**IntelliLight：智能交通灯控制的强化学习方法**

## 摘要

在我们的问题中，我们将环境E作为两条道路（以及该交叉路口的交通）的交叉点。有一个智能交通灯代理G.为了使符号更简单，我们用“N”，“S”，“W”，“E”分别代表北，南，西和东，并使用“红色”和“绿色”相应地代表红灯和绿灯。交通灯的设置被定义为相位（例如，西 - 东方向上的绿灯，其可以简化为Green-WE）。当灯光从绿色变为红色时，会有3秒黄灯，而其他方向仍然保持红色。因此，一个绿灯和随后的黄灯可以通过“绿色”一起表示。为了简化问题，我们假设交通灯只有两个阶段，即1）Green-WE，和2）Red-WE。由于现实世界设置的限制，交通信号灯只能按特定顺序改变（即1 - > 2 - > 1 - > 2 - > ......）。给定状态s（描述该交叉点附近的交通的位置和速度），代理G的目标是给出最佳动作a（即，是否将光改变到下一阶段），以便奖励r （即，交通的平稳性）可以最大化。

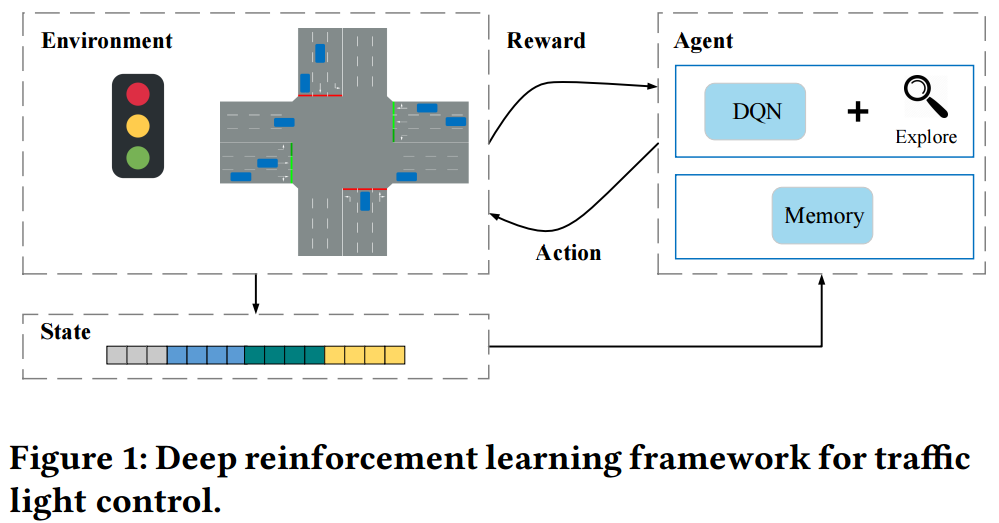
CCS概念

•计算方法→控制方法; •应用计算→运输;

**1引言**

交通拥堵变得越来越昂贵。 例如，根据福布斯2014年的一份报告，交通拥堵每年花费美国人1240亿美元[12]。 在欧盟，交通拥堵成本估计为其GDP的1％[7]。 改善交通状况可以提高城市效率，改善经济，并减轻人们的日常生活。

减少交通拥堵的一种方法是智能地控制交通灯。 如今，大多数交通信号灯仍然采用预先定义的固定时间计划[18,23]进行控制，而不是通过观察实际交通来设计的。 最近的研究根据实际交通数据提出了手工制定的规则[5,20]。 但是，这些规则仍然是预定义的，无法动态调整为w.r.t. 实时流量。



为了根据实时交通动态调整交通灯，人们一直在使用强化学习技术[13,22,24]。由于两个关键挑战，传统的强化学习很难应用：（1）如何表现环境; （2）如何建立环境与决策之间的相关性。为了应对这两个挑战，最近的研究[15,22]已经应用深度强化学习技术，例如深度Q学习（DQN），用于交通灯控制问题。图1说明了深度强化学习框架的基本思想。环境由交通灯相位和交通状况组成。 State是环境的特征表示。代理将状态作为输入并学习模型以预测是否“保持交通灯的当前阶段”或“改变当前阶段”。该决定被发送到环境并且奖励（例如，通过交叉路口的车辆数量）被发送回代理。代理因此更新模型并进一步基于新状态和更新模型对下一时间戳做出新决策。在这样的框架中，交通状况可以被描述为图像，并且这样的图像被直接作为基于CNN的模型的输入，以丰富手工制作的环境特征。

最近的深层强化学习方法为交通灯控制问题取得了有希望的进展。 我们的方法通过做出几个重要的新贡献来扩展这一工作范围：

1.实际交通数据的实验。

 如今，正在从各种来源收集越来越多的交通数据。在中国，许多大城市都安装了配备人工智能的交通监控摄像头，以实时监控交通状况。这种实时交通数据使我们能够在现实世界中实施强化学习。然而，据我们所知，现有的研究都没有使用真实的交通数据来测试他们的方法。相反，他们使用流量模拟，而这种模拟并不反映现实世界的流量。例如，当前研究中的模拟模型通常假设车辆以恒定速率到达，但实际交通随着时间的推移是高度动态的。在我们的论文中，我们测试了从中国济南的1,704台监控摄像机获得的大规模实际交通数据的方法，为期31天（详见实验部分）。在该数据集中，有超过4.05亿辆车辆记录和超过1100万辆独特车辆板块。我们对这样大的真实数据集进行了全面的实验。

2.对政策的解释。

量化交通灯控制性能的常用措施是检查总体奖励，这可以通过几个因素来定义，例如车辆的等待时间和通过交叉路口的车辆数量。但是，现有的研究很少从模型中学习政策。在某些情况下，奖励可能会产生误导。可能有不同的政策有相同的奖励，但一个比另一个更合适。以图2为例。假设南北方向只有流量，流量每120秒就有80秒。对于南北方向的绿灯，政策＃1为80秒，然后红灯为40秒，然后重复。政策＃2与政策＃1的不同之处在于，光线每隔10秒就会改变，而不是南北方向的40秒红灯。这两项政策都会产生相同的奖励，因为根据这两项政策，任何车辆都不会等待。但是，在实际情况下，策略＃1优先于策略＃2。在本文中，我们声称研究政策而不是简单地显示整体奖励是很重要的。在我们的实验中，我们展示了从不同情景下的实际流量中学到的几个有趣的策略（例如，高峰时段与非高峰时段，工作日与周末）。

3.相位门控模型学习。

 如前面深度强化学习框架中所述，代理将把作为模型输入的状态（即环境的表示）作为模型输入。环境通常包括当前的交通灯阶段和交通状况。例如，除交通灯阶段外，图3中两种情况的环境是相同的。以前的研究都将阶段作为一个特征[17,22]，以及许多其他特征（例如，不同车道的车辆数量，车辆的位置）。并且这一特征可能不会起到足以影响模型输出的作用。因此，对于这两种不同的情况，模型将做出相同的决定（即，保持或改变当前阶段）。但是，无论哪一个决定，这种决定对其中一个案件来说并不理想。因为在图3中，案例A希望保持阶段，案例B希望改变阶段。在本文中，我们提出了一种新的相位敏感（即相位门与记忆宫殿相结合）强化学习剂，它是一种导致卓越性能的关键组件。

