用于类增量语义分割的端点权重融合*

Jia-Wen Xiao [1*] Chang-Bin Zhang [1†] Jiekang Feng [2] Xialei Liu [1‡] Joost van de Weijer [3]
Ming-Ming Cheng [1]
[1] TMCC, CS, Nankai University [2] Tianjin University [3] Universitat Autonoma de Barcelona

摘要

类增量语义分割（CISS）专注于减轻灾难性遗忘以提高辨别力。以前的工作主要利用正则化（例如知识蒸馏）来维持当前模型中的先前知识。然而，单独的蒸馏通常会给模型带来有限的收益，因为只有新旧模型的表示被限制为一致。在本文中，我们提出了一种简单而有效的方法来获得对旧知识有很强记忆力的模型，称为端点权重融合（EWF）。在我们的方法中，包含旧知识的模型与保留新知识的模型以动态融合的方式融合，加强了对不断变化的分布中旧类的记忆。此外，我们分析了我们的融合策略与流行的移动平均技术 EMA 之间的关系，这揭示了为什么我们的方法更适合类增量学习。为了促进参数空间中距离较近的参数融合，我们使用蒸馏来增强优化过程。此外，我们在两个广泛使用的数据集上进行了实验，实现了最先进的性能。

1. 引言

作为一项基本任务，语义分割在视觉应用中发挥着关键作用 [10, 25]。以前的完全监督工作旨在分割训练集中定义的固定类别。然而，经过训练的分割模型有望在实际应用中识别更多的类。一种简单的解决方案是通过混合新旧数据在整个数据集上重新训练模型。尽管如此，这一策略将带来巨大的标签和培训成本。从迁移学习的角度来看 [22, 30]，另一个简单的解决方案是根据新添加的数据调整先前学习的模型。但该模型很快就会过度适应新类别，同时忘记以前的旧类别。这种现象也被称为灾难性遗忘 [35]。

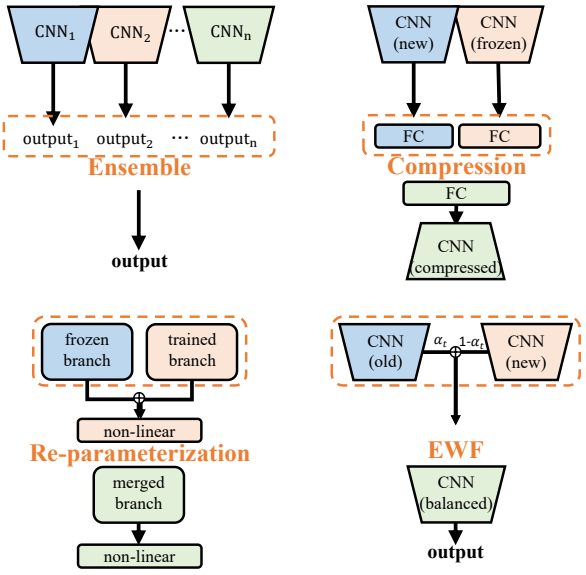


图 1: 增量学习的不同融合策略说明。集成方法利用多个模型来积累更多知识。压缩方法可以减小模型大小并将知识提炼到小型网络中。而重新参数化方法使用等效操作进行模型融合。我们的端点权重融合 (EWF) 提出了使用动态因子 α_t 进行模型添加而无需进一步训练。

为了在没有额外标签或训练成本的情况下缓解灾难性遗忘问题，类增量语义分割（CISS）[4, 16, 53]旨在优化保持对旧类的歧视和学习新类知识之间

*本文为 CVPR'22 论文 [52]的中文翻译版。
†前两位作者具有同等贡献。
‡通讯作者：刘夏雷 (xialei@nankai.edu.cn)。

的权衡。大多数作品 [4, 16, 17, 38]设计了正则化方法来保持记忆旧知识和学习新知识之间的平衡。我们观察到现有的作品仍然可能遭受灾难性遗忘，导致旧类的性能显着下降。在 CISS 场景中，不仅之前的数据由于隐私问题或数据存储限制而无法访问，而且新添加的数据集中旧类的区域被标记为背景，这进一步加剧了模型的过拟合。

此外，从旧模型训练新模型并将它们融合以获得最终模型是持续学习中的常见策略。如图1所示，我们将它们大致分为四类，模型扩展和融合两个阶段。一些方法 [27, 33, 43, 48]建议以增量步骤扩展模型并集成新旧输出，这会产生大量内存和推理成本。而一些作品应用压缩 [47, 48]将新旧模型压缩为参数较少的统一模型。然而，这些需要仅对新数据进行进一步训练，这可能会导致对新数据的偏见。随后，一些工作 [53, 56]探索知识解耦并通过重新参数化进行线性参数融合。然而，这是一种模块内融合策略，仅限于某些操作。作为最后一类，我们提出了端点权重融合（EWF），其形式是在新旧模型之间添加动态因子的参数，不需要进一步的训练和重新参数化，并且随着更多任务的进行而保持恒定的模型大小。

在这项工作中，我们将权重融合应用于 CISS 并提出了 EWF 策略，其目的是利用权重融合在新旧知识之间找到新的平衡。在增量训练过程中，我们选择当前任务训练轨迹的起点和终点模型。起点代表旧知识，终点代表新知识。在学习当前任务后，提出动态权重融合以实现有效的知识整合。我们通过两个模型相应参数的加权平均值来聚合它们。然而，对模型没有限制的训练过程会增加起点和终点之间的参数距离，限制了 EWF 策略带来的性能提升。为了克服这个缺点，我们通过知识蒸馏方案进一步增强了 EWF 策略 [16, 17, 53]，这可以大大增加两点模型的相似度并提高 EWF 的效率。总而言之，本文的主要贡献是：

- 我们提出了一种端点权重融合策略，该策略不需要进一步训练的成本，并且保持模型大小相同。它可以有效地在新旧类别之间找到新的平衡，缓解灾难性遗忘
- 我们的方法可以轻松地与多种最先进的方法集成。在多个长序列的 CISS 场景中，可以将基线性能提升 20% 以上。
- 我们对各种 CISS 场景进行了实验，证明我们的方法在 PASCAL VOC 和 ADE20K 上均达到了最先进的性能。

