SERE: 探索自监督 Transformer 的特征自关系

Zhong-Yu Li, Shanghua Gao, Ming-Ming Cheng

摘要—通过自监督表征学习训练卷积神经网络（CNN）已经被证明对于视觉任务是有效的。作为 CNN 的替代方案，vision transformer（ViT）利用空间自注意力和通道级前馈网络，拥有了强大的表征能力。最近的研究表明了自监督学习有助于释放 ViT 的巨大潜力。尽管如此，大多数工作都遵循为 CNN 设计的自监督学习策略（如样本的实例级判别），却忽略了 ViT 的特性。我们观察到，空间和通道维度的关系建模使 ViT 区别于其他网络。为了加强这一特性，我们通过特征自关系（SElf-RElation（SERE））建模来训练自监督 ViT 模型。具体来说，我们并不是仅仅使用图像不同视角的特征嵌入进行自监督学习，而是利用特征自关系，即空间/通道维度的自关系进行自监督学习。基于自关系的学习进一步强化了 ViT 的关系建模能力，使其产生更强的表征，进而稳定提升模型在多个下游任务中的性能。我们的源代码可以在以下链接获得：https://github.com/MCG-NKU/SERE.

Index Terms—特征自关系，自监督学习，视觉 Transofrmer

1 引言

通过有监督的方式训练神经网络已经在许多视觉任务上取得了巨大的成功，但收集人工标注产生了巨大的成本 [1], [2], [3]。而从无标注的图像中学习视觉表征 [4], [5], [6], [7], [8] 已经被证明是有监督训练的有效替代方案，例如通过自监督的方式预训练的卷积神经网络（CNN）已经实现了与有监督 CNN 相当甚至更好的性能 [9], [10]。近年来，vision transformers (ViT) [11], [12] 在许多视觉任务上展现了比 CNN 更强的表征能力。早期的工作已经将为 CNN 设计的自监督学习方法应用在了 ViT 中，并揭示了自监督 ViT [13], [14], [15] 的巨大潜力。为 ViT 设计的典型的自监督学习方法（如 DINO [13] 和 MoCoV3 [15]）将图像的不同视角输入到 ViT 网络，以生成特征表示。接着自监督学习，如对比学习 [15], [16], [17] 和自聚类 [13], [18]，通过以下假设训练 ViT 网络：图像的不同视角有相似的表征。然而，目前广泛使用的特征表示仍局限于为 CNN 设计自监督方法所使用的特征嵌入，例如图像特征级别的嵌入 [6], [7], [19] 和像素级别的特征嵌入 [20], [21]。现有的方法却很少关注 ViT 的特性，例如自关系建模能力。我们想知道与 ViT 相关的其他形式的特征表示是否能够有益于自监督 ViT 的训练。

我们旨在利用 ViT 的特性来改进自监督 ViT 的训练。ViT 分别使用多头自注意力（MHSA）和前馈网络（FFN）在空间和通道维度上建模特征关系 [11], [22], [23]。MHSA 通过提取图像块之间的关系来聚合空间信息，增强了语义相似的图像块之间的空间关系（见图 1(c)）。FFN 将来自不同通道的特征结合在一起，隐式地在通道维度上建模特征的自关系。例如，图 1(c) 表明 ViT 的通道学习了多样的视觉模式，并且不同通道之间存在不同程度的关系。特征自关系建模使 ViT 具有强大的表征能力。受此启发，我们使用特征自关系作为自监督学习中新的特征表示。

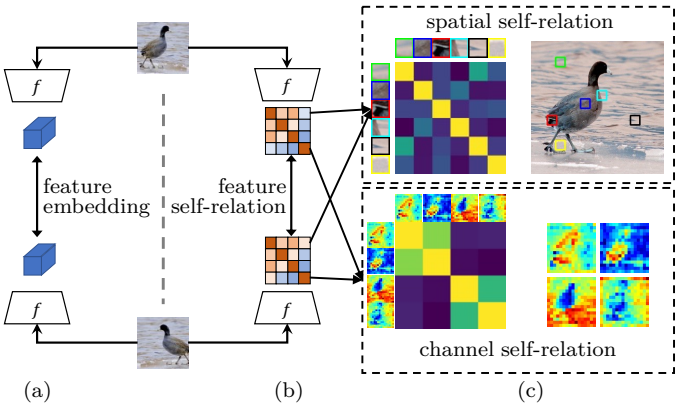


图 1. 使用特征嵌入和我们提出的特征自关系的自监督学习示意图。(a) 典型的自监督学习方法处理图像视图的特征嵌入。(b) 我们提出建模特征自关系，度量来自一个视图内不同维度上的关系。(c) 两种具体形式的自关系，即空间自关系和通道自关系。对于空间自关系，我们选择了 6 个不同颜色的方框表示图像块（右上），并可视化它们的自关系（左上）。对于通道自关系，我们展示了 4 个通道的可视化特征图（右下）以及相应的自关系（左下）。

在这项工作中，我们提出利用特征自关系（SElf-RElation，SERE）来进行自监督训练，增强 ViT 的自关系建模能力。考虑到 MHSA 和 FFN 建模了空间关系和通道关系，我们提出使用空间自关系和通道自关系作为特征表示。空间自关系提取了图像中不同图像块之间的关系。通道自关系建模了不同通道之间的联系，其中每个通道编码了独特的语义信息。特征自关系是一种新的特征表示，且与现有的特征表示（如图像级别和图像块级别的特征嵌入）兼容。如图 1所示，我们可以轻松地将现有的自监督学习方法中的特征嵌入替换为我们所提出的特征自关系。我们说明了利用特征自关系可以稳

• The authors are with TMCC, CS, Nankai University, Tianjin 300350, China. S. Gao is the corresponding author (shgao@mail.nankai.edu.cn).

定地提升多个用于 ViT 的自监督学习方法，例如 DINO [13]、iBOT [18]、MoCoV3 [15]，并且提升模型在各种下游任务上的性能，例如目标检测 [2], [24]、语义分割 [3], [25]、半监督语义分割 [26]、图像分类 [1]。

据我们所知，我们是第一个在自监督学习中研究自关系的工作。我们的主要贡献总结如下：

- 我们提出利用 ViT 的自关系（SERE），即与 ViT 的关系建模属性相适应的空间自关系和通道自关系，作为自监督学习的特征表示。
- 我们提出的 SERE 方法可以兼容现有的自监督方法，并稳定提升了 ViT 在各种下游任务上的性能。

2 相关工作

2.1 自监督学习

自监督学习在不需要任何人工标注的情况下学习丰富的表征。早期的工作利用手工设计的代理任务，例如图片上色 [27], [28]、拼图 [29]、旋转预测 [30]、自编码器 [31], [32]、图像修复 [33] 、计数 [34]。这些方法基于启发性线索学习图像的表征，但只能达到有限的泛化能力。最近，自监督学习因采用了新形式的自监督方法而取得了重大突破，例如对比学习 [7], [35], [36], [37], [38], [39]、自聚类 [40], [41], [42] 、表征对齐 [5], [6], [43], [44], [45], [46]。这些方法直接利用特征嵌入作为特征表示来生成自监督信号。例如，其中许多方法使用图像级别的特征嵌入作为特征表示。而一些方法则探索使用更细粒度维度上的特征嵌入，例如像素 [20], [47]、图像块 [48], [49]、物体 [21] 和区域 [21], [50] 维度。然而，这些表示仍然是图像不同区域的特征嵌入。与仅对特征嵌入施加约束的方法相比，我们进一步将特征嵌入转换为自关系，作为新的表示维度，从而增加了对特征嵌入之间关系的约束。自关系为自监督训练提供了丰富的信息，并与 ViT 的关系建模属性相适配，进一步提高了 ViT 的表示质量。同时，自关系与基于嵌入的方法相兼容，可以显著提升多种自监督学习方法的性能。

2.2 自监督 vision Transformer

Transformer 模型已经推广到计算机视觉领域 [11], [51]，并在许多任务上取得了最先进的性能，例如图像分类 [12]、语义分割 [51], [52]、目标检测 [53]。由于缺乏归纳偏置，训练 ViT 需要更多的数据和技巧 [11], [54]。最近的工作致力于使用自监督学习方法 [16], [55], [56], [57] 来训练 ViT，以满足 ViT 对数据的需求，同时保持低标注成本。其中许多基于实例判别的方法使用特征嵌入作为自监督学习的特征表示。例如，Chen 等人 [15] 和 Caron 等人 [13] 分别使用图像级别的嵌入来实施对比学习和自聚类。Zhou 等人 [18] 则使用图像块级别的特征嵌入进行自蒸馏。然而，这些方法仍然遵循最初为 CNN 设计的基于实例鉴别的代理任务，该代理任务通过最大化正样本之间的相似性来学习具有变换不变性的表示。ViT 中的新属性可能有助于自监督训练，然而这些方法忽略了这一性质。在本文中，我们探索了空间自关系和通道自关系，它们更适合对 ViT 的训练。

