1.1 研究背景

目标跟踪任务以视频为输入，要模型识别出每帧的目标物体，并且关联前后帧中的同一物体，并赋予唯一的TrackID，且此ID在整个序列中保持不变。所有出现的目标都要能够及时找到，并且目标位置要尽可能与真实位置一致。

近年来，仅使用RGB图像进行目标跟踪的技术已取得许多研究突破。然而，基于RGB图像单模态的目标跟踪技术存在较大挑战。在光照变化、阴影、遮挡等复杂场景下，在夜晚、雨雪天气等应用环境中，RGB图像的信息可能会受到严重干扰，导致目标跟踪性能下降。

RGB-T（RGB-Thermal，可见光与热红外）目标跟踪，旨在结合RGB图像和热红外图像的信息来实现对目标物体的持续、准确跟踪。RGB图像提供丰富的颜色、纹理等视觉信息，适合在光照良好的环境下进行目标识别与跟踪。相比之下，热红外图像能够捕捉到物体的热辐射信息，对光照条件不敏感，因此在夜间或光照不足的环境中表现出色。

RGB-T目标跟踪技术通过融合RGB和热红外两种模态的信息，能够互补彼此的不足，提高在复杂场景下的跟踪稳定性和准确性。这种技术广泛应用于视频监控、自动驾驶、无人机追踪等领域，对于提升这些系统的智能化水平和实用性具有重要意义。

1.2 研究现状

1.2.1 基于非深度学习的传统RGB-T目标跟踪技术

基于非深度学习的传统RGB-T目标跟踪技术包括基于手工特征的传统方法、基于稀疏表示的方法、基于滤波器的方法等[1]。

Han Xu等[2]的DRF模型，根据信息源对源图像进行分解，引入了图像融合的解纠缠表示方法，缓解唯一信息提取不恰当的问题。该模型通过其手工特征方法，融合这些不同类型的表示，然后再将融合的表示输入预训练的生成器，生成融合结果。虽然该模型在后续使用了基于深度学习的结果生成器；但在输入生成器之前，模型先使用手工特征的手段融合RGB和红外图像模态。这种基于手工特征的方法，存在特征表示能力不够的问题，跟踪性能也有一定的局限性。

Lin Li等[3]研究了基于稀疏表示的融合可见光图像和红外图像的鲁棒目标跟踪。基于稀疏表示的跟踪器能有效面对噪声和误差。然而，该方法由于使用像素强度作为特征表示，在面对复杂场景时，鲁棒性却表现得较差。

Chengwei Luo等[4]提出了一种基于滤波器的鲁棒目标跟踪方法，包括了基于相关滤波器的跟踪（CFT）模块和基于直方图的跟踪（HIST）模块。相关滤波方法通过在滤波模板和搜索区域之间进行相关操作来生成响应图，然后使用峰值对目标的位置进行确定。基于相关滤波器的跟踪器在精度和速度之间实现了良好的平衡，但在复杂环境中的特征表示能力有限。

1.2.2 基于深度学习的RGB-T目标跟踪技术

基于深度学习的跟踪器通过训练，可以获得更鲁棒的特征表示，从而显著提高跟踪器的性能。然而，深度学习在执行RGB-T目标跟踪任务时，一般需要设计较大的网络结构，并需要大量的数据进行训练。

Lichao Zhang等[5]基于DiMP跟踪框架中设计了不同的融合策略，使用了大规模的训练数据集，包括9,335个视频，共1,403,359帧。该模型进一步分析多模态融合的有效性，考虑了几种融合机制，包括不同水平上的像素级、特征级和响应级融合。然而，这种大规模的模型在训练和运行的效率不高，且跟踪性能有提升的空间。

Transformer[6]以注意力机制为核心，针对序列型的输入数据，最开始用于NLP领域，近些年也应用在视觉跟踪领域，并取得较大突破。Xin Chen等[7]在孪生框架中设计了一个基于自我注意的自我增强模块和一个基于交叉注意的特征融合模块，以关注全局信息。一方面，基于siamese的框架将对象跟踪视为搜索区域和模板区域之间的相似性度量问题，让处理速度达到了实时性的需求；另一方面，Transformer可以通过其自注意机制捕获序列中的长距离依赖关系，其接受范围比传统的CNN更大。

此外，在RGB-T目标跟踪中，需要面对如热交叉、照明变化、比例变化、遮挡和快速运动等挑战。Yun Xiao等[8]研发了APFNet。研发者针对上述5个挑战属性，设计了5个分支，并设计了融合结构将分支融合。其实验表明，模型充分利用多模型挑战属性之间的信息，有效地实现了RGB-T目标跟踪。

参考文献

[1] H. Zhang et al., "A Comprehensive Review of RGBT Tracking," in IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1-23, 2024, Art no. 5027223, doi: 10.1109/TIM.2024.3436098.

[2] H. Xu, X. Wang and J. Ma, "DRF: Disentangled Representation for Visible and Infrared Image Fusion," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-13, 2021, Art no. 5006713, doi: 10.1109/TIM.2021.3056645.

[3] Li, L., Li, C., Tu, Z., Tang, J. (2018). A Fusion Approach to Grayscale-Thermal Tracking with Cross-Modal Sparse Representation. In: Wang, Y., Jiang, Z., Peng, Y. (eds) Image and Graphics Technologies and Applications. IGTA 2018. Communications in Computer and Information Science, vol 875. Springer, Singapore.

[4] Chengwei Luo, Bin Sun, Ke Yang, Taoran Lu, Wei-Chang Yeh, Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme, Infrared Physics & Technology, Volume 99, 2019, Pages 265-276, ISSN 1350-4495.

[5] Zhang, L.; Danelljan, M.; Gonzalez-Garcia, A.; van de Wei jer, J.; and Shahbaz Khan, F. 2019. Multi-modal fusion for end-to-end RGB-T tracking. In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2252–2261.

[6] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, 5998–6008.

[7] Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 8126–8135.

[8] Y. Xiao, M. Yang, C. Li, L. Liu, and J. Tang, “Attribute-based progressive fusion network for RGBT tracking,” in Proc. 36th AAAI Conf. Artif. Intell. (AAAI), 2022, vol. 36, no. 3, pp. 2831–2838.