本文的其余部分安排如下。 第2节讨论了文献。 第3节正式定义了问题。 该方法见第4节，实验结果见第5节。最后，我们在第6节中总结了这篇论文。

**2相关工作**

在本节中，我们首先介绍传统的交通灯控制方法，然后介绍使用强化学习的方法。

**2.1常规交通灯控制**

早期的交通灯控制方法大致可分为两组。 第一种是预定时信号控制[6,18,23]，其中根据历史交通需求确定所有绿色阶段的固定时间，而不考虑交通需求的可能波动。 第二种是车辆驱动的控制方法[5,20]，其中使用实时交通信息。 车辆致动方法适用于具有相对高的交通随机性的情况。 但是，这种方法在很大程度上取决于当前交通状况的手工规则，而没有考虑到未来的情况。 因此，它无法达到全局最优。

**2.2交通灯控制的强化学习**

最近，由于在以前的方法中无法处理动态多方向流量，更多的工作尝试使用强化学习算法来解决交通灯控制问题[13,17,24]。 通常，这些算法将道路上的流量视为状态，将光上的操作视为动作。 与固定时间和流量响应控制方法相比，这些方法通常表现出更好的性能。

[1,2,4,8,24]中的方法将状态设计为离散值，例如车辆的位置或等待车辆的数量。 然而，离散的状态 - 动作对值矩阵需要巨大的存储空间，这使得这些方法不能用于大的状态空间问题。

为了解决先前方法的可处理大的状态空间，最近的研究[15,22]提出使用连续状态表示来应用Deep Q学习方法。 这些研究学习Q函数（例如深度神经网络）来映射状态和动作以奖励。 这些工作在状态表示方面各不相同，包括手工艺特征（例如，队列长度[15,17]，平均延迟[10,22]）和图像特征[9,16,22]）它们在奖励设计方面也有所不同，包括 平均延迟[3,22]，平均旅行时间[16,22]和队列长度[15]。

但是，所有这些方法都假设相对静态的流量环境，因此远离实际情况。 此外，他们只关注奖励，忽略了算法对实际流量的适应性。 因此，它们无法解释学习光信号对应于交通的变化的原因。 在本文中，我们尝试在更真实的流量设置中测试算法，并添加除奖励之外的更多解释。

**3问题定义**

在我们的问题中，我们将环境E作为两条道路（以及该交叉路口的交通）的交叉点。有一个智能交通灯代理G.为了使符号更简单，我们用“N”，“S”，“W”，“E”分别代表北，南，西和东，并使用“红色”和“绿色”相应地代表红灯和绿灯。交通灯的设置被定义为相位（例如，西 - 东方向上的绿灯，其可以简化为Green-WE）。当灯光从绿色变为红色时，会有3秒黄灯，而其他方向仍然保持红色。因此，一个绿灯和随后的黄灯可以通过“绿色”一起表示。为了简化问题，我们假设交通灯只有两个阶段，即1）Green-WE，和2）Red-WE。由于现实世界设置的限制，交通信号灯只能按特定顺序改变（即1 - > 2 - > 1 - > 2 - > ......）。给定状态s（描述该交叉点附近的交通的位置和速度），代理G的目标是给出最佳动作a（即，是否将光改变到下一阶段），以便奖励r （即，交通的平稳性）可以最大化。

