HyperTea：一种基于超图的时序增强和对齐网络，用于运动红外小目标检测

Zhaoyuan Qi, Weihua Gao, Wenlong Niu, Jie Tang, Yun Li, Xiaodong Peng

***Abstract*—在实际应用场景中，由于目标体积小、强度弱以及运动模式复杂，移动红外小目标检测（MIRSTD）仍然十分具有挑战性。现有方法通常仅建模特征节点之间的低阶相关性，并在单一时间尺度内进行特征提取和增强。虽然超图在高阶相关学习中得到了广泛应用，但在 MIRSTD 中受到的关注有限。为了发掘超图的潜力并增强多时间尺度特征表示，我们提出了 HyperTea，该方法集成了全局和局部的时间视角，有效建模特征的高阶时空相关性。HyperTea 由三个模块组成：全局时间增强模块（GTEM）通过语义聚合和传播实现全局时间上下文增强；局部时间增强模块（LTEM）旨在捕获相邻帧之间的局部运动模式，然后增强局部时间上下文；此外，我们还开发了一个时间对齐模块（TAM）来解决潜在的跨尺度特征未对齐问题。据我们所知，HyperTea 是首个集成卷积神经网络（CNNs）、递归神经网络（RNNs）和超图神经网络（HGNNs）进行 MIRSTD 的工作，大大提升了检测性能。在 DAUB 和 IRDST 上的实验验证了其达到当前最佳水平（SOTA）的性能。我们的源代码在 https://github.com/Lurenjia-LRJ/HyperTea 上可用。**

***Index Terms*—Moving infrared small target detection (MIRSTD), hypergraph learning, temporal context enhancement, hypergraph-based temporal enhancement and alignment Network (HyperTea).**

I. 引言

红外小目标检测（IRSTD）已广泛应用于侦察、监视和救援等领域 [?] 。在许多实际场景中，无人机（UAVs）是主要的检测目标。然而，红外成像系统的固有特性给红外小目标检测带来了以下挑战：

- 小尺寸：较长的成像距离导致目标规模小、强度弱、对比度低以及信号与杂波比（SCR）低，使目标在复杂背景中容易被遮蔽。
- 空间分辨能力差：由于红外成像系统的固有特性，目标缺乏颜色和纹理特征，并且表现出模糊的边缘。这些因素显著降低了目标与杂波之间的空间分辨能力。
- 时空不稳定性：对于移动的红外小目标，由于距离变化、姿态调整或杂波的部分遮挡等因素驱动，其空间特征（例如，大小和强度）可能在连续帧中表现出动态变化。同时，它们的运动特征（例如，速度和轨迹）受到两个主要因素的影响：一个是目标的内在运动，另一个是成像平台的运动和视角变换。这两个因素共同导致了目标的非均匀和非线性运动模式，从而对特征表示学习提出了重大挑战。
- 目标-干扰时空耦合：背景干扰在两个关键维度上常常与目标具有惊人的相似性。空间上，它们在单帧内的特征与目标非常相似。时间上，它们的运动模式可能表现出与目标短期相似的特征，特别是在成像平台位置和方向变化时，目标和干扰相对于成像平台的相对运动变得高度类似。这种在单帧中的空间特征和连续帧间的运动趋势双重相似性导致了干扰和真实目标之间强烈的时空耦合。这样的耦合模糊了二者之间的辨别线索，从而显著增加了准确检测的难度。

为了应对这些挑战，已经提出了许多 IRSTD 方法，通常分为基于模型的方法和基于学习的方法。基于模型的方法利用从图像观察和分析中得出的先验知识来检测目标 [?], [?], [?] 。虽然在某些场景中表现良好 [?], [?] ，但它们对人工设计的特征提取器和超参数调优的高度依赖限制了适应性。

相比之下，基于学习的方法最近显示出卓越的检测性能和泛化能力 [?], [?], [?] 。这些方法主要依赖于卷积神经网络（CNNs） [?], [?] 、变压器网络（Transformers） [?], [?], [?] 、循环神经网络（RNNs） [?] 及其混合体。基于 CNN 的方法 [?], [?], [?] 可以有效提取局部特征，但受限于基于网格的表示，限制了其建模空间全局依赖性的能力。由于变压器通过将图像建模为完全连接的图来克服这一限制，一些方法使用它们来改善长距离建模 [?], [?], [?] 。然而，二次计算复杂度和由块嵌入引起的低级细节丢失可能限制其应用。基于 RNN 的方法由于其出色的序列依赖建模能力而获得越来越多的关注 [?], [?] 。尽管如此，它们在并行训练上存在困难，且计算成本较高，特别是对于长序列。最近，作为一种创新的基础网络架构，超图神经网络（Hypergraph Neural Networks, HGNNs）因其更高阶学习能力在一般目标检测中展现出巨大的潜力 [?], [?] ，但其在 IRSTD 中的应用尚待探索。

鉴于每种网络架构都有其内在的局限性和优势，而现有方法在跨不同时间尺度建模语义关联方面存在困难，我们自然想到通过利用网络的互补优势来探索一种新的 MIRSTD 范式，从而实现多时间尺度学习。

因此，一个关键问题是如何基于基本的网络架构构建检测流程。受 [?], [?] 的启发，我们的框架主要采用 RNN 来增强短期时间依赖性，因为它在建模序列依赖性方面表现优异，同时采用 CNN 来增强较长时间范围的时空上下文关系，因为它能够高效提取特征。HGNNs 被整合进来以赋予局部和全局特征更高阶的依赖表示能力。而 transformer 的核心组件——注意力机制，被用来实现跨时间的特征对齐。

为了充分利用各种基本网络架构的互补优势，并促进目标的跨时间特征表示学习，我们提出了一种基于超图的时间增强和对齐网络（HyperTea）。具体而言，我们的工作侧重于在全球和局部时间尺度上同时表示红外目标，特别是

Manuscript received ; revised . (Corresponding author: Wenlong Niu, Xiaodong Peng.)

Zhaoyuan Qi is with the Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: qizhaoyuan24@mails.ucas.ac.cn). Weihua Gao, Wenlong Niu, Jie Tang, Yun Li, and Xiaodong Peng are with the Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China (e-mail: gaoweihua22@mails.ucas.ac.cn, niuwenlong@nssc.ac.cn; tangjie@nssc.ac.cn; liyun02@nssc.ac.cn; Pxd@nssc.ac.cn)

www.xueshuxiangzi.com

利用超图来实现高阶相关性的学习。我们的 HyperTea 并不是简单的基本网络结构融合。相反，它捕捉高阶相关性，然后分别增强全局和局部时间上下文，同时还涉及跨尺度对齐。据我们所知，这是首次将 CNNs、RNNs 和 HGNNs 整合用于 MIRSTD 的工作。总之，我们工作的主要贡献总结如下。

1) 我们探索并提出一个开创性的三架构方案 HyperTea，以探索 MIRSTD 的多时间尺度特征表示学习，该方案首次将 CNNs、RNNs 和 HGNNs 集成到一个统一框架中。
2) 随着 HGNNs 首次应用于 IRSTD，我们提出两个模块分别增强全局和局部时间上下文。具体而言，基于 CNN 和 HGNN，开发了一个全局时间增强模块（GTEM），以从全局时间上下文中提取高阶语义关联并将其传播到每个帧以进行增强。此外，我们通过 HGNN 改进了 ConvLSTM，从而设计了局部时间增强模块（LTEM），以增强小目标的局部帧间关联。
3) 设计了一个时间对齐模块（TAM），以减少不同时间尺度之间特征的潜在错位，并放大目标与背景之间的语义差异。

II. 相关工作

A. 超图学习

如 [?], [?] 所示，超图在建模复杂的高阶关联方面展示了令人印象深刻的能力，并已应用于多个领域 [?], [?], [?] 。最近，通用超图神经网络 [?] 引入了一种空间方法，用于特征顶点之间的高阶信息传播，并实现了显著的检测性能。尽管有这些进展，超图学习在 IRSTD 中的应用仍有待探索。

