面向开放集识别的类独有语义重构

黄宏志，王煜，胡清华，程明明

本文摘要—开放集识别（Open Set Recognition，OSR）使深度神经网络（Deep Neural Networks，DNN）能够识别未知类别的样本，同时对已知类别的样本保持较高的分类精度。现有基于自动编码器（Autoencoder，AE）和原型学习的方法在处理这一具有挑战性的任务方面有着巨大的潜力。本文结合了自动编码器和原型学习两类方法的优势，提出了一种类独有语义重构（Class-Specific Semantic Reconstruction，CSSR）的新方法。具体地，CSSR 用类别独有的自动编码器所表示的流形取代了原型点。与传统的基于原型的方法不同，CSSR 在深度神经网络骨干的顶部为每个类别插入类别独有的自动编码器，在每个单独的自动编码器流形上对每个已知的类别进行建模，并通过自动编码器的重构误差度量类别的隶属度。同时，与通过自动编码器去重构原始图像不同，本文提出使用自动编码器去重构深度神经网络所学习的语义特征表示。通过端到端的学习，深度神经网络和自动编码器相互促进，从而学习到有区分性和有代表性的特征信息。多个数据集上进行的实验结果表明，本文所提出的方法在封闭集和开放集的识别中都取得了出色的表现。本文的方法非常简单，可以灵活地纳入到现有的主流框架中。本工作的代码发布在 https://github.com/xyzedd/CSSR 上。

关键词—分类，开放集识别，自动编码器，原型学习，类独有语义重构

1 背景介绍

传统的深度神经网络（Deep Neural Networks，DNNs）是基于封闭集假设进行训练的，即测试中出现的样本类别在训练过程中都已经被模型学习过。在现实世界的应用中，测试样本可能来自未知的类别 [41]。当遇到这样的未知样本时，传统深度神经网络会强制性地将其分类为某个已知类，这种错误的预测可能会在如医疗诊断和自动驾驶等某些关键场景下导致不可挽回的损失。

开放集识别（Open Set Recognition，OSR）通过使模型对已知类（即封闭集）的样本进行正确分类，并准确识别那些未知类（即开放集） [9] 来解决上述问题。开放集识别的主要挑战是在训练过程中没有未知类的信息，因此很难区分已知类和未知类（即减少开放空间风险） [32]。传统深度神经网络主要学习已知类的判别特征，并建模整个特征空间的分类面。而这会导致一个严重的问题：未知类的样本仍然位于某些特定的区域，因而会被以较高的置信度识别为已知类 [40]。因此，以前的许多工作都提出要学习已知类的紧致特征，以便模型能够分离封闭集和开放集空间。在这些方法中，基于自动编码器（Autoencoder，AE）的方法 [24], [25], [34], [35], [41] 和类原型（Prototype-like）的方法 [3], [4], [40] 是目前最具有竞争力的。

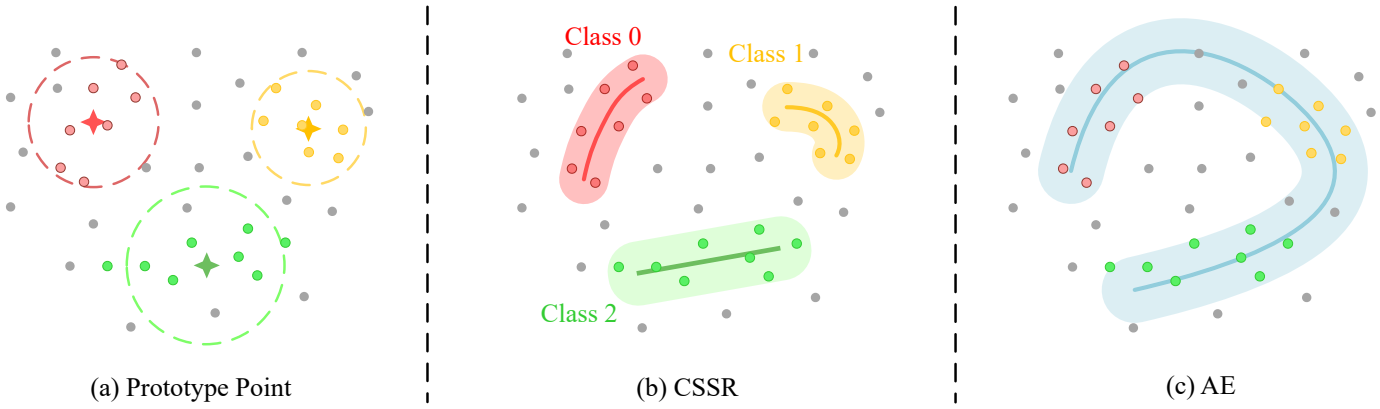
- 作者黄宏志、王煜和胡清华隶属于中国天津市天津大学智能与计算学部（邮编 *300350*）以及天津市机器学习重点实验室（邮编 *300350*）。王煜和胡清华同时隶属于信创海河实验室。程明明隶属于中国天津市南开大学计算机学院 *TKLNDST* 实验室（邮编 *300350*）。
- 本文的通讯作者是王煜和胡清华 {*wang.yu, huqinghua*}*@tju.edu.cn*。

- 这项工作得到了国家重点研发计划 *2018AAA0100400* 的部分支持、国家自然科学基金 *62106174*、*61732011* 和 *61925602* 的部分支持，以及中国博士后科学基金会 *2021TQ0242* 和 *2021M690118* 的部分支持。

正如图 1 (c) 中所述，基于自动编码器的方法通过重构原始输入图像来保留图像的大部分信息，从而能够学习到相关的隐空间特征。由于自动编码器可以在训练中学习重构已知类的图像，来自未知类的测试图像将导致高的重构误差，因此，可以用这种性质来识别未知类 [24], [34]。这可以被认为是在学习一个低维流型来适应已知样本的分布。为了对已知类进行分类，这些方法通过对原始图像进行像素级重构从而获得相对应隐空间特征来学习一个分类器。然而，这些方法仍然存在两个问题：（一）分类退化以及（二）开放空间风险的引入。

分类退化问题指的是使用由自动编码器学习到隐空间特征会损害分类器在封闭集分类的性能。这主要是因为一些不必要的分类信息（如背景信息）被保留下来，干扰了分类器对已知类识别的学习 [37]。开放空间风险引入指的是为适应已知样本而学习的连续流形可能会吞噬类间区域，从而引入开放空间风险（Open Space Risk）。引入开放空间风险的示意图如图 1 (c) 所示。

与基于自动编码器的方法不同，类原型的方法，如包括广义卷积原型学习 [40] 和最近提出的互换点学习 [4] 分别学习特定类的点来拟合从标记类或剩余类中提取出来的特征，从直觉上给出了一些建模方法。然而，原型学习框架在开放集识别任务上仍然面临巨大的挑战。主要的挑战是**类的代表性不足问题**，即只用一个点或很少的点不能充分地代表类。一方面，原型学习假设类的具体特征是高斯分布 [40]。然而，这在现实世界中是很少可以满足的，而这将引入开放空间的风险。另一方面，在原型学习框架中，类内特征被压缩到相当有限的点上。而这可能导致模型过滤掉某些有助于区分未知类的必要信息 [6]。



(a) Prototype Point (b) CSSR (c) AE

图 1. 使用原型类方法（a）、本文提出方法 CSSR 模型（b）和基于自动编码器（c）的开放集方法的比较。原型学习用一个或几个原型点对每个类进行建模，它的拟合能力有限，学习的是极端紧凑的特征表示。自动编码器用一个连续的低维流形对已知样本进行建模，但一些类间的开放区域被吞噬并被错误地分类到一个已知的空间。CSSR 模型在一个由自动编码器代表的单独流型上为每个已知类建模。它能很好地适应已知样本，同时释放出被单个自动编码器吞噬的区域。

为了解决上述基于自动编码器方法和类原型方法所遇到的问题，本文提出一种新方法类独有语义重构（Class-Specific Semantic Reconstruction，CSSR）以充分考虑了以上难题。具体来说，CSSR 模型通过用一个类别特定的自动编码器流型对每个已知的类进行建模从而减少开放空间的风险。在 CSSR 模型中，样本的特征是用深度神经网络提取的。然后，为每个已知类指定一个单独的自动编码器，将不同的类别投影到不同的流形上。自动编码器被插入到深度神经网络骨干的顶部用来重构语义特征，而非传统做法中对原始图像的重构。由于自动编码器流型是特定类的可学习特征，因此重构误差（也就是点到流型的距离）可当作 Logits 用于分类。在实现过程中，CSSR 模型最小化了交叉熵损失，其中代表标记类的自动编码器的重构误差被优化到最小。模型框架示意图请参考图 1 (b)。

本文提出的框架可以帮助解决现有方法中存在上述问题。与基于自动编码器的方法相比，我们提出的 CSSR 模型（1）通过丢弃不必要的信息和重构语义特征（而非原始图像）来解决分类退化的问题；（2）通过学习特定类的流形来释放了被吞噬的类间区域，从而解决开放空间风险问题。与类原型的方法相比，所提出的 CSSR 模型通过学习特定类的流形很好地处理了类代表不足的问题。这不仅打破了类的高斯假设，而且比使用单个特征点表示类别更能保留类的关键信息。

通过端到端的学习过程，类独有自动编码器和深度神经网络能够实现相互促进，在学习与类别高度相关的语义特征的同时，也可以识别开放空间。自动编码器倾向于将每个类与语义特征的子集联系起来，已知类的样本倾向于激活与其相关的特征，而不激活与其不相关的特征；对于未知类的样本，它们的语义特征则不会被激活，因为它们与已知类的任何特征都不相关，这一特性也被用于检测未知类。在多种数据集上进行的实验结果表明，所提出的方法明显优于其他目前最好的方法，并同时提高了封闭集和开放集识别的性能。

综上所述，本文有以下贡献：

1) 本文提出一个简单而有效的 CSSR 模型用于开放集识别。它为每个已知的类指定了一个单独的自动编码器，并将这些自动编码器插入到深度神经网络骨干网络的顶部，从而可以重构主干网络所学到的语义特征。CSSR 模型改善了拟合和特征学习能力，从而提高了开放集的性能。
2) 本文从理论上分析并解释了现有方法产生的开放空间风险的问题，并讨论了 CSSR 模型与现有方法之间的联系以全面理解 CSSR 模型。
3) 本文在多种设置下进行了实验，实验结果表明CSSR 模型可以大大超过多个基线方法，并在多个公共数据集上取得最先进的性能。基于F1 分数评价指标，CSSR 模型在开放集识别的任务上能够实现平均 8.3% 的提升。

2 相关工作

本文主要与开放集识别有关，特别是基于自动编码器和类原型的方法。开放集识别能很自然与其他一些经典问题联系起来，如分布外检测（Out-Of-Distribution，OOD）检测 [13] 和新颖性检测（Novelty Detection） [27]。本节将简要讨论分布外检测方法。