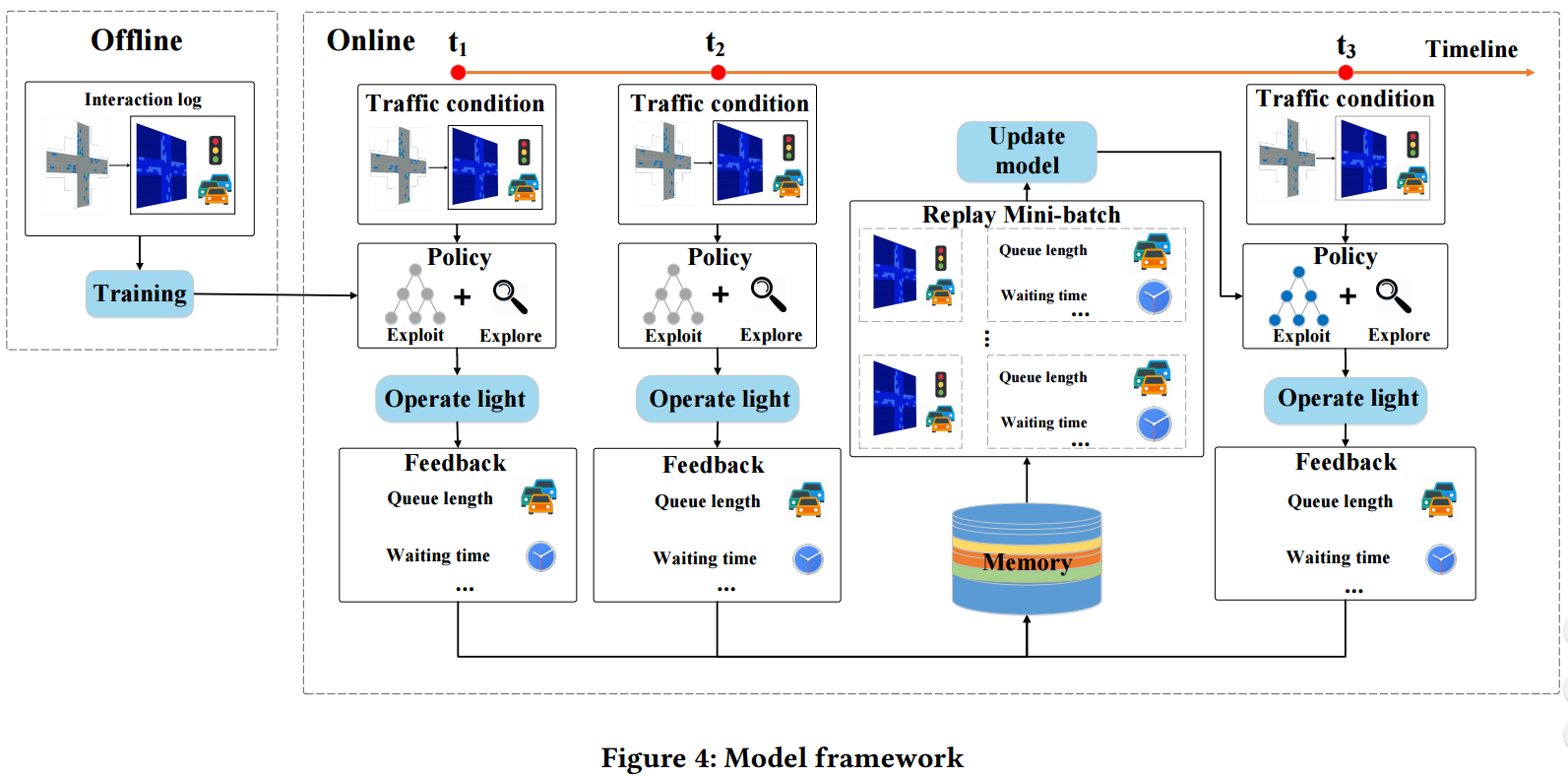
**4技术细节**

近年来交通灯控制由于其在调整交通中的重要作用而引起了很多关注。 当前的方法通常具有两类，常规方法和基于深度强化学习的方法。 传统方法通常依赖于先前的知识来为每个光相设置固定时间或设置改变规则。 这些规则易于动态改变流量。 强化学习方法通常将交通状况（例如，等待车辆的队列长度和更新的等待时间）作为状态，并尝试基于当前状态进行可以改善交通状况的动作。

然而，当前的方法并没有考虑实际情况中的复杂情况，因此可能导致陷入一种单一的行为。 这将导致在复杂交通情况下较差的交通调整性能。

在本节中，我们提出了一种深层加固交通灯代理来解决这个问题。 我们将在4.1节中首先介绍模型框架。 然后，我们在4.2节中展示了代理的设计。 我们在4.3节中进一步描述了网络结构。 另外，我们在4.4节描述了记忆宫殿。 请注意，尽管我们的模型设计用于具有两个阶段的四向交叉，但是将其扩展到其他类型的交叉点或多个阶段场景并不困难。

**4.1框架**



我们的模型由离线部分和在线部分组成（如图4所示）。我们提取了描述交通状况的五种特征作为状态（详见4.2节），并使用奖励来描述行动对交通的改善程度（详见4.2节）。在离线阶段，我们为灯光设置固定的时间表，让流量通过系统收集数据样本。在使用此阶段记录的样本进行培训后，该模型将被放入在线部分。在线阶段，在每个时间间隔Δt，交通信号灯将从环境中观察状态s并根据ε-贪婪策略结合探索采取行动a（即，是否将光信号改变到下一阶段）（即，具有概率ε）的随机动作和利用（即采取具有最大估计奖励的行动）。之后，代理G将观察环境并从中获得奖励。然后，元组（s，a，r）将被存储到存储器中。在几个时间戳（例如，图4中的t2）之后，代理G将根据存储器中的日志更新网络。

**4.2代理设计**

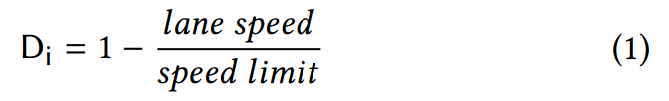
首先，我们介绍状态，行动和奖励表示。

**•状态：** 我们的状态是为一个交叉点定义的。 对于该交叉口处的每个车道i，状态组件包括队列长度Li，车辆数量Vi，车辆Wi的更新等待时间。 另外，该状态包括车辆位置M，当前阶段Pc和下一阶段Pn的图像表示。

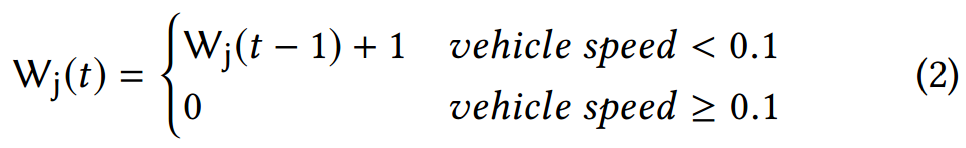
**•动作：**动作定义为a = 1：将交通灯改变为下一状态Pn，a = 0：保持当前状态Pc。

**•奖励：** 如等式3所示，奖励被定义为以下因素的加权和：

1. 在所有接近车道上的队列长度L的总和，其中L被计算为给定车道上的等待车辆的总数。速度小于0.1米/秒的车辆被认为是等待的。
2. 在所有接近车道上的延迟D的总和，其中车道i的延迟Di在等式1中定义，其中车道速度是车道i上的车辆的平均速度，并且速度限制是车道上允许的最大速度。车道 i：



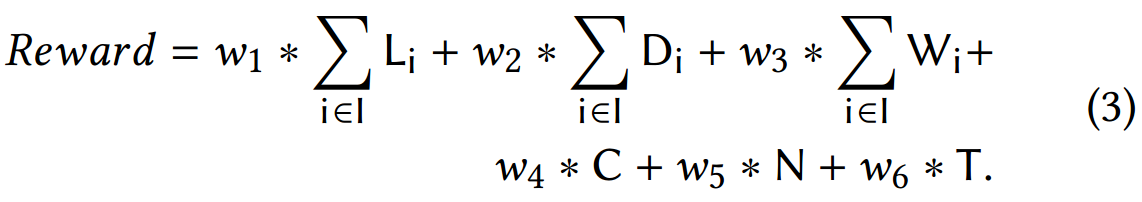
1. 所有临近车道的更新等待时间W的总和。 这相当于接近车道上所有车辆的总和。 等式2中车辆j的更新等待时间W在等式2中定义。注意，每当车辆移动时，车辆的更新等待时间被重置为0。 例如，如果车辆的速度从0s到15s为0.01m / s，从15s到30s为5m / s，从30s到60s为0.01m / s，当t = 15s时，Wj为15秒，0秒和30秒， 30和60年代相对。



（4）灯开关C的指示灯，其中C = 0用于保持电流相位，C = 1用于改变电流相位。

（5）在最后动作a之后的时间间隔Δt期间通过交叉路口的车辆总数N.

（6）在最后一次动作a之后的时间间隔Δt期间通过交叉路口的车辆T的总行驶时间，定义为车辆在接近车道上花费的总时间（以分钟为单位）。



因此，给定的交通状况的当前状态s，则代理的任务G是找到行动（改变或保持电流相位），其可以导致在长期运行的最大奖励R，以下的Bellman方程（方程 4）[21]。 在这种情况下，时间t的动作值函数q是下一个时间戳t + 1的奖励和最大潜在的未来奖励的总和。 通过这个未来的猜想，代理人可以选择更适合长期奖励的行动。



