SeqCSIST: 序列空间邻近红外小目标解混

翟曦盟 [1], **徐博涵** [2], **陈耀弘** [1], **王浩** [1], **郭克华** [3], **戴一冕** [4,5†]

[1] 西安光学精密机械研究所, 中国科学院大学, [2] 信息科学与工程学院, 河南工业大学, [3] 计算机科学与工程学院, 中南大学, [4]PCA Lab, VCIP, 计算机学院, 南开大学, [5]NKIARI, 福田, 深圳
[†] 通讯作者

由于光学透镜焦距和红外探测器分辨率的限制，远距离空间邻近红外小目标群在红外图像中通常以混叠光斑的形式出现。本文提出了一个新颖的任务--序列空间邻近红外小目标解混，旨在从高密度空间邻近红外小目标群中以亚像素定位的形式检测出所有目标。然而，实现如此精确的检测是一个极其困难的挑战。此外，缺乏高质量的开源数据集也限制了研究的发展。为此，我们贡献了一个开源生态系统，包括一个名为 SeqCSIST 的具有序列基准的数据集，和一个为这项特殊任务提供客观评估指标的工具包，同时还有 23 种相关方法的实现。我们还提出了可变形细化网络（DeRefNet），这是一个模型驱动的深度学习框架，它引入了时间可变形特征对齐（TDFA）模块，实现了自适应的帧间信息聚合。据我们所知，这项工作是首次在多帧范式内解决空间邻近红外小目标解混任务的尝试。在 SeqCSIST 数据集上的实验表明，我们的方法性能优于当前最先进的方法，平均精度均值（mAP）指标提升了 5.3%。我们的数据集和工具包可在 https://github.com/GrokCV/SeqCSIST 上获取。

日期: 8.30.2025
项目主页: https://github.com/GrokCV/SeqCSIST

GrokCV

1 引言

红外成像不受光照变化的影响，即使在具有挑战性的低光照条件下，也能高度可靠地捕捉关键场景和目标，这种独特的能力确保了在不同环境中的一致性能 Wang et al. (2024)。因此，红外成像在军事和民用领域中发挥着至关重要的作用，例如监视、边境保护和搜救行动。然而，由于预警探测系统的要求以及红外成像系统的分辨率限制，远距离目标往往以小目标的形式出现在视场中。因此，能够进行远距离红外小目标的精确探测是这些应用成功的关键 Dai et al. (2024)。

通常，红外小目标检测（IRSTD）专注于在广域范围内搜索和探测目标，但存在几个严重的局限：

1. **缺乏固有特征**：由于传感器与目标之间的距离较远，后者通常呈现为一个低对比度的实体 Zhang et al. (2022b)。缺乏纹理或形状等独特特征，进一步使检测过程复杂化。
2. **密集目标的混叠**：红外图像中密集排列的目标可能会导致信号混叠，使检测算法难以区分和识别单个目标。

为了解决这些挑战，深度学习方法 Dai et al. (2021) 利用高维空间中的端到端特征提取，系统地提高了检测精度和操作鲁棒性。这使得研究重点聚焦于多尺度特征融合，弥合了详细目标表示与上下文信息之间的差距。早期的工作，如密集嵌套注意力网络（DNANet）Li et al. (2022)，通过利用密集的跨层语义交互，实现了多尺度特征提取和精细的细节保留。此后的工作中，无论是通过语义分割中的多尺度渐进式融合，如 U-Net in U-Net（UIU-Net）框架 Wu et al. (2022)，还是基于回归的检测范式，如局部到全局融合网络（LoGoNet）Li et al. (2023b)，这两类方法都通过利用多尺度融合与任务特定的结构优化来增强检测性能。

尽管上述红外小目标检测方法取得了一定的成效并推动了该领域的发展，但它们的应用主要**局限于大视场且目标稀疏的场景**。这些传统红外小目标检测方法背后的假设是，图像中检测到的目标与其对应的现实世界实体之间存在精确的一对一映射，其表现形式可由图 1 的第一行给出。

如图 2 所示，由于传感器分辨率的限制，当目标的空间距离越来越近时，它们的能量在红外成像过程中逐渐发生混叠，**不可避免地将相邻目标叠加到同一个像素上**，导致在成像平面上形成模糊的光斑 Lin et al. (2011)。在这种情况下，上述红外小目标检测方法无法达到预期的检测性能。此外，以往的检测方法侧重于识别目标轮廓，而点目标通常是没有轮廓的。即使检测到密集目标的斑点轮廓，也无法确定目标的确切数量和位置。

为了填补该领域的关键空白，我们提出了**一个新颖的任务**：空间邻近红外小目标解混（Closely-Spaced Infrared Small Targe Unmixing）。必须强调的是，这项任务在性质和范围上与传统意义上的红外小目标检测有着本质上的不同，它是**红外小目标检测的下**

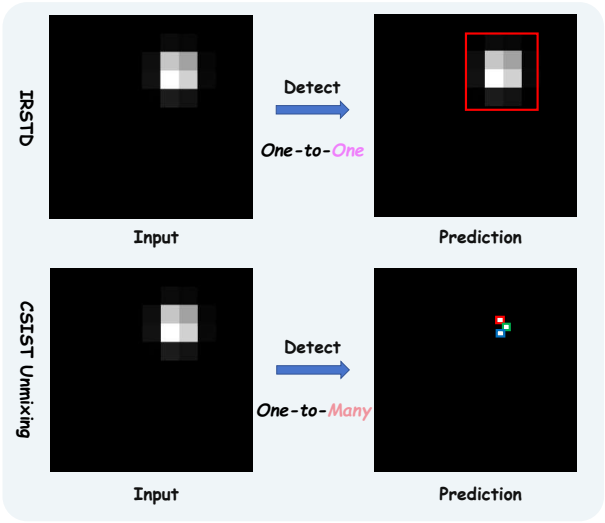

图 1 第一行说明了红外小目标检测方法假设检测到的目标与真实世界物体之间存在一一对应的关系。当待识别的图像中只有一个光斑时，检测结果对应于单个目标。相比之下，第二行所示的空间邻近红外小目标解混执行了更精细的检测，能够实现亚像素定位，并对光斑内潜在的子目标进行解混。

游任务。虽然红外小目标检测为小目标检测提供了一个框架，但它在处理涉及高密度目标的复杂场景时存在缺陷。空间邻近红外小目标解混通过明确解决此类场景，将这一挑战推向了更深层次，为在实际应用中实现更准确、更可靠的检测铺平了道路。与传统意义上的检测任务关注目标上下文和特征不同，这项新任务的特点如下：

1. **处理密集目标场景**：红外小目标检测专注于检测红外图像中稀疏的、孤立的目标。相比之下，空间邻近往外小目标解混处理的是具有紧密排列目标和能量混叠的场景，专注于解析视觉上无法区分的目标群。

2. **亚像素检测精度**：如图 1所示，与以边界框形式标记目标的红外小目标检测不同，空间邻近红外小目标解混将像素空间中的模糊光斑分解为不同的子目标，将任务从简单的检测转变为精确的亚像素定位。

这使得空间邻近红外小目标解混成为一项比传统红外小目标检测更具挑战性的任务。相应地，一个问题出现了：我们能否通过时空信息融合实现更好的空间邻近红外小目标解混效果？

我们的答案是肯定的。在本文中，我们首先构建了一个具有序列基准的数据集 SeqCSIST，这是我们开源生态系统的基础。SeqCSIST 数据集 5,000 个随机轨迹共 100,000 帧图像。每个轨迹包含 20 个连续的目标切片，大小为 11×11，其中混叠的目标数量从 2 到 4 不等。值得注意的是，SeqCSIST 是第一个专门用于序列空间邻近红外小目标解混研究的数据集。除了该数据集，我们还发布了一个工具包，作为我们开源生态系统的第二个板块。该工具包包括一个用于客观评估空间邻近红外小目标解混能力的指标，以及 23 种相关方法的实现。