B. 红外小目标检测

根据不同的驱动范式，IRSTD 方法可以分为基于模型的方法和基于学习的方法。基于模型的方法可以进一步分为基于滤波的方法 [?], [?] ，基于局部对比度的方法 [?], [?], [?], [?] ，以及基于低秩表示的方法 [?], [?] 。这些方法通常依赖于特定的假设、人工设计的特征提取器以及基于先验知识的超参数。这种依赖性往往限制了它们在复杂现实场景中的泛化能力 [?] 。相反，基于学习的方法由于其强大的特征表示和学习能力，能够实现更优越的检测性能。根据所用参考帧的数量，基于学习的方法可以进一步分为单帧和多帧方案。

作为一个重要的研究方向，单帧方案最近吸引了大量关注。基于 CNN 的方法通常建立在 U-Net [?] 和其变体之上，特征融合是代表性改进策略。例如，ACM [?] 通过不对称上下文调制实现不同层次的特征融合。DNANet [?] 通过高层和低层特征之间的逐步交互和增强实现更好的检测性能。为了实现多层次和多尺度表示学习，UIUNet [?] 将一个小型 U-Net 嵌入到一个较大的 U-Net 主干中。ISNet [?] 提取基础边缘特征，并通过跨层次特征融合提高检测性能。RDIAN [?] 利用多尺度卷积和方向引导的注意机制捕获更丰富的目标特征用于融合。最近，许多研究将 CNN 与 ViTs [?] 结合起来，以利用其在建模远程依赖中的显著优势。IR-TransDet [?] 探索全局图像、目标和邻近像素之间的关系。SCTransNet [?] 通过空间通道交叉变换器有效编码目标和背景之间的语义差异。受 ODE 启发，提出了 RKformer [?] 和 ABMNet [?] 。考虑到全监督的高标注成本，点监督方案 [?], [?], [?], [?], [?], [?] 和自监督方案 [?] 已成为可行的替代方案。最近，一些流行的网络架构如 Mamba [?] 、SAM [?] 也迁移到了 IRSTD。然而，缺乏运动信息限制了单帧方案在复杂动态背景下检测移动红外目标的能力。为了克服这个问题，最近提出了一些多帧策略。

根据主要网络架构，多帧方案可以进一步分为基于 CNN、CNN 与 transformer 结合以及 CNN 与 RNN 结合的方法。基于 CNN 的方案如 STDMANet [?] ，提出了时空差分多尺度注意力网络。DTUM [?] 使用方向编码卷积来提取目标运动特征。TMP [?] 并行提取时空特征并进行跨域特征融合。对于 CNN 与 transformer 结合的方法，ST-Trans [?] 通过改进的视频 Swin Transformer [?] 建模时空相关性。Tridos [?] 使用视频 Swin Transformer [?] 捕捉全局频域特征。为了利用 RNN 强大的时间建模能力，像 SSTNet [?] 这样的 CNN 与 RNN 方案基于改进的 ConvLSTM [?] 在跨切片运动背景下建模时空特征。此外，在弱监督学习方面，S2MVP [?] 通过使用少量标记的训练样本取得了显著的性能。为了利用语言提示中的先验知识，视觉语言检测框架 MoPKL [?] 正逐渐受到关注。

然而，目前的 IRSTD 方法尚未涉及跨时间尺度（局部-全局时间范围）上下文建模。此外，现有方法只能捕获特征节点之间的低阶相关性，而无法建模多个节点之间的相关性。如何建模节点之间的高阶相关性并利用它们增强运动上下文仍有待探索。

鉴于此，我们专注于跨时间尺度的目标特征增强和对齐，特别是使用 HGNNs 来学习高阶关联。为了直观和清晰，我们总结了 HyperTea 的独特优势如下：

与普通图相比，超图最重要的特点是超边可以与多个顶点关联，因此可以建模更复杂的高阶关系。一个超图可以定义为 $\mathcal{G}=\{\mathcal{V},\mathcal{E}\}$ ，其中 $\mathcal{V}$ 是顶点集，$\mathcal{E}$ 是超边集。通常，使用基于距离的方法来构建 $\mathcal{E}$ 。基于具有额外残差连接的空间域超图卷积，超图卷积单元（HCU）可以表示为：

$$\begin{aligned}\mathrm{HCU}(\boldsymbol{X}) &= \mathrm{HyperConv}(\boldsymbol{X},\boldsymbol{H}) \\ &= \boldsymbol{X}+\boldsymbol{D}_v^{-1}\boldsymbol{H}\boldsymbol{D}_e^{-1}\boldsymbol{H}^T\boldsymbol{X}\boldsymbol{\Theta}\end{aligned} \tag{1}$$

其中 H 表示 X 的关联矩阵，D_v 和 D_e 分别表示顶点和超边的对角度矩阵，Θ 是一个全连接层。

可以这样理解超图卷积。$\boldsymbol{H}^T$ 的乘法促进了从节点到超边的信息聚合，随后使用 $\boldsymbol{D}_e$ 进行归一化。随后，与 $\boldsymbol{H}$ 的相乘使信息从超边聚合回顶点，并通过 $\boldsymbol{D}_v$ 进行归一化。通过这个两阶段的信息传播，超图卷积可以建模顶点之间的高阶相关性。

我们旨在通过跨时间特征学习有效提高网络检测移动小目标的能力。为实现这一目标，我们提出了一个新的框架，如 ?? 所示。为了清晰起见，其简化工作流程在 ?? 中展示。

我们的 HyperTea 以 T 个连续帧 $\boldsymbol{I}_s=[\boldsymbol{I}_1,\boldsymbol{I}_2,\ldots,\boldsymbol{I}_T]$ 作为输入，并输出在关键帧 $\boldsymbol{I}_T$ 上的检测结果。按照 MIRSTD [?], [?], [?] 的常见范式，我们使用 Cross-Stage Partial Darknet (CSPDarknet) [?] 作为骨干网络。为了实现跨时间特征学习的策略，我们设计了三个关键模块：GTEM、LTEM 和 TAM。

对于管道，我们首先将每个帧输入具有共享权重的骨干网络，以获得多帧空间特征 $\boldsymbol{F}_s=\{\boldsymbol{F}_1,\boldsymbol{F}_2,\ldots,\boldsymbol{F}_T\}\in\mathbb{R}^{T\times C\times H\times W}$ ，其中 C 、H 和 W 分别表示特征的通道、高度和宽度。之后，GTEM 接收 $\boldsymbol{F}_s$ 作为输入，并输出全局时间上下文增强特征 $\boldsymbol{G}_{st}=\{\boldsymbol{G}_1,\boldsymbol{G}_2,\ldots,\boldsymbol{G}_T\}\in\mathbb{R}^{T\times C\times H\times W}$

www.xueshuxiangzi.com

。LTEM 也接收 $\boldsymbol{F}_s$ 作为输入，并输出局部时间上下文增强特征 $\boldsymbol{L}_{st} \in \mathbb{R}^{C\times H\times W}$ 。为了实现全局和局部时间上下文的聚合，我们随后将 $\boldsymbol{G}_{st}$ 和 $\boldsymbol{L}_{st}$ 输入 TAM，以减轻不同时间尺度上的错位。上述管道可以如下公式化:

$$\begin{cases} \boldsymbol{G}_{st} &= \text{GTEM}(\boldsymbol{F}_s) \\ \boldsymbol{L}_{st} &= \text{LTEM}(\boldsymbol{F}_s) \\ \boldsymbol{F}_{gl} &= \text{TAM}(\boldsymbol{G}_{st}, \boldsymbol{L}_{st}) \end{cases} \tag{2}$$

