掩码自编码器是高效的类增量学习器

翟江天 [1] 刘夏雷 [1,*] Andrew D. Bagdanov [2] 李珂 [3] 程明明 [1]
[1] VCIP, CS, 南开大学 [2] MICC, 佛罗伦萨大学 [3] 腾讯优图实验室

摘要

类增量学习（CIL）旨在按顺序学习新的类别，同时避免对于先前知识的灾难性遗忘。我们提出使用掩码自编码器（MAEs）作为 CIL 的高效学习器。MAEs 最初的设计目的是通过重构无监督学习来学习有用的表征，它们可以很容易地与监督损失集成以进行分类。此外，MAEs 还能从随机选取的块中可靠地重建原始输入图像，在 CIL 中，我们用它来更有效地存储来自过去任务的范例。我们还提出了一种双边 MAE 框架，用于从图像级和嵌入级的融合中学习，从而产生更高质量的重建图像和更稳定的表征。我们的实验证实，与 CIFAR-100、ImageNet-Subset 和 ImageNet-Full 的最新技术相比，我们的方法实现了更加卓越的性能。代码可在https://github.com/scok30/MAE-CIL获取。

1. 引言

在过去的十年中，深度学习对大多数计算机视觉任务产生了广泛而深远的影响。鉴于人类在其一生中不断学习的方式，人们自然而然地期望模型也能够积累知识并建立过去的经验，以增量式地适应新任务。但现实世界是动态的，这导致随着时间的推移，数据分布发生变化，而深度模型在学习新任务时往往会灾难性地遗忘旧任务 [26]。

类增量学习（CIL）旨在按顺序学习新的分类任务，同时避免灾难性遗忘 [2,25]。CIL 方法大致可分为三类 [5]，即基于排练的方法 [13,29,31]、基于正则化的方法 [15,17] 和基于架构的方法 [1,23,24]。其中基于排练的方法通过存储过去任务中的范例或生成合成样本进行重放，从而达到最先进的性能。

通常，在增量学习过程中只允许使用固定大小的内存。因此，从过去任务中存储的范例受到了限制。其他工作利用生成式网络 [35,38,43]（如 GANs）来合成来自旧任务的样本从而进行重放。虽然这些方法可以生成重放数据以减轻遗忘，但其典型缺点是生成的图像质量不高，而且遗忘同样可能发生在生成式模型中。在这项工作中，我们引入了掩码自编码器（MAEs）[11] 作为重放的基础模型。它只需要一小部分图像块就能重建整个图像，从而实现高效的范例存储。因此，与其它基于范例的方法相比，我们可以用同样有限的内存存储更多的范例。与之前的生成式方法相比，基于 MAE 的重放方法更稳定，因为它使用部分线索来推断全局信息，而全局信息与任务无关，在不同任务中遗忘较少。该方法通过固定的图像块缓解了 GANs 在不同任务中的不稳定生成效应。

掩码自编码器（MAE）[11] 最初为在自监督学习场景中学习更好的特征表征而提出。在这项工作中，我们将其视为高效的类增量学习器，同时提出了一种新颖的双边 Transformer 架构，用于在 CIL 中高效重放范例。我们的主要想法很简单：通过随机遮蔽输入图像的图像块训练模型来重构遮蔽的像素，MAEs 可以为 CIL 提供一种新的自监督表征学习范式，从而使模型学习到更多通用的表征，这对于 CIL 而言至关重要。此外，利用带有分类标签的监督目标还能提高无监督 MAE 的训练效率和模型稳健性 [18]。同时，遮蔽后的输入通过提供数据的一

*通讯作者 (xialei@nankai.edu.cn)

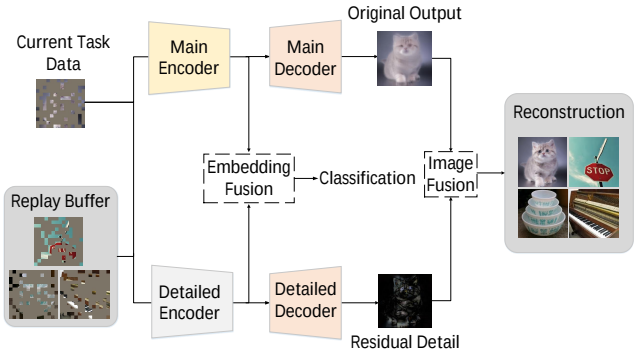


图 1. 本文提出的用于高效 CIL 的双边 MAE。重放缓冲区包含从过去任务图像中选取的随机图像块，这比存储整幅图像更有效。将这些数据与当前任务中遮蔽的输入数据相结合，MAE 能够从遮蔽的输入中同时学习图像分类和重建。为了进一步改善重建的图像的质量并学习表征，本文方法使用了嵌入级和图像级的融合来学习更加稳定的表征以及包含更多细节的重建图像，从而用于 CIL。

个随机子集可以在分类中起到很强的正则化作用。

在学习新任务时，MAE 能够从范例的稀疏采样块中粗略地重构图像。这个过程能够使框架生成重建后的重放图像，但仍存在两个问题：(i) 生成的图像纹理往往不够精细和逼真，这减少了重放数据的多样性；(ii) 在嵌入层面上，线性分类器缺乏来自低层次特征的信息。因此，我们为 CIL 引入了图像级和嵌入级融合的双边 MAE 框架（图1为方法概览）。将互补的详细图像和重建图像融合在一起，可以用详细、高质量的数据分布来丰富不充分的重放数据，从而减轻灾难性遗忘。两个分支的嵌入层面融合还能保持嵌入的稳定性和多样性，因此我们的框架能够在可塑性和稳定性之间取得更好的平衡。

本文的双边 MAE 框架的主要贡献分为三个方面：

- 本文提出了一种用于高效增量学习的 MAE 框架，该框架结合了自监督重建和重放数据生成的优势。
- 为了进一步提高重建图像的质量和学习效率，我们设计了一种新颖的包含两个互补分支的双边 MAE，以获得更好的重建图像和正则化表征。
- 在 CIFAR-100、ImageNet-Subset 和 ImageNet-Full 的不同 CIL 设置下，本文的方法取得了最先进的性能。