2. 相关工作

类别增量学习。 类别增量学习主要侧重于减轻灾难性遗忘，同时学习新类别所需的判别性信息。它在图像分类中最常被分析，这些技术可以大致分为三类 [11]。许多作品 [44, 45, 46]关注模型的结构属性（即基于结构的方法）。这个想法是冻结旧模型并扩展架构空间以学习新知识，这通常会导致模型的容量和内存大小不断增长。另一种方法是在增量学习过程中对模型进行正则化（即基于正则化的方法）[7, 8, 12, 17]，通过约束（例如知识蒸馏 [41, 42]或梯度惩罚 [26, 29]）加强记忆。这些方法给学习过程带来的成本可以忽略不计，但它们允许参数更新的自由度较小。其他一些方法 [1, 2, 28]提出通过排练来复习知识（即基于排练的方法）。他们存储旧数据并将其与新数据混合以重新训练模型 [5, 21, 55]。

类增量语义分割。 语义分割 [19]旨在为每个单个像素分配不同的类别，最近在类增量学习场景中引起了人们的关注 [4, 16]。然而，与分类问题相比，语义分割的数据需要更多的存储空间 [24, 49]。因此，最近的工作主要集中在利用蒸馏将旧知识转移到新模型，而不保存旧任务中的样本。MiB [4]提出对潜在类别进行建模以解决语义漂移问题。PLOP [16]应用特征蒸馏来限制表示能力。SDR [38]使用原型匹配来加强潜在空间的一致性。RC-IL [53]分析了条带池的缺点，并提出了基于平均池的蒸馏来支持训练。相反，SSUL [6]没有应用蒸馏，并提出修复特征提取器而不是更新其参数。此外，他们还引入了数千个额外数据来帮助生成伪标签。但是，简单地修复模型肯定会破坏可塑性和稳定性之间的平衡，并且在面对大量新数据时是不可持续的。另一方面，仅应用蒸馏会限制性能，因为它只能限制新数据的表示相同。总结上述思想并受到 RC-IL [53]的启发，我们设计了一种与上述不同的策略，配合蒸馏，通过模

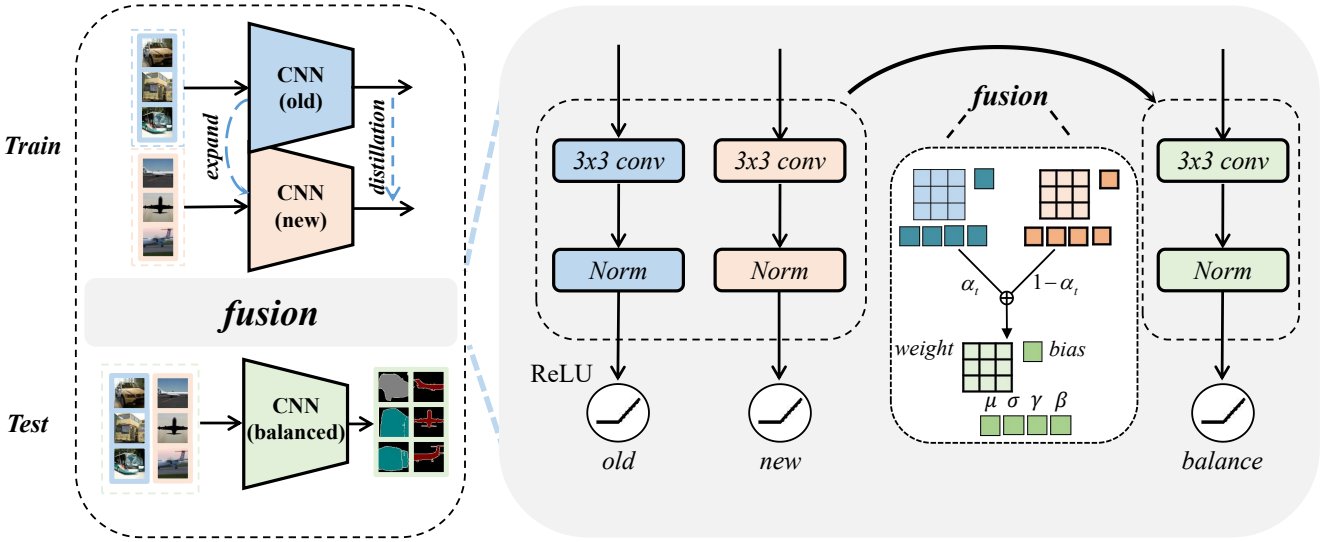


图 2: 我们的端点权重融合 (EWF) 框架的说明。在左侧，我们说明了训练和测试过程。通过知识提炼增强训练，并在训练后应用融合以积累所有可见的知识。在右侧，给定一个 3×3 卷积层和一个规范化层，使用动态权衡参数 α_t 将它们融合。请注意，我们的融合过程没有计算成本。

型权重融合维持新旧知识之间的平衡。

权重融合。 权重融合广泛应用于神经网络训练中以提高性能。它可以应用于线性和非线性运算。在线性模式下，RCM [27]探索了卷积的可加性并将其应用于多任务学习。ACNet [14]和 RepVGG [15]首先提出结构重新参数化，用于将多分支卷积-批量归一化串行序列合并到普通卷积层。RC-IL 巧妙地利用这一操作在持续学习中建立了表征补偿机制。在非线性模式中，通常使用权重平均来拉近不同模型之间的联系。BYOL [20]等著名方法使用 EMA [40]来提高知识转移效果或模型鲁棒性。从这一点出发，我们设计了一个达到新平衡的策略，该策略对当前的训练过程没有任何影响，充分释放模型的学习能力。

3. 方法

3.1. 前置知识

我们考虑一个多步骤训练过程，其中 T 个任务在语义分割的完全监督场景中由模型 f_θ 顺序学习，其中 f 由参数 θ 参数化。这里 f_{θ_t} 表示任务 t 的模型。每个任务包含一个数据集 $D_t = \{x_i, y_i\}$，其中 x_i 表示数据，y_i 表示相应的标签。任务 T 的训练标签空间表示为 $C_t \cup c_b$，其中 C_t 包括该任务中出现的所有类，c_b 代表背景。由于 $C_i \cap C_j = \emptyset$，不同任务的目标并不相同，容易导致灾难性遗忘。为了节省标注成本，仅对现阶段需要学习的类别进行标注。因此，c_b 不仅包含真实的背景，还包含过去和未来任务中出现的类。这使得模型 f_θ 的训练变得复杂，因为背景标签 c_b 可以引用不同任务中的不同类。这加剧了遗忘的严重性。