为了应对这项新任务的固有挑战，我们提出了**可变形细化网络**（**De**formable **Ref**inement **Net**work，**DeRefNet**）来实现有效的序列空间邻近红外小目标解混。该网络由三个主要模块组成：一个稀疏驱动的特征提取模块、一个位置编码模块和一个时间可变形特征对齐模块。与依赖通用 ResNet 骨干网络进行特征提取的传统方法不同，DeRefNet 充分考虑了目标的稀疏性先验，并通过非线性可学习和稀疏化变换实现了空间邻近红外小目标特征的有效提取。随后，为了实现更精细的亚像素目标定位，网络采用了位置编码模块来增强时间信息。最后，通过时间可变形特征对齐模块，在特征级别上进行多帧可变形对齐，实现了中间帧基于参考帧的动态细化，使得网络无需显式的运动估计和图像配准操作。值得注意的是，DeRefNet 是将多帧方法引入空间邻近红外小目标解混领域的开创性尝试。

本文的主要贡献可总结如下：

1. **新颖的任务**：我们提出了一项名为序列空间邻近红外小目标解混的新任务，它拓宽了红外小目标检测的定义。

2. **开源生态系统**：我们发布了一个用于空间邻近红外小目标解混的开源生态系统，包括 SeqCSIST 数据集和一个用于基准测试的工具包。

3. **端到端框架**：我们提出了 DeRefNet，一个具有时间可变形特征对齐的多帧模型驱动的深度学习架构。

2 相关工作

2.1 序列红外小目标检测

在过去几年中，序列红外小目标检测（SIRST 检测）已成为红外预警与跟踪系统中的一项关键技术。它能够实现对连续帧中小目标的检测与跟踪，同时因其高灵敏度和抗干扰能力而备受重视 Tong et al. (2024)。

传统的序列红外小目标检测方法，例如基于局部对比度的方法 Chen et al. (2013)、基于人类视觉系统的方法 Han et al. (2014) 以及基于低秩的方法 Feng et al. (2023)，虽然在特定条件下表现出有效性，但在应对复杂背景等常见挑战时仍存在困难。为解决这些局限性，研究人员开发了基于深度学习的方法，显著提升了检测的准确性与鲁棒性。在近期的工作中，基于记忆增强的全局-局部聚合（MEGA）网络 Chen et al. (2020) 是首个在红外目标检测中系统性地整合全局上下文信息与局部细节特征的方法。尽管该方法增强了关键帧的特征表示，但它在很大程度上依赖于经验性设计的结构。类似的，后来出现的时间

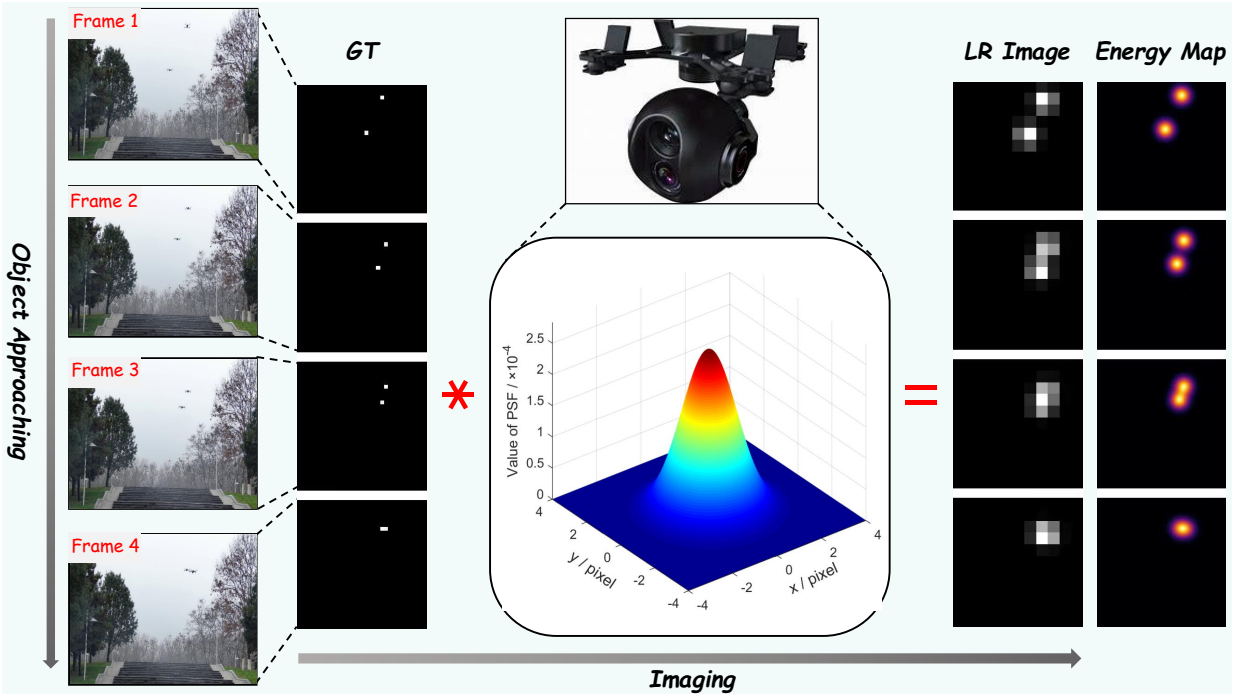

图 2 该示意图表明，随着远距离小目标彼此靠近，其能量混叠现象也愈发明显，以至于无法再通过视觉观察来区分目标的数量和具体位置。

感知全卷积神经网络（TFCST）Huang et al. (2024) 和背景估计框架（BEmST）Deng et al. (2024) 虽然推动了时空建模技术的发展，但仍沿用了传统的特征提取范式。

尽管这些方法在时空关系建模方面表现出有效性，但它们存在一个共同的局限性：主要侧重于设计复杂的检测头，且采用通用的网络主干。这种设计忽略了空间邻近红外小目标固有的稀疏特性，导致特征表示次优。此外，这些方法忽略了领域特定的先验知识，而这些先验知识本可以引导学习过程朝向更有效的特征提取方向发展。我们的工作通过以下两个亮点来解决这些局限性：

1. **下游任务**：序列空间邻近红外小目标解混与序列红外小目标检测是本质上不同的任务。空间邻近红外小目标解混专注于亚像素级别的定位，而序列红外小目标检测则侧重于识别目标轮廓。我们的工作通过为空间邻近红外小目标解混引入一个序列基准来强调这一区别。

2. **模型驱动的主干网络**：通过一个模型驱动的深度学习网络嵌入领域特定的先验知识，与通用的序列红外小目标检测框架相比，我们的框架能更好地应对序列空间邻近红外小目标解混的独特挑战。

2.2 深度展开

深度展开范式通过将迭代推理过程转化为类似于神经网络的逐层结构，结合了基于模型的方法与深度神经网络的优势。这种转换允许模型参数在不同层之间变化，从而实现更灵活的架构，并能通过基于梯度的方法进行高效优化 Zhang et al. (2020)。例如，基于模型驱动的交替方向乘子法（ADMM）算法，ADMM-Net 将每次迭代映射到一个神经网络层，并通过判别式学习来优化参数 Yang et al. (2018)。

近年来，深度展开在各种计算机视觉任务中的应用日益增多。例如将迭代收缩阈值算法（ISTA）重构为一个深度神经网络 ISTA-Net Zhang and Ghanem (2018)，从而同时利用了模型驱动方法和学习驱动方法的优势。虽然该方法解决了传统方法依赖非线性稀疏变换所带来的性能瓶颈，但在实际应用中处理多场景图像时缺乏灵活性。在此基础上，又出现了 ISTA-unfolding 深度网络 (ISTA-Net++)，这是一种用于压缩感知的灵活自适应深度部署网络 You et al. (2021)。它实现了在不同压缩感知率下的图像重建。