2. 相关工作

增量学习 在过去几年中，人们提出了各种增量学习方法 [2,5]。最近的研究大致可分为三类：基于重放的方法、基于正则化的方法和基于参数隔离的方法。基于重放的方法通过重放先前任务的训练样本缓解任务重现偏差。除了重放样本，BiC [36]、PODNet [8] 和 iCaRL [29] 采用蒸馏损失来防止遗忘并提高模型的稳定性。GEM [21]、AGEM [3] 和 MER [30] 通过修改当前训练样本的梯度来匹配旧样本，从而利用过去任务的范例。基于排练的方法可能会导致模型与存储样本过度拟合。

伪重放法通过重建旧数据进行重放。MeRGANs [35] 使用条件 GANs 来平衡旧样本和当前样本的生成。此外，dreaming 相关的方法如 DeepInversionVersion [41] 以及 AlwaysBeDreaming [33] 利用反向传播的信号生成与原始数据集相似的图像。

基于正则化的方法，如 LwF [17]、EWC [15] 和 DMC [44] 提供了学习更优表征的方式，同时为适应新任务保留足够的可塑性。基于参数隔离的方法 [24, 39] 针对每项任务使用具有不同计算图的模型。在不断扩张的模型的帮助下，新的模型分支能够有效减轻灾难性遗忘，其代价是带来了更多的参数和更大的计算开销。

自监督学习。 自监督学习 [7,27,37] 已被证明能帮助模型学习可泛化的特征，因此一个很自然的想法是将其应用于增量学习。早期的研究使用了一些前置任务，如块排列 [27] 或旋转预测 [10]。对比学习方法是对不同样本之间成对的相似性和不相似性进行建模 [4]。与此相比，MAEs [11] 通过从遮蔽后的输入中重建图像来学习特征表征。

、有一些工作提出了用于类增量学习的自监督范式。PASS [46] 结合旋转预测 [10] 来学习可跨任务迁移的表征。DualNet [28] 使用 BarlowTwins [42] 引入了一个” 慢” 任务，以规范” 快” 增量学习。在本文中，我们探索了用于生成重放数据的框架，框架同时使用了语义和细节级别的自监督，通过更丰富的重放数据和更通用的特征来减轻遗忘。

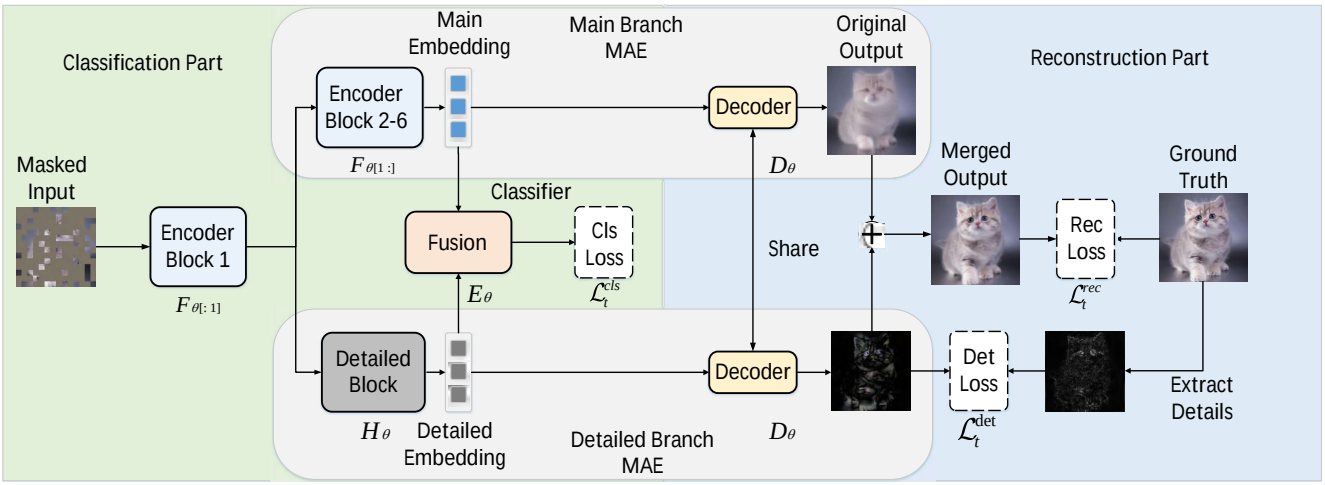


图 2. 本文提出的用于 CIL 的双边 MAE 总体框架。遮蔽后的输入经过两个分支，其中嵌入级的融合用于分类，图像级的融合用于重建。整张图像可以通过一小部分输入的图像块生成，同时重建的图像可以用于重放。

3. 用于 CIL 的连续 MAE

在本节中，我们首先定义了类增量学习问题和基本的 MAE 模型。然后，我们介绍了我们的增量学习框架以及框架所基于的双边 MAE 架构。

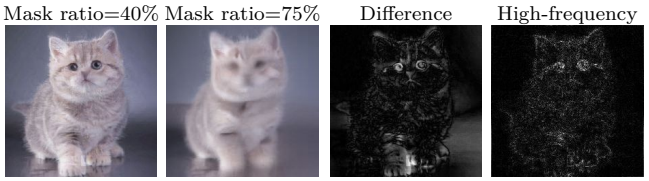
3.1. 方法序言

类增量学习问题。 CIL 旨在不断学习包含新类别的任务，同时避免或减轻对于旧任务的遗忘。在学习任务 $t \in \{1, 2, ..., T\}$ 的特定阶段，模型的训练仅能利用来自当前任务的数据 $\{(x_i^t, y_i^t)\}$，其中 x_i^t 表示任务 t 中的图像 i，y_i^t 表示对应的类别标签。一个 CIL 模型通常由一个特征提取器 F_θ 和一个常规的分类器 G_ϕ 构成，该分类器在遇到新任务时将进行扩张。在学习任务 $t+1$ 时，C_{t+1} 个新类将被添加到 G_ϕ。特征提取器 F_θ 首先将输入 x 映射到深度特征向量 $z = F_\theta(x) \in \mathbb{R}^d$，$d$ 为输出的特征表征的维度，之后统一的分类器 $G_\phi(z) \in \mathbb{R}^{C_{1:t}}$ 生成一个关于类别 $C_{1:t}$ 的概率分布，该分布将被用于预测输入图像 x。