最后，全球本地对齐的时间特征 $\boldsymbol{F}_{gl}$ 被输入检测头以获得检测结果。

C. 全局时间增强模块

增强全球时序上下文通过帧间依赖仍然是 MIRSTD 中的一个挑战性问题。以前的方法通常利用注意力机制来建模空间依赖 [?] 或跨通道语义依赖 [?] 。然而，由于注意力机制本质上将图像建模为一个完全连接的图，有限的表示能力阻碍了高阶依赖学习。最近，全球-局部语义交互已经成为特征增强的有效范式 [?], [?] 。受到这些启发，我们提出如 ?? 所示的 GTEM，包括时序特征聚合、方向偏好卷积、高阶关联学习和全局语义散射机制。

具体来说，我们首先通过两个卷积层聚合全局时间特征。

$$\boldsymbol{L}_{gt} = \text{Conv}^2_{3\times3}(\text{Concat}[\boldsymbol{F}_1, ..., \boldsymbol{F}_T])) \tag{3}$$

，其中 $\text{Conv}^2_{3\times3}$ 表示两个 3×3 基本卷积组件，具有批归一化和一个激活层。

在实际中，背景杂波的特征通常表现为相应于物理对象的条状或条形图案。因此，我们设计了一个方向优先卷积块（DPCB）来将这些判别特征编码到通道维度中，以增强特征表示，如图所示 ?? 。在实现方面，我们分别使用核尺寸为 1×5 ，5×1 ，1×1 的卷积来提取具有不同方向偏好的特征。然后使用 3×3 个基础的深度卷积组件来增强局部空间特征并扩展感受野。之后我们应用 3×3 卷积来融合提取的特征并获得 D_{gt} 。

$$\begin{cases} \boldsymbol{L}_h = \text{DWC}(\text{Conv}_{1\times5}(\boldsymbol{L}_{gt})) \\ \boldsymbol{L}_v = \text{DWC}(\text{Conv}_{5\times1}(\boldsymbol{L}_{gt})) \\ \boldsymbol{L}_s = \text{DWC}(\text{Conv}_{1\times1}(\boldsymbol{L}_{gt})) \\ \boldsymbol{D}_{gt} = \text{Conv}_{3\times3}(\text{Concat}[\boldsymbol{L}_h, \boldsymbol{L}_v, \boldsymbol{L}_s]) + \boldsymbol{L}_{gt} \end{cases} \tag{4}$$

其中 DWC 代表 3×3 深度卷积，$\boldsymbol{L}_h, \boldsymbol{L}_v, \boldsymbol{L}_s$ 分别表示具有水平方向偏好、垂直方向偏好和无方向偏好的特征。

随后，HCU 接收 D_{gt} 作为其输入，并生成更高阶的语义增强特征 G_h 。然后，为了通过 G_h 增强全局时间上下文，我们通过具有共享权重的卷积层在每个时间步传播 G_h 。由此可以得到全局时间增强特征 G_{st} ，如下所示:

$$\boldsymbol{G}_h = \text{HCU}(\boldsymbol{D}_{gt}) \tag{5}$$

$$\boldsymbol{G}_i = \text{Conv}^2_{3\times3}(\text{Concat}[\boldsymbol{F}_i, \boldsymbol{G}_h]), i = 1, ..., T \tag{6}$$

为了在关键帧上实现稳健且准确的检测，我们认为有效地建模局部运动模式是至关重要的，特别是在处理动态复杂场景时。考虑到 convLSTM [?] 及其变体在时空表示学习中的强大能力，并结合 HGNNs 进行高阶学习的优势，我们设计了超图卷积 LSTM 及其基本单元（HCCell）。在 HCCell 中，可以在不同的门之间在一个共同的语义空间中传递高阶信息，这极大地增强了其局部时空特征学习能力。

具体而言，我们对 ConvLSTM 节点做出了以下改进。如 ?? 所示，我们首先通过 1×1 卷积将输入 $\mathcal{X}_t$ 和隐藏状态 $\mathcal{H}_{t-1}$ 投射到相同的语义空间。然后，我们通过 HCU 进行高阶语义增强以获得增强的特征 $\mathcal{L}_h$ 。随后，我们应用带有残差连接的卷积来实现对 $\mathcal{L}_h$ 的语义调制，并获得 $\mathcal{F}_t$ 。之后，我们沿通道维度拆分 $\mathcal{F}_t$ ，然后应用相应的激活函数，最终获得四个门。HCCell 的计算模型在数学上遵循以下公式:

$$\begin{cases} \mathcal{L}_h = \text{HCU}(\text{Conv}_{1\times1}([\mathcal{X}_t, \mathcal{H}_{t-1}])) \\ \mathcal{F}_t = \text{Conv}_{1\times1}(\mathcal{L}_h) + \mathcal{L}_h \\ i_t, f_t, o_t, g_t = \text{Act}(\text{Chunk}(\mathcal{F}_t)) \\ \mathcal{C}_t = f_t \circ \mathcal{C}_{t-1} + i_t \circ g_t \\ \mathcal{H}_t = o_t \circ \tanh(\mathcal{C}_t) \end{cases} \tag{7}$$

，其中 Act 使用 sigmoid 函数激活 i_t, f_t, o_t ，对 g_t 使用 tanh 函数，$\circ$ 表示 Hadamard（元素）积。

我们的方法利用不同的分支来探索红外目标的全局和局部时空特征表示。然而，目标与成像平台之间的相对运动以及学习路径差异常常导致全局和局部时空特征之间的特征不对齐。为了解决这个问题，我们提出了如 ?? 所示的 TAM，它由全局-局部时间注意力模块（GLTA）和通道空间聚合模块（CSAM）组成。

1) 全局-局部时序注意力：具体来说，我们首先对全局时序上下文增强特征 $\boldsymbol{G}_{st}$ ，局部时序上下文增强特征 $\boldsymbol{L}_{st}$ 进行补丁嵌入和层归一化处理。

$$\hat{\boldsymbol{G}}_{st} = \text{LN}(\text{EMD}(\boldsymbol{G}_{st})) \tag{8}$$

$$\hat{\boldsymbol{L}}_{st} = \text{LN}(\text{EMD}(\boldsymbol{L}_{st})) \tag{9}$$

随后，我们使用 1×1 卷积将 $\boldsymbol{G}_{st}$ 和 $\boldsymbol{L}_{st}$ 投射到具有相同维度的语义空间中。然后，为了融合不同时间尺度的特征，我们的 GLTA 使用 $\boldsymbol{L}_{st}$ 作为查询，$\boldsymbol{G}_{st}$ 作为键和值来实现细粒度的特征融合。

$$\boldsymbol{Q} = \text{Conv}_{1\times1}(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(\hat{\boldsymbol{L}}_{st}))) \tag{10}$$

$$\boldsymbol{K} = \text{Conv}_{1\times1}(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(\hat{\boldsymbol{G}}_{st}))) \tag{11}$$

$$\boldsymbol{V} = \text{Conv}_{1\times1}(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(\hat{\boldsymbol{G}}_{st}))) \tag{12}$$

$$Attn = \text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V} \tag{13}$$

，其中 $\sqrt{d}$ 是一个归一化的缩放因子。通过这种方式，任何两个像素都可以进行交互并捕获它们之间的跨时间依赖关系，而不仅仅局限于当前的时间尺度。

此外，为了保留更多的原始信息，我们还结合了残差块，将不同时间尺度的关键帧特征嵌入到查询结果 $\boldsymbol{R}$ 中:

$$\boldsymbol{R} = \text{LN}(\boldsymbol{G}_T + \text{UP}(\text{LN}(\text{Conv}_{1\times1}(Attn) + \hat{\boldsymbol{L}}_{st}))) \tag{14}$$

，其中 LN 表示层归一化，UP 表示重建层，由一个上采样层和一个卷积层组成。

2) 通道空间聚合模块：最近，利用通道和空间全局信息在特征增强中已成为一种常见做法。受 [?], [?] 的启发，我们设计了 CSAM 以进一步优化空间和通道权重的选择。

