

考虑类间差异损失的中华白海豚个体识别

张东晓¹⁾, 袁梦¹⁾, 祝茜²⁾, 王先艳^{3,4)*}

¹⁾ (集美大学理学院 厦门 361021)

²⁾ (山东大学海洋学院 威海 264209)

³⁾ (自然资源部第三海洋研究所 厦门 361005)

⁴⁾ (福建省海洋生态保护与修复重点实验室 厦门 361005)
(wangxianyan@tio.org.cn)

摘要: 为了提升中华白海豚个体识别的准确率, 提出基于类间差异损失的细粒度识别模型. 首先基于不同类别图像的特征之间具有差异性的事实, 设计类间差异损失函数; 其次在 VGG16 的基础上, 根据类别数调整部分卷积层的通道数, 并设计 2 种全连接方式用于识别不同规模的个体; 最后将所提损失函数与已有损失进行组合, 激励网络学习到更具区分度的特征. 选取在厦门湾拍摄的 2 177 幅中华白海豚图像, 人工标注为 30 头个体作为数据集进行实验的结果表明, 所提损失函数可以将准确率提升 1.05 个百分点, 达到 98.65%, 且比主流的细粒度识别算法至少高出 0.9 个百分点.

关键词: 类间差异损失; 细粒度; 中华白海豚; 互通道损失

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.19983

Identification of *Sousa Chinensis* Considering the Inter-Class Loss

Zhang Dongxiao¹⁾, Yuan Meng¹⁾, Zhu Qian²⁾, and Wang Xianyan^{3,4)*}

¹⁾ (School of Science, Jimei University, Xiamen 361021)

²⁾ (Marine College, Shandong University, Weihai 264209)

³⁾ (Third Institute of Oceanography, Ministry of Natural Resources, Xiamen 361005)

⁴⁾ (Fujian Provincial Key Laboratory of Marine Ecological Conservation and Restoration, Xiamen 361005)

Abstract: To improve the accuracy of individual identification of *Sousa Chinensis*, a fine-grained identification model based on the loss of inter-class difference is proposed. Firstly, based on the fact that the features of different types of images are different, a loss function of inter-class difference is designed; Secondly, the channel number of several convolution layers about VGG16 is adjusted according to the number of categories. And two full connection modes are designed for identification of large-scale and small-scale individuals respectively; Finally, the proposed loss and the existing loss are combined to stimulate the network to learn more distinguishing features. 2 177 images of *Sousa Chinensis* taken in Xiamen Bay are selected and manually labeled as 30 individuals. Experiments on this dataset show that the proposed loss function can improve the accuracy rate by 1.05 percentage points to 98.65%, which is at least 0.9 percentage points higher than the mainstream fine-grained identification algorithms.

Key words: inter-class loss; fine-grained; *Sousa Chinensis*; mutual channel loss

收稿日期: 2022-09-22; 修回日期: 2023-02-20. 基金项目: 国家自然科学基金(42076159); 福建省自然科学基金(2020J01710, 2021J06031); 集美大学国家基金培育计划(ZP2020063). 张东晓(1980—), 男, 博士, 副教授, 硕士生导师, 主要研究方向为机器视觉、数字图像处理、智能信息处理; 袁梦(2000—), 女, 在校学生, 主要研究方向为机器视觉、数字图像处理; 祝茜(1966—), 男, 博士, 教授, 博士生导师, 主要研究方向为海洋哺乳动物保护生物学; 王先艳(1982—), 男, 博士, 研究员, 硕士生导师, 论文通信作者, 主要研究方向为海洋哺乳动物种群生态学、行为学及栖息地保护管理.

中华白海豚(简称白海豚)是我国国家一级重点保护水生哺乳动物,也是濒危野生动植物物种国际贸易公约(CITES)附录 I 物种(CITES, 2011),被国际自然保护联盟(IUCN)红色名录列为“易危”物种(IUCN, 2018). 白海豚地方种群的数量及分布一直以来备受关注^[1-3],探讨其数量变化的动态和原因,有助于研究人员进一步掌握和预测种群动态信息,促进白海豚保护^[4]. 但是,传统的个体识别需要工作人员具备较强专业背景和专业技能,且通常耗时耗力. 所以,借助图像识别技术对白海豚进行识别和确认,是一项十分有意义的工作.

早期的标记重捕法^[5]是一种基于图像识别技术的非侵入、非破坏的种群采样方法. 当前,借助深度神经网络强大的特征提取能力,通过深度学习技术对白海豚开展个体辨识,虽然取得了一定的进展^[6],但这种通用图像识别模型往往难以取得进一步的突破. 这是因为该识别任务仅能够分辨物体大类,而对白海豚个体识别需要借助背鳍的颜色、斑点分布和缺刻等细微特征,这对算法提出了更高的要求.

对于这种具有细微差异的个体识别问题,细粒度图像识别技术^[7]行之有效,如用于对鱼类^[8]和一般野生动物^[9]进行识别,以及对服装图像进行分类^[10]. 细粒度图像识别较为朴素的思路是先找到前景对象和局部区域^[11],然后提取这些区域的特征,最后用所得特征训练识别模型. Zhang 等^[12]基于目标检测获得具有区分度的特征区域; Lin 等^[13]设计 2 个特征提取器获取图像中标注区域的特征,并对它们进行融合,可以很好地提取到标注区域的细粒度特征. 但是,这类方法需要标注出区分度位置,标注的工作量很大.

为了解决上述问题,一些学者利用弱监督学习定位信息量丰富的区域,较常见的做法是在网络中引入自注意力模块. Ji 等^[14]提出注意力卷积二值神经网络(attention convolutional binary neural tree, ACNet)网络,使用注意力变换模块强化网络,获取有区分度的特征; Ding 等^[15]提出注意力金字塔卷积神经网络(attention pyramid convolutional neural network, AP-CNN),将通道注意力(channel-wise attention, CWA)和空间注意力引入特征金字塔结构中,以精准地定位不同尺度的高分度区域.

除了使用自注意力机制,也有学者提出一些新颖的思路. Wang 等^[16]设计一种动态位置感知网络结构,使用图卷积将位置编码模块学习到的位置信息与具有判别力的视觉内容对齐,消除了独

立使用位置信息可能引入非判别信息的问题.

