通过异构自监督学习增强表征

李钟毓，尹博文，刘永祥，刘丽，程明明

Abstract—融合来自不同架构的异构表征已促进多种视觉任务发展，例如某些混合网络结合了Transformer与卷积结构。然而在自监督学习领域，此类异构架构间的互补特性尚未得到充分挖掘。为此，我们提出异构自监督学习框架（**HSSL**），通过强制基础模型向架构异构的辅助头学习，在不改变模型结构的前提下以表征学习方式赋予基础模型新特性。为全面理解HSSL，我们在包含基础模型与辅助头的多种异构组合上开展实验，发现基础模型的表征质量随架构差异增大而提升。这一发现促使我们提出：快速确定最佳辅助头的搜索策略，以及若干简单有效的模型差异增强方法。本框架兼容多种自监督方法，在图像分类、语义分割、实例分割和目标检测等下游任务中均取得卓越性能。代码与数据集详见https://github.com/NK-JittorCV/Self-Supervised/

Index Terms—自监督学习, 异构架构, 表征学习

1 引言

自监督学习无需昂贵标注即可学习丰富表征的优势已获广泛验证。这一成功主要归功于不同代理任务的设计，特别是实例判别任务 [1], [2], [3], [4]与掩码图像建模 [5], [6], [7]。通过将这类方法适配于卷积神经网络 [8], [9]、视觉Transformer [2], [8], [10], [11]及Swin Transformer [12]等不同架构，研究者在图像分类 [13]、语义分割 [14], [15]与目标检测 [16]等下游任务中均取得了显著提升。

由于计算机制的差异，不同神经网络架构学习到的表征具有揭示其本质特性的独特属性，例如全局与局部建模能力。已有研究 [17], [18], [19], [20]证实不同架构的特性存在互补性，补充材料第1节通过对照实验进一步验证了多架构组合相对于单一架构的优越性。现有方法 [12], [21], [22], [23]主要通过架构设计来利用这种互补性，而本文则在不改变模型架构的前提下，通过表征学习的方式实现互补优势的融合。

受上述分析启发，我们提出异构自监督学习方法 (HSSL)，该方法能通过融合其他架构的特性来增强模型性能。具体而言，在预训练阶段，模型由基础模型和辅助头组成，其中辅助头采用与基础模型异构的架构。这种异构性使得辅助头能够提供基础模型所缺失的特征表征。为使基础模型获得这些缺失特征，我们促使基础模型的表征模仿辅助头的表征，如Fig. 1所示。当预训练完成后，基础模型已整合新特性，此时可移除辅助头。

为进行全面分析，我们考察了基础模型与辅助头之间的多种异构组合，发现基础模型的改进程度与其和辅助头之间的差异性呈正相关关系。差异性越显著，意味着辅助头能提供更多基础模型所缺失的特征表征，从而放大基础模型的性能增益。这一发现使得特定基础模型能够选择最匹配的辅助头。我们提出了一种快速搜索策略，可同步评估所有候选辅助头与同一基础模型进行异构表征学习的效果。因此能够快速确定最优的辅助头配置。此外，我们还对选定的辅助头进行针对性调整，通过扩大其与基础模型的差异性来进一步提升性能。

我们提出的HSSL可以应用于不同的自监督学习方案，例如对比学习 [24]、自聚类 [2]和掩码图像建模 [5]，因此与多种自监督训练方法正交 [2], [5], [10], [24]。在包括图像分类 [13]、语义分割 [14]、半监督语义分割 [25], [26]、实例分割 [16]和目标检测 [14], [16]在内的多种下游任务中，HSSL无需改变网络结构即可为不同架构带来显著提升。

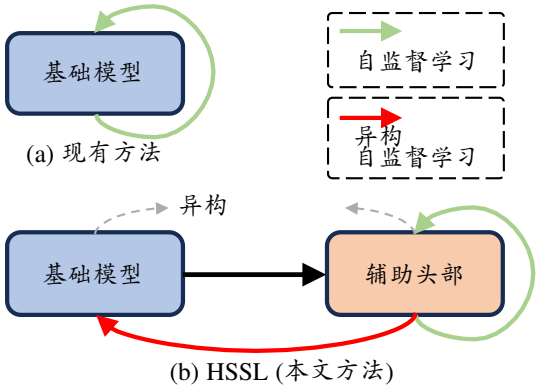


Fig. 1. 异构自监督学习 (HSSL)示意图。(a) 通用自监督学习方法使基础模型自我监督。(b) 本文方法在架构与基础模型异构的辅助头部指导下监督基础模型，使基础模型学习到新特性。

我们的主要贡献总结如下：

- 我们提出异构自监督学习方法，使基础模型能够学习不同架构的特征。
- 通过大量实验，我们发现基础模型与辅助头之间的差异与基础模型的改进呈正相关，并提出一种快速搜索策略来为特定基础模型寻找最合适的辅助头。
- 所提出的表征学习方式与现有自监督方法兼容，并在各种下游任务中持续提升性能。

2 相关工作

2.1 自监督学习

自监督学习能够在无监督环境下学习丰富的表征，降低标注数据的收集成本。早期方法设计了多种可生成自由监督信号的代理任务，例如着色 [27], [28]、拼图游戏 [29]、旋转预测 [30]、自编码器 [31], [32]、图像修复 [33]和计数 [34]。

自监督学习近年来的成功主要归功于实例判别 [35], [36], [37], [38], [39]和掩码图像建模 [5], [40], [41], [42], [43]方法。新范式如相关图像建模 [44]和损坏图像建模 [45]的提出，进一步丰富了该领域。

实例判别方法。通过随机图像增强生成多个视图并对其表征进行对齐 [46], [47], [48], [49], [50]。该框架已扩展出多种

- 李钟毓、尹博文和程明明任职于中国天津的南开大学*VCIP & TBI*中心。
- 程明明同时任职于深圳市福田区南方科技大学人工智能研究院。
- 刘永祥和刘丽任职于中国长沙的国防科技大学电子科学与技术学院。
- 本工作部分得到国家重点研发计划（*2021YFB3100800*）和国家自然科学基金（*62225604*，*62376283*）资助。计算资源由南开大学超算中心提供。

损失函数形式，包括对比学习 [15], [51], [52], [53], [54]、特征对齐 [55], [56], [57]、聚类分配 [58], [59], [60], [61], [62]、冗余减少 [63], [64]、排序 [65]和关系建模 [8], [66]。

这些方法已应用于图像级 [52], [67], [68], [69]和密集级 [70], [71], [72], [73]任务，并在卷积神经网络 [9]、视觉Transformer [2]和Swin Transformer [12]等架构上展现出广泛适应性。然而现有方法常忽视不同架构间的潜在互补性。本文提出HSSL框架，通过异构自监督学习方案利用不同架构的互补特性。此外，本方法与现有自监督技术具有正交性。

掩码图像建模。基于掩码图像建模（MIM）的方法 [74], [75], [76], [77]通过未掩码图像块重建被掩码区域，着重于空间上下文学习。研究者们探索了多种重建目标以获取具有不同特性的表征。例如，基于像素的重建 [5], [41], [78], [79], [80], [81]通常能产生强非线性表征。为赋予表征更强的语义信息，更多类型的重建目标被采用，如人工设计的HOG特征 [82], [83], [84]、频域信息 [80], [85]、掩码位置 [86]、在线网络特征 [66], [87], [88], [89], [90]、离散化标记 [40], [91], [92]，或多目标组合 [91], [93], [94]。最新研究 [66]还通过现成的预训练模型重建表征，尤其在使用大规模数据集时 [95]表现出色。当采用在线网络时 [90], [96]，部分工作 [6], [10], [97], [98], [99], [100]进一步结合掩码建模与实例判别的优势以提升性能。此外，除重建目标外，部分研究 [74], [77], [79], [101]还探索了不同掩码策略以促进高层表征学习。

与实例判别类似，掩码图像建模技术 [10], [102], [103], [104], [105], [106], [107], [108]已被成功应用于视觉Transformer [109]、ConvNext-V2 [102]及Swin等多种架构。这些进展凸显了利用架构多样性提升基于MIM的表征学习能力的潜力。

2.2 神经网络异构性

异构神经网络通过融合多种架构类型 [20], [23], [114]，能够产生互补特征并促进各类视觉任务，包括语义分割 [17], [115], [116]、目标检测 [117], [118]、图像分类 [18], [119], [120]以及图像质量评估 [121]。现有方法主要通过设计新架构来实现特征互补。例如Wu等研究者 [120]将卷积与注意力机制结合于同一架构，从而获得更优的分类精度。相较之下，我们通过表征学习使特定架构构建的网络能够学习其他任意架构的特征，且无需改变网络结构。因此，所提方法在融合不同架构特征方面具有更高灵活性。

部分研究 [19], [122]尝试利用架构互补性来改进自监督学习。具体而言，这些方法使ViT与ResNet相互引导。但除这对组合外，它们缺乏对不同架构间互补性的系统分析与理解。相比之下，我们研究了包括但不限于ViT和ResNet的多种架构，并对互补性如何促进自监督学习进行了机理分析。研究发现模型差异越大带来的提升越显著，这为设计针对性辅助头模块来引导特定模型提供了理论依据。

3 方法

在Section 3.1节中，我们回顾了现有的自监督方法。接着在Section 3.2节中，我们描述了提出的异构自监督学习方法，并论证了其与现有方法的兼容性。Section 3.3节通过实验证明，性能提升源自异构架构的互补特性。Section 3.4节分析了优质辅助头的特征，发现更大的模型差异能带来更多收益。基于这一发现，我们在Section 3.5节提出快速搜索策略来为特定基础模型选择最佳辅助头，并在Section 3.6节提出若干简单有效的方法来扩大模型差异以获得更大提升。

