BEiTv2 的词元变化程度高于掩码比例。当 50% 的词元被掩码时，只有 1.41% 和 3.97% 的词元保持不变。这表明无法保证图像块与视觉词元之间的局部对应关系。

此外，我们的模型在预训练期间达到了最高的词元预测准确率，这进一步证明了我们的模型更适用于掩码

后预测框架，能够基于视觉上下文预测视觉词元。相比之下，BEiT 和 BEiTv2 未能学习到足够的图块间关系，无法通过图块间相关性进行掩码推理。

6 结论与未来工作

采用参数化分词器模型的现有基于词元的掩码图像建模方法面临训练成本高的问题，而基于像素的掩码图像建模则需要使用冗余解码器来对齐原始像素。在这项工作中，我们提出了一种新颖的质心中心化视觉Transformer预训练框架，利用基于质心的掩码图像建模来建模图像。借助 K-means 聚类算法，我们的方法效率很高，构建分词器所需的训练数据和时间极少。我们的 CCViT 在分类和分割等下游任务中表现出色。进一步分析表明，由于具有局部不变性，我们的基于质心的分词器更适合上下文依赖架构预训练。未来，我们计划探索蒸馏方法以取得更好的结果，因为这与我们的贡献正交。

参考文献

1. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.：data2vec：一种用于语音、视觉和语言自监督学习的通用框架。载于：ICML，第162卷，第1298–1312页（2022） 2. Bao, H., Dong, L., Piao, S., Wei, F.：BEiT：图像Transformer的BERT预训练。载于：ICLR（2022） 3. Caron, M., Bojanowski, P., Joulin, A., Douze, M.：用于视觉特征无监督学习的深度聚类。载于：ECCV，第132–149页（2018） 4. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.：通过对比聚类分配进行视觉特征的无监督学习。载于：NeurIPS，第33卷，第9912–9924页（2020） 5. Caron, M., Touvron, H., Misra, I., J´egou, H., Mairal, J., Bojanowski, P., Joulin, A.：自监督视觉 Transformer 中的新兴特性。载于：CVPR（2021） 6. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.：从像素进行生成式预训练。载于：ICML，第1691–1703页（2020） 7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.：用于视觉表征对比学习的简单框架。载于：ICML，第1597–1607页。PMLR（2020） 8. Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., Wang, J.：用于自监督表征学习的上下文自编码器。arXiv预印本 arXiv:2202.03026（2022） 9. Chen, X., He, K.：探索简单的孪生表征学习。载于：CVPR，第15750–15758页（2021） 10. Chen, X., Xie, S., He, K.：自监督视觉 Transformer 训练的实证研究。载于：ICCV，第9640–9649页（2021） 11. Chen, Y., Liu, Y., Jiang, D., Zhang, X., Dai, W., Xiong, H., Tian, Q.：SdAE：自蒸馏掩码自编码器。载于：ECCV，第108–124页（2022） 12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.：ImageNet：大型层次化图像数据库。载于：CVPR，第248–255页（2009）