在训练任务 t 时，模型的目标是尽量减少当前任务的损失，同时不降低之前任务的性能。减少遗忘的常用技术是保留一小部分先前任务的训练样本。令 ϵ 为先前任务样本的缓冲区。CIL 任务中一个关键的问题是重放数据的数量有限制。相比于当前任务 t 的全部数据而言，只有少量的旧任务类别样本是可用的（常用的设置是每一个类存储 20 个样本），这会带来新旧任务之间训练不平衡的问题。

一种用于分类的 MAE 框架。 MAE 首先将输入图像 x 裁剪为不重叠的图像块，我们将一张完整图片 x 的图像块数目定义为 N_f。在完成分块后，MAE 随机将 N_f 个图像块中的一部分进行遮蔽，遮蔽的比例为 $r \in [0, 1]$，剩余 $N = \lfloor N_f \times (1-r) \rfloor$ 个图像块。随后，这些采样后大小为 $K \times K$ 的像素块通过一个 MLP 被映射为 D 维的视觉嵌入。其与一个类别标记拼接后，得到大小为 $\mathbb{R}^{(N+1)\times D}$ 的张量。在对原始的块进行位置编码后，该输入将被送入 MAE transformer 编码器。这项操作保持了嵌入的形状不变。输出的类别标记嵌入能够通过交叉熵损失 $\mathcal{L}_t^{\text{cls}}$ 用于分类，如图2所示。

对于 MAE 解码器，可学习的掩码标记被插入到嵌入中以代替遮蔽的图像块，同时 MAE 编码器的输出形状从 $\mathbb{R}^{(N+1)\times D}$ 变为 $\mathbb{R}^{(N_f+1)\times D}$。虽然解码器不会被用于分类，但其有助于网络将图像级别的重建监督反向传播到嵌入级别。这可以稳定图像嵌入并有利于优化整个过程。此外，解码后重建的图像可提供更丰富、更高质量的重放数据。为了限制计算量，我们使用单层 Transformer 块进行解码。通过编码器得到的额外分类损失可以加快收敛速度，提高训练过程中的重构效率。输入图像 x 与重建图像



$\hat{x}_1$ (Eq. 9) $\hat{x}_2$ (Eq. 10) $\hat{x}_1 - \hat{x}_2$ (Eq. 11) x'' (Eq. 6)

图 3. 一个关于从不同遮蔽比例 r_1 和 r_2 的重建结果中提取细节图像的示例。第三幅图像来自前两幅图像的差值，最后一幅图像是从第三幅图像中提取的高频分量。

$\hat{x}$ 的均方误差被用作重建损失函数 $\mathcal{L}_t^{\text{rec}}(x, \hat{x})$。

3.2. 利用 MAE 高效存储范例

每个任务训练完成后，我们保存一小部分样本图像并对其进行随机遮蔽。通过保持相同的存储容量，我们就能为每个类保存更多的重放数据，因为每个样本占用的空间更少。例如，相比于传统的基于重放的方法而言，以 0.75 作为遮蔽比例能够使我们保存四倍数量的（可重建的）样本。

令 S 和 P 表示图像和块的尺寸。我们的编码器将输入图像切分为 $\frac{S}{P} \times \frac{S}{P}$ 个图像块。对于每个没有被遮蔽的图像块，我们保存其 2D 的索引 (i, j)。我们仅需一个字节来存储索引，因为索引的范围小于 255。两个用于存储 2D 索引的额外字节与存储的图像块相比可以忽略不计。遮蔽比例为 0.75 的大小为 224× 224 的图像仅仅占据 36.75KB 的存储空间。对于 $P = 16$ 的情况，保存的图像块的数量为 $(1 - 0.75) \times (\frac{224}{16})^2 = 49$，索引占据的存储空间仅为 98B。

3.3. 双边 MAE 融合

为进一步提高重建质量和嵌入的多样性，我们提出了一种双分支 MAE，以学习全局和细节的图像分类以及重建的知识。我们在图2中阐述了整体框架。嵌入层面的双边融合旨在提高表征的多样性。图像级的重构学习可为 CIL 提供高质量的重放数据和稳定的自监督。

嵌入融合。 在以下内容中，我们使用 $F_{\theta[:1]}$ 与 $F_{\theta[1:]}$ 表示 Transformer 编码器中第一个以及之后的块。令 H_θ 和 E_{theta} 表示图2中的细节块和嵌入融合模块，这些结构为标准的 MLP 层以及注意力块。分类损失的计算公式为：

$$f = F_{\theta[:1]}(\text{mask}(x, r)) \quad (1)$$

$$z = E_\theta(F_{\theta[1:]}(f), H_\theta(f)) \quad (2)$$

$$\mathcal{L}_t^{\text{cls}}(x, y) = \mathcal{L}_{\text{ce}}(G_\phi(z), y), \quad (3)$$

其中 $\text{mask}(x, r)$ 表示将比例为 r 的随机遮蔽操作施加到图像 x，f 为第一个编码器块提取到的嵌入，这也是双边 MAE 两个分支的输入，$G_\phi(z)$ 为用于交叉熵损失的预估类别分布。

基于细节损失的图像融合。 对于细节头和相应的损失，我们发现在频域中工作更容易使网络关注高频细节，而这正是细节分支应该重建的。我们定义了一个频率掩蔽函数 $M(\cdot)$，它能将参数（一个图像块）转换到频域，然后使用一个围绕原点的圆形掩蔽器将低频分量掩蔽。如图2所示，MAE 解码器由模型的两个分支共享，因为它们具有相似的重建任务以及相同的输入和输出形状。令 D_θ 表示共享的解码器，之后两个分支图像级别的输出以及重建损

Algorithm 1 我们提出的双边 MAE 的伪代码

Input: The number of task T, training samples $D_t = \{(x_i, y_i)\}_t$ of task t, model Θ^0, replay buffer ϵ, and masking ratios r, r_1, r_2.