3.1 预备知识

HSSL可以通过不同形式实现，例如实例判别和掩码图像建模。本文主要采用实例判别框架作为示例说明。我们首先简要回顾实例判别的通用框架。给定图像x，通过不同数据增强生成其不同视图x_1和x_2。它们的表征z_1和z_2分别由教师网络和学生网络提取。随后，实例判别方法最大化z_1与z_2之间的相似性。具体而言，损失函数存在不同形式 [2], [10], [123]，我们将其抽象表示为$\mathcal{L}(z_1, z_2)$。

3.2 异构监督

将现有方法 [2], [10]采用的主干网络称为基础模型，HSSL通过引入一个与基础模型架构不同的辅助头，为基模型补充其缺失的特征表征。整体流程如Fig. 2所示。为简化表述，我们将教师分支和学生分支的基础模型/辅助头分别记为f_1/h_1和f_2/h_2。给定输入x_1和x_2，基础模型提取特征表示$z_1^b = f_1(x_1)$与$z_2^b = f_2(x_2)$。随后辅助头将这些表征作为输入，输出$z_1^h = h_1(z_1^b)$和$z_2^h = h_2(z_2^b)$。由于异构架构提取的z_1^h/z_2^h与z_1^b/z_2^b存在差异，z_1^h/z_2^h包含了z_1^b/z_2^b所缺失的部分特征。通过损失函数$\mathcal{L}(z_1^h, z_2^b)$促使$z_1^h$与$z_2^b$相互靠近，基础模型可从中学习这些缺失的特征。

同时，为确保辅助头能学习到有意义的特征，我们还将教师模型和学生模型中辅助头提取的表征进行对齐，即使用损失函数$\mathcal{L}(z_1^h, z_2^h)$。基础模型与辅助头进行联合预训练，总损失函数$\mathcal{L}$定义如下：

$$\mathcal{L} = \mathcal{L}(z_1^h, z_2^b) + \mathcal{L}(z_1^h, z_2^h). \tag{1}$$

在预训练阶段，辅助头以串联方式连接在基础模型末端，仅需少量层数即可学习有效特征。因此额外增加的训练时间和内存开销可以忽略不计。预训练完成后，我们将移除辅助头，仅保留基础模型。

与不同**SSL**方法的兼容性。所提出的HSSL方法可适配多种自监督学习（SSL）框架，包括MoCo [24]、DINO [2]、iBOT [10]及MAE [5]，如Tab. 6所示。当与不同方法结合时，公式Equ. (1)定义的损失函数会呈现不同形式。对于基于聚类的算法 [2], [10]，表征会通过投影头和softmax函数转换为K维概率分布，此时损失函数定义为：

$$\mathcal{L} = -\sum_{i=1}^{K}(z_1^h)_i \log((z_2^b)_i) - \sum_{i=1}^{K}(z_1^h)_i \log((z_2^h)_i), \tag{2}$$

为简化表述省略了投影头与softmax函数。此外，其他形式的损失函数也可以与HSSL结合使用，例如对比学习中的InfoNCE [124] [2]以及掩码图像建模中的重建损失 [5]。更多细节请参阅补充材料第6节。

不同架构的分析。为了验证所提出方法HSSL的有效性，我们评估了不同辅助头对基础模型的影响。在此分析中，我们旨在探索不同架构对HSSL的影响。因此，我们选择ResNet [111]、PoolFormer [113]、ResMLP [112]、ConvNext [110]、ViT [109]和Swin [12]作为辅助头，因为它们具有多样化的架构。例如，ResNet [111]是基于局部卷积的经典卷积网络，而ConvNext [110]进一步采用了大核卷积。ViT [109]是基于全局自注意力的Transformer网络，而Swin [12]则在Transformer架构中集成了局部注意力机制。此外，PoolFormer [113]和ResMLP [112]采用了超越卷积与Transformer架构的差异化建模机制，即池化与空间MLP。如Tab. 1所示，使用辅助头能持续提升基础模型在所有配对中的表现[1]。值得注意的是，与基础模型异构的辅

1. 为节省计算成本，Section 3与Section 5所有实验均采用包含ImageNet-1K [13]中300个类别的ImageNet-S_{300}数据集 [25]。

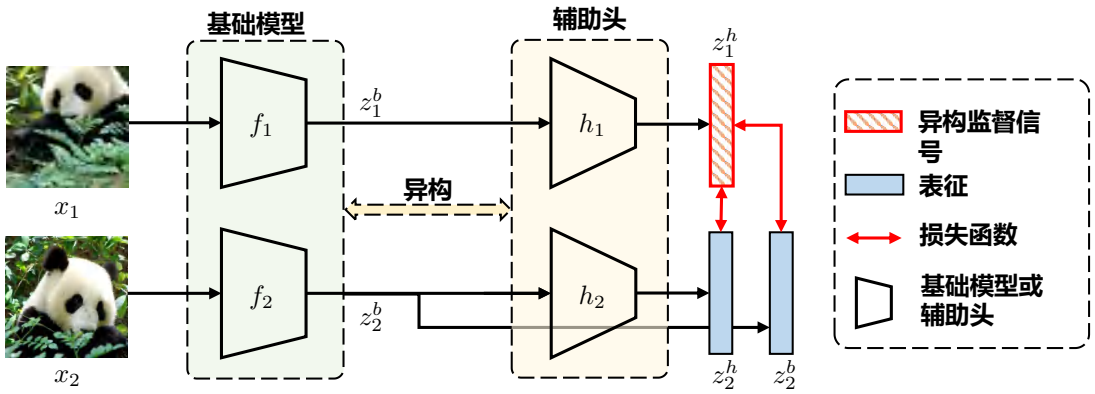


Fig. 2. 本文提出的HSSL 框架。 基础模型与辅助头的架构具有异构性。 辅助头提取的表征同时监督两个网络。 基础模型和辅助头可采用任意架构，例如ViT [109]、Swin [12]、ConvNext [110]、ResNet [111]、ResMLP [112] 及PoolFormer [113] 等。

TABLE 1
不同辅助头对各基础模型的影响

		基础模型			
		ViT		ResNet	
		Top-1	Top-5	Top-1	Top-5
	基线	67.5	84.4	63.2	84.3
辅助头	ViT [109]	68.0	84.7	64.0	84.3
	Swin [12]	69.4	85.9	63.9	84.4
	PoolFormer [113]	70.1	86.3	63.9	84.5
	ResNet [111]	71.7	86.9	63.5	84.3
	ResMLP [112]	72.6	87.8	64.4	84.9
	ConvNext [110]	72.7	87.6	63.7	84.4

TABLE 2
弱辅助头同样能增强强基础模型。 未声明辅助头的实验表示对应基础模型的基线结果。

基础模型	辅助头	Top-1
ViT [109]	-	67.5
ViT [109]	ResNet [111]	71.7
ViT [109]	ResMLP [112]	72.6
ResMLP [112]	-	58.0
ResMLP [112]	ViT [109]	59.6
Swin [12]	-	72.8
Swin [12]	PoolFormer [113]	73.7
Swin [12]	ResMLP [112]	73.4

助头 比同构辅助头带来更显著的性能提升。 以ViT作为基础模型时， ViT同构辅助头仅带来0.5%的Top-1准确率提升，而ConvNext异构辅助头则实现4.2%的Top-1准确率增益。 这些实证结果验证了所提HSSL方法的必要性。 我们还研究了相对较弱的辅助头是否能增强更强的基础模型。 Tab. 2 展示了积极的结果。 例如， 较弱的 PoolFormer [113] 将 Swin [12] 基础模型的 Top-1 准确率提高了 0.9%。 这表明我们的 HSSL 方法对不同模型架构具有鲁棒性，并且能在不同设置下带来一致的性能提升。

3.3 异构的效果

虽然HSSL 在不同基础模型和辅助头的组合中都有效， 但我们进一步探究了辅助头如何增强基础模型。 具体而言，我们观察到 辅助头可以解决部分基础模型无法处理的样本。 为说明这一点， 我们首先定义集合B_1、B_2和H， 分别包含： 通过基线方法（DINO [2]）预训练的基础模型能正确解决的样本、 通过HSSL 预训练的基础模型能正确解决的样本、 以及通过HSSL 预训练的辅助头能正确解决的样本。 同时， U表示包含数据集中所有样本的集合。 那么$H \cap (U - B_1)$包含的样本是： 辅助头能解决但基线预训练基础模型无法处理的样本。 这些样本的数量定义如下：

$$N_s = |H \cap (U - B_1)|. \tag{3}$$

TABLE 3
辅助头解决了基础模型（ViT）无法处理的样本。 sIoU和N_s分别表示HSSL与原始版本在正确样本上的重叠程度以及HSSL新增的正确样本数量。具体细节见Section 3.3节。

辅助头	Top-1	N_s	sIoU
ViT [109]	68.0	792	59.5
Swin [12]	69.4	854	60.7
PoolFormer [113]	70.1	904	60.8
ResNet [111]	71.7	1061	67.8
ResMLP [112]	72.6	1270	72.9
ConvNext [110]	72.7	1278	70.2

以ViT作为基础模型， 我们在Tab. 3中展示： 辅助头能够解决部分超出基础模型能力的样本。 更重要的是， 能解决更多基础模型未解样本的辅助头， 会给基础模型带来更显著的性能提升。