13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.：BERT：用于语言理解的深度双向 Transformer 预训练。arXiv 预印本 arXiv:1810.04805 （ 2018 ） 14. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N., Guo, B.：PeCo：用于视觉 Transformer 的 BERT 预训练的感知码本。arXiv预印本 arXiv:2111.12710（2021） 15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 等：一张图像胜过16×16个词：用于大规模图像识别的 Transformer。载于：ICLR（2021） 16. El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., Grave, E.：自监督预训练是否需要大规模数据集？arXiv预印本 arXiv:2112.10740（2021） 17. Esser, P., Rombach, R., Ommer, B.：驾驭 Transformer，实现高分辨率图像合成。载于：CVPR，第12873–12883页（2021） 18. Gao, L., Callan, J.：Condenser：一种用于稠密检索的预训练架构。载于：EMNLP，第981–993页（2021） 19. Gidaris, S., Singh, P., Komodakis, N.：通过预测图像旋转进行无监督表征学习。载于：ICLR（2018） 20. Grill, J.B., Strub, F., Altch´e, F., Tallec, C., Richemond, P., Buchatskaya, E., Do- ersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., 等：Bootstrap Your Own Latent——一种新的自监督学习方法。载于：NeurIPS，第33卷，第21271–21284页（2020） 21. He, K., Chen, X., Xie, S., Li, Y., Doll´ar, P., Girshick, R.：掩码自编码器是可扩展的视觉学习器。载于：CVPR，第16000–16009页（2022） 22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.：用于无监督视觉表征学习的动量对比。载于：CVPR，第9729–9738页（2020） 23. Johnson, J., Douze, M., J´egou, H.：利用 GPU 进行十亿规模的相似性搜索。《IEEE大数据汇刊》**7**(3)，535–547（2019） 24. Larsson, G., Maire, M., Shakhnarovich, G.：将着色作为视觉理解的代理任务。载于：CVPR（2017年7月） 25. Li, J., Zhou, P., Xiong, C., Hoi, S.：无监督表征的原型对比学习。载于：ICLR（2021） 26. Li, X., Ge, Y., Yi, K., Hu, Z., Shan, Y., Duan, L.Y.：mc-BEiT：用于图像BERT预训练的多选离散化。载于：ECCV，第231–246页（2022） 27. Noroozi, M., Favaro, P.：通过解决拼图任务进行视觉表征的无监督学习。载于：Leibe, B., Matas, J., Sebe, N., Welling, M.（编）ECCV，第69–84页（2016） 28. van den Oord, A., Vinyals, O., Kavukcuoglu, K.：神经离散表征学习。载于：NeurIPS，第6306–6315页（2017） 29. Peng, Z.：MAE-pytorch。https://github.com/pengzhiliang/MAE-pytorch（2021） 30. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.：BEiT v2：使用向量量化视觉分词器进行掩码图像建模。arXiv预印本 arXiv:2208.06366（2022） 31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等：从自然语言监督中学习可迁移视觉模型。载于：ICML（2021） 32. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.：零样本文本到图像生成。载于：ICML，第8821–8831页（2021） 33. Rolfe, J.T.：离散变分自编码器。载于：ICLR（2017） 34. Tan, H., Lei, J., Wolf, T., Bansal, M.：VIMPAC：通过掩码词元预测和对比学习进行视频预训练。arXiv预印本 arXiv:2106.11250（2021）

35. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.：用于自监督视觉预训练的掩码特征预测。载于：CVPR，第14668–14678页（2022） 36. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.：通过非参数实例判别进行无监督特征学习。载于：CVPR，第3733–3742页（2018） 37. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.：面向场景理解的统一感知解析。载于：ECCV论文集，第418–434页（2018） 38. Yang, S., Ge, Y., Yi, K., Li, D., Shan, Y., Qie, X., Wang, X.：RILS：语言语义空间中的掩码视觉重建（2023） 39. YM., A., C., R., A., V.：通过同步聚类和表征学习实现自标注。载于：ICLR（2020） 40. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.：通过ADE20K数据集实现对场景的语义理解。国际计算机视觉杂志 **127**(3)，302–321（2019） 41. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.：iBOT：采用在线分词器的图像BERT预训练。载于：ICLR（2022）

A 预训练超参数

表5. CCViT 在 ImageNet-1K 上预训练的超参数。

超参数	**取值**
令牌块层	12
像素块层	2
层缩放	0.1
图像块大小	16×16
位置嵌入	相对
预训练轮数	300
批量大小	2048
优化器	AdamW
优化器动量	$\beta_1 = 0.9, \beta_2 = 0.98$
峰值学习率	1.5e-3
最低学习率	1e-5
学习率调度	余弦延迟
预热周期	10
梯度裁剪	3.0
随机失活	✗
路径丢弃	✗
随机路径	✗
权重衰减	0.05
数据增强	RandomResizedAndCrop
图像分辨率	224×224

B 图像分类微调超参数

表 6. 在 ImageNet- 1K 上进行 CCViT 图像分类微调的超参数。

超参数	**取值**
位置嵌入	相对
微调轮数	100
批量大小	1024
优化器	AdamW
优化器动量	$\beta_1 = 0.9, \beta_2 = 0.999$
峰值学习率	2e-3
最低学习率	1e-6
学习率调度	余弦延迟
逐层学习率衰减	0.65
预热周期	20
梯度裁剪	✗
标签平滑	0.1
随机失活	✗
随机深度	0.1
权重衰减	0.05
数据增强	RandAug（9，0.5）
图像分辨率	224×224
随机擦除	0.25
混合增强	0.8
CutMix	1.0