Output: model Θ^T

```
for t ∈ {1, 2, ..., T} do
    Θ^t ← Θ^{t−1}
    R_t ← ReconstructOldSamples (ε_t, r)
    while not converged do
        (x, y) ← Sample (R_t, D_t)
        (L_t^cls, L_t^rec) ← BilateralMAE (x, y)
        (x̂_1, x̂_2) ← MaskAndReconstruct (x, r_1, r_2)
        L_t^det ← ComputeDetailLoss (x̂_1, x̂_2)
        train Θ^t by minimizing L_t from Eq. 12
    end while
end for
```

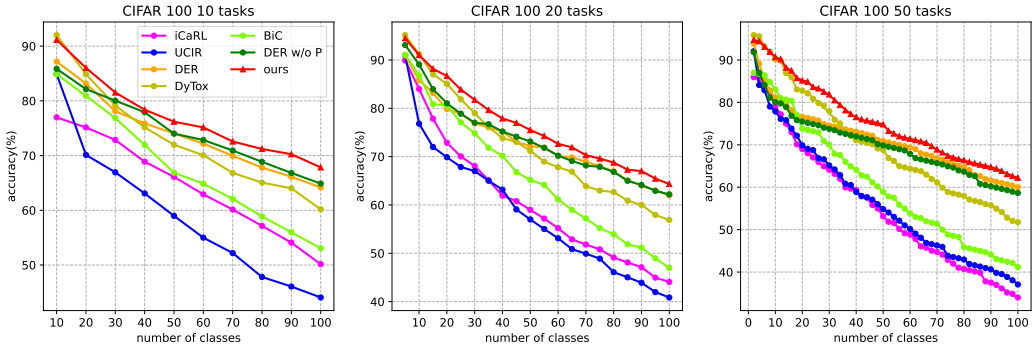


图 4. 在 CIFAR-100 数据集 10 个、20 个和 50 个任务场景下增量任务性能的变化。

Method	N=10			N=20			N=50		
	Avg ↑	Last ↑	F ↓	Avg ↑	Last ↑	F ↓	Avg ↑	Last ↑	F ↓
iCaRL [29]	65.27	50.74	31.23	61.20	43.75	32.40	56.08	36.62	36.59
UCIR [12]	58.66	43.39	35.67	58.17	40.63	37.75	56.86	37.09	38.13
BiC [36]	68.80	53.54	28.44	66.48	47.02	29.30	62.09	41.04	34.27
PODNet [8]	58.03	41.05	41.47	53.97	35.02	36.70	51.19	32.99	40.42
DER w/o P [40]	75.36	65.22	15.02	74.09	62.48	23.55	72.41	59.08	26.73
DER [40]	74.64	64.35	15.78	73.98	62.55	23.47	72.05	59.76	26.59
DyTox [9]	75.47	62.10	15.43	75.10	59.41	21.60	73.89	57.21	24.22
Ours	79.12	68.40	12.17	78.76	65.22	14.39	76.95	63.12	18.34

表 1. CIFAR-100 数据集 10 个、20 个和 50 个任务场景下的平均准确率（%）、最后阶段准确率（%）以及遗忘程度 F（%）。

失 $\mathcal{L}_{\square}^{\text{rec}}$ 可以表示为：

$$f = F_{\theta[:1]}(\text{mask}(x, r)) \quad (4)$$
$$x' = D_\theta(F_{\theta[1:]}(f)) \quad (5)$$
$$x'' = \text{ifft2}(M(D_\theta(H_\theta(f)))) \quad (6)$$
$$\hat{x} = x' + x'' \quad (7)$$
$$\mathcal{L}_t^{\text{rec}} = \mathcal{L}_{\text{mse}}(x, \hat{x}), \quad (8)$$

其中 x' 和 x'' 分别为主要的和残差的细节输出，ifft2 为逆快速傅立叶变换。

细节损失 $\mathcal{L}_t^{\text{det}}$ 还利用了频率掩蔽函数 M 来比较细节分支的输出与输入图像的两个 MAE 重建结果之间的差值：

$$\hat{x}_1 = D_\theta(F_\theta(\text{mask}(x, r_1))) \quad (9)$$
$$\hat{x}_2 = D_\theta(F_\theta(\text{mask}(x, r_2))) \quad (10)$$
$$\mathcal{L}_t^{\text{det}} = ||M(D_\theta(H_\theta(f))) - M(\hat{x}_2 - \hat{x}_1)||_1, \quad (11)$$

其中 $\hat{x}_1$ 和 $\hat{x}_2$ 是两张使用了不同遮蔽比例 r_1 和 r_2 的重建图像（从计算图剥离）。残差 $\hat{x}_2 - \hat{x}_1$ 被用作在频域中损失 $\mathcal{L}_t^{\text{det}}$ 内的细节分枝。图3举例说明了这一点。

$\mathcal{L}_t^{\text{cls}}$、重建损失 $\mathcal{L}_t^{\text{rec}}$ 以及细节损失 loss $\mathcal{L}_t^{\text{det}}$ 的加权和构成了训练的总损失：

$$\mathcal{L}_t = \lambda_{\text{cls}}\mathcal{L}_t^{\text{cls}} + \lambda_{\text{rec}}\mathcal{L}_t^{\text{rec}} + \lambda_{\text{det}}\mathcal{L}_t^{\text{det}}. \quad (12)$$