我们认可这些尝试的宝贵贡献，但它们主要集中在如图像重建等低层视觉任务上。相比之下，我们的工作将焦点转移到空间邻近红外小目标解混这一高层视觉任务上，而这类方法在该领域鲜有探索。我们的主要贡献体现在以下两个关键方面：

1. **在新任务中的开创性应用**：我们率先将深度展开范式引入序列空间邻近红外小目标解混任务，为该领域未来设计优化驱动的神经网络开辟了

新的可能性。

2. **动态场景适应性**：与依赖固定采样位置的传统深度展开方法不同，我们的模型能根据输入图像动态调整采样位置，从而更好地适应多帧空间邻近红外小目标场景。

3 SeqCSIST 数据集

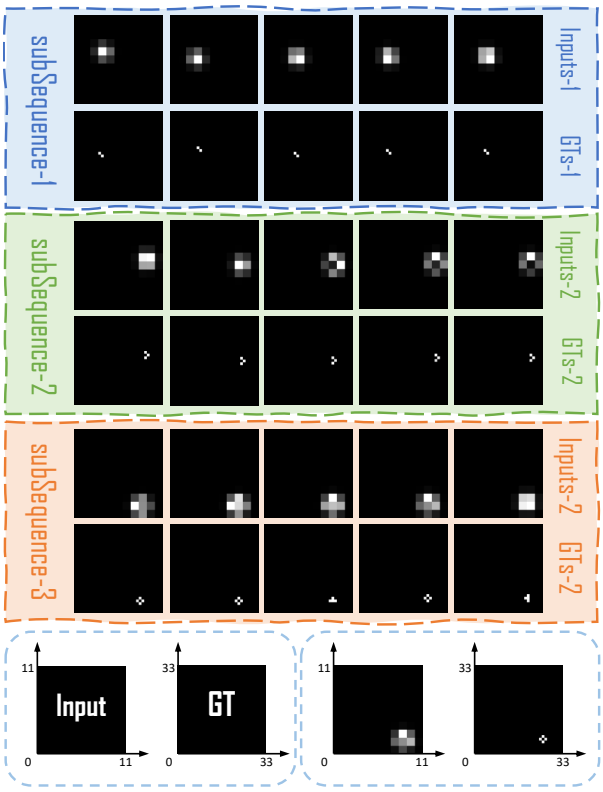

图 3 SeqCSIST 数据集包含两个组成部分：一是光学系统捕获的 11×11 低分辨率图像，二是真实标签（GT）的具体坐标。图中的每一组（由两行图像构成）代表一个子序列，即序列中的部分连续帧。在每个子序列中，第一行展示了 11×11 的低分辨率图像，其对应的真实标签（GT）可视化结果则显示在第二行。为便于清晰地观察目标位置，真实标签的可视化结果被上采样至 33×33 分辨率。

由于光学镜头的焦距和红外探测器分辨率的限制，远距离目标在红外成像平面上通常表现为低强度区域，且仅占几个像素 Liu et al. (2024)。在成像过程中，光学衍射会将点目标的能量分布到相邻像素。该能量分布对应于一个艾里斑，其捕获了大约 84% 的总能量 Reagan and Abatzoglou (1993)，并且在数学上可以用二维高斯点扩散函数（PSF）来近似 Lin et al. (2012)。艾里斑的半径与点扩散函数的标准差呈正相关。首先，点扩散函数的标准差受传感器 f 数和探测波长的影响，用于衡量能量的弥散程度。其次，艾里斑的半径约为点扩散函数标准差的 1.91 倍。此外，艾里斑的半径，即瑞利单位（R），定义了传感器的分辨率极限。当目标间距小于 1R 时，会发生能量混叠。因此，邻近目标（CSOs）在成像平面上表现为像素簇 Macumber et al. (2005)，使得光学系统无法将其分辨为独立的光点。该效应如图 2 所示。

在本研究中，为模拟真实环境下的远距离红外小目标序列成像，我们采用 84% 能量集中的定义作为分辨率标准，扩散方差设为 0.5 像素。我们将训练图像尺寸设置为 11×11，**每张图像的目标数量在 2、3 和 4 之间随机变化。目标强度值在 [220, 250] 范围内随机分布**，并且成像平面（一张 11×11 的图像）上所有目标的坐标和强度值都记录在相应的 XML 文件中。值得注意的是，目标点沿二次函数、圆形和直线等随机曲线运动。

SeqCSIST 数据集遵循以下约束：

1. **空间分布**：所有目标都相对于**一个没有强度值的参考点**进行定位，其方向编号与目标编号相对应（在上下左右四个方向中随机选择）。确保**目标点相对于参考点的方向**（从参考点指向目标点）**在帧间运动中保持不变**。

2. **初始位置与运动方向**：参考点在第一帧的初始坐标记为 $(x, f(x))$，其中 x 表示 x 轴上的任意整数，不包括与图像边界重合的值。**参考点沿轨迹的运动方向由其初始 x 坐标决定**：若 $x < 5$，则该点沿轨迹在 x 轴正方向上前进；否则，在 x 轴负方向上前进。

3. **子目标约束**：每个目标点到参考点的初始距离设置为 0.3 像素，这意味着每个目标点之间的最小距离为 $0.3\sqrt{2}$ 像素，这确保了所有目标点都保持在单个像素的空间范围内，并且解混后的目标位于不同的像素上。

4. **帧间运动规则**：参考点在连续帧中沿轨迹移动，如果其初始 x 坐标小于 5，则沿 x 轴正方向前进 $+0.04$ 像素，否则沿 x 轴负方向后退 -0.04 像素。在此基础上，**每个目标点相对于前一帧沿远离参考点的方向移动一个随机距离**，该距离值在 $(0, +0.0014)$ 像素范围内。

制定这些规则的总体目标是在空间分辨率、目标运动模式和能量重叠复杂性方面，紧密复现真实的红外小目标检测过程。下面，我们解释每条规则的设计动机及其在实现逼真模拟中的作用：

- **初始目标距离设为 $0.3\sqrt{2}$ 像素**：这确保了所有目标都被限制在单个像素区域内。经过点扩散函数（PSF）处理以模拟光学衍射后，这将导致混合的能量簇，从而忠实地再现实际红外图像中邻近小目标的空域混合现象。

- **固定的相对运动方向**：在一个序列中为目标保持一致的运动方向，有助于模拟一个“目标群”沿连续轨迹移动的场景，这在多个目标协同运动的真实世界场景中很常见。

- **每帧移动 0.04 像素**：这模拟了亚像素级别的运动，这是利用时间信息辅助能量解混的一个关键因素。选择足够小的移动步长是为了保持帧间的时间相关性，这反映了目标运动的自然连续性。
- **四个离散的运动方向**：从四个可能的方向中进行选择，模仿了在实际环境中遇到的随机但物理上合理的目标运动方向分布，为数据集增加了逼真的可变性。

遵循以上原则，数据集的最终视觉效果如图 3所示。

SeqCSIST 数据集包含 5,000 条轨迹，总计 100,000 帧。每条轨迹包含 20 帧，每 5 个连续帧组成一个序列（每个序列的初始帧在轨迹中有一个对应的序列号，从 0 到 15，因此总共有 16 个序列）。该数据集被划分为三个子集：70% 用于训练（3500 条轨迹），15% 用于验证（750 条轨迹），以及 15% 用于测试（750 条轨迹）。

