专题综述

生物特征模板保护研究与展望

安徽大学 张慧 王华彬 金哲 李学俊

一、引言

随着互联网和硬件设备快速发展，生物特征识别技术的识别精度和速度已满足实际应用需求，被广泛应用于交通监管、门禁管理和移动应用等领域。相比于传统密码、智能卡等方式，生物特征作为身份凭证，具有唯一性、稳定性、便捷性、安全性等优势，但与此同时，生物特征被视为个人敏感信息，一旦泄露或被非法使用，将给用户带来不可挽回的后果。生物特征模板保护（Biometric Template Protection, BTP）作为可信生物特征识别（Trustworthy Biometrics）的关键技术是目前的研究热点。

1.1 生物特征识别

生物特征识别是基于个人的生理(例如人脸、虹膜、指纹和掌纹等)或者行为特征 (例如步态、语音和笔迹等)进行用户身份的识别。通用的生物特征识别框架主要包括 5 部分，如图 1 所示。传感器（Sensor）用于读取用户的生物特征信号，例如相机、指纹采集器等，通过传感器采集的信号质量影响着最终的识别性能。生物信号经过特征提取器(Feature Extractor)转换成生物特征并存储在数据库(Database)中。匹配器(Matcher)用于对认证时产生的查询生物特征与数据库中注册得到的生物特征模板进行比对,比对结果输入决策模块(Decision Module)，从而得到决策结果。

由于深度学习、硬件加速设备等技术的快速发展，目前生物特征识别的精度和速度已达到了很高的水平，并应用于很多刚需场景中。因此，关注生物特征识别的

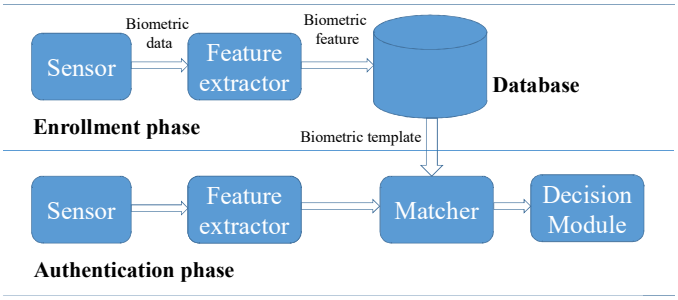


图 1 通用的生物特征识别框架示意图

安全和隐私性问题，提高生物特征安全，保障用户隐私，始终是推动生物特征识别技术持续发展的关键之处。

1.2 生物特征模板保护

生物特征模板保护技术通过不可逆变换或者加密原始生物特征以生成安全的生物特征模板，这有利于避免直接存储原始生物特征数据所带来的安全和隐私泄露风险。在注册阶段，用户的生物特征以安全模板的形式存储在数据库中；在认证阶段，用户的实时生物特征经过同样的模板保护过程生成特征模板，并与数据库中存储的模板进行匹配操作，根据系统预先设置的阈值和二者的相似性程度返回用户身份的认证结果。

生物特征模板保护主要分为可撤销生物特征(Cancelable Biometrics， CB)和生物特征加密(Biometric Cryptosystems，BC)两个方向。CB 方案中，原始生物特征经过不可逆变换得到可撤销的模板，并在变换域内完成模板的匹配操作，但基于可撤销模板不可能或很难还原出原始生物特征，从而保护生物特征安全。BC 方案则结合密码学原语从生物特征生成或绑定密钥，

通过验证恢复密钥的准确性实现用户合法身份的认证。

1.3 亟待解决的关键问题

虽然目前已经有很多模板保护方案提出，但生物特征模板保护技术的发展仍受到生物特征的内在特性和实际应用场景的限制，生物特征的安全性和识别系统的实用性无法保证。主要概括为以下三个关键问题：

(1) 噪声干扰致使不可完全再现：受光照、角度和环境等因素的影响，用户的同一生物模态因大量随机噪声的存在致使每次提取的特征都存在差异，不能完全再现。所以用户在认证时输入的生物特征和数据库中注册存储的特征模板不完全相同，二者进行匹配操作时得到的相似度可能低于阈值，从而降低生物特征识别的性能。而基于生物特征继而进行的模板保护操作，往往以性能降低为代价提高生物特征的安全性，这会进一步增大系统性能降低的程度。因此克服生物特征的噪声干扰，平衡识别性能和生物特征安全，设计具有容错性的安全生物特征识别系统是推进该领域研究的一个重要挑战。

(2) 外部因子泄露威胁系统安全：为赋予生物特征模板如密码一样灵活重置的能力，降低生物特征泄露致使用户永久不可复用该生物特征的风险，一般采用生物特征和外部因子(如令牌)结合的方式生成不可逆的、可撤销的生物特征模板。但这种可撤销方案使得生物特征的安全与外部因子的安全息息相关，当系统受到令牌丢失攻击(Lost token attack)时，攻击者可能从丢失的令牌中分析出原始生物特征或者伪装成合法用户，威胁系统安全，降低识别性能。因此，如何设计一种更具鲁棒性的可撤销生物特征模板方案至关重要。

(3) 攻击与对抗技术的此消彼长：针对生物特征识别系统的攻击无处不在，攻击方法层出不穷，例如字典攻击(Dictionary Attacks)、关联攻击(Correlation Attacks)、爬山攻击(Hill Climbing Attacks)和重建攻击(Reconstruction attacks)等。即使始终有对抗攻击的模板保护方案提出，但面对复杂的应用环境和无法预知的攻击威胁，生物特征的保护方案在攻击者面前无能为力。攻击与对抗技术此消彼长的状态致使生物特征模板保护仍处于不断发展的阶段。

二、可撤销生物特征方案研究现状

根据所采用的变换方式是否可逆，BC 方案分为两类：不可逆变换和生物特征盐析。基于不可逆变换的 BC 方案通过变换原始生物特征模板，使转换模板无法反转。经典算法有 IoM 哈希[1]、布隆过滤法(Bloom Filter)以及基于布隆的改进算法。生物特征盐析则通过设置一些人工模式(例如随机噪声)与生物特征模板进行混合操作，从而保护原始生物特征。经典算法有生物哈希(Biohashing)以及基于生物哈希的扩展。这些方案在一定程度上提高了生物特征安全，但大多数可撤销技术都以严重的性能下降为代价提高安全性，无法保持与原始生物特征相比的合理精度。基于此，下面介绍近期的代表性工作，以解决目前 CB 方案中普遍存在的难点问题。

2.1 LIoM 哈希算法

针对系统对高性能和匹配速度的要求，Dong 等人[2]提出了用于大规模开集人脸识别的基于 IoM 哈希的模板保护方案：LIoM（Index-of-max hashing by learning）哈希。LIoM 哈希是基于 IoM 哈希设计的紧凑人脸特征表示算法，通过汉明距离实现高效的匹配操作，IoM 哈希的不可逆变换过程确保用户的隐私性保护。为验证算法的有效性，结合多种融合策略在大规模人脸数据集上进行了全面评估。

2.2 免对齐的可撤销虹膜特征识别方案

特征对齐是存在于 CB 方案中的一个重要问题。由于很多生物特征(如虹膜)无法精准对齐，为保持合理精度匹配过程需要进行移位打分操作。但 CB 方案的使用转变了原始特征空间，该匹配策略不再适用。基于此，Lee 等人[3]基于方向梯度直方图提出了一种随机增强梯度直方图算法（Random Augmented Histogram of Gradients，R·HoG）用于虹膜特征模板保护，将未对齐的 irisCode 转换为对齐健壮的可撤销模板，提高了模板匹配的效率和生物特征系统的安全性。

2.3 无令牌的可撤销生物特征识别方案

由于 CB 方案通常被设计用来保护具有两个输入因子的生物特征模板，即：生物特征识别和用于模板替换

的令牌，一旦令牌丢失，受保护模板极易遭受安全攻击和隐私侵犯。Lee 等人[4]提出了一种单因子可撤销生物特征识别方案，即扩展特征向量(Extended Feature Vector，EFV)哈希算法，该算法只需要一个生物特征作为输入，并利用与生物特征数据分离的置换密钥作为匹配的标识符，从而实现安全的单因子认证。

2.4 单因子可撤销生物特征认证方案

双因子可撤销的生物特征认证方法引入额外因子即令牌化因子带来了隐私和安全威胁问题。针对这一问题，孔小景等人[5]提出了一种唯一二值数据生物特征作为输入因子的单因子可撤销生物识别方法，即 WSE 哈希算法。WSE 哈希算法满足不可逆性，可撤销性，不可链接性以及精确性这 4 个可撤销的生物特征模板保护标准，也抵御了 3 种方式的安全性攻击测试。同时 WSE 哈希算法也可以扩展到二值向量形式表示的虹膜、面部特征、掌纹和静脉等生物特征识别。另外，算法安全性，如碰撞攻击、差分攻击等攻击方式，是未来研究方向。

三、生物特征加密方案研究现状

BC 方案的设计思想是出于对密钥的保护，同时也保证了生物特征的安全。BC 方案主要分为两类：密钥绑定和密钥生成。密钥绑定是采用生物特征与密钥进行绑定，产生公开的辅助数据，认证时结合生物特征与辅助数据以释放密钥。目前具有代表性的方案有模糊承诺(Fuzzy Commitment)，模糊保险箱(Fuzzy Vault)等。密钥生成则是基于生物特征生成或提取出密钥，认证时比对密钥以认证用户身份。代表方案有模糊提取器(Fuzzy Extractor)、安全骨架(Secure Sketch)等。但所有的 BC 方案都存在一个问题，就是提取的生物特征质量影响着恢复密钥的正确率。

3.1 基于密钥绑定的指纹细节点保护方案

基于生物特征类内差异大的特性，目前 CB 方案都依赖纠错码来提高系统的容错能力。但纠错码的纠错能力有限，很难平衡安全性和性能要求。因此，Jin 等人[6]提出了一种非纠错码密钥绑定方案以及基于指纹细节点的可撤销的不可逆变换。该方案不局限于二元特征和匹配器，可应用于多种生物特征表示。具体的密钥绑定和密钥恢复过程如图 2 所示。

