2019 年数学建模华为题
无线智能传播模型

1 无线信道建模背景

随着 5G NR 技术的发展，5G 在全球范围内的应用也在不断地扩大。运营商在部署 5G 网络的过程中，需要合理地选择覆盖区域内的基站站址，进而通过部署基站来满足用户的通信需求。在整个无线网络规划流程中，高效的网络估算对于精确的 5G 网络部署有着非常重要的意义。无线传播模型正是通过对目标通信覆盖区域内的无线电波传播特性进行预测，使得小区覆盖范围、小区间网络干扰以及通信速率等指标的估算成为可能。由于无线电波传播环境复杂，会受到传播路径上各种因素的影响，如平原、山体、建筑物、湖泊、海洋、森林、大气、地球自身曲率等，使电磁波不再以单一的方式和路径传播而产生复杂的透射、绕射、散射、反射、折射等，所以建立一个准确的模型是一项非常艰巨的任务。

现有的无线传播模型可以按照研究方法进行区分，一般分为：经验模型、理论模型和改进型经验模型。经验模型的获得是从经验数据中获取固定的拟合公式，典型的模型有 Cost 231-Hata、Okumura 等。理论模型是根据电磁波传播理论，考虑电磁波在空间中的反射、绕射、折射等来进行损耗计算，比较有代表性的是 Volcano 模型。改进型经验模型是通过在拟合公式中引入更多的参数从而可以为更细的分类场景提供计算模型，典型的有 Standard Propagation Model（SPM）。

在实际传播模型建模中，为了获得符合目标地区实际环境的传播模型，需要收集大量额外的实测数据、工程参数以及电子地图用来对传播模型进行校正。此外无线 LTE 网络已在全球普及，全球几十亿用户，每时每刻都会产生大量数据。如何合理地运用这些数据来辅助无线网络建设就成为了一个重要的课题。

近年来，大数据驱动的 AI 机器学习技术获得了长足的进步，并且在语言、图像处理领域获得了非常成功的运用。伴随着并行计算架构的发展，机器学习技术也具备了在线运算的能力，其高实时性以及低复杂度使得其与无线通信的紧密结合成为了可能。

在本届数学建模竞赛中，希望参赛者能够对机器学习的工作方式有一定掌握并站在设备供应商以及无线运营者的角度，通过合理地运用机器学习模型（不限定只使用这种方法）来建立无线传播模型，并利用模型准确预测在新环境下无线信号覆盖强度，从而大大减少网络建设成本，提高网络建设效率。

2 无线传播模型建模方法简介

在传统的无线传播模型的建立过程中，往往首先需要对传播场景进行划分，每一个场景对应一个传播经验模型。然而，经验模型在实际使用中往往不够精确，所以仍然需要通过采

集大量的工程参数以及实际平均信号接收功率（Reference Signal Receiving Power，RSRP）测量值进行经验模型公式的修正。从所述过程中可以看到，传播模型建立本质上是一个函数拟合的过程，即通过调整传播模型的系数，使得利用传播模型计算得到的路径损耗值与实测路径损耗值误差最小。所以当工程参数、地理位置信息、特定地理位置测量点的 RSRP 已知的情况下，该问题可以归类为一个监督学习问题。

与传统经验模型需要额外人力物力进行校正相比，是否可以利用采集的历史数据并利用机器学习技术，得到一套合适的机器学习模型用以对不同场景下信道传播路径损耗进行准确预测，成为一个非常有价值的研究方向。

本题为参赛队伍提供统一的数据集。各参赛队伍可以自行将数据集拆分为训练集、测试集以及验证集，将其用于 AI 算法模型的训练及测试。算法的目的在于通过寻找工程参数、地理环境等因素与平均信号接收功率（RSRP）之间的映射模型（理论与实践表明 RSRP 是工程参数、地理环境等因素的随机函数），从而能够在新的环境中快速预测特定地理位置的 RSRP 值。

赛题提供的训练数据集包含多个小区的**工程参数数据**、**地图数据**和 **RSRP 标签数据**，其格式为 csv 格式（Comma-Separated Values, 逗号分隔值格式）。数据集的结构以及对应数据的含义将会在下节中详细阐述。

