基于网络服务邻近性的强化学习智能体域适应

Kaushik Dey, Satheesh K. Perepu爱立信研究院（人工智能）印度{deykaushik,perepu.satheesh.kumar}@ericsson.com

Pallab Dasgupta, Abir Das印度理工学院克勒格布尔分校印度{pallab,abir}@cse.iitkgp.ac.in

arXiv:2303.01013v1 [cs.LG] 2 Mar 2023

摘要—**无线网络中服务需求具有动态演变的特性，这促使电信行业考虑采用智能、自适应的强化学习（RL）智能体，以管理日益丰富的网络服务。随着未来采用6G网络，预计将涌现许多新型服务，其中有时会由网络外部的应用定义这些服务。若不进行域适应，为管理特定服务类型而训练的强化学习智能体可能并不适合管理另一种服务类型。我们提出一种简单的启发式方法，用于评估新服务与现有服务之间的邻近度，并证明通过明确定义的域适应流程，与新服务最接近的服务所对应的强化学习智能体能够迅速适应新服务类型。我们的方法使经过训练的源策略能够在动力学发生变化的新情境中进行适配，而无需重新训练新策略，从而显著提升计算效率和成本效益。面对快速演变的服务类型，这类领域自适应技术或将很快为基于强化学习的通用服务管理奠定基础。**

索引词—**无线网络、服务管理、强化学习**

I. 引言

在大型无线网络的服务管理中引入自动化智能控制正变得无处不在，并且普遍预计这将成为管理未来6G网络中需要支持的大量服务类型的关键推动因素。在当前5G网络中，服务通常被大致归入少数几种类型，例如会话视频（计算机视觉）、超可靠低时延通信 （超可靠低时延通信）和大规模物联网 （海量物联网），但预计此类服务类型的数量将在不久的将来大幅增加。

现有文献和实践已证明，使用强化学习（RL）智能体以最优方式控制网络服务参数具有诸多优势 [6, 8,17]。强化学习智能体可以针对特定服务和无线电环境进行训练，从而以最优方式控制空中接口或核心网络中的参数，自动满足网络的预期和需求。此类强化学习智能体可以部署在编排层、核心网络层，甚至 gNodeB 等边缘节点上[1]。本研究假设强化学习策略部署在核心网络层或编排层，并使用单一策略管理一组 gNodeB。

[1]5G中的gNodeB在功能上等同于传统系统中的基站，负责与其覆盖区域（即小区站点）内的用户设备（UE）进行无线通信。

然而，仅靠这种方法可能还不够，因为在不久的将来，从 5G 迈向 6G 将进一步加剧网络各个部分对原生 AI 功能的需求 [18, 21]。随后，外部应用将动态生成新的服务类型，网络中的消费模式也会发生变化。随着移动模式的变化以及 gNodeB 之间相关切换的发生，底层无线电环境也将随之改变 [19]。在城市场景中，无线电环境可能会迅速变化（例如，每 10 分钟或更短时间变化一次）。

因此，强化学习策略可能需要频繁重新训练。但从能耗和计算资源的角度考虑，尽管所消耗的服务类型或流量分布发生变化，也不适合对强化学习智能体进行重新训练。另一方面，在某一特定环境（例如用户设备（UE）在各个gNodeB之间的分布）中训练的智能体，在用户设备分布发生变化的不同环境中可能无法实现最佳性能。因此，必须采用能够使强化学习智能体适应不同服务类型和环境的人工智能算法。这种在有限或无需重新训练的情况下适应不同领域、新服务或环境条件的能力，可能是未来网络服务管理中有效部署智能体的关键竞争差异化因素 [19]。用强化学习的术语来说，我们希望能够将在一个领域中学习到的策略，通过在另一个相关领域中获得的相应观测和动作进行适应，使得下一状态的观测相似或足够接近 [20]。

本文借鉴强化学习文献中研究的领域自适应技术，提出了一种将针对某类服务训练的强化学习智能体适配到另一类服务的方法。我们证明，适配的速度和质量取决于对原始服务与目标服务之间接近度的度量。这使我们能够采用一种启发式方法，选择可快速且可靠地适配到目标服务的源策略。本文的主要贡献如下：

1) 我们首次提出一种利用循环一致生成对抗网络（循环生成对抗网络）在电信领域的不同服务间实现强化学习策略泛化的方法。结果表明，与现有方法相比，我们的方法的样本效率高出 5–6 倍（图 6a、7a）。2) 我们验证了所提出的方法适用于无线电环境因移动模式而频繁变化的情况。该方法能够实现快速

适配，且从计算资源和样本效率的角度来看成本可控。与重新训练相比，我们的方法的样本效率几乎高出 8–9 倍（表 II），并且能够在零接触场景中实现策略适配，因为在新环境中不需要奖励信号。

3）我们提出了一种启发式方法，可在启动对应关系映射之前，预先判断域适应的效果。域适应并非在所有情况下都能取得良好效果，而我们提出的启发式方法能够成功识别此类分布变化。此外，该指标还可指导我们选择哪项服务作为源服务，以便将其适配到更多新服务。

本文组织如下：第 II 节概述域适应领域的相关工作。第 III 节阐述域适应的必要性，并介绍所提出的方法，同时详细说明用于估计服务邻近性的启发式方法。第 IV 节介绍实验设置，第 V 节展示采用所提方法获得的结果。第 VI 节给出总结性意见和未来研究方向。

II. 强化学习域适应相关工作

在强化学习领域，现有文献提出了多种将策略适配到新领域的方法。这些方法大致可分为以下三类：

1）表征学习。为了使两个领域对齐，该方法尝试找到相对于源领域和目标领域的不变表征 [2，7，12]。这种方法基于以下假设：根据两个环境的共性学习得到的策略，不会受到目标领域相对于源领域发生的变化的影响。然而，如果无法通过创建表征增强来解释两个领域之间的差异，该方法便会面临挑战。此外，当两个环境的动力学不同时（如本文涉及的三种不同服务），该技术还需要配对数据，而收集这类数据既成本高昂，又耗时。