如图所示 Figure 1 ，给定输入张量 $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$ ，CSAM 首先通过 1×1 卷积调整通道维度，并采用 3×3 和 5×5 深度卷积以增强局部空间信息。然后，CSAM 将

www.xueshuxiangzi.com

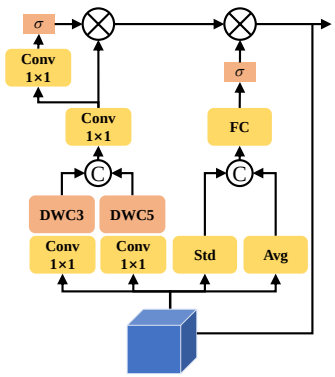

Fig. 1: 我们提出的 CSAM 的详细信息。

通道维度压缩到 1 以获得空间注意力图 $\boldsymbol{X}_{sa}$ 。上述过程可表示为:

$$\boldsymbol{X}_s = f_1^c([f_3^{dwc}(f_1^c(\boldsymbol{X})), f_5^{dwc}(f_1^c(\boldsymbol{X}))]) \tag{15}$$

$$\boldsymbol{X}_{sa} = f_{sa}^c(\boldsymbol{X}_s) \odot \boldsymbol{X}_s \tag{16}$$

其中 f_1^c 表示 1×1 卷积，f_3^{dwc} 和 f_5^{dwc} 分别代表 3×3 和 5×5 深度卷积，f_{sa}^c 代表输出通道为 1 且采用 sigmoid 激活函数的 1×1 卷积，$\odot$ 是广播的哈达玛积。

然后，为了更好地估计不同通道的权重，我们计算每个通道的最大值和标准差，进一步使用可学习参数来获得通道权重 $\boldsymbol{X}_{ca}$ ，如下所示:

$$\boldsymbol{X}_{ca} = fc([\max(\boldsymbol{X}), \mathrm{std}(\boldsymbol{X})]) \tag{17}$$

，其中 fc 表示具有 sigmoid 激活函数的线性变换层。最后，CSAM 通过整合互补的空间和通道信息以及残差拼接来丰富特征的表示。

$$\boldsymbol{X}_{sc} = \boldsymbol{X}_{ca} \odot \boldsymbol{X}_{sa} + \boldsymbol{X} \tag{18}$$

遵循一般检测范式 [?], [?] ，传统的损失函数可以定义如下:

$$\mathcal{L} = \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}} + \lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{obj}}\mathcal{L}_{\mathrm{obj}} \tag{19}$$

其中 $\mathcal{L}_{\mathrm{reg}}$ 是边界框回归损失，$\mathcal{L}_{\mathrm{cls}}$ 是分类损失，$\mathcal{L}_{\mathrm{obj}}$ 是目标概率损失。λ_{reg} 、λ_{cls} 、λ_{obj} 是平衡损失条款的超参数。对于 $\mathcal{L}_{\mathrm{cls}}$ 和 $\mathcal{L}_{\mathrm{obj}}$ ，我们使用二元交叉熵损失函数。超参数和 $\mathcal{L}_{\mathrm{reg}}$ 遵循 [?] 。

III. 实验

A. 数据集和评估指标

为了全面验证我们 HyperTea 的有效性和优越性，我们在两个数据集，DAUB [?] 和 IRDST [?] 上进行了实验。数据集的划分遵循 [?] 。

在评估指标方面，我们遵循目标检测范式中广泛采用的做法 [?] ，并采用精确度（Pr）、召回率（Re）、F1-得分以及在 IoU 阈值为 0.5 时的平均平均精度（mAP $_{50}$ ）。这些指标定义如下:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{20}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{21}$$

$$\mathrm{F1} = \frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}, \tag{22}$$

，其中 TP、FP 和 FN 分别表示真实正例（正确检测到的目标）、假正例（被错误识别为目标的背景区域）和假负例（遗漏的目标）的数量。F1-得分作为精确度和召回率的调和平均值，提供了一个平衡的措施，可以同时评估检测器在尽量减少漏检和误报方面的能力。

B. 实现细节

在所有实验中，输入帧分辨率被调整到 512 × 512。所有方法均在 50 个周期内训练，批量大小为 4。我们将初始学习率设置为 0.01，并采用动量为 0.937 的 SGD 作为优化器，权重衰减为 5×10^{-4} ，学习率衰减因子为 0.1。对于非极大值抑制，IoU 阈值设定为 0.65，而置信度阈值为 0.001。所有实验均在一台 Nvidia GeForce 4090D GPU 上进行。

C. 与其他方法的比较

1）定量比较：考虑到基于学习的方法在检测性能上通常比基于模型的方法优越得多，我们选择了一些最新的基于学习的 IRSTD 方法进行比较。对于基于分割的方法，我们遵循结合检测器的模式 [?], [?] 。

在 Table I 中清晰地展示了不同检测方法在两个数据集上的定量表现。根据 Table I 中的数据，可以识别出三个重要发现。首先，单帧方法的检测性能通常比多帧方法差；然而，当数据集中时间关联难以利用时，它们可能优于某些多帧方法。例如，在 DAUB 上，SOTA 单帧方法 MSHNet 实现了 85.25 % 的 mAP $_{50}$ 和 92.56 % 的 F1 分数，低于在 Table I 中列出的任何多帧方法。这是因为单帧方法仅依赖空间特征，从而限制了它们自适应捕捉目标时间动态的能力。然而，在 IRDST 上，领先的单帧方法 DNANet 达到了 71.39 % 的 mAP $_{50}$ 和 84.86 % 的 F1 分数，超过了多帧方法 TMP 和 SSTNet。这可能归因于 IRDST 中比 DAUB 更复杂的时间关联。为了量化帧间相似性，我们使用了均方误差（MSE）指标，计算出 DAUB 的平均 MSE 为 32.95，而 IRDST 为 112.03。IRDST 中更高的 MSE 表明时间关系的更大复杂性，这表明在这样的条件下，一些多帧方法有效利用时间关联能力的局限性可能导致其表现被单帧方法超越。

其次，我们的 Hypertea 在两个数据集上的大多数评估指标中都取得了最佳性能，尤其是在 mAP $_{50}$ 和 F1 得分方面。例如，在 DAUB 上，Hypertea 取得了最高的 mAP $_{50}$ 为 95.59 % 和最高的 F1 得分 98.23 %。此外，在 IRDST 上，我们的 Hypertea 仍然实现了最高的 mAP $_{50}$ 为 76.42 % 和 F1 得分 87.62 %，远远超过了之前的最先进（SOTA）方法 Tridos，其提供的 mAP $_{50}$ 为 72.54 %，F1 得分 85.63 %。

第三，我们的 HyperTea 展现出更强的时间特征提取能力，使其更适合复杂成像场景。由于 DAUB 中的样本在时间域上相对稳定，所有多帧方法在 DAUB 上的表现比在 IRDST 上更好。例如，之前的多帧 SOTA 方法 Tridos 在 DAUB 上达到 mAP $_{50}$ 为 91.52 % 和 F1 得分为 96.17 %，但在 IRDST 上仅达到 mAP $_{50}$ 为 72.54 % 和 F1 得分为 85.63 %。然而，在 DAUB 上，我们的 HyperTea 相比 Tridos 实现了 4.07 % 的 mAP $_{50}$ 提升和 2.06 % 的 F1 得分提升。在 IRDST 上，HyperTea 的提升仍然显著，相比 Tridos，mAP $_{50}$ 和 F1 得分分别提高了 3.88 % 和 1.99 %。这些对比表明，与其他方法相比，HyperTea 能够更有效地适应动态时间场景，并提取复杂的时间特征进行检测。

