arXiv:2305.04440v2 [cs.CV] 4 Mar 2024

开箱即用的视觉变换器：少样本类别无关计数的意外基线

Zhicheng Wang、Liwen Xiao、Zhiguo Cao、Hao Lu

图像处理与智能控制重点实验室，教育部；人工智能与自动化学院，华中科技大学，中国武汉 430074 zhicheng wang@hust.edu.cn

摘要

类别无关计数（CAC）旨在给定少量示例，从查询图像中计数感兴趣的物体。通常，这项任务通过分别提取查询图像和示例的特征，再匹配它们的特征相似度来完成，从而形成“提取然后匹配”的范式。本文表明，CAC 可以简化为“提取并匹配”的方式，尤其是借助视觉变换器（ViT），在其中，特征提取和相似度匹配通过自注意力同时执行。我们从解耦的视角揭示了这种简化的原理。由此得到的模型称为 CACViT，它将 CAC 流程简化为单个预训练纯视觉变换器。此外，为补偿纯视觉变换器中的缩放和归一化所导致的尺度信息与数量级信息损失，我们提出了两种有效策略，分别用于尺度嵌入和数量级嵌入。在 FSC147 和 CARPK 数据集上的大量实验表明，CACViT 在效果（误差降低 23.60%）和泛化能力方面均显著优于最先进的类别无关计数方法，这表明 CACViT 为 CAC 提供了一个简洁而强大的基线。代码即将发布。

引言

目标计数旨在估计查询图像中目标的数量。大多数以往的目标计数方法都针对特定领域，例如人群（Zhang 等人，2015；Shu 等人，2022；Zou 等人，2021）、植物（Lu 等人，2017；Madec 等人，2019）和汽车（Onoro-Rubio 和 L´opez-Sastre，2016）。它们通常需要大量特定类别的训练数据，才能学习出良好的模型（Wang 等人，2020）。相比之下，类别无关计数（CAC）旨在仅凭少量示例估计任意类别的计数值。由于有望推广到未见场景，并减少对特定类别训练数据的依赖，这一方向近来备受关注（Lu、Xie 和 Zisserman，2019；Ranjan 等人，2021；Shi 等人，2022；Liu 等人，2022）。

类别无关计数由 Lu 等人首次提出（Lu、Xie 和 Zisserman 2019），该问题默认被表述为模板匹配问题，从而形成了提取-再-匹配范式。以往的模型（Ranjan 等人，2021；Shi 等人，2022；Lin 等人，2022）使用共享的卷积神经网络进行查询图像和示例的特征提取，因为卷积神经网络的自底向上特征提取方法能够适应尺寸完全不同的图像。鉴于交叉注意力机制能够在注意力图上标记响应，一些模型（如 CounTR（Liu 等人，2022））采用交叉注意力来匹配查询图像与示例的特征。然而，在 CounTR 中，查询特征和示例特征分别由视觉 Transformer和卷积神经网络进行嵌入，匹配部分则通过额外的交叉注意力阶段实现。这种策略引入了大量冗余和特定于任务的设计，与任务无关基础模型的发展趋势并不一致。

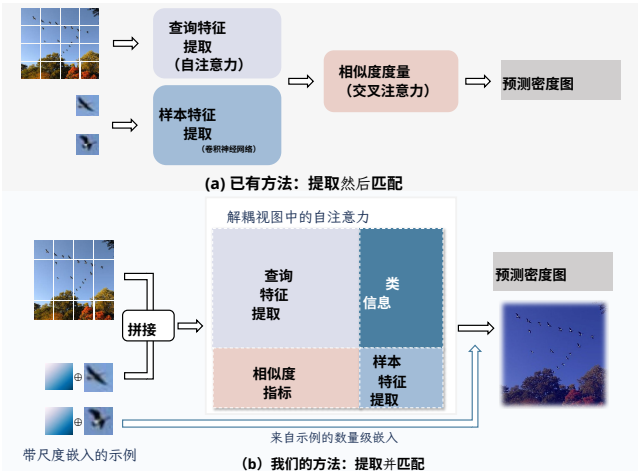


图1：已有方法与我们方法的高层思路。（a）以视觉Transformer为基础的先前类别无关计数框架遵循“先提取再匹配”的范式，针对查询图像和示例使用不共享的特征提取器（例如，视觉Transformer和卷积神经网络），并在特征提取后进行交叉注意力等匹配；（b）我们基于视觉Transformer的框架遵循“提取并匹配”的范式，在解耦视图中使用自注意力，并额外采用感知宽高比的尺度嵌入（SE）和数量级嵌入（ME），以弥补视觉Transformer中尺度的信息损失。

近年来，计算机视觉界见证了纯 ViT 在大型多模态架构中取得的巨大成功（Touvron 等，2023年；Yu 等，2022年）。很快，大量关于更好地将 ViT 适配到下游视觉任务的研究

版权所有 © 2024，人工智能促进协会（www.aaai.org）。保留所有权利。

应运而生，例如目标检测（Li 等，2022年；Lin 等，2023年）、姿态估计（Xu 等，2022年、2023年）和图像抠图（Yao 等，2023年）。作为一项模板匹配任务，类别无关计数本质上适合利用 ViT 及其注意力机制；然而，关于将 ViT 适配到 CAC 任务的研究却很少。

在这项工作中，我们分享了一些见解：普通视觉Transformer中的注意力机制能够提取查询图像和示例的特征，并对它们进行特征匹配。将查询和示例的令牌拼接在一起，并输入普通视觉Transformer后，视觉Transformer中的自注意力过程可以分为两组自注意力和两组交叉注意力。前两组自注意力用于提取查询图像和示例的特征，而后两组交叉注意力则包含查询图像与示例之间的匹配过程。因此，无需多个特征提取器或额外的匹配后处理，它便形成了一种新颖的提取与匹配范式。与已有方法相比，在这一范式中，从查询特征到示例的额外注意力能够为查询图像提供更多类别信息，从而更好地感知目标。基于这一思路，我们提出了一个用于类别无关计数的框架，该框架主要由单个预训练视觉Transformer构成，验证了普通视觉Transformer用于 CAC 任务的可行性。