C 语义分割微调的超参数

表 7. CCViT 在 ImageNet-1K 上预训练的超参数。

超参数	**取值**
位置嵌入	相对
微调步数	160K
批量大小	16
优化器	AdamW
优化器动量	$\beta_1 = 0.9, \beta_2 = 0.999$
峰值学习率	8e-5
最低学习率	0
学习率调度	线性
逐层学习率衰减	0.9
预热步数	1500
随机失活	✗
随机深度	0.1
权重衰减	0.05
输入分辨率	512×512

D 质心分词器的消融研究

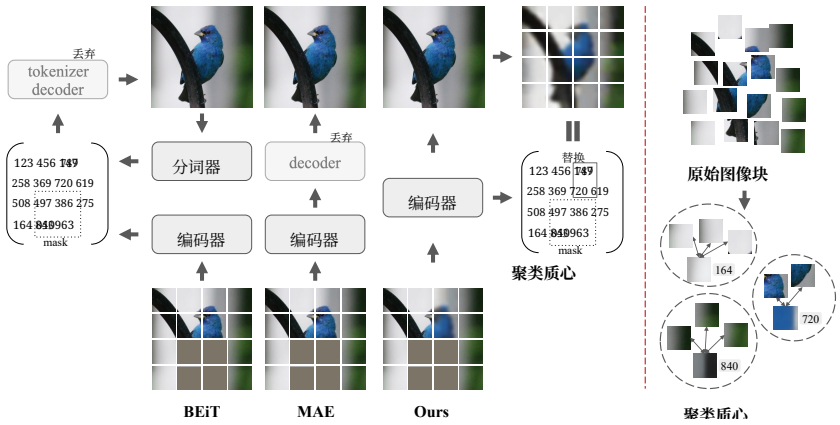


图 3。 BEiT、MAE 与我们的方法之间的预训练架构比较。

图 3 展示了我们基于质心的 CCViT、基于标记的 BEiT 和基于像素的 MAE 在预训练架构方面的主要对比。我们的输入由掩码图像块、替换图像块（质心）和原始图像块组成。我们的模型架构不包含冗余解码器，进一步简化了流程。我们将词元和像素都用作预训练目标。

E 质心分词器的消融研究

表 8. BEiT“框架”下 ImageNet-1K 图像分类的微调结果。†：将基线模型与 BEiTv2 进行比较并不公平，因为后者采用了一个额外的、用于图像块聚合的视觉Transformer分支。此外，其分词器是通过从强大的 CLIP 监督信号中蒸馏得到的。

方法	轮次	ViT-B	ViT-L
BEiT	800	83.2	85.2
BEiTv2†	1600	85.5	87.3
mc-BEiT	800	84.1	85.6
MAE	400	83.1	84.5
Ours	**300**	**84.2**	**86.0**

在此设置下，我们进一步消融研究视觉分词器的有效性。我们在整体的 BEiT 风格框架上报告了不同基线的结果。对于我们的 CCViT，我们移除像素 ViT 块，仅使用词元作为预训练目标，以匹配 BEiT 框架。对于 MAE，由于其预训练目标基于像素，我们报告了基于 BEiT 框架 [29] 的实现结果。[2] 结果见表 8。与其他视觉分词器相比，我们的分词器是最有效的方法。请注意，BEiTv2 由强大的 CLIP 模型监督，并采用了不同的架构，因此直接比较并不公平。

F 在 ADE20K 上进行中间微调

表9. 在 ADE20K 上进行中间微调，该模型在 ImageNet-1K 分类任务上进行预训练和微调。

方法	监督	轮次	平均交并比
BEiT	DALL-E	800	45.6
mc-BEiT	VQGAN	800	47.0
Ours	质心	300	48.4
+ Intermediate fine-tuning			
BEiT	DALL-E	800	47.7
mc-BEiT	VQGAN	800	50.8
Ours	质心	**300**	**51.6**

根据 BEiT，在 ImageNet-1K 上进行中间微调可以进一步提升 ADE20K 上的性能。因此，我们进行了这项实验

[2]基于 BEiT 框架的 MAE GitHub 实现：https://github.com/pengzhiliang/MAE-pytorch。

并将其与 BEiT 和 mc-BEiT 进行了比较。我们首先立即在 ImageNet-1K 上微调 100 个训练轮次，然后按常规在 ADE20K 上进行语义分割微调。表 9 汇报了性能结果。我们提出的 CCViT 方法在平均交并比上达到 51.6%，比我们的仅预训练模型提升了 $+3.2\%$，并显著优于先前方法 BEiT 和 mc-BEiT，其平均交并比分别为 $+3.9\%$ 和 $+0.8\%$。