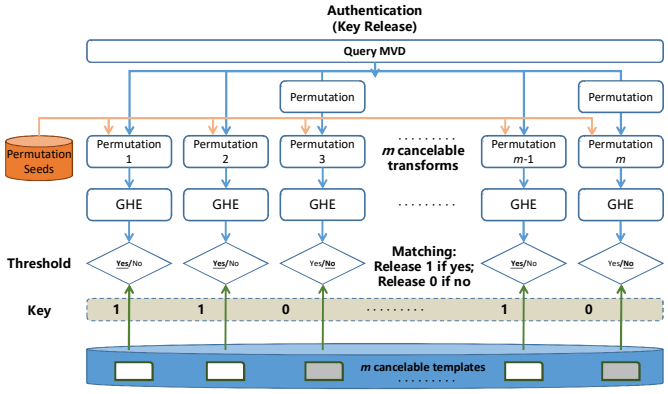


图 2 密钥绑定和密钥恢复示意图

3.2 基于对称密钥环加密的 BC 方案

由于生物特征提取时存在各种噪声干扰，生物特征识别系统出现模糊性而识别性能下降或不工作。为提高系统的容错性和安全性，Lai 等人[7]把密钥绑定视为对称的加解密问题，提出了基于对称密钥环加密(Symmetric Keyring Encryption，SKE)的 CB 方案。SKE 方案由 RV 密钥对、过滤机制和 Shamir 的秘密共享方案组成，理论分析和实验结果显示该方案可扩展到其他生物特征，并可以抵抗多种安全攻击。具体的方案概述如图 3 所示。

3.3 基于深度哈希网络的多光谱掌纹模糊承诺方案

在生物识别密码系统中，所生成的生物密钥通常结合单向函数进行严格保护，但这很难平衡模板的大小和准确性。Wu 等人[8]采用深度哈希网络，生成存储量小、匹配复杂度低的二值模板。根据鉴别能量，对模板中的比特优选后，减少了模板数据量和构造模糊承诺的计算复杂度。

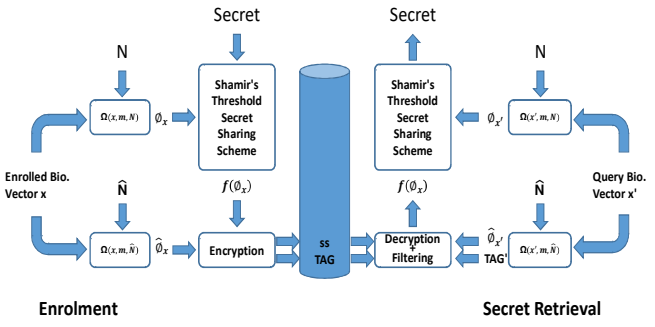


图 3 SKE 方案概述图

四、生物特征模板的攻击与对抗

随着黑客技术的不断进步，针对生物特征模板的攻击呈现出层出不穷的趋势。为增强生物特征识别的安全性和用户隐私保护，研究攻击算法并设计对抗方案尤为重要。

4.1 基于 GAN 生成器的深度人脸特征重构

基于深度卷积神经网络(CNN)的人脸识别目前已经达到了很高的识别精度，但从深度学习模型中提取的特征(深度特征)的安全性和私密性问题常常被忽视。基于此，Dong 等人[9]提出了在不访问 CNN 网络配置的情况下基于 GAN 生成器从深层特征重构人脸图像的方案，具体的框架概述如图 4 所示。该方案使用 GAN 生成器同时发挥优化目标的人脸分布约束和人脸生成器的作用，实现了高相似度和高视觉质量的人脸图像重构。该工作中所伪造的人脸图像揭示了当前人脸识别系统的安全和隐私风险，对手可能伪造人脸图像非法访问人脸识别系统。因此，需要结合模板保护方案和反欺骗检测手段保护生物特征，规避隐私数据泄露风险。

4.2 增强的掌纹重构攻击

目前很多重建攻击重构的原像普遍存在自然性、完整性和视觉质量差等问题。基于此，Sun 等人[10]采用了两种策略，一是邻域范围修改约束，减少图像质量的恶化；二是挑选重要的像素，打包修改，既增强了修改对优化的影响，同时降低了引入过多的图像不自然痕迹，从而得到两种增强的重构攻击用于掌纹识别。

4.3 基于风格迁移技术的掌纹重构攻击

攻击者基于跨数据库攻击重建的图像可以有效的攻击其他生物特征系统。为实现在线跨数据库攻击，并保证重构图像的高质量，Yang 等人[11]提出了两种新颖的风格转移技术，从特征模板重构恢复原始图像，并用于攻击基于编码的掌纹识别系统。提出的重构方法揭示了基于纹理编码的掌纹识别方法的脆弱性，因此，为了提高生物特征识别系统安全和用户隐私，有效的攻击防御技术和重构检测方法成为不可或缺的一项研究。

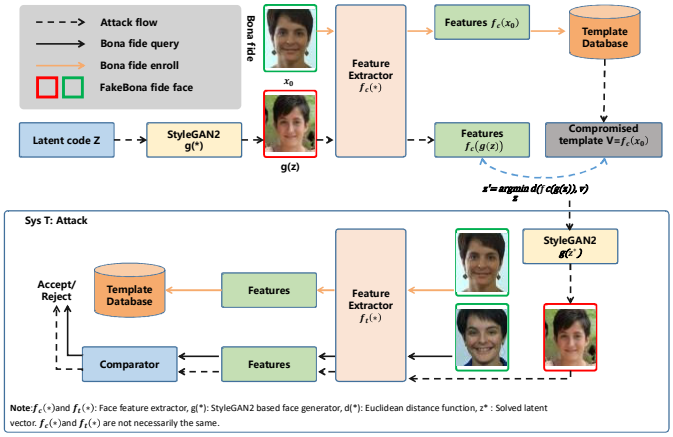


图 4 人脸图像重构框架图

五、总结与展望

本文主要论述了生物特征模板保护技术在可撤销生物特征和生物特征加密两个方向的研究进展，并针对目前该领域所存在的关键问题，重点介绍了具有代表性的研究工作。随着生物特征识别技术的不断普及，生物特征模板保护作为一项必要且有意义的工作必将持续发展。基于此，本文对未来的发展进行展望：

（1）无令牌的可撤销生物特征模板保护方案设计

目前无令牌 CB 方案比较欠缺，可以借鉴各学科领域研究方法设计安全无令牌 CB 方案，打破目前生物特征模板保护的发展瓶颈，提高生物特征认证系统安全性。

（2）多模态生物特征融合策略设计

多模态生物特征识别系统已被证明具有更高的性能和安全性优势，但特征级融合作为系统中的关键环节仍面临挑战，需要设计兼容不同生物特征信号的融合策略，提高系统对不同类型和维度生物特征的使用能力和常见攻击的抵抗能力。

（3）重构攻击和对抗方案设计

为了维持性能和安全的平衡，可撤销生物特征模板保护方案具有相对距离保持属性，但这容易招致基于相对距离保持的攻击，攻击者试图基于模板重构原像。因此，需要研究基于相对距离保持的攻击框架，设计对抗策略，从而在复杂多变的应用环境保证生物特征识别系统的安全运行。

责任编委 储珺

参考文献

[1] Jin, Z., Hwang, J. Y., Lai, Y. L., Kim, S., & Teoh, A. B. J. (2017). Ranking-based locality sensitive hashing-enabled cancelable biometrics: Index-of-max hashing. IEEE Transactions on Information Forensics and Security, 13(2), 393-407.

[2] Dong, X., Kim, S., Jin, Z., Hwang, J. Y., Cho, S., & Teoh, A. B. J. (2020). Open-set face identification with index-of-max hashing by learning. Pattern Recognition, 103, 107277.

[3] Lee, M. J., Jin, Z., Liang, S. N., &Tistarelli, M. (2022). Alignment-Robust Cancelable Biometric Scheme for Iris Verification. IEEE Transactions on Information Forensics and Security, 17, 3449-3464.

[4] Lee, M. J., Jin, Z., & Teoh, A. B. J. (2018, December). One-factor cancellable scheme for fingerprint template protection: Extended Feature Vector (EFV) Hashing. In 2018 IEEE international workshop on information forensics and security (WIFS) (pp. 1-7). IEEE.

[5] 孔小景，李学俊，金哲，周芃，陈江勇. 一种单因子的可撤销生物特征认证方法.自动化学报, 2021, 47(5): 1159-1170.

[6] Jin, Z., Teoh, A. B. J., Goi, B. M., & Tay, Y. H. (2016). Biometric cryptosystems: a new biometric key binding and its implementation for fingerprint minutiae-based representation. Pattern Recognition, 56, 50-62.

[7] Lai, Y. L., Hwang, J. Y., Jin, Z., Kim, S., Cho, S., & Teoh, A. B. J. (2019). Symmetric keyring encryption scheme for biometric cryptosystem. Information sciences, 502, 492-509.

[8] Wu, T., Leng, L., & Khan, M. K. (2022). A multi-spectral palmprint fuzzy commitment based on deep hashing code with discriminative bit selection. Artificial Intelligence Review, 1-18.

[9] Dong, X., Miao, Z., Ma, L., Shen, J., Jin, Z., Guo, Z., & Teoh, A. B. J. (2022). Reconstruct Face from Features Using GAN Generator as a Distribution Constraint. arXiv preprint arXiv:2206.04295.

[10] Sun, Y., Leng, L., Jin, Z., & Kim, B. G. (2022). Reinforced Palmprint Reconstruction Attacks in Biometric Systems. Sensors, 22(2), 591.

[11] Yang, Z., Leng, L., Zhang, B., Li, M., & Chu, J. (2022). Two novel style-transfer palmprint reconstruction attacks. Applied Intelligence, 1-18.