另一种新颖的思路是通过专门设计的损失函数,激励网络学习到精细化的特征. Zhang 等^[17]设计了广义 Triplet 损失,与传统的 Softmax 损失函数一起微调网络. Chang 等^[18]进一步指出,无需设计过于复杂的网络结构和训练机制,通过精巧的损失函数就能达到较好的细粒度识别效果,通过设计互通道(mutual channel, MC)损失函数可以激励网络按类别将特征分组,使得同一组的特征刻画同一个类别,且同类别的特征尽可能具有多样性. 实验结果表明,MC 损失确实实现了预期目标,但其只考虑同类特征的多样性,忽视了类间特征之间的差异性,而类间差异(inter-class differences, ICD)在图像识别特别是细粒度识别中更重要.

为了弥补 MC 损失在 ICD 方面的缺陷,本文在 MC 损失的基础上,进一步设计能够描述特征之间差异性的损失函数,激励网络学到更具区分度的特征. 在丰富细粒度图像识别理论研究的同时,进一步提升白海豚的识别率.

1 MC 损失

针对细粒度图像识别问题,Chang 等^[18]设计了 MC 损失函数,与经典的交叉熵(cross entropy, CE)损失一起构成总损失

$$L'(\mathbf{F}) = L_{\text{CE}}(\mathbf{F}) + \mu \cdot L_{\text{MC}}(\mathbf{F}) \quad (1)$$

其中, $\mathbf{F}$ 表示输入图像的特征,提取自成熟的卷积神经网络; $L_{\text{CE}}(\mathbf{F})$ 是 CE 损失,激励网络关注全局判别区域的特征; $L_{\text{MC}}(\mathbf{F})$ 是 MC 损失,引导网络关注局部判别区域的特征; μ 用于调节 $L_{\text{CE}}(\mathbf{F})$ 和 $L_{\text{MC}}(\mathbf{F})$ 的权重.

CE 损失较为经典,本文不再赘述. MC 损失由判别力组件和多样性组件构成. 判别力组件的核心作用是诱导网络按照类别将特征 $\mathbf{F}$ 自动分为 c 组 $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_c$, 使得第 i 组特征 $\mathbf{F}_i$ 专门负责第 i 类的识别任务; 其中, c 表示类别数,且每组特征的通道数均相等,记为 n . 多样性组件用来诱导组内特征尽可能具有多样性,使其关注到目标的不同区域.

为了叙述方便,将 $\mathbf{F}$ 的每个特征通道按行优先的方式重排为 $W \cdot H$ 维列向量,其中, W 和 H 分别表示特征的宽度和高度.

计算判别力组件之前,需要先计算每组特征归于对应类别的置信值. 定义映射 g , 将特征组

F_i 映射为预测成第 i 类的置信值, 即

$$g(F_i) = \frac{1}{W \cdot H} \sum_{k=1}^{W \cdot H} \underbrace{\max_{j=1,2,\dots,n}}_{\text{CCMP}} \underbrace{[F_i \cdot M_i]_{k,j}}_{\text{CWA}} \quad (2)$$

其中, $[.]_{k,j}$ 表示矩阵的第 k 行、第 j 列元素; M_i 表示 n 阶对角阵, 对角线元素中有一半被随机置为 0, 剩下一半为 1. 通过 M_i 可以激励网络, 让每个通道包含尽可能多的判别信息, 实现针对 F_i 的 CWA. 跨通道最大池化 (cross-channel max pooling, CCMP)^[19] 可以保留特征通道的峰值, 并发现有利于细粒度识别的信息. 全局平均池化 (global average pooling, GAP)^[20] 对空间维度的信息进行平均, 可以获得第 i 类预测结果的置信值.

令 $g(F) = (g(F_1), g(F_2), \dots, g(F_c))^T$, 并对其进行 Softmax 归一化, 得到各类预测概率值; 然后计算概率值与类别向量 y 的 $L_{\text{CE}}(F)$, 得到判别力组件

$$L_{\text{dis}}(F) = L_{\text{CE}}(y, S(g(F))) \quad (3)$$

其中, $S(\cdot)$ 是 Softmax 归一化函数, 定义为

$$S(x) = \frac{1}{\sum_{i=1}^m e^{x_i}} \cdot (e^{x_1}, e^{x_2}, \dots, e^{x_m}) \quad (4)$$

其中, $x = (x_1, x_2, \dots, x_m)^T$ 表示 m 维向量.

多样性组件定义为

$$L_{\text{div}}(F) = \frac{1}{c} \sum_{i=1}^c h(F_i) \quad (5)$$

其中, $h(F_i)$ 用来衡量 F_i 的多样性, 定义为

$$h(F_i) = \sum_{k=1}^{W \cdot H} \max_{j=1,2,\dots,n} [S(F_i)]_{k,j} \quad (6)$$

$S(F_i) = (S(F_i^{(1)}), S(F_i^{(2)}), \dots, S(F_i^{(n)}))$ 表示对 F_i 的每列单独使用 Softmax 所得的结果.

多样性组件的作用是隐式地发现图像中不同的鉴别区域. 由式(5)可知, $L_{\text{div}}(F)$ 越大, 模型在不同的通道越有可能关注到差异性区域, 所以训练时希望多样性组件的值趋于变大. 与此相反, 判别力组件的训练目标是值趋于变小, 所以定义

$$L_{\text{MC}}(F) = L_{\text{dis}}(F) - \lambda \cdot L_{\text{div}}(F) \quad (7)$$

其中, λ 用于调节判别力组件和多样性组件的权重.

2 本文模型

2.1 ICD 损失

图像识别的关键是特征, 同类图像的特征要有一致性, 不同类图像的特征要有差异性. 已有的

算法通常将注意力放在第一点上, 即尽可能提取使同类图像具有一致性的特征. Chang 等^[18]提出的基于判别力组件和多样性组件的 MC 损失也是这样的思路, 该损失使模型可以直接在通道上学习局部可区分特征, 且每个特征通道都关注不同的可判别区域; 但是这 2 个组件均未关注不同类图像特征之间的差异性.

在诸如白海豚个体识别的细粒度问题中, 由于不同白海豚个体之间的差别较为细微, 激励模型针对不同白海豚个体提取到具有差异化的特征显得尤为重要. 因此, 本文在 MC 损失的基础上考虑 ICD, 使不同白海豚个体的特征尽可能有差异.