为了让视觉Transformer更好地适应特定的 CAC 任务，我们在模型设计中引入了更多与 CAC 任务密切相关的考量。具体而言，我们发现，该架构中的某些限制或操作（如调整尺寸和 softmax 归一化）会导致尺度信息以及计数值数量级信息的丢失。首先，示例必须调整尺寸以适配视觉Transformer的输入，这会在匹配过程中造成尺寸歧义。以往基于卷积神经网络的模型（Shi 等人，2022）尝试通过为示例添加尺度嵌入来弥补尺度信息；然而，它们忽略了长宽比信息，而这对于比例异常的类别至关重要。现有文献在很大程度上忽视了这一点。其次，经过 softmax 函数处理的注意力图能够表示查询图像中物体的相对分布，因此会削弱模型对物体数量的感知。我们通过恢复归一化注意力图中的数量级顺序来解决这一问题。所提出的尺度嵌入和数量级嵌入都易于实现。将尺度信息和数量级信息融入纯粹的视觉Transformer架构后，我们得到了一个出人意料地简单而高效的 CAC 基线模型。该模型称为 CACViT，充分利用了视觉Transformer中的自注意力机制，同时针对这一架构在该任务中的缺陷进行了调整。

在公开基准数据集 FSC147（Ranjan 等人，2021）上的实验表明，CACVit 的表现大幅优于此前最佳方法；就平均绝对误差而言，其在验证集和测试集上的相对误差分别降低了 19.04% 和 23.60%。我们还在汽车计数数据集 CARPK（Hsieh、Lin 和 Hsu，2017）上展示了该方法的跨数据集泛化能力。此外，我们还进行了大量消融研究，以验证所提出方法的合理性。

总而言之，我们的贡献主要有三点：

- 一种新颖的提取与匹配范式：我们证明，在类别无关计数中可以同时进行特征提取与匹配；
- CACViT：一种简单而强大的基于视觉Transformer的基线模型，在FSC-147基准上取得了新的最先进性能；
- 我们提出两种有效策略，将尺度、长宽比和数量级信息嵌入 CACViT。

相关工作

类别无关计数包含两个主要组成部分：特征提取和特征匹配。我们首先回顾以往计数模型中的这两个组成部分，随后讨论其他领域中联合进行特征提取与匹配的方法。

类别无关计数中的特征提取。计数任务中的特征提取研究最初始于类别特定计数（Abousamra 等人，2021；Cao 等人，2018；He等人，2021；Idrees 等人，2018；Laradji 等人，2018；Cheng 等人，2022）。在类别特定计数中，大多数工作旨在应对数量变化和尺度变化带来的挑战。对于类别无关计数，除上述挑战外，特征提取的核心还包括构建统一匹配空间。为了获得统一匹配空间，以往大多数工作（Ranjan 等人，2021；Shi 等人，2022；You 等人，2023）都对查询图像和示例使用共享的基于 CNN 的特征提取器。CounTR（Liu 等人，2022）首次将视觉Transformer用于类别无关计数中的特征提取，并为查询图像（使用视觉Transformer）和示例（使用卷积神经网络）采用了不同的特征提取器。因此，该方法使用两阶段训练方案来统一特征空间。

类别无关计数中的特征匹配。 与特征提取相比，类别无关计数中的匹配策略受到更多关注。匹配的关键点包括以下两方面：1）对外观变化的鲁棒性；2）表征数量级的能力。早期尝试中使用了朴素的内积（Ranjan 等人，2021；Yang 等人，2021），但它对待计数对象的外观变化不够鲁棒。Shi 等人 （Shi 等人，2022）开发了一种双线性匹配网络（BM-Net），将固定的内积扩展为可学习的双线性相似度度量，相较于内积提升了鲁棒性。近期基于视觉Transformer的模型 CounTR（Liu 等人，2022）使用交叉注意力进行匹配，乍看之下，这似乎是基于 Transformer 的方案的自然选择。然而，我们表明，在简单的 ViT 模型 CACViT 中，我们可以通过自注意力在提取特征的同时进行特征匹配。

联合特征提取与匹配。 对于模板匹配和多模态任务，特征提取与匹配是两个主要组成部分。在跟踪和检测任务中，MixFormer 网络（Chen 等，2022年）和 FCT 网络（Han 等，2022年）被提出用于增强目标对象与图像之间的相关性，从而为定位头获得增强特征。在多模态任务中，ViLT（Kim、Son 和 Kim，2021年）在特征提取阶段加强文本与图像之间的交互，从而产生高效的多模态特征，