王华彬

安徽大学计算机科学与技术学院副教授，研究方向：模型识别与信息处理，医疗图像处理，虚拟现实。

Email: wanghuabin@ahu.edu.cn

金哲

安徽大学人工智能学院教授，研究方向：可信人工智能，模式识别与安全，深度学习。

Email: jinzhe@ahu.edu.cn

李学俊

安徽大学计算机科学与技术学院教授，研究方向：边缘计算与智能软件，医学人工智能，工业互联网。

Email: xjli@ahu.edu.cn

热点追踪

基于自监督学习的单目深度估计方法

中科院自动化研究所 周正铭 董秋雷

一、摘要

单目深度估计旨在从单幅输入图像中估计场景的深度，其对于三维重建、场景理解等任务有着重要的意义。由于在真实场景中获取稠密而准确的深度真值是很困难的，基于自监督学习的单目深度估计受到了广泛的关注。近年来，尽管自监督单目深度估计方法取得了较好的表现，但如何进一步缩小自监督与有监督方法之间的差距并提升其精度仍然是一个开放性的问题。针对这一问题，本文一方面从训练约束的角度，考虑如何更有效地同时利用两种现有的自监督深度约束训练模型；另一方面从网络结构的角度，考虑如何使网络学到对于深度估计更有效的特征。具体地，我们提出了一种感知遮挡的由粗到细的自监督单目深度估计方法，称为 OCFD-Net；提出了一种自蒸馏特征聚合模块，并在此基础上提出自蒸馏聚合网络，称为 SDFA-Net。相关成果分别被 ACM MM 2022 和 ECCV 2022 录取。

二、引言

基于自监督学习的单目深度估计旨在使用没有深度真值的样本训练一个深度神经网络，使其可以从单幅输入图像中预测稠密的深度图。现有方法一般采用不同视角拍摄的同一场景的多幅图像作为训练数据，将深度估计任务转化为图像重建任务，并使用重建损失训练网络(如图 1 所示)。根据训练数据的来源，现有自监督单目深度估计方法可以分为采用视频序列训练的方法和采用双目图像训练的方法。其中，采用视频序列训练的方法[1,2]在训练阶段以视频序列中的连续帧作为训练样本。由于连续帧之间的相机运动是未知的，这些方法在训练阶段除了估计深度图之外，还需要估计图像之间的相机运动情况。采用双目图像训练的方法[3,4]在训练阶段以双目相机拍摄的图像对作为训练样本。由于拍摄双目

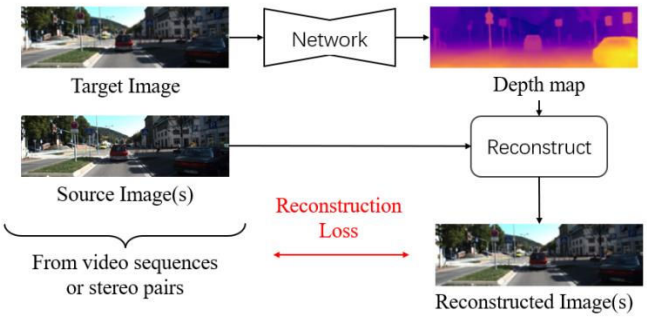


图 1 基于自监督学习的单目深度估计基本原理

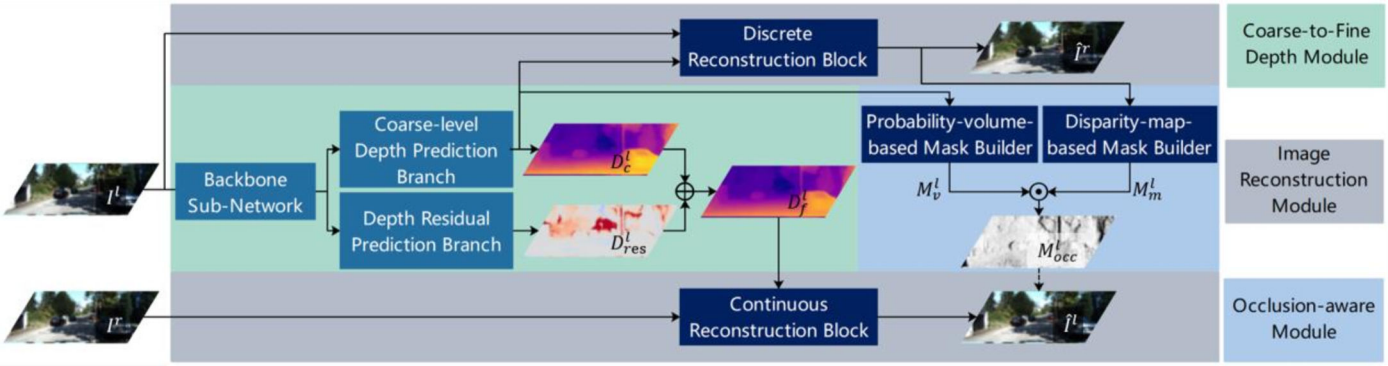


图 2 感知遮挡的由粗到细自监督单目深度估计网络 OCFD-Net 的结构示意图

图像的相机相对位置是固定的，且双目系统中的视差和深度具有确定的转换关系，这些方法只需要预测图像对应的深度图或视差图。

尽管近年来基于自监督学习的方法得到了广泛的研究且取得了较好的表现，但如何进一步提升自监督单目深度估计方法的精度仍然是一个开放性的问题。针对这一问题，一方面，本文提出了一种感知遮挡的由粗到细自监督单目深度估计网络，称为 OCFD-Net[5]。该模型通过分别估计粗粒度深度和场景深度残差的方法，结合了连续[2]和离散[4]两种深度约束的优势，并通过一个遮挡感知模块缓解训练中的遮挡问题对深度结果造成的负面影响。另一方面，本文提出了一种基于自蒸馏的特征聚合模块，并基于此模块设计了一种新的单目深度估计网络，称为 SDFA-Net[6]。在自蒸馏特征聚合模块中，采用三个分支分别预测三个特征偏移图，用于在自蒸馏条件下对待融合的多尺度特征进行细化。实验结果表明，我们提出的 OCFD-Net 和 SDFA-Net 在室外驾驶场景数据集上的表现超越了绝大多数现有的方法。

三、正文

为了有效地在自监督单目深度估计中利用连续和离散两种深度约束，我们首先通过对比实验分析了两种深度约束各自的优点和不足。分析结果表明：离散深度约束有助于保留更多深度细节信息，且可以使模型取得较高的精度，但使用离散深度约束训练的模型难以在平坦区域生成平滑的深度估计结果；连续深度约束有助于保持深度结果的平滑性，但其使得估计结果的精度相对较低。基于上述分析，我们提出了感知遮挡的由粗到细自监督单目深度估计网络 OCFD-Net(如图 2 所示)。该

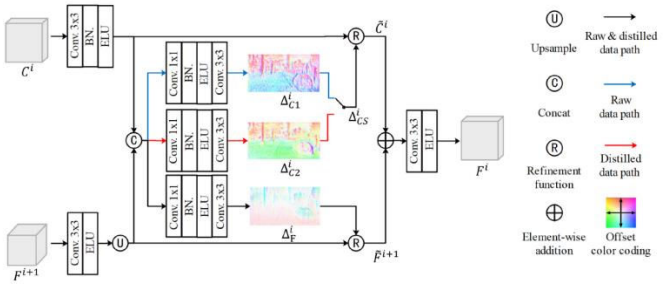


图 3 自蒸馏特征聚合模块

网络分为三个模块，其中由粗到细的深度估计模块用于从输入图像估计深度图。该模块通过一个骨干子网络从输入图像中提取特征。以该特征为输入，分别用粗粒度深度预测分支预测一个离散表示的粗粒度深度图，用深度残差预测分支预测一个连续表示的场景深度残差图。最终将两部分相加得到细粒度的深度图。图像重建模块通过预测的深度结果进行图像重建，从而对网络进行自监督训练。具体地，通过离散形重建的方式，引入离散深度约束对粗粒度深度进行训练；通过连续形重建的方式，引入连续深度约束，对深度残差进行训练。遮挡感知模块用于通过估计的深度结果计算图像中潜在的遮挡区域，并输出表示遮挡概率的 Mask。具体的，该模块分别基于深度概率体和视差图生成两个遮挡 Mask，并将两个遮挡 Mask 逐像素相乘得到遮挡概率 Mask。在训练时，我们采用重建损失和平滑正则同时训练粗粒度和细粒度的深度结果。其中，在对细粒度深度结果进行训练时，我们基于遮挡概率 Mask 减小遮挡区域的重建损失，并加大相应区域的平滑正则，来缓解遮挡问题对深度估计结果造成的负面影响。