**4.3网络结构**

为了根据状态和行动来估计奖励，代理需要学习深度Q网络Q（s，a）。

在现实世界中，流量非常复杂，需要单独考虑许多不同的情况。 我们将在Example 4.1中说明这一点。

例4.1。我们仍然假设一个简单的交叉点，这里有两个光转换：1）Green-WE，和2）Red-WE。是否改变交通信号灯的决策过程包括两个步骤。第一步是从交通状况（例如，等待多少辆车，每辆车等待多长时间）到部分奖励的映射。该映射的示例可以是r = -0.5×L-0.7×W。无论绿灯亮在哪条车道上，这都由不同的阶段共享。然后，为了确定动作，代理应该在不同阶段观察不同车道的交通。例如，如图3（a）所示，当红灯处于NS方向时，NS方向上的更多等待流量（即，第一步中的较低奖励）将使光趋于改变（因为通过将该车道上的灯光从红色变为绿色，此车道上的更多车辆可以通过该交叉路口），而WE方向上的更多等待交通（即，第一步中的较低奖励）将使得灯光趋于保持。当红灯位于WE方向时，表壳正好相反。因此，光照阶段应该对特征进行明确选择。

在先前的研究中，由于在复杂交通条件下近似Q函数的模型的简化设计，代理在区分不同阶段的决策过程方面存在困难。 因此，我们在此提出一种可以明确地明确考虑不同情况的网络结构。 我们将这种特殊的子结构称为“相位门”。

我们的整个网络结构如图5所示。图像特征是从交通状况的观察中提取出来的，并被送入两个卷积层。 这些层的输出与四个明确挖掘的特征，队列长度L，更新的等待时间W，阶段P和总车辆数V连接。然后将连接的特征馈送到完全连接的层以学习来自交通状况的映射 潜在的奖励。 然后，对于每个阶段，我们设计一个单独的学习过程，从奖励到制定决策Q（s，a）的价值。 通过由相控制的门选择这些单独的过程。 如图5所示，当阶段P = 0时，左分支将被激活，而当阶段P = 1时，右分支将被激活。 这将区分不同阶段的决策过程，防止决策偏向某些行为，并增强网络的适应能力。

**4.4记忆宫和模型更新**

代理将定期从内存中获取样本并使用它们来更新网络。 通过在偶尔添加新数据样本并删除旧样本来维护该存储器。 这种技术被称为经验重播[19]，并已广泛应用于强化学习模型。

但是，在实际的流量设置中，不同车道上的流量可能会非常不平衡。 由于先前的方法[9,10,15,22]将所有状态 - 动作 - 奖励训练样本存储在一个存储器中，因此该存储器将由在不平衡设置中最频繁出现的阶段和动作支配。 然后，将学习代理以很好地估计这些频繁的相位 - 动作组合的奖励，但忽略其他不太频繁的相位 - 动作组合。 这将导致学习的代理人对不频繁的阶段 - 动作组合做出错误的决定。 因此，当不同车道上的交通显着不同时，这些不平衡的样本将导致在较不频繁的情况下表现较差。

受记忆宫理论[11,14]在认知心理学中的启发，我们可以通过使用不同的记忆宫殿来解决这种不平衡，以实现不同的相位 - 动作组合。 如图6所示，针对不同相位 - 动作组合的训练样本被存储到不同的存储器宫殿中。 然后将从不同的宫殿中选择相同数量的样本。 这些平衡样本将防止不同的相位动作组合干扰彼此的训练过程，从而提高网络的拟合能力以准确地预测奖励。

**5实验**

在本节中，我们使用合成和实际流量数据进行实验。 我们通过与其他方法进行比较来展示全面的定量评估，并展示了一些有趣的案例研究。

**5.1实验设置**

实验在模拟平台SUMO（Simulation of Urban MObility）2上进行。 SUMO为道路网络设计，交通量模拟和交通灯控制提供灵活的API。具体而言，SUMO可以控制交通管理，根据交通管理系统（由交通管理员获得）进行交通管理。

合成数据实验的环境是一个四向交叉点，如图2所示。交叉口与四条150米长的路段相连，每条路有三条入口和三条出口通道。 这部分实验中的交通灯包含两个阶段：（1）绿色WE（WE上的绿灯，SN上的红灯），（2）红色WE（WE上的红灯，SN上的绿灯）。 请注意，当绿灯在一个方向上时，另一个方向上有红灯。 此外，绿灯之后是3秒黄灯，然后变为红灯。 尽管这是现实世界场景的简化，但是可以以类似的方式进一步进行更多类型的交叉点（例如，三向交叉点）和更复杂的光相位（例如，具有左转相位）的研究。

**5.2参数设置**

我们方法的参数设置和奖励系数分别如表2和表3所示。 我们发现，只要Δt在5秒到25秒之间，动作时间间隔Δt对我们模型的性能影响最小。

**5.3评估指标**

我们使用以下措施评估不同方法的性能：

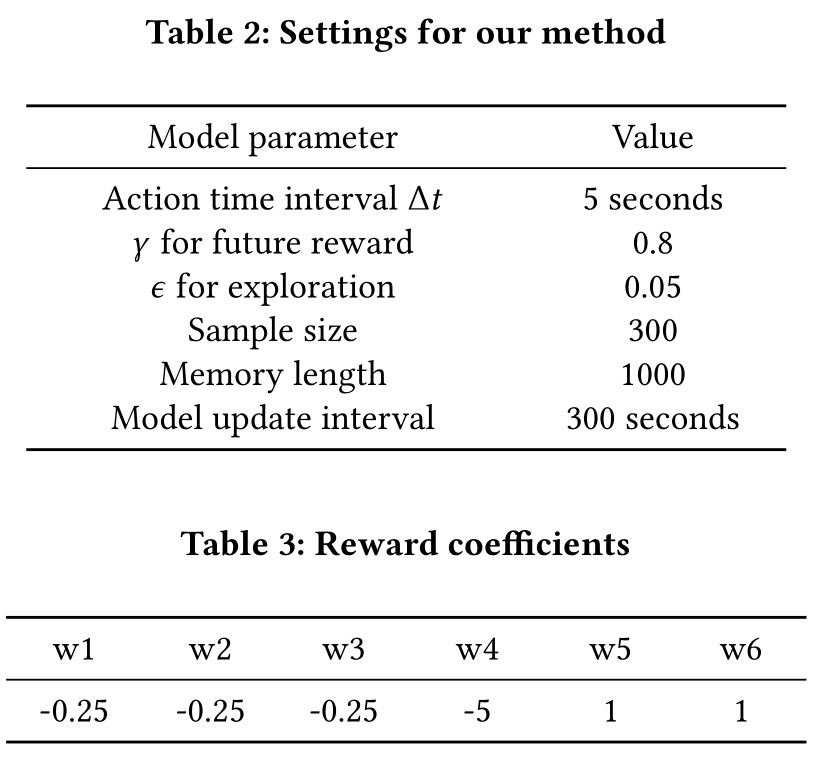
•奖励：平均奖励。在等式3中定义，奖励是几个术语（正面和负面术语）的组合，因此，奖励的范围是从-∞到∞。在特定配置下，当所有车辆自由移动而没有任何停止或延迟时，将有奖励的上限。