2.3 掩码图像建模

与我们的工作同期的基于掩码图像建模（MIM）[14], [33], [58], [59] 的自监督学习方法已经成为自监督 ViT 中实例判别（ID）的一种主流替代方法。MIM 根据未被遮挡的部分重建被遮挡的图像块。MIM 具有不同形式的重建目标，例如离散标记器 [58], [60]、原始像素 [14], [57], [61], [62], [63]、HOG 特征 [64]、图像块表示 [18] 等。与 ID 相比，MIM 中的图像块级重建增强了细粒度表征 [18], [59]。而我们与 MIM 的不同之处在于，我们提出的 SERE 显示地增强了模型建模空间关系的能力。实验还表明，SERE 可以超越并补充各种基于 MIM 的方法。此外，我们增强了模型建模通道之间关系的能力，而这是 MIM 所缺少的。

2.4 vision Transformer 的性质

最近的研究表明，ViT 在许多视觉任务 [12], [52], [65] 上取得的显著成功依赖于其强大的建模空间关系的能力。Dosovitskiy 等人 [11] 和 Kim 等人 [23] 发现注意力机制可以关注到图像中与语义相关的区域。Raghu 等人 [22] 揭示了 ViT 深层所编码的表征仍会保留很强的空间信息。他们还观察到在 ViT 中，图像块和语义相似的区域之间存在很强的相关性。Caron 等人 [13] 发现自监督 ViT 捕捉到比监督 ViT 更明确的语义区域。这些观察表明 ViT 具有强大的建模关系的能力，这与 CNN 的模式匹配机制有很大不同。在这项工作中，我们显式地将空间和通道特征自关系用于自监督学习，以此增强这种能力。

2.5 关系建模

关系建模具有不同形式，如成对关系和注意力等。这些方式已经促进了各种视觉任务的发展，例如知识蒸馏 [66], [67], [68], [69], [70], [71]、度量学习 [72]、语义分割 [73], [74], [75]、无监督语义分割 [76]、目标定位 [77], [78], [79]、对比学习 [80]、遮挡图像建模 [81]、特征聚合 [82] 、纹理描述符 [83], [84]。在自监督学习中，早期的工作 [85] 提出通过计算整个 batch 中的通道关系，即 batch 关系，来利用关系建模。与之相比，我们探索了自关系，即一张图像内部特征的空间或通道关系，这与 ViT 的关系建模属性相适配。

3 方法

3.1 概述

在这项工作中，我们关注于基于实例判别的自监督学习范式 [4], [13]。首先，我们简要回顾常见的基于实例判别的自监督学习的框架。给定一张未标注的图像 x，通过不同的随机数据

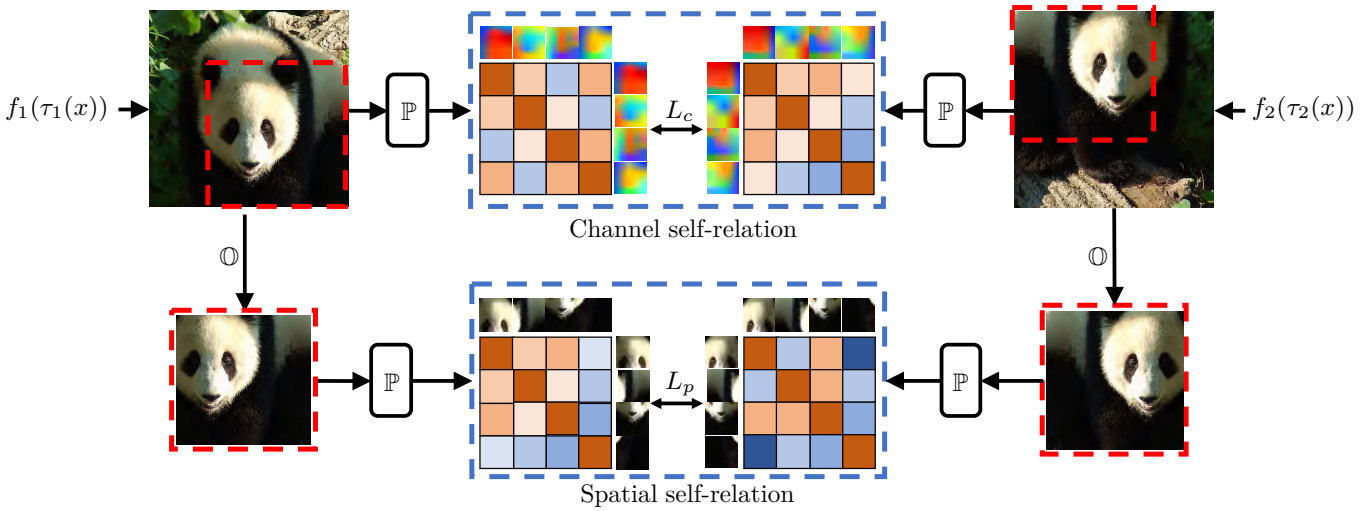


图 2. 我们的方法从空间和通道维度建模自关系。给定一张图像 x，通过两种随机的数据增强生成两个视图。这里的图像块代表了由编码器提取的特征嵌入。特征嵌入经过变换 $\mathbb{P}$ 转化为空间或通道自关系。L_p 和 L_c，即式 (3) 和式 (5) 中定义的损失函数，使不同视图的自关系之间保持一致。空间自关系只考虑重叠区域中的特征。$\mathbb{O}$ 表示式 (2) 中在两个视图之间的重叠区域提取特征的操作，其中红色虚线框表示重叠区域。

增强生成多个视图，例如使用增强操作 τ_1 和 τ_2 生成两个视图 $\tau_1(x)$ 和 $\tau_2(x)$。在假设图像的不同视图包含相似信息的前提下，大多数实例鉴别方法的主要思想是最大化不同视图编码的共享信息。首先，将两个视图送入编码器网络以提取特征嵌入 $r_1 \in \mathbb{R}^{C\times HW}$ 和 $r_2 \in \mathbb{R}^{C\times HW}$，其中 $H \cdot W$ 表示图像块数量，C 表示通道数量。根据不同自监督学习方法的训练目标，使用 $\mathbb{P}$ 来转换特征嵌入，获得不同的特征表示，例如图像级别和图像块级别的嵌入。不同的自监督优化目标利用特征表示来计算损失，如下所示：

$$L_I = \mathrm{R}(\mathbb{P}(r_1), \mathbb{P}(r_2)), \tag{1}$$

其中 R 表示视图直接的一致性，并可以定义为多种形式，例如对比损失 [7]、非对比损失 [6] 和聚类损失 [4]。

在这项工作中，我们的主要关注点是探索 $\mathbb{P}$ 的新形式。受 ViT 中的关系建模属性启发，我们不直接使用特征嵌入，而是在多个维度上利用特征自关系作为 ViT 自监督学习的特征表示。在接下来的部分，我们将介绍两种特定的自关系，即空间自关系和通道自关系。

3.2 Spatial Self-relation

以往的研究 [11], [13], [22], [23] 已经观察到 ViT 通过 MHSA 模块可以建模局部图像块之间的关系。与此同时，对于许多密集预测任务 [20], [21]，例如目标检测和语义分割，建模更准确的空间关系至关重要。因此，我们使用空间自关系进行自监督学习，以此来增强 ViT 的空间关系建模能力。在接下来的部分中，我们首先详细介绍了 $\mathbb{P}$ 的细节，该变换将 ViT 编码的特征嵌入转化为空间自关系。然后，我们以空间自关系作为特征表示，提供了相应的自监督损失。

计算空间自关系 给定 ViT 骨干网络生成的特征嵌入 $r_1 = f_1(\tau_1(x)) \in \mathbb{R}^{C\times HW}$ 和 $r_2 = f_2(\tau_2(x)) \in \mathbb{R}^{C\times HW}$。一个由

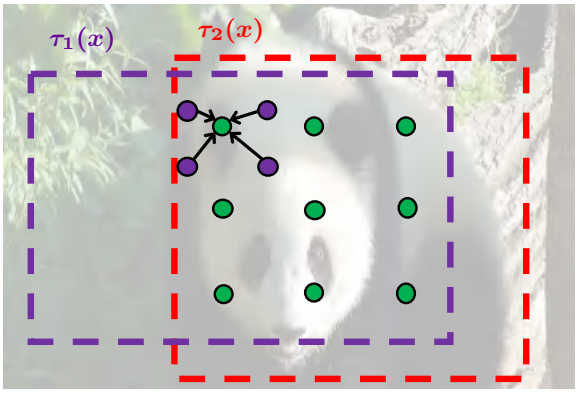


图 3. 空间自关系的区域对齐采样操作。$\tau_1(x)$ 和 $\tau_2(x)$ 是图像的不同视图，虚线框表示它们在原始图像中的区域。绿色的点表示在重叠区域中均匀采样的点。紫色的点表示 ViT 中的图像块特征。

Batch 归一化层 [86] 和 ReLU 激活层 [87] 组成的投影头 h_p 进一步处理这些特征并获取 $p_1 = h_p(r_1)$ 和 $p_2 = h_p(r_2)$。然后，我们分别计算它们的空间自关系。