本文针对白海豚个体进行识别, 所以也将类别描述为具体的白海豚个体, 如上文中的第 i 类即表示第 i 头白海豚; 此外, 本文用于白海豚识别的 ICD 思想也适用于一般的细粒度图像识别问题.

为了全面地利用组内各通道信息, 本文在通道方向对 F_i 进行平均, 即对 F_i 执行跨通道平均池化 (cross-channel average pooling, CCAP). 为了使不同特征组处于同一尺度, 使用 Softmax 对 CCAP 所得向量进行归一化, 并将归一化结果记为 $\bar{F}_i$, 称为第 i 头白海豚的综合特征. 该计算过程表示为

$$\bar{F}_i = S \left(\underbrace{\text{avg}_{j=1,2,\dots,n} [F_i]_{:,j}}_{\text{CCAP}} \right) \quad (8)$$

其中, $[.]_{:,j}$ 表示矩阵的第 j 列. 由式(4)可知, $\bar{F}_i$ 的分量均大于 0, 且和为 1.

为了与损失的最小化保持一致, 本文通过最小化特征之间的相似度来最大化差异. 为此针对各综合特征 $\bar{F}_1, \bar{F}_2, \dots, \bar{F}_c$, 首先计算它们两两之间的余弦相似度; 然后对所有相似度进行平均, 作为新的损失来激励网络学习更具区分度的特征, 即

$$L_{\text{dif}}(F) = \frac{2}{c(c-1)} \sum_{i=1}^c \sum_{j=i+1}^c \frac{\bar{F}_i \cdot \bar{F}_j}{\|\bar{F}_i\| \cdot \|\bar{F}_j\|} \quad (9)$$

其中, $\bar{F}_i \cdot \bar{F}_j$ 是 $\bar{F}_i$ 和 $\bar{F}_j$ 的内积, $\|\cdot\|$ 是向量的 2 范数.

由于 $\bar{F}_i$ 和 $\bar{F}_j$ 中每个分量均大于 0, 因此 $\bar{F}_i \cdot \bar{F}_j > 0$, 从而 $L_{\text{dif}}(F) > 0$. 此外, 在 $\bar{F}_i$ 各元素和为 1 的条件下, 不难证明 $\|\bar{F}_i\| \geq 1/\sqrt{W \cdot H}$; 而在成熟卷积神经网络中, 最后所得特征的空间维度往往不是很高, 即 $W \cdot H$ 不是很大. 因此在训练过程中, 无需考虑式(9)中分母是否接近 0.

本文将式(9)称为 ICD 损失函数, 并与 $L_{\text{MC}}(F)$ 和 $L_{\text{CE}}(F)$ 一起构成本文模型的总损失

$$L(F) = L_{\text{CE}}(F) + \mu \cdot L_{\text{MC}}(F) + \gamma \cdot L_{\text{dif}}(F) \quad (10)$$

其中, μ 和 γ 分别用来调节 $L_{MC}(F)$ 和 $L_{dif}(F)$ 的权重.

2.2 模型结构设计

由于白海豚的个体差异集中在背鳍, 本文基

于 YOLO^[21] 训练背鳍检测模型, 并基于此提出用于白海豚识别的模型结构, 包含 2 部分: 图 1 所示为用于提取特征的卷积神经网络结构; 图 2 所示为损失计算模块.

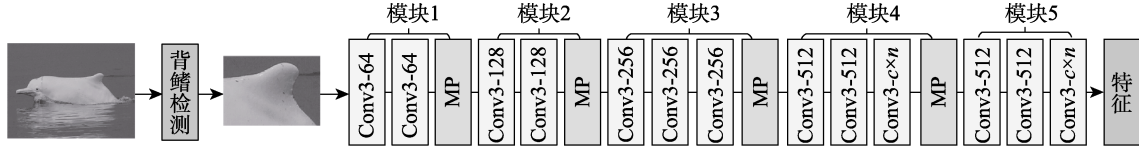


图 1 特征提取的结构

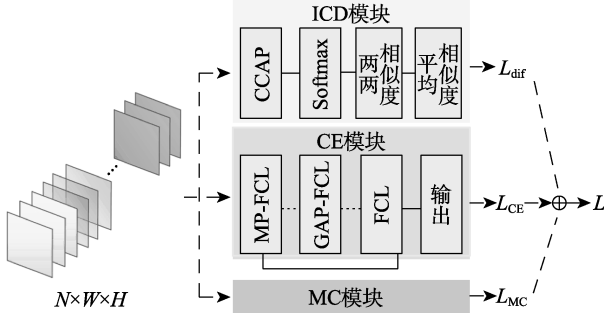


图 2 损失计算模块的结构

图 1 中, 针对背鳍图, 本文以 VGG16^[22]为基础设计用于提取特征的网络, 包含 5 个模块, 共 13 个卷积层和 4 个最大池化(max pooling, MP)层. 所有隐藏层的激活函数均采用 ReLU^[23], 卷积层均使用多个堆叠的 3×3 卷积核, 池化层均使用 2×2 池化. 因为输入图像的大小需要调整到 224×224 像素, 所以在 4 次池化之后, 特征的空间维度降为 14×14 , 即所提特征的空间维度是 14×14 .

由于特征组 F_i 的通道数是 n , 且有 c 组特征, 因此特征的通道数为 $c \cdot n$. 为此, 对 VGG16 的模块 5 进行微调, 将最后的卷积层的通道数设为 $c \cdot n$; 相应地, 对模块 4 也进行调整, 使它的卷积层结构与模块 5 保持一致, 实现模块的堆叠. 模块 1~模块 3 沿用了 VGG16 的结构.

与式(10)对应, 图 2 所示的损失计算模块包含用于计算 $L_{dif}(F)$, $L_{CE}(F)$ 和 $L_{MC}(F)$ 的子模块. 其中, 左侧是图 1 所示卷积神经网络输出的特征图 $F \in \mathbb{R}^{cn \times 14 \times 14}$.

在进行损失计算时, 将特征 F 同步送入 ICD 模块、CE 模块和 MC 模块, 可以得到 $L_{dif}(F)$, $L_{CE}(F)$ 和 $L_{MC}(F)$. 在 CE 模块中, 不需调整特征维度; 在另外 2 个模块中, 需要先将特征组 F_i 的每个通道重排成列向量.