G 不同分词器的性能

表10. 不同分词器的性能。

方法	训练	推理	内存	参数量	浮点运算次数
BEiT	天数	15.7±2.5毫秒	21.5G	98M	3.3T
BEiTv2	天	74.8±0.6ms	13.2G	109M	2.3T
Ours	**158秒**	**9.3±0.5毫秒**	**3.4G**	**0**	**805M**

表11. 不同图像噪声下未变化词元比例的比较。

方法	高斯噪声			高斯模糊		
	1	10	25	0.5	1	2
BEiT	88.02	32.54	9.31	61.18	25.32	6.93
BEiTv2	95.03	57.43	24.02	83.52	61.29	0.08
Ours	**98.94**	**88.61**	**72.28**	**96.38**	**86.39**	**66.72**

本节在表10和表11中展示不同分词器的性能。

在分词器训练速度方面，基于质心的分词器仅需在单块 RTX 3090 NVIDIA图形处理器上耗时约 158 秒即可构建完成，而 BEiT 和 BEiTv2 的分词器则需要数天[3]。在分词器推理速度方面，在相同环境（单块 RTX 3090）下，使用相同的批量大小（包含 64 张图像），我们的模型仅需 9.3 毫秒，而 BEiT 和 BEiTv2 的分词器分别需要 15.7 毫秒和 74.8 毫秒。值得注意的是，我们的分词器占用的 GPU 内存非常少，而且可以将速度扩展到更大的批量大小，从而充分利用 GPU 的并行能力，进一步扩大我们的速度优势。我们还计算了不同分词器的浮点运算次数；得益于非参数化设计，我们的质心分词器大幅优于其他分词器。

我们评估了不同分词器在不同图像噪声下的抗噪性。我们通过测量未变化词元比例来展示结果

[3] 这取决于硬件环境。BEiTv2分词器在单张 RTX 3090 NVIDIA图形处理器上需要训练 7 天。



图4. 表4和表11中不同图像噪声的可视化。

见表11。我们的质心分词器始终显著优于其他参数化分词器，尤其是在施加较强噪声时。此外，图4展示了所探索的不同噪声类型。

H 像素重建

在本节中，我们展示了 CCViT 能够利用视觉上下文进行语义预测。我们探讨了仅使用预训练编码器进行图像重建、而不使用任何额外模型的可行性。以往基于像素的掩码图像建模方法使用解码器，并将原始像素作为预训练目标来预测掩码图像块；而基于词元的掩码图像建模方法则需要额外的分词器模型。仅利用编码器输出的质心和词元概率向量，CCViT 就能轻松重建被掩码的部分。请注意，掩码图像建模方法并非旨在实现令人满意的图像修复效果。

我们比较了 BEiT [2], MAE [21], 和 CCViT 的重建示例，三者分别代表基于词元、基于像素和基于质心的掩码图像建模。基于像素的 MAE 使用预训练模型的直接输出，因为其预训练目标是原始像素。基于词元的 BEiT 使用额外的分词器解码器，并将预测的掩码词元作为输入。基于质心的 CCViT 则直接使用预测的掩码词元对应的质心，从而无需额外的参数化模型。我们展示了 CCViT 的纯词元版本和完整版本的结果：前者仅以索引令牌作为预训练目标，后者则同时以词元和像素为目标。即 重建示例见图 5。由于 MAE 往往会破坏未掩码图像块，为缓解图像块之间明显的边界问题，我们将未掩码图像块替换为真实值，并对 CCViT 的输出进行高斯模糊处理以供可视化。

在这种情况下，图像中的大部分“青椒”都被掩码了，只留下茎可见。BEiT 的重建结果与真实值相差甚远。MAE 只是粘贴周围的图像，并在茎下方补上两个橙子。我们的 CCViT 在形状预测方面表现更好。仅令牌 CCViT 的预测甚至更好。这表明，尽管 CCViT 的预训练严格仅使用