2) 元学习。这里，模型在任务分布 [1], 上进行训练，使其能够利用所学知识，从同一分布中学习新任务。深度元强化学习 [16]将元学习扩展到强化学习场景，在该场景中，模型通过学习策略，在新环境中选择最优动作，从而学会适应新任务。这类方法已被证明有助于模型在接触新环境时，以较少的样本学习近乎最优的动作轨迹。然而，这些技术通常在模型使用任务分布而非单一任务进行训练时表现出色。因此，这类方法可能并不特别适合我们在本研究中尝试的一对一迁移。

3) 对应关系学习。与学习不变表征不同，这里我们着重学习状态之间的映射，并对齐动作，使不同领域之间的下一状态转移相对应。这类方法 [20]可通过循环生成对抗网络 [23]实现，且不需要不同环境之间配对的状态-动作元组。最重要的是，无需了解目标环境中的奖励函数。文献 [20] 中的方法依靠学习不同输入模态、物理参数和形态之间的对应关系。在我们的实验中，所有服务的数据包传输方式各不相同，而且海量物联网服务与超可靠低时延通信服务的关键绩效指标（关键绩效指标）也有所不同。因此，控制参数、分组优先级或最大比特率 （最大比特率）的任何变化都会对关键绩效指标产生不同影响。因此，我们主要处理的是不同环境之间物理参数或动力学发生变化的情形，而适应后的策略必须在新环境中找到近乎最优的状态-动作对应关系。

使用循环一致性损失的循环生成对抗网络传统上用于执行图像到图像转换和重建。然而，循环生成对抗网络也已应用于图像中的域适应 [3] ，随后又通过在仿真图像与真实世界图像之间进行图像转换，应对从仿真到现实的挑战 [4]。近来，研究人员开始探索将 CycleGAN 用于强化学习，其中利用强化学习场景一致性损失 [9] 来学习一种任务感知转换，使其不受图像 Q值的影响。

上述所有研究都主要将循环生成对抗网络用于视觉对应关系，直到 [20] 将其用于对齐不同的动力学和跨物理对齐。然而，迄今为止，此类方法尚未在电信或网络领域中尝试过；在这些领域中，数据包传输的特性会因服务而异，甚至对每项服务的预期也可能通过不同的关键绩效指标来衡量。

在多维动作设置中，一个回合单步的时长可能取决于所采取的动作类型，这使得建立对应关系更具挑战性。在我们的案例中，优先级动作需要约 40 秒才能在模拟器中产生稳定效果，而最大比特率动作不到 2 秒便可显现效果。还需考虑到，在网络环境中，数据收集（由 3 个元素组成的元组样本）与 MUJOCO [14] 或视觉仿真环境 [5] 相比也是一项成本高昂的过程。因此，必须基于少得多的数据建立对应关系，且所提出的方法需要证明其相较于现有技术的样本效率。

III. 方法概述

我们来看看以下可能需要进行领域适配的情况。强化学习智能体使用优先级和最大比特率(MBR)作为控制参数。

1) 移动模式会在一天中发生变化。因此，分布在各个 gNodeB 上的用户设备(UE)也会随之变化，从而改变流量模式。

2) 可能会引入新服务。在我们的研究中，我们考察了三种不同类型的服务，即交互式视频（CV）、超可靠低时延通信（URLLC）和海量物联网（mIoT）。首先，我们仅使用一种服务（例如 URLLC）训练策略。然后，我们尝试将 URLLC 策略适配到其他服务类型，例如 mIoT 或 CV。

在我们的工作中，我们提出了一种方法：不进行重新训练，而是在两个环境之间建立状态-动作对的对应关系。在建立对应关系的过程中，我们不会使用来自新环境的任何奖励信号。这种适配速度更快、计算成本更低，因此更适合部署在计算能力有限的边缘设备上。此外，由于无需进行人工奖励工程，整个过程无需人工干预即可轻松自动化。

在本研究中，我们使用了一个网络仿真器，可模拟三种服务：（i）交互式视频（CV），（ii）超可靠低时延通信（URLLC），以及（iii）海量物联网（mIoT）。我们可以针对每种服务设置特定关键绩效指标（KPI）的意图或预期。URLLC 和海量物联网服务的预期以关键绩效指标分组错误率（PER）表示，而计算机视觉服务则通过关键绩效指标体验质量（QoE）进行衡量。为便于分析，我们选择了两个算法可调整的控制参数，即最大比特率（MBR）和分组优先级。

A. 域适应的必要性

为开展研究，我们采用一种名为 PPO [11] 的强化学习（RL）技术，在每个源域中生成最优策略，并将其直接应用于目标域。例如，源域可以是超可靠低时延通信，目标域则可以是海量物联网或计算机视觉。图 1 和图 2 展示了一项实验：在海量物联网和计算机视觉环境中测试超可靠低时延通信策略。显然，该策略的表现远不理想，无法达到海量物联网或计算机视觉各自设定的目标。

同样，在某个环境中针对分布于若干 gNodeB 的用户设备分布训练得到的超可靠低时延通信策略，当用户设备分布发生变化时，表现也会不佳。图 9 展示了这样一种情况：由于移动性，分布于各 gNodeB 的用户设备从高斯分布变为均匀分布。