3.2. 端点权重融合 (EWF)

如图1所示，现有的模型扩展融合方法在增强模型对旧知识的记忆方面有其自身的缺点。因此，为了更好地保留旧知识，同时提高模型的学习能力，我们引入了端点权重融合策略。考虑到步骤 t 的训练过程，我们选择起点模型和终点模型。上一步的最终模型 θ_{old} 被认为是旧知识的最佳容器。而当前任务训练后的模型 θ_{new} 包含新类别的判别信息。此外，我们引入了另一个参数 用于端点权重融合，旨在以一定的比例融合两组参数。这个比率可以看作是一个平衡因素。因此，端点权重融合的操作可以写为

$$\theta_{balanced} = \alpha_t \theta_{new} + (1 - \alpha_t)\theta_{old} \quad (1)$$

其中 $\theta_{balanced}$ 表示该任务的最终模型，$\theta_{old}, \theta_{new}$ 的定义如上。此外，增量类别（记为 N_{new}）和原始类别（记为 N_{old}）的数量在一定程度上代表了对 θ_{new}

和 θ_{old} 的重视程度。对于 α_t 的选择，我们需要在当前的学习步骤中考虑 N_{new} 和 N_{old}。我们决定用与 N_{new} 和 N_{old} 相关的方程代替常数比率。详细来说，这个公式可以表示为：

$$\alpha_t = \sqrt{\frac{N_{new}}{N_{new} + N_{old}}} \tag{2}$$

它可以用于不同的任务和场景。它能够以动态因子适应 CISS 中的每项任务。我们在图2中说明了我们方法的主要思想。

增强 EWF 的知识蒸馏。在实践中，没有任何约束的训练会显着增加不同模型之间的距离并打破模型表示的相似性。这意味着它会恶化模型的遗忘，从而进一步削弱模型区分新类别的能力。此外，保证两个遥远模型之间的低误差线性路径是一个过于强烈的假设。因此，选择一个没有约束的训练模型作为终点可能是有害的，因此，我们引入知识蒸馏来增强端点权重融合方法模型的兼容性。知识蒸馏是防止模型遗忘的常用技术。如上所述，我们利用蒸馏来支持我们的策略，限制两个端点之间的距离并迫使它们相似。一般来说，持续学习中使用的蒸馏主要分为两类，即基于特征的蒸馏和基于逻辑的蒸馏。它们可以表示为：

$$\begin{aligned} L_{FD} &= \frac{1}{|D|} \sum_{(x_i,y_i)\sim D} ||\Psi_{old}(x_i) - \Psi_{new}(x_i)||^2 \\ L_{LD} &= \frac{1}{|D|} \sum_{(x_i,y_i)\sim D} KL(\Phi_{old}(\Psi_{old}(x_i)), \Phi_{new}(\Psi_{new}(x_i))) \end{aligned} \tag{3}$$

Ψ_{old}/Φ_{old} 和 Ψ_{new}/Φ_{new} 分别表示旧的和新的特征提取器/分类器，D 是增量学习步骤的相应数据集。在 CISS 中，[4, 16]分别提出了两种流行的蒸馏损失（即 UNKD、POD）。前者是基于逻辑的蒸馏，后者是基于特征的蒸馏。它们可以很容易地集成到我们的方法中。

关于 EMA 与 EMA 的讨论。与我们的方法类似的更新策略是使用 EMA [40]来更新模型。这里我们将讨论我们的方法 EWF 和 EMA 之间的区别，以及 EWF 的优点。EMA 策略在训练过程中维护一个移动平均模型，并使用该模型替换最终模型进行推理。移动平均模型可以表示为：

$$v^i = \beta v^{i-1} + (1-\beta)\theta^i \tag{4}$$

其中 v_i 表示前 i 个模型的移动平均值，θ_i 表示第 i 次迭代后的模型。β 为移动平均参数，通常设置为 0.9/0.99。不难看出，EMA 是一种迭代操作，由于模型之间的距离较小，更容易实现。由于我们使用 SGD 作为优化器，因此如下式：如式5所示，EMA 和 EWF 的结果可以表示为式6中的 v_n 和 $\theta_{balanced}$. 注意，学习率不影响分析结果，因此我们将学习率设置为 1。更详细的推导在附录中。

$$\begin{aligned} \theta^n &= \theta^{n-1} - \nabla_L(\theta^{n-1}) \\ &= \theta^1 - \sum_{k=1}^{n} \nabla_L(\theta^k) \end{aligned} \tag{5}$$

$$\begin{aligned} v^n &= \theta^1 - \sum_{k=1}^{n-1} (1-\beta^{n-k}) \nabla_L(\theta^k) \\ \theta_{balanced} &= \theta^1 - \alpha \sum_{k=1}^{n-1} \nabla_L(\theta^k) \end{aligned} \tag{6}$$

由于 β 通常设置为 0.9/0.99，由于 α 按式1设置，α 总是小于 β。根据式6, EMA 对早期的梯度给予更多的权重，然后在后续迭代中逐渐减弱梯度的影响。相反，EWF 赋予不同部分的梯度相对均匀的权重。此外，为了更好地观察增量学习步骤的稳定性，我们计算了 θ_1 和 θ_i 之间的表示相似度。根据我们在增量步骤中的观察，如图3所示，θ_1 和 θ_i 之间的表示相似度先减小然后增大，并在后续阶段逐渐稳定。这表明，为了学习新知识，表示会在蒸馏和交叉熵损失的作用下首先崩溃，然后恢复。那么 EMA 将更多的梯度集中在早期崩溃的过程上，而我们的方法更关注恢复后有用的梯度信息。从这个角度来看，我们的方法理论上比 CISS 中的 EMA 更好。

整体框架。如上所述，为了记住旧任务中的知识，我们借用知识蒸馏来加强模型的记忆。为了学习新类别的区分，我们使用交叉熵损失来优化模型。一般来说，目标由下式给出：

$$\min_{\theta_t} \mathcal{L}_{CE}(\theta_t) + \mathcal{L}_{KD}(\theta_t; \theta_{t-1}) \tag{7}$$