3 训练数据集简介

训练数据集一共包括了多个文件，**每个文件代表一个小区内的数据**。文件的命名方式为 train_id.csv，其中 id 为小区的唯一标识，例如 train_1003501.csv 表示唯一标识为 1003501 的小区数据。

文件的每一行代表小区内固定大小的测试区域的相关数据，行数不定(根据小区大小不同，面积越大的小区行数越多，反之亦然)，列数则固定为 18 列，其中前 9 列为站点的**工程参数数据**；中间 8 列为**地图数据**；最后 1 列是用于训练的 **RSRP 标签数据。**下表显示了其中一行数据作为样例：

Table 1：训练数据样例

工程参数数据								
Cell Index	Cell X	Cell Y	Height	Azimuth	Electrical Downtilt	Mechanical Downtilt	Frequency Band	RS Power
1003501	393621.9	3394449	35	300	6	4	2585	13.2
地图数据								
Cell Altitude	Cell Building Height	Cell Clutter Index	X	Y	Altitude	Building Height	Clutter Index	
524	32	1	392800	3395210	524	0	5	
RSRP 标签数据								
RSRP								

-90.5								

下面介绍三部分中每一列的具体含义。

3.1 工程参数数据

工程参数数据记录了某小区内站点的工程参数信息，共有 9 个字段。各字段对应含义如 Table 所示。

Table 2：工程参数数据的字段含义

字段名称	含义	单位
Cell Index	小区唯一标识	-
Cell X	小区所属站点的栅格位置，X 坐标	-
Cell Y	小区所属站点的栅格位置，Y 坐标	-
Height	小区发射机相对地面的高度	m
Azimuth	小区发射机水平方向角	Deg
Electrical Downtilt	小区发射机垂直电下倾角	Deg
Mechanical Downtilt	小区发射机垂直机械下倾角	Deg
Frequency Band	小区发射机中心频率	MHz
RS Power	小区发射机发射功率	dBm

为了方便数据处理，地图进行了栅格化处理，每个栅格代表了 5m × 5m 的区域（如下图 Fig.1 所示），其中（Cell X，Cell Y）记录了站点所在栅格的左上角坐标。其他的工程参数(Height, Azimuth, Electrical Downtilt, Mechanical Downtilt)如图 Fig.2 所示，其中机械下倾角(Mechanical Downtilt)是通过调整天线面板后面的支架来实现的，是一种物理信号下倾；而电下倾角(Electrical Downtilt)是通过调整天线内部的线圈来实现的，是一种电信号下倾。实际的信号线下倾角是机械下倾角和电下倾角之和。