与图像级别嵌入不同，不同视图的空间自关系之间的监督应该在处于同一空间位置的图像块之间计算。然而，由于数据增强中的随机裁剪和翻转，p_1 和 p_2 在空间维度上不对齐。为了解决不对齐问题，我们使用了一个区域对齐的采样操作 $\mathbb{O}$ [26]，该操作从 p_1 和 p_2 的重叠区域中均匀采样 $H_s \times W_s$ 个点。[1] 如图3所示，我们定位原始图像中的重叠区域，并将该区域分割成 $H_s \times W_s$ 个网格，这些网格与 ViT 中的图像块直接不需要对齐的。对于每个网格的中心，我们计算它在每个视图的特征图中的空间坐标并通过双线性插值采样其在不同视图中的特征。操作 $\mathbb{O}$ 的详细信息在补充材料中展示。

1. 在这项工作中，由于自关系的正交性，我们将提出的空间自关系与现有方法相结合。由于现有方法不限制不同视图必须重叠，因此我们只将空间自关系损失函数应用到具有重叠区域的视图中。

对于一个视图，例如 $p_1 \in \mathbb{R}^{C\times HW}$，我们计算空间自关系 $\mathbb{A}_p(p_1) \in \mathbb{R}^{H_sW_s\times H_sW_s}$ 如下：

$$\mathbb{A}_p(p_1) = \text{Softmax}\left(\frac{\mathbb{O}(p_1)^T \cdot \mathbb{O}(p_1)}{\sqrt{C}} \Big/ t_p\right), \tag{2}$$

其中，$\mathbb{O}(p_1) \in R^{C\times H_sW_s}$ 是在重叠区域中采样的特征，T 表示矩阵转置操作，t_p 是控制 Softmax 函数的锐度的温度参数。在空间自关系矩阵中，每一行表示一个局部图像块与其他图像块的关系，通过 Softmax 函数进行归一化以生成概率分布。

基于空间自关系的自监督训练。空间自关系可以用于许多形式的自监督训练表示。为简单起见，我们以非对称非对比的自监督训练损失 [5], [6] 为例。具体损失函数如下所示：

$$L_p = \text{R}_e(\not{G}(\mathbb{A}_p(p_1)), \mathbb{A}_p(g_p(p_2))), \tag{3}$$

其中，R_e 是交叉熵损失，$\not{G}$ 是梯度截断操作。g_p 是用于非对称非对比损失 [5], [6] 的预测头，包括一个全连接层、一个 Batch 归一化层和一个 ReLU 层。

多头空间自关系。 在 ViT 中，MHSA 通过将特征分成多个组，执行多个并行的自注意操作。研究者已经观察到不同的组可能会关注不同的语义模式 [13]。受此启发，我们沿通道维度将特征嵌入分成 M 组，并在每组内计算空间自关系，为每个视图获取 M 个空间自关系。默认情况下，我们设置 $M = 6$，如表12所示。

3.3 通道自关系

在神经网络中，每个通道表示图像中的某种模式。不同的通道编码不同的模式 [88], [89]，使神经网络有强大的表示能力。ViT 中的 FFN [11] 将通道间的模式合并，并隐式地建模通道之间的关系 [88]，即一个通道中编码的模式与其他通道编码的模式之间存在不同程度的相关性，如图 2所示。这一机制启发我们将通道自关系作为为自监督学习的特征表示，以增强通道维度上的自关系建模能力。具体来说，我们将 ViT 的特征嵌入转化为通道自关系，然后使用通道自关系作为自监督学习的特征表示。

生成通道自关系表示。我们首先给出将特征嵌入转化为通道自关系的变换 $\mathbb{P}$。如式 (2) 所示，给定两个视图的特征嵌入（即 r_1 和 r_2，一个结构与 h_p 相同的投影头 h_c 处理这些嵌入并得到 $c_1 = h_c(r_1)^T$ 和 $c_2 = h_c(r_2)^T$。之后，我们分别为每个视图计算通道自关系。对于一个视图，例如 $c_1 \in \mathbb{R}^{HW\times C}$，我们计算其通道自关系 $\mathbb{A}_c(c_1) \in \mathbb{R}^{C\times C}$，如下所示：

$$\mathbb{A}_c(c_1) = \text{Softmax}\left(\frac{c_1^T \cdot c_1}{H \cdot W} \Big/ t_c\right), \tag{4}$$

其中 Softmax 函数将自关系的每一行归一化以得到概率分布；t_c 是控制概率分布的锐度的温度参数。

基于通道自关系的自监督学习。通道自关系也可以作为一种新形式的特征表示，并应用于不同的自监督损失函数。与式 (3) 中基于空间自关系的非对比损失类似，我们给出使用通道自关系的非对比损失如下所示：

$$L_c = \text{R}_e(\not{G}(\mathbb{A}_c(c_1)), \mathbb{A}_c(g_c(c_2))), \tag{5}$$

其中 R_e 是交叉熵损失；g_c 是结构与式 (3) 中的 g_p 相同的预测头。该损失函数强化了不同视图间通道自关系的一致性，从而增强了模型的通道自关系建模能力。与空间自关系不同，我们不需要考虑不同视图之间的空间不对齐问题。因为我们强化的是通道自关系的一致性，不是通道特征的一致性，而式 (4) 中定义的通道自关系没有空间维度。

3.4 实现细节

损失函数。默认情况下，我们将我们提出的空间/通道自关系和图像嵌入作为自我监督损失的特征表示，因为这些形式表达了图像特征的不同性质。总的损失函数如下所示：

$$L = L_I + \alpha L_p + \beta L_c, \tag{6}$$

其中，α 和 β 为空间和通道损失的权重，L_I 是基于图像级嵌入的损失，例如 DINO [13] 中的基于聚类的损失。我们在表 8中表明了仅使用我们提出的自关系便可以实现与图像级嵌入相当甚至更好的性能。结合这三种表示形式可以进一步获得更高的表征质量，这表明自关系是图像级嵌入的有效补充。为了提高训练效率并进行公平比较，我们使用 multi-crop 增强技术来生成全局和局部视图。对于局部视图，我们遵循 [4], [13] 来计算每个全局和局部视图之间的损失，但忽略局部视图之间的损失。

网络架构。我们使用 Vision Transformer [11] 作为编码器网络。根据 [7], [13]，两个视图 $\tau_1(x)$ 和 $\tau_2(x)$ 的表示 r_1 和 r_2 是通过一个动量更新的编码器网络 f_1 和编码器网络 f_2 提取的。在训练期间，编码器网络 f_2 的参数 θ_2 通过梯度下降进行更新。编码器网络 f_1 的参数 θ_1 按照以下方式进行更新：$\theta_1 = \lambda\theta_1 + (1-\lambda)\theta_2$，其中 $\lambda \in [0, 1]$ 是动量系数。根据 DINO [13]，λ 被设置为 0.996，并且在训练过程中通过余弦调整策略逐渐增加到 1.0。相应的，我们将 f_1 和 f_2 后的投影头分别表示为 h_p^1/h_c^1 和 h_p^2/h_c^2。h_p^1/h_c^1 的参数也是通过 h_p^2/h_c^2 的参数进行动量更新的，这遵循 f_1 参数的更新方案。在预训练后，我们仅使用编码器网络进行下游任务的迁移学习。

4 实验

在这一部分，我们验证了将空间和通道自关系用于自监督学习的效果。我们在章节 4.1 中提供了预训练设置。在章节 4.2中，我们在多个下游任务中将我们的方法与现有方法进行比较，并展示了一致的提升。在章节 4.3 中，我们进行了消融实验。

表 1
ImageNet-1K 数据上的分类任务，一级 ImageNet-S 上的半监督语义分割任务。对于 ImageNet-S，我们报告了验证集和测试集上的 mIoU。"PT" 表示加载自监督预训练的权重进行初始化，"FT" 表示加载在 ImageNet-1K 微调后的权重进行初始化。

	Backbone	Epochs	Classification ImageNet-1K		Segmentation PT		FT	
			Top-1	Top-5	val	test	val	test
DINO [13]	ViT-S/16	100	79.7	95.1	35.1	34.4	54.6	54.4
+SERE	ViT-S/16	100	80.9	95.5	36.9	36.0	57.3	56.2
iBOT [18]	ViT-S/16	100	80.9	95.4	38.1	37.8	57.9	57.4
+SERE	ViT-S/16	100	81.5	95.8	41.0	40.2	58.9	57.8
iBOT [18]	ViT-B/16	100	83.3	96.6	48.3	47.8	62.6	63.0
+SERE	ViT-B/16	100	83.7	96.7	48.6	48.2	63.0	63.3

表 2
语义分割、目标检测和实例分割任务的迁移学习。AP^b 表示目标检测 (DET) 中的检测平均精度（AP），AP^m 表示实例分割 (SEG) 中的分割平均精度（AP）。

	VOC SEG		ADE20K SEG	
	mIoU	mAcc	mIoU	mAcc
DINO [13]	77.1	87.5	42.6	53.4
+SERE	79.7	88.8	43.8	54.6

	COCO DET			COCO SEG		
	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
DINO [13]	46.0	64.9	49.7	40.0	62.0	42.8
+SERE	46.6	65.9	50.2	40.5	62.9	43.5