为了使得网络能在自监督方式下学到对于深度估计更有效和准确的特征，我们提出了自蒸馏特征聚合模

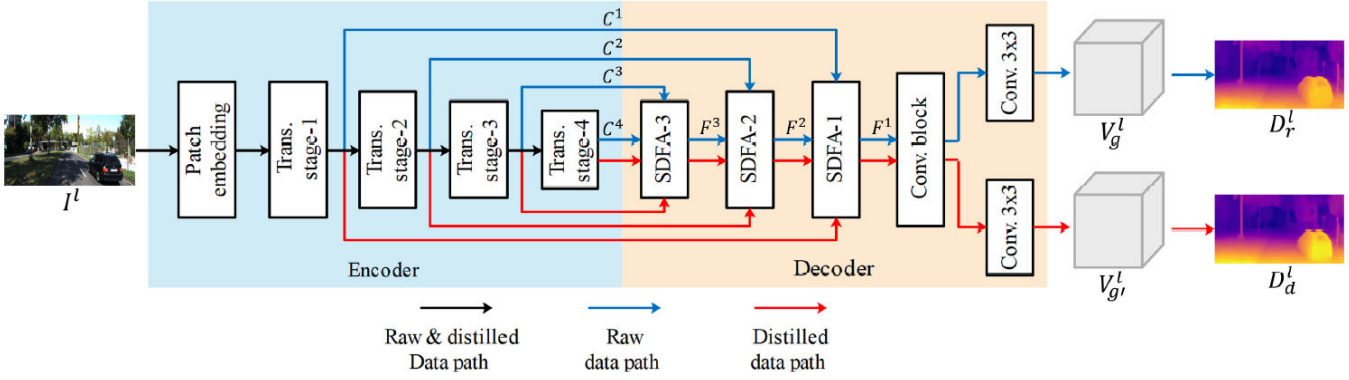


图 4 自蒸馏特征聚合网络 SDFA-Net 的结构示意图

块(如图 3 所示)用于融合两个不同尺度的特征，并保持其上下文一致性。该模块受到语义分割任务中特征对齐模块[7]的启发，使用可学习的偏移向量来融合不同尺度的特征。考虑到在自监督学习的过程中，图像重建损失存在一定的歧义性，可能使得模块无法通过自监督方式学习到准确的特征偏移图，进而造成深度结果的误差，自蒸馏特征聚合模块采用了三个分支来学习三个不同的特征偏移图。其中一个特征偏移图被用于细化小尺度的特征；其余两个特征偏移图被共同用于细化大尺度的特征，并分别通过图像重建损失和自蒸馏损失进行训练。进一步地，我们以改进的 Swin-transformer 作为编码器，以自蒸馏特征聚合模块作为解码器设计了用于单目深度估计的自蒸馏特征聚合网络 SDFA-Net(如图 4 所示)。SDFA-Net 的解码器中存在两条前向传播的数据通路，分别称为原始数据通路和蒸馏数据通路。在不同的数据通路中，会使用自蒸馏特征聚合模块中相应的偏移图预测分支来学习特征偏移图。为了有效地训练所提出的模型，我们将训练中的每次迭代分为三个步骤：第一步使用编码器提取多尺度特征，并用解码器中的原始数据通路预测深度图，采用图像重建损失和平滑正则作为损失函数；第二步使用网络中的蒸馏数据通路从编码器得到的多尺度特征中估计深度，并从第一步估计的深度结果中选择置信度较高的部分作为伪标签，计算蒸馏损失。第三步对损失进行反向传播来训练网络。

四、实验结果

表 1 展示了我们提出的 OCFD-Net 和 SDFA-Net 在 KITTI 室外驾驶场景数据集中的深度估计结果。可以看出在绝大多数指标下本文所提出的两个方法都取得了最好的表现。图 5 进一步展示了两个方法深度估计的可视化结果。从图中可以看出两个方法都能较好地保留场景中的深度细节信息，例如在细小的物体处以及物体的边缘处等。

为了验证本文所提出方法的有效性，图 6 中展示了 OCFD-Net 预测的粗粒度深度图(第二行)，深度残差图(第三行)和细粒度深度图(第四行)的可视化结果。对于深度残差图，红色表示残差为正，蓝色表示为负。可以看到深度残差使得细粒度深度在物体边缘处更加准确，在图像中的平坦区域结果更加平滑。图 7 展示了 SDFA-

表 1 OCFD-Net，SDFA-Net 和其他算法在 KITTI 数据集上的深度估计结果

Method	PP.	Abs. Rel. ↓	Sq. Rel. ↓	RMSE ↓	logRMSE ↓	A1 ↑	A2 ↑	A3 ↑
Monodepth2	✓	0.107	0.849	4.764	0.201	0.874	0.953	0.977
MonoResMatch	✓	0.111	0.867	4.714	0.199	0.864	0.954	0.979
DepthHints	✓	0.096	0.710	4.393	0.185	0.890	0.962	0.981
DBoosterNet-e		0.095	0.636	4.105	0.178	0.890	0.963	0.984
SingleNet	✓	0.094	0.681	4.392	0.185	0.892	0.962	0.981
FAL-Net	✓	0.093	0.564	3.973	0.174	0.898	0.967	**0.985**
EPCDepth	✓	0.091	0.646	4.207	0.176	0.901	0.966	0.983
EdgeOfDepth	✓	0.091	0.646	4.244	0.177	0.898	0.966	0.983
PLADE-Net	✓	0.089	0.590	4.008	0.172	0.900	0.967	**0.985**
OCFD-Net		0.091	0.576	4.036	0.174	0.901	0.967	0.984
OCFD-Net	✓	0.090	0.563	4.005	0.172	0.901	0.967	0.984
SDFA-Net		0.090	0.538	3.896	0.169	0.906	**0.969**	**0.985**
SDFA-Net	✓	**0.089**	**0.531**	**3.864**	**0.168**	**0.907**	**0.969**	**0.985**

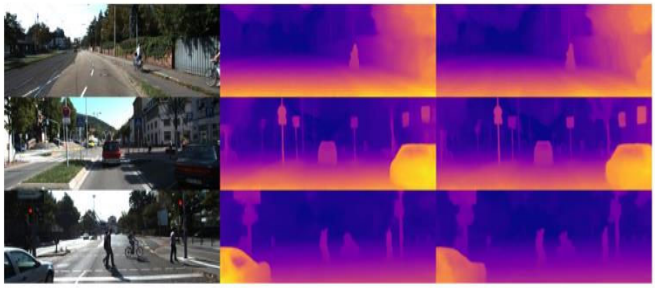


图 5 OCFD-Net 和 SDFA-Net 在 KITTI 数据集上的深度估计可视化结果

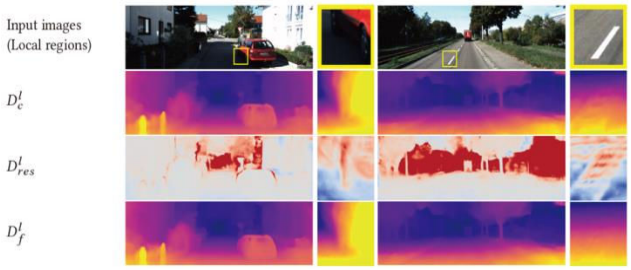


图 6 OCFD-Net 预测的粗粒度深度，深度残差和细粒度深度的可视化结果

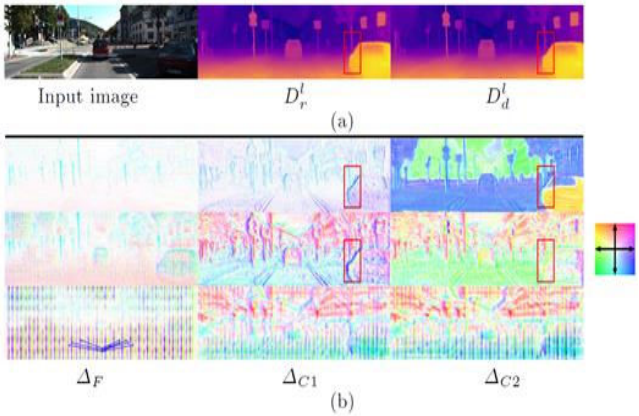


图 7 (a)SDFA-Net 通过不同数据通路预测的深度可视化结果 (b)自蒸馏特征聚合模块中特征偏移图可视化结果

Net 中使用不同数据通路预测的深度结果，可以看到使用蒸馏通路预测的深度图((a)中第三列)更加准确，尤其是在物体的边缘处。此外，我们可视化了 SDFA-Net 中多个自蒸馏特征聚合模块中学习到的特征偏移图，其中第一列的偏移图用于细化小尺度特征，第二、三列的偏移图用于分别在原始和蒸馏通路中细化大尺度特征。可以看到相较于原始通路中的特征偏移图，蒸馏通路中的的特征偏移图在物体边缘处更准确。

责任编委 崔海楠

参考文献

[1] Zhou, Tinghui, Matthew Brown, Noah Snavely, and David G. Lowe. "Unsupervised learning of depth and ego-motion from video." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1851-1858. 2017.

[2] Godard, Clément, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. "Digging into self-supervised monocular depth estimation." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3828-3838. 2019.

[3] Garg, Ravi, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. "Unsupervised cnn for single view depth estimation: Geometry to the rescue." In European conference on computer vision, pp. 740-756. Springer, Cham, 2016.

[4] GonzalezBello, Juan Luis, and Munchurl Kim. "Forget about the lidar: Self-supervised depth estimators with med probability volumes." Advances in Neural Information Processing Systems 33 (2020): 12626-12637.

[5] Zhou, Zhengming, and Qiulei, Dong. "Learning Occlusion-Aware Coarse-to-Fine Depth Map for Self-Supervised Monocular Depth Estimation" In Proceedings of the 30th ACM International Conference on Multimedia, pp. 6386–6395. 2022.

[6] Zhou, Zhengming, and Qiulei Dong. "Self-distilled feature aggregation for self-supervised monocular depth estimation." In European Conference on Computer Vision, pp. 709-726. Springer, Cham, 2022.

[7] Huang, Zilong, Yunchao Wei, Xinggang Wang, Wenyu Liu, Thomas S. Huang, and Humphrey Shi. "Alignseg: Feature-aligned segmentation networks." IEEE Transactions on Pattern Analysis and Machine Intelligence 44, no. 1 (2021): 550-557.

