赛题三：视觉+几何+语义：多源异构数据协同的视频目标跟踪

一、赛题背景

随着人工智能与计算机视觉技术的快速发展，视觉目标跟踪作为智能感知领域的核心技术之一，在自动驾驶、智能安防、人机交互、无人机搜救等场景中具有重要应用价值。然而，实际复杂场景下的目标跟踪仍面临诸多挑战：例如目标遮挡、光照变化、背景干扰、快速运动等问题，单一模态（如可见光 RGB）数据易受环境噪声影响，导致跟踪精度下降。多传感器感知技术的发展为解决这一问题提供了新思路。距离深度（Depth）数据可提供目标的空间几何信息，增强对尺度变化和遮挡场景的鲁棒性；文本描述则能引入语义先验，辅助目标身份与属性的精准定位。如何高效融合多源异构数据的互补优势，突破复杂场景下的跟踪瓶颈，已成为学术界与工业界共同关注的焦点。

基于以上背景与需求，本赛题旨在推动复杂场景下准确鲁棒的视觉目标跟踪算法发展，鼓励参赛者通过多源数据融合和视觉目标跟踪理论方法的创新改进，有效提升视觉目标跟踪性能。本赛题将提供包含可见光 RGB 图像序列、深度 Depth 图像序列和文本数据的目标跟踪数据集。参赛队伍需要设计基于三模态数据输入的视觉目标跟踪算法，充分利用多源异构数据互补信息（视觉+几何+语义），实现复杂场景下目标的精准跟踪。本赛题主要考察参赛学生的问题分析、数据处理、深度网络构建、算法及编程、数据可视化等能力。

二、赛题应用场景

多源异构数据协同的视频目标跟踪对于自动驾驶、智能安防、人机交互等诸多实际应用场景都具有重要应用价值。例如，在智能安防场景中，当警方需在地铁站密集人流中追踪一名携带危险物品的嫌疑人时，系统需融合可见光 RGB 摄像头（捕捉衣着颜色、外观等信息）、深度传感器（定位空间位置）及文本线索（如“手持黑色行李箱的人”等描述的语义信息），实时锁定目标。当目标突然蹲下藏匿物品时，可见光 RGB 图像可能因视角部分遮挡影响，深度数据仍可通过几何轮廓跟踪目标，同时结合文本语义验证行为异常，最终实现精准拦截。该场景能够直观体现多源数据协同应对遮挡、形变、干扰的核心作用，与赛题“视觉+几何+语义”深度融合的目标高度契合。

三、赛题任务

本赛题旨在利用多源数据融合技术和视觉目标跟踪技术，实现复杂场景下视觉目标准确稳定跟踪，即对三模态数据输入进行处理分析，实现给定目标的持续准确跟踪。

（一）任务描述

在视觉目标跟踪中，给定目标的初始状态（视频初始帧中的位置和尺寸），参赛

者需要对可见光 RGB 图像序列、Depth 图像序列和文本描述三种数据输入进行分析和处理，从而预测出目标在后续图像帧中的位置和尺寸，实现目标跟踪。本赛题提供的数据集中可见光 RGB 图像和深度 Depth 图像在空间和时序上都是对齐的，文本数据是描述目标在视频第一帧中的表观状态。任务给定目标在第一帧的目标位置和尺寸，要求参赛者设计算法对三模态数据输入进行处理分析，实现目标的高效准确定位跟踪。

（二）任务输入输出说明

数据集标签说明：目标在每一帧的标签格式为[x, y, w, h]，表示的是在图像中目标的矩形边界框，其中 x, y 为左上角坐标，w, h 是目标矩形框的宽和高。

算法输入：RGB 图像序列、Depth 图像序列、文本描述、目标在第一帧中的矩形边界框[x1, y1, w1, h1]。其中 RGB 图像为三通道 UINT8 格式，Depth 图像为单通道 UINT16 格式，文本 TXT 文件为一句描述目标的英文句子。

算法输出：目标在序列每一帧图像中的位置和尺寸，即每一帧图像中目标的矩形边界框：[[x1, y1, w1, h1], [x2, y2, w2, h2], ..., [xn, yn, wn, hn]]。