•队列长度：平均队列长度随时间的变化，其中时间t的队列长度是所有接近通道上的L（在4.2节中定义）的总和。较小的队列长度意味着所有车道上的等待车辆较少。

•延迟：随时间的平均延迟，其中时间t的延迟是所有接近通道的D（在等式1中定义）的总和。较低的延迟意味着所有通道的速度更高。

•持续时间：车辆在接近车道上所花费的平均行程时间（以秒为单位）。它是最重要的衡量标准之一，也是最重要的衡量标准。较小的持续时间意味着车辆通过交叉路口的时间较少。

总之，更高的奖励表示该方法的更好性能，并且较小的队列长度，延迟和持续时间表示流量较少被阻塞。



**5.4比较方法**

为了评估模型的有效性，我们将模型与以下基线方法进行比较，并调整所有方法的参数。 然后我们报告他们的最佳表现。

•固定时间控制（FT）。 固定时间控制方法使用预先确定的周期和相位时间计划[18]，并在交通流量稳定时广泛使用。

•自组织交通灯控制（SOTL）[5]。 该方法根据当前的交通状态控制交通灯，包括重叠时间和在红灯处等待的车辆数量。 具体地说，当等待车辆的数量高于手动调节的阈值时，交通信号灯将改变。

•交通灯控制（DRL）的深度强化学习。 在[22]中提出，该方法应用DQN框架来为交通路口选择最佳的灯光配置。 具体而言，它仅依赖于原始交通信息作为图像。

除基线方法外，我们还考虑了模型的几种变体。

•IntelliLight（基础）。 在第4.2节和第4.3节中使用相同的网络结构和奖励功能。这种方法没有记忆宫和相门。

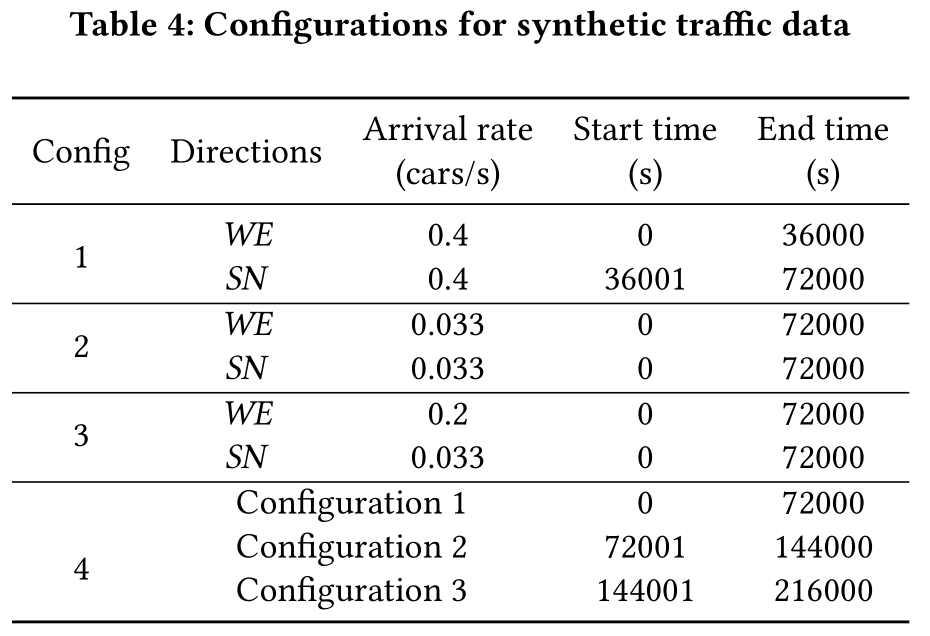
•IntelliLight（Base + MP）。 通过将心理学中的Memory Palace添加到IntelliLight - Base，我们将来自不同阶段和时间的样本存储在不同的记忆中。

•IntelliLight（Base + MP + PG）。 这是添加两种技术（Memory Palace和Phase Gate）的模型。

**5.5数据集**

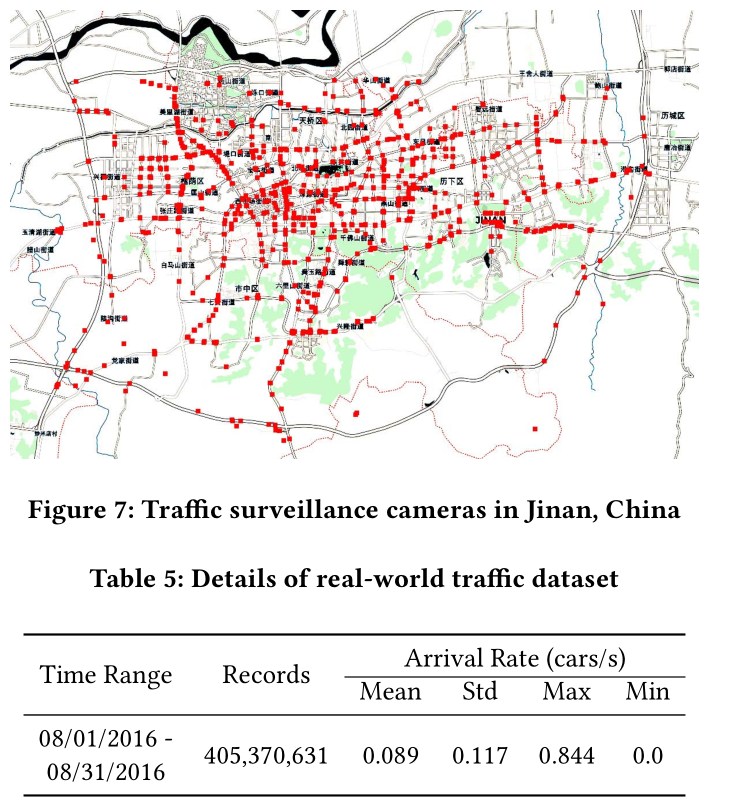
**5.5.1合成数据。**

在我们实验的第一部分中，合成数据与四种流量设置一起使用：简单的流量变化（配置1），同样稳定的流量（配置2），不稳定的流量（配置3）和复杂的流量（配置4） 前三种配置的组合。 如表4所示，车辆的到达是通过具有一定到达率的泊松分布产生的。