2.1 开放集识别

早期的开放集识别工作利用了传统的机器学习方法。他们使用分类器产生的分数，从而可以通过测量样本和已知类之间的分数相似度来识别未知样本 [1], [16]。例如，Scheirer 等人 [32] 采用支持向量机识别已知类，使用极值分布来检测未知类别。最近，研究人员开始利用深度神经网络强大的特征学习能力进行未知类别的检测。

一些研究人员为开放集识别设计或采用了一些独特的分类层 [2], [13], [46]。一个最直接的方式是利用最大 SoftMax

概率拒绝一些不自信的预测 [13]。Bendale 等人 [2] 证明了 SoftMax 概率并不稳定，并提出用 OpenMax 函数代替 SoftMax 函数。该方法重新分配 SoftMax 的分数，以显式地获得未知类的置信分数。Zhou 等人 [46] 提出了占位符学习（Placeholder Learning）的概念，通过为未知类保留分类器占位符来校准过度自信的预测。

基于自动编码器的深度神经网络方法： Zhang 等人 [44] 提出重构误差中包含有用的可区分性信息，并提出使用稀疏特征来模拟开放集识别问题。Yoshihashi 等人 [41] 设计了 CROSR 方法，该方法使用隐空间特征对封闭集分类器进行训练和未知检测。Oza 和 Patel [24] 提出了一种二阶段的 C2AE 方法。该方法首先训练用于封闭集识别的编码器，然后将其参数固定住，并加入类条件信息来训练用于未知检测的解码器。Sun 等人 [34] 使用变分自动编码器强制不同的隐空间特征近似不同的高斯模型进行未知检测。他们随后开发了 CPGM [35]，将鉴别信息添加到概率生成模型中。Perera 等人 [26] 将原始图像和重构的图像送入分类网络，当重构的图像与原始输入的图像一致时，预测结果将会是有高置信度的。然而，如本文在小节 1 中所述，基于自动编码器的方法存在两个问题：（一）从像素级图像重构中学习到的特征包中包含不必要的背景信息，这可能会同时损害在封闭集和开放集中的性能。（二）自动编码器学习一个连续的流形来适应已知的样本，这可能会吞噬类间区域，引入了开放空间风险。

类原型的深度神经网络方法： Yang 等人 [40] 提出了广义卷积原型学习，它用一个面向开放世界的原型模型取代了封闭世界中假设的 SoftMax 分类器。Chen 等人 [4] 提出了互换点学习（Reciprocal Point Learning，RPL），它根据互换点的差异性（Otherness）将样本分类为已知或未知。随后，RPL 被进一步改进为 ARPL [3]。该模型整合了一个额外的对抗性训练策略，通过产生对抗性（看起来为“真”）的训练样本来提高模型对已知和未知类别的区分度。由于缺乏拟合能力和代表性的多样性，类原型的方法的可用性是有限的。本文中，我们应用类别独有的自动编码器来解决这个问题。

2.2 分布外检测

正如 Hendrycks 和 Gimpel 首次提出的那样，分布外检测主要目的是检测不属于训练集分布的样本。一些方法考虑了训练期间可以获得分布外样本的问题 [8], [19]–[21], [29], [47]。然而，这与我们的任务并不一致，即在训练期间只有分布内的数据可以获得。虽然分布外检测问题可以通过可用的已知离群点（Outlier）来简化，但开放集识别在实际中更常见且更具有现实意义。此外，“特别设计”的模型也可以从离群点暴露中获益，如 OpenGAN [17] 可以使用或不使用用于辅助的分布外数据。在下文中，我们主要关注训练没有额外分布外数据的模型。

全监督方法： 在与开放集识别相似的问题设置下，这些方法在分类任务的基础上建立分布外检测器。有些方法寻求更好的分数函数，包括最大 SoftMax 概率 [13]、最大 Logits 分数 [12] 以及能量分数 [21]。Vyas 等人 [38] 单独使用“留一法”的分类器集合来模拟分布外数据进行训练，即在训练中迭代式地留出一个类别作为离群点数据。Sastry 和 Oore [31]提出用格莱姆矩阵（Gram Matrix）来描述活动模式，并通过跟训练数据的格莱姆矩阵比较来计算元素层面的偏差，从而来对分布外隶属度进行评分。

自监督方法： 自监督方法近期较新地被学者们提出，通过自我监督来学习良好的特征实现准确的分布外检测。Golan 和 El-Yaniv [10] 和 Hendrycks 等人 [14] 考虑了预测图像变换的任务（如将图像旋转到 0°、90°、180° 和 270°），这种方法在 [36] 也作为辅助任务使用。自监督对比学习在无监督的特征学习中展现了相当大的成功，并且也可以应用于分布外检测中 [33], [36], [39]，通过对比学习获得的特征在分布内和分布外数据之间有明显的规律。尽管这些方法是为无标签的环境设计的，但也被扩展到监督学习中。

3 预备知识

开放集识别： 给定一组 n 个有标记的实例，$\mathcal{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$，其中 $y_i \in \{1, ..., m\}$ 是已知类的相应标签。对于开放集识别，目标是从 $\mathcal{X}$ 中学习一个模型，将测试样本分为 $m+1$ 个类，即 m 个已知类其中一个或由 $m+1$ 索引的未知类。

自动编码器： 自动编码器以无监督的方式学习一组数据的有效特征表示。通过串联在一起的编码器 f 和解码器 g 的瓶颈结构，自动编码器被迫将高维输入特征压缩到低维隐空间 H 从而能够充分地重构原始输入，也就是使每个输入样本 $\|\mathbf{x} - g(f(\mathbf{x}))\|_2^2$ 的重构误差最小。解码器 g 学习一个流形 $V = \{g(\mathbf{h}) | \mathbf{h} \in H\}$，而编码器 f 学习一个从原始特征空间到流形 V 的映射。基于自动编码器的开放集识别方法将流形 V 与已知类别样本的分布相适应，其中重构误差是输入样本和流形 V 之间的距离。现有的方法重构了整个原始图像，并使像素层面的重构误差最小。然而，拟合背景像素（类别无关的信息）对封闭集和开放集的识别都是无益的。Zhang 等人 [45] 也证明了这一点，他在文中指出，在隐空间特征上建立流形密度估计比在原始图像上效果更好。因此，我们在由骨干网络提取的隐空间上建立自动编码器。

原型和互换式学习： 通过为每个类别 i 定义类别独有的点集合 U_i，原型学习 [40] 将测试样本分配给最近的原型点，而远离所有原型点的样本被视为来自未知类别。从严格地数学角度上看，对封闭集和开放集的识别建模如下:

$$
\begin{aligned}
p(y = i | \mathbf{x}, \mathcal{B}, U) &\propto \left(- \min_{\mathbf{u} \in U_i} \|\mathcal{B}(\mathbf{x}) - \mathbf{u}\|_2^2 \right), \\
p(unknown | \mathbf{x}, \mathcal{B}, U) &\propto \min_i \min_{\mathbf{u} \in U_i} \|\mathcal{B}(\mathbf{x}) - \mathbf{u}\|_2^2,
\end{aligned} \tag{1}
$$

其中 $\mathcal{B}$ 是骨干网络，从输入 $\mathbf{x}$ 中提取隐空间特征。相反，互换学习 [4] 利用特定类别的互换点集 U_i 来学习差异性

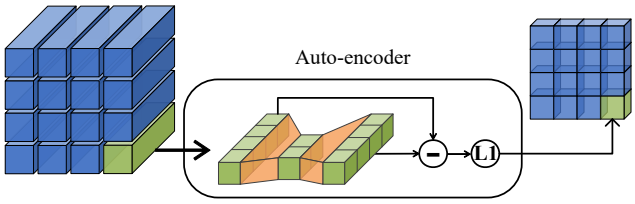
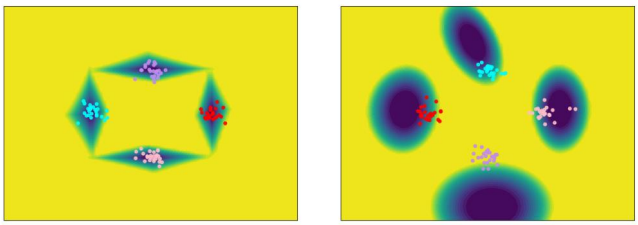


图 2. 单个自动编码器的结构。在示例中，我们在像素层面上进行重构操作。它将语义特征图作为输入并输出每个像素的重构误差。



(a) CSSR (b) CSSR MSE

图 3. 用 (a) CSSR 模型和 (b) 带有均方误差的 CSSR 模型的开放空间比较。我们设置了一个简单的实验，从不同的高斯分布中产生的四个类别用于训练。然后，我们测试了整个特征空间的未知检测。鲜黄色和深蓝色的区域分别对应于不同方法所识别的开放空间和封闭空间。

（Otherness），而不是隶属度（Belongingness），并认为靠近所有互换点的样本是未知的。这可以表示为

$$p(y=i|\mathbf{x},\mathcal{B},U) \propto \sum_{\mathbf{u}\in U_i} \|\mathcal{B}(\mathbf{x})-\mathbf{u}\|_2^2,$$
$$p(known|\mathbf{x},\mathcal{B},U) \propto \max_i \max_{\mathbf{u}\in U_i} \|\mathcal{B}(\mathbf{x})-\mathbf{u}\|_2^2. \quad (2)$$

在训练阶段，该模型对 SoftMax 归一化类概率的交叉熵损失进行优化。然而，仅仅优化判别损失是无效的。因此，两种方法都提出了不同的正则化约束来解决开放空间风险，从而实现更好的训练效果。原型框架提出了一个生成损失 $\mathcal{L}_{pl}$（也叫原型损失），这是一个在混合高斯密度假设下的最大似然正则化：

$$\mathcal{L}_{pl}(\mathbf{x},y;U,\mathcal{B}) = \min_{\mathbf{u}\in U_y} \|\mathcal{B}(\mathbf{x})-\mathbf{u}\|_2^2. \quad (3)$$

对于互换学习框架，开放空间风险由特征到互换点距离的约束性方差来约束，形式化为：

$$\mathcal{L}_{rp}(\mathbf{x},y;U,\mathcal{B}) = \sum_{u\in U_y} \|d(\mathcal{B}(\mathbf{x}),u)-R_y\|_2^2, \quad (4)$$

其中 $d(\cdot,\cdot)$ 是一个距离函数，而 R_y 是一个特定类的可学习余量。

$\mathcal{L}_{pl}$ 和 $\mathcal{L}_{rp}$ 都引入了额外的紧凑性约束。在这项研究中，我们观察到，优化单一判别性交叉熵损失会导致特征分布和原型点之间的分布不一致。详细的讨论请参考小节 4.2.1。