四、数据集及数据说明

（一）数据来源

本赛题提供的数据集为自行采集的数据，其中可见光 RGB 图像序列和 Depth 深度图像序列是使用 ZED 双目相机采集，文本描述是根据目标在 RGB 图像序列的第一帧中呈现出的表观由人为进行文本描述标注。赛题所提供数据集在采集过程中涉及到的受试者均获得了同意。数据未经过归一化等预处理操作。

（二）数据规模

本赛题提供的训练集包含 1000 个序列，每个序列原长 600 帧，按每隔 10 帧抽取 1 帧进行标注，因此训练集共包含带目标标注框的 60,000 个 RGB-Depth 图像对，1000 条描述目标初始状态的文本。验证集包含 50 个序列，共包含 11,800 个带目标标注框的 RGB-Depth 图像对，50 条文本。测试集初赛、复赛和总决赛的测试集各包含 50 个 RGB-Depth 图像序列，50 条文本描述。示例数据可以从链接 https://pan.baidu.com/s/1rsSJJ-PWXS1_g6m8ym_FaQ?pwd=4k8a 获取，正式数据将在报名后开放下载。

（三）数据格式

RGB 图像为三通道 UINT8 的 JPEG 格式图像；Depth 图像为单通道 UINT16 的 PNG 格式图像；文本数据 TXT 格式文件，其中包含一句描述目标的英文句子。每个序列的目标框标签为一个 TXT 文件，文件的每一行是目标在一帧图像中的位置和尺寸[x, y, w, h]。

五、算法设计要求

（一）算法类型

本赛题对算法类型不做严格限制，鼓励参赛者采用深度学习的视觉目标跟踪算法，

如基于卷积神经网络或 Transformer 神经网络的视觉目标跟踪网络，也可以结合其他机器学习算法。

（二）创新性

本赛题鼓励参赛队伍提出创新算法框架或者在现有算法基础上做出创新改进，以提高复杂场景下基于多源异构数据的视觉目标跟踪算法性能。例如，参赛队伍设计新的多源异构数据信息融合方法和视觉目标跟踪网络架构提升跟踪性能。

六、性能指标要求

本赛题使用成功率曲线的下面积 AUC 作为评估参赛者解决方案性能的指标。AUC 的具体计算方式如下：

首先，计算预测目标框和真实标签框之间的重叠率（Overlap Rate）：

$$R_t = \begin{cases} \dfrac{P_t \cap G_t}{P_t \cup G_t}, if\ P_t \neq \emptyset\ \&\ G_t \neq \emptyset \\ 1,\ if\ P_t = \emptyset\ \&\ G_t = \emptyset \\ 0,\ otherwise \end{cases}$$

其中，P_t和G_t分别是第 t 帧图像中算法预测目标框和真实标签目标框。当目标在第 t 帧中可见时，R_t会测量预测值与真实边界框之间的交并比 （Intersection over Union, IoU） 值。如果目标超出视野范围或被完全遮挡，即 G_t为空，如果P_t预测也为空，则R_t的值将被赋为 1。对于其他情况，R_t的值为 0。

然后，计算不同阈值下的成功率（Success Rate），阈值分别为从 0 到 1，每隔 0.05 取一个值，即$\theta \epsilon \{0,\ 0.05, 0.1, 0.15, \ldots,\ 0.95, 1\}$，在阈值$\theta_i$下的成功率方式为：

$$u_t(\theta_i) = \begin{cases} 1, if\ R_t > \theta_i \\ 0, otherwise \end{cases},\quad SR(\theta_i) = \frac{1}{T}\sum_{1}^{T} u_t(\theta_i)$$

其中R_t根据上述公式计算得出，θ_i为阈值，$u_t(\theta_i)$为第 帧预测是否成功的指标。如果 R_t大于阈值θ_i，则$u_t(\theta_i)$取值为 1，否则取值为 0。阈值θ_i下的成功率 $SR(\theta_i)$为成功的帧数量占比。

最后，根据不同阈值θ_i下的成功率 $SR(\theta_i)$值，可以绘制成功率 SR 和θ的曲线，AUC 为曲线的下面积。

七、功能要求

（一）准确性

算法模型实现面向多源异构数据的视觉目标跟踪时，需要具备较高的准确性，确保在较为复杂的场景下也能准确跟踪目标，减少跟踪漂移的情况。准确性以 AUC 指标进行衡量。

（二）鲁棒性

面对不同质量的 RGB、Depth 和文本数据，算法应能稳定运行，输出可靠的跟踪结果。在一个或者多个模态数据成像质量较差时，算法也不应出现大幅性能波动，保持对目标跟踪的准确性和稳定性。