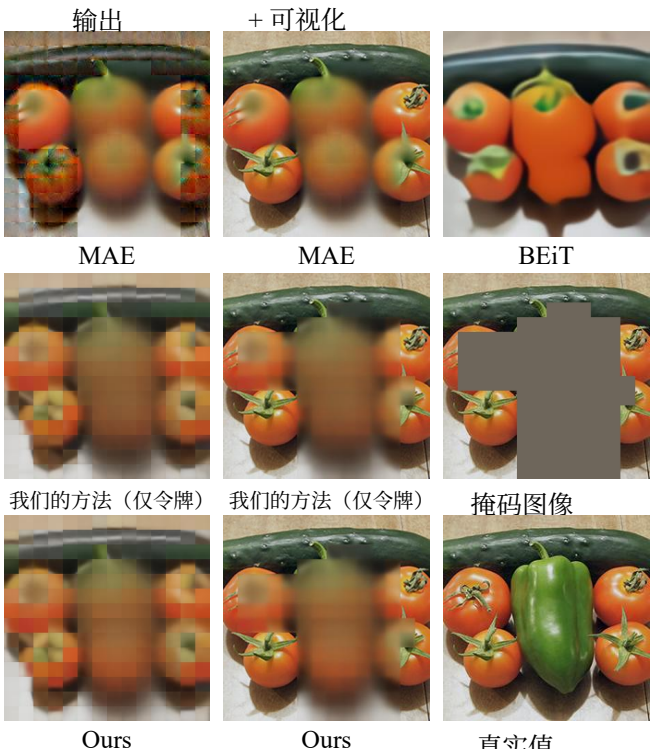


图5. 通过 BEiT、MAE 和我们的 CCViT，从 ImageNet-1K *validation* 数据集中获得的重建示例。

质心令牌目标，且没有明确引入像素目标，但 CCViT 仍能通过质心令牌隐式预测像素，并学习质心像素与索引之间的相关性。更多重建示例见图6。

I 质心可视化

与基于词元的掩码图像建模不同，后者难以明确地可视化每个词元所代表的单一语义；我们在图7中展示了全部8192个质心。每个质心都是一个 16×16 图像块，同时也代表一个索引令牌。