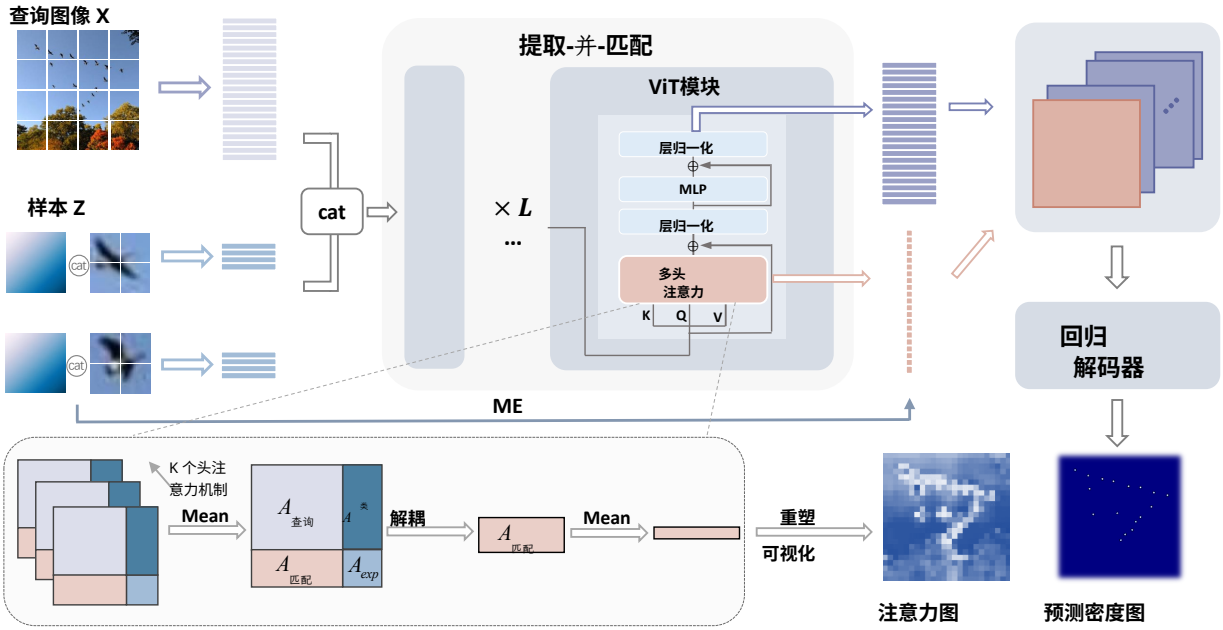


图 2：CAC 视觉变换器（CACViT）的框架。查询图像和带有尺度嵌入的样本被分割成图像块，形成令牌。随后，将展平后的令牌串联起来并输入变换器编码器。之后，将查询图像的输出特征与来自带有数量级嵌入（ME）的注意力图的相似度度量串联起来进行回归。最后，回归解码器预测密度图。值得注意的是，注意力图与密度图相似。

有助于提升下游任务的性能。据我们所知，我们是首个在类别无关计数（CAC）中同时考虑特征提取与匹配的研究，并对这一范式在 CAC 中的可行性进行解耦分析，从而简化类别无关计数任务的工作流程。

类无关计数视觉变换器

本节首先介绍我们用于类别无关计数（CAC）的框架概述，该框架名为类无关计数视觉变换器（CACViT）。随后，我们分析纯自注意力层为何能够替代以往的架构。在构建 ViT 风格模型后，我们发现直接应用 ViT 会导致尺度信息丢失，并针对尺度补偿提出了定制解决方案。

方法概述

CACViT 的流程如图 2 所示。查询图像 $\mathcal{X}$ 和对应的示例 $\mathcal{Z}$（调整大小至固定的示例尺寸）被划分为令牌。随后，将它们拼接起来，并通过自注意力层进行处理。之后，将查询图像的输出特征与最后一个注意力图中的相似度图进行拼接，形成最终的密度图。采用 ℓ_2 损失对密度图进行监督。

一种解耦的自注意力视角

我们遵循 CACViT 中自注意力的标准表述。然而，同时存在查询和样本令牌时，我们对类别无关计数中的自注意力有不同的理解。

我们首先重新审视变换器中的自注意力。给定一个输入令牌序列 $\mathcal{I}$，我们对其进行归一化，并通过线性层将其转换为 $\mathbf{Q}$、$\mathbf{K}$ 和 $\mathbf{V}$ 三元组。然后，我们采用缩放点积注意力机制，计算查询$\mathbf{Q}$ 和键 $\mathbf{K}$之间的注意力值。每个输出令牌都是所有令牌的加权和，并以注意力值作为权重，表示为：

$$\texttt{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \texttt{softmax}(\mathbf{Q}^\top\mathbf{K}/\sqrt{\texttt{D}})\mathbf{V}\,, \quad (1)$$

其中注意力图 $A = \text{softmax}(\mathbf{Q}^\top\mathbf{K}/\sqrt{\text{D}})$。

在同时存在查询词元和示例图像词元的情况下，我们有 $A \in \mathbb{R}^{(M+M_z)\times(M+M_z)}$，其中 M 和 M_Z 分别表示查询图像的词元数量和示例的词元数量。值得注意的是，借助自注意力块（Vaswani等人，2017年）中的跳跃连接（He等人，2016年），查询词元和示例图像词元的物理含义得以在各层中保留，这是我们的方法和分析的先决条件。

直接考虑 A 时，很难直观地理解自注意力机制如何同时充当特征提取器和匹配器。在这里，我们从解耦的视角重新审视自注意力在类别无关计数（CAC）中的作用。

首先，我们可以将令牌间注意力图 A 拆分为以下 4 个子注意力图：

$$A = \begin{bmatrix} A_{query} & A_{class} \\ A_{match} & A_{exp} \end{bmatrix}, \quad (2)$$

其中 $A_{query} \in \mathbb{R}^{M\times M}$、$A_{class} \in \mathbb{R}^{M\times M_Z}$、$A_{match} \in \mathbb{R}^{M_Z\times M}$，以及 $A_{exp} \in \mathbb{R}^{M_Z\times M_Z}$。