www.xueshuxiangzi.com

TABLE I: 在四个指标上与最新技术进行比较：mAP $_{50}$ 、精确度、召回率和 F1。

Scheme	Methods	Publication	IRDST				DAUB			
			mAP $_{50}$	Pr	Re	F1	mAP $_{50}$	Pr	Re	F1
single-frame	ACM [?]	WACV 2021	65.31	82.11	80.70	81.40	83.41	93.11	90.65	91.86
	ALC [?]	TGRS 2021	53.10	77.30	69.03	72.93	79.03	90.41	88.43	89.40
	ISNet [?]	CVPR 2022	66.30	84.24	79.80	81.96	83.02	93.52	89.99	91.72
	AGPCNet [?]	TAES 2023	67.10	85.40	79.74	82.47	82.43	90.17	92.62	91.38
	MTUNet [?]	TGRS 2023	67.20	86.35	78.70	82.35	77.21	93.40	83.42	88.13
	DNANet [?]	TIP 2023	71.39	88.36	81.63	84.86	84.44	97.37	87.34	92.08
	RDIANet [?]	TGRS 2023	68.70	85.29	81.86	83.54	80.72	88.29	92.95	90.56
	MSHNet [?]	CVPR 2024	71.04	87.92	81.89	84.80	85.25	99.30	86.67	92.56
	SCTransNet [?]	TGRS 2024	68.41	82.93	83.08	83.01	79.97	94.34	85.99	89.97
	RPCANet [?]	WACV 2024	68.80	84.41	82.29	83.33	78.90	82.85	96.93	89.34
	SeRankDet [?]	TGRS 2024	69.50	86.16	81.38	83.70	84.67	94.60	90.47	92.49
multi-frame	DTUM [?]	TNNLS 2023	71.80	85.12	85.36	85.24	90.46	97.86	93.58	95.67
	SSTNet [?]	TGRS 2024	70.91	87.73	81.98	84.76	89.07	97.12	92.72	94.87
	TMP [?]	ESWA 2024	69.12	86.57	80.99	83.68	85.37	94.23	91.74	92.96
	Tridos [?]	TGRS 2024	72.54	87.82	83.63	85.63	91.52	96.83	95.52	96.17
	HyperTea(Ours)	-	76.42	91.18	84.33	87.62	95.59	98.14	98.31	98.23

2) 可视化比较: 为了直观地展示不同方法的检测性能，我们选择了三个具有代表性的场景进行演示，如 Figures 2 to 4 所示。可以观察到，我们的 HyperTea 始终能够准确地检测到移动的红外小目标。相比之下，其他方法通常会遭遇漏检或误报的问题。

例如，在 Figure 2 上，在 IRDST 数据集上，我们的 hypertea 可以在复杂场景中精确检测目标，即使存在大量云杂波。相比之下，许多其他方法如 ACM、DNANet、RDIAN、SCTransNet、DTUM、SSTNet 和 TMP 则倾向于将杂波误分类为目标。此外，在 Figure 3 中，当目标被云杂波包围时，一些方法无法避免杂波干扰，并错误地将杂波识别为目标，包括 DNANet、RDIAN、SCTransNet、DTUM 和 TMP。同时，如在 Figure 4 中所示，在 DAUB 数据集上，一些方法难以在复杂背景中检测极小目标，例如 ISNet、MSHNet、SeRankDet 和 TMP。然而其他方法则导致误报，包括 ACM、AGPCNet、DTUM 和 Tridos。综上所述，这些典型场景的可视化结果与 Table I 中呈现的定量性能高度一致，从而证明了我们的 Hypertea 的优越性。

3) PR 曲线比较：为了直观地评估不同方法的整体性能，我们在 DAUB 和 IRDST 上绘制了两条 PR 曲线，如 Figure 5 所示。可以很容易地观察到，我们的 Hypertea 的 PR 曲线几乎包络了两个数据集上其他方法的曲线。在 DAUB 上，Hypertea 的曲线在竞争对手之上始终占据右上区域；显著的是，即使在相对较高的召回率下，它仍然保持高精度。在 IRDST 上，Hypertea 的 PR 曲线几乎总是位于其他方法的曲线之上。在 PR 曲线分析中，曲线越靠近右上角，表明检测的精度和召回率之间的平衡越优越。因此，这两组 PR 曲线共同表明，与其他方法相比，我们的 Hypertea 实现了最佳的整体检测性能。

4) 模型复杂度比较：我们在 Table II 中将 Hypertea 的模型复杂度与 15 种代表性方法进行了比较，重点关注参数数量（Params）、GFlops 和每秒帧数（FPS）。从表格中可以得出两个关键观察结果。

其一是，Hypertea 在参数和 GFlops 上略有增加。在多帧方法中，DTUM 的参数最少，为 2.79M，而 Hypertea 为 13.75M，低于 TMP (16.41M) 和 Tridos (14.13M)。在 GFlops 方面，TMP 在多帧方法中取得了最佳结果，为 92.85 %，而 Hypertea 为 148.66。一个可能的原因是 TMP 基于单一时间尺度，而 Hypertea 跨局部和全局时间尺度学习特征表示，并需要跨尺度对齐，这需要额外的计算。然而，考虑到 Hypertea 显著的性能优势，这些复杂性成本是可以接受且值得的。

另一个观察是，Hypertea 在性能和 FPS 之间实现了出色的平衡。由于基于时空域的多帧方法通常在性能上优于单帧方法，但其 FPS 通常较低。例如，DNANet 是目前最先进的单帧方法，其 F1 得分为 84.86 % ——与 SSTNet 的 84.76 % 大致相当，但远低于 Hypertea 的 87.62 % 。并且其 FPS 为 12.13 也低于 Hypertea 的 14.52。此外，在多帧方法中，Hypertea 的 FPS 最高，甚至超过了一些单帧方法，如 DNANet 和 SCTransNet。这突显了其在检测性能和推理效率之间的优越平衡。

TABLE II: 在 IRDST 上的推理复杂性比较。

Methods	Frames	mAP $_{50}$ ↑	F1 ↑	Params ↓	GFlops ↓	FPS ↑
ACM [?]	1	65.31	81.40	2.89M	24.14	51.55
ALC [?]	1	53.10	72.93	2.92M	23.95	56.23
AGPCNet [?]	1	67.10	82.47	14.85M	366.37	15.72
MTUNet [?]	1	67.20	82.35	9.13M	67.42	48.79
ISNet [?]	1	66.30	81.96	3.46M	265.87	26.82
DNANet [?]	1	71.39	84.86	7.19M	135.00	12.13
RDIANet [?]	1	68.70	83.54	2.71M	50.67	45.64
MSHNet [?]	1	71.04	84.80	6.55M	69.78	25.73
SCTransNet [?]	1	68.41	83.01	35.74M	147.02	14.22
RPCANet [?]	1	68.80	83.33	3.17M	377.48	26.40
SeRankDet [?]	1	69.50	83.70	111.37M	1147	18.54
DTUM [?]	5	71.80	85.24	2.79M	103.73	12.85
SSTNet [?]	5	70.91	84.76	11.95M	123.60	12.97
TMP [?]	5	69.12	83.68	16.41M	92.85	14.02
Tridos [?]	5	72.54	85.63	14.13M	130.72	13.24
HyperTea1(Ours)	5	76.42	87.62	13.75M	148.66	14.52

D. 消融研究

1) 不同组件的效果：为了研究 HyperTea 中每个组件的有效性，我们在 DAUB 和 IRDST 上进行了系列消融研究，如 Table III 所示。对于 TAM 的消融实验，我们使用 1×1 卷积来调整通道数量，旨在验证模块的有效性，同时尽量减少对网络结构的改变。