我们进一步探究在辅助头的指导下，基础模型能否处理$H \cap (U - B_1)$集合中的样本。通过HSSL 方法预训练后，基础模型和辅助头都能处理超出基线方法能力范围的样本。这些样本可分别表示为基础模型的$B_2 \cap (U - B_1)$和辅助头的$H \cap (U - B_1)$。值得注意的是，这两个子集之间存在显著重叠，其重叠程度可通过以下公式量化：

$$\text{sIoU} = \frac{|B_2 \cap (U - B_1) \cap H \cap (U - B_1)|}{|B_2 \cap (U - B_1)|}. \tag{4}$$

如Tab. 3所示，当使用ViT作为基础模型时，不同辅助头获得的sIoU值。例如， 采用ConvNext [110] 作为辅助头时重叠率达到70%。高度重叠表明基础模型的改进主要源于互补性和异质性。

3.4 模型差异分析

如Tab. 1所示， 不同辅助头对特定基础模型会产生不同效果。对于基于Transformer的基础模型ViT， 采用ConvNext作为辅助头比其他结构更为合适。当基础模型为ResNet时，使用ResMLP和ViT作为辅助头可以补充全局建模能力并带来更显著的性能提升。

上述观察促使我们深入探究优秀辅助头的构成要素。通过研究不同架构组合，我们发现基础模型与辅助头之间的差异越大，越能为基础模型带来性能增益。这一现象启发我们

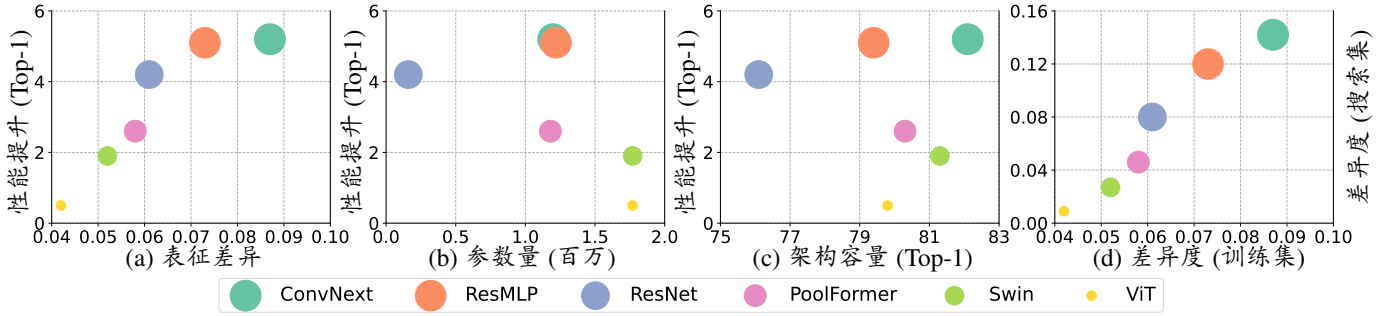


Fig. 3. 在(a)-(c)中，我们可视化了基础模型 (ViT-S/16)的性能提升与三个因素之间的关系： (a) 基础模型与辅助头之间的表征差异； (b) 单层辅助头的参数量； (c) 用于构建辅助头的架构容量。 对于各架构的容量指标， 我们采用各架构原始论文中报告的ImageNet-1K监督分类准确率作为参考。 在(d)中， 我们展示了通过搜索与单独检验各辅助头所得差异度之间的一致性趋势。 所有图中，点的大小与对应辅助头带来的性能提升呈正相关。

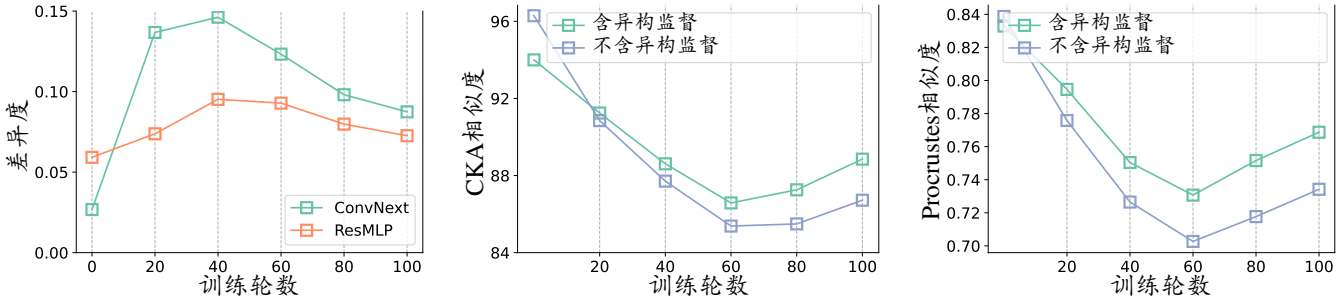


Fig. 4. 预训练过程中基础模型与辅助头之间差异或相似性的训练动态。 左图： 当使用ConvNext [110] 或ResMLP [112] 作为辅助头且使用ViT [109] 作为基础模型时， 基础模型与辅助头之间的差异$\mathcal{D}$ （定义见公式Equ. (5)）。 中图： 基础模型与辅助头之间的特征级CKA相似性。 右图： 基础模型与辅助头之间的特征级Procrustes相似性。

在Section 3.5节提出快速确定特定基础模型最佳辅助头的搜索策略，并在Section 3.6节提出若干简单有效的方法来扩大模型差异。

模型差异。 在异构自监督学习过程中， 辅助头会学习到基础模型本身缺失的部分特征。 也就是说，基础模型与辅助头之间存在表征差异， 即z_1^b与z_1^h之间的差异。 以基于自聚类的方法 [2], [10]为例， Section 3.1中定义的z_1^b 表示在K维上的概率分布。 我们采用Kullback-Leibler散度来度量该差异：

$$\mathcal{D} = -(z_1^b)^T \log(\frac{z_1^h}{z_1^b}), \tag{5}$$

其中z_1^b和z_1^h均提取自预训练后的教师网络。

差异越显著，改进越明显。 以ViT-S/16作为基础模型为例，在Fig. 3 (a)中， 我们展示了其从每个辅助头学习时的改进情况 以及与每个辅助头之间的差异。 可以观察到， 改进与差异之间存在正相关关系。 更显著的差异意味着辅助头学习到了 基础模型中缺失的更多特征， 从而促使基础模型补充更多特征。

为了进一步确认性能提升是否源于异构性， 我们分析了其他影响因素， 包括辅助头的参数量以及 构建辅助头所用架构的容量。 在此过程中，我们采用 ImageNet-1K数据集上的监督分类准确率 [13] （由各架构原始论文报告） 作为架构容量的参考指标。

如Fig. 3(c)和(d)所示， 这两个因素与性能提升 均无正向相关性。 例如， ViT [109] 虽比ResNet [111] 具有更大容量，但作为辅助头时ResNet比ViT更适用。 这些结果表明： 性能提升并非来自更强的辅助头， 而是源于异构性本身。

模型差异的动态变化。 基于差异分析， 我们研究了预训练过程中辅助头如何影响基础模型。 Fig. 4详细展示了 基础模型与辅助头之间的交互关系， 分别呈现了二者间的差异度$\mathcal{D}$（如Equ. (5)所定义）、 CKA相似度和Procrustes相似度。 从Fig. 4可以看出，训练过程中差异度呈现先上升后下降的趋势。 值得注意的是，异质监督 会显著放大差异度并降低相似度， 如中间和右侧面板所示。 这一观察结果表明，异质监督 能够促进基础模型从辅助头中学习。 此外，左侧面板显示， 当采用ViT作为基础模型时， 使用ConvNext作为辅助头比ResMLP 会产生更大的差异度。 这与先前分析一致， 表明更大的差异度可能 带来更显著的性能提升。 自然地，模型差异为我们提供了为特定基础模型选择最优辅助头的可能性。

3.5 寻找合适的辅助头部

一个合适的辅助头部 能够提供特定基础模型所缺失的更多特征， 从而更好地补充基础模型并产生更高的改进效果。 然而， 在无监督设置下， 缺乏标注数据来评估每个辅助头部。 受模型差异与改进效果正相关关系的启发， 我们采用模型差异度量， 通过无标签方法为特定基础模型确定最合适的辅助头部。 但由于候选辅助头数量庞大， 逐个测试候选方案十分耗时。 因此， 我们提出了一种高效的搜索策略， 通过一次快速训练即可找到 与基础模型差异最大的辅助头。

快速搜索策略。 与标准HSSL 架构不同，后者仅采用单一辅助头， 我们在训练过程中并行排列所有候选辅助头，如Fig. 5所示。 这种设计使得每个辅助头能够独立进行 异构自监督学习而互不干扰。 假设存在 N个候选辅助头， 每个头对应不同的架构。 对于输入x_1和x_2， 我们首先将它们送入 教师分支和学生分支的基础模型 分别生成表征$z_1^b =$

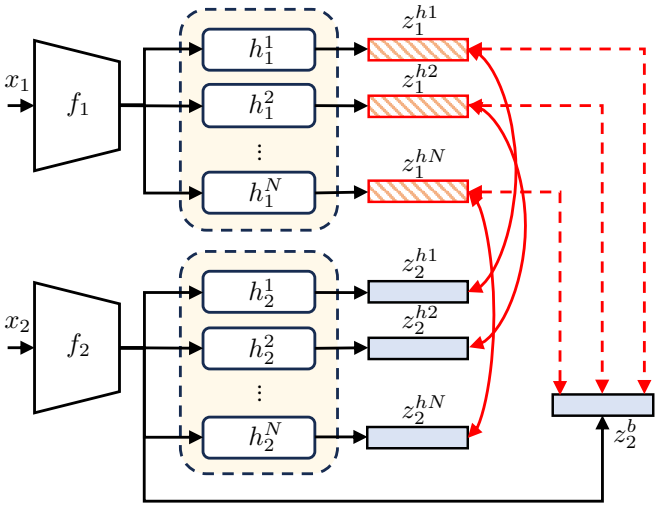


Fig. 5. 快速搜索策略示意图。给定N个不同的架构，我们构建N个不同的辅助头，其中$h^i_{1/2}$表示使用第i个架构构建的辅助头。下标1和2分别表示教师分支和学生分支。图中，红色虚线和实线对应于公式Equ. (6)中第一项和第二项的损失。为清晰起见，图中省略了投影头。