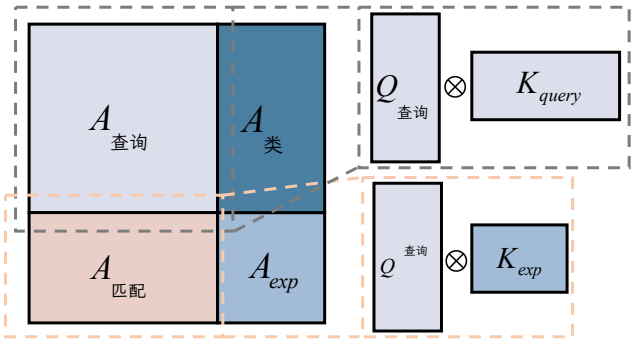


图3：CACViT 中自注意力的解耦视图。左上角的 A_{query} 可视为查询图像的自注意力。左下角的 A_{match} 可解释为查询图像与示例之间的交叉注意力，尽管其实现方式是自注意力。

如图3所示， A_{query} 可以看作查询图像的自注意力图，表示图像内部与其自身相关的最重要信息。A_{query} 对查询图像进行特征提取，其作用与先前工作中的视觉Transformer或CNN骨干网络相同。同样， A_{exp} 表示示例图像的自注意力图，其作用与先前工作中对示例图像进行特征提取相同。

A_{match} 图3所示的内容可视为查询图像与示例之间的交叉注意力，因此可以取代以往基于视觉Transformer的方法中使用的交叉注意力层，这些方法将查询图像作为查询，将示例作为键。这样， A_{match} 便可执行特征匹配。总之，与以往工作相比，CACViT中的自注意力以简单而自然的方式执行特征提取和匹配。

除了匹配和提取之外，我们还获得了额外的 A_{class}。需要注意的是， A_{class}的含义不同于 A_{match}。在 A_{class} 中，每一行 i 表示从查询词元 i 到示例图像词元的注意力。在浅层中，由于自注意力能力较弱，包含前景物体的查询词元可以通过额外的示例图像词元更快地获取相应的类别信息。相反，背景从示例图像词元获得的注意力较少。因此，在浅层中，A_{class}^{T} 类似于 A_{match}，主要关注前景物体。随着层数加深，自注意力能力逐渐增强。对于包含前景信息的查询词元，由于查询图像中存在重复物体，其自注意力会强于包含背景信息的查询词元。此外，由于查询图像对自身及示例的总注意力保持不变，包含背景信息的查询词元对示例的注意力会强于包含前景信息的查询词元。因此，转置后的 A_{class} 将关注背景区域。

解耦注意力的可视化结果如图 4 所示，这验证了我们的分析。更详细的分析
见补充材料。

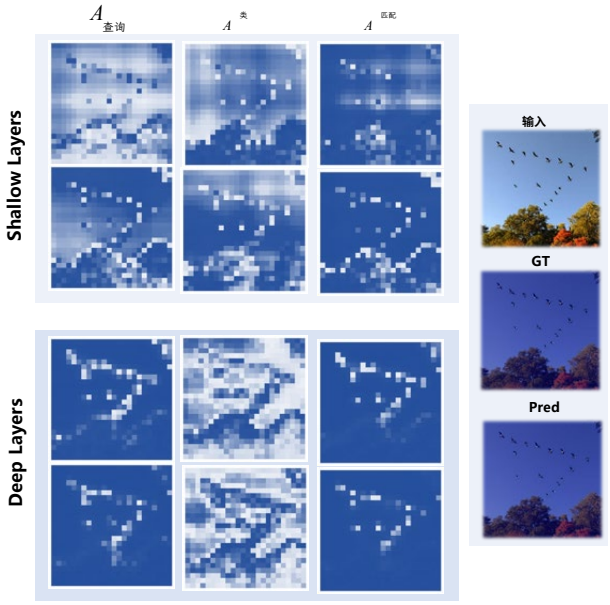


图4：不同层中解耦视图下注意力图的可视化结果。(a) A_{query} 和 A_{match} 突出前景并抑制背景。(b) 在浅层中，A_{class} 更关注前景；而在深层中，A_{class} 则突出背景。

尺度与幅度先验

尽管视觉Transformer中的自注意力机制适用于 CAC 任务，但该结构中的某些限制或功能可能导致信息损失。首先，必须调整示例的大小以适配固定大小的令牌。如图5所示，物体的尺度大小和宽高比差异很大。因此，调整大小后的示例会丢失尺度信息，并使匹配过程变得复杂。其次，用于归一化注意力图的 softmax 会削弱表达物体数量的能力。如图6所示，对于分布相同但计数值不同的查询图像，由于归一化，注意力图会非常相似，从而导致数量级信息损失。

为了弥补损失并使我们的 CACViT 更适用于 CAC 任务，我们引入了具有宽高比感知能力的尺度嵌入和数量级嵌入。

宽高比感知尺度嵌入。 虽然以往的研究已探讨了尺度层级信息（Shi 等人，2022），但宽高比信息却常常被忽视。在此，我们提出了一种进一步考虑宽高比的尺度嵌入方法。

嵌入过程如图7所示。具体而言，给定一个样本 z_k，其原始宽度和高度分别为 W_k 和 H_k。调整大小后，我们得到一个固定尺寸的样本 $z_k \in \mathbb{R}^{W_z \times H_z \times C}$，其中 W_z和 H_z 分别是每个样本的固定高度和宽度。首先，为了保留样本的宽度信息，我们创建一个在0和 W_k之间包含 W_k 个点的线性向量。然后，我们可以广播该向量，得到宽度图$\hat{W}_k \in \mathbb{R}^{W_z \times H_z}$。同理，我们可以得到高度