4 方法介绍

正如前文指出的那样，原型学习存在两个问题：（一）特定类的原型点集合的拟合能力是有限的；（二）非多样性的特征表示对开放集检测来说是不够充分的。为了提高原型学习框架的拟合能力和特征学习能力，本文将利用自动编码器的能力，生成原型流形来拟合已知类。在本节中，我们首先介绍了模型的主要架构，并分析了所提出的模型如何解决开放空间的风险。然后，我们描述了本文将要提出的未知检测策略。最后，本文将讨论我们的模型和现有方法之间的联系。

4.1 类独有语义重构

我们用针对每个类别 i 的自动编码器来取代特定类的点集 U_i，并将该自动编码器表示为 $\mathcal{A}_i$。如 Fig. 2 所示，它将隐空间特征 $\mathbf{z}$ 作为输入，并输出重构的图像 $\hat{\mathbf{z}}=\mathcal{A}_i(\mathbf{z})$。然后，我们通过计算其一范数 $\mathcal{L}_1$ 的重构误差为：

$$d(\mathbf{z},\mathcal{A}_i) = \|\mathbf{z}-\mathcal{A}_i(\mathbf{z})\|_1. \quad (5)$$

跟原型学习一样，基于重构误差，我们的框架可以估计出类的归属性。鉴于样本 $(\mathbf{x},c)\in\mathcal{X}$，我们让 $p(y=i|\mathbf{x}) \propto (-d(\mathbf{z},\mathcal{A}_i))$ 来学习原型流形。通过应用 SoftMax 将 Logits 归一化，最终的概率可以定义为：

$$p(y=i|\mathbf{z},\mathcal{A}) = \frac{e^{-\gamma d(\mathbf{z},\mathcal{A}_i)}}{\sum_{j=1}^{m} e^{-\gamma d(\mathbf{z},\mathcal{A}_j)}}, \quad (6)$$

其中 γ 是一个控制概率分配硬度的超参数。考虑到一个理想的解决方案，即最大限度地提高真实标签类别的输出概率，自动编码器首先应该学习最小距离映射（从特征到流形），以最小化真实标签类别的重构误差。同时，自动编码器的流形也应该学习保持彼此之间的距离，以使除对应于真实标签的自动编码器外的重构误差最大化。

如前文的前置知识小节所述，特定类别的自动编码器 $\mathcal{A}_i$ 定义了类 i 的流形 V_i。在上述理想情况下，最大化 $p(y=c|\mathbf{x},\mathcal{A})$ 可以被认为是具有无限个原型点（流形 V_i）的原型学习。假设 $d(\mathbf{z},\mathcal{A}_c)$ 是最小化的，重构误差可以近似表示为

$$d(\mathbf{z},\mathcal{A}_c) = \|\mathbf{z}-\mathcal{A}_c(\mathbf{z})\|_1 \approx \min_{\mathbf{v}\in V_c} \|\mathbf{z}-\mathbf{v}\|_1. \quad (7)$$

这个点赋值过程等同于公式 (1)。点集 U_c 被 V_c 取代，平方的二范数 $\mathcal{L}_2$ 准则被一范数 $\mathcal{L}_1$ 准则取代。

4.2 解决开放空间的风险

4.2.1 拟合已知的类别

在原型点学习中，均方误差（Mean Square Error，MSE）被用来作为距离测量。我们观察到，在使用均方误差时，原型点和语义特征之间的分布不一致很可能也会被学习起来。图 3(b) 中的例子直观地显示了在原型和真实标签分布之间出现了差距这种现象。我们还观察到，这种差距是由模型的过度拟合，而不是欠拟合造成的。由公式 (3) 给出的原型损失正如作为 GCPL 工作中的正则化项，试图明确地减少这种差距，从而可以解决开放空间风险。我们接下来分析均方误差是如何导致上述的不一致的，而我们在公式 (5) 中使用的平均绝对误差（Mean Absolute Error，MAE），则可以保持一致性。

在下面的分析中，我们考虑最简单的原型学习形式，其中 $|U_i| = 1$，$\mathbf{u}_i \in U_i$ 代表 i 类别的唯一原型点。因此，$d(\mathbf{z}, U_i)$ 被简化为 $\|\mathbf{z} - \mathbf{u}_i\|$，而类别 i 的 SoftMax 概率为：

$$p(y = i|\mathbf{z}, U) = \frac{\exp(-\|\mathbf{z} - \mathbf{u}_i\|)}{\sum_j \exp(-\|\mathbf{z} - \mathbf{u}_j\|)}. \tag{8}$$

考虑到 $\mathbf{z} = \mathbf{u}_c + \boldsymbol{\varepsilon}$，我们分析了 $p(y = c|\mathbf{z}, U)$ 与 $\boldsymbol{\varepsilon}$ 分别使用平均绝对误差和均方误差的变化情况。原型学习的目的是让 $p(y = c|\mathbf{z}, U)$ 在样本正好位于其原型点时达到最大，即 $\mathbf{z} = \mathbf{u}_c$。当偏移量 $\boldsymbol{\varepsilon} = \mathbf{z} - \mathbf{u}_c$ 变大时，这个概率应该减少。然而，以下定理表明，使用均方误差不能达到上述目的。

定理 1. *给定 $d(\mathbf{z}, U_i) = \|\mathbf{z} - \mathbf{u}_i\|_2^2$，同时假设 $\mathbf{u}_i \neq \mathbf{u}_j$ for $\forall i \neq j$，则存在 c 以及 $\boldsymbol{\varepsilon} \neq 0$ 满足 $p(y = c|\mathbf{u}_c, U) < p(y = c|\mathbf{u}_c + \boldsymbol{\varepsilon}, U)$.*

定理 2. *给定 $d(\mathbf{z}, U_i) = \|\mathbf{z} - \mathbf{u}_i\|_1$，对于每一个 $c, \boldsymbol{\varepsilon}$，$p(y = c|\mathbf{u}_c, U) \geq p(y = c|\mathbf{u}_c + \boldsymbol{\varepsilon}, U)$ 都成立.*

这两个定理的证明可以在补充文件中找到。让 c 是相对于 $\mathbf{z}$ 的标签。那么，定理 1 表明，在使用均方误差的情况下，$\mathbf{z} = \mathbf{u}_c$ 不一定是使 $p(y = c|\mathbf{z}, U)$ 最大化，并且使交叉熵损失最小化的最佳方案，从而导致上述的不一致分布。相反，平均绝对误差严格满足 $p(y = c|\mathbf{u}_c, U) \geq p(y = c|\mathbf{u}_c + \boldsymbol{\varepsilon}, U)$，如定理 2 所示；因此当 $\mathbf{z} = \mathbf{u}_c$ 时，取最大 $p(y = c|\mathbf{z}, U)$ 和最小交叉熵损失。这保证了语义特征和原型点的分布的一致性。在图 3(a) 中说明的实验表明，原型是很好的拟合，而且开放空间的风险得到了很好的控制。

4.2.2 类别相关特征的学习

在骨干网络和特定类别自动编码器的联合优化中，学习过程发生在两方面。除了自动编码器被训练成适合已知类别外，骨干网络也同时被训练成使提取的特征接近于特定类别的流形。这个框架解决了章节中提到的两个问题，其中多样性是学习良好的特征所需要的。这种多样性自然体现在位于流形表面的特征上。同时在流形的垂直方向上，重构误差迫使特征变得紧凑。这些特性使得语义特征与类别相关。每个类别都由全局特征的一个子集描述并与之相关，而一个样本往往只激活与其类别对应的相关特征。

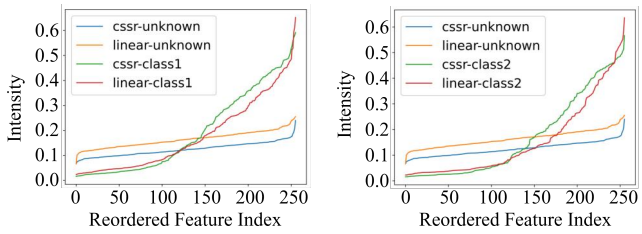


图 4. CSSR 模型针对特定类别特征激活的可视化。我们将 CIFAR10 中的六个类别作为已知的类别，其余四个类别为未知类别。请注意，每个特征的幅度在六个已知类别中是归一化的；为了更好的可视化，我们将每条曲线排序为递增的顺序。

我们在这里举一个直观的例子来解释特定类别的特征如何更容易学习。考虑一种情况，自动编码器具有最简单的形式：每个编码器是一个特定类别特征子集的元映射（Identity Mapping），而解码器相同地映射回这些特征。激活与类相关的特征不会导致重构误差，而激活与类不相关的特征则会导致较大的重构误差。为了减少这种情况下的重构误差，骨干网络学会了只激活与类相关的特征。对于联合优化，骨干网络作为一个更强大的拟合，被用来降低自动编码器的拟合复杂性（使自动编码器尽可能简单）。最后，模型最终提取了特定类别的特征。这一特性也被用来检测未知的类，这将在小节 4.4 中介绍。

图 4 显示了已知和未知类别的平均激活强度（绝对激活值）。已知类在一些特定的特征上显示出强烈的激活，而在其他特征上则几乎没有激活。然而，未知类在所有特征上的激活值都很低，因为它们没有经过训练，没有与任何特定的特征相关联。与普通线性分类层相比，CSSR 模型提取的特征在几个方面具有优势：（1）CSSR 模型的类相关特征的贡献更均匀，而不是集中在少数特征上；（2）CSSR 模型的类不相关特征的激活程度较低；（3）CSSR 模型的未知类激活强度明显低于普通线性分类器，表明已知类和学习的语义特征之间的关联更强。

最近，一种基于特征学习的新颖性检测方法 [36] 也观察到，经过良好学习的特征使得特征幅度可以直接用于区别分布内数据和分布外数据。此外，面对训练过程中分布外检测数据可以获取到的问题，Dhamija 等人 [8] 提出了一种损失函数，明确地减少分布外数据的激活幅度。而对于 CSSR 模型来说，这一特性使其在训练期间不需要访问分布外数据就能隐式地实现同样的目的。