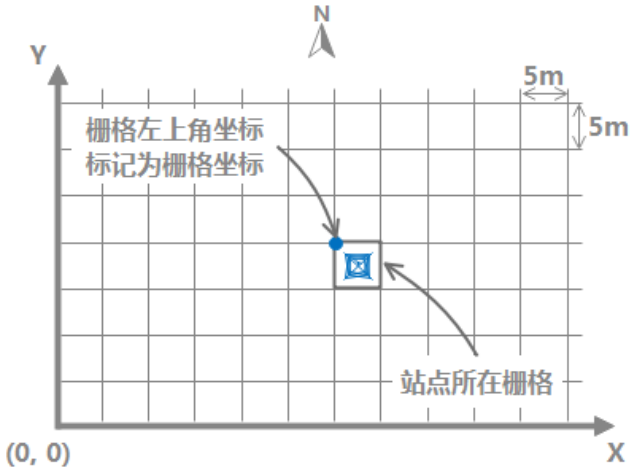


Fig. 1：栅格化地图的坐标说明

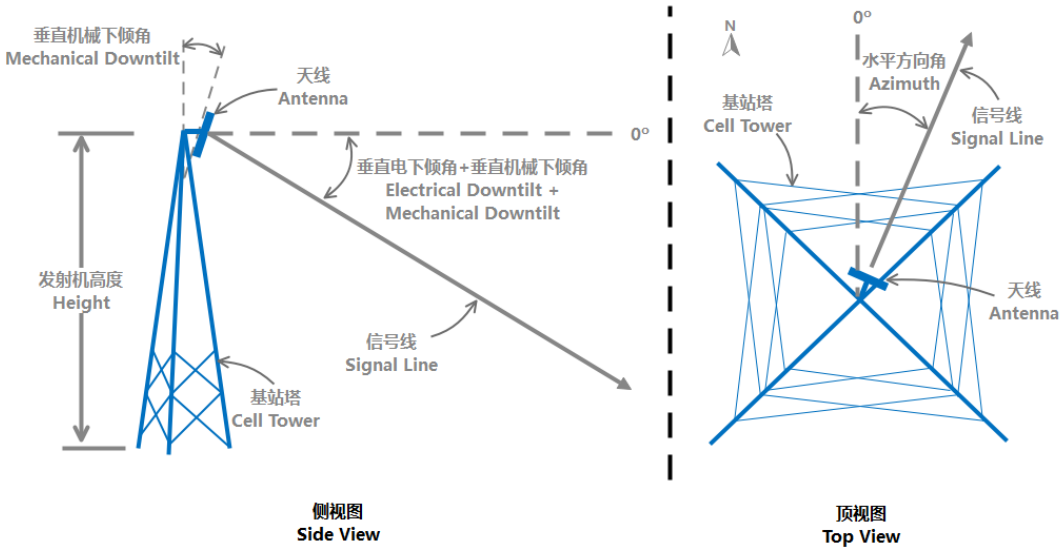


Fig. 2：工程参数数据含义说明

3.2 地图数据

地图数据记录地形地貌等信息，共有 8 个字段，各字段对应含义如 Table 所示。考虑地图类型的多样性和复杂性，城区、农村、湖泊等实际地物被抽象为数字，这些数字称为地物类型名称编号（Clutter Index），在

Table 中可以看到地物类型名称编号所对应的实际地物类型。

Table 3：地图数据的字段含义

字段名称	含义	单位
Cell Building Height	小区站点所在栅格(Cell X, Cell Y)的建筑物高度，若该栅格没有建筑物，则为 0	m
Cell Altitude	小区站点所在栅格(Cell X, Cell Y)的海拔高度	m
Cell Clutter Index	小区站点所在栅格(Cell X, Cell Y)的地物类型索引	-
X	栅格位置，X 坐标	-
Y	栅格位置，Y 坐标	-
Building Height	栅格(X,Y)上的建筑物高度，若该栅格没有建筑物，则为 0	m
Altitude	栅格(X,Y)上的海拔高度	m
Clutter Index	栅格(X,Y)上的地物类型索引	-

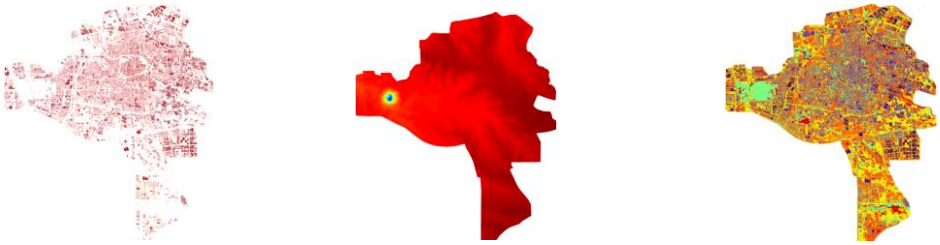
Table 4：地物类型名称的编号含义

Clutter Index	含义	Clutter Index	含义
1	海洋	**11**	城区高层建筑（40m~60m）
2	内陆湖泊	**12**	城区中高层建筑（20m~40m）
3	湿地	**13**	城区<20m 高密度建筑群

4	城郊开阔区域	**14**	城区<20m 多层建筑
5	市区开阔区域	**15**	低密度工业建筑区域
6	道路开阔区域	**16**	高密度工业建筑区域
7	植被区	**17**	城郊
8	灌木植被	**18**	发达城郊区域
9	森林植被	**19**	农村
10	城区超高层建筑（>60m）	**20**	CBD 商务圈

与工程参数数据一样，地图数据也进行了栅格化处理，每个栅格代表了 5m × 5m 的区域，其中（X，Y）记录了地图所在栅格的左上角坐标。

在明确了地图存储格式之后，可以针对不同的参数对地图进行可视化处理。如 Fig. 3 所示，Fig. 3a-c 分别根据栅格坐标以及房屋高度、海拔高度和地物类型索引作为特征对地图进行可视化处理。通过可视化处理，可以对地图数据有一个更为直观的了解。