我们方法的伪代码在算法1中给出。

4. 实验结果

4.1. 性能指标与实现

数据集以及设置。 我们在三个数据集上进行了实验：CIFAR-100 [16]、ImageNet-Subset 和 [6]，以评估我们方法的性能。对于 CIFAR-100 和 ImageNet-Subset，我们分别在包含 10 个、20 个和 50 个任务的场景中进行了测试，每个任务的类别数相同。我们评估了 ImageNet-Full 的 10 任务设置，其中每个任务都包含 100 个新类别。为了衡量训练期间完成

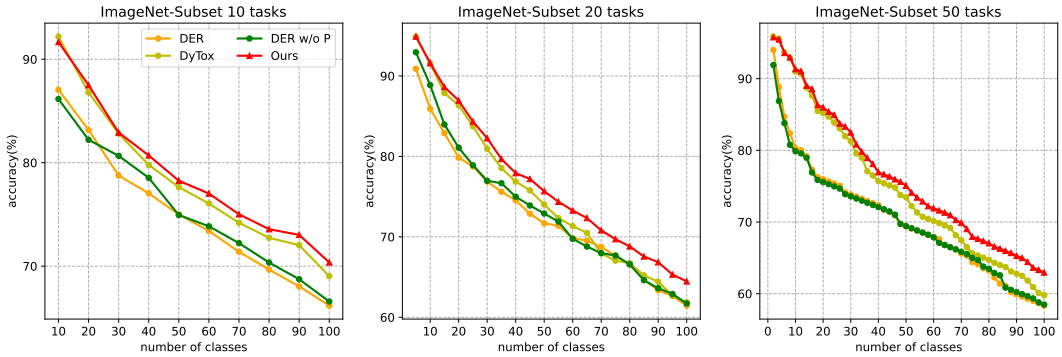


图 5. 在 ImageNet-Subset 数据集增量任务性能的变化。

Method	N=10			N=20			N=50		
	Avg ↑	Last ↑	F ↓	Avg ↑	Last ↑	F ↓	Avg ↑	Last ↑	F ↓
BiC [36]	64.96	55.07	31.32	59.40	49.35	34.70	53.75	44.56	40.23
PODNet [8]	63.44	51.75	35.63	55.11	45.37	41.70	51.72	42.94	44.65
DER w/o P [40]	77.18	66.70	14.86	72.70	61.74	20.76	70.44	58.87	24.20
DER [40]	76.12	66.06	15.09	72.56	61.51	20.46	69.77	58.19	25.35
DyTox [9]	77.15	69.10	14.66	73.13	61.87	17.32	71.51	60.02	20.54
Ours	79.54	70.29	12.04	75.20	64.40	14.89	74.42	62.87	17.22

表 2. 在 ImageNet-Subset 数据集 10 个、20 个和 50 个任务场景下的平均准确率（%）、最后阶段准确率（%）以及遗忘程度 F（%）。

Method	top-1		top-5	
	Avg ↑	Last ↑	Avg ↑	Last ↑
iCaRL [29]	38.40	22.70	63.70	44.00
Simple-DER	66.63	59.24	85.62	80.76
DER w/o P [40]	68.84	60.16	88.17	82.86
DER [40]	66.73	58.62	87.08	81.89
DyTox [9]	71.29	63.34	88.59	84.49
Ours	74.76	66.15	91.43	87.13

表 3. ImageNet-Full 数据集 10 个增量任务场景下的结果。

所有任务后的总体准确率，我们报告了每个任务后所学任务的平均准确率，以及在增量学习结束时所有任务的准确率。

实现细节。　对于所有的数据集我们使用了相同的网络。模型从零开始训练以防止数据泄露，批量大小为 1024，使用初始学习率为 1e-4 和带有余弦衰减的 Adam [14]。式12的损失权重设置为 $\lambda_{cls} = 0.01$，$\lambda rec = 1.0$，$\lambda det = 1.0$。遮蔽比例设置为 $r = 0.75$，$r_1 = 0.75$ 以及 $r_2 = 0.4$。每一个任务训练 400 轮。对于文献中基于范例的方法，我们为每个类别存储 20 个样本（这是常见的做法）。

对于编码器我们使用了 5 个 Transformer 块，解码器使用了 1 个 Transformer 块。所有的 Transformer 块拥有相同的编码维度 384 以及 12 个自注意力头。这种设计不同于原始的 MAE，因为其更加轻量。我们保存图像块所占用的内存量与其它每类存储 20 幅完整图像的方法相同。例如，我们选择 80 幅图像，使用 0.75 的遮蔽比例随机保存每幅图像中 25% 的图像块（因此只占用与 20 幅完整图像相同的空间）。细节区块使用 3 层 MLP 实现，维度为 384。有关网络结构的更多细节在补充材料中给出。

4.2. 与最先进方法的对比

在本节中我们将目前最先进的方法与我们的方法进行了对比，包括 DER [40] 和 DyTox [9]。在所有的图和表中，“DER w/o P” 表示不含剪枝的 DER [40]，因此其在不同任务中能够拥有更多参数。

Method	Replay	Reconstruction	Bilateral	Avg	Last
Baseline				73.40	62.31
Variants	✓			75.88	64.35
	✓	✓		77.48	66.54
	✓	✓	✓	79.12	68.40

表 4. 在包含 10 项任务的 CIFAR-100 设置中，对我们提出的方法的每个组成部分进行的消融实验。Replay 表示使用 MAE 生成的数据进行重放，Reconstruction 表示应用自监督重建损失，Bilateral 表示引入 MAE 的细节分支。

DyTox [9] 同样使用了 Transformer 结构，我们使用其官方代码库以复现结果。

CIFAR-100。 我们在表1中给出了平均准确率（Avg）、最后一个任务后的准确率（Last）以及平均遗忘程度（F）。显然，在每种设置下，我们的方法都远优于其他方法。对于较长的任务序列，我们的双边 MAE 可从自监督重建和更丰富的重放数据中获益，与其他方法相比，它的遗忘率要低得多。总体准确率曲线见图4。在使用相同大小的重放存储空间时，我们的方法在所有三种场景下的最后一项任务后的准确率均比 DyTox 高出约 6%。