4 方法

4.1 概述

需要强调的是，本文提出的解混（unmixing）任务与遥感领域的高光谱解混（hyperspectral unmixing）在根本上有所不同。后者通常涉及物理混合模型下的端元提取和丰度估计，而本文提出的任务更侧重于**目标分离与亚像素定位**。

序列空间邻近红外小目标解混的目标是从一个奇数帧的视频序列中，对中间帧的目标进行亚像素定位与解混。给定 $2N+1$ 个视频帧 $\{L_{t-N},\ldots,L_{t+N}\}$，其中 $L_t \in \mathbb{R}^{C\times H\times W}$ 表示中间帧，其余的参考帧为中间帧的目标解混提供补充的时空信息。低分辨率帧 $\{L_{t-N},\ldots,L_{t+N}\}$ 通过模型 DeRefNet 进行处理，得到一个预测的解混响应 $H_t \in \mathbb{R}^{C\times cH\times cW}$，其中 c 是解混比率。

$$H_t = f_{DeRefNet}(L_{t-N},\ldots,L_{t+N}). \tag{1}$$

DeRefNet 的具体网络架构如图 4所示。该框架主要由两部分组成：一个用于从输入序列中捕获深层特征表示的特征提取模块，以及一个时间可变形特征对齐（TDFA）模块，该模块自适应地将参考帧与中间帧对齐，以确保精确的时空对应。值得注意的是，为了更好地捕捉帧间的关系，我们在特征提取后的结果中加入了时间特征。

为了模拟真实的光学成像条件，维度为 $cm\times cm$ 的真实标签（GT）图像通过采样矩阵 Φ 进行下采样，生成 $m\times m$ 的输入帧，其中 c 表示下采样比率。如图 4所示，DeRefNet 通过处理输入序列中大小为 $m\times m$ 的图像来对中间帧进行目标解混。随后，这些低分辨率帧首先通过一个初始化矩阵 Q_{init} 被上采样到 $cm\times cm$。然后，特征提取模块捕获所有帧的空间特征，产生的特征被分为 $2N$ 个参考帧特征 $H_{t-N}^{(k)},\ldots,H_{t-1}^{(k)},H_{t+1}^{(k)},\ldots,H_{t+N}^{(k)}$ 和一个中间帧特征 $H_t^{(k)}$。提取出的特征被增添了时间信息，然后经过一个二维卷积模块。最后，时间可变形特征对齐模块在参考特征和中间特征之间执行时空对齐，以指导目标解混过程，最终生成代表解混后目标响应的 H_t。

4.2 采样与初始化

为了模拟从高分辨率目标分布到低分辨率测量的光学退化过程，我们基于一个连续域的点扩散函数（PSF）模型来显式地构建下采样矩阵 Φ。Φ 中的每个元素量化了一个亚像素目标对特定低分辨率传感器像素的贡献。形式上，一个中心位于 (x,y) 的像素对一个位于 (x_t,y_t)、亮度为 a_i 的亚像素目标的响应被建模为一个二维高斯扩散函数：

$$\mathrm{PSF}(x,y;x_t,y_t) = a_i\cdot\frac{1}{2\pi\sigma^2}\exp\left(-\frac{(x-x_t)^2+(y-y_t)^2}{2\sigma^2}\right). \tag{2}$$

考虑到传感器的积分效应，像素响应 $r(x,y)$ 通过在像素区域上对 PSF 进行积分来计算：

$$r(x,y) = \iint_{\text{pixel}} \mathrm{PSF}(u,v;x_t,y_t)\,du\,dv. \tag{3}$$

该积分使用 Python 中的 `dblquad` 方法进行数值计算，其中每个低分辨率像素接收来自一个密集的模拟亚像素目标网格的贡献。所有目标产生的像素响应被组织成一个矩阵 $\Phi\in\mathbb{R}^{m\times n}$，其中每一行对应一个低分辨率传感器元件，每一列对应高分辨率网格中的一个特定亚像素位置。矩阵 Φ 有效地捕捉了在真实光学成像系统中发生的空间变化的模糊和积分过程。

根据上述原理，成像平面上获得的图像是低分辨率图像，它是由真实标签（GT）高分辨率图像经过光学系统退化得到的，表示为 Eq. (4)：

$$L = \Phi H_{\mathrm{GT}}. \tag{4}$$

本工作的目标是从低分辨率测量值 L 中恢复高分辨率的真实标签 H_{GT}，这类似于解决一个压缩感知（CS）问题。受传统方法 ISTA 的启发，DeRefNet 首先对低分辨率图像 L 进行初始化。具体来说，初始化方法遵循 Zhang and Ghanem (2018)，其中使用一个线性映射 Q_{init} 将低分辨率的 $m\times m$ 图像 L_i 映射到一个高分辨率的 $cm\times cm$ 图像 $H^{(0)}$。

我们采用最小二乘映射来初始化每个高分辨率估计 $H^{(0)}$，主要基于三个关键原因：首先，通过将网络锚定到一个具有物理意义的起点，我们防止了在早期反向传播步骤中出现梯度爆炸或消失，并确保了训练的稳定性；其次，这种选择通过模仿 ISTA 的解析初始化，保留了其可解释性和收敛保证，而简单的卷积或零初始化会导致收敛速度变慢和鲁棒性降低；第三，嵌入低分辨率输入 L 与真实标签 H_{GT} 之间

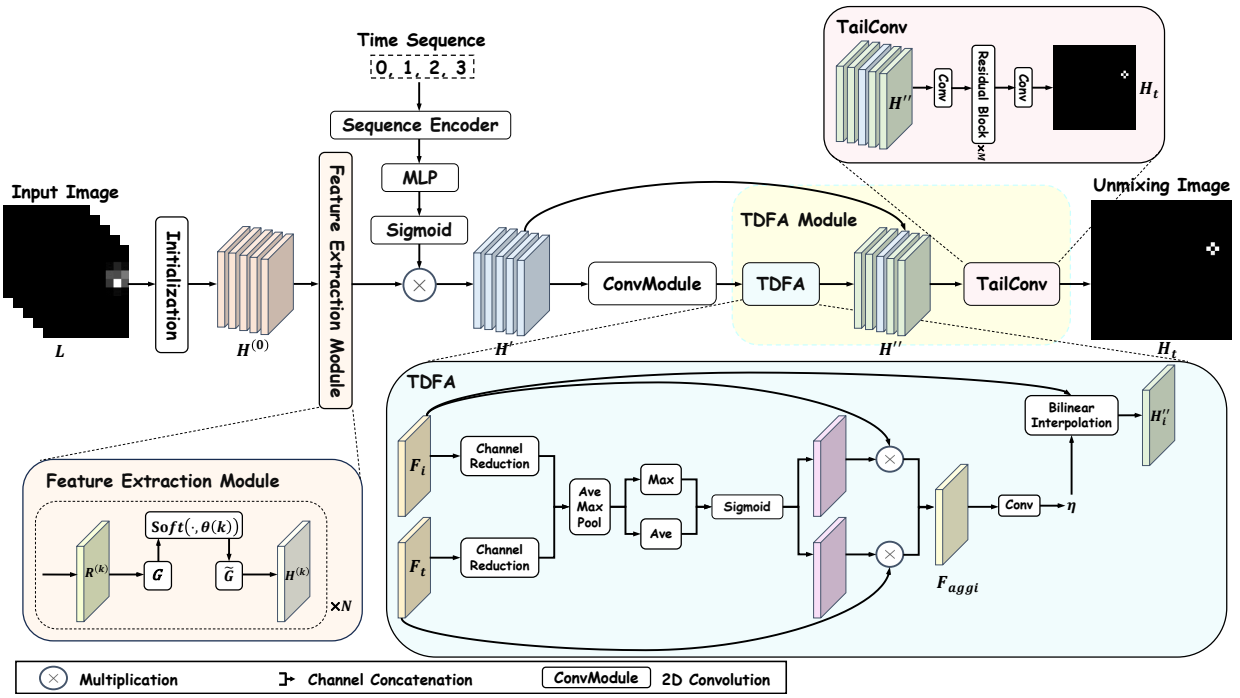

图 4 所提出的 DeRefNet 方法的框架由两个主要部分组成。第一部分是特征提取模块，用于从初始化后的图像中提取空间特征；随后，将时间信息融入到提取的特征图中。第二部分是时间可变形对齐模块，它将参考帧对齐到中间帧。