周正铭

中科院自动化研究所硕士研究生。主要研究方向为深度估计、三维计算机视觉等。
Email: zhouzhengming2020@ia.ac.cn

董秋雷

中科院自动化研究所研究员。主要研究方向为三维计算机视觉、模式识别等。
Email: qldong@nlpr.ia.ac.cn

热点追踪

基于图注意力双线性池化的鲁棒性 RGB-T 跟踪

南京邮电大学 江晨风 康彬 周全

一、研究背景

随着多媒体技术的蓬勃发展，热红外摄像机已经成为一种经济实惠的摄像机。该摄像机可以捕捉温度高于绝对零度的目标发射的热红外辐射，适用于夜间监视。将 RGB 与热红外摄像机联合使用优点如下：(1)热红外摄像机具有很强的抗照度变化能力，可以在光照条件较差的情况下为 RGB 摄像机提供强有力的支持；(2)RGB 摄像机将有助于解决基于热红外摄像机的监控所面临的热交叉难题。因此，结合 RGB 和热特征的 RGB- T 跟踪可以有效应对恶劣天气挑战[1]。在 RGB-T 跟踪中，RGB 和热视频序列是成对获得的。其关键思想是利用 RGB 和热信息的互补性进行高效的多模型融合。

近些年来，研究者们开发了许多先进的方法进行多模型融合，例如基于粒子融合的 RGB- T 跟踪器[2, 3]，建立多图融合模型[4, 5]，求解统一优化问题 [6, 7]，用于 RGB- T 跟踪的密集卷积神经网络[8]，多适配器卷积网络[9]等等，其中后两种方法采用了深度卷积神经网络技术。与手工特征相比，深度卷积神经网络可以更好的提取深度语义信息，对目标进行鲁棒表示。因此，近年来深度学习技术在 RGB-T 跟踪方面表现出了巨大的潜力。然而，现有的基于 CNN 的 RGB-T 跟踪器通常将多层卷积特征图视为层次上的整体特征，忽略了 RGB 与热目标之间的部分特征相互作用。这可能会明显降低具有挑战性的视频对的跟踪精度。更严重的是，RGB 与热目标的少量有用信息可能在空间域上部分匹配甚至不匹配。在这种情况下，简单地将多个深层特征作为整体特征进行多模型融合，可能会产生不可避免的负面影响。

针对上述问题，本文提出了一种简单有效的面向四流的 Siamese 网络(FS-Siamese)用于 RGB-T 跟踪，其中四流的特征嵌入可分为范例嵌入对和候选嵌入对。通过基于图注意力的双线性池化模块，可以分别融合两个嵌入对，生成增强样本和增强候选，用于生成后续的相似图。对于双线性池化，其在异构部分信息融合方面表现出优于传统线性融合策略的性能。尽管双线性池化获得了一定的性能提升，但它无法区分深度特征图中元素的重要性。鉴于这些观察结果，我们在双线性池化中引入了共同注意力机制，将多模型池化描述为一个多图学习问题。由于目标外观可能发生剧烈变化，因此有必要在基于图注意力的双线性池化模块中引入一种有效的更新策略。目前最先进的更新策略[10, 11]只关注于探索当前目标特征与先前目标特征之间的时间相关性，而忽略了一个事实，即在线探索目标与其周围背景环境之间的空间相关性对于定位相似度最高的候选对来说是非常重要的。因此，我们设计了一种基于元学习的更新策略，以有效地更新基于图注意力的双线性池化模块的全连接层。这为利用类别信息在线更新样例语义表示提供了途径。本文的主要贡献如下:

(1)将基于注意力的双线性池化问题描述为一个多图学习问题。我们将图注意力网络和外积整合为一个统一的结构，使多个图在联合学习的同时实现有效局部信息交互。这样可以有效地消除目标对融合过程中的干扰。

(2)传统的面向多流跟踪网络只融合不同流的目标回归结果，在融合目标嵌入时没有探究成对关系。为了克服这一限制，我们提出了一种基于图注意的双线性池化的四流导向网络结构，用于有效融合多源嵌入对。

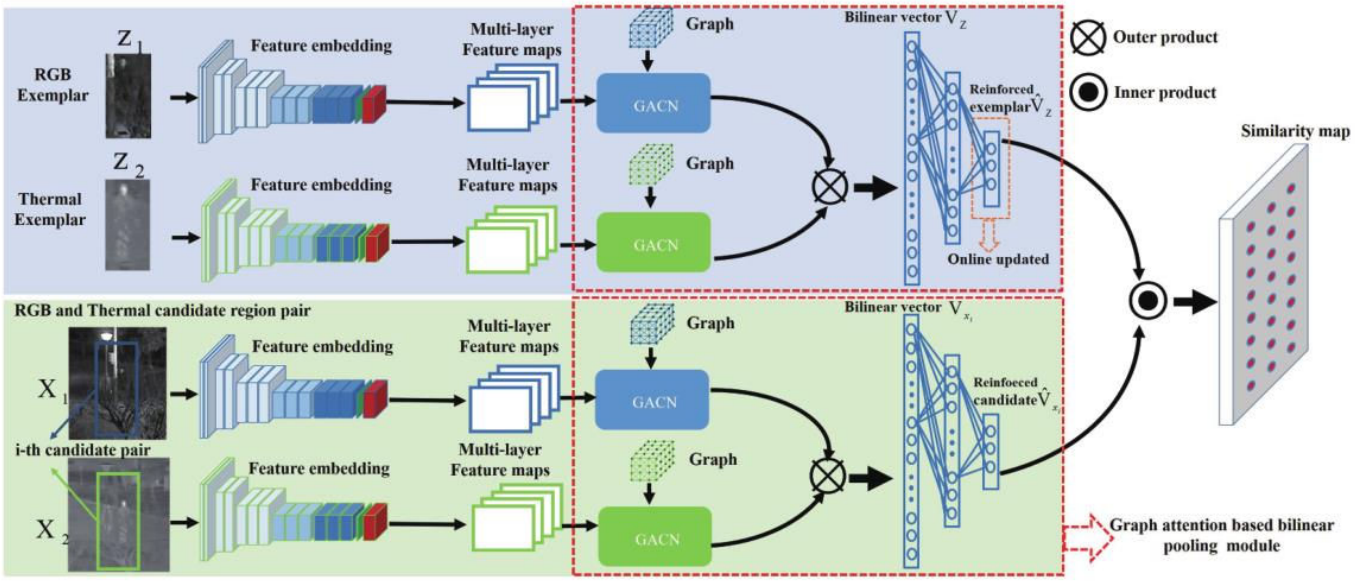


图 1 所提出面向四流的 Siamese 网络(FS-Siamese)结构示意图。

(3)将元学习应用于基于双线性池化的图注意力更新，利用类别信息在线限制样本学习到与当前跟踪结果相似的语义表示，有助于区分增强样本和增强候选样本。

(4)在 GTOT、RGBT234、CUB-2002011、FGVC-aircraft 和 Cars 数据集上的大量实验表明，基于图注意力的双线性池化模块可以有效地融合 RGB-T 跟踪中的多域多层特征图，同时可以扩展到其他多模型融合任务。

二、FS-Siamese网络结构介绍

1. 网络结构概述

我们的面向四流的 Siamese 网络(FS-Siamese)结构如图 1 所示，整体网络包含四个嵌入流。两个流用于嵌入目标范例(目标模板)对$\boldsymbol{Z_1}$和$\boldsymbol{Z_2}$。另外两个流用于在搜索区域内嵌入候选对($\boldsymbol{X_i^1}$和$\boldsymbol{X_i^2}$)。特征嵌入后，通过基于图注意力的双线性池化对样本嵌入对和第i个候选嵌入对分别进行强化融合。这可为内积计算提供一个局部强化的目标外观表征。在传统的 Siamese 网络中，目标位置的准确性依赖于样本和候选目标之间的相互关联。相比之下，我们的网络结构可以给出更准确的相似度计算结果。这是因为我们采用了基于图注意力的双线性池化模块，充分利用了多源嵌入对中固有的部分特征交互。

2. 基于图注意力的双线性池化模块

双线性池化是一种很有前途的模型，它可以克服线性池化的局限性，因为它使用外积来探索特征通道之间的成对相关性。假设我们得到两个域特征映射张量$\boldsymbol{F^1} \in \mathbb{R}^{N\times K\times C}$，$\boldsymbol{F^2} \in \mathbb{R}^{N\times K\times C}$ (N和K为单个特征映射的长度和宽度，C为特征映射通道的个数)。利用外积将两个张量的位置相乘，并将所有积集合在一起，最终可以得到双线性向量$\boldsymbol{V} \in \mathbb{R}^{c^2\times 1}$。由于特征图中的单个元素对应原始图像中的某个块，如果将目标块视为局部形式，双线性池化中的外积实际上可以探索两个图像域中局部形式之间的结构关系。这样我们就可以使用条件部分信息来表示目标外观。将张量$\boldsymbol{F^1}$和$\boldsymbol{F^2}$重新化为矩阵形式$\widetilde{\boldsymbol{F}}^{\boldsymbol{1}} \in \mathbb{R}^{NK\times C}$和$\widetilde{\boldsymbol{F}}^{\boldsymbol{2}} \in \mathbb{R}^{NK\times C}$，则双线性池化向量可以表示为:

$$\boldsymbol{V} = bilinear\left(\widetilde{\boldsymbol{F}}^{\boldsymbol{1}}, \widetilde{\boldsymbol{F}}^{\boldsymbol{2}}\right) = vec\left(\left(\widetilde{\boldsymbol{F}}^{\boldsymbol{1}}\right)^{\boldsymbol{T}}\widetilde{\boldsymbol{F}}^{\boldsymbol{2}}\right)$$

其中$\widetilde{\boldsymbol{F}}^{\boldsymbol{1}} = [\tilde{f}_1^1, \ldots, \tilde{f}_i^1, \ldots, \tilde{f}_C^1]$，$\widetilde{\boldsymbol{F}}^{\boldsymbol{2}} = [\tilde{f}_1^2, \ldots, \tilde{f}_i^2, \ldots, \tilde{f}_C^2]$，向量$\boldsymbol{V}$中的第$(j-1)C+i$)个元素记为$\boldsymbol{V}_{(j-1)C+i} = \left(\tilde{f}_i^1\right)^{\boldsymbol{T}}\tilde{f}_j^2$。$bilinear(\cdot)$表示双线性运算。向量$\tilde{f}_i^1$ (或$\tilde{f}_j^2$)中的每个元素表示图像块的条件局部形式。上式表明，每个局部表示具有同等的重要性，但却忽略了一个事实，对于多模型融合的$\widetilde{\boldsymbol{F}}^{\boldsymbol{1}}$和$\widetilde{\boldsymbol{F}}^{\boldsymbol{2}}$，其贡献实际上是不同的。

基于此，我们设计了一个基于图注意力的双线性池化模块来开发共同注意力机制。$\boldsymbol{V}$的元素被重新表述为:

$$\boldsymbol{V}_{(j-1)C+i} = \left(\tilde{\boldsymbol{f}}_i^1\right)^{\boldsymbol{T}}\boldsymbol{W}_{ij}\tilde{\boldsymbol{f}}_j^2$$