4.1 预训练设置

除非另有说明，我们采用 ViT-S/16 作为骨干网络。DINO [13] 被选为我们的主要基线方法。模型由 AdamW [90] 优化器进行训练，学习率为 0.001，batch 大小为 512。我们在 ImageNet-1K [1] 数据集上进行 100 epochs 的预训练以进行性能比较。在消融实验中，我们使用 ImageNet-S_{300} 数据集 [26] 以节省训练成本。与 [13] 一致，我们采用多尺度裁剪训练方案，其中采用 2 个分辨率为 224×224 的全局视图，以及 4 个分辨率为 96×96 的局部视图。全局视图的裁剪比例在 0.35 和 1.0 之间。局部视图的裁剪比例在 0.05 和 0.35 之间。对于空间自关系，式 (2) 中的 $\mathbb{O}$ 操作的 H_s/W_s 被设置为 13/13 (用于全局视图) 或 6/6 (局部视图)。空间自关系中的头数 M 默认设置为 6。对于编码器网络, 式 (2) 中的 t_p 和式 (4) 中的 t_c 分别设置为 0.5 和 0.1。对于动量编码器，我们将 t_p 和 t_c 设置为 1.0 和 1.0。式 (6) 中的 α 和 β 分别设置为 1.0 和 1.0。

对于 iBOT [18]，我们使用了 10 个局部视图以进行公平比较。我们将图像的全局视图裁剪比例设置在 0.4 和 1.0 之间，局部视图的裁剪比例被设置在 0.05 和 0.4 之间。在优化过程中，我们还使用参数为 0.3 的梯度裁剪。在式 (6) 中，α 和 β 为 0.2 和 0.5。此外，我们提供了使用 ViT-B/16 作为骨干网络的实验，并在附录中展示了预训练和微调的详细信息。

表 3
对比更长的预训练时间。

(a) 在 ADE20K 数据集上的语义分割结果。

	Backbone	Epochs	mIoU	mAcc
iBOT [18]	ViT-S/16	800	45.4	56.2
+SERE	ViT-S/16	100	45.8	56.8
iBOT [18]	ViT-B/16	400	50.0	60.3
+SERE	ViT-B/16	200	50.0	60.9

(b) 在 ImageNet-1K 数据集上的分类结果。

	Backbone	Epochs	Top-1	Top-5
iBOT [18]	ViT-S/16	300	81.1	-
+SERE	ViT-S/16	100	81.5	95.8

表 4
ImageNet-1K 上的半监督分类。我们使用 1%/10% 的训练标签微调模型，并使用 100% 的验证标签评估模型。

	1%		10%	
	Top-1	Top-5	Top-1	Top-5
DINO [13]	52.1	77.8	70.0	89.8
+SERE	55.9	81.0	71.5	90.6

4.2 性能和分析

我们通过将预训练模型迁移到图像级分类任务和密集预测的下游任务来验证自关系对自监督学习的有效性。除非另有说明，模型在 ImageNet-1k 上进行了 100 轮的预训练。为了更容易理解，使用自关系表示法进行预训练的模型标记为 SERE。

ImageNet-1K 分类。我们比较了在 ImageNet-1K 数据集上进行完全微调的分类性能。当使用 ViT-S/16 时，预训练模型使用 AdamW 优化器和 512 的 batch 大小进行 100 epochs 的微调。初始学习率设置为 1e-3，并按系数 0.65 的逐层衰减。在 5 epochs 的 warm up 后，学习率按照余弦衰减策略逐渐衰减到 1e-6。我们报告了在 ImageNet-1k 验证集上的 Top-1 和 Top-5 准确率。如 表 1所示，SERE 在 Top-1 准确率上比 DINO 和 iBOT 分别提高了 1.2% 和 0.6%。与经过 300 epochs 训练的 iBOT 相比，SERE 在 Top-1 准确率上提高了 0.4%，而预训练时间只有 1/3，如 3 (b) 所示。此外，使用 ViT-B/16，SERE 在 Top-1 准确率上较 iBOT 提高了 0.4%，如 1所示。这些结果表明，SERE 增强了 ViT 的与类别相关的表征能力。

ImageNet-1K 半监督分类。我们还以半监督方式评估分类性能。按照 [18] 的设置，我们在 ImageNet-1K 数据集上使用 1% 和 10% 的训练标签微调预训练模型 (1000 epochs)。我们使用 AdamW 优化器，以 1024 的 batch 大小和 1e-5 的学习率来训练模型。表 4报告了 ImageNet-1K 验证集上的 Top-1 和 Top-5 准确性。SERE 在使用 1% 和 10% 标签情况下均取得

表 5
分类任务上的迁移学习。我们在多个数据集上对预训练模型进行微调，并报告 Top-1 准确率。

	$Cifar_{10}$	$Cifar_{100}$	$INat_{19}$	Flwrs	Cars
DINO [13]	98.8	89.6	76.9	97.8	93.5
+SERE	98.9	90.0	77.5	98.0	93.5

表 6
与在 ImageNet-1K 数据集上与基于掩码图像建模的方法相比较。† 表示有效的预训练 epochs [18]，它考虑了预训练期间实际使用的图像数量。‡ 表示模型在 ImageNet-1K 上进行了 200 epochs 的微调，而其他模型则进行了 100 epochs 的微调。

	Architecture	Pre-training Epochs†	Top-1
DINO [13]	ViT-S/16	300	79.7
MAE‡ [14]		800	80.9
iBOT [18]		400	80.9
DINO [13]+SERE		300	80.9
iBOT [18]+SERE		400	81.5
BEiT [58]	ViT-B/16	800	83.2
MAE [14]		800	83.3
iBOT [18]		400	83.3
iBOT [18]+SERE		400	83.7

了更高的准确性。在仅有 1% 标签时，Top-1 准确性显著提高了 3.8%，展现了我们的方法在半监督学习下的优势。

在 ImageNet-S 上进行半监督语义分割。 ImageNet-S [26] 数据集扩展了 ImageNet-1K 数据集，其为所有的验证图像以及部分训练图像上提供像素级语义分割标注。在 ImageNet-S 数据集上评估语义分割可以避免预训练和微调数据集之间潜在的领域偏移影响。我们使用 ImageNet-S 训练集中的语义分割标注对模型进行微调，并在 ImageNet-S 的验证集和测试集上评估性能。对于 ViT-S/16 模型的初始化，我们可以使用自监督的预训练权重 (ImageNet-S_{PT}) 或在 ImageNet-1K 数据集上微调后的权重 (ImageNet-S_{FT})。我们使用一个随机初始化的 1×1 卷积层作为分割头部。我们使用 AdamW 优化器对模型进行 100 epochs 的微调，batch 大小为 256，权重衰减为 0.05。学习率初始为 5e-4，并按系数 0.5 逐层衰减。在经过 5 epochs 的 warm up 后，学习率按余弦衰减策略逐渐下降到 1e-6。再训练中，图像被缩放并裁剪为 224×224；测试时，图像的分辨率沿短边缩放为 256。

如表 1所示，与 DINO 和 iBOT 相比，当使用自监督预训练的权重初始化模型时，SERE 将验证集的 mIoU 提高了 1.8% 和 2.9%。当加载微调后的权重进行初始化时，SERE 相对于 DINO/iBOT 提高了 2.7%/1.0% 的 mIoU。我们可以得出结论，SERE 增强了关系建模能力，使 ViT 具有更强的与形状相关的表示能力。

在分类任务上进行迁移学习。为了评估在分类任务上的迁移能力，我们在不同的数据集上微调了预训练模型。这些数据集包括 CIFAR [91]、Flowers [92]、Cars [93]、iNaturalist19 [94]。训练细节总结在补充材料中。如表 5所示，SERE 在 Top-1 准确率上表现优于 DINO，表明 SERE 有助于分类任务上的迁移学习。

表 7
将 SERE 与多种自监督学习方法结合。模型在 $ImageNet\text{-}S_{300}$ 数据集上进行 100 epochs 的预训练。

	VOC SEG		$ImageNet\text{-}S^{PT}_{300}$	
	mIoU	mAcc	val	test
MoCov3 [15]	65.7	78.7	24.0	24.8
+SERE	67.5	80.6	29.1	29.9
DINO [13]	68.1	81.1	28.8	29.6
+SERE	73.5	84.7	41.2	42.0
iBOT [18]	74.5	85.5	41.5	42.0
+SERE	75.9	86.3	45.3	45.6

在语义分割上进行迁移学习。我们还评估了在语义分割任务上的迁移学习性能，包括 PASCAL VOC2012 [25] 和 ADE20K [3] 数据集上的语义分割。以 ViT-S/16 为骨干网络，我们使用 UperNet [95] 作为分割模型。在 PASCAL VOC2012 和 ADE20K 数据集上，我们根据 [18] 的训练设置对模型进行了 20k 和 160k 步的微调，并设置 batch 大小为 16。如表 2所示，对于 PASCAL VOC2012 数据集，自关系相较于 DINO 分别提高了 2.6% 和 1.3% 的 mIoU 和 mAcc。在 ADE20K 数据集上，与 DINO 相比，mIoU 和 mAcc 分别提高了 1.2% 和 1.2%。如表 3(a) 所示，SERE 甚至可以以较少的预训练时间超过 iBOT。因此，语义分割任务受益于 SERE 更强的自关系表示能力。