的最优线性关系，显著缩小了有效的参数搜索空间，从而加速了网络收敛。最终的初始化形式如下：

$$Q_{\text{init}} = \arg\min_{Q} \|QL - H\|^2 = L^T(LL^T)^{-1}H, \quad (5)$$

这里 $L = [L_{t-N}, \ldots, L_{t+N}]$ 且 $H = [H_{t-N}, \ldots, H_{t+N}]$。从几何角度看，这将 L 投影到由 H_{GT} 张成的子空间上，同时最小化了重建偏差——这是一种有原则的替代方案，相较于经验性地放大高频伪影的启发式 CNN 初始化方法。这种闭式解 $L^T(LL^T)^{-1}H$ 保留了 ISTA 的代数结构，并确保网络的第一层从最优的线性解混算子开始。

如 Eq. (6) 所示，对每个输入 L_i 初始化 $H_i^{(0)}$ 建立了一个仿射变换，在潜空间中保持了线性可分性。这与 ISTA 在早期迭代中对凸性的要求相符，确保了后续网络层是在改进而非重塑基本的映射关系。

$$H_i^{(0)} = Q_{\text{init}} L_i. \quad (6)$$

4.3 特征提取模块

为了实现利用稀疏性先验的特定任务特征提取，我们设计了一个空间特征提取模块，该模块在一个宿主网络架构中融入了深度展开范式。与依赖堆叠残差块的传统目标检测网络不同，我们的方法提取了更具语义意义的特征。该模块通过 Eq. (7) 将初始化的 $cm \times cm$ 特征图从 $H_{t-N}^{(0)}, \ldots, H_{t+N}^{(0)}$ 转换为 $H_{t-N}^{(k)}, \ldots, H_{t+N}^{(k)}$：

$$H_{t-N}^{(k)}, \ldots, H_{t+N}^{(k)} = f_{FE}(H_{t-N}^{(0)}, \ldots, H_{t+N}^{(0)}). \quad (7)$$

具体而言，我们方法的基础在于经典的 ISTA 算法，该算法通常通过以下两个步骤解决 CS 问题：

$$R_i^{(k)} = H_i^{(k-1)} - \rho\Phi^T(\Phi H_i^{(k-1)} - L_i), \quad (8)$$

$$H_i^{(k)} = \arg\min_{H_{\text{GT}i}} \frac{1}{2}\|H_{\text{GT}i} - R_i^{(k)}\|_2^2 + \lambda\|\Psi H_{\text{GT}i}\|_1, \quad (9)$$

这里，ρ 是梯度下降的步长，Ψ 是预定义的手工设计变换矩阵，$H_i^{(k)}$ 表示第 i 张图像的第 k 次迭代结果，其中 $i \in \{t-N, \ldots, t+N\}$。

深度展开范式将此框架扩展为一个包含 N 个阶段的深度架构，每个阶段对应一次 ISTA 迭代。由于在处理更复杂的非正交（甚至非线性）变换 Ψ 时求解 $H_i^{(k)}$ 存在困难，且 ISTA 通常需要大量迭代才能达到理论最优值，这通常会带来巨大的计算成本，深度展开范式 Zhang and Ghanem (2018) 用一个可训练的非线性变换函数 $G(\cdot)$ 替代了手工设计的矩阵 Ψ。给定 $G(\cdot)$ 的可逆性，其左逆 $\tilde{G}(\cdot)$ 定义为 $\tilde{G}(\cdot) \circ G(\cdot) = \mathbf{I}$。于是，上述的 Eq. (9) 变为：

$$H_i^{(k)} = \arg\min_{H_{\text{GT}i}} \frac{1}{2}\|H_{\text{GT}i} - R_i^{(k)}\|_2^2 + \lambda\|G(H_{\text{GT}i})\|_1. \quad (10)$$

在这里，$R_i^{(k)}$ 是 $H_i^{(k-1)}$ 在第 k 次迭代中的直接重建结果，我们期望在 Eq. (10) 中，$R_i^{(k)}$ 与 $H_{\mathrm{GT}i}$ 之间的差异能被最小化。因此，$R^{(k)}$ 满足以下定理 Zhang and Ghanem (2018)：

$$\|G(H_{\mathrm{GT}i}) - G(R_i^{(k)})\|_2^2 \approx \alpha \|H_{\mathrm{GT}i} - R_i^{(k)}\|_2^2. \quad (11)$$

通过此定理，Eq. (10) 可以被优化为：

$$H_i^{(k)} = \arg\min_{H_{\mathrm{GT}i}} \frac{1}{2}\|G(H_{\mathrm{GT}i}) - G(R_i^{(k)})\|_2^2 + \theta \|G(H_{\mathrm{GT}i})\|_1, \quad (12)$$

其中 λ 和 α 被合并为 θ，从而得到其闭式解形式：

$$G(H_i^{(k)}) = \mathrm{soft}(G(R_i^{(k)}), \theta). \quad (13)$$

给定 $G(\cdot)$ 的逆，可以高效地以闭式解计算出 $H_i^{(k)}$：

$$H_i^{(k)} = G^{-1}(\mathrm{soft}(G(R_i^{(k)}), \theta)). \quad (14)$$

在这里，θ 是一个收缩阈值，并且是一个可学习的参数。每个阶段都有其自己的 $G(\cdot)$ 和 $\tilde{G}(\cdot)$，并且它们都是可学习的过程。因此，上述方程可以转换为：

$$H_i^{(k)} = \tilde{G}^{(k)}\left(\mathrm{soft}\left(G^{(k)}(R_i^{(k)}), \theta^{(k)}\right)\right). \quad (15)$$

经过第 k 次迭代后，得到最终的特征提取结果 $H_i^{(k)}$。因此，我们得到了 Eq. (7)。

值得注意的是，该架构不仅利用了稀疏性先验，还实现了所有参数的端到端学习，从而获得了比传统通用骨干网络更优越的特征提取能力，正如我们的消融研究所证明的那样。

4.4 时间可变形特征对齐模块

为了增强跨多帧的特征表示，我们融入了时间信息并开发了一种专门的对齐方法，以高效捕捉帧间关系。与追踪完整运动轨迹的传统光流方法 Dosovitskiy et al. (2015) 不同，我们的方法通过一个基于注意力的架构，选择性地将关键参考帧特征与中间帧对齐，从而在保留关键信息的同时显著降低了计算复杂性。**我们首先通过一个可学习的编码过程，用时间信息来增强空间特征**。具体来说，时间信息的固定位置编码是通过位置编码模块实现的。为了更好地利用时间信息 t，我们增加了一个 MLP 层，以实现网络对时间信息的学习和利用，然后使用 Sigmoid 函数得到 T。

$$T = Sigmoid(MLP(Encoder(t))). \quad (16)$$

其中 t 代表时间信息，Sigmoid 函数确保了适当的缩放。然后，这个时间编码 T 通过逐元素相乘的方式与提取的空间特征 $H^{(k)}$ 相结合：

$$H^{'} = H^{(k)} T. \quad (17)$$

为了进一步增强表示能力，我们通过一个二维卷积层扩展了 $H^{'}$ 的通道维度，生成了特征图 F，它包含一个中间特征 F_t 和其他参考特征 F_i，其中 $i \in \{t-N, \ldots, t-1, t+1, \ldots, t+N\}$。**这些特征作为 TDFA 模块的输入**。