其中，共同注意权重矩阵$\boldsymbol{W}_{ij}$的目的是表明$\tilde{\boldsymbol{f}}_i^1$和$\tilde{\boldsymbol{f}}_j^2$ 中元

素之间的相关性。

本文的研究动机是将目标嵌入、共同注意权重矩阵估计和特征嵌入融合集成到一个统一的端到端网络结构中。为此，本文提出的基于图注意力的双线性池化模块将图注意力卷积网络和外积相结合，可以有效地利用消息传递在 RGB 和热图像中定位信息块，且计算复杂度低。基于图注意的双线性池化模块有关公式描述如下。

基于矩阵分解，$\boldsymbol{W}_{ij}$可分解为：$\boldsymbol{W}_{ij}=\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{D}_{ij}\boldsymbol{P}$，其中$\boldsymbol{D}_{ij}$是对角矩阵，可以进一步分解为两个对角矩阵$\boldsymbol{D}_{ij}=\boldsymbol{S}_i^{\,T}\boldsymbol{S}_i$。定义$\boldsymbol{D}_i=\boldsymbol{S}_i\boldsymbol{P}$，$\boldsymbol{D}_j=\boldsymbol{S}_j\boldsymbol{P}$。因此：

$$\boldsymbol{V}_{(j-1)C+i}=\left(\tilde{\boldsymbol{f}}_i^1\right)^{\boldsymbol{T}}(\boldsymbol{D}_i)^{\boldsymbol{T}}\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{P}\boldsymbol{D}_j\tilde{\boldsymbol{f}}_j^2=\left(\boldsymbol{P}_i\tilde{\boldsymbol{f}}_i^1\right)^{\boldsymbol{T}}\left(\boldsymbol{P}_j\tilde{\boldsymbol{f}}_j^2\right)$$

定义$\tilde{\boldsymbol{f}}_i^1=\boldsymbol{P}^{\boldsymbol{T}}\hat{\boldsymbol{f}}_i^1$，所以：

$$\boldsymbol{P}_i\tilde{\boldsymbol{f}}_i^1=(\boldsymbol{P}\boldsymbol{D}_i\boldsymbol{P}^{\boldsymbol{T}})\hat{\boldsymbol{f}}_i^1。$$

$\boldsymbol{D}_i$是方阵，可以进一步分解。假设$\boldsymbol{P}$是拉普拉斯矩阵的特征向量，$(\boldsymbol{P}\boldsymbol{D}_i\boldsymbol{P}^{\boldsymbol{T}})\hat{\boldsymbol{f}}_i^1$可以看作是图卷积。同样的，$\boldsymbol{D}_j$也可以使用图卷积进行更新。

基于上述分析，设$\boldsymbol{G}(\hat{\boldsymbol{F}}^1,\hat{\boldsymbol{A}}^1)$和$\boldsymbol{G}(\hat{\boldsymbol{F}}^2,\hat{\boldsymbol{A}}^2)$分别为RGB和热特征映射张量的属性图，其中$\hat{\boldsymbol{F}}^i$ $(i=1,2)$中的行表示为第i个图中的节点，$\hat{\boldsymbol{A}}^{\mathrm{i}}$为编码节点对之间的成对相似度的相邻矩阵。基于双线性池化的多图学习问题表述为：

$$\boldsymbol{V}=bilinear\left(\boldsymbol{G}(\hat{\boldsymbol{F}}^1,\hat{\boldsymbol{A}}^1),\boldsymbol{G}(\hat{\boldsymbol{F}}^2,\hat{\boldsymbol{A}}^2);\boldsymbol{\Theta}\right)$$

其中图$\boldsymbol{G}(\hat{\boldsymbol{F}}^1,\hat{\boldsymbol{A}}^1)$和图$\boldsymbol{G}(\hat{\boldsymbol{F}}^2,\hat{\boldsymbol{A}}^2)$可以被图卷积神经网络学习，$\boldsymbol{\Theta}=\{\boldsymbol{\Theta}^1,\boldsymbol{\Theta}^2\}$定义为图卷积神经网络的参数集，$bilinear(\cdot)$是指利用外积动态聚合两个图卷积神经网络的双线性算法。

我们建立图注意力卷积网络，以实现在没有任何先验知识的情况下进行图学习。具体来说：

$$\boldsymbol{P}_i\tilde{\boldsymbol{f}}_i^1=\sigma\left(\Sigma_{k\in N(i)}\eta(i,k)\hat{\boldsymbol{f}}_k^1\right)$$

其中$\eta(i,k)$表示节点i和k之间边界的权值，$\sigma(\cdot)$是激活函数，$N(i)$表示节点i的邻域集。根据$\boldsymbol{V}$的表达式，我们自适应地学习$\eta(i,k)$来估计$\boldsymbol{P}_i\tilde{\boldsymbol{f}}_i^1$。

同样，$\boldsymbol{P}_j\tilde{\boldsymbol{f}}_j^2$也可以用相似的方法估算。$\eta(i,k)$的计算方法如下：

$$\eta(i,k)=\frac{exp\left(LeakyReLU\left(\boldsymbol{\beta}^T\left[\boldsymbol{U}\hat{\boldsymbol{f}}_i^1\parallel\boldsymbol{U}\hat{\boldsymbol{f}}_k^1\right]\right)\right)}{\Sigma_{s\in N(i)}exp\left(LeakyReLU\left(\boldsymbol{\beta}^T\left[\boldsymbol{U}\hat{\boldsymbol{f}}_i^1\parallel\boldsymbol{U}\hat{\boldsymbol{f}}_s^1\right]\right)\right)}$$

式中$\boldsymbol{\beta}$为单层前馈神经网络的参数向量，$\boldsymbol{U}$为表示$\tilde{\boldsymbol{A}}$和$\hat{\boldsymbol{A}}$关系的参数矩阵。$\parallel$为串联算子，$LeakyReLU(\cdot)$为激活函数。

3. 更新策略

我们希望将基于图注意的双线性池化结果的更新重新表述为一个单样本学习问题。因为当前的跟踪结果实际上是正样本，与样本相似度较低的候选样本可视为负样本。无论当前的候选样本发生了多大的变化，范例和当前的跟踪结果应该仍然具有相同的类别。因此，我们可以将类别信息纳入到$\hat{\boldsymbol{V}}_z$的在线更新中，其中$\hat{\boldsymbol{V}}_z$是样本生成双线性向量后的全连通层。

为了实现元学习的目标，我们定义与$\hat{\boldsymbol{V}}_z$相似度最高的第i个候选池化结果$\hat{\boldsymbol{V}}_{x_i}$作为正样本$\boldsymbol{C}_1$，与$\hat{\boldsymbol{V}}_z$相似度最低的第$j$个候选池化结果$\hat{\boldsymbol{V}}_{x_j}$作为负样本$\boldsymbol{C}_2$。在分类中，我们引入参数向量$\boldsymbol{\phi}$，用于微调语义表征来更新样本。其在线训练损失函数定义为：

$$J(\boldsymbol{\phi})=-log\,\mathcal{P}\left(y=1\mid\hat{\boldsymbol{V}}_z\right)$$

式中$\mathcal{P}\left(y=1\mid\hat{\boldsymbol{V}}_z\right)$定义为：

$$\mathcal{P}\left(y=1\mid\hat{\boldsymbol{V}}_z\right)=\frac{exp\left(-\left\|f\left(\hat{\boldsymbol{V}}_z;\boldsymbol{\phi}\right)-\boldsymbol{C}_1\right\|^2\right)}{\sum_{k=1}^{2}exp\left(-\left\|f\left(\hat{\boldsymbol{V}}_z;\boldsymbol{\phi}\right)-\boldsymbol{C}_k\right\|^2\right)}$$

其中，$f(\cdot)$表示经过整个网络结构处理后的输出函数，参数向量$\boldsymbol{\phi}$包含特征嵌入模块，基于图注意力的双线性池化模块以及最后的内积计算这三部分所涉及到的所有参数。

三、实验结果

1. 定量跟踪实验

为了测试我们的网络结构的效率，我们在两个广泛使用的 RGB-T 数据集上进行了大量的实验:GTOT 和 RGBT234。具体结果可见于图 2 和图 3。主要是通过两个客观指标进行定量评估：精度图和成功图。精度图表示不同定位误差阈值下的累计位置误差，定位误差定义