MC 模块详见文献[18], 本文不再赘述, 此处重点阐述 ICD 模块和 CE 模块. 由式(9)可知, 在计算 $L_{dif}(F)$ 时, 先对每个特征组 F_i 做 CCAP, 沿着通道维度合并特征, 得到对应的综合特征; 然后使用 Softmax 进行归一化, 得到归一化的综合特征 $\bar{F}_i$; 再针对所有综合特征, 使用余弦相似度分别度量两两之间的相似性; 最后对所有相似度进行平均, 得到 $L_{dif}(F)$.

计算 $L_{CE}(F)$ 时, 遵循通用卷积神经网络的做法, 如 VGG^[22], 对提取到的特征 F 做 MP, 将其空间维度降到 7×7 . 事实上, 本文使用 MP 也是出于对数据量的考虑. 由于种种因素, 厦门湾的白海豚种群数量骤减, 由 20 世纪 90 年代的不足 100 头减少到目前成年个体已不超过 50 头^[4]. 因此在类别数较少的情况下, 需要进一步降低特征的维度, 避免出现拟合现象.

通常, 对 MP 后的特征有 2 种处理方式: 一是将特征中每个通道上的每个位置视为 1 个神经元, 所有神经元构成 1 个全连接层(fully connected layer, FCL); 另一种是使用 GAP 将每个通道的平均值视作 1 个神经元, 所有通道对应的神经元组成 1 个 FCL. 为了叙述方便, 本文将第 1 种方式称为 MP-FCL, 第 2 种方式称为 GAP-FCL. 显然, MP-FCL 的神经元个数为 $49 \cdot N$, GAP-FCL 的神经元个数为 N . 在种类较多的情况下, 可以优先使用 GAP-FCL; 而在种类较少的情况下, 优先选择 MP-FCL. 本文将在实验中比较这 2 种方式的效果.

输出层的神经元个数是 c , 在输出层之前再增加 1 个 FCL, 其神经元个数记为 N_h . N_h 的取值由经验确定: 如果使用 MP-FCL, 那么 N_h 大一些; 如果使用 GAP-FCL, 那么 N_h 小一些. 具体数值依据这个原则由经验来确定, 本文实验中, 在 $c = 30$ 的情况下, 使用 MP-FCL 时 $N_h = 512$, 使用 GAP-FCL 时 $N_h = 60$.

3 实验结果与分析

3.1 实验环境

本文在 Ubuntu 18.04 系统中搭建实验环境: Pytorch1.2+CuDNN7.0+Python3.6; 服务器的核心配置为 Intel(R) Xeon(R) Silver 4110 CPU, 64 GB 内存, Tesla P100 显卡, 16 GB 显存.

训练模型时使用了 SGD 优化器, 初始学习率为 0.1, 在 150 轮次时调整为 0.01, 225 轮次时调整为 0.001, 共训练 300 轮.

白海豚数据是 2010—2015 年在厦门湾拍摄的, 均已人工标注 ID 信息, 包含 30 头个体, 共有白海豚图像 2177 幅. 为了叙述方便, 将其记作 DD30

数据集.

3.2 对比实验和消融实验

将 DD30 数据集按照 7 : 3 随机划分为训练集和测试集, 最终训练集有 1509 幅图像, 测试集有 668 幅图像.

3.2.1 GAP-FCL 和 MP-FCL 的对比

公平起见, 针对 GAP-FCL 和 MP-FCL 均单独调参, 各自选择最优的超参数. MP-FCL 的最优参数为 $n=3$, $\gamma=2$; GAP-FCL 的最优参数为 $n=3$, $\gamma=1$. 式(10)中的 μ 和式(7)中的 λ 均参考文献[18], 设置为 $\mu=0.005$, $\lambda=10$. 在训练集和测试集的准确率及损失随训练轮次的变化情况如图 3 所示.

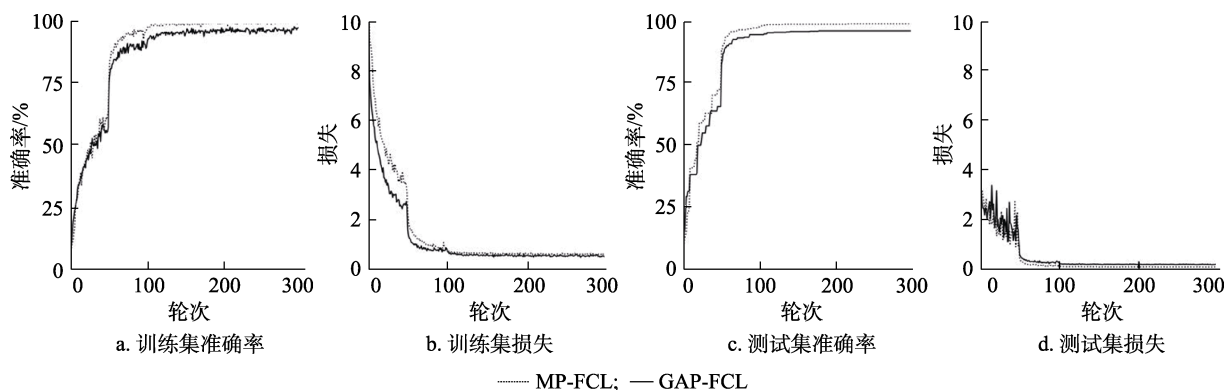


图 3 GAP-FCL 和 MP-FCL 的准确率和损失变化曲线

从图 3 可以看出: (1) 在训练过程中, GAP-FCL 和 MP-FCL 均已收敛, 而且测试集的准确率和损失与训练集的趋势基本吻合, 但是 GAP-FCL 在训练集上的准确率有轻微振荡, 而 MP-FCL 较为平稳, 特别是训练 100 轮以后, 准确率和损失均保持平稳. (2) 测试集的损失均低于训练集的, 而训练集的准确率均高于测试集的准确率, 这似乎与常理不符. 这是由于为了展示本文提出的损失模块, 在训练集上计算损失时使用式(10), 记录的是总损失; 为了描述模型在识别时真正的损失(即预测结果的误差), 在测试集上只计算了 CE 损失, 而 CE 损失是式(10)的一部分, 因此导致测试集的损失低于训练集的损失.