ImageNet-Subset 与 ImageNet-Full. 我们在表2和表3中分别报告了我们的方法 ImageNet-Subset 和 ImageNet-Full 的性能。在包含 10 个、20 个和 50 个任务的设置中，我们的方法在最后一个任务后的准确率绝对增益分别比 DyTox [9] 高出 1.19%、2.53% 和 2.85%。每个阶段的平均准确率较高，遗忘率较低，这也证明了我们的方法在减轻遗忘方面的有效性。我们同样在图5中说明了 ImageNet-Subset 的性能变化。在第一个任务中，我们的方法与 DyTox 的准确率相似，但在后面的任务中，我们的方法超过了所有其他方法，尤其是在长任务序列中。在更大规模的 ImageNet-Full 中，我们的双边 MAE 在所有指标上都明显超过其他方法约 3%。

4.3. 消融实验

不同组成部分的消融实验。 我们的双边 MAE 包括自监督重建任务、重放数据生成以及用于图像级和嵌入级融合的双边 MAE 分支。我们在表4中对这三个因素进行了分析。我们方法中的这三个主要部分具有不同的功能，它们相互配合，使性能比基线提高了约 6%。我们观察到：(a) 更高质量的重放数据对性能有直接贡献，采用 $r = 0.75$ 的掩码比率能以与基线相同的存储成本获得 4 倍的重放数据。(b) 重建损失是一种有效的自我监督，能使平均准确率提高约 2%。(c) 双边架构通过提高重放数据生成质量以及引入图像和嵌入级监督取得良好的效果。

r	Data Source	Avg	Last
0.60	Generated	77.50	67.37
0.75	Generated	79.12	68.40
0.90	Generated	77.12	67.02
N/A	Real	79.57	68.87

表 5. 对遮蔽率和生成数据质量的消融实验。实验在包含 10 项任务的 CIFAR-100 设置上进行，我们以百分比的形式报告了 top-1 准确率。Data Source 表示重放是生成的还是真实的。在最后一行中，我们重放了部分真实图像，其存储容量与使用比例为 $r_1 = 0.75$ 时相当。

Domain	Avg	Last
Spatial	77.45	65.93
Frequency	79.12	68.40

表 6. 对于细节头的消融实验。实验在包含 10 项任务的 CIFAR-100 设置上进行，我们以百分比的形式报告了 top-1 准确率。Domain 表示我们的损失应用于哪一个域。

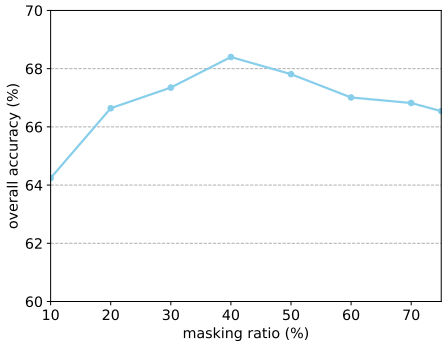


图 6. 关于 $\mathcal{L}_t^{\text{det}}$ 中遮蔽比例 r_2 的消融实验。另一个遮蔽比例 r_1 设置为 75% 以作为参考。

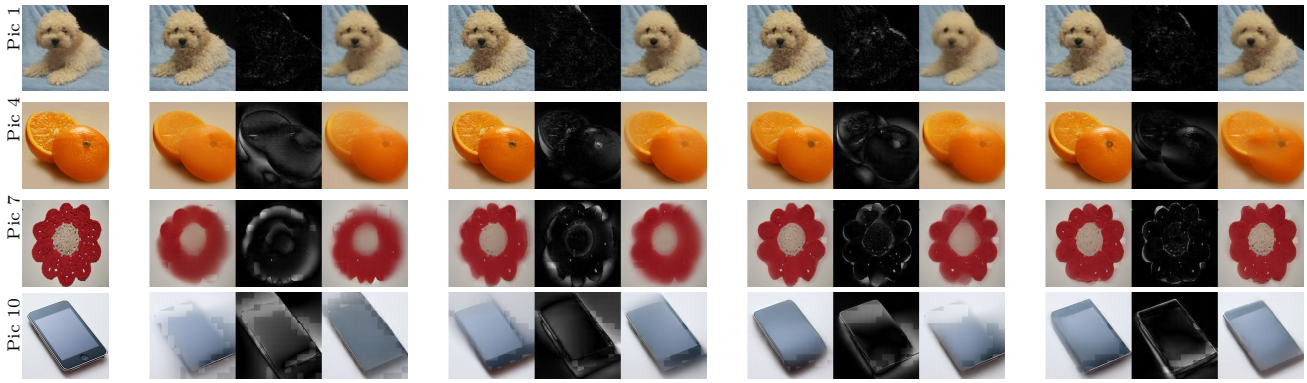


图 7. 在包含 10 个任务的设置中对 ImageNet 子集中的图像进行重建得到的结果。从任务 1、4、6 和 10 中选取的四幅原始图像如左图所示。其余各栏显示的是使用我们的双边 MAE 组合（左）、仅使用我们 MAE 的细节分支（中）和仅使用我们 MAE 的主分支（右）重建的图像。

遮蔽比例。 MAE [11] 的一个关键参数是遮蔽比例。在 r 的选择上需要权衡：过大的 r（如 0.95）会导致重建效果不佳，从而影响重放数据的质量，造成更严重的遗忘。然而，过小的 r 产生的额外重放数据量有限（例如，当 r 为 0.10 时，我们只能承受约 11% 的额外重放数据）。The results in Table 5 show that $r = 0.75$ is a good trade-off for our bilateral MAE. 表5中的结果表明，对于我们的双边 MAE 而言，$r = 0.75$ 是一个很好的折衷。为了验证生成的重放数据的质量，结果中还包含使用原始图像代替生成图像进行重放的准确率。表5第 2 行和第 4 行的结果显示，我们的方法获得了高质量的图像，与重放真实图像相比，准确率相差不到 0.5%。