在物体检测和实例分割上进行迁移学习。我们使用 Cascade Mask R-CNN [24] 来评估在目标检测和实例分割任务上的迁移学习性能。根据 [18] 的设置，模型在 COCO train2017 数据集 [2] 上进行 12 epochs 的微调，batch 大小为 16。在 COCO val2017 数据集上，我们报告了检测 AP (AP^b) 和分割 AP (AP^m) 的结果。与 DINO 相比，SERE 的 AP^b 提高了 0.6%，AP^m 提高了 0.5%，这些结果表明 SERE 有助于模型准确地定位和分割物体。

与掩码图像建模 (MIM) 对比。我们还证明了我们提出的方法 SERE 在优于并补充各种基于掩膜图像建模（MIM）的方法。如表 6所示，SERE 可以显著增强基于对比学习的方法 (例如 DINO)。DINO+SERE 以更少的预训练/微调时间，获得了与基于 MIM 的方法 (iBOT 和 MAE) 相媲美的性能。与此同时，SERE 和 MIM 可以相互补充。例如，iBOT 通过与 SERE 集合可以进一步提高 0.4% 的 Top-1 准确率。此外，图 4中的定

表 8
使用不同特征表示进行自监督训练。L_I、L_p 和 L_c 分别表示使用图像级嵌入 [13]、空间自关系和通道自关系的损失函数。在对下游任务进行微调时，三种损失均不适用的模型是随机初始化的。

L_I	L_p	L_c	VOC SEG		ImageNet-S^{PT}_{300}	
			mIoU	mAcc	val	test
✗	✗	✗	25.6	35.7	0.2	0.2
✓			68.1	81.1	28.8	29.6
	✓		71.5	83.0	23.7	23.7
		✓	61.4	75.6	22.5	22.3
✓	✓		70.7	82.6	33.3	34.5
✓		✓	69.8	82.9	36.5	38.3
	✓	✓	71.5	83.3	30.6	30.3
✓	✓	✓	73.5	84.7	41.2	42.0

表 9
PASCAL VOC 数据集上的分割 F-measure [96]。F-measure 忽略了语义类别。

	L_p	$L_p + L_I$	$L_p + L_I + L_c$
IoU	87.1	86.7	87.7

性结果显示，与 iBOT 相比，SERE 产生的注意力图更精确，噪声更少。这些结果强烈证实了与 MIM 相比 SERE 的有效性。

与更多自监督学习方法合作。自关系表示与现有的特征表示是正交的。因此，它可以集成到各种自监督学习方法中。为了证明这一点，我们将 SERE 与 MoCo v3 [15]、DINO 和 iBOT 结合起来，即将这些方法的自监督部分用作式 (6) 中的 L_I。我们在 ImageNet-S_{300} 数据集上进行了 100 epochs 的预训练，以节省计算成本，而其他训练设置与基线方法保持一致。如表 7所示，使用 SERE 可以一致地提升不同的基线方法，说明了 SERE 的的泛化性。例如，对于在 Pascal VOC 数据集上的语义分割，SERE 将 MoCo v3 的 mIoU 提高了 1.8%，mAcc 则提高了 2.0%。对于在 ImageNet-S_{300} 数据集上的半监督语义分割，SERE 比 MoCo v3 提高了 5.1% 的 mIoU。

4.3 消融研究

为了节省计算成本，我们用两个全局视图在 ImageNet-S_{300} 数据集 [26] 上对所有模型进行了 100 epochs 的预训练。预训练后，我们在 PASCAL VOC 数据集上进行了语义分割的评估，并在 ImageNet-S_{300} 数据集上进行了半监督语义分割的评估。

空间和通道自关系的影响。我们比较了不同的特征表示在自监督学习中的有效性，包括我们提出的空间/通道自关系与 DINO 使用的图像级特征嵌入。如表 8所示，与特征嵌入相比，

表 10
自关系与图像块级嵌入合作。DINO+ 表示将使用图像块级嵌入的聚类损失添加到 DINO [13] 中。

DINO	DINO+	SERE	VOC SEG		ImageNet-S^{PT}_{300}	
			mIoU	mAcc	val	test
✓			68.1	81.1	28.8	29.6
	✓		72.6	84.3	40.0	40.4
		✓	73.5	84.7	41.2	42.0
	✓	✓	75.0	86.1	44.8	46.0

表 11
与基于 batch 关系的损失，即 Barlow [85] 方法，进行比较。

	VOC SEG		ImageNet-S^{PT}_{300}	
	mIoU	mAcc	val	test
Barlow [85]	69.5	82.2	33.2	32.9
SERE	69.8	82.9	36.5	38.3

空间自关系在 PASCAL VOC 数据集上提高了 3.4% 的 mIoU 和 1.9% 的 mAcc。这些结果表明，使用空间自关系来训练自监督 ViT 可以进一步增强 ViT 的空间关系建模能力，这有益于密集预测任务。而通道自关系虽然不如其他两种表示形式，但仍然提高了 ViT 的表示质量。使用通道自关系预训练的模型在分割和分类任务上的表现要显著优于随机初始化的模型。

与图像级别嵌入协作。我们验证了自关系与图像级别嵌入之间的正交性，如表 8所示。当与图像级别特征嵌入结合时，空间自关系和通道自关系在 PASCAL VOC 数据集上分别提高了 2.6% 和 1.7% 的 mIoU。在 ImageNet-S_{300} 数据集上，与特征嵌入相比，空间自关系和通道自关系分别提高了 4.5% 和 7.7% 的 mIoU。同时使用这三种特征表示则进一步提高了所有任务的性能，这表明自关系与图像级别特征嵌入在自监督学习中是正交且互补的。

在 L_I 和 L_c 之间的合作。如表 8所示，仅仅 L_p 在 PASCAL VOC 数据集上可以实现优于 $L_p + L_I$ 或 $L_p + L_c$ 的性能。然而，使用 $L_p + L_I + L_c$ 的性能优于 L_p。这种现象是因为图像级别嵌入（L_I）和通道自关系（L_c）存在各自的局限性，而它们的协同作用可以缓解这些局限性。具体如下：1）关于 L_c，建模通道自关系需要有意义且多样化的通道特征作为基础。然而，仅依赖 L_c 可能不能充分优化通道特征，并导致模型崩溃 (例如每个通道编码可能相同的特征)。相比之下，L_I 有助于学习多样化且有意义的通道特征，因此解决了上述 L_c 的局限性。2）L_I 会对空间特征产生负面影响。我们通过检查忽略语义类别的 F-measure [96] 来验证这一点。如表 9所示，在 $L_p + L_I$ 与 L_I 比较时，F-measure 下降，表明 L_I 损害了空间特征。我们假设 L_I 使表征在空间维度上的区分性弱于 L_p。然而，通过同时使用 L_c，我们促进了更准确的空间表征

表 12
不同头数 M 对空间自关系的影响。

M	VOC SEG		ImageNet-S_{300}^{PT}	
	mIoU	mAcc	val	test
1	72.4	84.0	38.7	39.3
3	72.7	84.8	38.9	39.4
6	73.5	84.7	41.2	42.0
12	73.4	85.1	40.8	41.7
16	72.5	84.3	39.3	39.8

表 13
式 (2) 和式 (4) 中不同 t_p 和 t_c 的影响。

t_p	t_c	VOC SEG		ImageNet-S_{300}^{PT}	
		mIoU	mAcc	val	test
0.50	0.50	72.0	84.2	36.7	36.7
0.50	0.10	73.5	84.7	41.2	42.0
0.50	0.01	70.4	82.7	33.6	34.6
1.00	0.10	70.2	83.1	36.7	38.2
0.50	0.10	73.5	84.7	41.2	42.0
0.10	0.10	73.7	85.0	39.9	40.8

学习，从而缓解了使用 L_I 引起的缺点。

与图像块级别嵌入协同工作。我们还验证了自关系表示与图像块级别嵌入之间的正交性，如表 10所示。我们使用图像块级别嵌入实现聚类损失，并将该损失添加到 DINO 得到基线方法 DINO+。DINO+ 在不同的指标上较 DINO 有显著提升，说明了使用图像块级别嵌入的有效性。与 DINO+ 相比，自关系在 PASCAL VOC 和 ImageNet-S 数据集上分别将 mIoU 提高了 0.9% 和 1.2%。协同使用这两种表示带来了进一步的提升，如在 PASCAL VOC 和 ImageNet-S 数据集上较 DINO+ 分别提高了 2.4% 和 4.8% 的 mIoU。这些结果表明，自关系与块级别嵌入在自监督的 ViT 中是互补的。

自关系与 batch 关系的比较。相关工作 Barlow [85] 在整个 batch 内建模通道关系，即 batch 关系。相比之下，我们提出的 SERE 在单个图像内计算自关系。为了验证自关系相对于 batch 关系的优势，我们分别使用这两种关系对 ViT-S/16 进行预训练。如表 11所示，与 batch 关系相比，自关系在 PASCAL VOC 和 ImageNet-S_{300} 数据集上分别将 mIoU 提高了 0.3% 和 3.3%。这些结果表明，自关系比 batch 关系更适合于 ViT 的自监督训练。