图5：具有不同尺度等级和宽高比等级的图像示例。第一行和第二行分别展示了 FSC147 数据集中尺度和宽高比的变化。

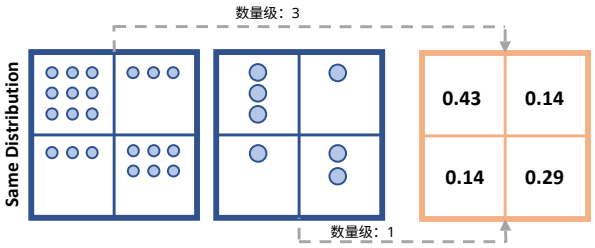


图6：幅值嵌入的目标。对于分布相同但计数值不同的查询图像，由于归一化的作用，其理想注意力图应当相同，这导致了数量级信息的损失。

映射 $\hat{}\, H_k \in \mathbb{R}^{W_z \times H_z}$。最后，我们将 $\hat{}\, W_k$ 和 $\hat{}\, H_k$ 相加，得到尺度嵌入 $S_k = W_k + H_k$。然后，我们可以将其与调整大小后的样本 z_k 拼接，以恢复尺度信息。需要注意的是，通过这种方式，我们还利用尺度嵌入为每个图像块编码了位置信息。

幅值嵌入。 数量级信息可以通过图像尺寸与样本尺寸的比值进行粗略表示。具体而言，我们假设一个样本的宽度为 W_k，高度为 H_k。考虑到图像块的尺寸为 $W_p \times H_p$ 像素，以样本 z_k 为条件的图像块的最大容量可以表示为：

$$ME_k = \frac{W_p \times H_p}{W_k \times H_k}. \tag{3}$$

如果我们有 K 个示例，就可以根据这些示例计算平均幅值嵌入 ME。然后，将该嵌入与注意力图中的相似度分数相乘，从而得到最终的相似度图。

实验

实现细节

为展示我们方法的优越性，我们在 FSC147（Ranjan 等人，2021）上进行了大量实验，该数据集是首个大规模类别无关计数数据集。与 Shi 等人，2022 类似，我们还采用平均绝对误差（平均绝对误差）和均方误差（MSE）作为评估指标。

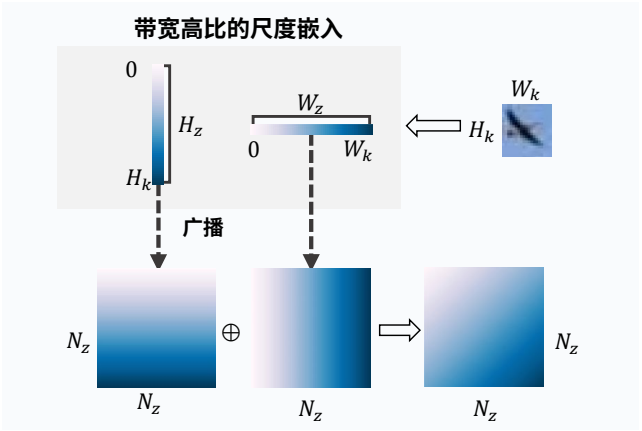


图7：说明如何为样本计算考虑纵横比的尺度嵌入。对于宽度为 W_k、高度为 H_k的样本，将其调整为固定宽度 W_z 和固定高度 H_z。

网络结构： 网络以大小为 384×384 的图像作为输入，首先将其划分为大小为 16×16 的图像块。每个样本的大小为 64×64，随后将其划分为大小为 16×16 的图像块。我们的特征提取器采用平均绝对误差（He 等，2022年）进行预训练，由 12 个变换器编码器块组成，隐藏维度为 768，且每个多头自注意力层包含 12 个注意力头。随后采用另外 3 个隐藏维度为 512 的变换器块，以增强特征并降低维度，从而进行上采样。我们的回归解码器由 4 个隐藏维度为 256 的上采样层组成，与 CounTR（Liu 等人，2022年）中的设置相同。

训练细节： 为确保公平比较，我们采用与 CounTR（Liu 等人，2022）相同的数据增强、测试时裁剪和归一化方法。我们使用 AdamW（Loshchilov 和 Hutter，2017）作为优化器，批量大小为 8。模型训练 200 个周期，学习率为 $1e-4$，权重衰减率为 0.05，并进行 10 个周期的预热。我们的模型在 NVIDIA GeForce RTX 3090 上训练和测试。请注意，CACViT 在单个 GPU 上训练 12 小时约需 14GB 内存。

与最先进方法的比较

定量结果。 本节将我们的 CACViT 与 FSC-147 数据集上的最先进方法进行比较。如表1所示，在 1-shot 和 3-shot 设置下，CACViT 的表现均显著优于所有对比方法。在 1-shot 设置下，与 CounTR 相比，CACViT 的验证平均绝对误差相对降低了 13.23%，测试平均绝对误差相对降低了 28.52%。此外，CACViT 将验证均方根误差降低了 17.45%，将测试均方根误差降低了 66.76%。在 3-shot 设置下，与 CounTR 相比，CACViT 的验证平均绝对误差相对降低了 19.04%，测试平均绝对误差相对降低了 23.60%。还可以看到，与 CounTR 相比，CACViT 将验证均方根误差降低了 23.84%，将测试均方根误差降低了 46.33%，这验证了我们方法的鲁棒性。

模型	骨干网络	分辨率	样本数	验证集平均绝对误差	验证集均方根误差	测试平均绝对误差	测试均方根误差
FamNet+ (Ranjan 等人，2021)	ResNet50	384	1	26.55	77.01	26.76	110.95
BMNet+ (Shi 等人，2022)	ResNet50	[384,1584]	1	17.89	61.12	16.89	96.65
CounTR（Liu 等人，2022）	ViT-B 和卷积神经网络	384	1	13.15	49.72	12.06	90.01
CACViT（我们的方法）	ViT-B	384	1	**11.41**	**41.04**	**8.62**	**29.92**
FamNet+ （Ranjan 等人，2021）	ResNet-50	384	3	23.75	69.07	22.08	99.54
RCAC（Gong 等人，2022年）	ResNet-50	384	3	20.54	60.78	20.21	81.86
BMNet+ （Shi 等人，2022年）	ResNet-50	[384,1584]	3	15.74	58.53	14.62	91.83
SAFECount（You 等人，2023年）	ResNet-18	512	3	15.28	47.20	14.32	85.54
SPDCN（Lin 等，2022年）	ResNet-18	576	3	14.59	49.97	13.51	96.80
CounTR（Liu 等人，2022年）	ViT-B 和 卷积神经网络	384	3	13.13	49.83	11.95	91.23
CACViT（我们的方法）	ViT-B	384	3	**10.63**	**37.95**	**9.13**	**48.96**