EWF 的整体算法如 Alg 1 所示。

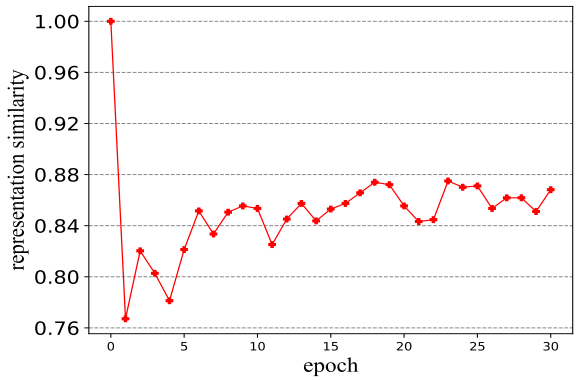


图 3: 训练前的旧模型 θ^1 与正在训练的新模型 θ^i 之间的表征相似度。我们使用的相似度度量是余弦相似度。我们提取新旧网络的特征提取器的中间结果，并计算它们之间余弦相似度的平均值。

Algorithm 1 Pseudo Code for EWF

Require: f, θ_0, T, D_T and learning rate γ

$t \leftarrow 1$

while $t \leq T$ do

Initialize N_{new}, N_{old}

$\theta^1 \leftarrow \theta_{t-1}$

$\alpha_t \leftarrow \sqrt{\frac{N_{new}}{N_{new}+N_{old}}}$

$i \leftarrow 1$

while not converged do

Sample mini-batch $\{x_i, y_i\} \sim D$

$\theta^{i+1} \leftarrow \theta^i - \gamma \nabla_{L_{CE}+L_{KD}}(f_{\theta^i})$

$i \leftarrow i + 1$

end while

$\theta_{old} \leftarrow \theta^1, \theta_{new} \leftarrow \theta^i$

$\theta_{balanced} \leftarrow \alpha_t \theta_{new} + (1 - \alpha_t)\theta_{old}$

$\theta_t \leftarrow \theta_{balanced}$

$t \leftarrow t + 1$

end while

4. 实验

我们演示实验协议、场景和训练细节。此外，我们通过定量和定性实验评估我们的算法。

4.1. 实验设置

4.1.1 协议

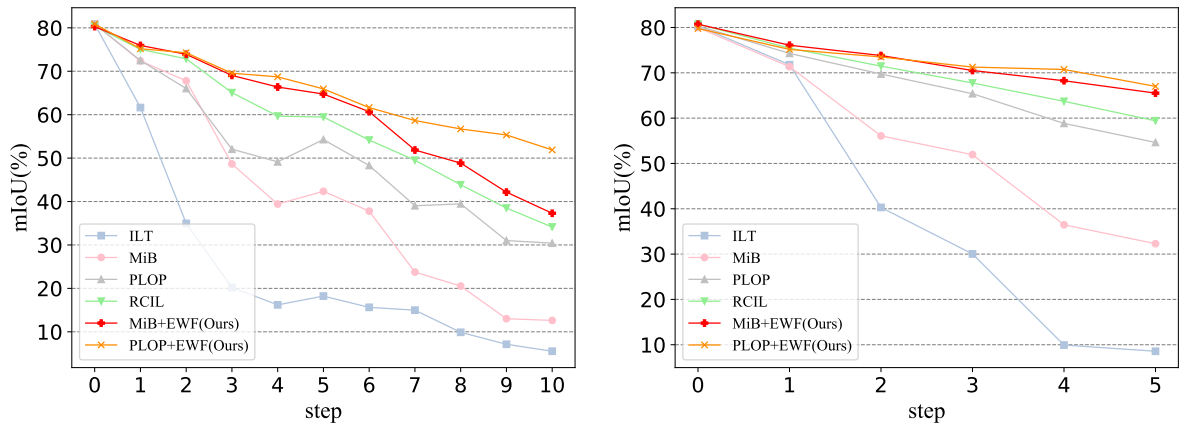
一般来说，CISS 的训练分为 T 个步骤，每个步骤代表一个任务，每个步骤中标记的类是不相交的。我们采用与其他工作一样的重叠设置，其中当前的训练数据可能包含在先前步骤中标记为背景的潜在类别。重叠设置更现实，因此我们仅像以前的方法一样评估此设置 [6, 16]。继现有工作 [4, 16, 53]之后，我们在两个广泛使用的分割数据集 PASCAL VOC 2012 [18]和 ADE20K [54]上进行了实验。PASCAL VOC 2012 数据集 [18]包含 10,582 个训练图像和 1449 个验证图像，具有 20 个对象类和背景类。ADE20K 数据集 [54]由 150 个类组成，包含 20、210 个训练图像和 2000 个验证图像。继之前的工作 [4, 16, 53]之后，X −Y 表示 CISS 的不同设置。在 X−Y 设置中，模型可以在初始步骤中识别 X 个类别，然后应该在后续的每个步骤中学习 Y 个新添加的类别。在每个步骤中，只有当前任务数据可用于训练。我们在 PASCAL VOC 2012 [18]上进行实验，有四种设置：15-1、10-1、5-3 和 19-1。在 ADE20K [54]上，我们验证了我们的方法在三种设置（100-5、100-10 和 100-50）上的有效性。

4.1.2 实施细节

继现有工作 [4, 16, 53]之后，我们应用 Deeplab-v3 [9]作为我们的分割模型，并以 ResNet-101 [23]作为主干。我们还在主干中使用就地激活的批量归一化 [3]。在我们的实验中，我们使用了一些数据增强，包括水平翻转和随机裁剪。EWF 的比率 t 定义为方程 2。使用 SGD 优化器，我们在每个步骤中训练模型 30 (PASCAL VOC 2012) 和 60 (ADE20K) epoch，批量大小为 24。我们将第一个训练步骤的初始学习率设置为 0.01，将下一个连续训练步骤的初始学习率设置为 0.001 学习步骤。所有实验均在 4 个 RTX 2080Ti GPU 上进行。学习率随着多计划而衰减。在训练过程中，我们使用 20% 的训练集作为验证，并报告原始验证集上的平均交并集 (mIoU)。