因此，本文提出一种域适应技术，使得在一个环境中训练的策略能够适配并调整到另一个环境或另一项服务。

B. 领域自适应方法

为实现跨领域对齐，我们提出了一种在无序状态和动作对之间学习对应关系的方法。我们训练状态对应关系模型和

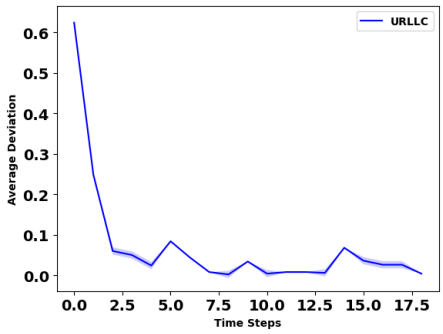


图 1：URLLC 服务上的 URLLC 训练性能

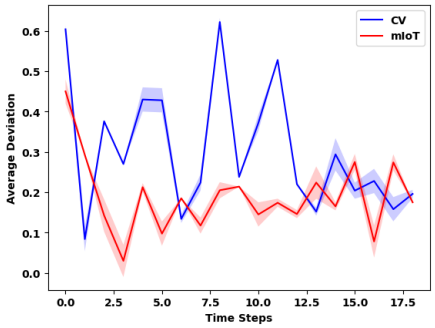


图 2：URLLC 策略在海量物联网和计算机视觉上的性能

动作对应模型，并结合动力学循环一致性 [20], ，以对齐不同领域中后续状态的序列。我们在源领域和目标领域中均使用由当前状态、动作和后续状态组成的元组。在源环境中训练得到的策略用于收集两个领域中的元组批次。随后，通过循环一致生成对抗网络（Cycle-GAN）[3, 22, 23]对这些无序元组对进行对齐。

我们通过一个用例来描述该方法：策略在 URLLC 服务上进行训练，并在海量物联网服务上进行域适应。

1) 在超可靠低时延通信环境中执行训练好的 URLLC 策略多个回合，以收集以下形式的三元组：（状态、动作、下一状态）：

$$\tau^u = (s_t^u, a_t^u, s_{t+1}^u)$$

2) 还会在移动物联网环境中执行训练好的 URLLC 策略，但不进行任何适配，以收集以下形式的三元组：

$$\tau^m = (s_t^m, a_t^m, s_{t+1}^m)$$

3) 下一步，为了对齐并建立状态-动作对之间的对应关系，我们创建了一个观测对齐函数，G：$S^u \rightarrow S^m$，以及一个动作对齐函数，H：$S^u \times A^u \rightarrow A^m$ 及其逆函数 H^{-1}：$S^m \times A^m \rightarrow A^u$。因此，给定超可靠低时延通信和海量物联网之间的初始对应状态，所选择的动作应使后续状态

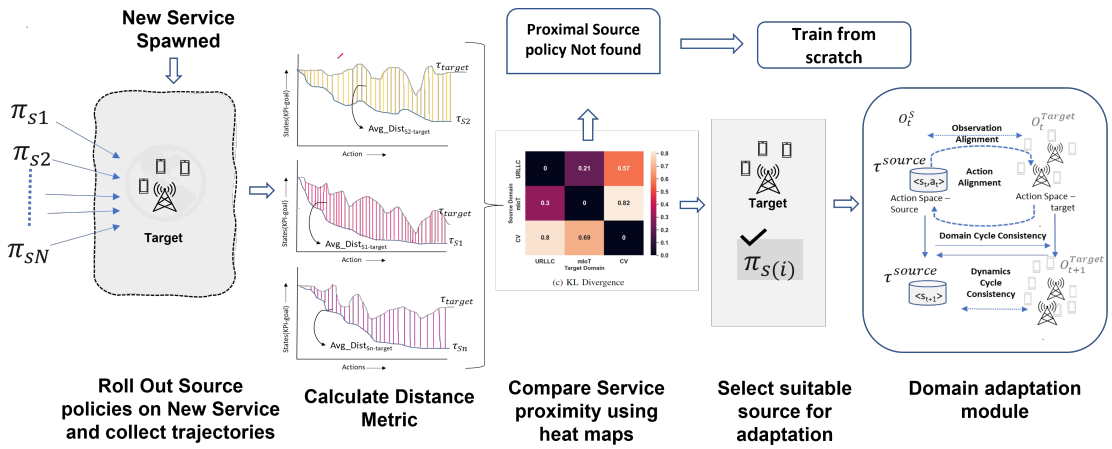


图 3：所提出的领域自适应方法概览

也保持对应，即若 $S_t^u \leftrightarrow S_t^m$ 且 $S_{t+1}^u \leftrightarrow S_{t+1}^m$，则 $(s_t^u, a_t^u) \leftrightarrow (s_t^m, a_t^m)$，我们将其称为动作对应。通过对抗训练实现不同领域间的状态映射，并通过循环一致性损失 [24]，实现动作对应，从而使通过 H 获得的动作能够经由 H^{-1} 映射回来。

4）引入领域循环一致性损失，$Loss_{domclc}$，通过该损失使用 $H^{S^u,\ A^u}$ 来获得 $\tilde{a}^m$。然后应用 $H^{-1}(S^m,\ \tilde{a}^m)$，以得到预期等于 A^u 的 $\tilde{a}^u$ 。

5）除了状态与动作对应之外，我们还需要动力学对应，它通过动力学循环一致性函数引入，用于映射两个领域间的转移动力学。给定分别由 G 和 H 得到的 S_m^t 和 A_m^t，我们可以创建一个映射，使得 $T_m(\tilde{s}_m^t,\ \tilde{a}_m^t) \leftrightarrow s_u^{t+1}$，其中 T_m 是 mIoT 领域中的转移函数，使得 $T_m(s_m^t,\ a_m^t) \rightarrow s_m^{t+1}$。因此，$T_m(\tilde{s}_m^t,\ \tilde{a}_m^t)$ 应映射到 $G:(s_u^{t+1})$，这也可以表示为一个循环一致性函数。

6) 上述目标通过损失函数 [20]共同优化，表达式如下：

$$\begin{aligned} Loss_{full} \quad = \quad & \lambda_0(Loss_{dyncyc}(G,H)) + \\ & \lambda_1(Loss_{adv}(H, D_A^m) + Loss_{adv}(H^{-1}, D_A^u) \\ & \quad + Loss_{domclc}(H, H^{-1})) + \\ & \lambda_2(Loss_{adv}(G, D_{S_m})) \end{aligned}$$