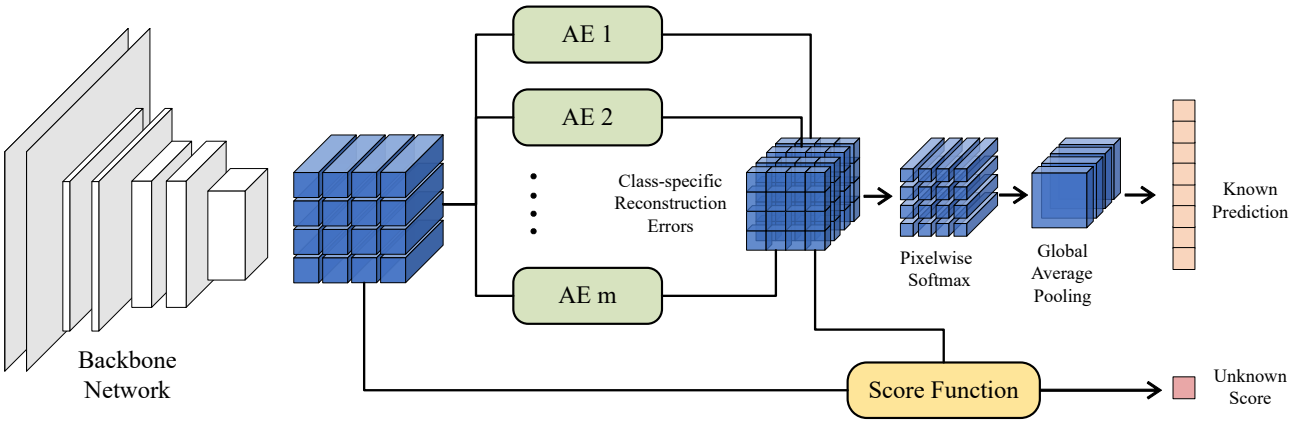


图 5. 本文所提出的模型的总体结构。骨干网络（$\mathcal{B}$）将图像作为输入并提取其语义特征图 $\mathbf{z}$。自动编码器（$\mathcal{A}_c$）针对 c 类，编码并重构像素语义特征 $\mathbf{z}$。随后，我们将类独有自动编码器的像素重构误差作为 Logits，并对 Logits 乘以 γ 进行像素层面的 SoftMax。然后，采用全局平均法，将按像素计算的预测值平均成为对整个图像的一般预测值。对于**未知推断**，该模型使用与预测的类别和语义特征相对应的像素重构误差作为输入，并为图像的未知性打分。最后，RSSC 模型会确定一个阈值来确保 95 % 的已知样本被正确接受；如果样本的未知得分低于阈值，则被拒绝。

4.3 整体框架

图 5 说明了本文所提方法的训练和推理过程。模型的框架主要包括两个模块：用于学习隐空间特征的骨干网络 $\mathcal{B}$ 和用于分类已知类和检测未知类的特定自动编码器 $\{\mathcal{A}_i\}_{i=1}^m$。

为了充分利用从 $\mathcal{B}$ 中提取的语义特征图 $Z = \mathbf{z}_{ij}$，我们平等地处理每个像素的隐空间特征。然后，通过对各像素的预测进行平均化来进行全局预测：

$$p(y = i|Z, \mathcal{A}) = \frac{1}{|Z|} \sum_{\mathbf{z} \in Z} p(y = i|\mathbf{z}, \mathcal{A}). \tag{9}$$

显然，$p(y = i|Z, \mathcal{A})$ 的总和为 1，因为 $p(y = i|\mathbf{z}, \mathcal{A})$ 的总和单独为 1。最后，通过梯度下降使真实类别 c 的负对数概率（Negative Log-probability）最小化来训练该模型，具体方法如下：

$$\mathcal{L} = -\log p(y = c|Z, \mathcal{A}). \tag{10}$$

由于每个像素的隐空间特征都集中在输入图像的一个局部区域，这种操作可以看作是通过软裁剪扩增原始输入图像并集成预测扩增图像。在测试阶段，这是一种自然的测试增强技术，从而提高预测性能。上述操作可以通过 1×1 的卷积、像素层面的 SoftMax 和全局平均池化来实现。此外，我们通过使用线性编码器和解码器来实现自动编码器。为了简单起见，本文使用 `tanh` 激活函数且无偏置项（Bias）。

基于 [4] 中的想法，所提出的方法也可以扩展到互换学习框架中。特定于类的自动编码器可以用来估计其他性，以代替互换点集。唯一的修改是，我们设置一个负的超参数 γ。然后，类似于公式 (7)，假设 $d(\mathbf{z}, \mathcal{A}_c)$ 被最大化了，重构误差可以被近似为：

$$d(\mathbf{z}, \mathcal{A}_c) = \|\mathbf{z} - \mathcal{A}_c(\mathbf{z})\|_1 \approx \max_{\mathbf{v} \in V_c} \|\mathbf{z} - \mathbf{v}\|_1. \tag{11}$$

这个过程等同于公式 (2)。在本文的其余部分，我们将 CSSR 模型的互换版本称为 RCSSR 模型。

4.4 未知类别的检测

假设给定测试样本的语义特征 Z 和预测标签 c。我们从两个不同的角度来构建未知类别检测的打分函数（Score Function）：（一）重构误差和（二）特定类别的特征统计信息。

4.4.1 基于重构误差的打分函数

一个自然的想法是利用重构误差来检测未知类别。正如在小节 4.2.2 中提到的，未知样本会导致语义特征不被激活。我们观察到，激活不良的语义特征会导致较低的重构误差，从而导致无法检测出未知的类别。由于我们通过使用线性解码器和由 `tanh` 激活的线性编码器来实现自动编码器，所以考虑了 $\|\mathbf{z}\|_1$ 和 $\|\mathbf{z} - \mathcal{A}_c(\mathbf{z})\|_1$ 之间的近似线性关系。因为编码器 f 和解码器 g 是线性函数，所以它们满足 $f(\lambda\mathbf{z}) = \lambda f(\mathbf{z}), g(\lambda\mathbf{z}) = \lambda g(\mathbf{z})$ 。对于 `tanh` 激活，假设 $f(\mathbf{z})$ 位于零附近，则 $\tanh(\lambda\mathbf{z}) \approx \lambda \tanh(\mathbf{z})$。因此有：

$$\|\lambda\mathbf{z} - \mathcal{A}_c(\lambda\mathbf{z})\|_1 \approx \lambda\|\mathbf{z} - \mathcal{A}_c(\mathbf{z})\|_1,$$

同时我们通过除以 $\|\mathbf{z}\|_1$ 来去除比例因子 λ。对于未知样本来说，$\mathbf{z}$ 位于零附近的要求可以很容易满足。此外，我们通过乘以一个额外的 $\|\mathbf{z}\|_1$ 项来考虑特征幅度的未知检测能力。

具体来说，我们将第一个打分函数定义如下。对于 CSSR 模型来说，已知类别应该具有较低的相对重构误差，同时具有较高的特征量级。这由以下公式给出：

$$s_{p1}(\mathbf{z}, c) = -\frac{d(\mathbf{z}, \mathcal{A}_c)}{\|\mathbf{z}\|_1^2}. \tag{12}$$

同时，对于 RCSSR 模型来说，已知类别的样本应该具有较高的相对重构误差和较高的特征量级：

$$s_{r1}(\mathbf{z},c)=\frac{d(\mathbf{z},\mathcal{A}_c)}{\|\mathbf{z}\|_1}\times\|\mathbf{z}\|_1=d(\mathbf{z},\mathcal{A}_c). \tag{13}$$

与封闭集分类过程类似，通过充分运用像素层面的特征，我们对特征图 Z 进行像素层面的打分，各个分数相加并平均作为整个图像的最终分数。也就是说，$s_*(Z,c)=\frac{1}{|Z|}\sum_{\mathbf{z}\in Z}s_*(\mathbf{z},c)$，其中 s_* 代表 s_{p1} 或 s_{r1}（注意，我们用 $*$ 代表选择 CSSR 模型或 RCSSR 模型）。

与原始的原型学习相比，s_{p1} 由于其特征学习能力而有所不同。然而，s_{r1} 具有与互换学习中的公式 (2) 相同的形式。其中在公式 (11) 和预测函数 $c=\arg\max_i d(\mathbf{z},\mathcal{A}_i)$ 中进行了近似。

4.4.2 基于激活规律的打分函数

虽然通过特征幅度，CSSR 模型学习类相关特征之前已经被使用过，但我们考虑通过一阶和二阶统计，对类特定的激活规律建立一个更微妙的模型。低阶统计信息已被应用于分布外检测中 [29]。然而，统计信息作为分布外检测分类器的输入特征，它需要分布外样本进行训练。相反，我们通过采集统计信息来直接制定打分函数。

假设语义特征图 Z_i 和待预测类 c_i 是由训练集的样本 $\mathbf{x}_i$ 得到的。由于我们只关注特征的激活强度，我们通过计算其元素的绝对值对所有的特征图进行预处理；在下面的讨论中，我们假设特征图已经被预处理。由于需要类别独有模式，因此根据预测的类别将特征图分为不同的集合，即特征集 $\mathcal{Z}^c=\{Z_i|c_i=c,i=1,2,...,n\},c=1,...,m$。

对于一阶统计，我们首先采取了特定类别的平均激活强度：

$$\mu_i=\sum_{Z\in\mathcal{Z}^i}\sum_{\mathbf{z}\in Z}\frac{1}{|\mathcal{Z}^i||Z|}\mathbf{z}. \tag{14}$$

为了考虑不同特征的激活强度的不同尺度，我们进一步应用了类别归一化：

$$\tilde{\mu}_i=\frac{\mu_i}{\sum_j\mu_j}, \tag{15}$$

其中向量除法是按元素层面进行的。为了检测未知性，激活已知类所激活的特征的样本更有可能是同一个已知类。具体来说，每个特征的激活强度由 $\tilde{\mu}_c$ 加权，未知性由各特征的加权平均强度来评估。同时，我们还进行了像素层面的分数积分。从形式上看，打分函数定义为

$$s_2(Z,c)=\sum_{\mathbf{z}\in Z}\frac{1}{|Z|}\mathbf{z}^\top\tilde{\mu}_c. \tag{16}$$