八、开发环境

（一）软件环境

1. 编程语言

本赛题不强制编程语言的使用，但推荐使用 Python 语言的 3.6 及以上版本，因其丰富的数据处理、科学计算库和深度学习框架支持，如 NumPy、Pandas、Matplotlib 等用于数据处理和可视化的库、OpenCV 等图像数据处理的库、Pytorch、TensorFlow 等深度学习框架库。

2. 深度学习框架

推荐使用 TensorFlow2.x 或 PyTorch1.x，这两个框架在深度学习领域广泛应用，具有高效的计算性能和丰富的 API，便于模型的搭建、训练和部署。

3. 计算资源

本赛题不对参赛队伍使用的计算资源做限制。参赛者可使用本地工作站或云端计算平台进行开发和训练。本地工作站需配备 NVIDIA GPU（如 RTX 系列及以上）以加速深度学习计算；云端平台可选择阿里云天池、腾讯云 TI 平台、百度 AIStudio 等，这些平台提供了多种配置的计算资源，方便参赛者根据需求灵活选择。

（二）硬件环境

本赛题对参赛队伍使用的 CPU 型号、内存大小、GPU 型号等硬件资源不做强制性要求，参赛队伍可根据自身需求自由配置。

九、成绩评价

（一）输入数据格式要求

参赛者算法应能正确读取赛题数据，包括三通道 UINT8 格式的 RGB 图像、单通道 UINT16 格式 Depth 图像、TXT 格式的文本数据，以及保存目标矩形边界框[x, y, w, h]的 TXT 文件。

（二）输出数据格式要求

算法对每个序列数据进行目标跟踪，要求以 TXT 文件输出目标在序列每一帧图像中的位置和尺寸，即每一帧图像中目标的矩形边界框：[[x1, y1, w1, h1], [x2, y2, w2, h2], ..., [xn, yn, wn, hn]]。

（三）成绩计算

成绩将根据算法输出结果与真实标注数据对比，依据性能评估指标 AUC 进行打

分。

十、解题思路

（一）数据预处理

在数据训练过程中可以进行一些数据增强的预处理操作，例如对 RGB 图像和 Depth 图像进行归一化操作、进行图像旋转、缩放、翻转、裁剪等操作扩充训练数据集，增强模型的泛化能力。

（二）特征提取

参赛队伍应选取合适的深度学习网络结构提取 RGB 图像、Depth 图像以及文本数据的特征，通过训练并结合机器学习方法如特征选择等实现高判别性和高鲁棒性的数据特征表示。

（三）模型训练

参赛队伍可以选择合适的深度学习框架（如 TensorFlow、PyTorch）搭建模型，设置合理的训练参数，如学习率、迭代次数、批量大小等。在训练过程中，利用验证集数据对模型进行评估和调优，防止模型过拟合。

（四）多模型决策融合

参赛队伍可尝试将多个不同结构或训练阶段的模型进行融合，如采用投票法或加权平均法等，综合多个模型的预测结果，提高最终跟踪预测的准确性。根据性能评估指标，参赛队伍可对模型进行针对性优化，如调整模型结构、增加训练数据量等。

十一、参考资源

（一）书籍

1. 《深度学习》，由 IanGoodfellow 、YoshuaBengio 和 AaronCourville 撰写，系统介绍了深度学习的基础概念、模型架构和训练方法，对理解和应用神经网络有很大帮助。

2. 《Python 深度学习》，作者 François Chollet，使用 Python 进行深度学习的探索实践，涉及计算机视觉、自然语言处理、生成式模型等应用。书中包含 30 多个代码示例，步骤讲解详细透彻。

（二）在线课程

1. Coursera 上的“DeepLearningSpecialization”课程，由吴恩达教授授课，涵盖了深度学习的多个关键领域，包括神经网络基础、卷积神经网络、循环神经网络等，课程内容丰富且理论性强。

2. Bilibili 上的“动手学深度学习 Pytorch 版”课程，讲解了如何使用 Python 语言 Pytorch 深度学习框架实现深度神经网络，课程内容的应用实践性强。

（三）学术论文

[1] Zhou L, Zhou Z, Mao K, et al. Joint visual grounding and tracking with natural language specification[C]//Proceedings of the IEEE/CVF conference on computer vision

and pattern recognition. 2023: 23151-23160.