频域中的细节损失。 我们通过将嵌入从空域转换到频域来实现细节损失。这样做的目的是为了集中处理高频信息，这与 MAE 细节分支的学习目标相匹配。如表6所示，进行这样的转换是有益的，因为它在最后一项任务中带来了超过 2% 的增益。

细节损失中关于 r_1 和 r_2 的消融实验。 在所有实验中，我们将 r_1 设为 0.75 作为参考，同时改变用于计算细节损失真值的 r_2。对 r_2 的权衡是，如果 r_2 较大，则按照遮蔽比例 r_1 和 r_2 重建的结果差异较小，因此，监督信号中关于细节损失的信息很少，细节分支的影响也会减小。另一方面，过小的 r_2（如 0.10）会保留主分支重建图像的大部分残留部分，这可能会导致对主分支的监督较弱，减慢其训练速度。我们在图6中展示了一系列 r_2 值的结果。这些结果表明，约 0.40 的 r_2 值可以很好地为我们的 MAE 细节分支提供监督。

Method	Parameters (M)	Avg ↑	Last ↑	F ↓
DER w/o P	112.27	75.36	65.22	15.02
DyTox	10.73	75.47	62.10	15.43
Ours (MLP size = 1536)	12.89	79.12	68.40	12.17
Ours (MLP size = 768)	9.35	78.36	67.52	12.90

表 7. 模型大小的比较。我们比较了两个版本的双边 MAE 模型和对比模型。实验在包含 10 个任务的 CIFAR-100 上进行。

模型与范例的大小。 为了比较不同方法的有效性，通常使用参数数量相同或相似的模型，并使用相同数量的范例。在我们的方法中，我们对原始 MAE 进行了调整，使其更加轻量，参数数量与 DyTox 相当甚至更少，如表7（最后一行）所示。我们将屏蔽率默认设置为 75%，每类保存 80 个范例，因此模型和范例的存储大小相似，因为我们存储的图像块所需的空间与仅使用 20 个范例的基线完全相同。

关于有效缓冲区大小的消融实验 在表8中，我们使用相同的缓冲区大小，通过遮蔽 DyTox 中输入图像的图像块将我们的方法与 DyTox 进行了比较。所有三行都使用相同大小的内存来存储范例。在使用 80 个遮蔽率为 75% 的范例时，DyTox（最后一行）的

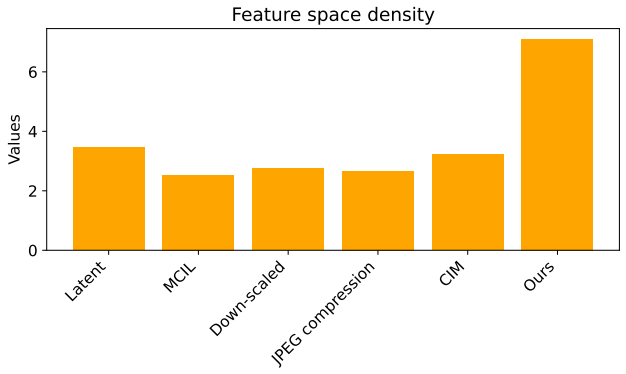


图 8. 对于特征空间密度 π 的比较。

性能比使用 20 个完整图像范例时更好。它的性能仍然劣于我们的方法，这表明我们的性能提升并不仅仅来自于额外的范例，还来自于将 MAE 与细节分支整合到我们的双边架构中。

重建分析。 我们在图7中展示了在 ImageNet-Subset 的包含 10 个任务的设置中进行图像重建的结果。左栏显示的是从任务 1、4、7 和 10 中随机选取的图像。我们的双边 MAE 以与任务无关的方式学习重建图像，这有助于在学习之前就为未来的任务生成合理的结果。我们 MAE 的细节分支学习重建高频细节，以补充主分支。主分支的结果有时缺乏特定样本的特征，但在我们提出的细节分支的帮助下，重建结果更加准确，并能提供更好的重放数据。

Method	Buffer	Memory size	25% Patches	Images	Acc(%)
Ours	80	1x	✓		68.40
DyTox	20	1x		✓	62.10
DyTox†	80	1x	✓		65.46

表 8. 在内存使用量相同的情况下对于有效缓冲区大小的消融实验。Dytox† 表示直接对存储的图像范例应用遮蔽比例 $r = 75\%$，以便使 DyTox 的范例数量和存储大小与我们的设置相同。

更通用的表征有助于 CIL。 我们按照 PASS [46] 的方式计算了不同方法的特征空间密度指标 [32]：$\pi = \pi_{intra}/\pi_{inter}$，其中 π_{intra} 表示同一类别中的平均余弦相似度，π_{inter} 表示不同类别中的平均余弦相似度。特征空间密度的增加与数据偏移情况下更强的泛化能力相关 [46]。随后我们比较了所有任务训练后的特征空间密度，如上图8所示。很明显，我们的方法得出的密度显著高于其他方法。

Metric	Memory	Avg↑	Last↑
Latent replay (CVPRW'20) [19]	-	62.44	51.30
MCIL (CVPR'20) [20]	60	63.25	53.12
Down-scaled (TNNLS'21) [45]	60	67.04	55.40
JPEG compression (ICLR'22) [34]	60	72.34	61.32
CIM (CVPR'23) [22]	60	75.30	63.05
Ours	60	79.12	68.40

表 9. 在 CIFAR-100 包含 10 个任务的设置中，我们的双边 MAE 框架与其它节省内存的方法进行了比较。Memory 表示每个类所需的存储空间（以 KB 表示）。