Method	15-1 (6 steps)			10-1 (11 steps)			5-3 (6 steps)			19-1 (2 steps)		
	0-15	16-20	all	0-10	11-20	all	0-5	6-20	all	0-19	20	all
LwF [31] (TPAMI2017)	6.0	3.9	5.5	8.0	2.0	4.8	20.9	36.7	24.7	53.0	8.5	50.9
ILT [36] (ICCVW2019)	9.6	7.8	9.2	7.2	3.7	5.5	22.5	31.7	29.0	68.2	12.3	65.5
SDR [39] (CVPR2021)	47.3	14.7	39.5	32.4	17.1	25.1	-	-	-	69.1	32.6	67.4
RCIL [?] (CVPR2022)	70.6	23.7	59.4	55.4	15.1	34.3	63.1	34.6	42.8	77.0	31.5	74.7
MiB [4] (CVPR2020)	38.0	13.5	32.2	12.2	13.1	12.6	57.1	42.5	46.7	71.2	22.1	68.9
MiB+EWF (Ours)	78.0	25.5	65.5	56.0	16.7	37.3	69.0	45.0	51.8	77.8	12.2	74.7
PLOP [16] (CVPR2021)	65.1	21.1	54.6	44.0	15.5	30.5	25.7	30.0	28.7	75.4	37.3	73.5
PLOP+EWF (Ours)	77.7	32.7	67.0	71.5	30.3	51.9	61.7	42.2	47.7	77.9	6.7	74.5
Joint	79.8	72.6	78.2	79.8	72.6	78.2	78.2	78.0	78.2	76.9	77.6	77.4

表 1: 针对不同类增量分割场景，在 Pascal VOC 2012 数据集上最后一步的 mIoU(%)。



(a) PASCAL VOC 上的 10-1 场景

(b) PASCAL VOC 上的 15-1 场景

图 4: 10-1 (a) 和 15-1 (b) 设置下每个步骤都 mIoU (%)。

4.2. 与竞争方法的比较

在本节中，我们将我们的方法应用于 MiB [4] 和 PLOP [16]。此外，我们还将我们的方法与 LwF [31]、ILT [37]、SDR [38]和 RCIL [53] 进行比较。

PASCAL VOC 2012。 在此数据集中，我们使用与 [4, 16, 53]相同的实验设置，我们使用类增量学习设置 15-1,10-1,5-3,19-1 进行实验。如表1所示，我们报告了最终任务的实验结果。从结果中，我们可以观察到，在更具挑战性的设置（例如，15-1、10-1、5-3）上，我们的方法大幅提高了 MiB 和 PLOP 的结果。例如，在 15-1 上在 1 设置下，我们的算法对于 PLOP 和 MIB 分别获得了 12.4% 和 33.3% mIoU 的性能增益。此外，在最长的学习序列，10-1 设置上，我们的方法一致获得了较大的增益，提高了 PLOP 和 MIB 的性能。MIB 分别提高了 21.4% 和 24.7%。在表1中，我们还报告了不同设置下新旧类的性能。我们的方法对旧类实现了显着的性能提升。这表明我们的 EWF 策略可以显着增强模型的内存旧知识的记忆（通过成功地对抗遗忘）。在更具挑战性的设置中，例如 15-1、5-3、10-1，我们的方法还提高了新类的性能。这表明 EWF 可以在新类上实现高可塑性步骤，并且所提出的网络动态加权组合允许在可塑性和稳定性之间实现良好的权衡。我们还在图4中展示了持续学习过程中的动态性能变化。很明显，随着学习步骤的增加，我们的方法与最佳基线之间的差距越来越大，并且我们的方法在不同设置下的曲线（15- 1 和 5-3）在大部分学习轨迹中都位于顶部。

		15-1 (PASCAL VOC 2012)			10-1 (PASCAL VOC 2012)		
Method	Auxiliary Data	0-15	16-20	all	0-10	11-20	all
RECALL [34] (ICCV2021)	GAN / Web-Data	65.7	47.8	62.7	59.5	46.7	54.8
ST-CIL [50] (TNNLS2022)	Unlabeled Data	71.4	40.0	63.6	-	-	-
SSUL [6] (NeurIPS2021)	Saliency Map	77.3	36.6	67.6	71.3	45.9	59.2
SSUL + EWF (Ours)	Saliency Map	77.9	38.9	68.6	72.4	47.4	60.5