对于二阶统计，受到 Sastry 和 Oore 工作 [31] 的启发，我们利用格莱姆矩阵（Gram Matrices）来模拟特征间的相互关联性。令 $F\in\mathbb{R}^{D\times|Z|}$ 为特征强度矩阵，特征图 Z 中的像素向量沿着列拼接，D 为特征维度。第 i 个样本的格莱姆矩阵由 $G=FF^\top$ 定义。格莱姆矩阵中的元素 G 描述了相应的两个特征（以行和列为索引）同时被激活的可能性有多大。我们对特征图类的格莱姆矩阵进行平均化，具体来说，就是作为特征某一元素同时在不同特征出现的概率方差。然后，对于一个测试样本，我们计算它的格莱姆矩阵，并通过使用预先计算的模板与它的预测标签的元素相乘之和对其不可知性进行评分。这个过程可以被表述为：

$$G^c=\frac{1}{|\mathcal{Z}^c|}\sum_{Z\in\mathcal{Z}^c}G(Z), \tag{17}$$

$$s_3(Z,c)=Sum(G^c\odot G(Z)), \tag{18}$$

其中 $Sum(\cdot)$ 是一个对矩阵元素求和的函数，$\odot$ 是矩阵元素层面（Elementwise）的乘法。我们观察到，上述操作相当于将像素特征 $\mathbf{z}$ 扩展到二阶多项式空间，即 $\mathbf{z}\mathbf{z}^\top=[z_iz_j]_{D\times D}$。格莱姆矩阵可以写成像素级扩展特征的总和 $G=\sum_{\mathbf{z}\in Z}\mathbf{z}\mathbf{z}^\top$；因此，$G^c$ 代表扩展特征空间的一阶统计。此外，打分函数可以看作是像素层面的评分和积分，即 $s_3(Z,c)=\sum_{\mathbf{z}\in Z}Sum(G^c\odot\mathbf{z}\mathbf{z}^\top)$。除了格莱姆矩阵的主要定义外，这里还可以选择 [31] 中提出的扩展高阶格莱姆矩阵，即：

$$G=\left(F^pF^{p\top}\right)^{\frac{1}{p}}, \tag{19}$$

其中矩阵的幂是按元素层面计算的。我们发现，高阶格莱姆矩阵可以略微提高性能；在本文的其余部分，我们默认设置 $p=8$。

4.4.3 整体打分函数

在整合上述打分函数之前，我们对各个分数进行归一化处理，使不同打分函数的度量衡一致。具体来说，对于每一个打分函数 s_*，我们得到的分数是针对**随机扩增**训练样本的，以减少对训练样本过拟合带来的过度自信的影响。然后，可以计算平均值和标准差，并进行归一化。这个过程可以表示为：

$$\tilde{s}_*(Z,c)=\frac{s_*(Z,c)-E(s_*)}{Std(s_*)}, \tag{20}$$

其中 $E(s_*)$ 和 $Std(s_*)$ 分别代表预先计算的关于 s_* 的平均值和标准差。现在我们通过线性组合来整合所有三个打分函数，以获得我们的最终打分函数:

$$s_{all}(Z,c)=w_1\times\tilde{s}_{*1}(Z,c)+w_2\times\tilde{s}_2(Z,c)+w_3\times\tilde{s}_3(Z,c), \tag{21}$$

其中 s_{*1} 代表 s_{p1} 或 s_{r1}，w_1、w_2, 和 w_3 是每个检测分数的权重。

这个过程在图 6 中得到了说明。我们将未知类检测的过程总结如下三个步骤：（一）在训练集上**不使用数据增强**运行整个模型，收集语义特征的一阶和二阶特征统计。这个步骤使用非增强的训练样本，保留了已知类别的完整信息。（二）在训练集上**使用数据增强**运行整个模型，收集各个分数的归一化参数。这一步使用增强的训练样本可以减少过度自信的影响。（三）使用公式 (21) 中的综合分数进行推理。为了拒绝

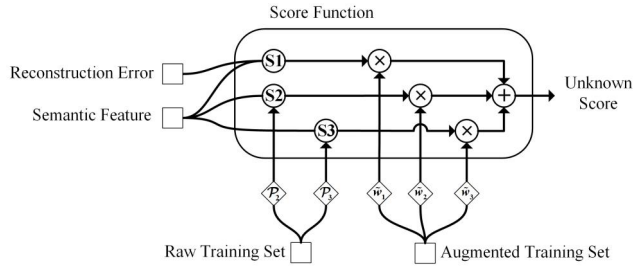


图 6. 未知类别推理过程。在对测试样本进行推理之前，我们从原始训练集获得一阶和二阶统计数据（分别对应图中的 $\mathcal{P}_2$ 和 $\mathcal{P}_3$）。然后，从增强的训练集中获得单个分数的归一化系数（$\frac{1}{Std(s_*)}$）。最终的分数权重由归一化系数和预定义权重的乘积计算，即 $\tilde{w}_* = \frac{w_*}{Std(s_*)}$，形成打分函数 $s_{all}(Z,c) = \sum_i \tilde{w}_i \times s_i(Z,c)$。

未知类样本，我们采取了一个阈值，保证 95% 的已知样本被接受。

4.5 与当前方法的联系

4.5.1 与基于自动编码器方法的联系

基于自动编码器的开放集识别方法 [24], [25], [34] 同时学习分类任务和重构任务。最近的工作通过多任务学习或多阶段学习来实现这一目标。除了章节中提到的自动编码器引起的吞噬问题外，我们从训练框架的角度分析了现有的方法。

多任务学习 [25], [34] 通过共享共同知识来学习联合特征，共同优化分类任务和重构任务。因此，这些方法在封闭集分类上的表现往往比普通分类器略差。

多阶段训练方法 [24] 在分类任务上训练一个编码器，然后训练解码器的同时冻结编码器的参数。虽然这种策略可以保证在封闭集分类的性能，但在训练解码器时可能会遇到麻烦。因为编码器被训练来提取具有区分性的分类特征，而这些特征可能不包含足够的像素信息来重构细节。

正如本文在小节 1 中所说的，将封闭集分类和图像重构任务分开限制了整个模型的性能，因为编码的背景信息损害了分类任务，而这两个任务之间的权衡是不可避免的。本文的方法从两个方面避免了上述问题：（1） CSSR 模型重构的是语义特征，而不是原始图像，这就避免了重构不必要的背景像素，而专注于与类别相关的代表性语义特征。（2） CSSR 模型对重构和分类任务进行连接，其中重构任务是根据分类损失来训练的。因此，两个任务的性能之间存在着正相关，而不是由于权衡两个任务的性能而产生的负相关。

4.5.2 与表示学习方法的联系

Chen 等人 [5] 最近提出了一个新的特征学习框架，称为 SimSiam 模型。我们观察到，CSSR 模型和 SimSiam 模型有着相似的架构。SimSiam 模型将两个随机增强的视图作为输入，并通过一个共享的骨干网络来处理它们。然后，一个预测多层感知机（Multilayer Perceptron，MLP）头对一个视图的输出进行转换，并将其与另一个匹配。学习目标是使两个输出向量的负余弦相似度最小。SimSiam 模型中的预测多层感知机头是一个瓶颈结构（正如作者所建议的），它类似于一个自动编码器。通过确保两个视图是相同的，SimSiam 模型学习最大化原始语义特征和重构语义特征之间的余弦相似度。对于 CSSR 模型中的每个自动编码器，对于某些类别的样本，重构误差最小化（类似于正数对），对于其余的样本，重构误差最大化（类似于负数对）。它所学习的特定类别特征，是由骨干网络提取的共享特征的子空间。这解释了 CSSR 模型是如何增强学习类相关语义特征的，并使特征强度直接作用于区分未知的类。

表 1. 不同方法在未知类别检测任务上的 AUROC 指标比较。最佳性能值用**粗体**突出显示。

方法	SVHN	CIFAR10	CIFAR+10	CIFAR+50	TinyImageNet
CROSR [41]	89.9	88.3	91.2	90.5	58.9
C2AE [24]	92.2	89.5	95.5	93.7	74.8
MLOSR [25]	95.5	84.5	89.5	87.7	71.8
CGDL [34]	93.5	90.3	95.9	95.0	76.2
GFROSR [26]	93.5	83.1	91.5	91.3	64.7
GCPL [40]	92.6	82.8	-	-	-
ARPL [3]	96.7	91.0	**97.1**	95.1	78.2
Plain Softmax	88.6	67.7	81.6	80.5	57.7
OSRCI [22]	91.0	69.9	83.8	82.7	58.6
PROSER [46]	94.3	89.1	96.0	95.3	69.3
CSSR	**97.9**	91.3	96.3	96.2	**82.3**
RCSSR	97.8	**91.5**	96.0	**96.3**	81.9

5 实验部分

5.1 实验细节

由于 CSSR 模型只修改了分类层，各种骨干网络可以交替使用来实现 CSSR 模型。跟 Chen 等人的工作 [4] 一样，我们选择用 Wide-ResNet [43] 来训练小规模的数据集，其深度、宽度和 Dropout 概率本文中分别设置为 40、4 和 0，即 WRN40-4。然而，对于更大规模的数据集（比如 TinyImageNet），我们用 ResNet18 [11] 代替骨干网络从而提高效率。在训练阶段，我们使用了随机梯度下降优化器，动量（Momuntum）设为 0.9。模型共训练 200 个迭代次数（Epoch），批次大小（Batch-Size）固定为 128。学习率最初被设置为 0.4，然后在第 130 次和第 190 次迭代中下降 10 倍。我们在所有的实验中都设置了 $|\gamma| = 0.1$。自动编码器是用线性编码器和线性解码器实现的。为了使自动编码器的嵌入空间 H 有界，我们使用 `tanh` 作为激活函数。在 ResNet18 和 WRN40-4 架构中，自动编码器的嵌入空间维度被设置为 64。分数积分权重全部设置为 1。以前的方法使用数据增强技术来改善开放集的识别效果。按照先前工作的设置 [26], [46]，我们在 [7] 中应用了一个简单

表 2. CIFAR-10 数据集上的开放集分类结果，测试阶段增加了多种未知数据集。

方法	IMGN-C	IMGN-R	LSUN-C	LSUN-R
Plain Softmax	63.9	65.3	64.2	64.7
CROSR [41]	72.1	73.5	72.0	74.9
GFROSR [26]	75.7	79.2	75.1	80.5
C2AE [24]	83.7	82.6	78.3	80.1
CGDL [34]	84.0	83.2	80.6	81.2
PROSER [46]	84.9	82.4	86.7	85.6
CSSR	92.9	90.9	**94.1**	93.5
RCSSR	**93.3**	**91.5**	94.0	**94.0**

的数据增强技术。我们考虑了 RandAugment 中使用的一个变换子集，即亮度、颜色、均衡、旋转、锐度、剪切和对比度。对于每个输入图像，最多两种变换被取样并应用到图像上。

除了原型 CSSR 模型外，本文还实现了 RCSSR 模型，并对其进行了评估，以进行综合比较。

5.2 与最先进的方法比较

5.2.1 未知类别的检测

本实验采用了 [22] 中定义的评价协议。本实验中使用了五个图像数据集：SVHN [23]、TinyImageNet [28]、CIFAR10 [18]、CIFAR+10 和 CIFAR+50。对于 SVHN 和 CIFAR10，6 个类被随机抽出作为已知类，其余 4 个类被用作未知类。对于 TinyImageNet，20 个类被抽样作为已知类，其余 180 个类作为未知类。对于 CIFAR+M 数据集，模型是以 CIFAR10 中的四个非动物类作为已知类进行训练的，而 CIFAR100 数据集中的 M 动物类被随机选择为未知类。一个与阈值无关的指标，即接受者操作特征曲线下的面积（Area Under the Receiver Operating Characteristic，AUROC）被用作评价指标。它是通过改变阈值，将真阳性率（True Positive Rate）与假阳性率（False Positive Rate）相比较来计算的。如果已知类和未知类完全可分，则 AUROC 值为 1。和 [22] 一样，我们对五个随机试验的结果进行了平均。