TDFA 模块分两个互补的阶段运行，以实现参考帧与中心帧之间的最佳对齐。与尝试明确追踪完整运动轨迹的光流方法不同，我们通过选择性注意力实现可调整的帧间对齐，**这是一种隐式的可学习策略，它利用稀疏性先验，仅关注高激活区域——即那些具有最强特征的区域**。在第一阶段，我们实现了一个选择性注意力架构，动态地结合参考帧和中间帧的特征：首先，参考帧和中间帧的特征通过选择性注意力过程被动态地结合起来。

$$F_{aggi} = SelectiveAttention(F_i, F_t). \quad (18)$$

这个过程首先通过卷积减少两个特征图的通道维度，然后将它们拼接起来。接着，我们提取最大池化和平均池化特征，并通过一个 Sigmoid 函数进行激活。认识到中心帧在我们的处理流程中的首要地位，我们策略性地对注意力进行加权：将平均池化权重应用于中心帧特征，而将最大池化权重应用于参考帧，以强调它们最显著的特征。最终得到的加权特征沿着通道维度进行拼接，创建聚合特征图 F_{aggi}。

第二阶段通过评估参考帧与聚合特征 F_{aggi} 的相关性，自适应地从参考帧中提取特征，这使得模型能够有效地将参考帧的补充信息与整体特征表示对齐和整合。该过程为每个参考特征图 F_i 预测采样参数 η：

$$\eta = conv(F_{aggi}). \quad (19)$$

这里，$\eta = \{\Delta p_k\}$ 包含了内容相关的偏移矩阵，用于调整采样窗口以更好地捕捉聚合特征 F_{aggi} 的复杂结构。利用这些学习到的偏移，我们通过双线性插值 $\phi(\cdot)$ 将每个参考特征图 F_i 调制成一个对齐后的表示 $H_i^{''}$。

$$H_i^{''} = \phi(F_i, \eta) \quad (20)$$

对于 $H_i^{''}$ 中的任何空间位置 P，其特征响应计算如下：

$$H_i^{''}(p) = \sum_{k=1}^{K^2} \omega_k \cdot F_i(p + p_k + \Delta p_k). \quad (21)$$

这里，K^2 是网格中总的采样点数量，ω_k 代表分配给第 k 个采样点的权重，p_k 表示固定的采样偏移，而 Δp_k 是根据输入特征调整固定采样位置的动态偏移，即 $\eta = \{\Delta p_k\}$。请注意，Δp_k 可以是小数，为了处理这种情况，我们遵循 Dai et al. (2017) 中描述的方法，使用双线性插值来精确计算这些非整数位置的特征值。

最后，将对齐后的 $2N$ 个参考特征 $H_i^{''}$ 和一个中间帧特征 $H_t^{(k)}$ 通过尾部卷积进行聚合，以获得解混后的图像 H_t。为了增强网络的表示能力和优化稳定性，

这个聚合过程包含了一个由 n 个残差块组成的级联结构，这在有效加深架构的同时，通过跳跃连接缓解了梯度消失问题。

4.5 损失函数

我们的 DeRefNet 的损失由三部分组成：约束损失、对齐损失和回归损失。具体来说，约束损失确保了在特征提取模块内部的每次迭代中，内部过程 $\tilde{G}(\cdot) \circ G(\cdot) = \mathbf{I}$ 的可逆性。其目的是减少初始化图像与提取的特征图像之间的差异，具体公式如 Eq. (22) 所示：

$$\mathcal{L}_{\text{constraint}} = \frac{1}{XY} \sum_{i=1}^{X} \sum_{i=1}^{L} \left\| \tilde{G}^{(k)} \left(G^{(k)}(s_i) \right) - s_i \right\|_2^2 . \tag{22}$$

这里 X 是训练集的大小，L 是阶段数，Y 是超分辨率图像特征的大小，s_i 代表真实标签（GT）。

对齐损失指的是将参考帧与中间帧对齐时产生的损失。减少对齐损失可以确保在时间可变形特征对齐块内，对齐后的参考帧与中间帧之间的差异最小化。对齐损失的具体公式如下：

$$\mathcal{L}_{\text{align}} = \frac{1}{2N} \sum_{i=t-N, i \neq t}^{t+N} \left\| H_i^{''} - H_t^{'} \right\|_1 . \tag{23}$$

$$\mathcal{L}_{\text{regression}} = \frac{1}{(T-4) * (M/T)} \sum_{k=1}^{M/T} \sum_{i=2}^{T-2} \| H_{ki} - s_{ki} \|_2^2 , \tag{24}$$

这里，T 代表一条轨迹的帧数，H_{ki} 和 s_{ki} 分别代表第 k 条轨迹的第 i 帧的重建结果及其对应的 GT，M/T 代表轨迹的数量。

结合以上三种损失，得到 DeRefNet 模型的总损失如下：

$$\mathcal{L}_{all} = \beta \mathcal{L}_{constraint} + \gamma \mathcal{L}_{align} + \zeta \mathcal{L}_{regression} \tag{25}$$

其中 β、γ 和 ζ 是各项损失的权重系数，我们设置 β 和 γ 为 0.01，ζ 为 1。为了验证这些超参数的选择，我们进行了一项网格搜索实验，包含了六种不同的权重配置：

- **A 组 (默认)**: $\zeta = 1$, $\beta = 0.01$, $\gamma = 0.01$ —平衡的监督；用作基线。
- **B 组**: $\zeta = 1$, $\beta = 0.05$, $\gamma = 0.05$ —削弱辅助约束；评估回归损失的主导作用。
- **C 组**: $\zeta = 1$, $\beta = 0.2$, $\gamma = 0.2$ —放大辅助信号；测试更强的正则化是否能改善泛化能力。
- **D 组**: $\zeta = 1$, $\beta = 0.2$, $\gamma = 0.05$ —强调变换约束，同时保持对齐约束较弱。
- **E 组**: $\zeta = 1$, $\beta = 0.05$, $\gamma = 0.2$ —强调时间对齐；测试其对整体准确率的贡献。
- **F 组**: $\zeta = 0.5$, $\beta = 0.1$, $\gamma = 0.1$ —降低主损失的权重，以观察辅助损失占主导地位时的影响。

表 1 损失权重配置的网格实验：平衡主回归损失与辅助约束以实现最优空间邻近红外小目标检测

组别	CSO-mAP	AP_{05}	AP_{10}	AP_{15}	AP_{20}	AP_{25}
A	**51.55**	**1.00**	**14.40**	**54.90**	90.40	97.10
B	50.82	0.70	11.00	51.50	92.20	98.60
C	50.44	0.70	10.40	50.50	92.00	98.60
D	50.54	0.70	10.50	50.70	92.10	98.80
E	51.11	0.80	11.40	51.90	92.50	98.90
F	50.86	0.80	10.90	51.10	**92.60**	**99.00**

实验结果表明，我们的默认配置（A 组）取得了最佳的综合性能。A 组的优越性能验证了我们超参数选择的合理性：(1) 设置 $\zeta = 1$ 确保回归损失在学习时间边界中保持其主要作用；(2) 为辅助损失设置适中的权重 $\beta = \gamma = 0.01$，提供了足够的正则化，而不会压倒主要目标。B-F 组的结果显示，无论是削弱辅助约束（B 组）还是过分强调它们（C-D 组），都会导致关键指标性能下降，而降低主损失权重（F 组）尽管在宽松的时间阈值上略有改善，却牺牲了精确定位能力。该消融实验证实了我们平衡的方法能够最优地利用主监督和辅助正则化。