表 2: Pascal VOC 2012 15-1 和 10-1 相交设置下最后一步的 mIoU(%)。

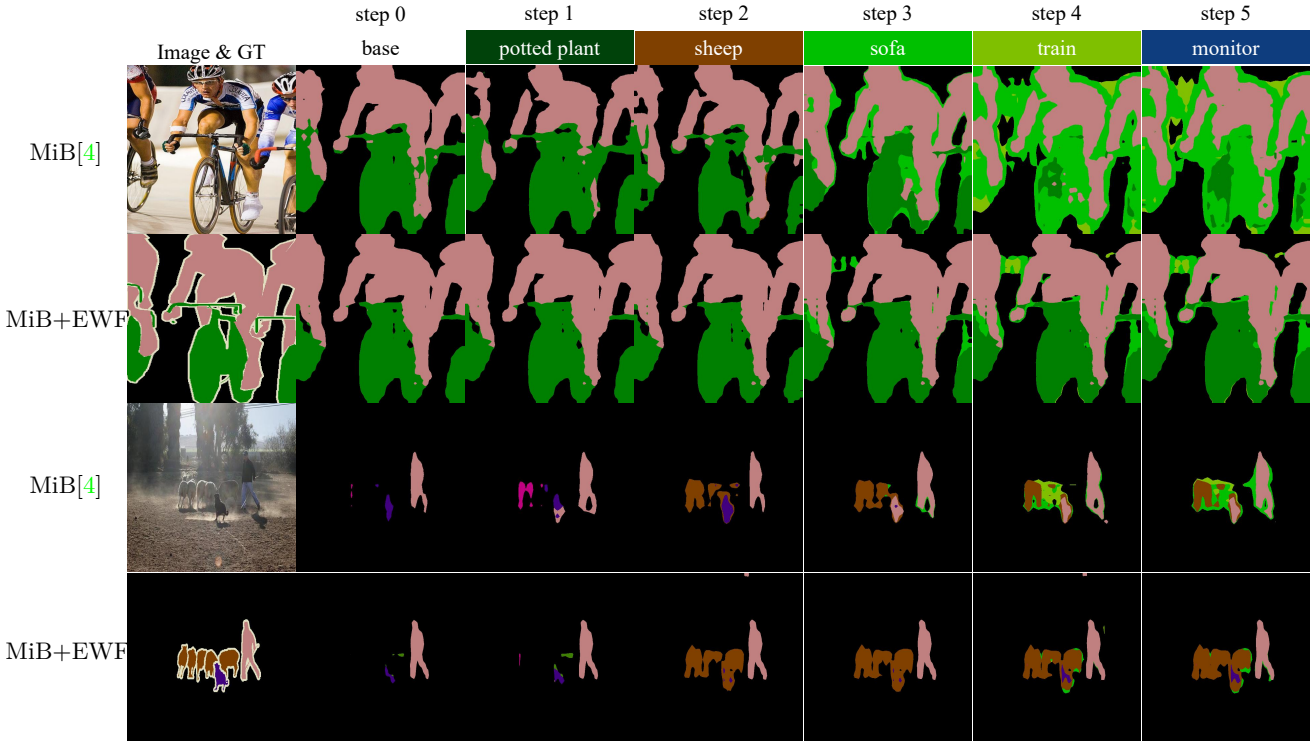


图 5: 不同方法之间的定性比较。所有预测结果均来自 15-1 重叠设置的最后一步。在初始步骤中，学习了 15 个类别，并逐步学习了 5 个任务，类别包括盆栽、绵羊、沙发、火车和监视器。

与引入辅助数据的方法的比较。 值得注意的是，有几种方法 [6, 34, 51]引入了不同形式的辅助数据来辅助持续的语义分割，帮助模型构建更好的伪标签或增强对旧知识的记忆。RECALL [34]学习生成模型或从网络爬行数据中检索以进行重放，而 SSUL [6]利用显着对象检测器（在具有 5000 个图像的 MSRA-B 数据集 [32]上进行训练）来生成显着图作为辅助数据。ST-CIL [51]利用带有伪标签的未标记数据集作为基本事实。即使我们的算法旨在处理无额外数据的场景，为了进一步证明我们方法的有效性，我们将我们的方法集成到 SSUL 中，并将其与表2中的上述方法进行比较。请注意，SSUL 完全冻结了主干网，并且很难直接应用我们的方法。因此，当我们将我们的方法应用于 SSUL 时，我们学习了部分骨干网络参数。具体来说，我们将 SSUL 中主干的第二阶段设置为模型融合的可学习参数。SSUL 在使用辅助数据的这些方法中获得了最好的性能，并且我们应用于 SSUL 的方法可以进一步将所有类别的性能提高约 1%。

可视化。 如图5所示，我们比较了基于 MiB 的方法，并显示了来自基本任务（步骤 0）的样本和来自步骤 2 的样本。从顶部两行开始，人员和自行车类别很大程度上保留了我们的方法，但它逐渐忘记了

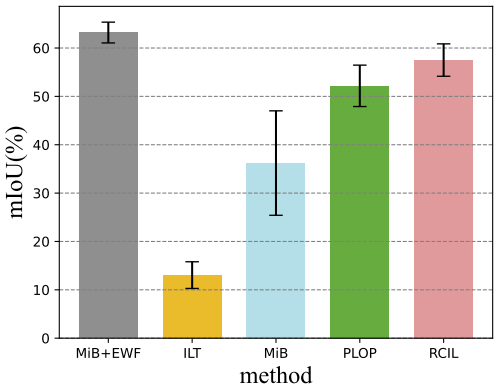


图 6: 关于不同顺序的稳健性的说明。

基线 MiB。从底部两行开始，它显示了一个包含旧类和新羊类的示例。我们的方法可以在融入新知识的同时更好地保留旧知识。

ADE20K。 为了进一步评估我们方法的有效性，我们在 ADE20K 数据集上进行了实验。在表3中，我们显示了设置 100-50、100-10 和 100-5 的实验结果。如表3所示，我们的方法在此数据集上取得了优异的性能。特别是在最具挑战性的设置 100-5 和 100-10 上，我们的方法分别比 MiB [3] 提高了 6.2% 和 3.0%。它还在 100-5 设置上大幅超越了最先进的方法 RC-IL。这表明我们的 EWF 对大规模数据集也有效。

4.3. 消融实验

在这一部分中，我们论证并分析了权重融合的有效性及其动态因子选择。我们使用 MiB [4] 进行消融实验。

融合策略。 在表4中，我们展示了不同融合策略的性能。这些实验是在 PASCAL VOC 2012 上使用设置 15-1 进行的。大多数这些方法通过融合上一步模型中存在的信息来执行持续学习。其中，EMA [40]在训练期间用移动平均值更新一组存储的参数，并将其用于推理。模型集成在推理过程中以平均方式融合先前模型的预测和新的预测。None 表示仅使用蒸馏和交叉熵训练的模型（即 MiB [4]）。与无融合方法相比，EMA 和模型集成分别具有 5.1% 和 5.0% 的性能提升。而我们的 EWF 在简单融合策略的基础上有了 33.3% 的提升，这再次验证了 Sec.3.2 中的讨论。

融合因子选择。 在这一部分中，我们对融合因子的选择进行了实验。为了证明我们的参数选择策略的优势，我们在融合过程中使用一些固定值作为平衡参数，并比较我们的方法和他们的方法之间的差异来评估其优势。具体来说，我们选择三个场景（即 15-1、10-1、5-3）进行实验，并计算三个场景下不同策略的平均 mIoU 来衡量最终性能。由于平衡因子 $[0, 1]$，我们取 0.2、0.4、0.6 和 0.8 作为固定值来与我们的参数选择策略进行比较。如表5所示，对于 15-1 设置，与其他固定参数相比，我们的方法达到了最高性能。此外，虽然我们的方法略低于其他设置上固定参数的最高性能，但不同设置上最高性能的参数差异很大。这意味着为所有场景选择固定参数是不现实的，并且对算法有害。重要的是，我们的方法在所有其他固定值中达到了最高的平均性能，这表明我们的策略可以轻松应用于不同的设置，而无需手动调整超参数。