本文使用不同的架构比较了与我们的方法相关的框架：基于自动编码器的方法 [24]–[26], [34], [41]、类原型的方法 [3], [40]和最新的两种方法 [22], [46]。实验结果请见表 1，除 CSSR 模型外的其他数值均来自于 [3], [25], [40], [46]。除了在 CIFAR+10 上略微落后于 ARPL 外，CSSR 模型在其余五个数据集上的表现都优于其他方法，尤其是在 SVHN（+1.2%）、CIFAR+50（+1.0%）和 TinyImageNet（+4.1%）上。

5.2.2 开放集识别

除了检测未知类之外，开放集识别还需要对已知类进行联合分类，同时拒绝未知类。我们遵循 Yoshihashi 等学者 [41] 设计的实验设置，模型在整个 CIFAR10 上被训练为已知类。在测试阶段，其他数据集ImageNet [30] 和 LSUN [42]的样本被用作未知数。这两个数据集被进一步裁剪或调整大小，以确保它们具有与已知样本相同的图像大小；选择 10000 个样本（以保持与 CIFAR10 测试集的一致性）形成 ImageNet-Crop（IMGN-C）、ImageNet-Resize（IMGN-R）、LSUN-Crop（LSUN-C）和 LSUN-Resize（LSUN-R）四个子数据集。为了进行公平的比较，我们使用了 Liang 等人 [20] 发布的数据集版本。在性能评估上，通过 11 个类（包括 10 个已知类和 1 个未知类）的宏观平均 F1 分数进行，实验结果请见表 2。除 CSSR 模型以外的数值来自于 [25], [34], [46]。可以观察到，CSSR 模型以很大的优势（平均值为 8.3%）超过了现有方法。

5.2.3 分布外检测

在这一节中，我们按照 Chen 等人的实验设置，与分布外检测环境下的方法进行比较。我们还在这个实验中比较了 CSI [36] 和 OpenGAN [17]。CSI 和本文的方法有一个类似的想法，即利用学习好的图像表示来检测分布外样本。为了保持比较的公平性，我们在评估期间禁用了 CSI 的测试增强技术。对于 OpenGAN，我们使用基线模型训练的骨干作为其特征编码器，并直接使用相应的测试未知数据集作为其分布外检测验证集进行模型选择。我们考虑了两对具有挑战性的分布外检测基准方法 [13]，同时包括三个常用数据集：CIFAR10、CIFAR100 和 SVHN。模型在 CIFAR10 上训练，而 CIFAR100 和 SVHN 在测试阶段分别作为近分布外检测和远分布外检测数据集。其中，重叠的类别已从 CIFAR100 中删除。除了 AUROC 指标之外，本文按照 Chen 等人的工作 [3]，还使用了其他几个评价指标：

- **检测准确率（DeTection ACCuracy，DTACC）：** 这个指标代表了在所有可能的阈值上最大的已知/未知分类精度。在计算准确率时，假定阳性和阴性样本在测试集中出现的概率相同。
- **精度-记忆曲线下面积（Area Under the Precision-Recall Curve，AUPR）：** 该曲线绘制了在阈值不同的情况下，精度 $TP/(TP+FP)$ 对召回率 $TP/(TP+FN)$ 的比值，其中 TP、FP 和 FN 分别表示真阳性、假阳性和假阴性（False Negative）。AUPR 进一步计算为 AUIN 和 AUOUT，其中在和不在分布样本分别被设定为阳性。

如表 3 所示，我们将结果与 ARPL [3] 的结果进行了比较。最先进的几个方法在两个分布外数据集上都取得了相似的性能。对于近分布外检测数据集，CSSR 模型的表现与 APRL 相当，并且主要明显优于基于原型点的开放集识别（GCPL），而 RCSSR 模型则以 2.3% 的增量优于 APRL。对于远分布外数据集，我们观察到 CSSR 模型和 RCSSR 模型具有相似的性能；它们都以很大的优势超过了传统的类原型的方法。

表 3. 在多种指标下，区分 CIFAR10 与近分布外数据集 CIFAR100 和远分布外数据集 SVHN 的性能。

方法	In:CIFAR10 / Out:CIFAR100				In:CIFAR10 / Out:SVHN			
	DTACC	AUROC	AUIN	AUOUT	DTACC	AUROC	AUIN	AUOUT
SoftMax	79.8	86.3	88.4	82.5	86.4	90.6	88.3	93.6
GCPL [40]	80.2	86.4	86.6	84.1	86.1	91.3	86.6	94.8
RPL [4]	80.6	87.1	88.8	83.8	87.1	92.0	89.6	95.1
ARPL [3]	83.4	90.3	91.1	88.4	91.6	96.6	94.8	98.0
CSI [36]	84.4	91.6	92.5	90.0	92.8	97.9	96.2	99.0
OpenGAN [17]	84.2	89.7	87.7	89.6	92.1	95.9	93.4	97.1
CSSR	83.8	92.1	89.4	89.3	**95.7**	**99.1**	98.2	**99.6**
RCSSR	**85.3**	**92.3**	**92.9**	**91.0**	**95.7**	**99.1**	**98.3**	**99.6**

5.3 消融实验

本节将分析 CSSR 模型的不同组成部分和打分函数在模型中的贡献。我们首先比较了该模型的各种架构。

数据集： 本文在 CIFAR10 上训练模型。在 CIFAR10 的实验中，我们将 CIFAR10 中的所有 10 个类作为已知类，然后，在 SVHN、LSUN-Resize、ImageNet-Resize、LSUN-Fix（LSUN-F）和 ImageNet-Fix（IMGN-F）上进行测试。LSUN-Fix / ImageNet-Fix 两个数据集包含由 Tack 等人 [36] 制作的在 LSUN / ImageNet 数据集中随机采样和调整大小的图像，这两个数据集比 Liang 等人 [20] 发布的原始版本更具挑战性。

消融项： （一）**分类层：** 我们将传统的分类模型与普通的线性分类层作为基线方法进行比较，为了进行公平的比较，我们将骨干网络和超参数固定。（二）**分类策略：** 我们使用了本文提出的像素层面预测策略（像素层面的 SoftMax，然后进行平均池化，简写为 SM-AP）或普通预测策略（平均池化，然后进行 SoftMax, 简称为 AP-SM）。像素层面预测策略对训练和测试都有影响。（三）**重构误差测量：** 我们用均方误差或平均绝对误差（默认）来测量 CSSR 模型的重构误差。

消融实验结果显示在表 4 中。该表说明了以下情况：（一）作为一个非线性分类层，CSSR 模型略微提高了封闭集的性能。（二）像素层面分类法略微提高了封闭集的分类性能，而很大程度上提高了未知检测的性能。（三）使用平均绝对误差的距离度量普遍优于均方误差，证明均方误差是检测未知样本的一个好选择。

接下来，分析了不同评分函数的效果。我们通过固定训练好的 CSSR 模型和指定不同的评分函数进行决策比较。本实验中利用了 ImageNet30（由 Hendrycks 等人 [14] 介绍的 ImageNet 的一个子集），其中 10 个类被采样为已知，其余 20 个类为未知。在所有的实验中，类的划分保持不变，为了简单起见，按字母顺序选择前 10 个类作为已知类。为了适应 ImageNet-30，它具有更高的图像分辨率，RandomCrop 被 RandomResizedCrop 所取代，在训练 ImageNet 时遵循标准的数据扩增。一个带有线性分类层的普通 ResNet18 被视为基线方法。对六种不同的打分函数进行了比较。相对重构误差（Relative Reconstruction Error，RRE，$\frac{d(\mathbf{z},\mathcal{A}_c)}{|\mathbf{z}\|_1}$）、特征幅度（Feature Magnitiude，FM，$|\mathbf{z}\|_1$）、s_{*1}（s_{p1} 用于 CSSR 模型、s_{r1} 用于 RCSSR，最大 SoftMax 概率用于基线方法），s_2（公式 (16)），s_3（公式 (18)）和 s_{all}（公式 (21)）。

实验结果如表 5 所示，我们观察到以下几点：（一）普通线性分类层学习的特征与类的关系较小；特征大小对检测未知类不敏感，基于特征的打分函数对未知类的判别能力较差。（二）CSSR 模型提高了特征学习能力，也提高了两个基于特征的打分 s_2、s_3。（三）对于 s_{*1}，整合 RRE 和 FM，显著提高了 CSSR 模型和 RCSSR 模型的开放集性能。（四）像素预测和平均绝对误差善于提高所学特征的质量，从而提高基于特征的打分函数的性能。

尽管公式 (21) 中的分数融合不能保证改善所有个体的分数，但它近似地保持了最佳的个体分数。我们证明了这三个打分函数在不同的数据集上有不同的表现。保留三个打分函数的最佳性能可以提高不同数据集的整体性能同时减少差异。在图 7 中，我们进一步展示了不同的评分函数在未知检测实验中的表现以证明在不同数据集下的性能变化，如图 1 所示。例如，s_3 在 CIFAR+10 和 CIFAR+50 中获得了优势，但在 SVHN 和 TinyImageNet 中却没有。然而，融合后的分数至少是第二好的，同时标准偏差是最小的。

5.4 深入分析

5.4.1 封闭集性能

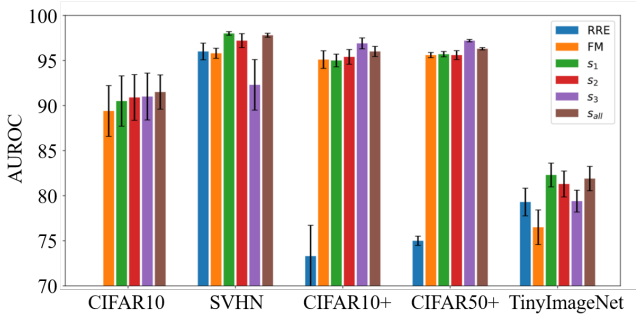
保持封闭集的性能对开放集的识别同样重要。这里，我们比较了 CIFAR10 上的封闭集性能。我们禁用了额外的数据增强技术使得实验比较公平。对于比较方法，我们采用了他们原始论文中的报告结果。考虑到实现的差异，我们还提供了我们实现下的基线方法结果。如表 6 所示，CSSR 模型在封闭集分类上展现了卓越的性能。具体上，CSSR 模型略微提高了 0.2 %（CSSR）和 0.5%（RCSSR）的性能，表明更好的表示学习提高了 CSSR 模型的封闭集性能。

表 4. 对多种架构的消融研究。第一行指定了用作未知类的数据集，Close Acc 代表封闭集测试的准确性（%）。对于未知检测，我们提供 AUROC 值进行评估。