从 Table III 中可以得出两个关键发现。首先，每个组件在一定程度上都有助于提高检测性能，并且当所有组件一起应用时能够达到最佳性能。例如，在 IRDST 上，没有任何组件的基线仅能达到 66.80 % 的 mAP $_{50}$ 和 81.94 % 的 F1 分数。单独应用 LTEM 后，mAP $_{50}$ 提升至 69.42 % ，F1 分数提高到 83.55 % 。单独应用 GTEM 时，mAP $_{50}$ 和 F1 分数进一步分别提升到 71.44 % 和 84.99 % 。

www.xueshuxiangzi.com

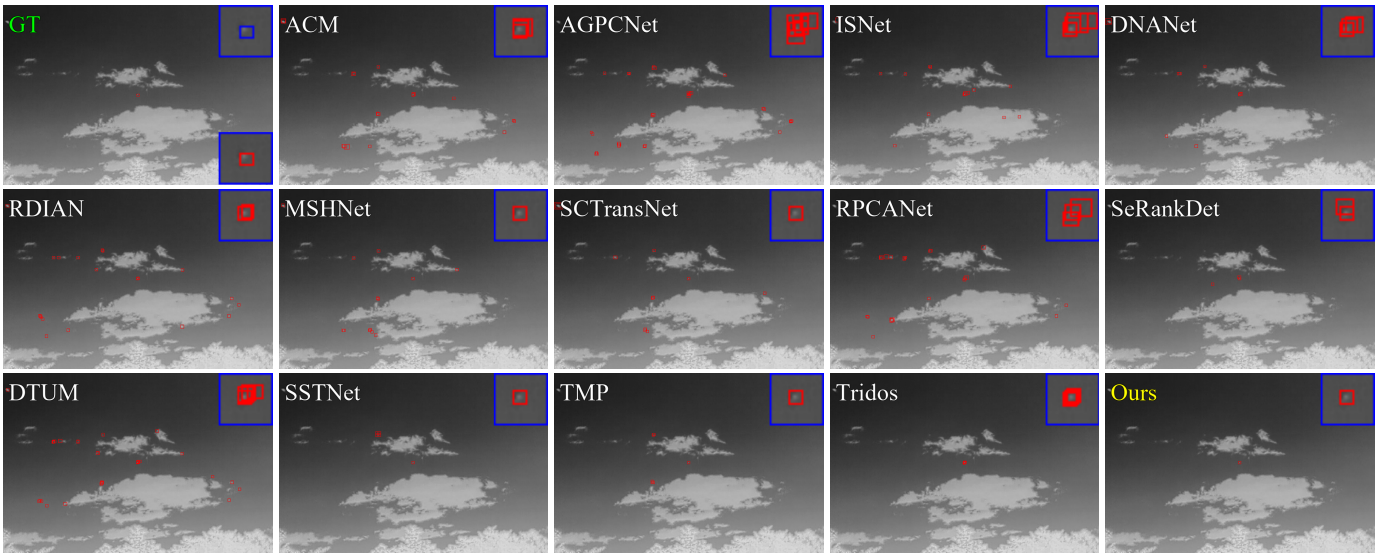

Fig. 2: 对 IRDST 进行 14 种方法的可视化比较，使用 72/266.bmp。GT 是指真实值。红色和蓝色框分别表示检测到的目标和放大的检测区域。

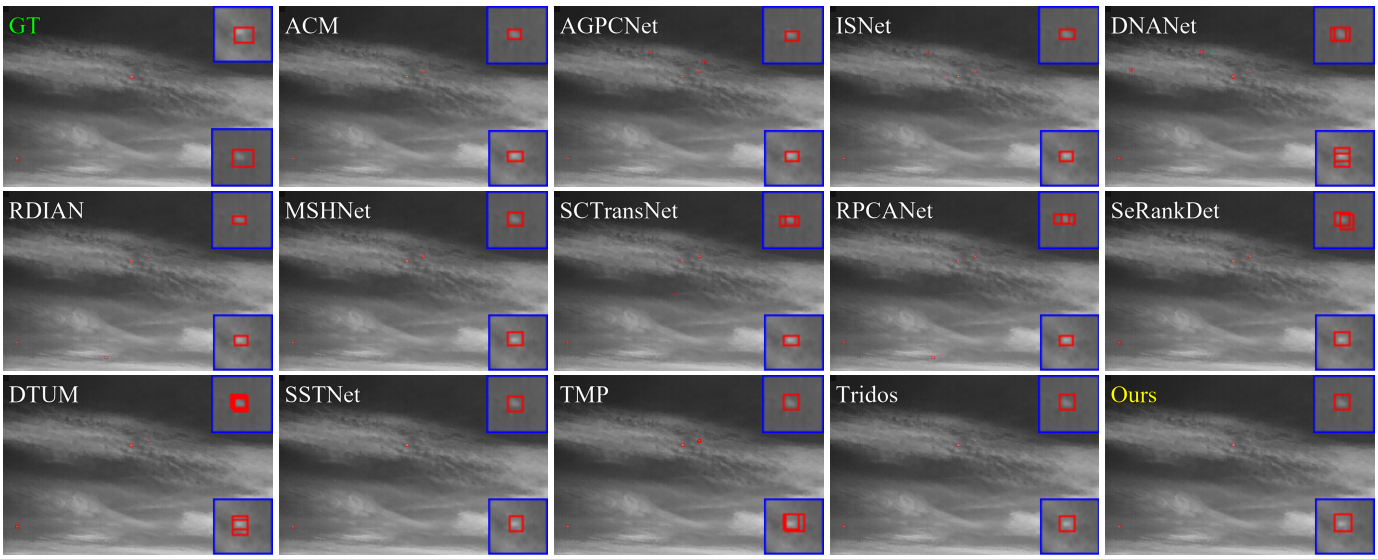

Fig. 3: 在 IRDST 上对 14 种方法进行了可视化比较，使用的是 6/14.bmp。GT 是地面实况。红色和蓝色的框分别表示检测到的目标和放大的检测区域。

此外，TAM 的性能随着不同输入而变化。当同时输入由 LTEM 和 GTEM 增强的特征时，DAUB 上的 mAP $_{50}$ 和 F1 分数分别上升至 95.59 % 和 98.23 % ；在 IRDST 上，它们达到 76.42 % 和 87.62 % ，实现了最佳性能。其次，简单的 1×1 卷积无法有效实现跨时间特征的融合。例如，在 IRDST 上，未使用 TAM 的模型仅能达到 69.07 % 的 mAP $_{50}$ 和 83.25 % 的 F1 分数，这低于单独使用 LTEM 的模型（mAP $_{50}$ ： 69.42 % ， F1： 83.55 % ）和单独使用 GTEM 的模型（mAP $_{50}$ ： 71.44 % ， F1： 84.99 % ）。

2) DPCB 和 HCU 在 GTEM 中的效果: 为了探讨 HCU 和 DPCB 对 GTEM 的潜在贡献，我们进行了一组消融实验，如 Table IV 所示。首先，DPCB 通过融合多方向特征提供了更稳健的特征表示，从而提高了检测性能。例如，在 DAUB 上，使用 DPCB 可以将 mAP 从 91.06 % 提高到 92.52 %，F1 从 95.91 % 提高到 96.68 %。同样，在 IRDST 上，HCU 可以将 mAP 从 69.67 % 提高到 71.84 %，F1 从 83.75 % 提高到 85.11 %。这些效果可以归因于 HCU 在建模高阶相关性方面的强大能力。