其中，λ_0、λ_1 和 λ_2 为用于平衡各项损失的常数。

7) 在大幅缩小的样本空间上执行上述步骤（与从头训练相比），即可基于原始 URLLC 策略获得适用于 mIoT 领域的域适应策略。

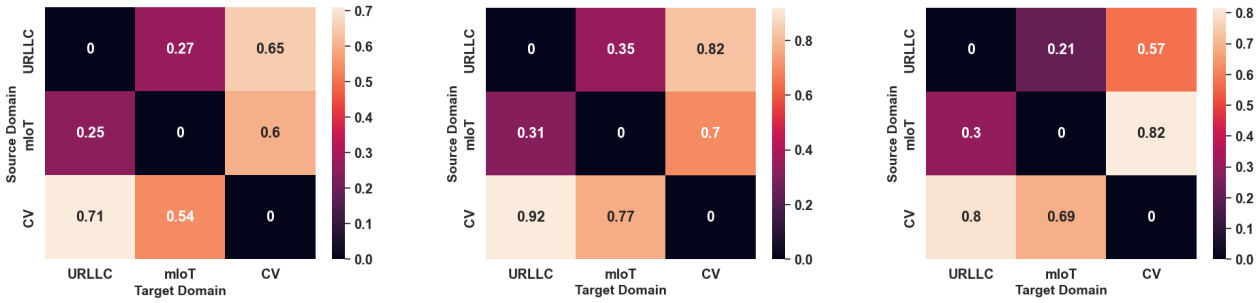
域适应方法概述如图 3 所示。$S^{source(1..n)}$ 表示每个都有训练好的策略 $\pi_{s(1...N)}$ 可用的源状态。S^{target} 表示新建或变更后的域中的状态，该域的产生原因可能是新增服务，或因移动性及跨 gNodeB 的切换变化而导致无线电环境改变。此处采用相同的方法：首先通过服务邻近度启发式选择合适的源 $S^{source(i)}$ ，然后应用观测对齐函数，如下所示：$G:\ S^{source_{(i)}} \rightarrow\ S^{target}$；动作对齐函数为 $H:\ S^{source_{(i)}} \times A^{source_{(i)}} \rightarrow A^{target}$，其逆函数为 $H^{-1}:\ S^{target} \times A^{target} \rightarrow A^{source_{(i)}}$ 。其余域适应过程与上文详述的过程类似。

C. 服务邻近性的启发式方法

我们在第五节介绍了关于域适应的研究。我们发现，选择源策略是更快获得更优策略的重要步骤。鉴于现有服务种类繁多，谨慎地选择源策略以适配新服务至关重要。我们提出了一种基于服务邻近性的启发式方法来进行选择。

如果两个领域相似，那么在相似状态下采取相同动作最优的可能性就很高。这两个领域中相似状态之间的距离也会较小，因此我们可以将该距离用作启发式依据。为了计算两个领域中相似状态之间的距离，宽松双模拟距离 [13]这一概念已得到广泛应用。与双模拟密切相关的马尔可夫决策过程同态理论，有助于构建依赖状态的动作映射 [10, 15]。这两种方法都已广泛应用于迁移学习。然而，宽松双模拟度量需要两个环境中的奖励信息，以及转移概率之间的坎托罗维奇距离。对于新领域，这两种度量所需的信息通常都无法预先获得。此外，奖励构造可能需要投入奖励工程，而这往往需要人工完成且耗时。因此，鉴于我们假设新领域没有可用信息，这些方法都不适用。

我们的启发式方法通过在目标域上运行源策略来计算。我们考虑源策略所采取的动作序列



(a) 欧几里得距离 (b) 曼哈顿距离 (c) KL 散度

图 4：源域与目标域性能之间的热图

从初始状态 S_0 出发，并记录在源域 A 中访问的状态轨迹，其形式为 $deviation_A = \tau_A - Target_A$。随后，我们在目标域中从 S_0 开始执行相同的动作序列，并记录在目标域中访问的状态轨迹 $deviation_B = \tau_B - Target_B$。之后，我们计算 $deviation_A$ 与 $deviation_B$ 之间的距离。此处采用的 $deviation$ 度量与第 V 节中强调的目标条件因子的状态定义一致。源域与目标域之间的接近度由多条轨迹上的平均距离决定。此处的多条轨迹对应于两个领域中的不同初始状态 S_0，和/或不同领域中的不同目标值。

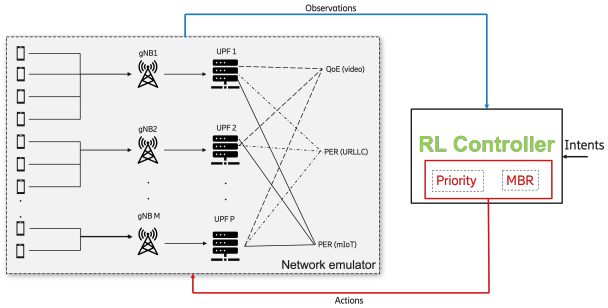


图 5：网络仿真器概览

我们计算目标域与各个源域之间的三种距离，即欧几里得距离、曼哈顿距离和*KL* 散度。欧几里得距离、曼哈顿距离和 KL 散度的热图分别如图 4a、图 4b 和图 4c 所示。为验证我们的方案，我们考虑了已知的服务：URLLC、mIoT 和 CV；这些服务的数据充足，且已训练好相应策略。计算热图时，我们考虑这些服务的所有有序对，每次将其中一个视为源域，另一个视为目标域。

从这些图中可以明显看出，与涉及 CV 的有序对相比，〈URLLC、mIoT〉和〈mIoT、URLLC〉的启发式距离更小。这与领域知识相符，并且第 V 节{此处原文连写：theresultspresentedinSectionV}中呈现的结果进一步印证了这一点。