表1：在 FSC-147 数据集上与最先进的类别无关计数方法进行比较。表格上半部分展示了 1-shot 设置下的结果，下半部分展示了 3-shots 设置下的结果。CounTR（Liu 等人，2022）需要采用两阶段训练方案（He 等人，2022）。请注意，我们的 1-shot CACViT 即使在 3-shots 设置下，也优于所有先前方法。

方法	低 MAE	低 RMSE	高 MAE	高 RMSE
BMNet+	6.23	29.96	23.21	127.62
SPDCN	5.06	31.53	22.25	133.59
CounTR	5.44	36.96	18.56	124.11
CACViT	**3.14**	**6.26**	**15.33**	**69.54**

表2：与最先进的类别无关计数方法在低密度和高密度两个子测试集上的比较。两个子测试集的目标数量范围分别为8–37和37–3701，且每个子集约包含600张图像。

值得注意的是，1-shot 设置在测试集上的表现甚至优于 3-shot 设置，这与我们的直觉相悖。这一现象源于密集环境中标注质量的影响，更多细节请参阅补充材料。

不同密度下的性能。为进一步验证模型在3-shot设置下对不同目标密度的适应性能，我们将测试集划分为低密度和高密度两个子测试集，其目标数量范围分别为8–37和37–3701，且每个子集约包含600张图像。如表2所示，我们的方法在低密度和高密度环境下均优于其他方法，这验证了我们方法的鲁棒性。

运行成本评估。为验证我们方法的效率，我们将模型大小和浮点运算次数（FLOPS）与 CounTR（Liu 等人，2022）进行比较，后者采用了与 CACViT 相同的骨干网络。结果见表 3。请注意，CounTR 的论文并未报告微调所用的训练轮数，因此我们从其官方 GitHub 页面获取了相关数据。我们的训练设置与 CounTR 类似。不过，1）与 CounTR 相比，我们没有在 FSC147 数据集上进行自监督预训练（He 等，2022）；2）我们不需要额外的卷积神经网络来提取示例的特征。显而易见，CACViT 能够大幅减少训练轮数

方法	模型大小（M）	GFLOPs	训练轮数
CounTR 预训练	112M	27G	300
CounTR 微调	99M	84G	1000*
CACViT（我们的方法）	99M	89G	200

表3：模型大小和 FLOPs 的比较。* 表示我们从其官方项目页面获取了 #epochs，该信息未在论文中提及。

从 1300 轮降至 200 轮，并将训练策略简化为单阶段，同时保持模型大小和 FLOPS 与 CounTR 相当。

定性分析。为了进行定性分析，我们将 CACViT 与最先进的基于卷积神经网络的模型 BMNet+以及最先进的基于视觉Transformer的模型 CounTR 进行了比较。如图 8 所示，在不同场景中，包括包含多个类别的图像、密集环境和色彩丰富的环境，CACViT 的表现均优于其他方法。对于受到其他类别干扰的困难样例（第 1 列和第 2 列），我们的方法能够更好地学习待计数物体的特征。

消融研究

我们在 FSC147 上进行了消融研究，并在表4中给出了定量结果。我们采用 CACViT 的简化版本作为基线，该版本仅由基于自注意力的标准特征提取和匹配模块组成，并对注意力图中的 A_{class} 部分进行屏蔽（B1 和 B6）。

额外类别信息 A_{class}。通过比较表4中的 B6 和 B7，可以观察到，背景抑制使测试集上的平均绝对误差相对降低了 15.44%，均方根误差相对降低了 23.12%。性能提升表明，示例能够为查询特征提供有效信息。

宽高比感知尺度嵌入。如表 4 所示，宽高比感知尺度嵌入在

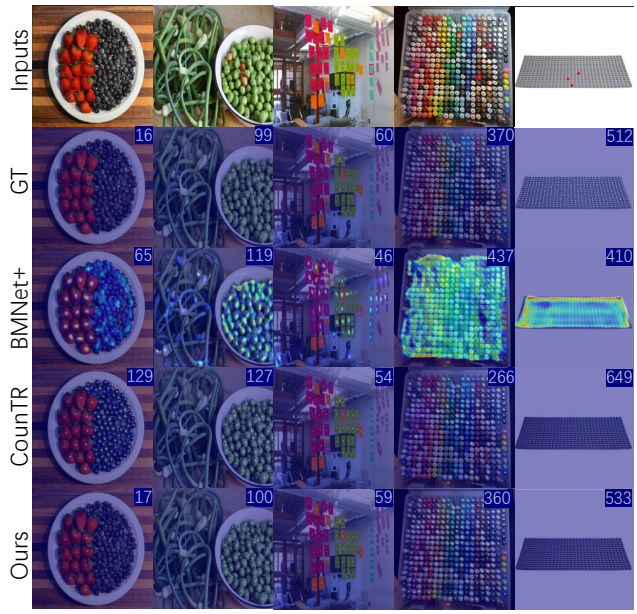


图8：FSC147 数据集上的定性结果。所选输入展示了不同的挑战，包括包含多个类、密集环境和彩色环境的图像。我们的方法始终优于先前方法，定位更加精准。

No.	CLS	SE	ME	验证集平均绝对误差	验证集均方根误差
B1				13.00	48.37
B2	✓			12.40	48.62
B3	✓	✓		10.61	40.42
B4	✓		✓	12.32	42.42
B5	✓	✓	✓	**10.62**	**37.95**
No.	CLS	SE	ME	测试平均绝对误差	测试均方根误差
B6				13.67	117.60
B7	✓			11.56	90.41
B8	✓	✓		10.89	78.03
B9	✓		✓	11.61	83.06
B10	✓	✓	✓	**9.13**	**48.96**