5 实验

5.1 实验设置

评价指标：为评估模型在序列空间邻近红外小目标解混任务中的性能，我们采用了 **CSO 平均精度均值 (CSO-mAP)**，这是一个为本任务定制的评估指标。

评估过程首先将每个预测分为真正例（TP）或假正例（FP），并按照 COCO 的惯例创建一个二进制列表。在此列表中，TP 预测由 1 表示，而 FP 预测由 0 表示。这种二进制表示构成了构建精确率-召回率（PR）曲线的基础。通过系统地调整正例预测的置信度阈值，可以导出多对精确率和召回率值，这些值共同描绘出 PR 曲线。平均精度（AP）通过计算 PR 曲线下的面积得出，它提供了对模型在不同置信度水平下行为的详细评估，捕捉了精确率和召回率之间的平衡。为进一步规范评估，我们使用 CSO 平均精度均值（CSO-mAP），该指标聚合了在不同距离阈值 δ_k 下的 AP 分数。该指标为在序列空间邻近红外小目标解混任务背景下比较模型性能提供了一个一致的框架。

训练设置：作为目标检测的下游后处理任务，我们的方法处理的是从检测结果中裁剪出的目标切片，而非原始输入图像。在整个实验中，初始化的超分辨率比率 c 设置为 3。输入的目标切片尺寸为 11×11，真实标签（GT）图像尺寸和输出的解混图像尺寸均为 33×33。为了每次能加载一条完整的

轨迹，实验中的批处理大小设置为 20。为提高数据利用率并获得更好的模型性能，每次处理 5 个连续帧（$1-5, 2-6, ..., 16-20$）。因此，每条轨迹包含 $20-4=16$ 个序列，而不是 $20 \div 5 = 4$ 个序列。最后，残差块的数量 n 设置为 5。在实验中所有涉及空间特征提取的部分，均使用 32 通道的卷积。在优化方面，我们采用 Adam 优化器和 MMEngine 框架来执行网络，学习率始终保持在 10^{-4}。

5.2 消融研究

1) 对早期特征提取的影响：为了评估深度展开范式相对于传统堆叠残差块的有效性，我们使用多帧特征提取范式 Tian et al. (2020) 进行了一项受控的消融研究。通过将网络头部的残差堆叠结构替换为我们的深度展开架构，我们分离出了这一架构选择对特征提取性能的影响。

如表 2 所示，深度展开方法在所有评估指标上都持续优于基线。在多个 IoU 阈值下，性能增益尤为明显，这证明了我们的方法在不同精度要求下的鲁棒性。

这种持续的改进可归因于两个关键因素。首先，与通用的残差堆叠不同，我们的深度展开架构融入了任务特定的稀疏性先验，从而能够进行更具针对性的特征提取，这与红外图像的固有特性相符。其次，通过将超分辨率过程移至流程的前端，我们的方法有效解决了在红外场景中小目标普遍存在的混合问题。

在较低 IoU 阈值下观察到的尤其显著的改进，凸显了深度展开架构在区分邻近目标方面的增强能力——这在检测冲突频繁发生的密集场景中是一项至关重要的能力。这些结果提供了有力的证据，表明融入了适当先验知识的任务特定架构，其性能优于通用的特征提取方法，即使后者已针对类似任务进行了广泛优化。

表 2 特征提取模块与堆叠残差块结构的性能对比

主干网络	CSO-mAP	AP_{05}	AP_{10}	AP_{15}	AP_{20}	AP_{25}
残差块	47.96	0.50	8.60	43.80	89.30	97.50
深度展开	**50.27**	**0.70**	**11.40**	**50.80**	**90.60**	**97.90**

2) 对可变形对齐的影响：为了系统地评估可变形对齐（DA）模块相对于传统光流法的优越性，我们对这两个模块进行了性能对比评估。如表 3 所示，两个模块的性能在多个阈值下进行了量化比较。

消融实验的结果显示，采用基于插值的可变形采样的可变形对齐模块，在不同阈值下的目标检测性能显著优于传统光流法。与光流法相比，它在多个 IoU 阈值下始终获得更高的平均精度（AP），这揭示了后者在捕捉空间邻近红外小目标场景中位移信息方面的固有局限性。值得注意的是，得益于其能够有效适应复杂运动模式和空间变化的动态感受野调整机制，可变形模块即使在较低的 IoU 阈值下也能保持稳健的 AP 性能。这表明当目标距离非常近时，该模块仍能保持其优越性。

3) 对时间编码器的影响：如表 3的第二行和第三行所示，为了凸显时序信息的重要性，我们在多帧解混框架的基础上增加了一个时间编码器。

结果表明，增加时间编码器能够提升模型在多个 IoU 阈值下的性能。具体来说，虽然两种模型都使用可变形对齐来改善空间特征提取，但时间编码器的加入进一步优化了运动表示，从而实现了更鲁棒和具有时间感知的预测。值得注意的是，在较高的 IoU 阈值下，AP 分数得到了一致的提升，这表明时间编码有助于模型更有效地捕捉序列依赖性和运动模式。这表明时间编码器通过引入更丰富的时间上下文，弥补了静态空间对齐的局限性，最终在动态环境中带来了性能的提升。

4) 对动态可变形对齐的影响：为了提升性能并降低模型权重，我们用选择性注意力结构替换了可变形堆叠结构，创建了动态可变形对齐（DDA）模块。如表 3的第二行和第四行所示，实验结果支持了这一意图：实验 2（使用可变形堆叠结构）的 FLOPs 为 6.44G，而实验 4（使用选择性注意力机制）仅用 6.14G 的 FLOPs 就取得了更优的性能。这种计算成本的降低表明，选择性注意力方法在提升模型性能的同时，也简化了架构以提高效率。通过动态地优先处理显著特征，它增强了特征表示和对齐精度，从而带来了更鲁棒的检测结果。这些优势凸显了所提出的模块作为传统可变形堆叠结构的一种轻量级而强大的替代方案。

5) 跨域泛化能力验证：为了评估我们提出的 DeRefNet 在域漂移条件下的鲁棒性和泛化能力，我们使用一个模拟真实部署场景的混合数据集进行了一项补充实验。

具体来说，我们将合成的 SeqCSIST 数据集中的 5,000 条目标轨迹叠加到 5,000 个真实世界的红外背景序列上。这些真实背景采集自多样的场景，包括山脉、森林、建筑、塔楼和大气环境，因此展现出比合成域中更复杂的噪声模式和背景纹理。这个混合数据集在引入真实变化的同时，保留了真实的解混标注，从而可以进行有意义的跨域评估。

我们使用与在 SeqCSIST 上相同的训练协议，在这个混合数据集上训练和测试了 DeRefNet。如表 4中的对比结果所示，当应用于混合数据集时，DeRefNet 的 mAP 仅下降了 4.07

尽管目前尚无公开的真实标注解混数据集，这限制了全面的真实世界评估，但本实验提供了一个有意义的中间验证步骤。在未来的工作中，我们计划探索域自适应策略，以进一步增强跨数据集的可迁移性。

6) 对高斯加性噪声的鲁棒性：为了评估 DeRefNet 对传感器和热噪声的抵抗能力，我们在保持网络架

表 3 系统性模块化消融研究：核心组件对 DeRefNet 性能及计算成本的影响评估

深度展开	光流法	DA	时间编码器	DDA	参数量	FLOPs	CSO-mAP	AP_{05}	AP_{10}	AP_{15}	AP_{20}	AP_{25}
✓	✓	-	-	-	/	/	50.55	0.70	11.20	49.50	92.50	98.80
✓	-	✓	-	-	0.23 M	6.44 G	50.67	0.80	11.90	50.90	91.50	98.30
✓	-	✓	✓	-	/	/	51.39	0.80	12.40	52.20	92.70	98.80
✓	-	-	-	✓	0.28 M	6.14 G	51.09	0.80	12.00	52.00	92.30	98.40