IV. 实验设置

为了测试电信服务的域适应，我们构建了一个定制仿真器，用于模拟 5G 网络的运行。网络仿真器的概览如图 5 所示。该仿真器支持三种服务，即计算机视觉、超可靠低时延通信和海量物联网。这些服务承载的流量各不相同，因此会产生不同类型的流量模式。在仿真器中，用户设备被建模为通过用户面功能和 gNodeB 与应用层进行流量的发送和接收。为简化分析并清晰呈现从结果中得出的推论，我们假设用户设备在任意时刻只使用一种服务。

重申一下，我们为上述每种服务各选取了一个关键绩效指标：计算机视觉对应体验质量，超可靠低时延通信和海量物联网对应分组错误率。我们选取两个变量，即优先级和最大比特率（MBR），来控制这些关键绩效指标。优先级和 MBR 的值都与用户设备可获得的下行吞吐量成正比。

在本研究中，我们选择在服务级别控制优先级，即为服务选择的优先级值将应用于所有使用该服务的用户设备。不过，我们决定在用户设备级别控制最大比特率。这些变量所考虑的取值范围如表 I 所示。

表 I：关键绩效指标的选定范围

KPI	范围
QoE	$[1, 5]$
丢包率	[0%, 100%]
优先级	$\{0, \cdots, 10\}$
MBR	[0, 5] 兆字节每秒

我们的目标是将针对某项服务训练的策略适配到其他服务，或适配到变化后的无线电环境。我们还通过在仿真器中于后台运行另一项服务，并随机更改其优先级和最大比特率值，来模拟噪声。这会迫使策略不断调整最优控制参数，并防止简单的贪心方法达到最优。

五、结果与讨论

本节展示以下实验的结果。

1) 跨服务的域适应。在这里，我们在一种服务中进行训练，并观察其如何适配其他服务。我们考虑超可靠低时延通信、海量物联网和计算机视觉之间的所有配对。2) 跨两种无线电环境的域适应。在这里，我们考虑不同 gNodeB 间用户设备分布的差异。我们采用超可靠低时延通信服务作为测试平台。

A. 跨服务类型的域适应

在这里，我们在每个源域上训练一个强化学习策略。对于计算机视觉领域，目标是将体验质量维持在目标值；而对于超可靠低时延通信和 mIoT 领域，目标是将 分组错误率 维持在目标值。换言之，体验质量是计算机视觉的受控变量，而分组错误率是超可靠低时延通信和 mIoT 的受控变量。所有这些领域的动作空间相同，即服务的优先级以及服务中用户设备的最大比特率。对于所有领域，所考虑的观测空间如下：

$$\text{Observation Space} = \langle K, |K - target|/K \rangle \tag{1}$$

其中， K 是测得的受控变量当前值， $target$ 是目标关键绩效指标值。这种状态空间表示方式有两个优点：(1) 有助于构建目标条件学习，使算法能够在状态空间中实现任意期望的目标值；(2) 鉴于体验质量和分组错误率的测量尺度略有不同，这种表示方式可在状态空间中实现伪归一化。动作空间为

$$\text{Action Space} = [\text{Prio}, \text{MBR}_1, \text{MBR}_2, \cdots, \text{MBR}_N] \tag{2}$$

其中，MBR_i 是为 i^{th} 用户设备选择的最大比特率值， N 是服务中的用户设备数量，Prio 是为该服务选择的优先级值。在所有这些场景中，我们假设每项服务内$N = 4$ 。所考虑的奖励为

$$\text{Reward} = -|K - target| \tag{3}$$

其中， K 是受控变量的当前值。

对于每种服务类型，我们训练一个单智能体强化学习策略，以获得最大的全局奖励。现在，我们的目标是使用对抗训练，将学习到的策略适配到一个目标域，即另外两种服务类型之一。为了评估适配后策略的优劣，我们将目标域上训练的强化学习策略作为黄金参考。我们还确定适配后策略达到98%成功率所需的样本数量，也就是说，相对于黄金目标策略的偏差为2%。

我们使用第 II 节中讨论的领域自适应方法，将每个源策略适配到另外两个目标域。具体而言，我们准备以下策略进行比较：

1) 从头开始为目标域学习策略。我们收集目标域的样本，并在每收集 100 个样本后暂停训练。每次暂停期间，我们对训练好的模型进行 8 个回合的评估。对于这组回合，我们计算模型相对于目标域预训练黄金策略的平均偏差，以及偏差的标准差。我们持续上述过程，直到测试回合中的平均偏差达到 2%。本实验旨在研究从头开始训练的策略的学习进程。需要指出的是，本实验是为了与域适应进行比较；对于新服务，黄金策略并非先验存在。

2）通过从源域进行适配来学习策略。在此，我们使用第 II 节中讨论的方法，在目标域上适配源策略。我们使用随机权重初始化对应模型，并利用这些对应模型测量源策略在 8 个回合中的性能。随后，我们分别从源域和目标域收集 100 个样本，更新对应模型，并测量其在目标域上 8 个回合中的性能。对于每 100 个样本，重复这一过程，直到测试回合中的平均偏差达到 2%。

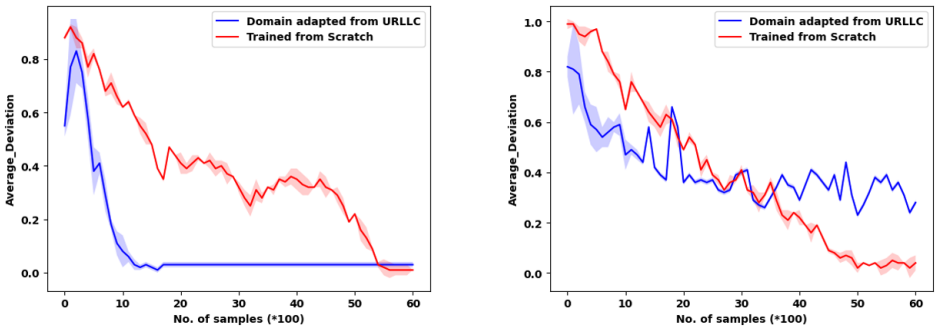
现在，我们概述分别以超可靠低时延通信、海量物联网和计算机视觉为源域时的结果。

1）场景 *1*：超可靠低时延通信作为源域：首先，我们训练一个强化学习策略，使体验质量维持在目标值。需要注意的是，目标是达到该目标值，而不是以任意幅度超过它。在存在随机噪声的情况下，这一目标有助于得到不会在资源累积方面盲目贪婪的策略。该策略在超可靠低时延通信服务上的性能如图 1 所示。从图中可以明显看出，该策略在超可靠低时延通信服务上进行测试时表现良好。