[2] Zhu J, Lai S, Chen X, et al. Visual prompt multi-modal tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 9516-9526.

[3] Zhu X F, Xu T, Liu Z, et al. UniMod1K: towards a more universal large-scale dataset and benchmark for multi-modal learning[J]. International Journal of Computer Vision, 2024, 132（8）: 2845-2860.

十二、提交要求

（一）算法代码

参赛队伍应提交完整的算法代码，包括数据预处理、模型训练、预测推理等各个环节的代码。代码需使用符合规范的 Python 语言编写，具备清晰的注释和文档说明，以便评审人员理解和运行。

（二）测试结果文件、模型文件与环境

参赛队伍应提交测试的结果文件、训练好的模型文件（由于模型可能较大，请提供模型下载链接），并提供模型的加载和使用说明，包括所需的运行环境、依赖库等信息。模型文件应能够在指定的测试环境中正常运行并输出预测结果。

（三）技术报告

参赛队伍应提交详细的技术报告，内容包括算法设计思路、模型架构图、实验设置（如训练参数、数据增强方法等）、性能分析（对主要指标和次要指标的详细分析）以及算法的创新点和不足之处的分析。技术报告格式采用 PDF，页数不限。

十三、更新与答疑

赛题可能会进行更新，为了回复参赛者在参赛过程中遇到的问题，赛题将单独设立选手答疑群（选手正式报名后可看到）。

十四、比赛流程及奖项设置

（一）报名阶段

参赛者在比赛官方网站上完成报名注册，提交个人或团队信息，获取初赛数据下载链接。

（二）初赛阶段

参赛者利用赛事方提供的训练数据集进行算法模型设计，利用提供的初赛测试集进行相应方法的验证与调试。初赛阶段参赛者每天提交结果的次数不限，但是初赛排行榜每隔 1 小时刷新一次。

（三）复赛（省赛）阶段

初赛结束后进入复赛阶段，开放复赛数据下载链接。仅有初赛阶段提交有效结果的参赛团队可以进入复赛。复赛期间，参赛者利用赛事方提供的复赛阶段数据进行算

法模型调试，提交对复赛测试数据的测试结果。复赛阶段持续 3 天，每个参赛队伍每天仅能提交 2 次。复赛排行榜每隔 1 小时刷新一次。

（四）复赛（省赛）成绩公布

在比赛官方网站上公布复赛成绩。以进入复赛参赛团队数量作为计奖基数，按照不超过大赛省赛设奖比例，评选出复赛一、二、三等奖（颁发省赛获奖证书）。评选复赛奖过程中，参赛者提交的算法性能低于赛事方提供的基线参考分数的判定为无效成绩，不予授奖。复赛一、二等奖晋级参加国赛总决赛。

（五）决赛阶段

1. 决赛线上评选。决赛阶段开放决赛数据下载链接，参赛者利用赛事方提供的决赛阶段数据进行算法模型调试，提交对决赛测试数据的测试结果。决阶段持续 1 周，每个参赛队伍每天仅能提交 1 次。决赛排行榜每隔 1 小时刷新一次。依据决赛排行榜结果，以进入决赛参赛团队数量作为计奖基数，按照不超过大赛国赛设奖比例，评选出国赛一等奖候选名单及国赛二、三等奖获奖名单（颁发国赛二、三等奖证书）。

2. 决赛作品提交。国赛一等奖候选团队在规定时间内提交技术文档、算法代码和模型文件、演示视频、补充材料等。提交截止后，不再接受任何形式的修改和补充。

3. 决赛评审阶段。由专业评审团队对国赛一等奖候选参赛团队的参赛作品进行结果复现与审核。评审过程中如有疑问，可要求参赛者进行解释说明。

4. 线下总决赛。国赛一等奖候选团队在规定时间内提交完善后的技术文档、算法代码和模型文件、演示视频、补充材料，参加国赛线下总决赛复核答辩，最终依据算法性能得分和线下答辩得分确定国赛一等奖获奖名单及其排名（未参加线下复核答辩视同放弃奖项）。国赛一等奖颁发荣誉证书。

十五、联系方式

赛项交流 QQ 群：1011080681

邮　　箱：xuefeng.zhu@jiangnan.edu.cn

报名官网：www.aicomp.cn