TABLE 4
以ViT作为基础模型时多个辅助头的协作情况。'$\mathcal{D}$'表示辅助头与基础模型之间的差异度。

辅助头	$\mathcal{D}$	Top-1	Top-5
ResMLP	7.3e-2	72.6	87.8
ConvNext	8.7e-2	72.7	87.6
ConvNext+ResMLP	11.0e-2	73.7	88.2

$f_1(x_1)$和$z_2^b = f_2(x_2)$。接着，在教师分支中，N个辅助头进一步处理z_1^b并生成异构表示$\{z_1^{hi} \mid i \in [0, N-1]\}$。类似地，学生分支生成$\{z_2^{hi} \mid i \in [0, N-1]\}$。对于第$i$个辅助头，我们如式Equ. (1)所示定义损失函数：

$$\mathcal{L}^{hi} = \mathcal{L}(z_1^{hi}, z_2^b) + \mathcal{L}(z_1^{hi}, z_2^{hi}). \tag{6}$$

所有辅助头的总体损失函数为：

$$\mathcal{L}^s = \frac{1}{N}\sum_{i=0}^{N-1}\mathcal{L}^{hi}. \tag{7}$$

实际应用中，式Equ. (6)中的特征在计算损失前会通过独立的投影头进行处理，这种做法在先前工作中已被广泛采用[2], [5]。在搜索过程中，我们为每个辅助头应用独立的投影头。为了最小化相互干扰，基础模型的表示$z^b_{1/2}$在与不同辅助头配对时，会被输入到不同的投影头中。为清晰起见，这些投影头在公式Equ. (6)和Equ. (7)中被省略。

使用公式Equ. (7)训练后，我们计算基础模型与每个辅助头之间的差异。对于第i个头，差异$\mathcal{D}_i$的计算方式为：

$$\mathcal{D}_i = -(z_1^b)^T \log(\frac{z_1^{hi}}{z_1^b}). \tag{8}$$

最终，我们选择差异最大的辅助头：

$$\arg\max_i \mathcal{D}_i, \tag{9}$$

其中第i个辅助头被识别为对基础模型最具互补性的辅助头。因此，可以快速确定基础模型的最佳辅助头。

搜索耗时分析。与通过多次训练逐个检验每个辅助头相比，所提出的搜索策略仅需单次训练。由于我们采用非常浅层的辅助头结构，训练过程中基础模型占据了绝大部分计算资源。实验结果表明，当存在六个辅助头时，同时训练所有辅助头（即本方案采用的搜索策略）所需时间仅为单独训练单个辅助头的1.4倍。因此，该搜索策略耗时约为$\frac{1.4\times1}{1\times6} \approx 23\%$，远低于逐个检验所有辅助头的传统方法。此外，我们通过实证研究发现仅需10%的训练数据即可完成有效搜索，这进一步显著降低了搜索耗时。

TABLE 5
辅助头中shortcut的分析。此处我们采用ViT作为基础模型，ConvNeXt作为辅助头。

首个shortcut	$\mathcal{D}$	Top-1	Top-5
保留	5.6e-2	71.0	86.8
移除	8.7e-2	72.7	87.6

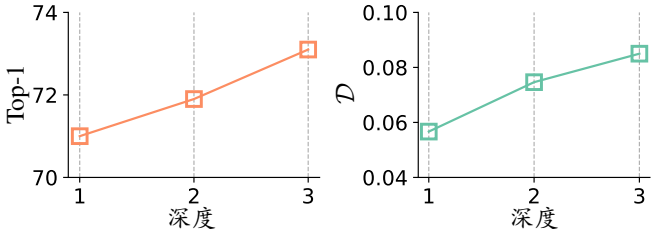


Fig. 6. 辅助头中网络深度的影响分析。我们采用ViT作为基础模型，ConvNext作为辅助头结构。

搜索结果分析。以ViT作为基础模型，我们分析了其与不同辅助头之间差异的相对关系。如Fig. 3 (b)所示，搜索过程中获得的相对关系与单独测试每个辅助头获得的结果一致，这验证了搜索策略的有效性。

3.6 扩大模型差异

Section 3.4节表明，更大的模型差异能带来更多的改进收益。受此观察启发，我们提出了三种简单但有效的技术来放大这种差异，从而进一步提升性能。

多辅助头的协同作用。基础模型仅能从特定辅助头中学习到有限的特征表示。受集成学习原理的启发，即组合多个模型可以带来性能提升，我们尝试将通过不同架构构建的多个辅助头相结合，以补充基础模型缺失的更多特征。具体而言，假设存在n个由不同架构组成的辅助头，我们分别在教师模型和学生模型中将其表示为$\{h_1^i|i \in [1,n]\}$和$\{h_2^i|i \in [1,n]\}$。基于基础模型产生的表示$z^b_{1/2}$，每个辅助头$h^i_{1/2}$对其进行处理并生成表示$h^i_{1/2}(z^b_{1/2})$。如Fig. 7(a)所示，这些表征通过以下方式组合：

$$z^{hc}_{1/2} = \text{concat}(\{h^i_{1/2}(z^b_{1/2})|i \in [1,n]\}), \tag{10}$$

其中concat表示沿通道维度的拼接操作。随后，我们将这些表征代入Equ. (1)式，得到新的损失函数如下：

$$\mathcal{L} = \mathcal{L}(z_1^{hc}, z_2^b) + \mathcal{L}(z_1^{hc}, z_2^{hc}). \tag{11}$$

与单一辅助头相比，多个辅助头能够为基模型提供更丰富的特征特性。如Tab. 4所示，当同时使用两个辅助头（即ConvNext [110] 和ResMLP [112] ）时，我们获得了比单独使用ConvNext或ResMLP更大的性能提升。这表明我们的HSSL方法可以通过多个辅助头的协同作用实现进一步的性能提升。

深化辅助头结构。多项研究[125], [126]发现，模型不同层级学习到的表征存在显著差异，且层间跨度越大表征差异

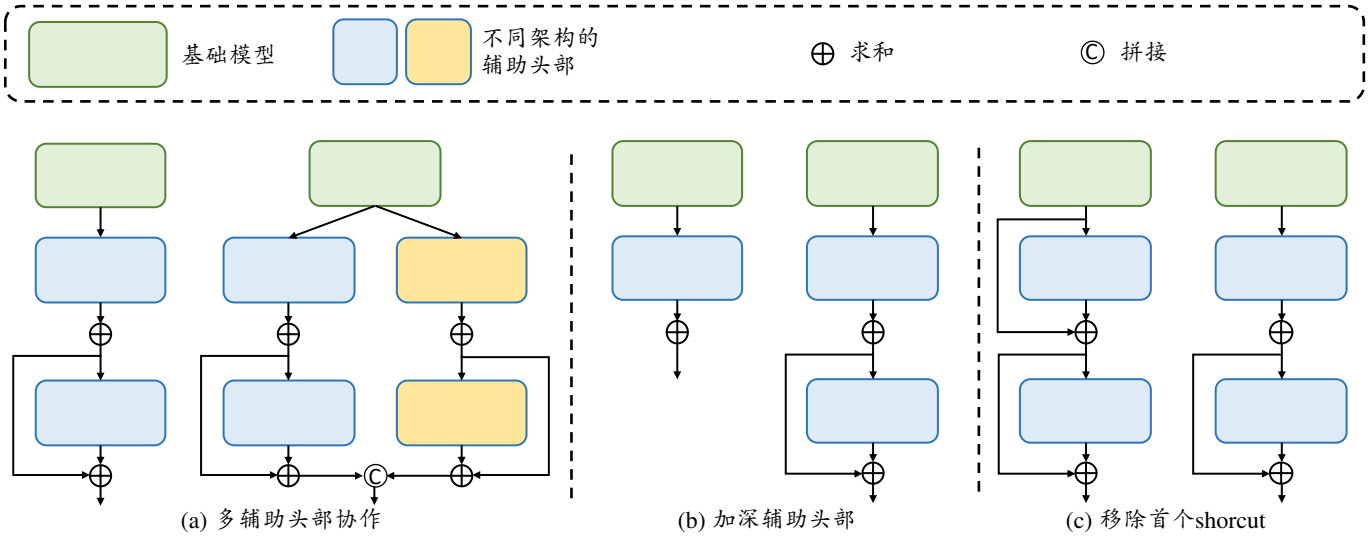


Fig. 7. 三种扩大模型差异的策略。每种策略中，左侧展示基线方法，右侧展示具体策略。

性越明显。与此同时，更深的网络结构能够学习到更具表现力的特征表示 [111]。基于此，我们通过深化辅助头结构使其学习与基础模型更具差异性的特征。具体实现方式是在辅助头上叠加更多模块，如 Fig. 7 (b) 所示。Fig. 6 中的实验结果表明，将辅助头从单层加深至三层后，表征差异显著扩大且带来更大性能提升。