表 1 所示为这 2 种 FCL 在测试集上最终的准确率对比. 可以看出, GAP-FCL 的准确率比 MP-FCL 低, 原因在于现有数据集中白海豚个体数量是 30, 所以 GAP-FCL 层的神经元个数 $N=30 \cdot n$, 比 MP-FCL 层的神经元个数小很多, 当 n 较小时, GAP-FCL 更容易导致模型出现欠拟合; 当然也可以选择较大的 n 来增加 GAP-FCL 层神经元个数,

但是这无益于改善识别效果, 因为 n 变大时, 意味着更多的通道参与刻画白海豚个体, 而过量的通道数容易导致过拟合, 进而降低识别准确率. 这一点在下文关于 n 的实验中也得到了证实. 同时这也正是上文中提到的如何选择 GAP-FCL 和 MP-FCL 的出发点: 当种类较多时, 可以优先使用 GAP-FCL, 而当种类较少时, 优先选择 MP-FCL. 针对白海豚数据, 考虑 MP-FCL 的准确率高于 GAP-FCL, 下文的实验均使用 MP-FCL.

表 1 全连接方式对准确率的影响 %

方法	训练集	测试集
MP-FCL	99.47	98.65
GAP-FCL	96.48	95.81

3.2.2 γ 的影响

将式(10)中 $L_{\text{diff}}(\mathbf{F})$ 的系数分别取 $\gamma=0.5, 1.0, 1.5, 2.0$ 和 2.5, 其余超参数与上一个实验保持一致的情况下, 识别准确率如表 2 所示. 可以看出, $\gamma=0.5, 1.0$ 和 2.0 时, 训练集的准确率基本相同, 但是测试集的准确率差别较大, $\gamma=2.0$ 时测试集

上的效果最好; $\gamma=1.5$ 时, 训练集和测试集的准确率都明显降低, 此时可能出现欠拟合; $\gamma=2.5$ 时, 训练集的准确率最高, 但是测试集的准确率大幅降低, 此时可能出现过拟合. 实验结果表明, $\gamma=2.0$ 时效果最好, 所以本文实验中将固定 $\gamma=2.0$.

表 2 γ 对识别准确率的影响 %

γ	训练集	测试集
0.5	99.47	97.31
1.0	99.47	98.20
1.5	98.87	97.16
2.0	99.47	98.65
2.5	99.53	97.60

3.2.3 特征组通道数 n 的影响

n 从 1~5 分别取值, 计算测试集的准确率, 结果如表 3 所示. 可以看出, 当 $n=1$ 时表现最差, 意味着对于每头白海豚只学习一个鉴别区域是不够的, 不足以区分不同的白海豚; 随着 n 增大, 鉴别区域也增多, 识别性能得到提升; 但是当 n 超过 3 时, 性能开始下降, 可能的原因是更多的鉴别区域导致模型过拟合. 综合考虑性能和计算成本, 将 n 设定为 3 较合理, 所以下文的实验中取 $n=3$.

表 3 n 对识别准确率的影响 %

n	准确率
1	34.43
2	96.86
3	98.65
4	97.90
5	96.71

3.2.4 针对组件的消融实验

针对多样性组件、判别力组件和 ICD 组件进行消融实验, 进一步考察 MC 损失和本文提出的 ICD 损失对识别效果的影响, 结果如表 4 所示. 其中, 用 A 表示多样性组件, B 表示判别力组件, C 表示 ICD 组件.

表 4 各种损失的消融实验对比 %

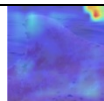
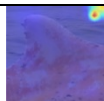
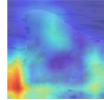
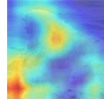
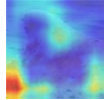
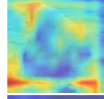
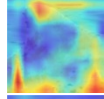
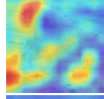
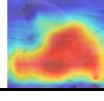
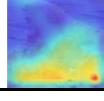
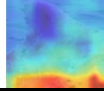
方法	准确率
CE 损失	90.57
CE 损失+A	91.17
CE 损失+B	92.06
CE 损失+C	91.92
CE 损失+B+A	97.60
CE 损失+B+C	97.75
CE 损失+A+C	88.47
CE 损失+B+A+C	98.65

注: 加粗表示最优结果.

从表 4 可以看出: (1) 当不使用上述组件而只用 CE 损失时, 准确率偏低, 结果不理想. (2) 在 CE 的基础上单独增加各组件, 均能小幅提升准确率, 其中, B 的提升幅度最大, 说明对特征类别进行约束(由组件 B 负责)是最重要的, 因为首先要确保特征能够根据类别被正确分组; 单独使用 C 的提升效果好于 A, 说明类间特征具有差异性(由组件 C 负责)比类内特征具有多样性(由组件 A 负责)更为重要. (3) 3 个组件两两组合时, 包含判别力组件的模式均大幅提升了准确率, 进一步验证了判别力组件的重要性, 也说明在特征被正确分组的前提下, 类间特征具有差异性和类内特征具有多样性均可以发挥巨大作用; 对比来看, 类间特征差异性的作用更大, 这与单独使用它们的结果是一致的. (4) 意外的是, 在 B 缺失时, A 和 C 出现互斥现象, 共用它们反倒使准确率降低, 这是因为当 B 缺失时, 无法确保特征能够根据类别被正确分组, 此时同时要求类内特征具有多样性(由组件 A 负责)和类间特征具有差异性(由组件 C 负责), 势必导致模型为了同时满足这 2 个条件, 突破 CE 损失对特征类别的约束, 造成较严重的特征错位, 以致效果变差; 至于这 2 个组件单独使用时并未拖累 CE 反倒有所裨益, 可能是因为单独使用时对 CE 损失尚不构成影响, 还能起到一定作用. (5) 3 个组件同时使用时效果最好, 表明 3 个组件相互配合可以达到最优, 同时也说明本文提出的 ICD 损失确实有助于提升识别效果, 改进了文献[18]的算法, 使得准确率提升 1.05%.

基于 Grad-CAM^[24]对单独使用上述组件以及综合使用的情况进行可视化, 结果如表 5 所示.