类顺序的稳健性。 在类增量语义分割场景中，模型遇到的类的顺序对于衡量算法的有效性具有重要意义。因此，为了验证我们的算法的有效性以及对不同类别顺序的鲁棒性，我们对五个不同类别顺序进行实验以计算平均性能及其标准差。如图6所示，我们的方法显着提高了性能，同时也显着增强了对不同类别顺序的鲁棒性。

5. 结论以及局限

在这项工作中，我们通过一种简单而有效的端点权重融合方法解决了类增量语义分割（CISS）问题。它通过现有的基于蒸馏的方法得到增强，并且可以轻松地与它们集成。动态参数融合策略被证明对于不同的设置是灵活的，并且避免了超参数的进一步调整。有趣的是，我们讨论了我们的方法和流行的权重融合方法 EMA 之间的关系，这揭示了为什么我们的方法在增量学习中更有效。实验结果表明，与基线相比，我们的方法可以获得显着的增益并实现卓越的性能。在未来的工作中，我们将进一步研究

	100-50 (2 steps)			100-10 (6 steps)							100-5 (3 steps)		
Method	1-100	101-150	all	1-100	101-110	111-120	121-130	131-140	141-150	all	1-100	101-150	all
ILT [36] (ICCVW2019)	18.3	14.8	17.0	0.1	0.0	0.1	0.9	4.1	9.3	1.1	0.1	1.3	0.5
PLOP [16] (CVPR2021)	41.9	14.9	32.9	40.6	15.2	16.9	18.7	11.9	7.9	31.6	39.1	7.8	28.7
RC-IL [?] (CVPR2022)	42.3	18.8	34.5	39.3	14.6	26.3	23.2	12.1	11.8	32.1	38.5	11.5	29.6
MiB [4] (CVPR2020)	40.7	17.7	32.8	38.3	12.6	10.6	8.7	9.5	15.1	29.2	36.0	5.6	25.9
MiB+EWF(Ours)	41.2	21.3	34.6	41.5	12.8	22.5	23.2	14.4	8.8	33.2	41.4	13.4	32.1
Joint	44.3	28.2	38.9	44.3	26.1	42.8	26.7	28.1	17.3	38.9	44.3	28.2	38.9

表 3: ADE20K 数据集上不同重叠持续学习场景的最后一步的 mIoU(%)。

Parameter Selection	Ours	0.2	0.4	0.6	0.8
15-1	65.6	65.6	63.7	60.1	53.3
10-1	37.3	39.5	31.8	22.7	14.3
5-3	51.8	38.0	51.2	56.1	52.9
Average	51.6	47.7	48.9	46.3	40.2

表 4: 动态参数平衡策略与固定平衡因子的比较。

Fusion strategy	$step_1$	$step_2$	$step_3$	$step_4$	$step_5$
None [4]	71.4	56.1	51.9	36.5	32.2
EMA [40]	74.0	59.6	59.8	40.9	37.3
model ensemble [13]	74.1	60.1	60.3	41.5	37.2
EWF (Ours)	76.1	73.8	70.5	68.3	65.6

表 5: 融合策略的消融实验。所有性能测试均在 PASCAL VOC 2012 15-1 设置下进行。

我们简单的 EWF 策略在 CISS 中如此有效的根本原因。此外，我们计划评估其他应用领域的增量学习策略。**致谢**。这项工作得到了国家自然科学基金委员会（No. 62225604 和 62206135）、中央高校基本科研业务费专项资金（南开大学，No. 63223050）和字节跳动有限公司的资助。我们感谢西班牙政府对项目 PID2019-104174GB-I00、TED2021-132513B-I00 的资助。

参考文献

[1] J. Bang, H. Kim, Y. Yoo, J.-W. Ha, and J. Choi. Rainbow memory: Continual learning with a memory of diverse samples. In IEEE Conf. Comput. Vis. Pattern Recog., 2021. 410

[2] E. Belouadah and A. Popescu. Il2m: Class incremental learning with dual memory. In Int. Conf. Comput. Vis., pages 583–592, 2019. 410

[3] S. R. Bulo, L. Porzi, and P. Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5639–5647, 2018. 413

[4] F. Cermelli, M. Mancini, S. R. Bulo, E. Ricci, and B. Caputo. Modeling the background for incremental learning in semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 9233–9242, 2020. 409, 410, 412, 413, 414, 415, 416, 417

[5] H. Cha, J. Lee, and J. Shin. Co2l: Contrastive continual learning. In Int. Conf. Comput. Vis., pages 9516–9525, 2021. 410

[6] S. Cha, Y. Yoo, T. Moon, et al. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. Advances in neural information processing systems, 34:10919–10930, 2021. 410, 413, 415

[7] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In Eur. Conf. Comput. Vis., 2018. 410

[8] A. Chaudhry, A. Gordo, P. Dokania, P. Torr, and D. Lopez-Paz. Using hindsight to anchor past knowledge in continual learning. In Proceedings of the AAAI conference on artificial intelligence, volume 35, pages 6993–7001, 2021. 410

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution,

and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell., 40(4):834–848, 2017. 413

[10] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. Frank Wang, and M. Sun. No more discrimination: Cross city adaptation of road scene segmenters. In Int. Conf. Comput. Vis., pages 1992–2001, 2017. 409

[11] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. IEEE transactions on pattern analysis and machine intelligence, 44(7):3366–3385, 2021. 410

[12] P. Dhar, R. V. Singh, K.-C. Peng, Z. Wu, and R. Chellappa. Learning without memorizing. In IEEE Conf. Comput. Vis. Pattern Recog., 2019. 410

[13] T. G. Dietterich. Ensemble methods in machine learning. In International workshop on multiple classifier systems, pages 1–15. Springer, 2000. 417

[14] X. Ding, Y. Guo, G. Ding, and J. Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In Int. Conf. Comput. Vis., October 2019. 411

[15] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun. Repvgg: Making vgg-style convnets great again. In IEEE Conf. Comput. Vis. Pattern Recog., 2021. 411

[16] A. Douillard, Y. Chen, A. Dapogny, and M. Cord. Plop: Learning without forgetting for continual semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., 2021. 409, 410, 412, 413, 414, 417