移除首个**shortcut**。在大多数架构中，默认会使用shortcut以确保收敛性。然而，在我们的HSSL 中，我们观察到辅助头第一层的shortcut 会减小基础模型与辅助头之间的差异。为说明这一论点，我们以两层辅助头为例，其中两层分别表示为 F_1和F_2。我们用z表示基础模型的输出。当保留第一个shortcut时，辅助头输出为$z + F_1(z) + F_2(z + F_1(z))$。相比之下，移除第一个shortcut后辅助头输出为$F_1(z) + F_2(F_1(z))$。可以看出前者直接将z（即基础模型的输出）加到辅助头的输出中，从而减小了模型差异。补充材料进一步展示了这一现象。因此，我们移除了第一个shortcut，如Fig. 7 (c)所示。该方法扩大了模型差异并带来更显著的改进，如Tab. 5所示。

4 实验

4.1 实验设置

我们将HSSL 与多种自监督方法相结合，包括MoCov3 [24]、DINO [2] 、AttMask [74] 、iBOT [10] 、MAE [5] 和MFF [81] 。对于每种方法，我们遵循其官方实现。在预训练阶段，我们采用ViT-S/16或ViT-B/16架构作为基础模型，除非另有说明，辅助头均使用 ConvNext架构。在辅助头中，默认深度设为3层，并移除第一层的快捷连接。预训练与微调的更多细节详见补充材料。

4.2 实验结果

ImageNet-1K图像分类实验。我们首先在ImageNet-1K上对基础模型进行全参数微调，并比较分类性能指标，如Tab. 6所示。采用ViT-B/16架构时，当预训练400轮次后，HSSL 方法的Top-1准确率较iBOT [10] 提升0.5%。当训练轮次增至600时，HSSL 可获得84.1%的Top-1准确率，甚至超越1600轮次的iBOT模型表现。我们还将HSSL 与基于掩码图像建模(MIM)的方法相结合，具体实现细节详见补充材料。在Tab. 6中，当ViT-S/16预训练400轮次后，HSSL 使MAE方法的Top-1准确率提升0.4%。相较于MFF [81] 方法，在ViT-B/16预训练300轮次后，我们的方法将Top-1准确率提高了0.3%。

我们还通过在ImageNet-1K数据集上使用k-NN和线性探测评估了HSSL 的有效性。如Tab. 6所示，HSSL 持续改进了包括基于实例判别的方法（如DINO [2] 和MoCo

2. 为公平比较，我们报告有效训练轮数 [10]，该指标考虑了预训练阶段实际使用的图像数量。对于采用10个局部裁剪的iBOT和DINO方法，有效训练轮数是实际轮数的四倍。

TABLE 6
将所提出的HSSL 方法与多种架构和框架相结合的性能对比。我们分别在完全微调、线性探测和k-最近邻评估协议下，报告了ImageNet-1K验证集上的Top-1准确率 [13]。†表示采用多裁剪策略 [2]，包含2个224×224全局裁剪和10个96×96局部裁剪。

	模型结构	Epochs[2]	Fine-tuning	Linear	k-NN
MAE [5]	ViT-S/16	400	80.4	-	-
MAE [5]+HSSL			80.8	-	-
MoCo [24]	ViT-S/16	200	-	65.3	57.4
MoCo [24]+HSSL			-	65.7	58.1
DINO [2]	ViT-S/16	200	-	67.9	61.2
DINO [2]+HSSL			-	70.8	65.5
iBOT [10]	ViT-S/16	200	-	71.3	65.2
iBOT [10]+HSSL			-	72.6	67.3
DINO [2]†	ViT-S/16	400	-	74.6	70.9
DINO [2]†+HSSL			-	75.7	72.5
iBOT [10]†	ViT-S/16	400	80.9	74.4	71.5
iBOT [10]†+HSSL			81.3	76.5	72.8
AttMask [74]†	ViT-S/16	400	-	76.1	72.8
AttMask [74]†+HSSL			-	76.7	73.1
iBOT [10]†	ViT-B/16	400	83.3	77.8	74.0
iBOT [10]†+HSSL			83.8	79.4	75.3
iBOT [10]†	ViT-B/16	1600	84.0	79.5	77.1
iBOT [10]†+HSSL		600	84.1	79.6	76.0
MFF [81]	ViT-B/16	300	83.3	-	-
MFF [81]+HSSL			83.6	-	-
DINO [2]	Swin-T	200	-	69.2	60.1
DINO [2]+HSSL			-	71.2	65.1
DINO [2]	PVT-Small	200	-	67.7	61.3
DINO [2]+HSSL			-	69.6	64.4

TABLE 7
基于ViT-B/16架构的现有方法对比分析 [109]。†表示使用了预训练感知码本进行标记化处理。

	Epochs[2]	Fine-tuning	Linear
MoCo [24]	600	83.2	76.7
DINO [2]	1600	83.6	78.2
SimMIM [41]	800	83.8	56.7
MAE [5]	1600	83.6	68.0
iBOT [10]	400	83.3	77.8
MaskFeat [83]	1600	84.0	-
BootMAE [94]	800	84.2	66.1
SdAE [90]	300	84.1	64.9
BEiT [40]	800	83.2	56.7
SiameseIM [88]	400	83.7	76.8
MOKD [122]	400	-	78.0
LocalMIM [84]	1600	84.0	-
MFF [81]	800	83.6	-
CIM [45]	300	83.3	-
ConMIM [87]	800	83.7	39.3
ccMIM [101]	300	83.6	66.9
ccMIM [101]	800	84.2	68.9
PeCo [92]†	300	84.1	-
SERE [8]	400	83.7	77.9
iBOT [10]+HSSL	400	83.8	79.4
iBOT [10]+HSSL	600	84.1	79.6
MFF [81]+HSSL	300	83.6	-

TABLE 8
语义分割任务对比实验。我们在ADE20K [127]数据集上使用ViT-B/16 [109]作为主干网络对UperNet [128]进行微调，实验设置遵循现有工作 [5], [10]。

	模型结构	Epochs[2]	mIoU
MoCo [24]	ViT-B/16	600	47.2
DINO [2]		1600	46.8
MAE [5]		1600	48.1
BootMAE [94]		800	49.1
SdAE [90]		300	48.6
BEiT [40]‡		800	45.6
SiameseIM [88]		400	49.6
MixedAE [129]		800	48.7
LocalMIM [84]		1600	49.5
MFF [81]		800	48.6
ConMIM [87]		800	46.0
ccMIM [101]		800	47.7
PeCo [92]†		300	48.5
SERE [8]		800	50.0
iBOT [10]	ViT-B/16	400	47.9
iBOT [10]		1600	50.0
iBOT [10]+HSSL		400	50.3
iBOT [10]	ViT-S/16	400	45.2
iBOT [10]		3200	45.4
iBOT [10]+HSSL		400	46.1

[24] ），以及将实例判别与MIM相结合的混合方法（如iBOT [10] 和AttMask [74] ）在内的多种方法。例如，当对ViT-B/16进行400轮预训练时，HSSL 将iBOT的线性探测准确率提高了1.6%。同时，HSSL 甚至可以用更少的训练轮数（600轮 vs. 1600轮）获得与iBOT相当的性能。这些结果表明，HSSL 增强了分类能力，并且与现有的表征学习方法正交。此外，Tab. 6强调HSSL 能够增强不同的Transformer架构，其改进范围超越了原始视觉Transformer [109]。例如，经过200轮预训练后，HSSL 分别将Swin-T [12]和PVT-Small [133]的线性探测准确率提高了2.0%和1.9%。

最后，我们直接将HSSL 与现有方法进行对比，如Tab. 7所示。相较于基于实例判别的方法，本方法经过600轮预训练后，在全微调和线性探测任务上分别达到84.1%和79.6%的Top-1准确率，显著优于需要1600轮预训练才能达到79.5%线性探测准确率的iBOT [10] 等方法。即使仅进行400轮的较短预训练，HSSL 在全微调和线性探测任务上仍分别以0.5%和1.6%的优势超越iBOT。尽管部分基于掩码建模的方法（如ccMIM [101]和BootMAE [94]）在微调任务中表现出竞争力，但其线性探测准确率较低，且如Tab. 8所示，在下游任务中的有效性可能受限。相比之下，HSSL 在线性探测、微调和下游任务中均展现出卓越性能。

TABLE 9
基于ViT-B/16的目标检测与实例分割性能对比 我们在COCO数据集 [16]上对模型进行微调 其中AP^m表示分割掩码平均精度 AP^b表示边界框平均精度

	模型结构	Epochs[2]	AP^m	AP^b
DINO [2]	ViT-B/16	1600	43.4	50.1
MAE [5]		1600	44.3	51.3
SERE [8]		400	43.8	50.7
MFF [81]	ViT-B/16	300	43.2	50.0
MFF [81]+HSSL		300	43.7	50.5
iBOT [10]	ViT-B/16	400	43.2	50.1
iBOT [10]+HSSL		400	44.0	51.0
iBOT [10]	ViT-B/16	1600	44.2	51.2
iBOT [10]+HSSL		600	44.3	51.4

TABLE 10
RAW图像目标检测中的跨域迁移学习。模型在AODRaw [130]数据集上进行微调，该数据集专门收集用于目标检测的RAW图像。除平均精度（AP）[16]外，我们还报告了在0.75和0.50交并比（IoU）阈值下的AP_{75}和AP_{50}指标。AP_s、AP_m和AP_l分别表示小、中、大尺寸目标的平均精度。

	Architecture	Epochs[2]	AP	AP_{50}	AP_{75}
DINO [2]	Swin-T [12]	200	28.9	45.7	30.2
DINO [2]+HSSL		200	29.5	46.5	30.8

TABLE 11
在更多图像分类基准上的迁移学习，包括CIFAR [131] 和 iNaturalist [132]。

	模型结构	Epochs[2]	$Cifar_{100}$	$INat_{18}$	$INat_{19}$
iBOT [10]	ViT-B/16	400	92.1	74.0	78.4
iBOT [10]		1600	92.2	74.6	79.6
iBOT [10]+HSSL		400	92.2	75.2	79.7