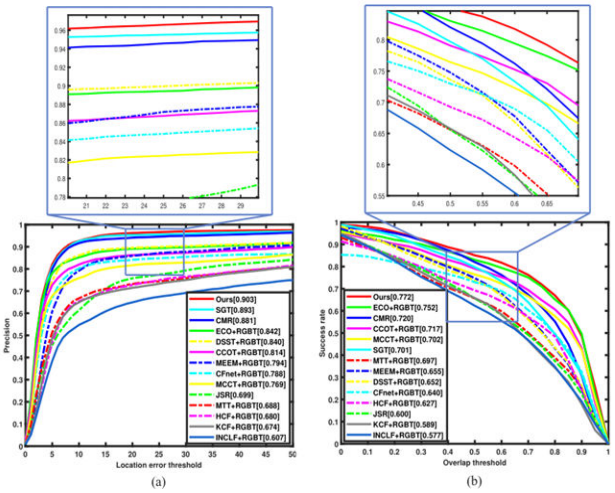


图 2　在 GTOT 数据集上的总体跟踪性能
(a)精度图，(b)成功图。

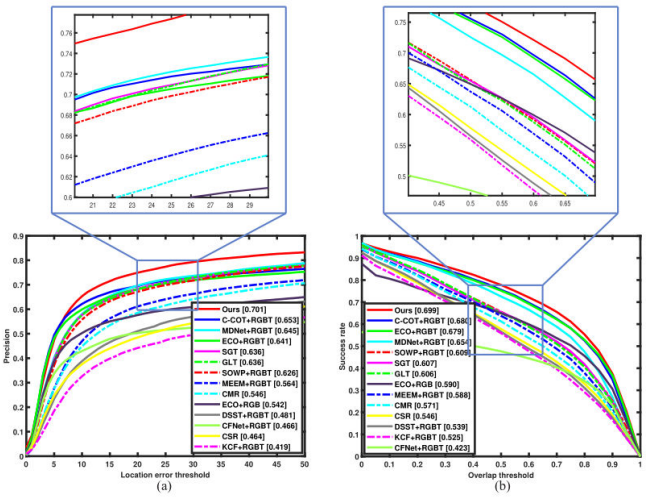


图 3　在 RGBT234 数据集的总体跟踪性能
(a)精度图，(b)成功图。

表 1　RGB-234 数据集中不同视频子集成功率的均值，最好的两个结果分别用红色和蓝色标注。

Attr. \ Meth.	Ours	ECO+RGBT	ECO+RGB	GLT	CFNet+RGBT	CMR	DSST+RGBT	CSR	KCF+RGBT	C-COT+RGBT	MDNet+RGBT	MEEM+RGBT	SOWP+RGBT	SGT	CMPP	self-SDCT+RGB
BC	56.1	52.0	49.9	50.7	28.8	37.6	42.5	34.1	37.5	50.2	48.5	52.8	50.7	50.3	53.8	44.3
CM	57.3	50.8	47.0	43.1	26.9	37.5	33.6	31.8	30.8	49.7	45.6	41.6	42.6	42.9	54.1	37.6
DEF	53.2	46.6	46.9	41.2	28.9	40.1	31.4	33.6	31.1	46.4	47.4	41.8	42.1	43.6	54.1	35.4
FM	54.6	45.6	45.3	41.9	28.6	42.2	30.3	34.7	27.3	45.3	47.6	46.3	45.5	45.2	50.8	36.3
HO	58.1	50.7	49.7	43.6	26.5	39.5	34.1	32.5	30.7	50.8	48.2	45.2	44.9	45.0	50.3	37.0
LI	58.6	52.6	54.0	48.2	32.9	34.4	43.2	34.7	39.1	53.9	43.8	48.1	49.6	45.8	58.4	48.3
LR	63.2	58.4	56.3	58.6	35.4	48.2	53.3	45.4	49.0	64.9	58.6	58.8	57.3	58.8	57.1	56.5
MB	54.5	55.2	49.5	43.3	23.6	37.9	34.1	29.5	29.9	49.9	46.7	37.8	43.2	41.1	54.1	36.1
NO	71.3	66.7	66.4	45.7	50.0	43.1	34.1	47.1	34.9	66.0	59.7	41.3	42.9	47.0	67.8	37.2
PO	62.7	62.2	61.6	50.7	42.6	45.5	43.0	43.1	40.8	62.1	57.6	49.2	52.2	51.3	60.1	43.5
SC	59.7	61.2	58.6	37.3	40.7	38.8	30.7	41.4	28.7	60.4	55.0	36.3	36.8	40.0	57.2	35.5
TC	67.2	70.0	62.4	51.5	41.1	55.0	34.2	37.6	34.7	71.9	59.5	51.9	51.6	57.2	58.3	38.8
Average	59.7	56.0	54.0	46.3	33.8	41.7	37.0	37.1	34.5	56.0	51.5	45.9	46.6	47.4	57.5	40.5

为跟踪包围框中心位置与人工标记的真实值之间的欧几里得距离，成功图反映了不同重叠阈值下的累计成功率(重叠分数大于 0.5 时的视频帧数)。由图 2 可以清晰看出，本方法的距离精度评分在 GTOT 数据集上比 ECO-RGBT 高出 5%以上。由于 ECO-RGBT 涉及热信息，其距离精度得分略高于 ECO-RGB。由图 3(a)可知，我们的方法的距离精度评分明显高于其他相比较的方法。同样，由图 3(b)，我们的方法在成功图中也获得了最好的结果。综上，得以验证我们的 FS-Siamese 网络在两个数据集上均表现出良好的性能。

此外，我们还在 RGB-234 数据集上测试了 12 个具有挑战性因素的影响，其结果如表 1 所示。从这次测试中我们可以清楚地看到，我们的方法在大多数有挑战性因素中获得了第一名。具体来说，重遮挡(HO)非常具有挑战性，因为只能从 RGB 和热目标中提取少量有用的信息。由于这个原因，最先进的跟踪方法，如 ECO, CMR 和 GLT 在这种情况下的跟踪性能很差。与传统方法相比，我们的方法成功率比顶级方法 CMPP 高 10%以上。除了 HO 之外，背景杂波(BC)、摄像机运动(CM)、快速运动(FM)、低照明(LI)和部分遮挡(PO)通常被认为是具有挑战性的场景，可作为验证跟踪精度的代表性测试。显然，与 CMPP 相比，我们的方法也能提高 6%以上的成功率。由于热交叉(thermal Crossover, TC)会严重干扰热目标的外观，池化模块在探索块关系时可能会受到更大的负面影响。由于这个原因，整体深度特征导向跟踪器(ECO+RGBT)提供了最好的成功率。表 1 的测试结果可知我们的方法可以有效地使用基于图注意力的双线性池化模块来增强有挑战性场景下的跟踪性能。

2. 定性跟踪实验

我们在图 4 中展示了定性跟踪性能。其中每个场景

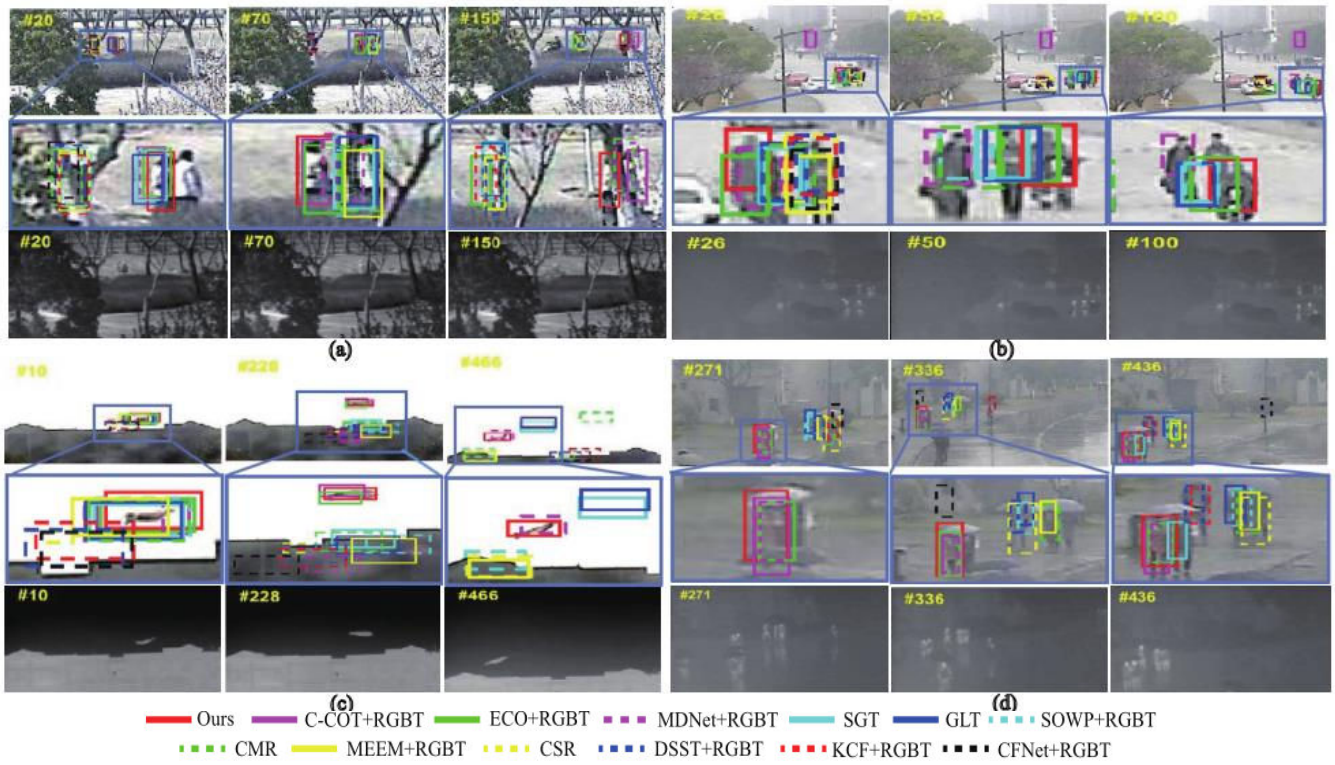


图 4 视频对定性结果(a) Diamond 视频对 (b) Elecbike3 视频对 (c) Kite4 视频对 (d) Manafterrain 视频对。

表 2 基于图注意力的双线性池化模块的消融实验设置。

Backbone	Outer product	GACN	Updating module	Methods
VGG-16	√	√	√	**Ours**
VGG-16	√	√		**Ours I**
VGG-16	√			**Ours II**

随机选择 3 个视频序列。运动目标在 diamond 序列中常被树干遮挡，即使是最先进的方法往往也会在严重遮挡后失去目标。从图 4(a)可以看出，无论局部遮挡还是重度遮挡，我们的方法都能跟踪目标。目标和相邻行人一起移动，造成图 4(b)中产生严重的背景杂波。在这种情况下，我们的方法可以做到与 ECO-RGBT 同样的效果，提供了良好的跟踪性能。如图 4(c)，在 kite 序列中，其他方法在第 300 帧后会开始有一定程度的漂移，而我们的方法仍然可以在整个视频帧中跟踪目标。图 4(d)为下雨情况下光照较低，通过该图可以看出，我们的方法可以有效地利用热信息来补充 RGB 序列。