(a) 建筑物高度 (b) 海拔高度 (c) 地物类型索引

Fig. 3：电子地图图像化示例

3.3 RSRP 标签数据

平均信号接收功率(RSRP)标签数据作为实际测量结果，在监督学习中用于和机器学习模型预测的结果作比较，共有 1 个字段，对应含义如 Table 所示。

Table 5：RSRP 标签数据表格的字段含义

字段名称	含义	单位
RSRP	栅格(X, Y)的平均信号接收功率，标签列	dBm

如 Fig. 4 所示，结合电子地图数据中的坐标和特征以及标签数据中的 RSRP 值，可以清晰地对信号功率分布进行可视化处理，从而明确辨识信号强弱覆盖区域。

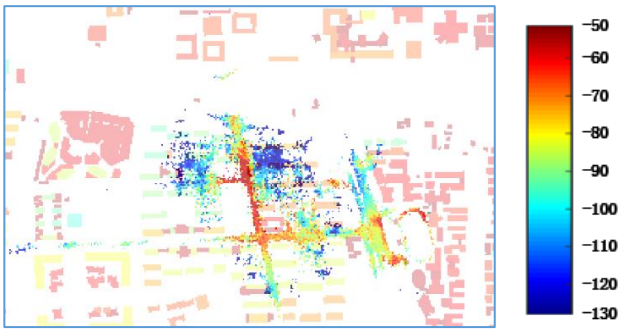


Fig. 4：标签数据的可视化处理

4 无线传播模型建模赛题

本赛题除在中国研究生数学建模竞赛网站上上交论文外，问题三需要在华为云平台上提交模型，不提交的队伍将被视为没有完成此题而不计入比赛成绩。

4.1 特征工程中的特征设计

高效的机器学习模型建立依赖于输入变量与问题目标的强相关性，因此输入变量也称为“特征”。特征工程的本质是从原始数据中转换得到能够最好表征目标问题的参数，并使得各个参数的动态范围在一个相对稳定的范围内，从而提高机器学习模型训练的效率。一般特征工程的典型技术有：

- 剔除失真、低质量数据；数据插值补齐；去除异常点；
- 连续数据离散化；数据去均值；幅度限制；方差限制。

高阶的特征工程需要充分利用与目标问题相关的专业知识。对于信道传播模型问题，可以如 Fig. 5 所示根据已知的几何位置来挑选合理的特征。例如，通过发射机相对地面的高度h_b、机械下倾角θ_{MD}、垂直电下倾角θ_{ED}，发射机所在栅格位置与目标栅格位置，可以得到栅格与发射机的距离d以及栅格与信号线的相对高度Δh_v，而Δh_v就可以作为一个特征。

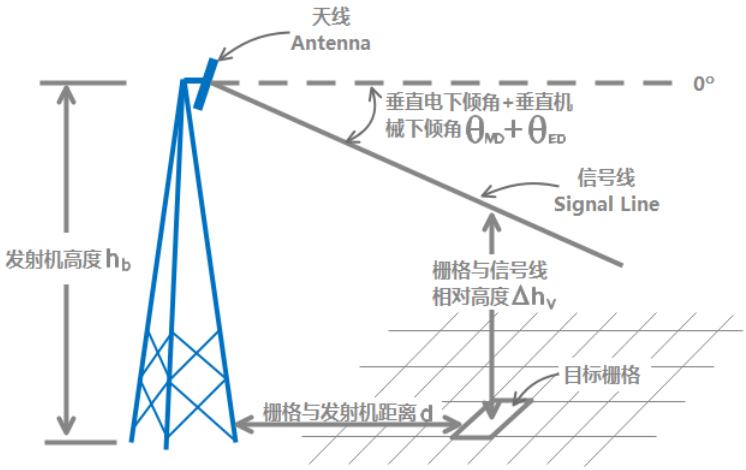


Fig. 5：根据目标栅格与发射机的地理位置关系提取特征

除了几何位置特征，传统经验信道模型中涉及的参数也可以纳入特征工程的考察范围。例如城市中的经典模型 Cost 231-Hata，其定义如下：

$$\mathrm{PL} = 46.3 + 33.9\log_{10} f - 13.82\log_{10} h_b - \alpha + (44.9 - 6.55\log_{10} h_{ue})\log_{10} d + C_m, \qquad (1)$$