图像分类中的迁移学习。除了ImageNet-1K之外，我们还将预训练的基础模型迁移到其他分类数据集上，包括CIFAR [131]和iNaturalist [132]。如Tab. 11所示，HSSL 在不同数据集上均带来了一致的性能提升，展现了卓越的迁移能力。

语义分割中的迁移学习。我们采用UperNet [128]作为分割模型来评估语义分割性能。如先前工作 [10]所述，我们在ADE20K数据集 [127]上对模型进行了微调。如Tab. 8所示，当ViT-B/16模型经过400轮预训练后，HSSL 取得了50.3%的mIoU指标。值得注意的是，HSSL 的表现优于iBOT [10] 方法，后者需要1600轮预训练才能达到相近的结果。当使用经过400轮预训练的ViT-S/16模型时，HSSL 同样以0.9%的mIoU优势领先于iBOT [10] 。这些结果充分证明了HSSL 在密集预测任务中的有效性。

实例分割的迁移学习。如 [10]所述，我们采用Cascade Mask-RCNN [134] 实现实例分割与目标检测任务。

TABLE 12
ImageNet-1K数据集上的半监督分类任务 [13]。我们采用线性分类器和k-近邻分类器，分别在1%/10%标注比例下报告Top-1准确率。

	模型结构	Epochs[2]	1%	10%
iBOT [10]	ViT-B/16	400	64.8	76.3
iBOT [10]+HSSL		400	66.1	76.8

TABLE 13
ImageNet-S [25]上的半监督语义分割。我们报告了验证集和测试集上的平均交并比(mIoU)。PT表示使用预训练权重初始化模型，FT表示分别使用完全微调后的权重初始化模型。

	模型结构	Epochs[2]	ImageNet-S$_{\mathrm{PT}}$		ImageNet-S$_{\mathrm{FT}}$	
			val	test	val	test
iBOT [10]	ViT-B/16	400	48.3	47.8	62.6	63.0
iBOT [10]		1600	50.5	50.1	-	-
iBOT [10]+HSSL		400	51.5	51.1	63.5	63.1

如Tab. 9所示，在仅进行400轮预训练的情况下，HSSL相较iBOT [10]在AP^{m}和AP^{b}指标上分别提升0.8%与0.9%。相较于MFF [81]方法，本方案在两项指标上均实现0.5%的性能提升。值得注意的是，HSSL同时显著降低训练成本，将所需预训练轮次从iBOT [10]的1600轮缩减至600轮即可获得更优性能。

跨域迁移评估。我们在AODRaw [130]数据集上进一步评估预训练模型性能，该数据集专为RAW域目标检测任务设计。RAW域与模型预训练的sRGB域存在显著域间差异。实验表明，当Swin-T主干网络进行200轮预训练时，HSSL的AP指标提升0.6%，充分验证其卓越的跨域泛化能力。

半监督学习。收集标注需要耗费巨大成本。半监督学习能够减少对昂贵标注的需求。因此，我们同时评估了HSSL在半监督分类和语义分割任务中的表现。对于半监督分类任务，我们遵循 [10]中的范式，使用部分标签对预训练基础模型进行微调。如Tab. 12所示，在使用1%和10%训练标签时，HSSL的Top-1准确率分别比iBOT [10]提高了1.3%和0.5%。在半监督语义分割任务中，我们在ImageNet-S数据集 [25]上对基础模型进行微调，该数据集包含919个类别和9190张标注图像。Tab. 13展示了验证集和测试集上的mIoU指标。可以观察到，HSSL在验证集和测试集上分别以4.7%和4.2%的mIoU显著优于iBOT [10]。

无监督语义分割。我们通过无监督语义分割任务评估预训练基础模型。训练阶段采用 [25]提出的流程，并在三个数据集 [25]上进行测试，即ImageNet-S$_{50}$、ImageNet-S$_{300}$和ImageNet-S数据集。如Tab. 14所示，在ImageNet-S数据集上，HSSL以1.8%的mIoU优势超越iBOT。半监督和无监督学习的结果表明，HSSL在缺乏标签的情况下仍能有效提升感知与识别性能。

时间与内存占用。Tab. 15展示了iBOT [10]与我们的HSSL所需的时间及内存占用。相较于基线方法，HSSL仅带来可忽略的计算开销增长，这是因为基础模型与辅助头之间的串行连接结构使得辅助头仅需少量层数即可提取有效表征。

5 消融实验与分析

为节省计算成本，我们在ImageNet-S$_{300}$数据集上进行了100个实际训练周期的模型预训练消融研究。默认情况下，我们将辅助头的深度设置为1。通过报告ImageNet上的k近邻分类准确率（Cls.）和ImageNet-S上的分割平均交并比（Seg.）来评估模型性能。

TABLE 14
ImageNet-S数据集上的无监督语义分割 [25]。919/300/50分别表示ImageNet-S/ImageNet-S$_{300}$/ImageNet-S$_{50}$数据集。我们遵循 [25]提出的流程和设置，并报告验证集和测试集上的mIoU指标。本实验中未采用多裁剪策略进行表征学习。

	数据集	模型结构	Epochs[2]	val	test
iBOT [10]	50	ViT-S/16	400	46.2	45.1
iBOT [10]+HSSL				54.4	54.5
iBOT [10]	300	ViT-S/16	200	22.2	22.4
iBOT [10]+HSSL				26.6	26.0
iBOT [10]	919	ViT-S/16	200	12.2	11.3
iBOT [10]+HSSL				14.0	13.6

TABLE 15
在配备8块GPU的机器上进行预训练时的时间与内存占用情况，批处理大小为256，并采用10个96×96的多尺度裁剪。

	模型结构	Epochs[2]	时间 (h)	空间 (G)
iBOT [10]	ViT-B/16	400	82.7	18.3
iBOT [10]+HSSL			94.5	21.4

TABLE 16
基础模型监督方式的消融实验。**B**和**A**分别表示基础模型和辅助头。**A**→**B**表示辅助头监督基础模型。**B**→**B**表示基础模型自我监督。

	Seg.	Cls.	
	mIoU	Top-1	Top-5
A→A	16.1	26.5	48.0
A→A + B→B	31.4	68.0	86.4
A→A + **A**→**B**	36.9	72.7	87.6

TABLE 17
辅助头结构设计的消融实验。

	Seg.	Cls.	
	mIoU	Top-1	Top-5
MLP	35.8	70.0	86.3
Token Mixer	36.3	70.1	86.4
MLP + Token Mixer	36.9	72.7	87.6

TABLE 18
共享投影与非共享投影的消融实验。

Shared proj.	Seg.	Cls.	
	mIoU	Top-1	Top-5
✔	35.8	72.3	87.5
✘	36.9	72.7	87.6

监督方式的影响。在连接辅助头之后，我们研究了是使用基础模型本身（同质监督）还是辅助头（异质监督）来指导基础模型。如Tab. 16所示，异质监督方式优于同质监督方式，实现了5.5%的mIoU提升和4.7%的Top-1准确率提升。这些结果验证了异质监督的重要性，它使得基础模型能够从辅助头学习互补特征。

辅助头的结构。我们采用统一框架来设计不同的辅助头，其中包含一个令牌混合器和一个多层感知机（MLP）模块。本文以ConvNext为例，评估了令牌混合器与MLP模块的作用效果。如Tab. 17所示，令牌混合器发挥着更关键的作

TABLE 19
辅助头并行与串行连接的消融实验。我们采用深度为3的串行连接方式。表中展示了相对于基线的时间与内存开销倍数。

	Seg.	Cls.		Computation cost	
	mIoU	Top-1	Top-5	time	memory
baseline	29.3	67.5	84.4	×1.00	×1.00
parallel	34.6	72.8	87.2	×2.53	×2.25
serial	37.1	73.9	88.4	×1.09	×1.12

TABLE 20
在使用ViT时利用不同粒度的异构自监督，以iBOT [10] 为例。

图像级	Patch级	Seg.	Cls.	
		mIoU	Top-1	Top-5
✘	✘	42.3	75.1	89.3
✔	✘	46.2	75.8	89.4
✔	✔	46.7	76.0	89.5

TABLE 21
与由深至浅策略（DTS） [25]的对比

	Seg.	Cls.	
	mIoU	Top-1	Top-5
baseline	29.3	67.5	84.4
+DTS	30.5	68.6	85.4
+HSSL	36.9	72.7	87.6

用， 相比MLP模块可带来0.6% mIoU与2.6% Top-1准确率的提升。

是否共享投影层。在计算损失函数之前，自监督学习方法通常需要通过投影头处理教师模型/学生模型的表征。Tab. 18探究了基础模型与辅助头之间是否应该共享投影层的问题。实验结果表明，不共享投影层可带来1.1% mIoU与0.4% Top-1准确率的优势。由于二者架构存在差异，基础模型与辅助头的表征会出现偏差，不共享投影层能为处理这种偏差提供更大灵活性。

辅助头的并行或串联连接方式。我们可以将辅助头与基础模型以串联或并联的方式连接。对于并联连接，我们使用完整的ConvNext-Tiny作为辅助头，直接接收图像作为输入。相比之下，串联连接仅需少量层数即可让辅助头提取丰富信息。如Tab. 19所示，并联结构的训练耗时约为串联连接的2.32倍。此外，采用串联连接不仅能获得优于并联结构的性能表现，还能实现更高的计算效率。