3. 消融实验

基于图注意力的双线性池化模块是我们的 FS-Siamese 网络的核心模块，主要包括三个部分:图注意力卷积网络(GACN)、外部乘积和更新模块。在测试中，我

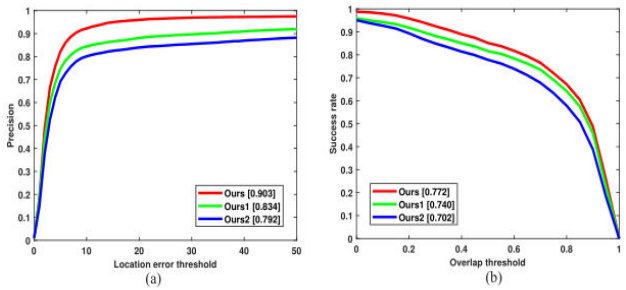


图 5 GTOT 数据集上基于图注意力的双线性池化模块消融试验。其中红色曲线对应实验设置中的方法 ours，绿色曲线对应实验设置中的方法 Ours I，蓝色曲线对应实验设置中的方法 Ours II。

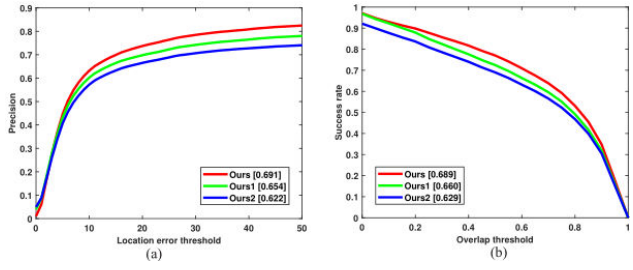


图 6 RGBT234 数据集上基于图注意力的双线性池化模块消融试验。不同颜色曲线实验设置规则同图 5。

们对 GTOT 和 RGBT234 数据集进行了消融实验，实验设置如表 2 所示，以展示不同部分的有效性。具体结果如图 5 和图 6 所示。从图 5 和图 6 中，我们可以看到两

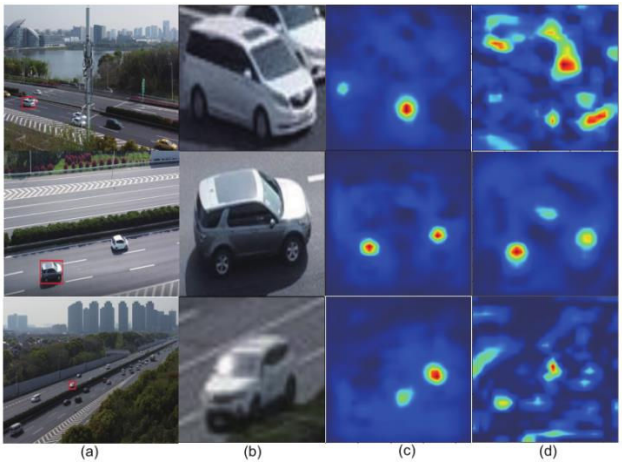


图 7　细粒度分类测试实验。 (a)测试图像，(b)局部放大图像，(c) MA-CNN+GACN 评估掩码，(d) MA-CNN 评估掩码。

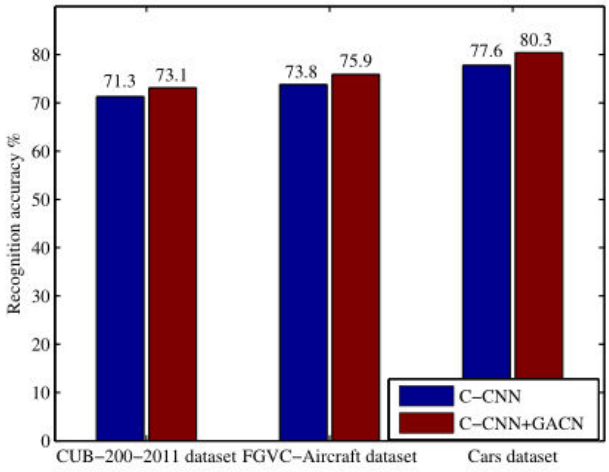


图 8　测试 GACN 普适性的细粒度识别。

个数据集上的精度图和成功图表明了我们基于图注意力的双线性池化模块的有效性。

在基于图注意力的双线性池化模块中，GACN 是及其重要的一部分，它可以通过探索局部特征交互来突出重要的图像块。基于此，我们设计了一个细粒度分类测试实验来展示 GACN 的有效性。具体来说，我们在 MA-CNN 网络的 Conv 层的末尾添加了 GACN。这样，目标嵌入就会更加关注信息丰富的图像块。详细测试实验如图 7 所示，图 7(c)和图 7(d)中评估的子区域掩码可以表明 GACN 的有效性。例如，第一行和第三行的局部放大图像分辨率较低，但 MA-CNN+GACN 仍能定位到信息子区域(如图 7(c))。相比之下，原始方法可能会在掩码中包含无信息的背景噪声(如图 7(d))。

此外，B-CNN 是细粒度识别中比较著名的一种方法，可以使用双线性池化融合两个网络结构的特征图。因此，我们将基于图注意力的双线性池化模块扩展到该方法，即“B-CNN+GACN”，以验证 FS-Siamese 创新的普适性。测试在三个细粒度识别数据集上进行:CUB-200-2011, FGVC-aircraft 和 Cars。从图 8 中我们可以清楚地看到，与原始 B-CNN 方法相比，B-CNN+GACN 可以明显提高 3%以上的识别精度。

四、总结

在本文中，我们提出了一个面向四流的 Siamese 网络(FS-Siamese)来有效地融合 RGB 和热信息。我们的网络得益于提出的基于图注意力的双线性池化模块，该模块可以采用共同注意力机制来探索 RGB 和热目标之间的部分特征相互作用。此外，我们采用元学习对双线性池化结果进行更新，可以对目标与其周围背景的空间关系进行在线更新。

在 GTOT 和 RGBT234 数据集上的大量实验表明，与最先进的 RGB 和 RGB- T 跟踪器相比，所提出的 FS-Siamese 网络可以提供更好的性能。

责任编委　王金甲

参考文献

[1] Y .Choi, N. Kim, S. Hwang, K.Park, J. S. Yoon, K.An, and I. S.Kweon, “Kaist multi-spectral day/night data set for autonomous and assisted driving,” IEEE Transactions on Intelligent Transportation Systems,vol. 19, no. 3, pp. 934–948, 2018.

[2] A.Leykin and R.Hammoud, “Pedestrian tracking by fusion of thermal-visible surveillance videos,” Machine Vision and Applications, vol. 21,no. 4, pp.587–595, 2010.

[3] M. Talha and R.Stolkin,"Particle filter tracking of camouflaged targets by adaptive fusion of thermal and visible spectra camera data," IEEE Sensors Journal, vol. 14, no. 1, pp.159–166, 2013.

[4] C.Li, N.Zhao, Y.Lu, C.Zhu, and J.Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in Proc.of the ACM international conference on Multimedia, pp.1856-1864,2017.

[5] C.Li, C. Zhu, J.Zhang, B.Luo, and J.Tang, "Learning local-global multi-graph descriptors for RGB-T object tracking," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 10, pp.2913–2926, 2019.

[6] C.Li, H.Cheng, S.Hu, X.Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," IEEE Transactions on Image Processing, vol. 25, no. 12, pp.5743–5756, 2016.

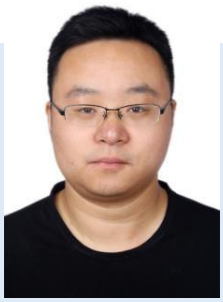
[7] X. Lan, M.Ye, R.Shao, B. Zhong, P . C. Y uen, and H. Zhou, "Learning modality-consistency feature templates: A robust rgb-infrared tracking system," IEEE Transactions on Industrial Electronics, vol. 66, no. 12, pp.9887–9897, 2019.

[8] Y . Zhu, C. Li, B. Luo, J. Tang, and X. Wang, "Dense feature aggregation and pruning for rgbt tracking," in Proc. of the ACM International Conference on Multimedia, pp.465-472,2019.

[9] C. Li, A. Lu, A. Zheng, Z. Tu, and J. Tang, "Multi-adapter rgbt tracking," in Proceedings of the IEEE International Conference on Computer Vision Workshops, pp.2262-2270,2019.

[10] Q.Wang, Z.Teng, J.Xing, J.Gao, and S.Maybank, "Learning attentions: residual attentional siamese network for high performance online visual tracking," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp.4854-4863,2018.

[11] Z. Zhu, Q. Wang, B. Li, W.Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in Proc. of the European Conference on Computer Vision, pp.101–117,2018.



周全

南京邮电大学副教授，硕士生导师。研究方向包括深度学习、模式识别和计算机视觉。江苏省"青蓝工程"青年骨干教师。中国计算机学会，图像图形学会高级会员，江苏省自动化学会模式识别专委会常务委员。IEEE 和 IAPR 高级成员。已主持国家自然科学基金，江苏省自然科学基金等 10 余项。已发表学术论文 70 余篇，包括 IEEE TIP、IEEE TITS、IEEE TMI、IEEE TNNLS 等。目前担任 70 多个 SCI 期刊审稿人，并担任 IEEE/SPIE ISAIR2019-2023、IEEE ICME2019 和 PRCV2022 区域主席。同时担任 Computer and Electrical Engineering 期刊编辑，以及 IEEE TMM、PR、MMTA 和 Visual Intelligence 等期刊的首席客座编辑。

Email: quan.zhou@njupt.edu.cn

康彬

南京邮电大学副教授，硕士生导师，一直从事计算机视觉、深度学习理论及应用研究工作，具体研究方向包括目标跟踪、细粒度识别及多模态信息融合等。主持参与了与深度学习应用相关多项科研项目，如：国家自然科学面上、青年基金以及军委科技委基础研发等。截至目前，共发表学术论文 40 余篇，其中发表的高水平论文包括 IEEE TIP、IEEE TNNLS、 IEEE TCSVT、IEEE TITS、AAAI 等。目前担任 IEEE TMM, IEEE SPL, IET Image Processing 等学术期刊审稿人，除此之外担任网络多媒体专委会秘书、IEEE ICC, GlobleCom 以及 WCSP 等通信领域知名国际会议技术委员会委员。

Email: kangb@njupt.edu.cn