多头效应。受 ViT 中的 MHSA 模块启发，我们使用多头空间自关系。表 12展示了不同头数 M 对空间自关系的影响。与单头版本相比，将 M 增加到 6 可以在 PASCAL VOC 数据集带来最大的性能提升，即 1.1% 的 mIoU。$M = 12$ 只实现了有限的额外增益，而 $M = 16$ 则出现了快速的性能下降。更多的头数可以建模多样化的空间自关系，但用于计算每个自关系的通道数量减少。头数过多则会导致自关系估计不准确，损害表示质量。因此，我们默认将头数设置为 6，以平衡空间自关系的多样性和质量。

表 14
当将 SERE 与 iBOT [18] 结合时，式 (6) 中不同 α 和 β 的影响。所有模型在 ImageNet-1K 上都进行了 100 epochs 的预训练。

α	β	Classification ImageNet-1K		Segmentation VOC		Segmentation ImageNet-S_{PT}	
		Top-1	Top-5	mIoU	mAcc	val	test
0.20	0.20	81.3	95.7	80.7	89.9	39.9	39.3
0.20	0.50	81.5	95.8	81.2	90.0	41.0	40.3
0.20	1.00	81.3	95.8	80.9	89.8	41.7	41.8
0.10	0.50	81.3	95.8	80.9	89.5	40.7	40.5
0.20	0.50	81.5	95.8	81.2	90.0	41.0	40.3
0.80	0.50	81.3	95.8	80.8	89.7	40.3	40.1

表 15
式 (3) 和式 (5) 中的非对称损失。

	VOC SEG		ImageNet-S_{300}^{PT}	
	mIoU	mAcc	val	test
DNIO baseline	68.1	81.1	28.8	29.6
+SERE symmetry	72.1	84.4	37.1	37.9
+SERE asymmetric	73.5	84.7	41.2	42.0

锐度效应。方程式 (2) 和式 (4) 中的温度项控制了自关系分布的锐度。小的温度值会使分布变得更加尖锐，而大的温度值则会使分布变得更加平滑。在表 13中，我们验证了温度对空间和通道自关系的作用。对于通道自关系，将温度从 0.1 降低到 0.01 会导致 PASCAL VOC 数据集上的 mIoU 从 73.5% 迅速下降到 70.4%。而将温度从 0.1 增加到 0.5 也会使 mIoU 从 73.5% 下降到 72.0%。因此，我们选择 0.1 作为通道自关系的默认温度。对于空间自关系，温度 0.5 表现比 1.0 更好，而将温度 0.5 和 0.1 的差异有限。我们将空间自关系的默认温度设置为 0.5，因为温度为 0.5 在大规模数据集 ImageNet-S 上实现了更好的性能。

损失函数权重的影响。方程（6）中的 α 和 β 分别确定了空间和通道自关系的相对重要性。表 14显示 SERE 对不同的 α 和 β 具有鲁棒性。在不同的权重中，$\alpha = 0.2$ 和 $\beta = 0.5$ 的组合在分类任务上实现了最佳性能，并在分割任务上具有竞争力的性能。因此，我们将这种组合作为默认设置。

非对称损失的影响。当使用图像级别嵌入作为特征表示时，非对称结构已被证明对非对比损失是有效的 [5], [6]。为了验证自关系表示是否也受益于非对称结构，我们在表 15中比较了基于非对称和对称结构的自关系损失。自关系在使用非对称和对称结构时都提升了 DINO 基线方法。对于 PASCAL VOC

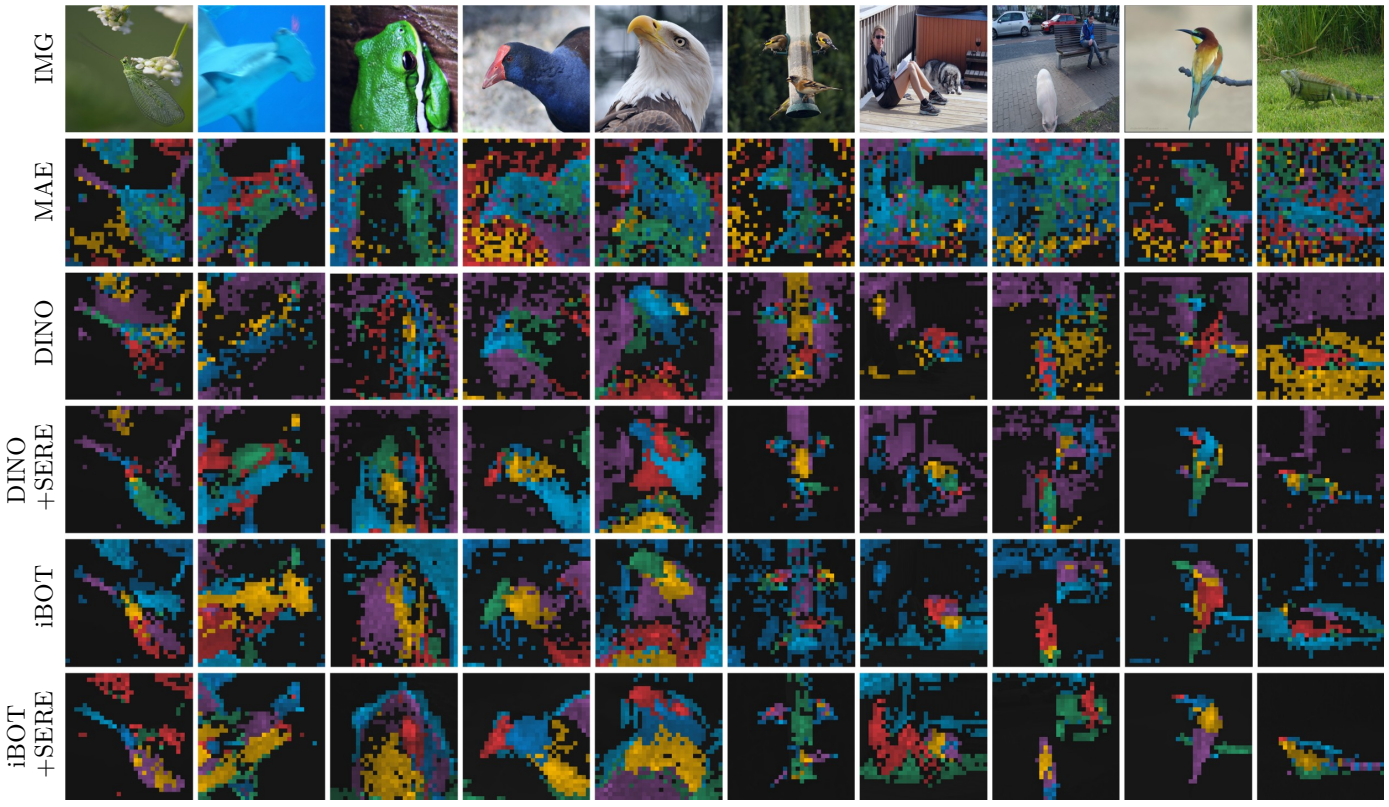


图 4. 可视化预训练 ViT-S/16 最后一个块的注意力图。我们提取了 CLS 令牌对其他图像块的注意力图。不同的颜色表示被不同头关注的区域。

表 16
自关系损失对 CNN 的作用。DINO 和 SERE 均使用 ResNet-50 网络进行训练。

	VOC SEG		ImageNet-S^{PT}_{300}	
	mIoU	mAcc	val	test
DINO (ResNet-50)	61.6	74.6	20.2	19.9
+SERE (ResNet-50)	62.5	75.0	20.9	20.7

和 ImageNet-S_{300} 数据集，对称结构在 mIoU 上优于 DINO，分别提高了 4.0% 和 8.3%。不对称结构进一步超越了对称结构，在 PASCAL VOC 和 ImageNet-S_{300} 数据集上分别提高了 1.4% 和 4.1% 的 mIoU。尽管不对称结构对自关系不是必不可少的，但它仍然有助于基于自关系的预训练。

对卷积神经网络的适应性。受 ViT 性质的启发，我们使用自关系进行自监督学习。然而，我们也想知道自关系表示是否能够有益于卷积神经网络（CNN）的自监督学习。为了验证这一点，我们分别使用 DINO 和 SERE 对 ResNet-50 [9] 进行预训练。训练细节见附录。如表 16所示，与 DINO 相比，SERE 在 PASCAL VOC 和 ImageNet-S_{300} 数据集中分别将 mIoU 提高了 0.7% 和 0.8%。尽管自关系是为 ViT 设计的，但自关系仍然可以改善 CNN 的表示质量。与在 ViT 上的改进相比，CNN 上的改进相对较小，这表明自关系更适合 ViT。