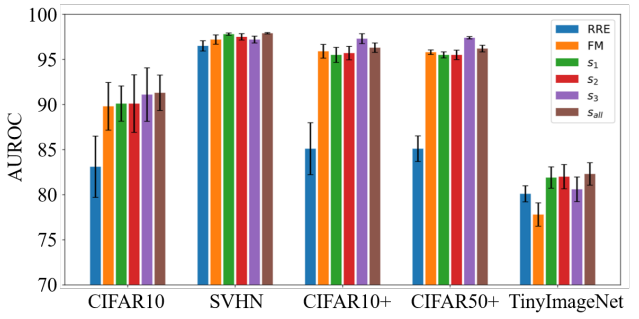
方法	Close Acc	SVHN	LSUN-R	IMGN-R	LSUN-F	IMGN-F	Average
Linear	96.77	97.0	95.3	94.2	92.3	93.2	94.4
Linear SM-AP	96.96	96.8	95.7	94.8	92.9	93.4	94.7
CSSR AP-SM	96.69	98.9	98.5	97.1	89.7	89.5	94.7
CSSR MSE	96.85	98.9	98.5	97.3	95.4	94.4	96.9
CSSR	96.86	99.1	98.8	97.5	96.2	95.3	**97.4**
RCSSR AP-SM	96.84	97.4	97.1	95.5	88.1	89.2	93.5
RCSSR MSE	96.82	98.7	97.8	96.5	92.0	91.1	95.2
RCSSR	97.02	99.1	99.1	98.1	96.0	95.0	97.3

表 5. 在 ImageNet-30 数据集上训练的消融模型的不同打分函数的 AUROC 比较。

方法	RRE	FM	s_{*1}	s_2	s_3	s_{all}
Linear	-	39.3	93.5	92.5	87.1	-
CSSR AG-SM	93.6	90.6	92.2	94.0	91.9	94.6
CSSR MSE	85.1	81.9	92.2	95.1	93.9	94.7
CSSR	91.3	91.5	94.8	95.5	94.6	95.3
RCSSR AG-SM	84.5	84.7	95.1	92.4	92.1	94.7
RCSSR MSE	89.7	81.3	95.0	94.2	93.7	94.6
RCSSR	86.8	90.2	95.1	94.6	94.4	95.0



(a) CSSR



(b) RCSSR

图 7. 未知检测实验中不同打分函数的表现。标准差由五种不同已知-未知划分的个随机试验计算。

表 6. CIFAR10 数据集上的封闭集性能比较。

方法	准确率
CROSR [41]	94.0
CGDL [34]	91.2
GCPL [40]	93.3
ARPL [3]	94.0
Our baseline	95.1
CSSR	95.3
RCSSR	95.6

5.4.2 开放度性能分析

作为代表开放集任务的复杂性的度量指标，开放度（Openness）[32] 定义为

$$Openness = 1 - \sqrt{\frac{2 * N_{train}}{N_{test} + N_{target}}}, \tag{22}$$

其中 N_{train} 是训练期间看到的已知类的数量，N_{test} 是测试期间将观察到的类的数量，N_{target} 是测试期间要识别的类的数量。利用常见的实验设置 [34], [46]，我们在 CIFAR100 上进行了实验，随机抽出 15 个类作为已知类。未知类的数量从 15 到 85 不等，这意味着开放度从 18% 到 49% 不等。16 个类别（15 个已知类和 1 个未知类）的识别性能通过分类精度进行评估。实验结果如图 8 所示。随着开放度的增加，CSSR 模型展现出良好的性能，而在使用普通线性分类层时，性能则迅速下降。

5.4.3 大规模数据集性能分析

为了在大规模分类任务上评估模型的性能，我们在 ImageNet-1000 上进行了实验 [30]。作为一个更具挑战性的数据集，ImageNet-1000 包括 1000 个类，有超过 100 万张训练图像和 5 万张验证图像。我们首先沿用了 Yang 等人使用的实验设置 [40]，其中前 100 个类被选为已知，其余 900 个为未知。我们采用 ResNet18 作为骨干网络。为了适应大规模的分类和节省参数，我们将自动编码器的嵌入空间的维度从 64 降低到 32。学习率仍然从 0.4 开始，但在 100 和 150 个迭代中下降

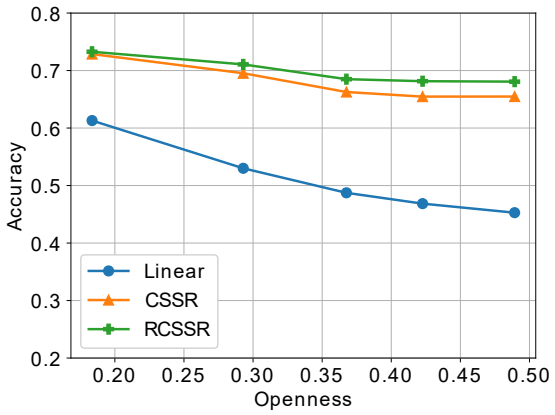


图 8. 不同开放度下CSSR 模型和基线方法的开放集识别精度。

表 7. 在划分的 ImageNet-1000 数据集的未知检测和开放集识别的性能评估结果。

	未知类检测		封闭集识别	
方法	AUROC	TNR@TPR95	Macro-F1	OSCR
SoftMax [40]	79.7	-	-	-
GCPL [40]	82.3	-	-	-
Our baseline	91.0	51.8	40.1	77.4
CSSR	**93.7**	**62.4**	**43.0**	**78.1**
RCSSR	93.1	58.4	40.7	77.6

了 10 倍，以进行充分的训练。考虑到 ImageNet-1000 所引入的更大、更复杂的语义空间，我们禁用了基于特征的评分函数 $w_2 = w_3 = 0$。其余的超参数与 ImageNet30 的实验中保持一致。除了考虑将模型的普通线性分类层作为我们的基线方法外，我们还采用了基于原型的原始方法 [40] 的报告结果来进行比较。请注意，在 [40] 中的实现利用了 ResNet50 作为骨干，这远比我们研究中使用的 ResNet18 更加强大。

为了更全面地评估模型，我们使用了两个额外的指标：（1）**TNR@TPR95** 是指在真阳性率（True Positive Rate，TPR）为 95% 的情况下，未知样本被正确拒绝的概率；（2）**开放集分类率（Open Set Classification Rate，OSCR）**该指标在 [8] 中定义，本文中也采用了该指标。我们用 δ 表示得分阈值。**正确分类比例（Correct Classification Rate，CCR）**是指在未知检测分数高于给定阈值 δ 的情况下被正确分类的已知样本的比例。**假阳性率（False Positive Rate，FPR）**是未知检测分数大于阈值 δ 的未知样本的比例。不同阈值下的正确分类比例和假阳性率值通过取正确分类比例对假阳性率曲线下的面积还原为一个特定的值。

实验结果请见表 7 。我们首先可以观察到，我们对基线方法的实现明显优于以前文献中的研究，表明对大规模数据集的开放集性能估计不足。提出的 CSSR 模型和 RCSSR 模型

表 8. 大规模开放集识别，其中 ImageNet-1000 的样本是已知的，iNaturalist 的样本作为未知的。

方法	AUROC	TNR@TPR95	DTACC
SoftMax	87.2	41.2	79.1
ARPL [3]	88.8	50.6	80.2
OpenGAN [17]	89.3	32.0	84.3
CSSR	93.8	70.7	86.5
RCSSR	**94.5**	**73.8**	**87.2**

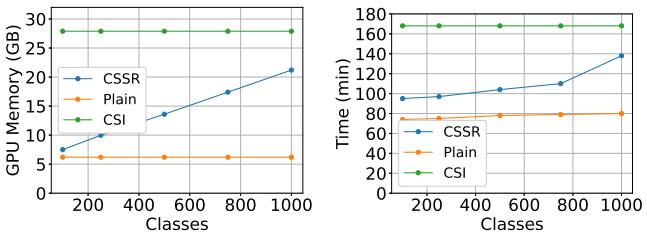
在检测未知类方面的表现优于基线方法，尤其是 CSSR 模型。我们还观察到，与未知检测的任务相比，RCSSR 模型在开放集识别方面的提升显得相对较小。这是由于 RCSSR 模型的封闭集性能的退化造成的。

此外，我们考虑了一个大规模的分类环境，来自 ImageNet-100 和 iNaturalist [15] 的样本分别作为已知类和未知类。除了基线方法，我们还比较了两种最先进的方法：ARPL [3] 和 OpenGAN [17]。为了实现高效的训练，我们对所有方法都使用了在 ImageNet-1000 上预训练的 ResNet18 骨干。为了训练 CSSR、APRL 和基线方法，我们固定了预训练的骨干网，并对分类层进行了 4 次微调。而对于 OpenGAN，预训练的骨干作为固定的特征编码器。请注意，OpenGAN 需要一个验证集，作为已知的未知数来选择适当的判别器进行未知推断。在这个实验中，我们仅取用测试集上表现最好的检查点来避免验证集的选择。如表 8 所示，CSSR/RCSSR 模型以显著的幅度超越了现有的方法。例如，在 AUROC 上改进了 5.2%，在 TNR@TPR95 上改进了 23.2%，在 DTACC 上改进了 2.9%。

表 9. 不同方法在 ImageNet-1000 数据集上训练的全部时间。$+x$h 值表明该方法除了预先训练一个普通模型外，还需要额外的 x 小时微调。

方法	训练策略	训练时间
Plain	From Scratch	140h
CSSR	From Scratch	225h
ARPL	Fix Backbone	+6h
OpenGAN	Fix Backbone	+1h
CSSR	Fix Backbone	+8h

一个可能的担心是：CSSR 模型的参数数量会随着类别数量的增加而线性增加。事实上，在 1000 个分类任务中，CSSR 模型的额外参数数量只有 CGDL [34] 的一半（CSSR 模型大约为 60M，CGDL 大约为 175M）。相对而言，一个主要的问题可能是 CSSR 模型的训练时间和 GPU 内存需求的增加。如图所示，与普通模型相比，我们观察到 CSSR 模型的训练时间增加了 50% 以上。值得注意的是，与最近的一些分布外检测方法相比，例如 CSI [36] 这个方法，CSSR 模型 [36] 仍然是轻量级的。为了使 CSSR 模型适用于极端大规模的分



(a) Training GPU memory (b) Training time for one epoch

图 9. GPU 内存的使用量 (a) 和一个迭代次数训练所消耗的时间量 (b) 的比较。这些数值是通过在 ImageNet-1000 上训练基于 ResNet18 的模型，批次大小为 128。当改变类的数量时，样本的数量保持不变。

类，正如我们在 ImageNet-1000 上所做的那样，我们可以先预训练一个普通的分类模型，然后在 CSSR 模型分类头上进行微调以减少资源消耗。如表格 9 所示，通过微调技术，CSSR 模型与其他最好方法相比展现出很小的效率差异。一个可行的方法是采用一个普遍存在的类的层次结构，将类从粗到细进行组织。CSSR 模型只需在粗粒度上执行，而在细粒度上使用简单的分类器。我们将这一问题保留在未来的工作中。