3) GLTA 和 CSAM 在 TAM 中的影响: 此外，我们研究 GLTA 和 CSAM 在 **??** 中的影响，从中可以得出三个不同的结论。首先，GLTA 有效地对齐了 LTEM 和 GTEM 增强的特征。例如，在 DAUB 上，使用 GLTA，将 mAP 从 89.74 % 提高到 92.13 %，F1 值从 95.19 % 提高到 96.61 %。其次，CSAM 可以进一步增强对齐的特征。例如，在 IRDST 上，使用 CSAM 将 mAP 从 73.30 % 提高到 76.42 %，并将 F1 值从 85.85 % 提高到 87.62 %。第三，单独使用 CSAM 获得的性能提升是有限的。例如，在 DAUB 上，单独使用 CSAM 仅使 mAP 从 89.74 % 提高到 89.81 %，F1 值从 95.19 % 提高到 95.23 %。这一现象也反映了不同时间尺度特征之间的失配对检测性能的严重影响。

www.xueshuxiangzi.com

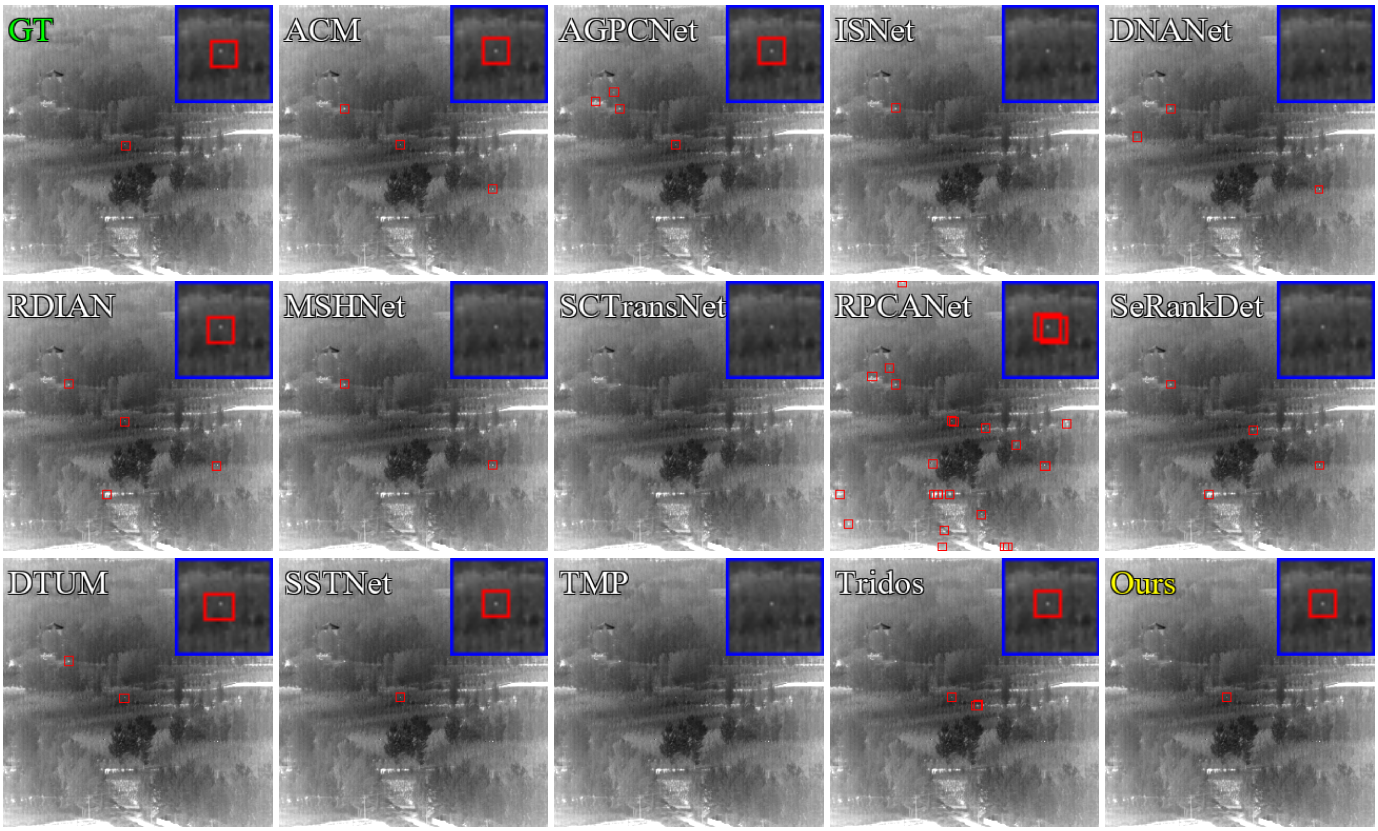

Fig. 4: 在 DAUB 上对 14 种方法的可视化比较，使用 21/433.bmp。GT 是地面实况。红色和蓝色框分别表示检测到的目标和放大的检测区域。

TABLE III: 不同组件的消融研究。TAM(A) 使用 $\boldsymbol{L}_{st}$ 作为查询，并使用 $\boldsymbol{F}_{s}$ 作为键 & 值。TAM(B) 使用 $\boldsymbol{F}_{T}$ 作为查询，并使用 $\boldsymbol{G}_{st}$ 作为键 & 值。

Settings	LTEM	GTEM	TAM	IRDST				DAUB			
				mAP $_{50}$	Pr	Re	F1	mAP $_{50}$	Pr	Re	F1
w/o All	-	-	-	66.80	84.95	79.13	81.94	82.63	95.51	87.32	91.23
w/ LTEM	✓	-	-	69.42	88.01	79.52	83.55	89.05	94.47	95.95	95.21
w/ GTEM	-	✓	-	71.44	88.28	81.93	84.99	89.79	96.97	93.82	95.37
w/ LTEM & GTEM	✓	✓	-	69.07	86.67	80.09	83.25	89.74	97.84	92.67	95.19
w/ TAM(A)	✓	-	✓	72.18	89.25	81.66	85.28	90.92	96.18	95.45	95.81
w/ TAM(B)	-	✓	✓	73.27	89.49	82.52	85.86	91.05	95.53	96.66	96.09
Ours	✓	✓	✓	76.42	91.18	84.33	87.62	95.59	98.14	98.31	98.23

TABLE IV: Ablation study on DPCB and HCU of GTEM.

Settings	IRDST				DAUB			
	mAP $_{50}$	Pr	Re	F1	mAP $_{50}$	Pr	Re	F1
GTEM w/o All	69.67	87.15	80.61	83.75	91.06	97.22	94.64	95.91
GTEM w HCU	71.84	88.76	81.76	85.11	92.33	97.59	95.55	96.56
GTEM w DPCB	71.99	89.07	81.78	85.27	92.52	96.13	97.24	96.68
GTEM	76.42	91.18	84.33	87.62	95.59	98.14	98.31	98.23

4) 层数和补丁大小在 LTEM 中的影响：如 Table V 所示，我们设计了六组实验来评估层数 (L) 和补丁大小 (ps) 对 LTEM 的影响。从这些实验中，我们可以得出以下三个观察结果。首先，当 L = 1 且 ps = 2 时，在 DAUB 和 IRDST 上的 mAP 分别可以达到 95.59 % 和 76.42 % 的峰值。在实验中，任何其他设置通常都无法使 HyperTea 达到最佳性能。其次，当层数增加时，HyperTea 在 DAUB 和 IRDST 上的表现明显下降。例如，在 DAUB 上，当 L 从 1 增加到 2 时，mAP $_{50}$ 从 95.59 % 下降到 88.07 %。这种现象的一个可能原因是，当层数大于 1 时，HyperTea 受到过拟合的影响。第三，当 ps 过小或过大时，性能下降。例如，在 DAUB 上，L = 1，当 ps 从 2 减少到 1 时，mAP $_{50}$ 从 95.59 % 下降到 92.31 %；当 ps 从 2 增加到 4 时，mAP $_{50}$ 从 95.59 % 下降到 93.04 %。这种现象的一个可能原因是，当 ps = 1 时，单个像素的语义特征不足以有效地建模长程依赖；当 ps = 4 时，过大的尺寸导致了低层次细节的丢失。只有当 ps = 2 时，才能在丰富特征表示的同时很好地保留低层次细节，从而实现最佳性能。