类别标记与补丁标记。部分方法 [8], [10], [74] 通过同时计算不同粒度层级的损失来实现优化。以同时考虑图像级和Patch级损失的iBOT [10] 为例，Tab. 20 展示了在不同粒度应用HSSL的效果差异。需要特别说明的是，当仅在图像级应用HSSL时，像素级自监督仅作用于教师模型与学生模型的基础架构之间。实验表明，基础模型通过单纯的图像级监督即可从辅助任务中学习到大部分有效信息。与此同时，引入像素级监督训练能带来额外的性能提升。这些结果表明，通过仅在图像级应用HSSL，我们能够显著节省计算资源开销。

与深度监督浅层方法的对比。深度监督浅层方法通过同构架构中深层对浅层的监督来增强浅层特征表示。如Tab. 21所示，该策略在ViT中仅带来轻微性能提升，这可能是因为ViT的深层与浅层特征高度相似 [135]，导致监督信号缺乏多样性。相比之下，异构自监督学习促使ViT学习多样化知识，在mIoU和Top-1准确率上分别较DTS方法显著提升了6.4%和4.1%。

6 结论

本文提出异构自监督学习方法 (HSSL)。具体而言，我们强制基础模型从与其架构异构的辅助头中学习，从而弥补基础模型自身缺失的特性。此外，我们发现基础模型与辅助头之间的差异程度与HSSL方法带来的性能提升呈正相关。这一正相关关系促使我们提出高效的搜索策略，可为特定模型寻找最合适的辅助头，并设计了若干简单有效的方案来扩大模型差异。实验表明HSSL与不同自监督学习方法具有正交性，在图像分类、语义分割、目标检测和实例分割等多种下游任务上均能提升模型性能。

局限性与未来工作。HSSL已成功与多种自监督学习方法集成以提升性能。然而，某些方法（例如 [66]中提出的方案）直接采用冻结预训练模型提取的表征作为学习目标，这类方法的适配过程仍存在挑战。未来工作可聚焦于开发更具普适性或专用性的方法，以更广泛地实现异构表征学习与各类方法的有效融合。此外，HSSL采用不同模型输出概率分布间的KL散度来衡量模型差异。虽然该指标特别适用于基于聚类的自监督学习，但对于输出表征的其他自监督学习方法，可探索采用CKA相似度 [136]等替代性指标进行差异评估。在搜索策略方面，未来研究还可致力于设计更精准高效的搜索策略，使其能够适配包括复杂结构模型在内的更广泛模型类型，从而进一步拓展方法的适用性与可扩展性。

REFERENCES

[1] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *ECCV*, 2018.
[2] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *IEEE ICCV*, 2021.
[3] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *ECCV*, 2020.
[4] J. Li, P. Zhou, C. Xiong, and S. C. Hoi, "Prototypical contrastive learning of unsupervised representations," in *ICLR*, 2021.
[5] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *IEEE CVPR*, 2022.
[6] Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, and J. Feng, "Contrastive masked autoencoders are stronger vision learners," *IEEE TPAMI*, vol. 46, no. 4, pp. 2506–2517, 2024.
[7] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *ICML*, 2022.
[8] Z.-Y. Li, S. Gao, and M.-M. Cheng, "Exploring feature self-relation for self-supervised transformer," *IEEE TPAMI*, 2022.
[9] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *NeurIPS*, 2020.
[10] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "ibot: Image bert pre-training with online tokenizer," in *ICLR*, 2022.
[11] 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 *NeurIPS*, 2020.
[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," *IEEE ICCV*, 2021.
[13] 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," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.
[14] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, pp. 303–338, 2009.

[15] 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 *IEEE ICCV*, 2021.

[16] 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 *ECCV*, 2014.

[17] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *IEEE ICCV*, 2019.

[18] H. Yan, Z. Li, W. Li, C. Wang, M. Wu, and C. Zhang, "Contnet: Why not use convolution and transformer at the same time?" *arXiv preprint arXiv:2104.13497*, 2021.

[19] 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 *NeurIPS*, 2021.

[20] B. Yin, X. Zhang, Z. Li, L. Liu, M.-M. Cheng, and Q. Hou, "Dformer: Rethinking rgbd representation learning for semantic segmentation," *arXiv preprint arXiv:2309.09668*, 2023.

[21] S. D'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in *ICML*, 2021.

[22] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*, 2020.

[23] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," in *IEEE CVPR*, 2022.

[24] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *IEEE ICCV*, 2021.

[25] S. Gao, Z.-Y. Li, M.-H. Yang, M.-M. Cheng, J. Han, and P. Torr, "Large-scale unsupervised semantic segmentation," *IEEE TPAMI*, 2022.

[26] B. Sun, Y. Yang, W. Yuan, L. Zhang, M.-M. Cheng, and Q. Hou, "Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation," *arXiv preprint arXiv:2306.04300*, 2023.

[27] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*, 2016.

[28] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *IEEE CVPR*, 2017.

[29] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, 2016.

[30] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *ICLR*, 2018.

[31] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *IEEE ICCV*, 2015.

[32] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *ICML*, 2008.

[33] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *IEEE CVPR*, 2016.

[34] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in *IEEE ICCV*, 2017.

[35] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *IEEE CVPR*, 2018.

[36] Y. Zhao, G. Wang, C. Luo, W. Zeng, and Z.-J. Zha, "Self-supervised visual representations learning by contrastive mask prediction," in *IEEE ICCV*, 2021.

[37] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations," in *IEEE ICCV*, 2021.

[38] S. A. Koohpayegani, A. Tejankar, and H. Pirsiavash, "Mean shift for self-supervised learning," in *IEEE ICCV*, 2021.

[39] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," *NeurIPS*, 2014.

[40] H. Bao, L. Dong, S. Piao, and F. Wei, "BEit: BERT pre-training of image transformers," in *ICLR*, 2022.

[41] 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 CVPR*, June 2022, pp. 9653–9663.

[42] J. Liu, X. Huang, J. Zheng, Y. Liu, and H. Li, "Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers," in *IEEE CVPR*, 2023.

[43] J. Chen, M. Hu, B. Li, and M. Elhoseiny, "Efficient self-supervised vision pretraining with local masked reconstruction," *arXiv preprint arXiv:2206.00790*, 2022.

[44] W. Li, J. Xie, and C. C. Loy, "Correlational image modeling for self-supervised visual pre-training," in *IEEE CVPR*, 2023.

[45] Y. Fang, L. Dong, H. Bao, X. Wang, and F. Wei, "Corrupted image modeling for self-supervised visual pre-training," in *ICLR*, 2023.

[46] B. Roh, W. Shin, I. Kim, and S. Kim, "Spatilly consistent representation learning," in *IEEE CVPR*, 2021.

[47] 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 *IEEE ICCV*, 2021.

[48] 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.

[49] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *ICML*, 2016.

[50] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *IEEE CVPR*, 2020.

[51] C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, T.-L. Liu, Y. Chen, and Y. LeCun, "Decoupled contrastive learning," in *ECCV*, 2022.

[52] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[53] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *IEEE CVPR*, 2021.

[54] 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 CVPR*, 2021.

[55] Y. Tian, X. Chen, and S. Ganguli, "Understanding self-supervised learning dynamics without contrastive pairs," in *ICML*, 2020.

[56] X. Chen and K. He, "Exploring simple siamese representation learning," in *IEEE CVPR*, 2021.

[57] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, "Whitening for self-supervised representation learning," in *ICML*, 2021.

[58] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, and C. C. Loy, "Online deep clustering for unsupervised representation learning," in *IEEE CVPR*, 2020.

[59] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *IEEE CVPR*, 2016.

[60] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *arXiv:2304.07193*, 2023.

[61] K. Zhu, M. Fu, and J. Wu, "Multi-label self-supervised learning with scene images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[62] M. Caron, P. Bojanowski, J. Mairal, and A. Joulin, "Unsupervised pre-training of image features on non-curated data," in *IEEE ICCV*, 2019.

[63] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *ICML*, 2021.

[64] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," in *ICLR*, 2022.

[65] N. Shvetsova, F. Petersen, A. Kukleva, B. Schiele, and H. Kuehne, "Learning by sorting: Self-supervised learning with group ordering constraints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 16 453–16 463.

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

[67] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "Self-labelling via simultaneous clustering and representation learning," in *ICLR*, 2020.

[68] L. Wang, F. Liang, Y. Li, H. Zhang, W. Ouyang, and J. Shao, "Repre: Improving self-supervised vision transformer with reconstructive pre-training," in *IJCAI*, 7 2022, pp. 1437–1443.

[69] K. Song, S. Zhang, Z. Luo, T. Wang, and J. Xie, "Semantics-consistent feature search for self-supervised visual representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[70] W. Zhang, J. Pang, K. Chen, and C. C. Loy, "Dense siamese network for dense unsupervised learning," in *ECCV*, 2022.

[71] P. O. O Pinheiro, A. Almahairi, R. Benmalek, F. Golemo, and A. C. Courville, "Unsupervised learning of dense visual representations," *NeurIPS*, 2020.

[72] J. Xie, X. Zhan, Z. Liu, Y. S. Ong, and C. C. Loy, "Unsupervised object-level representation learning from scene images," in *NeurIPS*, 2021.

[73] L. Huang, S. You, M. Zheng, F. Wang, C. Qian, and T. Yamasaki, "Learning where to learn in cross-view self-supervised learning," in *CVPR*, 2022.

[74] 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," in *ECCV*, 2022.

[75] 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 *NeurIPS*, 2021.

[76] 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 *ICLR*, 2022.

[77] H. Wang, K. Song, J. Fan, Y. Wang, J. Xie, and Z. Zhang, "Hard patches mining for masked image modeling," in *IEEE CVPR*, 2023.

[78] Z. Feng and S. Zhang, "Evolved part masking for self-supervised learning," in *IEEE CVPR*, 2023.