表 5 各组件对应的通道梯度热力图

组件	背鳍图	通道 1	通道 2	通道 3
A				
B				
C				
B+A+C				

从表 5 可以看出: (1) 当只加入多样性组件时, 只有第 1 个通道关注到了正确的鉴别区域, 其余 2

个通道均只关注到海面部分,无益于识别任务;当只加入判别力组件时,模型学习到的 3 个特征通道十分相似,不能关注到不同的鉴别区域;当只加入 ICD 损失时,对于同一头白海豚图像,特征通道关注的区域基本相同,也不能实现局部区域定位。(2) 加入 ICD 损失后关注到了更多的鉴别区域,说明 ICD 损失有助于学习到更多的鉴别区域;但是比只加入判别力组件时鉴别区域的范围广,导致鉴别区域不够精细,这也与表 4 中单独使用 ICD 组件的效果略逊于判别力组件是一致的。(3) 当同时使用 3 个组件时,3 个特征通道关注到不同的鉴别区域,如背鳍缺刻、曲线和斑点分布,进一步说明 3 个组件协同可以获得最好的识别效果。

3.2.5 与经典卷积神经网络的对比

选择 VGG^[22], ResNet^[25]和 GoogLeNet^[26-27]经典网络结构,采用迁移学习^[28]的策略,将 ImageNet 数据集上训练好的模型参数作为初始值,使用 DD30 数据集继续训练。训练时采用 2 种方式:一是冻结网络的特征层,只训练 FCL;二是所有层均参与训练。这些模型在测试集上的准确率对比如表 6 所示。可以看出,在冻结特征层的情况下,只有 VGG 网络的效果尚可,其余网络对 DD30 数据集均无效,说明网络的可训练参数不足,以致网络偏简单,导致模型欠拟合;当全部层均参与训练

时,识别效果有明显提升,在测试集的准确率均超过 90%,除了 ResNet18,其余均超过 95%,但是均不如本文效果,这可以从表 4 的最后一行看出。

对比表 4 和表 6 可以看出,本文只使用 CE 损失的准确率低于 VGG16,原因是本文的网络结构比 VGG16 少了通道数,如图 1 所示,模块 4 和模块 5 最后的卷积层通道数为 90(类别数 30 乘每类通道数 3),远远少于 VGG16 的 512,通道数变少使模型泛化能力变弱,导致准确率降低;另一方面,增加了一系列损失组件后,在未改变网络结构的前提下模型的泛化能力逐步提升,直至可以超越通道数更多的 VGG16。这也进一步说明,使用合适的损失函数确实能进一步提升模型的性能。

为了说明经典网络不存在训练不足的问题,对训练集和测试集上准确率和损失的变化情况进行实验。由于冻结特征层的效果较差,只展示所有层参与训练的结果,变化曲线如图 4 所示。从训练集的准确率和损失看,各模型已经收敛;从测试集的结果来看,这些模型的识别效果不是很理想,部分模型的准确率和损失波动较大,出现过拟合现象。这也说明基于卷积神经网络的通用图像识别模型在白海豚识别问题中具有一定的局限性。

3.2.6 与前沿算法的对比

本文挑选了近年来在各项视觉任务中表现优异的 ST(swin Transformer)^[29]及其升级版本 ST V2^[30];自监督预训练方法 SimMIM^[31]为视觉 Transformer 提供了很好的预训练模型,本文也与 SimMIM 和 ST 的结合进行比较;本文还选择了 ConvNeXt^[32],该结构借鉴 ST 的思想,在 ResNet 的基础上重新设计网络,在公开数据集上的表现超越了 ST;除此之外,本文还比较了细粒度算法 ACNet^[14]和 AP-CNN^[15]。各模型在训练集和测试集上的准确率对比如表 7 所示,这些指标均是用训练过程中得

表 6 经典卷积神经网络采用 2 种方式在测试集上的准确率对比 %

模型	冻结特征层	训练全部层
VGG16 ^[22]	89.37	95.80
VGG19 ^[22]	89.52	95.81
ResNet18 ^[25]	55.24	90.72
ResNet50 ^[25]	53.89	95.96
GoogLeNet ^[26]	39.07	95.66
InceptionV3 ^[27]	46.86	96.86

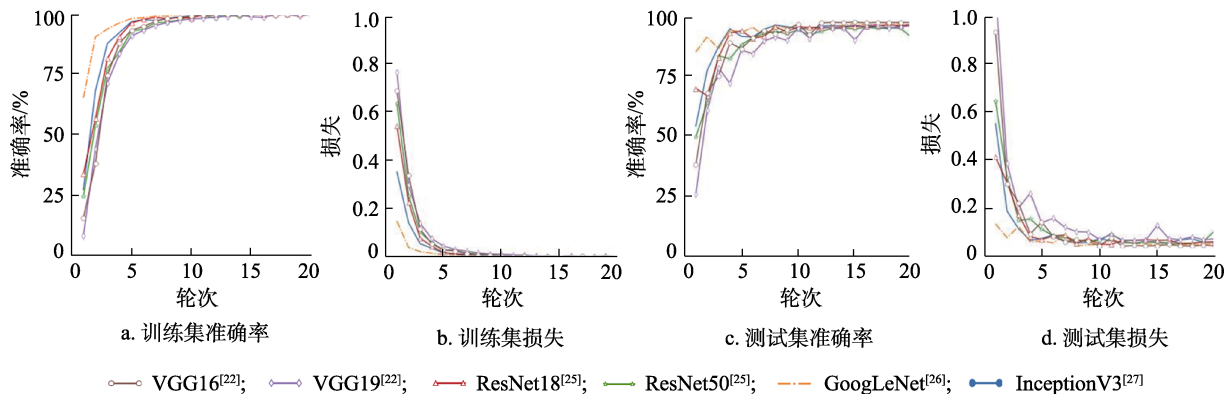


图 4 经典卷积神经网络的准确率和损失变化曲线

到的最优模型计算得出的. 其中, 训练对比模型时均加载了相关文献公开的预训练模型, 然后进行微调; 除了训练轮次, 其余参数均使用默认值; 训练轮次先设为 100, 如果损失和准确率趋于稳定, 那么停止训练, 否则依据收敛情况增加轮次继续训练, 直至指标趋于稳定; 训练与 ST 相关的 3 个模型时, 挑选了它们的最优版本; 此外, 2 种细粒度算法的骨干网络有 VGG19^[22]和 ResNet50^[25], 但是 VGG19 作为骨干网络的效果稍差些, 所以只展示 ResNet50 作为骨干网络的结果. 从训练集的准确率可知, 所有模型均已收敛; 从测试集的准确率看, ST 和 ACNet 略低, 没有超过 97%, 其余均超过 97%; 本文模型的准确率超出第 2 名将近 1 个百分点.