与其它高效重放方法的消融实验。 我们比较了我们的框架中 MAE 生成的重放样本与各种节省内存的方法生成的样本，这些方法分别基于隐重放（latent replay）、合成范例（synthesized exemplars）、缩放（down-scaling）、JPEG 图像压缩（JPEG image compression）和 CIM（前景提取和背景压缩）。所有这些方法都使用了相同的存储量（除了 Latent Replay 使用了一个带有 450 万参数的 GAN），而我们的方法则始终获得更高的性能。

5. 总结

在这项工作中，我们证明了掩码自动编码器是一种高效的增量学习器。我们的方法将随机图像块存储为范例，它可以仅利用部分信息重建高质量图像，以供重放。此外，我们还提出了一种新颖的双边 MAE 架构，可进一步提高嵌入多样性和重建质量。在样本存储成本相同的情况下，我们的双边 MAE 方法明显优于之前的最先进方法。

鸣谢 该项目得到国家自然科学基金（NO. 62225604，62206135）以及中央高校基本科研业务费专项资金资助（南开大学，070-63233085）。算力由南开大学超算中心提供。

参考文献

[1] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a net-

work of experts. In IEEE Conf. Comput. Vis. Pattern Recog., 2017. 409

[2] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. Neural Networks, 135:38–54, 2021. 409, 410

[3] Arslan Chaudhry, Marc' Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In Int. Conf. Learn. Represent., 2019. 410

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020. 410

[5] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. TPAMI, 2021. 409, 410

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE Conf. Comput. Vis. Pattern Recog., 2009. 413

[7] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In Int. Conf. Comput. Vis., 2015. 410

[8] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In Eur. Conf. Comput. Vis., 2020. 410, 413, 414

[9] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In IEEE Conf. Comput. Vis. Pattern Recog., 2022. 413, 414, 415

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In Int. Conf. Learn. Represent., 2018. 410

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In IEEE Conf. Comput. Vis. Pattern Recog., 2022. 409, 410, 416

[12] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In IEEE Conf. Comput. Vis. Pattern Recog., 2019. 413

[13] David Isele and Akansel Cosgun. Selective experience replay for lifelong learning. In AAAI, 2018. 409

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Int. Conf. Learn. Represent., 2015. 414

[15] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 2017. 409, 410

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. 413

[17] Zhizhong Li and Derek Hoiem. Learning without forgetting. In Eur. Conf. Comput. Vis., 2016. 409, 410

[18] Feng Liang, Yangguang Li, and Diana Marculescu. Supmae: Supervised masked autoencoders are efficient vision learners. arXiv preprint arXiv:2205.14540, 2022. 409

[19] Xialei Liu, Chenshen Wu, Mikel Menta, Luis Herranz, Bogdan Raducanu, Andrew D Bagdanov, Shangling Jui, and Joost van de Weijer. Generative feature replay for class-incremental learning. In IEEE Conf. Comput. Vis. Pattern Recog. Worksh., 2020. 417

[20] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In IEEE Conf. Comput. Vis. Pattern Recog., pages 12245–12254, 2020. 417

[21] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. Adv. Neural Inform. Process. Syst., 2017. 410

[22] Zilin Luo, Yaoyao Liu, Bernt Schiele, and Qianru Sun. Class-incremental exemplar compression for class-incremental learning. In IEEE Conf. Comput. Vis. Pattern Recog., pages 11371–11380, 2023. 417

[23] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In Eur. Conf. Comput. Vis., 2018. 409

[24] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In IEEE Conf. Comput. Vis. Pattern Recog., 2018. 409, 410

[25] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. IEEE Trans. Pattern Anal. Mach. Intell., 2022. 409

[26] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In Psychology of learning and motivation. Elsevier, 1989. 409

[27] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Eur. Conf. Comput. Vis., 2016. 410

[28] Quang Pham, Chenghao Liu, and Steven Hoi. Dualnet: Continual learning, fast and slow. Adv. Neural Inform. Process. Syst., 2021. 410

[29] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In IEEE Conf. Comput. Vis. Pattern Recog., 2017. 409, 410, 413, 414

[30] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In Int. Conf. Learn. Represent., 2019. 410

[31] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. Adv. Neural Inform. Process. Syst., 2019. 409

[32] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In Int. Mach. Learn., pages 8242–8252. PMLR, 2020. 417

[33] James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. Always be dreaming: A new approach for data-free class-incremental learning. In Int. Conf. Comput. Vis., 2021. 410

[34] Liyuan Wang, Xingxing Zhang, Kuo Yang, Longhui Yu, Chongxuan Li, Lanqing Hong, Shifeng Zhang, Zhenguo Li, Yi Zhong, and Jun Zhu. Memory replay with data compression for continual learning. In Int. Conf. Learn. Represent., 2022. 417

[35] Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. Adv. Neural Inform. Process. Syst., 2018. 409, 410

[36] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In IEEE Conf. Comput. Vis. Pattern Recog., 2019. 410, 413, 414

[37] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In IEEE Conf. Comput. Vis. Pattern Recog., 2018. 410

[38] Ye Xiang, Ying Fu, Pan Ji, and Hua Huang. Incremental learning using conditional adversarial networks. In Int. Conf. Comput. Vis., 2019. 409

[39] Ju Xu and Zhanxing Zhu. Reinforced continual learning. Adv. Neural Inform. Process. Syst., 2018. 410

[40] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In IEEE Conf. Comput. Vis. Pattern Recog., 2021. 413, 414

[41] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In IEEE Conf. Comput. Vis. Pattern Recog., 2020. 410

[42] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning, pages 12310–12320. PMLR, 2021. 410

[43] Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong gan: Continual learning for conditional image generation. In Int. Conf. Comput. Vis., 2019. 409

[44] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In WACV, 2020. 410

[45] Hanbin Zhao, Hui Wang, Yongjian Fu, Fei Wu, and Xi Li. Memory-efficient class-incremental learning for

image classification. TNNLS, 33(10):5966–5977, 2021. 417

[46] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In IEEE Conf. Comput. Vis. Pattern Recog., 2021. 410, 417