[17] A. Douillard, M. Cord, C. Ollion, T. Robert, and E. Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In Eur. Conf. Comput. Vis., volume 12365, pages 86–102, 2020. 410

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 413

[19] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857, 2017. 410

[20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33:21271–21284, 2020. 411

[21] T. L. Hayes, K. Kafle, R. Shrestha, M. Acharya, and C. Kanan. Remind your neural network to prevent catastrophic forgetting. In Eur. Conf. Comput. Vis., pages 466–483, 2020. 410

[22] K. He, R. Girshick, and P. Dollár. Rethinking imagenet pre-training. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4918–4927, 2019. 409

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE Conf. Comput. Vis. Pattern Recog., 2016. 413

[24] Z. Huang, W. Hao, X. Wang, M. Tao, J. Huang, W. Liu, and X.-S. Hua. Half-real half-fake distillation for class-incremental semantic segmentation. arXiv preprint arXiv:2104.00875, 2021. 410

[25] C. Häne, C. Zach, A. Cohen, and M. Pollefeys. Dense semantic 3d reconstruction. IEEE Trans. Pattern Anal. Mach. Intell., 39(9):1730–1743, 2017. 409

[26] A. Iscen, J. Zhang, S. Lazebnik, and C. Schmid. Memory-efficient incremental learning through feature adaptation. In Eur. Conf. Comput. Vis., pages 699–715, 2020. 410

[27] M. Kanakis, D. Bruggemann, S. Saha, S. Georgoulis, A. Obukhov, and L. Van Gool. Reparameterizing convolutions for incremental multi-task learning without task interference. In Eur. Conf. Comput. Vis., pages 689–707, 2020. 410, 411

[28] C. D. Kim, J. Jeong, S. Moon, and G. Kim. Continual learning on noisy data streams via self-purified replay. In Int. Conf. Comput. Vis., pages 537–547, 2021. 410

[29] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526, 2017. 410

[30] S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? In Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition, pages 2661–2671, 2019. 409

[31] Z. Li and D. Hoiem. Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell., 40(12):2935–2947, 2017. 414

[32] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. IEEE Transactions on Pattern analysis and machine intelligence, 33(2):353–367, 2010. 415

[33] Y. Liu, B. Schiele, and Q. Sun. Adaptive aggregation networks for class-incremental learning. In IEEE Conf. Comput. Vis. Pattern Recog., 2021. 410

[34] A. Maracani, U. Michieli, M. Toldo, and P. Zanuttigh. Recall: Replay-based continual learning in semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7026–7035, 2021. 415

[35] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In PsychologLearniny of learning and motivation, volume 24, pages 109–165. Elsevier, 1989. 409

[36] U. Michieli and P. Zanuttigh. Incremental learning techniques for semantic segmentation. In Int. Conf. Comput. Vis. Worksh., 2019. 414, 417

[37] U. Michieli and P. Zanuttigh. Incremental learning techniques for semantic segmentation. In Proceedings of the IEEE/CVF international conference on computer vision workshops, pages 0–0, 2019. 414

[38] U. Michieli and P. Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1114–1124, 2021. 410, 414

[39] U. Michieli and P. Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In IEEE Conf. Comput. Vis. Pattern Recog., 2021. 414

[40] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. SIAM journal on control and optimization, 30(4):838–855, 1992. 411, 412, 416, 417

[41] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning. In IEEE Conf. Comput. Vis. Pattern Recog., 2017. 410

[42] C. Simon, P. Koniusz, and M. Harandi. On learning the geodesic path for incremental learning. In IEEE Conf. Comput. Vis. Pattern Recog., 2021. 410

[43] P. Singh, P. Mazumder, P. Rai, and V. P. Namboodiri. Rectification-based knowledge retention for continual learning. In IEEE Conf. Comput. Vis. Pattern Recog., pages 15282–15291, 2021. 410

[44] P. Singh, V. K. Verma, P. Mazumder, L. Carin, and P. Rai. Calibrating cnns for lifelong learning. In Adv. Neural Inform. Process. Syst., volume 33, 2020. 410

[45] J. Smith, Y.-C. Hsu, J. Balloch, Y. Shen, H. Jin, and Z. Kira. Always be dreaming: A new approach for data-free class-incremental learning. In Int. Conf. Comput. Vis., 2021. 410

[46] V. K. Verma, K. J. Liang, N. Mehta, P. Rai, and L. Carin. Efficient feature transformations for discriminative and generative continual learning. In IEEE Conf. Comput. Vis. Pattern Recog., 2021. 410

[47] F.-Y. Wang, D.-W. Zhou, H.-J. Ye, and D.-C. Zhan. Foster: Feature boosting and compression for class-incremental learning. In European conference on computer vision, pages 398–414. Springer, 2022. 410

[48] S. Yan, J. Xie, and X. He. Der: Dynamically expandable representation for class incremental learning. In IEEE Conf. Comput. Vis. Pattern Recog., 2021. 410

[49] S. Yan, J. Zhou, J. Xie, S. Zhang, and X. He. An em framework for online incremental learning of semantic segmentation. In Proceedings of the 29th ACM international conference on multimedia, pages 3052–3060, 2021. 410

[50] L. Yu, X. Liu, and J. Van de Weijer. Self-training for class-incremental semantic segmentation. IEEE Transactions on Neural Networks and Learning Systems, 2022. 415

[51] L. Yu, X. Liu, and J. Van de Weijer. Self-training for class-incremental semantic segmentation. IEEE Transactions on Neural Networks and Learning Systems, 34(11):9116–9127, 2022. 415

[52] C.-B. Zhang, J. Xiao, X. Liu, Y. Chen, and M.-M. Cheng. Representation compensation networks for continual semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, 2022. 409

[53] C.-B. Zhang, J.-W. Xiao, X. Liu, Y.-C. Chen, and M.-M. Cheng. Representation compensation networks for continual semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7053–7064, 2022. 409, 410, 413, 414

[54] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In IEEE Conf. Comput. Vis. Pattern Recog., 2017. 413

[55] F. Zhu, X.-Y. Zhang, C. Wang, F. Yin, and C.-L. Liu. Prototype augmentation and self-supervision for incremental learning. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5871–5880, 2021. 410

[56] K. Zhu, W. Zhai, Y. Cao, J. Luo, and Z.-J. Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9296–9305, 2022. 410