5.4.4 失败分析

为了证明 CSSR 模型和带有线性分类层的普通分类模型之间的区别，我们选择了在 ImageNet30 实验中训练的模型（表 5）。然后，对于 CSSR 模型和普通线性模型，我们挑选了识别度最差的已知和未知样本进行可视化。具体来说，我们选择了在飞行器类别中检测分数最低的六个样本（在训练期间已知），以及在未知样本中分数最高的六个样本代表最差的检测失败案例。为了更好地进行比较，我们进一步约束未知样本，使其被预测为“飞行器”类别。图 10 种展示了选定的图像。对于已知样本的检测失败案例，CSSR 模型主要关注远处的图像，其中目标物体相对较小。然而，普通模型在分类平面太近的场合也会失败。由于已知类别的模型是在 CSSR 模型允许的范围内建立的，视觉变体对整体识别的影响较小。对于未知样本的失败，“帆船”造成的混乱最多，这可能是因为类似的背景和纹理。然而，在普通模型中，我们以为“坦克”这个类别不会被分类器混淆，然而这种错误却还是发生了。而这种错误在现实世界的应用中会导致严重问题。

6 总结与展望

在这项研究中，我们整合了自动编码器和类原型学习框架，提出了一个称为 CSSR 模型的端到端学习的开放集识别深度网络。CSSR 模型为每个已知的类别指定了一个单独的自动编码器，以替代传统原型学习框架中的类独有原型点集合。这些自动编码器被插入到骨干深度神经网络的顶部，从而可以学习重构图像的语义特征。这些类独有的自动编码器可以被认为是具有特定类别的可学习自动编码器流形的原型学习。我们还注意到，原型学习中常用的平均绝对误差距离有可能导致原型点和地面真实数据分布不一致。为了解决这个问题，我们使用了平均绝对误差距离来代替均方误差，并且可以保证原型点和数据点之间的一致性。此外，我们的框架还可以用互换点学习的思想改进为 RCSSR 模型。由于提出的方法可以学习提升特征表示，本文顺势探讨了各种基于特征的分数函数，包括一阶和二阶统计方法。

在多个数据集上进行的实验结果表明，所提出的方法优于其他最先进的方法。请注意，CSSR 模型只需要一个判别性质的分类损失函数以及作为一个分类器使用，而不需要其他损失函数。因此，CSSR 模型可以使用多种分类技术进一步实现轻松扩展。在现实世界的应用中，高维和大规模图像是开放集识别的两个典型挑战。我们计划在未来的工作中关注具有类分层结构的开放集识别问题。此外，我们的工作缺乏对用现有的已知离群点训练开放集分类器的关注。我们计划在未来的工作中通过利用图像背景来解决这个问题。

参考文献

[1] A. Bendale and T. Boult, “Towards open world recognition,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1893–1902.

[2] A. Bendale and T. E. Boult, “Towards open set deep networks,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1563–1572.

[3] G. Chen, P. Peng, X. Wang, and Y. Tian, “Adversarial reciprocal points learning for open set recognition.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[4] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, and Y. Tian, “Learning open set network with discriminative reciprocal points,” in *European Conference on Computer Vision*. Springer, 2020, pp. 507–522.

[5] X. Chen and K. He, “Exploring simple siamese representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.

[6] X. Chu, Y. Lin, X. Wang, X. Gao, Q. Tong, H. Yu, and Y. Wang, “Distance metric learning with joint representation diversification,” in *ICML 2020: 37th International Conference on Machine Learning*, vol. 1, 2020, pp. 1962–1973.

[7] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “Randaugment: Practical automated data augmentation with a reduced search space,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703.

[8] A. R. Dhamija, M. Günther, and T. E. Boult, “Reducing network agnostophobia,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 9157–9168.

[9] C. Geng, S.-j. Huang, and S. Chen, “Recent advances in open set recognition: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3614–3631, 2020.

[10] I. Golan and R. El-Yaniv, “Deep anomaly detection using geometric transformations,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 9758–9769.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

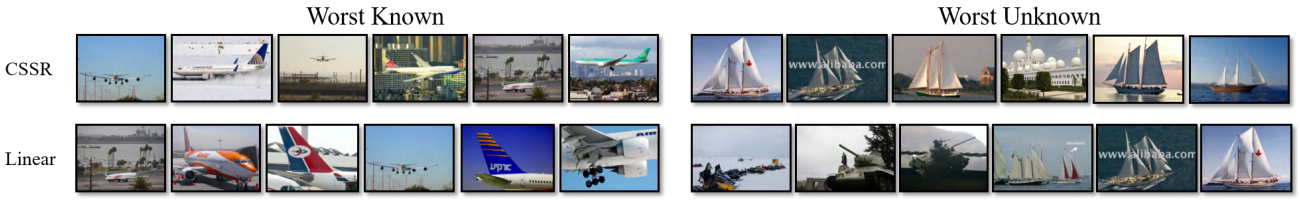


图 10. CSSR 模型或普通线性模型无法识别的已知和未知样本。

[12] D. Hendrycks, S. Basart, M. Mazeika, M. Mostajabi, J. Steinhardt, and D. Song, "A benchmark for anomaly segmentation," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[13] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *ICLR (Poster)*, 2016.

[14] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 15 663–15 674.

[15] G. V. Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. J. Belongie, "The inaturalist species classification and detection dataset," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

[16] P. R. M. Júnior, R. M. Souza, R. D. Werneck, B. V. Stein, D. V. Pazinato, W. R. Almeida, O. A. Penatti, R. D. Torres, and A. Rocha, "Nearest neighbors distance ratio open-set classifier," *Machine Learning*, vol. 106, no. 3, pp. 359–386, 2017.

[17] S. Kong and D. Ramanan, "Opengan: Open-set recognition via open data generation." *arXiv preprint arXiv:2104.02939*, 2021.

[18] A. Krizhevsky, "Learning multiple layers of features from tiny images." *Tech Report*, 2009.

[19] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 7167–7177.

[20] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *6th International Conference on Learning Representations*, 2018.

[21] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 21 464–21 475.

[22] L. Neal, M. L. Olson, X. Z. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images." in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 620–635.

[23] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Neural Information Processing Systems*, 2011, pp. 1–9.

[24] P. Oza and V. M. Patel, "C2ae: Class conditioned auto-encoder for open-set recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2307–2316.

[25] P. Oza and V. M. Patel, "Deep cnn-based multi-task learning for open-set recognition." *arXiv preprint arXiv:1903.03161*, 2019.

[26] P. Perera, V. I. Morariu, R. Jain, V. Manjunatha, C. Wigington, V. Ordonez, and V. M. Patel, "Generative-discriminative feature representations for open-set recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 814–11 823.

[27] P. Perera, R. Nallapati, and B. Xiang, "Ocgan: One-class novelty detection using gans with constrained latent representations," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2898–2906.

[28] H. Pouransari and S. Ghili, "Tiny imagenet visual recognition challenge," *CS 231N*, 2014.

[29] I. M. Quintanilha, R. de M. E. Filho, J. Lezama, M. Delbracio, and L. O. Nunes, "Detecting out-of-distribution samples using low-order deep features statistics," 2018.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[31] C. S. Sastry and S. Oore, "Detecting out-of-distribution examples with gram matrices," in *ICML 2020: 37th International Conference on Machine Learning*, vol. 1, 2020, pp. 8491–8501.

[32] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013.

[33] V. Sehwag, M. Chiang, and P. Mittal, "Ssd: A unified framework for self-supervised outlier detection," in *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

[34] X. Sun, Z. Yang, C. Zhang, K.-V. Ling, and G. Peng, "Conditional gaussian distribution learning for open set recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 480–13 489.

[35] X. Sun, C. Zhang, G. Lin, and K. V. Ling, "Open set recognition with conditional probabilistic generative models." *arXiv preprint arXiv:2008.05129*, 2020.

[36] J. Tack, S. Mo, J. Jeong, and J. Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 11 839–11 852.

[37] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning," *arXiv preprint arXiv:2005.10243*, 2020.

[38] A. Vyas, N. Jammalamadaka, X. Zhu, D. Das, B. Kaul, and T. L. Willke, "Out-of-distribution detection using an ensemble of self supervised leave-out classifiers," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 560–574.

[39] J. Winkens, R. Bunel, A. G. Roy, R. Stanforth, V. Natarajan, J. R. Ledsam, P. MacWilliams, P. Kohli, A. Karthikesalingam, S. Kohl, A. T. Cemgil, S. M. A. Eslami, and O. Ronneberger, "Contrastive training for improved out-of-distribution detection." *arXiv preprint arXiv:2007.05566*, 2020.

[40] H. M. Yang, X. Y. Zhang, F. Yin, Q. Yang, and C. L. Liu, "Convolutional prototype network for open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2358–2370, 2022.

[41] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, "Classification-reconstruction learning for open-set recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4016–4025.

[42] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[43] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *British Machine Vision Conference*, 2016.

[44] H. Zhang and V. M. Patel, "Sparse representation-based open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1690–1696, 2017.

[45] H. Zhang, A. Li, J. Guo, and Y. Guo, "Hybrid models for open set recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 102–117.

[46] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Learning placeholders for open-set recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4401–4410.

[47] E. Zisselman and A. Tamar, "Deep residual flow for novelty detection." 2020.

黄宏志（Hongzhi Huang） 于 2020 年获得天津大学计算机科学与技术专业学士学位，并在天津大学攻读硕士学位。他的研究兴趣是计算机视觉中的开放集识别和分布偏离检测。

王煜（Yu Wang） 分别于 2013 年和 2016 年、2020 年在天津大学获得通信工程学士学位、软件工程硕士学位和计算机应用与技术博士学位。他目前是天津大学的助理教授。他的研究兴趣是数据挖掘和机器学习，特别是在开放和动态环境下的多粒度学习，以及在计算机视觉和工业方面的应用。

胡清华（Qinghua Hu） 于 1999 年、2002 年和 2008 年分别在中国哈尔滨工业大学获得学士、硕士和博士学位。之后，他加入香港理工大学计算机系，担任博士后研究员。他于 2012 年成为天津大学的全职教授，现在是情报与计算学院的讲座教授和副院长。这些年来，他的研究兴趣集中在不确定性建模、多模态学习、增量学习和持续学习方面，受到国家自然科学基金和中国国家重点研发计划的资助。他在 IEEE TKDE, IEEE TPAMI, IEEE TNNLS 等发表了 300 多篇同行评审的论文。他是 ICMLC 2015 和 ICME 2021 最佳论文奖的获得者。他是 IEEE Transactions on Fuzzy Systems、 ACTA AUTOMATICA SINICA 和 ACTA ELECTRONICA SINICA 的副编辑。

程明明（Ming-Ming Cheng） 2012 年在清华大学获得博士学位，随后在牛津大学与 Philip Torr 教授合作了 2 年。他现在是南开大学的教授，领导媒体计算实验室。他的研究兴趣包括计算机视觉和计算机图形。他获得的奖项包括 ACM 中国新星奖，IBM 全球 SUR 奖等。他是 IEEE 的高级会员，是 IEEE TPAMI 和 IEEE TIP 的编委会成员。