其中 PL 定义为传播路径损耗（dB）、f为载波频率（MHz）、h_b为基站天线有效高度（m）、h_{ue}为用户天线有效高度（m）、α为用户天线高度纠正项(dB)、d 为链路距离（km）以及C_m为场景纠正常数（dB）。RSRP 与 PL 的关系为：

$$\mathrm{RSRP} = \mathrm{P_t} - \mathrm{PL}, \tag{2}$$

其中$\mathrm{P_t}$是小区发射机发射功率（dBm）(见 Table 2)。

问题一

请根据 Cost 231-Hata 模型以及下述数据集信息设计合适的特征，并阐述原因。

Table 6：数据集信息

工程参数数据								
Cell Index	Cell X	Cell Y	Height	Azimuth	Electrical Downtilt	Mechanical Downtilt	Frequency Band	RS Power
2	100	100	49m	45°	2°	2°	1800MHz	18.2 dBm
地图数据								
Cell Altitude	Cell Building Height	Cell Clutter Index	X	Y	Altitude	Building Height	Clutter Index	
47m	9m	11	500	500	9m	0m	1	
RSRP 标签数据								
RSRP								
-100 dBm								

4.2 特征工程中的特征选择

完成特征设计后，通常需要选择有意义的特征输入机器学习模型进行训练。对于不同方法构造出来的特征，需要从多个层面来判断这个特征是否合适。通常来说，可以从以下两个方面来选择特征：

- 特征是否发散：如果一个特征不发散，例如方差接近于 0，也就是说样本在这个特征上基本上没有差异，这个特征对于样本的区分并没有什么用。
- 特征与目标的相关性：这点比较显见，与目标相关性高的特征，应当优先选择。

问题二

基于提供的各小区数据集，设计多个合适的特征，计算这些特征与目标的相关性，并将结果量化、排序，形成如下的表格，并阐明设计这些特征的原因和用于排序的量化数值的计算方法。

Table 7：特征名称及其与目标的相关性

排序	特征名称	该特征与目标的相关性
1		
2		
…		

4.3 RSRP 预测

问题三

在设计和选择了有效的特征之后，就可以通过建立预测模型来进行 RSRP 的预测了。请各个参赛队根据自己建立的特征集以及赛题提供的训练数据集，建立基于 AI 的无线传播模型来对不同地理位置的 RSRP 进行预测。为研究生更明白本问题的目标，下面将分别介绍评审数据集、提交内容和线上代码评分方法。

4.3.1 评审数据集简介

线上代码评分系统将使用对参赛队保密的评审数据集来对模型进行评分，以便公平地测试各参赛队提交模型的实际泛化能力。评审数据集与训练数据集一样，一共包括了多个文件，**每个文件代表一个小区内的数据**。文件的命名方式为 test_id.csv，其中 id 为小区的唯一标识，例如 test_1003501.csv 表示唯一标识为 1003501 的小区数据。

评审数据集的文件中含有除了 RSRP 之外的前 17 个字段，与该 17 个字段对应的 RSRP 字段需要由研究生提交的模型代码程序预测生成。

4.3.2 提交内容

论文要以文字形式详细阐述 AI 模型的建模过程，包括模型的建立方法，参数的设置和训练的结果，特别是第三问要阐述清楚。

第三问需要提交完整的模型。针对每一个评审数据集的输入文件，模型输出要求也是一个文件，例如输入数据文件名为 test_123456.csv，则输出文件名必须为 test_123456.csv_result.txt。另外，输出文件的数量与输入文件必须一致，否则会以全 0 文件代替输出文件进行评分。例如，参赛队伍如果没有提交针对输入文件名为 test_123456.csv 的输出文件，系统在评分时会自动产生全零的 test_123456.csv_result.txt 进行评分。

每个输出文件内容的样例如下所示，

{"RSRP": [[-54.505], [-73.416], [-76.123], [-74.261], [-98.143]]}

其中方括号内的数字表示输入文件的每一行数据所对应的 RSRP 预测值，预测值的数量与输入文件的行数(表头除外)对应，例如上文的输出文件对应的输入文件应该是 5 行(表头除外)。如果输出文件的预测值少于输入文件的行数，则会以补 0 的形式将输出文件填满后进行评分；如果输出文件的预测值多余输入文件的行数，则会取输出文件的前 N 个预测值进行评分，其中 N 为输入文件的行数。