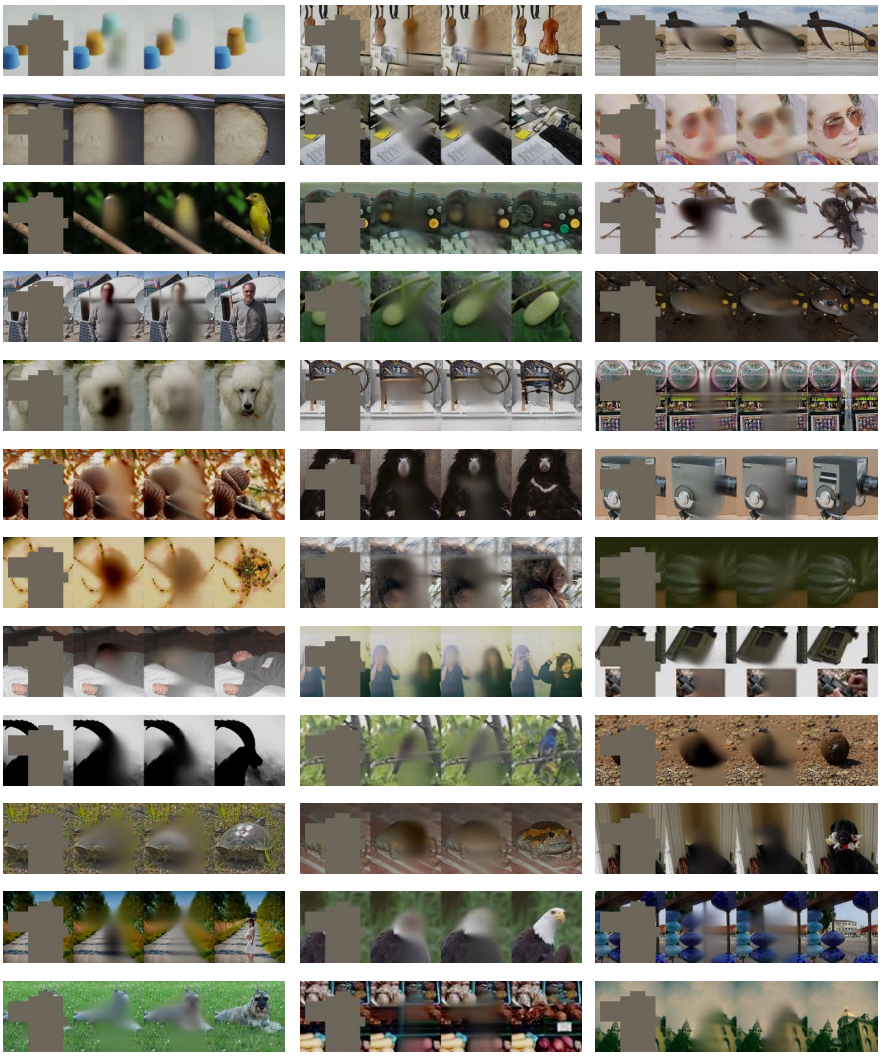


图6. ImageNet-1K *test* 图像的图像重建示例。对于每组四张图像，从左至右依次为掩码图像、MAE重建结果、我们的CCViT重建结果和真实值。掩码比例为50%。

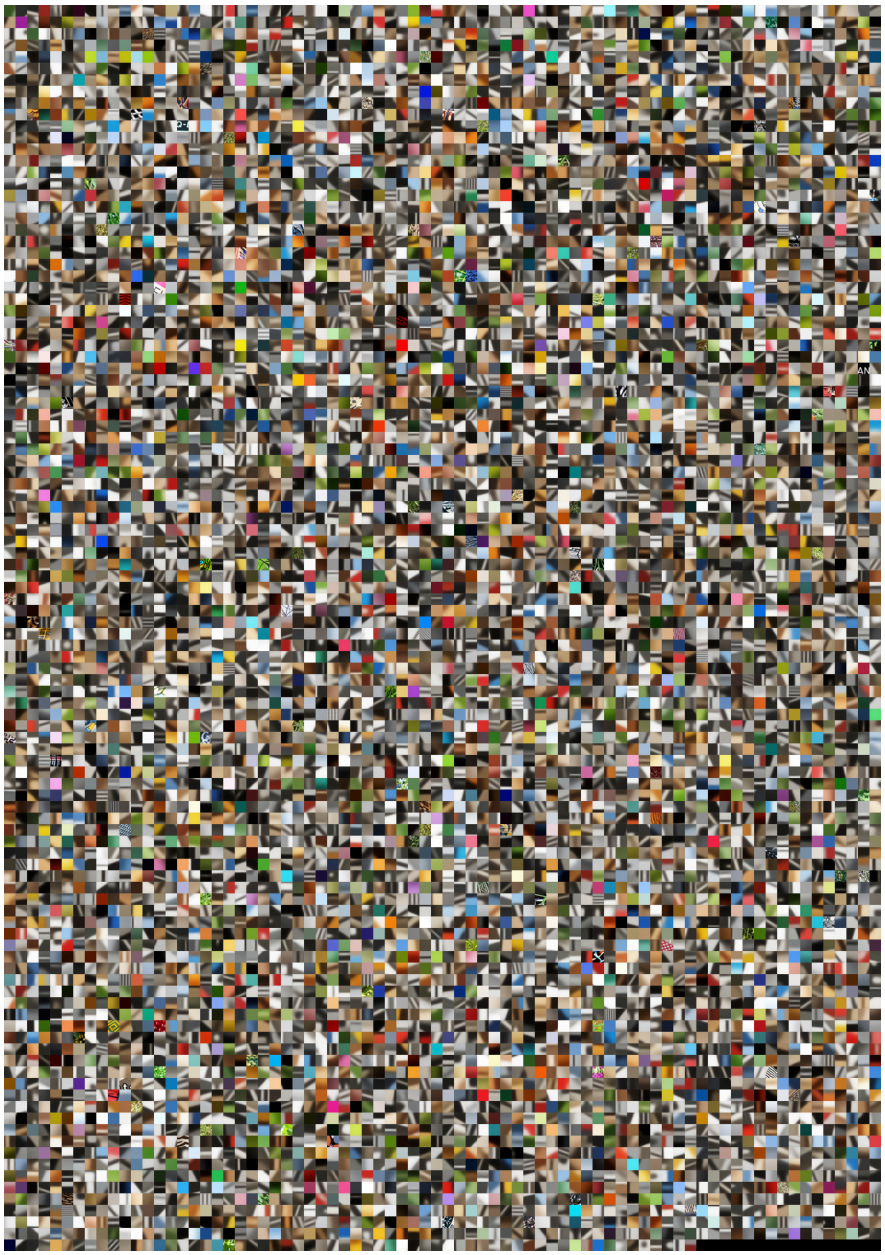


图 7. 8192 个质心的可视化。