**5.5.2真实世界的数据。**

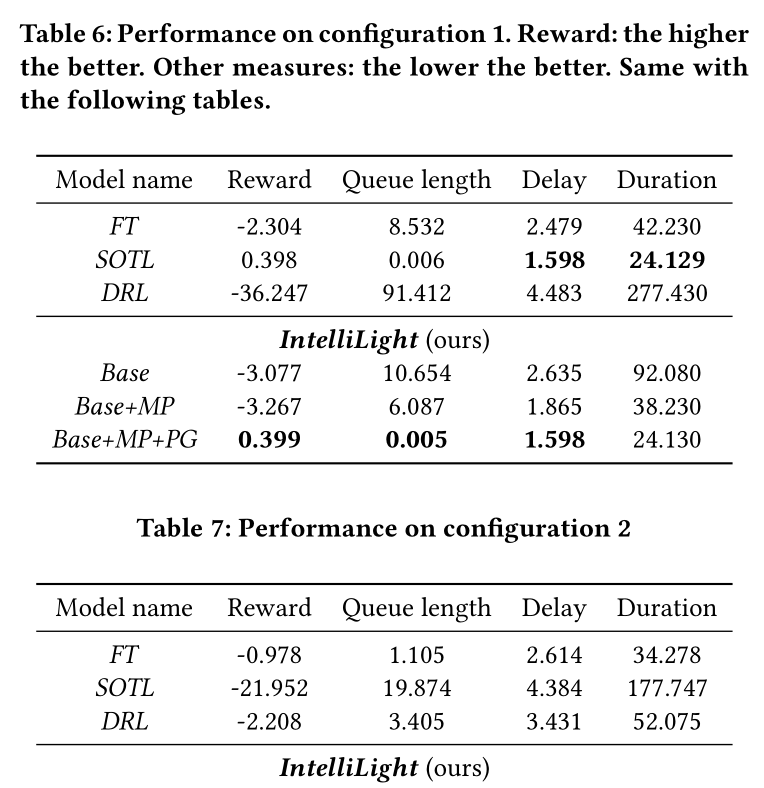
实际数据集由2016年1月8日至2016年8月31日期间的中国济南的1,704surveillancecameras收集。 这些摄像机的位置如图7所示。摄像机面向交叉口附近的车辆每秒聚集一次，数据集中的每条记录包括时间，摄像机ID和车辆信息。 通过用摄像机位置分析这些记录，车辆的轨迹在通过道路交叉点时被记录。 数据集涵盖935个位置，其中43个是四向交叉点。 我们使用通过24个交叉口的车辆数量作为实验的交通量，因为只有这些交叉口有连续的数据。 然后我们将这个真实世界的交通设置作为在线实验提供给SUMO。 从表5可以看出，不同道路上的交通流量在现实世界中是动态变化的。

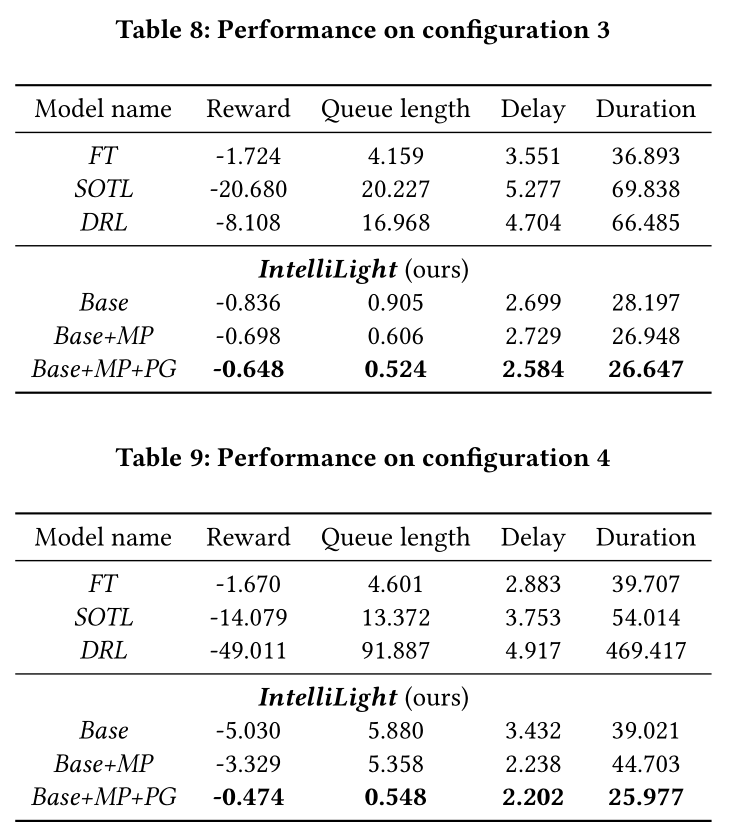


**5.6合成数据的表现**

**5.6.1与最先进的方法进行比较**

我们首先将不同的方法与不同的基线进行比较，然后根据不同的合成流量设置进行比较。 从表6,7,8和9中我们可以看到我们的方法比配置1,2,3和4中的所有其他基线方法表现更好。尽管某些基线在某些设置上表现良好，但它们只能在其他配置中实现（例如，SOTL在配置下实现了良好的效果1，几乎与其他方法一样） 这是因为我们的方法已经学会了将光保持到36000秒并在此之后切换光，并且SOTL的设计也表现得相似。因此，这两种方法的表现非常相似）。 相反，我们的方法IntelliLight在不同配置下表现出更好的性能。