4.3.3 线上代码评分方法

对于提交的预测 RSRP 值，将根据以下条件进行排序。

- 模型在评审数据集的评估下，**弱覆盖识别率** (PCRR：Poor coverage recognition rate) 必须大于等于 20%。

- 在 PCRR 精度达标后，再根据预测**均方根误差** (RMSE : Root mean squared error) 大小进行各参赛组的名次排序（RMSE 小者排名靠前）。

PCRR 和 RMSE 的介绍如下所示：

- **弱覆盖识别率 (PCRR : Poor coverage recognition rate)**

在进行预测的过程中如果可以有效识别弱覆盖区域，能够更好地帮助运营商精准规划和优化网络从而提升客户体验。因此，除 RMSE 为有效测试目标之外，弱覆盖识别准确率也是作为一项非常有价值的评价指标。

在本次建模比赛中，弱覆盖判决门限P_{th}的值定为-103 dBm。若 RSRP 预测值或实测值小于P_{th}则为弱覆盖并标记为 1，若大于等于P_{th}则为非弱覆盖并标记为 0。根据比较预测值和实测值得到的弱覆盖以及非弱覆盖的差别，可以对以下参数进行统计：

- True Positive（TP）： 真实值为弱覆盖，预测值也为弱覆盖；
- False Positive（FP）： 真实值为非弱覆盖，预测值为弱覆盖；
- False Negative（FN）：真实值为弱覆盖，预测值为非弱覆盖；
- True Negative（TN）： 真实值为非弱覆盖，预测值也为非弱覆盖。

Table 8：TP、FP、FN 和 TN 的定义

		真实结果	
		True（弱覆盖）	False（非弱覆盖）
预测结果	True（弱覆盖）	TP	FP
	False（非弱覆盖）	FN	TN

PCRR 综合考虑 Precision（准确率）和 Recall（召回率）的目标，其计算公式如下：

$$\mathrm{PCRR} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{3}$$

其中 Precision 可以理解为**预测结果为弱覆盖的栅格实际也是弱覆盖的概率**，其定义如下：

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

Recall可以理解为**真实结果为弱覆盖的栅格有多少被预测成了弱覆盖的概率**，其定义如下：

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

PCRR的计算代码可以参考以下程序

Table 9：PCRR 计算方法参考

```
def CaculatePcrr(y_true,y_pred):
    t = -103
    tp = len(y_true[(y_true < t)&(y_pred < t)])
    fp = len(y_true[(y_true >= t)&(y_pred < t)])
    fn = len(y_true[(y_true < t) & (y_pred >= t)])
```

```
    precision =tp/(tp+fp)
    recall = tp/(tp+fn)
    pcrr = 2 * (precision * recall)/(precision + recall)
    return pcrr
```

其中y_true为真实的RSRP标签列，y_pred为预测的RSRP标签列

- **均方根误差 (RMSE : Root mean squared error)**

RMSE 是评估预测值和实测值整体偏差的指标，其大小直观表现了仿真准确性。直接计算待评估数据的 RMSE，计算公式如下：

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P^{(i)} - \hat{P}^{(i)})^2} \tag{6}$$

其中$P^{(i)}$为参赛队机器学习模型对于第 i 组评审数据集的 RSRP 预测值，$\hat{P}^{(i)}$为第 i 组评审数据集的 RSRP 实际测量值。

4.3.4 模型提交与数据获取

组委会将为参赛队提供华为云 ModelArts 作为 AI 运算平台，训练数据集都存储在该平台上。参赛队伍可以将训练数据下载到本地展开训练，同时竞赛评审也利用华为云大赛平台进行。

本次竞赛线上部分的数据集获取、模型提交、评分与排名系统等详细内容请访问本次竞赛的华为云网站：

https://developer.huaweicloud.com/competition/competitions/1000013923/introduction

线上作品提交时间：9 月 21 日早上 9:00 - 9 月 23 日中午 12:00

参赛选手可以多次提交模型，每个队伍每天提交次数上限为 5 次。最终以其提交中最优成绩为准。