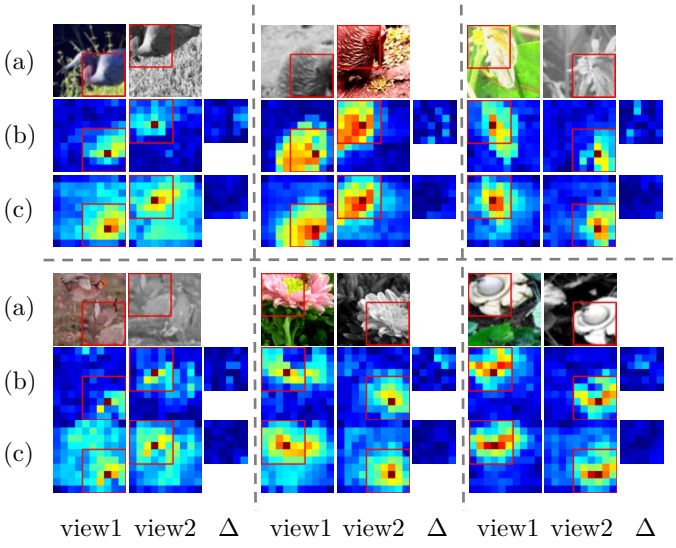


图 5. 两个视图的空间自关系之间的差异。(a) 每个图像的两个视图。(b) 由 DINO 生成的空间自关系。(c) 由 SERE 生成的空间自关系。View1 和 view2 表示从图像生成的两个视图的自关系。Δ 是重叠区域中自关系之间的差异，用红色框表示。我们在附录中提供了具体的可视化方法。

4.4 分析与可视化

自关系的不变性。在自监督学习中，学习对图像增强操作 (如缩放、平移和颜色抖动) 具有不变性的表征是非常重要的 [97], [98], [99], [100], [101], [102]。然而，现有方法主要关注特征嵌入的不变性，但并未考虑空间/通道关系的不变性，而这也是

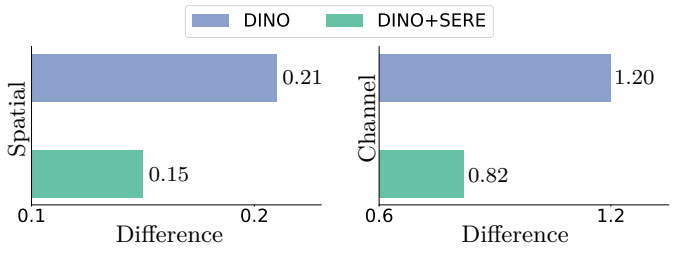


图 6. 在 ImageNet-S 验证集上两个视图间空间自关系（左侧）和通道自关系（右侧）的平均差异。我们在附录中展示了计算细节。

ViT 的重要属性之一。相比之下，我们提出的 SERE 可以增强空间/通道关系的不变性。为了验证这一点，我们测量了不同视图的自关系之间的平均差异。如图 6所示，我们观察到 SERE 在空间和通道维度中显著缩小了自关系差异。在图 5中的可视化结果也显示，SERE 预训练的模型在两个视图的重叠区域产生了较小的空间自关系差异。较小的差异意味着更高的不变性。因此，这些结果表明 SERE 预训练的 ViT 所捕获的自关系对图像增强操作有更强的不变性。

注意力图的可视化。在图 4中，我们可视化了 ViT 的最后一个块的注意力图。这些可视化结果表明 SERE 产生的注意力图比各种方法更精确且更少噪音，包括基于 MIM 的方法，如 MAE [14] 和 iBOT [18]。MAE 生成的注意力图比较嘈杂，几乎突出显示了图像中的所有位置。相比之下，SERE 的注意力图主要集中在语义物体上。例如，图 4的第三列显示 SERE 可以定位青蛙，但 MAE 主要关注背景。此外，与 iBOT 和 DINO 相比，SERE 生成的注意力图更准确地定位对象。例如，在图 4的第七和第八列中，SERE 发现了 iBOT 忽略的人物。

空间自关系与 MIM 的比较。空间自关系和 MIM 都作用于空间维度，但它们的效果显著不同。MIM 增强了标记级别的表示，而空间自关系则专注于提高对标记间关系的建模能力。我们通过以下几点来支持这一观点：1）如图 4所示，SERE 生成的注意力图比 MAE [14] 和 iBOT [18] 更精确且更少噪音。ViT 的注意力图可以反映模型对标记间关系建模的能力，因为注意力是以查询和键之间的标记级别关系计算的。因此，这一观察表明 SERE 提供了更强的捕获标记间关系的能力。2）如图 6所示，我们展示了 SERE 增强了空间自关系对不同图像增强操作的不变性。3）如表 6所示，SERE 相对于不同基于 MIM 的方法都实现了一致的改进，证实了 SERE 相对于 MIM 的有效性。例如，iBOT+SERE 较 iBOT 可以提升 0.4% 的 Top-1 准确率，如表 1所示。

5 结论

本文提出了一种基于特征自关系的自监督学习方案，以提高自监督 ViT 的关系建模能力。具体来说，我们提出使用特征的空间和通道自关系作为自监督学习的特征表示，而不是直接使用特征嵌入作为特征表示。自关系与特征嵌入正交，进一步增强了现有的自监督方法的性能。我们展示了特征自关系在细粒度水平上增强了自监督 ViT，有利于多个下游任务，包括图像分类、语义分割、目标检测和实例分割。

致谢。 This work is funded by NSFC (NO. 62225604, 62176130), and the Fundamental Research Funds for the Central Universities (Nankai Universitiy, 070-63233089). Computation is supported by the Supercomputing Center of Nankai University.

参考文献

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., "Imagenet large scale visual recognition challenge," Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[3] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in IEEE Conf. Comput. Vis. Pattern Recog., July 2017.

[4] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in Adv. Neural Inform. Process. Syst., 2020.

[5] X. Chen and K. He, "Exploring simple siamese representation learning," in IEEE Conf. Comput. Vis. Pattern Recog., June 2021.

[6] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Ávila Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - a new approach to self-supervised learning," in Adv. Neural Inform. Process. Syst., 2020.

[7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in IEEE Conf. Comput. Vis. Pattern Recog., June 2020.

[8] H. SUN and M. LI, "Enhancing unsupervised domain adaptation by exploiting the conceptual consistency of multiple self-supervised tasks," SCIENCE CHINA Information Sciences, vol. 66, no. 4, pp. 142 101–, 2023.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 770–778.

[10] S. Gao, Z.-Y. Li, Q. Han, M.-M. Cheng, and L. Wang, "Rf-next: Efficient receptive field search for convolutional neural networks," IEEE Trans. Pattern Anal. Mach. Intell., 2022.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," Int. Conf. Learn. Represent., 2021.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Int. Conf. Comput. Vis., 2021.

[13] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in Int. Conf. Comput. Vis., 2021.

[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in IEEE Conf. Comput. Vis. Pattern Recog., June 2022, pp. 16 000–16 009.

[15] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in Int. Conf. Comput. Vis., October 2021.

[16] Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, "Self-supervised learning with swin transformers," arXiv preprint arXiv:2105.04553, 2021.

[17] H. Lu, Y. Huo, M. Ding, N. Fei, and Z. Lu, "Cross-modal contrastive learning for generalizable and efficient image-text retrieval," Machine Intelligence Research, pp. 1–14, 2023.

[18] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "ibot: Image bert pre-training with online tokenizer," Int. Conf. Learn. Represent., 2022.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in International Conference on Machine Learning (ICML), 2020.

[20] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[21] O. J. Hénaff, S. Koppula, J.-B. Alayrac, A. van den Oord, O. Vinyals, and J. a. Carreira, "Efficient visual pretraining with contrastive detection," in Int. Conf. Comput. Vis., October 2021, pp. 10 086–10 096.

[22] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in Adv. Neural Inform. Process. Syst., A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[23] K. Kim, B. Wu, X. Dai, P. Zhang, Z. Yan, P. Vajda, and S. J. Kim, "Rethinking the self-attention in vision transformers," in IEEE Conf. Comput. Vis. Pattern Recog. Worksh., June 2021, pp. 3071–3075.

[24] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in IEEE Conf. Comput. Vis. Pattern Recog., June 2018.

[25] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," Int. J. Comput. Vis., vol. 88, pp. 303–338, 2009.

[26] S. Gao, Z.-Y. Li, M.-H. Yang, M.-M. Cheng, J. Han, and P. Torr, "Large-scale unsupervised semantic segmentation," arXiv preprint arXiv:2106.03149, 2021.

[27] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in Eur. Conf. Comput. Vis. Springer, 2016, pp. 649–666.

[28] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in IEEE Conf. Comput. Vis. Pattern Recog., July 2017.

[29] M. Noroozi and P. Favaro, "Unsupervised learning of visual representions by solving jigsaw puzzles," in Eur. Conf. Comput. Vis., 2016.

[30] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in Int. Conf. Learn. Represent., 2018.

[31] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in Int. Conf. Comput. Vis., December 2015.

[32] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in International Conference on Machine Learning (ICML), 2008.

[33] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in IEEE Conf. Comput. Vis. Pattern Recog., June 2016.

[34] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in Int. Conf. Comput. Vis., Oct 2017.

[35] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in IEEE Conf. Comput. Vis. Pattern Recog., June 2018.

[36] Y. Zhao, G. Wang, C. Luo, W. Zeng, and Z.-J. Zha, "Self-supervised visual representations learning by contrastive mask prediction," in Int. Conf. Comput. Vis., October 2021, pp. 10 160–10 169.

[37] C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, T.-L. Liu, Y. Chen, and Y. LeCun, "Decoupled contrastive learning," arXiv preprint arXiv:2110.06848, 2021.

[38] W.-C. Wang, E. Ahn, D. Feng, and J. Kim, "A review of predictive and contrastive self-supervised learning for medical images," Machine Intelligence Research, pp. 483–513, 2023.