表 7 7 种方法的准确率对比 %

方法	训练集	测试集
ST ^[29]	99.54	96.86
ST V2 ^[30]	99.27	97.16
SimMIM ^[31] +ST ^[29]	99.20	97.01
ConvNeXt ^[32]	99.67	97.01
ACNet ^[14]	99.73	96.11
AP-CNN ^[15]	100.00	97.75
本文	99.87	98.65

注. 加粗表示最优结果.

3.2.7 模型实际效果展示

使用最终训练好的模型对 DD30 所有白海豚图像进行预测, 共有 15 幅图像出现识别错误, 其预测概率排在前 5 名的情况如表 8 所示. 文件名中下划线前的数字是白海豚编号, 下划线后的数字是照片的序号, 不同序号对应不同的拍摄环境及拍摄日期.

从表 8 可以看出: (1) 前 13 幅图像中, 正确编号均出现在预测概率排名前 5 的结果中, 其中, 第 2 名的结果中, 有 6 幅图像是预测正确的, 接近 1/2; 第 5 名的结果中, 有 4 幅图像是预测正确的, 约占 1/3; 整体来看, 13 幅图像中预测概率较大和较小的情况几乎各占一半. 在实际使用时, 可以根据需求同时输出预测概率排在前 5 的预测结果, 再以人工进行二次确认, 有助于降低错误率. (2) 最后 2 幅图像较为异常, 这些编号均未出现在预测概率排在前 5 的结果中. 为了进一步确认该异常情况是模型本身的问题, 还是数据标注出错, 从预测概率排名前 5 的预测类别中随机挑选 1 幅图像在图 5 中展示, 其中, 第 1 幅图像的文件名分别为 XM019_044 和 XM019_045, 后面 5 幅依次为预测概率排名前 5 的预测结果.

从图 5 可以看出, 这 2 幅图像的预测结果基本一致, 表明它们对应同一头白海豚, 这与 2 幅图的

表 8 DD30 数据集中预测错误的结果对比

文件名	海豚编号	预测概率排名前 5 的结果				
		第 1 名	第 2 名	第 3 名	第 4 名	第 5 名
XM005_003	XM005	XM032	XM036	XM033	XM038	XM005
XM010_014	XM010	XM013	XM010	XM038	XM020	XM011
XM010_027	XM010	XM030	XM046	XM032	XM010	XM033
XM011_046	XM011	XM021	XM011	XM009	XM012	XM014
XM013_060	XM013	XM019	XM013	XM029	XM020	XM012
XM013_061	XM013	XM019	XM013	XM029	XM046	XM020
XM019_043	XM019	XM020	XM010	XM013	XM018	XM019
XM020_003	XM020	XM010	XM011	XM013	XM038	XM020
XM020_004	XM020	XM013	XM011	XM010	XM012	XM020
XM024_064	XM024	XM028	XM008	XM031	XM024	XM023
XM034_042	XM034	XM003	XM012	XM034	XM011	XM021
XM034_043	XM034	XM003	XM034	XM012	XM037	XM011
XM036_003	XM036	XM005	XM036	XM038	XM018	XM011
XM019_044	XM019	XM020	XM010	XM013	XM018	XM028
XM019_045	XM019	XM020	XM010	XM018	XM013	XM036

注. 加粗表示预测正确的白海豚编号.

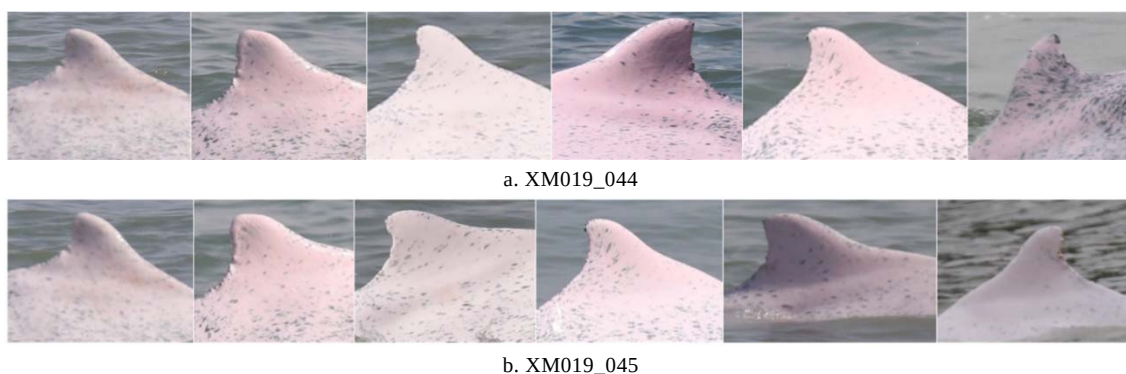


图 5 预测结果异常的背鳍图

编号信息一致,反映出模型具有较好的稳定性.仔细观察预测为 XM020 的图像,与目标白海豚的背鳍缺刻形状很接近,而且整体颜色及斑点分布也基本一致,经过专家进一步确认, XM019_044 和 XM019_045 与 XM020 是同一头白海豚,由于之前人工标注出现失误导致编号有误,进一步说明本文模型可以很好地辅助相关工作人员辨识和确认白海豚个体.在第 2 幅~第 5 幅背鳍图中,被预测的个体有的缺刻与目标一样,有的斑点分布与目标相似,有的形状与目标接近,说明该模型确实学到了关于白海豚的细粒度特征:缺刻、背鳍形状和斑点分布.

4 结 语

针对白海豚这种细粒度识别问题,本文关注不同个体的特征之间的差异性,设计了 ICD 损失函数,在成熟卷积神经网络的基础上构建了用于白海豚识别的模型.该模型可以为从事白海豚保护和研究的相关工作人员在白海豚个体识别和确认方面提供便利,节约时间和人力成本.

本文提出的网络结构可以直接用于其他动物种类的个体识别,只需对个体进行标注,即可训练相应的细粒度识别模型.

本文所提 ICD 损失具有普适性,除了 VGG,也可以相应地用在 ResNet, GoogLeNet 等网络结构中,丰富了关于细粒度图像识别的理论研究.未来将专门针对细粒度图像识别问题,探索将本文提出的损失函数用于视觉 Transformer, ConvNeXt 等新型网络结构中的可行性.

致谢 集美大学理学院客座教授刘榕和在数据预处理、模型训练等方面提出诸多宝贵建议,在此表示感谢!