在进行域适应之前，我们先来考察 URLLC 策略在海量物联网和计算机视觉服务上未经任何适配时的性能。性能曲线如图 2 所示。从图中可以看出，URLLC 源策略在两个目标域上的表现都不理想。这凸显了进行域适应的必要性。

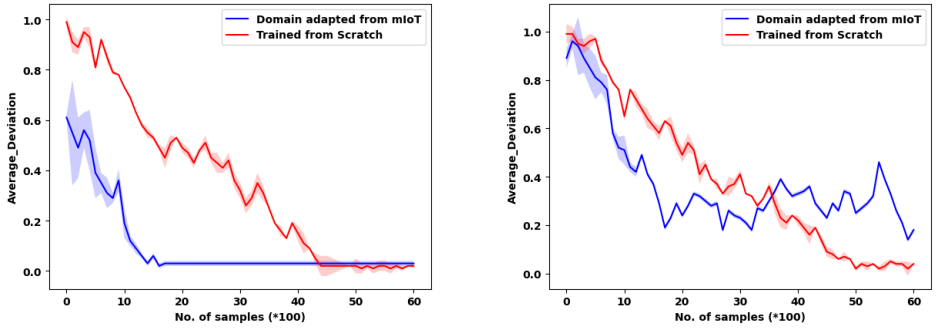
将从源 URLLC 策略适配而来的海量物联网和计算机视觉策略，与从头开始学习的相应策略进行比较，结果分别见图 6a 和图 6b。从图 6a 可以看出，在海量物联网目标域上，域适应策略约需 1200 个样本即可达到 2% 的目标平均偏差，而从头开始学习的策略则需要约 5500 个样本才能达到相同目标。这表明域适应方法具有较高的样本效率，这是实现快速适配的关键。有趣的是，使用循环生成对抗网络生成的对应关系进行此类适配时，不需要新环境的奖励结构。

图 6b 显示，对于计算机视觉域，从头开始训练的策略需要约 5000 个样本才能达到 2% 平均偏差的目标。另一方面，来自源域的域适应策略即使使用更多样本也未能达到预期目标，尽管它在初始阶段的学习速度明显快于从头开始训练的策略，并最终提升了相较于源策略的性能。因此



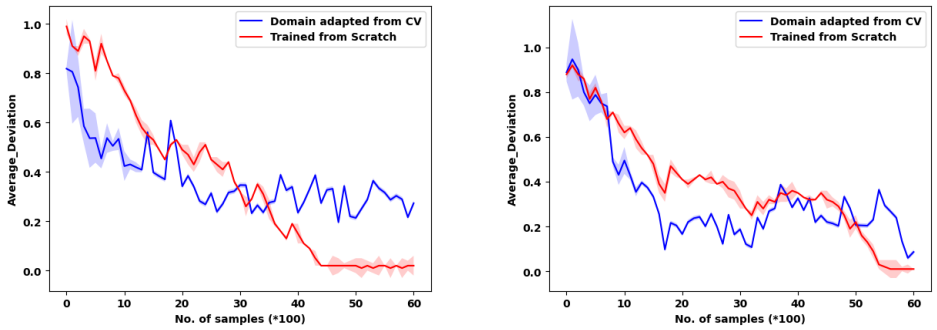
(a) 海量物联网　(b) 计算机视觉

图 6：比较以超可靠低时延通信为源领域的领域适配策略与从头学习的策略



(a) 超可靠低时延通信　(b) 计算机视觉

图 7：比较以海量物联网为源领域的领域适配策略与从头学习的策略



(a) 超可靠低时延通信　(b) 海量物联网

图 8：C 从计算机视觉作为源领域获得的领域适配策略与从头学习的策略的比较

在这里，域适配并未带来接近最优的性能，不同于海量物联网的情况。

本文给出了一个可能的解释，说明为什么从超可靠低时延通信领域到海量物联网领域的领域适配能够奏效，但从超可靠低时延通信领域到计算机视觉领域的领域适配却未能奏效。通常，在电信网络中，每项服务的流量模式会根据其传输的数据类型及传输方式而有所不同。对于超可靠低时延通信和海量物联网，我们可以看到相似的流量模式；而在计算机视觉中，我们可以看到完全不同的数据模式。多位领域专家也对此予以了确认三种服务之间的相似性或不相似性。这使我们推断，如果领域相似，域适应将带来更好的性能，并且与从头训练相比，样本效率会显著更高。另一方面，如果领域相似度较低，域适应可能需要更多样本，其表现也可能无法与从头训练相媲美。这促使我们采用第III-C节中提出的启发式方法。该方法特别指出了超可靠低时延通信与海量物联网之间的接近度，以及计算机视觉与这些服务之间缺乏接近度。

2) 场景-*2*：海量物联网作为源域：在这里，我们探讨将海量物联网源策略适配到超可靠低时延通信和计算机视觉。图 7a 比较了从头学习超可靠低时延通信策略与将海量物联网源策略适配到超可靠低时延通信的过程。从图中可以明显看出，与从头训练策略相比，获得域适应策略所需的样本数量更少。这与我们的启发式方法一致，该方法预测超可靠低时延通信与海量物联网之间具有接近度。

另一方面，图 7b 显示，将海量物联网源策略适配到计算机视觉并非良策，因为即使对应模型使用大量样本进行训练，域适应后的策略仍表现不佳。这再次印证了我们的启发式方法，该方法估计海量物联网与计算机视觉之间缺乏接近度。