[39] L. Wang, H. Xu, and W. Kang, "Mvcontrast: Unsupervised pretraining for multi-view 3d object recognition," Machine Intelligence Research, pp. 1–12, 2023.

[40] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in Eur. Conf. Comput. Vis., 2018.

[41] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, and C. C. Loy, "Online deep clustering for unsupervised representation learning," in IEEE Conf. Comput. Vis. Pattern Recog., June 2020.

[42] A. YM., R. C., and V. A., "Self-labelling via simultaneous clustering and representation learning," in Int. Conf. Learn. Represent., 2020.

[43] S. A. Koohpayegani, A. Tejankar, and H. Pirsiavash, "Mean shift for self-supervised learning," in Int. Conf. Comput. Vis., October 2021, pp. 10 326–10 335.

[44] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, "Whitening for self-supervised representation learning," in International Conference on Machine Learning (ICML), 2021, pp. 3015–3024.

[45] Y. Tian, X. Chen, and S. Ganguli, "Understanding self-supervised learning dynamics without contrastive pairs," in International Conference on Machine Learning (ICML), 2020.

[46] C. Ge, Y. Liang, Y. Song, J. Jiao, J. Wang, and P. Luo, "Revitalizing cnn attentions via transformers in self-supervised visual representation learning," in Adv. Neural Inform. Process. Syst., 2021.

[47] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in IEEE Conf. Comput. Vis. Pattern Recog., June 2021, pp. 16 684–16 693.

[48] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, Z. Li, and P. Luo, "Detco: Unsupervised contrastive learning for object detection," in Int. Conf. Comput. Vis., October 2021, pp. 8392–8401.

[49] Z. Dai, B. Cai, Y. Lin, and J. Chen, "Up-detr: Unsupervised pre-training for object detection with transformers," in IEEE Conf. Comput. Vis. Pattern Recog., June 2021, pp. 1601–1610.

[50] B. Roh, W. Shin, I. Kim, and S. Kim, "Spatilly consistent representation learning," in IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[51] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in Int. Conf. Comput. Vis., 2021.

[52] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in Adv. Neural Inform. Process. Syst., 2021.

[53] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2T: Pyramid pooling transformer for scene understanding," IEEE Trans. Pattern Anal. Mach. Intell., 2022.

[54] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in International Conference on Machine Learning (ICML). PMLR, 2021, pp. 10 347–10 357.

[55] C. Li, J. Yang, P. Zhang, M. Gao, B. Xiao, X. Dai, L. Yuan, and J. Gao, "Efficient self-supervised vision transformers for representation learning," in Int. Conf. Learn. Represent., 2022.

[56] P. Zhou, Y. Zhou, C. Si, W. Yu, T. K. Ng, and S. Yan, "Mugs: A multi-granular self-supervised learning framework," in arXiv preprint arXiv:2203.14415, 2022.

[57] Z. Li, Z. Chen, F. Yang, W. Li, Y. Zhu, C. Zhao, R. Deng, L. Wu, R. Zhao, M. Tang, and J. Wang, "MST: Masked self-supervised transformer for visual representation," in Adv. Neural Inform. Process. Syst., 2021.

[58] H. Bao, L. Dong, S. Piao, and F. Wei, "BEit: BERT pre-training of image transformers," in Int. Conf. Learn. Represent., 2022.

[59] S. Gao, P. Zhou, M.-M. Cheng, and S. Yan, "Towards sustainable self-supervised learning," arXiv preprint arXiv:2210.11016, 2022.

[60] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," arXiv preprint arXiv:2202.03026, 2022.

[61] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in IEEE Conf. Comput. Vis. Pattern Recog., June 2022.

[62] L. Wang, F. Liang, Y. Li, H. Zhang, W. Ouyang, and J. Shao, "Repre: Improving self-supervised vision transformer with reconstructive pre-training," arXiv preprint arXiv:2201.06857, 2022.

[63] S. Atito, M. Awais, and J. Kittler, "Sit: Self-supervised vision transformer," arXiv preprint arXiv:2104.03602, 2021.

[64] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," arXiv preprint arXiv:2112.09133, 2021.

[65] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in Eur. Conf. Comput. Vis. Springer, 2020, pp. 213–229.

[66] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in Int. Conf. Comput. Vis., October 2019.

[67] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in IEEE Conf. Comput. Vis. Pattern Recog., June 2019.

[68] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in Eur. Conf. Comput. Vis., 2018.

[69] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in Int. Conf. Comput. Vis., October 2019.

[70] X. Li, J. Wu, H. Fang, Y. Liao, F. Wang, and C. Qian, "Local correlation consistency for knowledge distillation," in Eur. Conf. Comput. Vis., 2020.

[71] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in Int. Conf. Learn. Represent., 2017.

[72] Y. Chen, N. Wang, and Z. Zhang, "Darkrank: Accelerating deep metric learning via cross sample similarities transfer," AAAI Conference on Artificial Intelligence (AAAI), vol. 32, no. 1, Apr. 2018.

[73] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., June 2019.

[74] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., June 2019.

[75] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in IEEE Conf. Comput. Vis. Pattern Recog., June 2022, pp. 12 319–12 328.

[76] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," in Int. Conf. Learn. Represent., 2022.

[77] O. Siméoni, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Unsupervised object discovery for instance recognition," in Winter Conference on Applications of Computer Vision, 2018.

[78] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce, "Localizing objects with self-supervised transformers and no labels," in Brit. Mach. Vis. Conf., November 2021.

[79] Y. Wang, X. Shen, S. X. Hu, Y. Yuan, J. L. Crowley, and D. Vaufreydaz, "Self-supervised transformers for unsupervised object discovery using normalized cut," in IEEE Conf. Comput. Vis. Pattern Recog., June 2022.

[80] M. Ki, Y. Uh, J. Choe, and H. Byun, "Contrastive attention maps for self-supervised co-localization," in Int. Conf. Comput. Vis., October 2021, pp. 2803–2812.

[81] I. Kakogeorgiou, S. Gidaris, B. Psomas, Y. Avrithis, A. Bursuc, K. Karantzalos, and N. Komodakis, "What to hide from your students: Attention-guided masked image modeling," arXiv preprint arXiv:2203.12719, 2022.

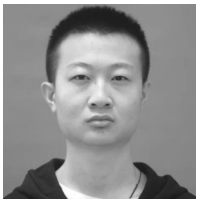
[82] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in Eur. Conf. Comput. Vis. Worksh., 2016, pp. 685–701.

[83] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in Adv. Neural Inform. Process. Syst., vol. 28, 2015.

[84] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in Int. Conf. Comput. Vis., December 2015.

[85] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," arXiv preprint arXiv:2103.03230, 2021.

[86] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International Conference on Machine Learning (ICML), 2015, pp. 448–456.
[87] A. F. Agarap, "Deep learning using rectified linear units (relu)," arXiv preprint arXiv:1803.08375, 2018.
[88] L. Liu, Q. Huang, S. Lin, H. Xie, B. Wang, X. Chang, and X. Liang, "Exploring inter-channel correlation for diversity-preserved knowledge distillation," in Int. Conf. Comput. Vis., October 2021, pp. 8271–8280.
[89] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Adv. Neural Inform. Process. Syst., 2012.
[90] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Int. Conf. Learn. Represent., 2019.
[91] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep. 0, 2009.
[92] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), pp. 722–729, 2008.
[93] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in Int. Conf. Comput. Vis. Worksh., 2013.
[94] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in IEEE Conf. Comput. Vis. Pattern Recog., June 2018.
[95] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in Eur. Conf. Comput. Vis., September 2018.
[96] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 569–582, 2015.
[97] S. Purushwalkam Shiva Prakash and A. Gupta, "Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases," Adv. Neural Inform. Process. Syst., vol. 33, 2020.
[98] M. Patrick, Y. M. Asano, P. Kuznetsova, R. Fong, J. a. F. Henriques, G. Zweig, and A. Vedaldi, "On compositions of transformations in contrastive self-supervised learning," in Int. Conf. Comput. Vis., 2021.
[99] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," in IEEE Conf. Comput. Vis. Pattern Recog., June 2020.
[100] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in Int. Conf. Learn. Represent., 2022.
[101] L. Ericsson, H. Gouk, and T. M. Hospedales, "Why do self-supervised models transfer? investigating the impact of invariance on downstream tasks," 2022.
[102] X. Wang, K. He, and A. Gupta, "Transitive invariance for self-supervised visual representation learning," in Int. Conf. Comput. Vis., 2017.



Zhong-Yu Li is a Ph.D. student from the college of computer science, Nankai university. He is supervised via Prof. Ming-Ming cheng. His research interests include deep learning, machine learning and computer vision.

Shanghua Gao is a Ph.D. candidate in Media Computing Lab at Nankai University. He is supervised via Prof. Ming-Ming Cheng. His research interests include computer vision and representation learning.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012, and then worked with Prof. Philip Torr in Oxford for 2 years. Since 2016, he is a full professor at Nankai University, leading the Media Computing Lab. His research interests include computer vision and computer graphics. He received awards, including ACM China Rising Star Award, IBM Global SUR Award, etc. He is a senior member of the IEEE and on the editorial boards of IEEE TPAMI and IEEE TIP.