[79] G. Li, H. Zheng, D. Liu, C. Wang, B. Su, and C. Zheng, "Semmae: Semantic-guided masking for learning masked autoencoders," in *NeurIPS*, 2022.

[80] J. Xie, W. Li, X. Zhan, Z. Liu, Y. S. Ong, and C. C. Loy, "Masked frequency modeling for self-supervised visual pre-training," in *ICLR*, 2023.

[81] Y. Liu, S. Zhang, J. Chen, Z. Yu, K. Chen, and D. Lin, "Improving pixel-based mim by reducing wasted modeling capability," in *IEEE ICCV*, 2023.

[82] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE CVPR*, 2005.

[83] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *IEEE CVPR*, 2022.

[84] H. Wang, Y. Tang, Y. Wang, J. Guo, Z.-H. Deng, and K. Han, "Masked image modeling with local multi-scale reconstruction," in *IEEE CVPR*, 2023.

[85] H. Liu, X. Jiang, X. Li, A. Guo, Y. Hu, D. Jiang, and B. Ren, "The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training," in *AAAI*, 2023.

[86] H. Wang, J. Fan, Y. Wang, K. Song, T. Wang, and Z. Zhang, "Droppos: Pre-training vision transformers by reconstructing dropped positions," in *NeurIPS*, 2023.

[87] K. Yi, Y. Ge, X. Li, S. Yang, D. Li, J. Wu, Y. Shan, and X. Qie, "Masked image modeling with denoising contrast," *ICLR*, 2023.

[88] C. Tao, X. Zhu, W. Su, G. Huang, B. Li, J. Zhou, Y. Qiao, X. Wang, and J. Dai, "Siamese image modeling for self-supervised vision representation learning," in *IEEE CVPR*, 2023.

[89] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *IEEE CVPR*, 2023.

[90] Y. Chen, Y. Liu, D. Jiang, X. Zhang, W. Dai, H. Xiong, and Q. Tian, "Sdae: Self-distillated masked autoencoder," in *ECCV*, 2022.

[91] 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," *IJCV*, vol. 132, no. 1, pp. 208–223, 2024.

[92] X. Dong, J. Bao, T. Zhang, D. Chen, W. Zhang, L. Yuan, D. Chen, F. Wen, N. Yu, and B. Guo, "Peco: Perceptual codebook for bert pre-training of vision transformers," in *AAAI*, 2023.

[93] P. Gao, Z. Lin, R. Zhang, R. Fang, H. Li, H. Li, and Y. Qiao, "Mimic before reconstruct: Enhancing masked autoencoders with feature mimicking," *IJCV*, vol. 132, no. 5, pp. 1546–1556, 2024.

[94] X. Dong, J. Bao, T. Zhang, D. Chen, W. Zhang, L. Yuan, D. Chen, F. Wen, and N. Yu, "Bootstrapped masked autoencoders for vision bert pretraining," in *ECCV*, 2022.

[95] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *IEEE CVPR*, 2023.

[96] xingbin liu, J. Zhou, T. Kong, X. Lin, and R. Ji, "Exploring target representations for masked autoencoders," in *ICLR*, 2024.

[97] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas, "Masked siamese networks for label-efficient learning," in *ECCV*, 2022.

[98] Z. Wu, Z. Lai, X. Sun, and S. Lin, "Extreme masking for learning instance and distributed visual representations," *arXiv preprint arXiv:2206.04667*, 2022.

[99] Z. Jiang, Y. Chen, M. Liu, D. Chen, X. Dai, L. Yuan, Z. Liu, and Z. Wang, "Layer grafted pre-training: Bridging contrastive learning and masked image modeling for label-efficient representations," in *ICLR*, 2023.

[100] Y. Shi, N. Siddharth, P. H. Torr, and A. R. Kosiorek, "Adversarial masking for self-supervised learning," in *ICML*, 2022.

[101] S. Zhang, F. Zhu, R. Zhao, and J. Yan, "Contextual image masking modeling via synergized contrasting without view augmentation for faster and better visual pretraining," in *ICLR*, 2023.

[102] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *IEEE CVPR*, June 2023, pp. 16 133–16 142.

[103] L. Huang, S. You, M. Zheng, F. Wang, C. Qian, and T. Yamasaki, "Green hierarchical vision transformer for masked image modeling," in *NeurIPS*, 2022.

[104] P. Gao, T. Ma, H. Li, Z. Lin, J. Dai, and Y. Qiao, "MCMAE: Masked convolution meets masked autoencoders," in *NeurIPS*, 2022.

[105] X. Li, W. Wang, L. Yang, and J. Yang, "Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality," *arXiv:2205.10063*, 2022.

[106] K. Tian, Y. Jiang, Q. Diao, C. Lin, L. Wang, and Z. Yuan, "Designing bert for convolutional networks: Sparse and hierarchical masked modeling," in *ICLR*, 2023.

[107] Y. Tian, L. Xie, Z. Wang, L. Wei, X. Zhang, J. Jiao, Y. Wang, Q. Tian, and Q. Ye, "Integrally pre-trained transformer pyramid networks," in *IEEE CVPR*, 2023.

[108] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "A unified view of masked image modeling," *arXiv preprint arXiv:2210.10615*, 2022.

[109] 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," *ICLR*, 2021.

[110] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *IEEE CVPR*, 2022.

[111] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.

[112] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, A. Joulin, G. Synnaeve, J. Verbeek, and H. Jégou, "Resmlp: Feedforward networks for image classification with data-efficient training," *IEEE TPAMI*, vol. 45, no. 4, pp. 5314–5321, 2023.

[113] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *IEEE CVPR*, 2022.

[114] C. Yang, Y. Wang, J. Zhang, H. Zhang, Z. Wei, Z. Lin, and A. Yuille, "Lite vision transformer with enhanced self-attention," in *IEEE CVPR*, 2022.

[115] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *IEEE CVPR*, June 2019.

[116] F. Yuan, Z. Zhang, and Z. Fang, "An effective cnn and transformer complementary network for medical image segmentation," *PR*, vol. 136, p. 109228, 2023.

[117] B. Yin, X. Zhang, Q. Hou, B.-Y. Sun, D.-P. Fan, and L. Van Gool, "Camoformer: Masked separable attention for camouflaged object detection," *arXiv preprint arXiv:2212.06570*, 2022.

[118] Y. Li, H. Mao, R. B. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *ECCV*, 2022, pp. 280–296.

[119] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," in *ICLR*, 2022.

[120] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *IEEE ICCV*, 2021.

[121] S. Lao, Y. Gong, S. Shi, S. Yang, T. Wu, J. Wang, W. Xia, and Y. Yang, "Attentions help cnns see better: Attention-based hybrid image quality assessment network," in *CVPRW*, 2022.

[122] K. Song, J. Xie, S. Zhang, and Z. Luo, "Multi-mode online knowledge distillation for self-supervised visual representation learning," in *IEEE CVPR*, 2023.

[123] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE CVPR*, 2020.

[124] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[125] Y. Li, Z.-Y. Li, Q.-S. Zeng, Q. Hou, and M.-M. Cheng, "Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation," in *ICML*, 2024.

[126] Z. Xie, Z. Geng, J. Hu, Z. Zhang, H. Hu, and Y. Cao, "Revealing the dark secrets of masked image modeling," in *IEEE CVPR*, 2023.

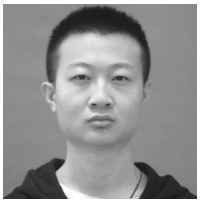
[127] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *IEEE CVPR*, 2017.

[128] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *ECCV*, September 2018.

[129] K. Chen, Z. Liu, L. Hong, H. Xu, Z. Li, and D.-Y. Yeung, "Mixed autoencoder for self-supervised visual representation learning," in *IEEE CVPR*, 2023.

[130] Z.-Y. Li, X. Jin, B. Sun, C.-L. Guo, and M.-M. Cheng, "Towards raw object detection in diverse conditions," *arXiv preprint arXiv:2411.15678*, 2024.

[131] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep. 0, 2009.
[132] 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 CVPR*, June 2018.
[133] 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 *IEEE ICCV*, 2021, pp. 568–578.
[134] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *IEEE CVPR*, June 2018.
[135] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *NeurIPS*, 2021.
[136] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *ICML*, 2019.



李钟毓 南开大学计算机学院博士研究生，师从程明明教授。研究方向包括深度学习、表征学习与计算机视觉。

尹博文 南开大学计算机学院博士研究生，侯淇彬教授课题组成员。研究方向为计算机视觉与多模态场景感知。

刘永祥 2004年获国防科技大学信息与通信工程专业博士学位，现为国防科技大学电子科学与技术学院教授。主要研究方向包括遥感图像分析、雷达信号处理、合成孔径雷达(SAR)目标识别、逆合成孔径雷达(ISAR)成像及机器学习。

刘丽 (Senior Member, IEEE) 2012年获国防科技大学博士学位，现为国防科技大学电子科学与技术学院教授。其论文在Google Scholar被引用逾11,000次，研究方向涵盖计算机视觉、机器学习、可信人工智能及合成孔径雷达。刘教授曾任IEEE TPAMI和IJCV客座编辑，现任IEEE TGRS、IEEE TCSVT等期刊副主编。

程明明 (Senior Member, IEEE) 2012年获清华大学博士学位，后于牛津大学Philip Torr教授课题组从事研究。2016年起任南开大学教授并创建媒体计算实验室。研究方向为计算机视觉与计算机图形学。曾获ACM中国新星奖、IBM全球杰出研究奖等荣誉，现任IEEE TPAMI和IEEE TIP编委。