3) 场景-*3*：计算机视觉作为源域：这里我们研究以计算机视觉为源域、以超可靠低时延通信和海量物联网为目标域的情况。结果分别如图 8a 和图 8b 所示。从图中可以明显看出，与从头开始训练的策略相比，这两种域适应的表现都较差。这也与我们的启发式方法一致，该方法表明计算机视觉与超可靠低时延通信和海量物联网都相距较远。

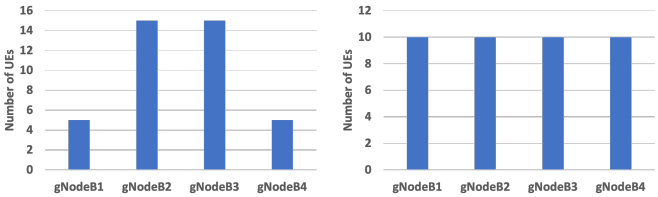
我们的结果支持所提出的用于估计服务邻近性的启发式方法，并证实了从邻近服务策略进行域适应的益处。与从头开始训练新策略相比，对邻近服务进行域适应具有更高的样本效率。此外，整个方法完全自动化，因为在新环境中不需要奖励信号。我们认为，域适应在 6G场景中将尤为有益，因为届时将有大量服务，其中一些会动态定义。每当出现新服务时，系统都可以即时自动地从邻近服务进行适配，而不必从头开始学习。

B. 不同无线电环境下的域适应

本节重点介绍底层无线电环境发生变化时域适应的优势。通常，城市中的移动模式相当动态，随着用户设备移动，gNodeB 之间的流量分布也会发生变化。当分布变化显著时，策略就需要进行适配。

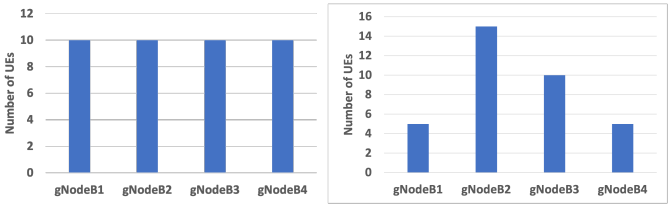
例如，图 9 展示了用户设备在各 gNodeB 之间的分布从高斯分布变为均匀分布。图 11a 对比了源策略在这两种环境中的表现，以及能够学习从高斯分布适配到均匀分布的域适应策略。从图中可以明显看出，原策略在新的分布下表现欠佳，而域适应策略的表现则好得多。鉴于无线电环境中经常出现此类情况，频繁重新训练通常难以实现，因此域适应似乎是可行的方向。

图 11b 展示了分布从均匀分布转变为泊松分布时的结果，如图 10 所示。同样，经过领域适配的策略在修改后的分布下表现良好。



(a) 现有分布（高斯分布） (b) 修改后的分布（均匀分布）

图 9：用户设备相对于 gNodeB 的分布



(a) 现有分布（均匀分布） (b) 修改后的分布（泊松分布）

图 10：用户设备相对于 gNodeB 的分布变化

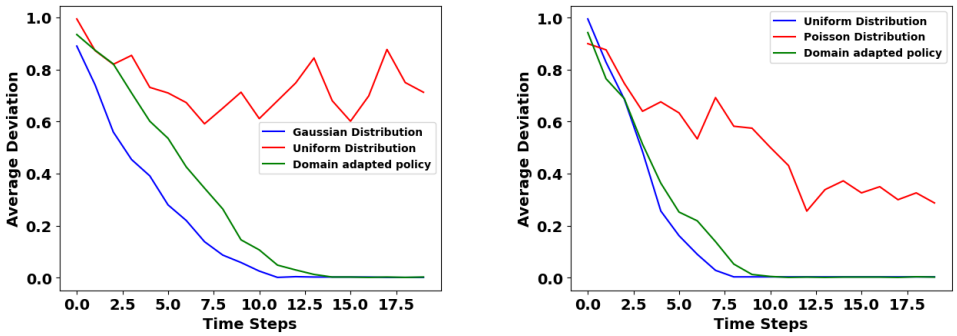
表 II 比较了经过领域适配的策略与从头训练得到的策略的训练表现。与从头训练的策略相比，经过领域适配的策略只需少得多的样本即可收敛到偏差低于 2%。这使得策略能够在新的环境中无需奖励信号即可快速自动适配。样本复杂度和计算需求的降低意义重大，推动了这类方法在边缘设备中的广泛应用。

VI. 结论

本文提出了域适应方法，以适应新的电信服务和无线电环境的变化。由于预计 6G 环境将包含许多服务，其中一些服务还可能动态演变，所提出的方法主张利用已训练的服务进行学习，从而快速制定出性能可接受的适配策略。该方法结合使用多个 CycleGAN，在旧服务与新服务之间构建对应网络，而不是针对新服务学习策略。此方法不需要新域的奖励信息或转移动力学。此外，我们还提出一种启发式方法，用于判断哪些现有策略可以复用，从而成功实现域适应。在定制仿真器上的实验结果表明，与针对新服务从头训练相比，所提出的方法效率更高。