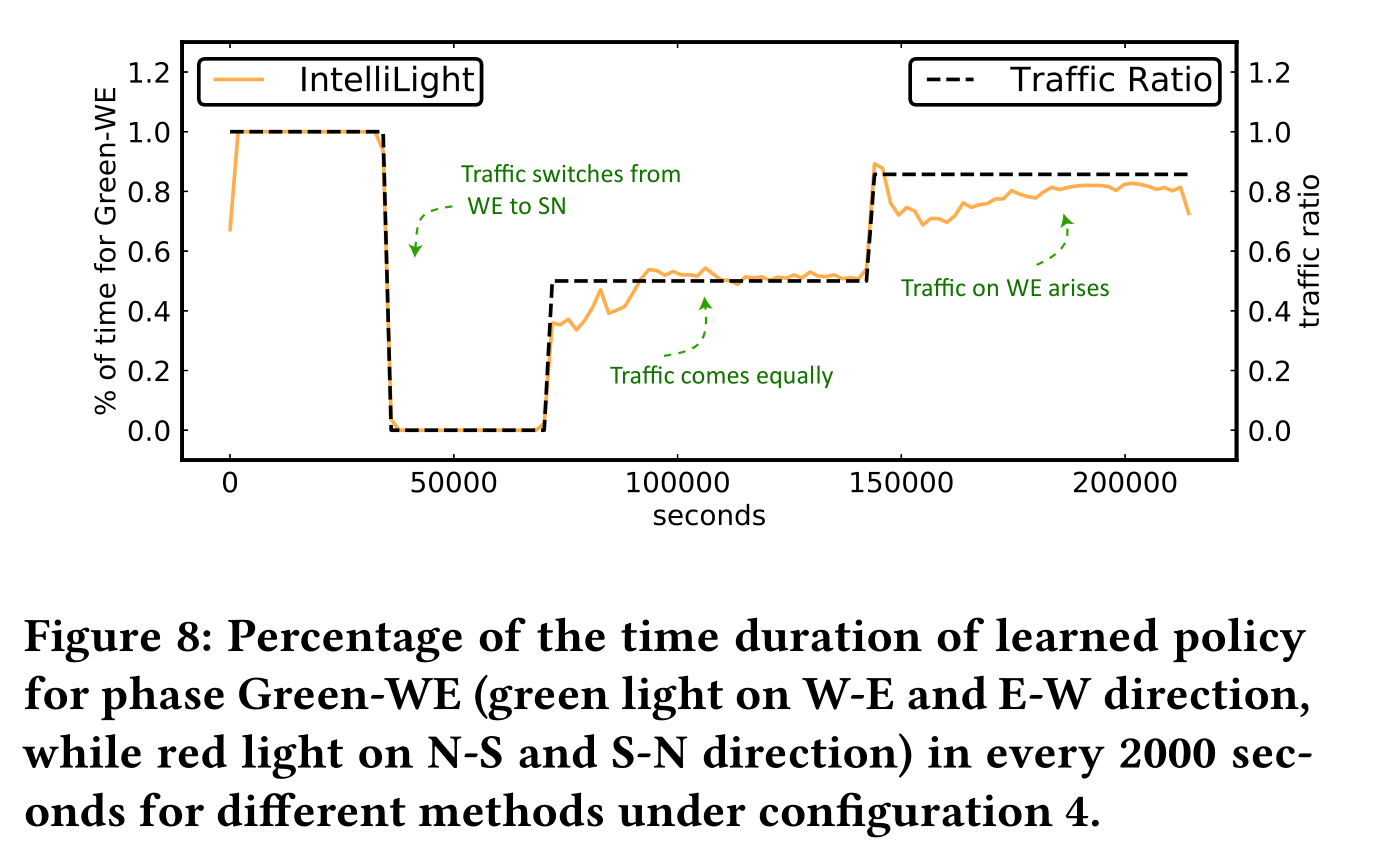


**5.6.2与我们提出的方法的变体进行比较**

  表6,7,8和9显示了我们提出的方法的变体的性能。 首先，我们可以看到添加Memory Palace有助于在配置3和4下获得更高的奖励，虽然它不会增加配置1和2下的奖励。这是因为对于简单（配置1和2），相位相对稳定很长一段时间（因为流量） 只来自一个方向或长时间不变化）。 因此，记忆宫殿无助于建立更好的模型来预测奖励。 在大多数情况下，进一步添加Phase Gate还可以缩短队列长度并获得最高奖励，从而证明了这两种技术的有效性。

**5.6.3学习信号的解释**

了解我们的方法学到了什么。 在动态交通条件下，我们显示了相位Green-WE的持续时间百分比（即WE方向上的绿灯，SN方向上的红灯），以及WE上的交通流量与来自所有方向的总交通流量的比率。 随着交通的变化，理想的交通灯控制方法将能够调整其相对于交通流量的持续时间并获得高回报。 例如，当交通从WE方向改变为SN时，交通灯代理预计将调整其相位持续时间，从给予WE绿灯到给予SN绿灯。 从图8中可以看出，IntelliLight可以随着流量的变化调整其相位持续时间。



**5.7现实世界数据的表现**

**5.7.1不同方法的比较**

在本节中，我们将我们的方法与实际数据的基线方法进行比较。 总体结果如表10所示。我们的方法IntelliLight在所有比较方法中获得最佳奖励，队列长度，延迟和持续时间，相对于最佳基线相对改善了32％，38％，19％和22％ 方法。 此外，我们的方法在多个交叉点（小标准偏差）具有相对稳定的性能。

**5.7.2关于实际交通的观察**

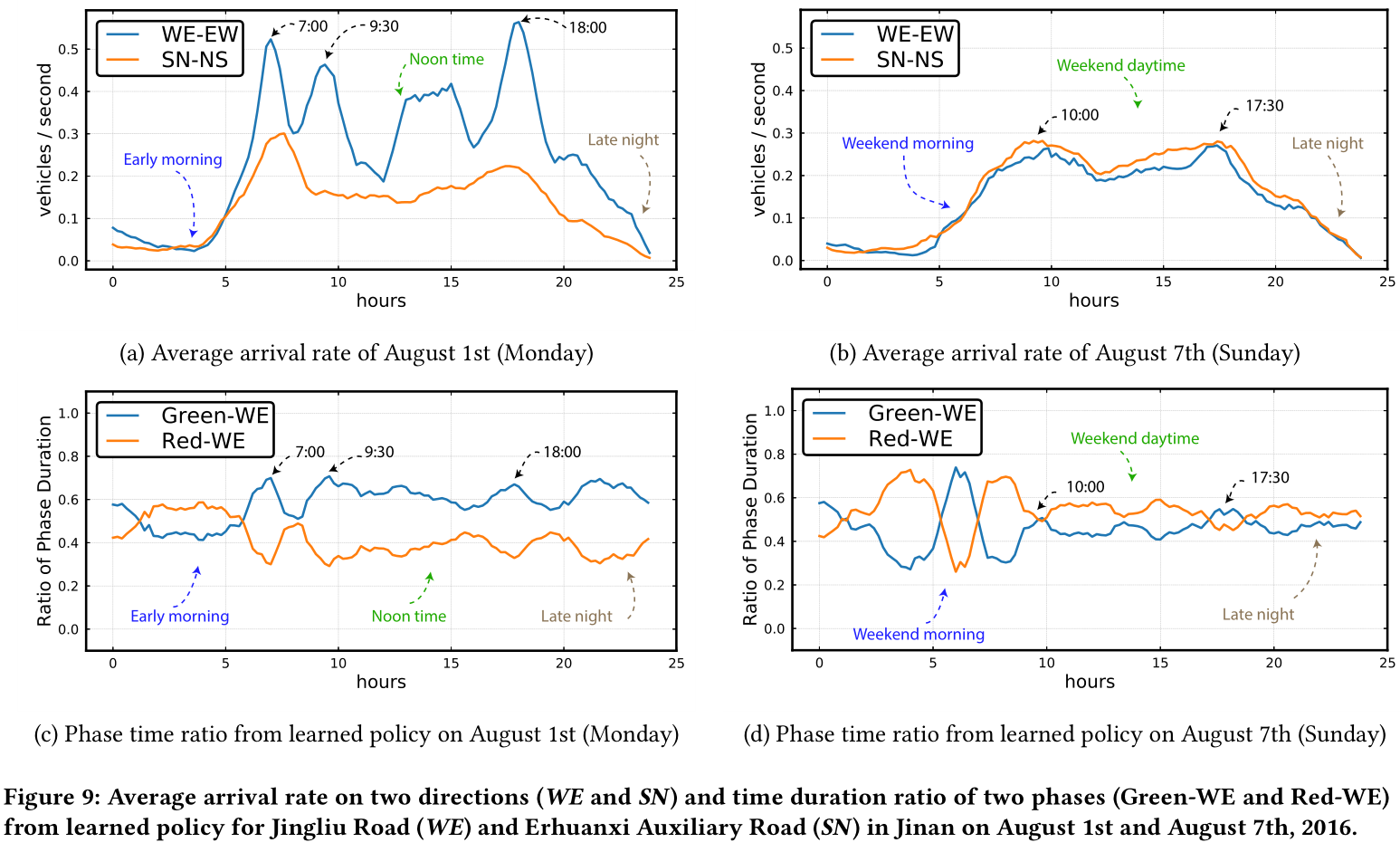
在本节中，我们将对从实际数据中学到的策略进行观察。 我们分析了在不同情况下京流路（WE方向）和二环西辅助路（SN方向）交叉口的交通信号灯政策：高峰时段与非高峰时段，工作日与周末，以及主要动脉与小动脉。