表4：FSC147 数据集上的消融研究。CLS 表示与先前方法相比，自注意力中的额外组件 A_{class}；SE 表示尺度嵌入，ME 表示数量级嵌入。

方法	微调后	MAE	RMSE
FamNet+	×	28.84	44.47
RCAC	×	17.98	24.21
SAFECount	×	16.66	24.08
BMNet+	×	10.44	13.77
CACViT（我们的方法）	×	**8.30**	**11.18**
FamNet+	✓	18.19	33.66
RCAC	✓	13.62	19.08
SAFECount	✓	5.33	7.04
BMNet+	✓	5.76	7.83
CounTR	✓	5.75	7.45
CACViT（我们的方法）	✓	**4.91**	**6.49**

表 5：在 CARPK 数据集上的泛化性能。所有模型均在 FSC147 数据集上进行预训练，且不包含“汽车”类。“微调”表示是否在 CARPK 数据集上对预训练模型进行进一步微调。

验证集上使平均绝对误差相对降低 14.44%，使均方根误差相对降低 16.87%；在测试集上使平均绝对误差相对降低 5.80%，使均方根误差相对降低 13.69%（比较 B2 与 B3 以及 B7与 B8）。这验证了宽高比感知尺度嵌入能够恢复示例的原始尺寸信息，因此对 CAC 任务至关重要。请注意，均方根误差上的表现进一步表明，SE 模块能够提升模型的鲁棒性。此外，我们在补充材料中与其他尺度嵌入方法进行了比较。

幅值嵌入。 B2 与 B4 以及 B7 与 B9 的比较表明，幅值嵌入使验证均方根误差提升了 6.2，测试均方根误差提升了 7.35。然而，平均绝对误差的性能提升并不明显，甚至略有下降。原因可能在于，调整大小后的样本提供了错误的物体尺寸先验，从而使模型感到困惑。

尺度嵌入与幅值嵌入的协同作用。 比较表4中的 B8 和 B10 可以发现，幅值嵌入使测试集上的平均绝对误差相对降低了 21.36%，均方根误差相对降低了 41.05%。值得注意的是，与不使用尺度嵌入的情况相比，性能提升显著得多，这表明尺度嵌入与幅值嵌入相结合能够进一步提升性能。尺度嵌入与幅值嵌入之间的协同作用可以分别为输入和输出提供尺度先验，因此 CACViT 能够始终捕捉尺度先验。

跨数据集泛化

我们在汽车计数数据集 CARPK（Hsieh、Lin 和 Hsu，2017年）上测试了模型的泛化能力。为了确保训练类别与测试类别不存在重叠，我们排除了 FSC147 中的“汽车”类别。结果如表 5 所示。可以观察到，我们的模型表现出较强的泛化能力。与 BMNet+（Shi 等人，2022年）相比，CACViT 的平均绝对误差（MAE）相对性能提升了 20.50%。

结论

在本工作中，我们提出了一种简单而高效的、基于视觉 Transformer的模型 CACViT，用于类别无关计数。具体而言，我们从解耦视角表明，视觉Transformer天然适用于类别无关计数任务。我们还针对类别无关计数提出了一种基于视觉Transformer的提取与匹配范式。随后，我们引入了具备宽高比感知能力的尺度嵌入和数量级嵌入，以弥补信息损失。我们的 CACViT 在 FSC147 上取得了当前最优的结果，并且我们还在 CARPK 上验证了其通用性。

参考文献

Abousamra、S.；Hoai、M.；Samaras、D.；以及 Chen、C. 2021 年。在拓扑约束下进行人群定位。载于*AAAI* 人工智能会议论文集，第 35 卷，第 872–881 页。

Cao, X.；Wang, Z.；Zhao, Y.；和 Su, F. 2018年。用于实现精准高效人群计数的尺度聚合网络。载于欧洲计算机视觉会议论文集（*ECCV*），734–750。Chen, Q.；Wu, Q.；Wang, J.；Hu, Q.；Hu, T.；Ding, E.；Cheng, J.；和 Wang, J. 2022年。MixFormer：跨窗口和维度混合特征。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），5249–5259。Cheng, Z.-Q.；Dai, Q.；Li, H.；Song, J.；Wu, X.；和 Hauptmann, A. G. 2022年。重新思考卷积网络在目标计数中的空间不变性。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，19638–19648。Gong, S.；Zhang, S.；Yang, J.；Dai, D.；和 Schiele, B. 2022年。对类内多样性具有鲁棒性的类别无关目标计数。载于计算机视觉—*ECCV 2022*：第*17*届欧洲计算机视觉会议，以色列特拉维夫，*2022*年*10*月*23*日至*27*日，论文集，第*XXXIII*部分，388–403。Springer。Han, G.；Ma, J.；Huang, S.；Chen, L.；和 Chang, S.-F. 2022年。基于 Fully Cross-Transformer 的少样本目标检测。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），5321–5330。He, K.；Chen, X.；Xie, S.；Li, Y.；Doll´ar, P.；和 Girshick, R. 2022年。掩码自编码器是可扩展的视觉学习器。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，16000–16009。He, K.；Zhang, X.；Ren, S.；和 Sun, J. 2016年。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，770–778。He, Y.；Ma, Z.；Wei, X.；Hong, X.；Ke, W.；和 Gong, Y. 2021年。用于无监督跨域人群计数的误差感知密度同构重建。载于*AAAI*人工智能会议论文集，第35卷，1540–1548。Hsieh, M.-R.；Lin, Y.-L.；和 Hsu, W. H. 2017年。基于空间正则化区域提议网络的无人机目标计数。载于*IEEE*国际计算机视觉会议论文集，4145–4153。Idrees, H.；Tayyab, M.；Athrey, K.；Zhang, D.；Al-Maadeed, S.；Rajpoot, N.；和 Shah, M. 2018年。用于密集人群计数、密度图估计和定位的组合损失。载于欧洲计算机视觉会议论文集（*ECCV*），532–546。Kim, W.；Son, B.；和 Kim, I. 2021年。ViLT：无需卷积或区域监督的视觉语言变换器。载于 Meila, M.；和 Zhang, T.，编，第*38*届国际机器学习会议论文集，第139卷，机器学习研究论文集，5583–5594。PMLR。