5) 超图距离阈值的影响：我们进一步进行了消融实验，以研究超图构建中使用的距离阈值的影响，结果如 Table VI 所示。可以观察到，当阈值为 8 时，DAUB 和 IRDST 上的 mAP $_{50}$ 分别达到峰值 95.59 % 和 76.42 %，过低和过高的阈值都会导致性能下降。这是因为较高的阈值导致更密集连接的超图，其中噪声节点被混入超边，可能导致特征缺乏判别力。相反，较低的阈值导致超图更稀疏，无法充分利用超图的高阶学习能力。因此，我们的 HyperTea

www.xueshuxiangzi.com

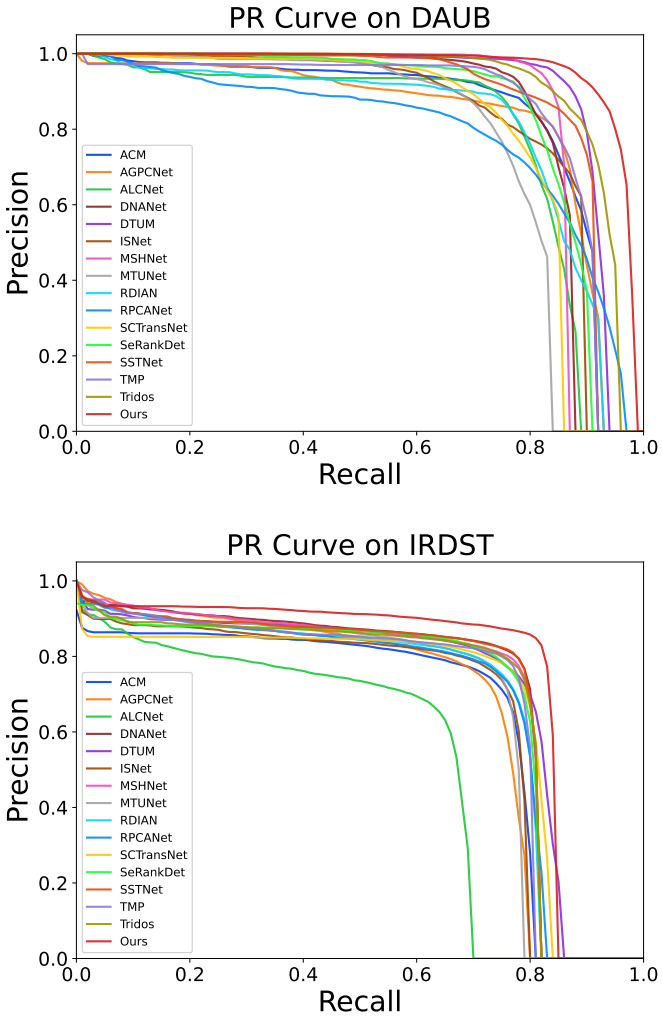

Fig. 5: 16 种代表性检测方法在 DAUB 和 IRDST 数据集上的 PR 曲线。

TABLE V: With different layer number L and patch size ps of LTEM.

L	ps	IRDST				DAUB			
		mAP $_{50}$	Pr	Re	F1	mAP $_{50}$	Pr	Re	F1
1	1	71.83	88.78	81.74	85.11	92.31	95.90	97.64	96.76
1	2	76.42	91.18	84.33	87.62	95.59	98.14	98.31	98.23
1	4	69.80	88.54	79.66	83.87	93.04	96.67	97.22	96.94
2	1	70.27	87.96	80.84	84.25	89.68	93.22	97.41	95.27
2	2	71.69	88.70	81.49	84.94	88.07	92.47	96.28	94.34
2	4	71.06	89.10	80.20	84.41	92.34	95.95	97.83	96.88

TABLE VI: With different hypergraph distance threshold.

Settings	IRDST				DAUB			
	mAP $_{50}$	Pr	Re	F1	mAP $_{50}$	Pr	Re	F1
6	72.14	88.36	82.24	85.19	93.25	95.99	98.22	97.09
7	72.96	89.01	82.85	85.82	92.86	96.36	97.93	97.14
8	76.42	91.18	84.33	87.62	95.59	98.14	98.31	98.23
9	71.15	87.99	81.66	84.71	88.54	92.09	97.62	94.78
10	70.31	87.04	81.32	84.08	92.33	94.99	98.22	96.58

是使用距离阈值 8 构建的，其确切值是根据经验和实验结果确定的。

6) 时间窗口大小 T 的影响: 我们进行了多组实验以不同的时间窗口大小来研究时间窗口 T 对检测性能的影响，如 Figure 6 所示。

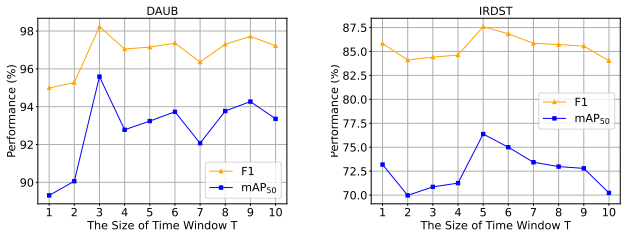

Fig. 6: 时间窗口大小 T 对 HyperTea 的影响。

首先，可以观察到 T 对检测性能的影响在 DAUB 和 IRDST 之间有所不同。在 DAUB 上，HyperTea 在 T 为 = 3 时实现了峰值 mAP $_{50}$ 和 F1 值。然而，在 IRDST 上，峰值 mAP $_{50}$ 和 F1 值在 T 为 = 5 时达到。这两个峰值对应的 T 值差异主要是由于数据集之间的差异。如前所述，DAUB 和 IRDST 的平均 MSE 值分别为 32.95 和 112.03，意味着 IRDST 的时间动态更加复杂，使得学习其时间特征更具挑战性，因此需要更多的帧。相比之下，DAUB 的时间动态相对稳定，因此由 3 帧提供的时间上下文就足以进行检测。

此外，适当的 T 对 IRSTD 至关重要。例如，在 DAUB 上，当 T $\leq$ 3 时，增加 T 会显著改善性能。当 T $>$ 3 时，检测性能先降低然后波动，呈现出整体趋于稳定的趋势。这与 DAUB 的时间动态特性一致：对于较小的 MSE，3 帧对于检测已经足够，较长时间的上下文提供的收益有限，甚至可能因为特征冗余和杂波干扰而导致性能下降。

7) 与 Hyper-YOLO 的比较：为了进一步验证我们 HyperTea 的优势，我们将其与 Hyper-YOLO 进行比较，如表 VII 所示。可以观察到，我们的 HyperTea 在两个数据集上都显著领先于 Hyper-YOLO。例如，在 DAUB 数据集上，HyperTea 达到 mAP $_{50}$ 为 95.59 %，精度（Pr）为 98.14 %，召回率（Re）为 98.31 %，F1 值为 98.23 %，这些数据远优于 Hyper-YOLO。这组量化比较表明，我们提出的 HyperTea 的主要优势在于充分利用基本网络结构之间的互补优势，并在不同的时间尺度上进行特征增强和对齐，而不仅仅是简单地将超图模块集成到检测框架中。

TABLE VII: PERFORMANCE COMPARISONS WITH Hyper-YOLO.

Methods	IRDST				DAUB			
	mAP $_{50}$	Pr	Re	F1	mAP $_{50}$	Pr	Re	F1
Hyper-YOLO	69.90	86.67	81.59	84.05	88.44	96.45	92.74	94.56
Ours	76.42	91.18	84.33	87.62	95.59	98.14	98.31	98.23

IV. 结论

为了通过跨时空表示学习移动红外微小目标来提高性能，我们提出了 HyperTea 网络架构，该架构结合了 CNN、RNN 和 HGNN。我们首先设计了一种 GTEM 来增强全局时间背景。然后，我们开发了 LTEM 来增强关键帧所在的局部时间背景。此外，我们利用 HGNN 赋予 GTEM 和 LTEM 更高阶学习能力。为了解决潜在的特征错位问题，我们设计了 TAM 来促进不同时间尺度特征的融合。对 DAUB 和 IRDST 的比较实验显示了我们的 HyperTea 的明显优越性。在大多数指标上，它优于现有的最先进的（SOTA）方法，具有适度的计算复杂度和 FPS。消融研究进一步表明，所有设计的组件都对提高检测性能有显著

www.xueshuxiangzi.com

贡献。在实验中，我们还发现数据集的时间相似性（通过 MSE 测量）对检测性能有着关键影响，这应为使用多帧输入的采样策略和模块设计提供见解。未来，设计更有效的时间特征采样、捕获和增强方案值得持续探索。

www.xueshuxiangzi.com