**1.高峰时段与非高峰时段**

图9（a）显示了星期一两个方向（WE和SN）的平均交通流量。 在这一天，在大多数情况下，WE方向上的流量比SN大，在此期间理想的交通灯控制方法预计会给WE方向提供更长的时间。 从图9（c）可以看出，相位Green-WE的持续时间的比率（即WE上的绿灯，而SN上的红灯）通常大于0.5，这意味着时间的变化，我们的方法长期为 我们 在高峰时段（大约7：00,9：30和18:00），从我们的方法中学到的政策也使我们在WE上的绿灯时间比非高峰时段更长。 在清晨，SN上的车辆到达率大于WE上的率，我们的方法自动给出更长的SN时间。 这表明我们的方法可以智能地适应不同的交通状况。

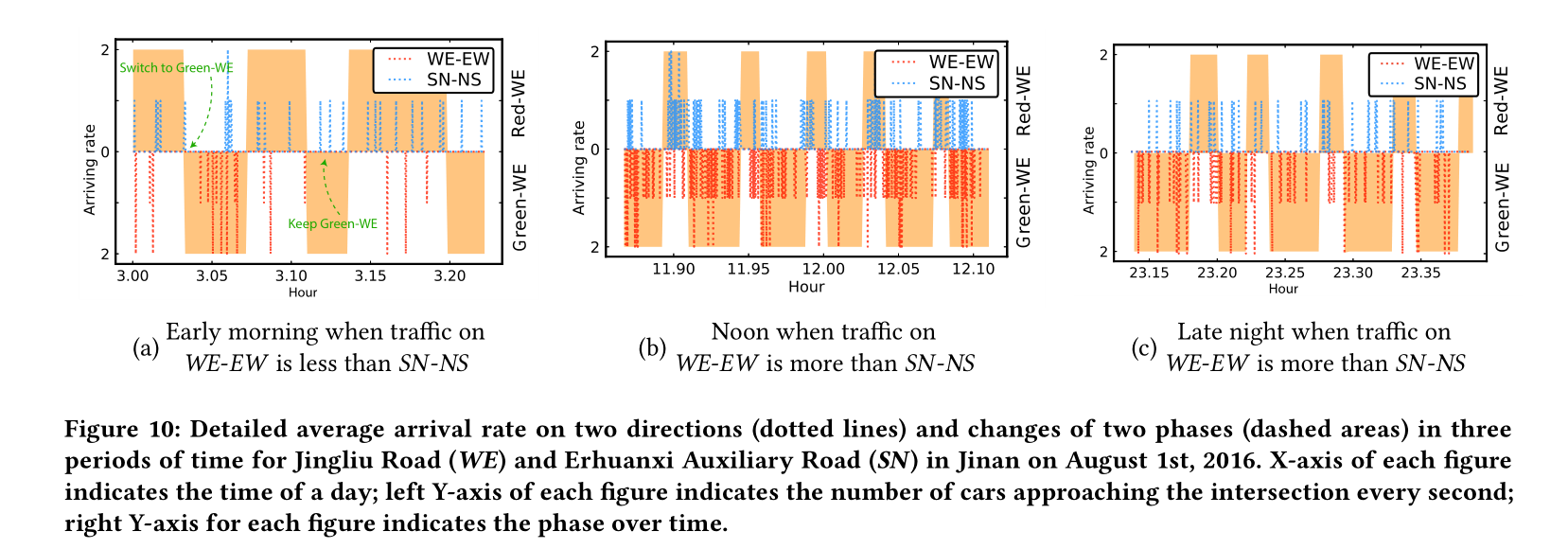
**2.平日与周末**

与工作日不同，周末显示有关交通状况和交通灯控制政策的不同模式。 我们的政策在周末白天给予较少的绿灯（在SN上更多的绿灯）比在工作日给出的绿灯更少。 这是因为在图9（b）中，在周末白天，SN上的流量比WE上的流量多，而在工作日，SN上的流量小于WE。 此外，通过比较图9（a）和图9（b），我们可以看到周一深夜时间WE和SN的流量相似，使得持续时间Green-We的比率接近0.5。



**3.主要动脉与轻微动脉**

主要干道是一段时间内交通量较大的道路，预计绿灯时间较长。在没有关于主干道的先验知识的情况下，使用我们的方法学习的交通灯控制策略更喜欢给予主要的动脉绿灯（包括保持绿灯的主动脉，并且为了主动脉保持绿灯的绿灯）。详细的，我们看看8月1日的三个时间段（3：00,12：00和23:30）。从图9（a）可以看出，WE方向的道路是主要道路，因为WE上的交通量通常比SN上的交通量大。如图10所示，虚线表示在两个不同方向上每秒到达的车辆的数量。除了到达率，我们还绘制了阶段的变化（虚线区域）。从图10（a）可以看出：1）Red-WE阶段的总时间段比Green-WE长，这与此时的交通量兼容。 2）虽然SN的交通量大于WE，但是从绿色WE到红色WE的交通信号灯通常不会被SN方向上的等待车辆触发。相反，在图10（b）和图10（c）中，从Green-WE到Red-WE的变化通常是由SN方向上的等待车辆触发的。这主要是因为在这段时间内，WE的道路是主要道路，而交通信号灯往往偏向Green-WE阶段。

**六，结论**

在本文中，我们使用精心设计的强化学习方法解决了交通灯控制问题。我们使用合成和现实世界的实验进行了广泛的实验，并证明了我们提出的方法优于现有技术方法的优越性能。此外，我们还展示了深入的案例研究和观察，以了解代理如何适应不断变化的流量，作为对奖励的量化衡量的补充。这些深入的案例研究有助于为现实世界的应用程序生成流量规则。

我们也承认我们当前方法的局限性，并且想指出几个重要的未来方向，以便更适用于更适用的世界。首先，我们将两相交通信号灯转换为多相交通灯，这将涉及更复杂但更现实的状态转换。其次，我们的论文讨论了简化的一个交叉案例，而现实世界的道路网络要比这复杂得多。尽管一些研究试图通过使用多个强化学习因子来解决多交叉问题，但是他们没有明确考虑不同交叉点之间的相互作用（即，一个交叉点的相位如何影响附近交叉点的状态）并且它们仍然是仅限于少量的十字路口。最后，我们的方法仍在模拟框架上进行测试，因此模拟了反馈。最终，应进行实地研究以了解现实世界的反馈并验证拟议的强化学习方法。