参考文献(References):

- [1] Liu Wenhua, Huang Zongguo. Distribution and abundance of Chinese white dolphins (*Sousa chinensis*) in Xiamen[J]. Acta Oceanologica Sinica, 2000, 22(6): 95-101(in Chinese)
(刘文华, 黄宗国. 厦门中华白海豚的分布和数量[J]. 海洋学报, 2000, 22(6): 95-101)
- [2] Wu Fuxing. Investigations on distribution and population size of Chinese white dolphin (*Sousa chinensis*) along the coastal waters of Fujian province, People's Republic of China[D]. Jinan: Shandong University, 2013(in Chinese)
(吴福星. 福建沿海中华白海豚的分布与种群数量调查[D]. 济南: 山东大学, 2013)
- [3] Zhong Mingding. Research on population ecology of Indo-Pacific humpback dolphin (*Sousa chinensis*) in Xiamen bay[D]. Xiamen: Third Institute of Oceanography, Ministry of Natural Resources, 2021(in Chinese)
(钟铭鼎. 厦门湾中华白海豚种群生态学研究[D]. 厦门: 自然资源部第三海洋研究所, 2021)
- [4] Wang Xianyan, Zhong Mingding, Wu Fuxing, et al. Indo-Pacific humpback dolphins (*Sousa chinensis*) in Xiamen bay: threats and conservation[J]. Journal of Applied Oceanography, 2021, 40(1): 154-162(in Chinese)
(王先艳, 钟铭鼎, 吴福星, 等. 厦门湾中华白海豚: 面临的威胁与保护对策[J]. 应用海洋学学报, 2021, 40(1): 154-162)
- [5] Keith M, Peddemors V M, Bester M N, et al. Population characteristics of Indo-Pacific humpback dolphins at Richards bay, South Africa: implications for incidental capture in shark nets[J]. South African Journal of Wildlife Research, 2002, 32(2): 153-162
- [6] Bogucki R, Cygan M, Khan C B, et al. Applying deep learning to right whale photo identification[J]. Conservation Biology, 2019, 33(3): 676-684
- [7] Wei X S, Song Y Z, Mac Aodha O, et al. Fine-grained image analysis with deep learning: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 8927-8948
- [8] Zhang Yating. Research on fish fine-grained image recognition algorithm based on deep learning[D]. Qingdao: Qingdao University of Science & Technology, 2021(in Chinese)

- (张亚婷. 基于深度学习的鱼类细粒度图像识别算法研究[D]. 青岛: 青岛科技大学, 2021)
- [9] Bain M, Nagrani A, Schofield D, *et al.* Count, crop and recognise: fine-grained recognition in the wild[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop. Los Alamitos: IEEE Computer Society Press, 2019: 236-246
- [10] Wu Miaomiao, Liu Li, Fu Xiaodong, *et al.* Fine-grained clothing image classification by style feature description[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(5): 780-791(in Chinese)
(吴苗苗, 刘骊, 付晓东, 等. 款式特征描述符的服装图像细粒度分类方法[J]. 计算机辅助设计与图形学学报, 2019, 31(5): 780-791)
- [11] Luo Jianhao, Wu Jianxin. A survey on fine-grained image categorization using deep convolutional features[J]. Acta Automatica Sinica, 2017, 43(8): 1306-1318(in Chinese)
(罗建豪, 吴建鑫. 基于深度卷积特征的细粒度图像分类研究综述[J]. 自动化学报, 2017, 43(8): 1306-1318)
- [12] Zhang N, Donahue J, Girshick R, *et al.* Part-based R-CNNs for fine-grained category detection[C] //Proceedings of the 13th European Conference on Computer Vision. Heidelberg: Springer, 2014: 834-849
- [13] Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 1449-1457
- [14] Ji R Y, Wen L Y, Zhang L B, *et al.* Attention Convolutional binary neural tree for fine-grained visual categorization[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10465-10474
- [15] Ding Y F, Ma Z Y, Wen S G, *et al.* AP-CNN: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification[J]. IEEE Transactions on Image Processing, 2021, 30: 2826-2836
- [16] Wang S J, Li H J, Wang Z H, *et al.* Dynamic position-aware network for fine-grained image recognition[C] //Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 2791-2799
- [17] Zhang X F, Zhou F, Lin Y Q, *et al.* Embedding label structures for fine-grained feature representation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1114-1123
- [18] Chang D L, Ding Y F, Xie J Y, *et al.* The Devil is in the channels: mutual-channel loss for fine-grained image classification[J]. IEEE Transactions on Image Processing, 2020, 29: 4683-4695
- [19] Goodfellow I J, Warde-Farley D, Mirza M, *et al.* Maxout networks[C] //Proceedings of the 30th International Conference on Machine Learning. Lille: PMLR Press, 2013: 1319-1327
- [20] Lin M, Chen Q, Yan S C. Network in network[OL]. [2022-09-22]. <https://arxiv.org/abs/1312.4400>
- [21] Redmon J, Divvala S, Girshick R, *et al.* You only look once: unified, real-time object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 779-788
- [22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2022-09-22]. <https://arxiv.org/abs/1409.1556>
- [23] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90
- [24] Selvaraju R R, Cogswell M, Das A, *et al.* Grad-CAM: visual explanations from deep networks via gradient-based localization[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 618-626
- [25] He K M, Zhang X Y, Ren S Q, *et al.* Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778
- [26] Szegedy C, Liu W, Jia Y Q, *et al.* Going deeper with convolutions[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 1-9
- [27] Szegedy C, Ioffe S, Vanhoucke V, *et al.* Inception-v4, inception-ResNet and the impact of residual connections on learning[C] //Proceedings of the 31st AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 4278-4284
- [28] Yosinski J, Clune J, Bengio Y, *et al.* How transferable are features in deep neural networks?[C] //Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 3320-3328
- [29] Liu Z, Lin Y T, Cao Y, *et al.* Swin transformer: hierarchical vision transformer using shifted windows[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 9992-10002
- [30] Liu Z, Hu H, Lin Y T, *et al.* Swin transformer V2: scaling up capacity and resolution[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 11999-12009
- [31] Xie Z D, Zhang Z, Gao Y, *et al.* SimMIM: a simple framework for masked image modeling[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 9643-9653
- [32] Liu Z, Mao H Z, Wu C Y, *et al.* A ConvNet for the 2020s[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 11966-11976