表 II：达到 98% 性能所需的样本数

场景	域适应	从头训练
高斯分布到均匀分布	600	5000
均匀分布到泊松分布	500	4600



(a) 高斯分布到均匀分布　　(b) 均匀分布到泊松分布

图11：在源分布（蓝色）和新分布（红色）上执行源策略的对比，以及在新分布（绿色）上执行领域自适应源策略的对比。分布变化模拟了底层无线电环境的变化。

参考文献

[1] Chelsea Finn、Pieter Abbeel 和 Sergey Levine。用于深度网络快速适配的模型无关元学习。发表于国际机器学习会议，第 1126–1135 页。PMLR，2017 年。[2] Abhishek Gupta、Coline Devin、YuXuan Liu、Pieter Abbeel 和 Sergey Levine。通过强化学习学习不变特征空间以迁移技能。*arXiv* 预印本 *arXiv:1703.02949*，2017 年。[3] Judy Hoffman、Eric Tzeng、Taesung Park、Jun-Yan Zhu、Phillip Isola、Kate Saenko、Alexei Efros 和 Trevor Darrell。Cycada：循环一致对抗域适应。发表于国际机器学习会议，第 1989–1998 页。Pmlr，2018 年。[4] Stephen James、Paul Wohlhart、Mrinal Kalakrishnan、Dmitry Kalash- nikov、Alex Irpan、Julian Ibarz、Sergey Levine、Raia Hadsell 和 Konstantinos Bousmalis。通过仿真到仿真实现仿真到现实：利用随机化到规范化自适应网络实现数据高效的机器人抓取。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 12627–12637 页，2019 年。[5] Michał Kempka、Marek Wydmuch、Grzegorz Runc、Jakub Toczek 和 Wojciech Ja´skowski。Vizdoom：用于视觉强化学习的 Doom 人工智能研究平台。发表于*2016 IEEE*计算智能与游戏会议（*CIG*），第 1–8 页。IEEE，2016 年。[6] Talha Ahmed Khan、Asif Mehmood、Javier Jose Diaz Ravera、Afaq Muhammad、Khizar Abbas 和 Wang-Cheol Song。利用机器学习实现基于意图的网络切片编排与资源保障。发表于*NOMS 2020-2020 IEEE/IFIP*网络运营与管理研讨会，第 1–2 页。IEEE，2020 年。[7] Guoqing Liu、Chuheng Zhang、Li Zhao、Tao Qin、Jinhua Zhu、Jian Li、Nenghai Yu 和 Tie-Yan Liu。用于强化学习的基于回报的对比表示学习。*arXiv* 预印本 *arXiv:2102 10960*，2021 年。。[8] Satheesh K Perepu、Jean P Martins、Kaushik Dey 等。用于蜂窝网络中基于意图的服务保障的多智能体强化学习。*arXiv* 预印本 *arXiv:2208.03740*，2022 年。[9] Kanishka Rao、Chris Harris、Alex Irpan、Sergey Levine、Julian Ibarz 和 Mohi Khansari。RL-CycleGAN：强化学习感知的仿真到现实。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 11157–11166 页，2020 年。[10] Balaraman Ravindran 和 Andrew G Barto。近似同态：马尔可夫决策过程非精确最小化的框架。发表于国际知识型计算机系统会议，2004 年。[11] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*，2017 年。[12] Pierre Sermanet、Corey Lynch、Yevgen Chebotar、Jasmine Hsu、Eric Jang、Stefan Schaal、Sergey Levine 和 Google Brain。时间对比 [13] Jonathan Taylor、Doina Precup 和 Prakash Panagaden。近似马尔可夫决策过程同态中的性能损失上界。发表于 D. Koller，网络：从视频中进行自监督学习。载于*2018 IEEE*国际机器人与自动化会议（*ICRA*），第1134–1141页。IEEE，2018年。D. Schuurmans、Y. Bengio和L. Bottou编，神经信息处理系统进展，第21卷。Curran Associates, Inc.，2008年。[14] Emanuel Todorov、Tom Erez和Yuval Tassa。Mujoco：用于基于模型的控制的物理引擎。载于*2012 IEEE/RSJ*智能机器人与系统国际会议，第5026–5033页。IEEE，2012年。[15] Elise van der Pol、Daniel Worrall、Herke van Hoof、Frans Oliehoek和Max Welling。马尔可夫决策过程同态网络：强化学习中的群对称性。神经信息处理系统进展，33：4199–4210，2020年。[16] Jane X Wang、Zeb Kurth-Nelson、Dhruva Tirumala、Hubert Soyer、Joel Z Leibo、Remi Munos、Charles Blundell、Dharshan Kumaran和Matt Botvinick。学习强化学习。*arXiv*预印本 *arXiv:1611.05763*，2016年。[17] Yiming Wei、Mugen Peng和Yaqiong Liu。面向6G的意图驱动网络：见解与挑战。《数字通信与网络》，6(3)：270–280，2020年。[18] Wen Wu、Conghao Zhou、Mushu Li、Huaqing Wu、Haibo Zhou、Ning Zhang、Xuemin Sherman Shen和Weihua Zhuang。面向6G无线网络的AI原生网络切片。《*IEEE*无线通信》，（）：–，2022年。[19] Wen Wu、Conghao Zhou、Mushu Li、Huaqing Wu、Haibo Zhou、Ning Zhang、Xuemin Sherman Shen和Weihua Zhuang。面向6G无线网络的AI原生网络切片。《*IEEE*无线通信》，（）：–，2022年。[20] Qiang Zhang、Tete Xiao、Alexei A Efros、Lerrel Pinto和Xiaolong Wang。利用动力学循环一致性学习用于控制的跨域对应关系。*arXiv*预印本 *arXiv:2012.09811*，2020年。[21] Zhengquan Zhang、Yue Xiao、Zheng Ma、Ming Xiao、Zhiguo Ding、Xianfu Lei、George K Karagiannidis和Pingzhi Fan。6G无线网络：愿景、需求、架构与关键技术。《*IEEE*车载技术杂志》，14(3)：28–41，2019年。[22] Long Zhou、Joshua D Schaefferkoetter、Ivan WK Tham、Gang Huang和Jianhua Yan。利用CycleGAN进行低剂量FDG PET图像去噪的监督学习。《医学图像分析》，65：101770，2020年。[23] Jun-Yan Zhu、Taesung Park、Phillip Isola和Alexei A. Efros。使用循环一致对抗网络进行非配对图像到图像转换。载于*IEEE*国际计算机视觉会议（*ICCV*）论文集，2017年10月。[24] Jun-Yan Zhu、Taesung Park、Phillip Isola和Alexei A Efros。使用循环一致对抗网络进行非配对图像到图像转换。载于*IEEE*国际计算机视觉会议论文集，第2223–2232页，2017年。