Laradji, I. H.; Rostamzadeh, N.; Pinheiro, P. O.; Vazquez, D.; and Schmidt, M. 2018. 斑点在哪里：通过定位和点监督进行计数。载于欧洲计算机视觉会议（*ECCV*）论文集，547–562。Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022. 探索用于目标检测的纯视觉Transformer骨干网络。arXiv:2203.16527。Lin, W.; Yang, K.; Ma, X.; Gao, J.; Liu, L.; Liu, S.; Hou, J.; Yi, S.; and Chan, A. B. 2022. 用于样例引导的类别无关计数的尺度先验可变形卷积。Lin, Y.; Yuan, Y.; Zhang, Z.; Li, C.; Zheng, N.; and Hu, H. 2023. DETR不需要多尺度或局部性设计。arXiv:2308.01904。Liu, C.; Zhong, Y.; Zisserman, A.; and Xie, W. 2022. CounTR：基于Transformer的通用视觉计数。*arXiv*预印本 *arXiv:2208.13721*。Loshchilov, I.; and Hutter, F. 2017. 解耦权重衰减正则化。*arXiv*预印本 *arXiv:1711.05101*。Lu, E.; Xie, W.; and Zisserman, A. 2019. 类别无关计数。载于计算机视觉—*ACCV 2018*：第*14*届亚洲计算机视觉会议，澳大利亚珀斯，*2018*年*12*月*2*日至*6*日，修订精选论文，第*III*部分 *14*，669–684。Springer。Lu, H.; Cao, Z.; Xiao, Y.; Zhuang, B.; and Shen, C. 2017. TasselNet：通过局部计数回归网络对自然场景中的玉米雄穗进行计数。《植物方法》，13(1)：1–17。Madec, S.; Jin, X.; Lu, H.; De Solan, B.; Liu, S.; Duyme, F.; Heritier, E.; and Baret, F. 2019. 使用深度学习技术从高分辨率RGB图像中估计穗密度。《农业与森林气象学》，264：225–234。Onoro-Rubio, D.; and L´opez-Sastre, R. J. 2016. 利用深度学习实现无透视畸变目标计数。载于计算机视觉—*ECCV 2016*：第*14*届欧洲计算机视觉会议，荷兰阿姆斯特丹，*2016*年*10*月*11*日至*14*日，论文集，第*VII*部分 *14*，615–629。Springer。Ranjan, V.; Sharma, U.; Nguyen, T.; and Hoai, M. 2021. 学习计数万物。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，3394–3403。Shi, M.; Lu, H.; Feng, C.; Liu, C.; and Cao, Z. 2022. 表征、比较与学习：一种用于类别无关计数的相似性感知框架。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，9529–9538。Shu, W.; Wan, J.; Tan, K. C.; Kwong, S.; and Chan, A. B. 2022. 频域中的人群计数。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，19618–19627。Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucu- rull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet,

X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poul- ton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Ro- driguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2：开放基础模型与微调聊天模型。arXiv:2307.09288。Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. 注意力就是你所需要的一切。神经信息处理系统进展，30。Wang, Q.; Gao, J.; Lin, W.; and Li, X. 2020. NWPU-crowd：用于人群计数与定位的大规模基准数据集。*IEEE*模式分析与机器智能汇刊，43(6)：2141–2149。Xu, Y.; Zhang, J.; ZHANG, Q.; and Tao, D. 2022. ViTPose：用于人体姿态估计的简易视觉Transformer基线。载于 Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., 编，神经信息处理系统进展，第35卷，38571–38584。Curran Associates, Inc. Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2023. ViTPose++：用于通用人体姿态估计的视觉Transformer基础模型。arXiv:2212.04246。Yang, S.-D.; Su, H.-T.; Hsu, W. H.; and Chen, W.-C. 2021. 类别无关的小样本目标计数。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，870–878。Yao, J.; Wang, X.; Yang, S.; and Wang, B. 2023. ViTMatte：利用预训练的纯视觉Transformer提升图像抠图效果。arXiv:2305.15272。You, Z.; Yang, K.; Luo, W.; Lu, X.; Cui, L.; and Le, X. 2023. 利用相似性感知特征增强进行小样本目标计数。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，6315–6324。Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; and Yan, S. 2022. MetaFormer实际上就是你所需要的视觉模型。arXiv:2111.11418。Zhang, C.; Li, H.; Wang, X.; and Yang, X. 2015. 通过深度卷积神经网络进行跨场景人群计数。载于*IEEE*计算机视觉与模式识别会议论文集，833–841。Zou, Z.; Qu, X.; Zhou, P.; Xu, S.; Ye, X.; Wu, W.; and Ye, J. 2021. 从粗到细：通过对抗评分网络实现领域自适应人群计数。载于第*29*届*ACM*国际多媒体会议论文集，2185–2194。