表 4 DeRefNet 在 SeqCSIST 和混合数据集上的性能，其中“混合”指将模拟的空间邻近红外小目标合成到真实世界的红外场景中

数据集	CSO-mAP	AP_{05}	AP_{10}	AP_{15}	AP_{20}	AP_{25}
SeqCSIST	**51.55**	**1.00**	**14.40**	**54.90**	**90.40**	**97.10**
混合数据集	47.48	0.70	11.00	46.30	84.80	94.60

表 5 综合噪声容忍度评估：DeRefNet 在不同标准差的高斯加性噪声下的性能分析

σ	CSO-mAP	AP_{05}	AP_{10}	AP_{15}	AP_{20}	AP_{25}
0 (无噪声)	**51.55**	**1.00**	**14.40**	**54.90**	**90.40**	**97.10**
2	49.09	0.80	12.40	50.50	86.60	95.30
3	48.40	0.80	11.80	48.00	86.00	95.30
4	48.41	0.90	12.20	48.00	85.80	95.20
5	47.23	0.80	11.40	46.10	84.10	93.90

构不变的情况下，用不同强度（$\sigma = 2$–5）的零均值高斯噪声对输入帧进行综合扰动。该噪声范围对应于红外成像系统典型的信噪比（33–42 dB）。我们排除了 $\sigma < 1$（太弱，无法对模型构成挑战）和 $\sigma > 10$（不切实际的失真），将重点放在 $\sigma = 2$–5 作为代表性的中等水平噪声。

如表 5所示，即使在中等噪声($\sigma = 2$)下，DeRefNet 的 mAP 仅下降 2.46 这些结果表明，我们的 DeRefNet 模型能够有效抵抗信号层面的干扰，并在真实的噪声条件下保持良好的泛化性能。

7) 对动态目标数量的适应性：为了评估 DeRefNet 在超出其 2-4 个共现目标的训练分布场景中的鲁棒性，我们进行了有针对性的实验，以评估 DeRefNet 在目标密度增加下的性能。我们将每个序列的目标范围从 2-4 个扩展到 2-8 个，根据领域专业知识，这代表了近距离红外场景的一个现实上限。包含 5-8 个目标的序列引入了显著的挑战，包括增加的目标间能量混叠、空间干扰和计算复杂性。实验方案如下：

- 生成了每个序列包含 2-8 个目标的扩展 SeqCSIST 数据集
- 保持了相同的训练协议和超参数
- 使用相同的指标进行直接比较

表 6 多目标解混性能评估：DeRefNet 在 SeqCSIST 序列数据处理中可变目标数量密度下的有效性分析

数据集	CSO-mAP	AP_{05}	AP_{10}	AP_{15}	AP_{20}	AP_{25}
2-4 个目标	**51.55**	**1.00**	**14.40**	**54.90**	**90.40**	**97.10**
2-8 个目标	49.79	0.70	12.60	50.60	87.90	97.10

如表 6所示，结果揭示了关于 DeRefNet 对增加的目标密度的适应性的几个重要发现：1) 当目标密度从 2-4 范围增加到 2-8 范围时，DeRefNet 的 CSO-mAP 下降了 1.76%（从 51.55% 降至 49.79%）。虽然这代表了中等的性能下降，但它表明模型在更具挑战性的条件下仍保持了合理的检测能力。2) 在中低 IoU 阈值的情况下，虽然解混精度有所降低，但指标的一致性仍能维持。在 AP_{20} 处性能有中度下降 (-2.5%)，而 AP_{25} 保持稳定 (97.10%)，这表明当模型实现高置信度检测时，精度得以保持。

这些发现证实了我们的时间解混框架的两个关键优势：1) 通过鲁棒的特征表示学习，有效泛化到训练分布边界之外，以及 2) 在动态目标数量条件下，同时保持解混保真度和空间定位精度。这些结果特别突出了该架构对目标密度变化本身不可预测的现实世界应用的适用性。

5.3 与先进方法的比较

为验证模型的有效性，我们在 SeqCSIST 数据集上将我们的 DeRefNet 与其他方法进行了比较，包括 ISTA Gregor and LeCun (2010) 和 ISTA-Net Zhang and Ghanem (2018) 等。具体的实验结果如表 7 所示。

从表 7 中的数据，我们可以对不同模型在不同 IoU 阈值下的 AP 性能进行详细的分析和比较。实验结果表明，**DeRefNet** 在 SeqCSIST 数据集上表现出卓越的目标解混能力，在 mAP 方面优于其他方法。我们分**三个**部分进行了实验，具体结果如下：

1) 与传统的模型驱动优化方法相比（例如 ISTA 和 BID），DeRefNet 在空间邻近红外小目标场景中的性能显著优于它们，实现了 51.55 的 mAP，远高于 ISTA（10.72）和 BID（14.40）。这一巨大差异凸显了纯模型驱动方法在有效处理序列空间邻近红外小目标解混方面的局限性。相比之下，DeRefNet 通过结合模型驱动和数据驱动的策略，在解析重叠目标和适应不同场景条件方面表现出色，在解混和定位精度上具有显著优势。

表 7 DeRefNet 性能的系统性基准分析：在 **SeqCSIST** 数据集上与包括经典优化、超分辨率网络和深度展开架构在内的多种方法类别的对比评估。

方法	FPS ↑	参数量 ↓	FLOPs ↓	CSO-mAP ↑					
				CSO-mAP	AP_{05}	AP_{10}	AP_{15}	AP_{20}	AP_{25}
传统优化									
ISTA Daubechies et al. (2004)	0.1		398.57 M	10.72	0.14	1.97	8.74	18.22	24.53
BID Levin et al. (2007)	0.1		10.89 M	14.40	0.00	3.00	13.00	26.00	30.00
图像超分辨率									
SRCNN Dong et al. (2015)	102,961	15.84 K	0.35 G	49.64	1.40	16.30	51.20	85.00	94.30
GMFN Li et al. (2019)	855	2.80 M	27.53 G	50.94	0.70	11.90	51.20	92.10	98.80
DBPN Haris et al. (2018)	7,109	1.96 M	4.75 G	50.40	0.80	12.50	51.20	90.00	97.40
SRGAN Ledig et al. (2017)	12,965	35.31 M	40.27 G	26.96	0.30	3.90	19.40	46.90	64.30
BSRGAN Zhang et al. (2021)	1,528	36.06 M	0.27 T	33.21	0.40	6.10	27.50	57.20	74.90
ESRGAN Wang et al. (2018)	1,024	50.45 M	0.38 T	36.86	0.40	6.00	30.30	66.80	80.70
RDN Zhang et al. (2018)	919	22.31 M	53.97 G	49.61	0.70	10.60	48.20	90.40	98.20
EDSR Lim et al. (2017)	11,476	0.39 M	0.99 G	50.19	0.60	10.30	48.80	92.20	99.00
ESPCN Shi et al. (2016)	144,901	54.75 M	22.73 K	47.18	1.60	15.30	46.60	80.30	92.00
TDAN Tian et al. (2020)	259	0.59 M	2.18 G	47.96	0.50	8.60	43.80	89.30	97.50
深度展开									
LIHT Wang et al. (2016)	253	21.10 M	0.42 G	6.36	0.10	1.00	4.30	10.40	16.00
LAMP Metzler et al. (2017)	7,172	2.13 M	86.97 G	9.09	0.10	1.50	6.50	15.00	22.30
ISTA-Net Zhang and Ghanem (2018)	4,052	0.17 M	4.09 G	48.95	0.70	11.20	49.70	87.70	95.40
FISTA-Net Xiang et al. (2021)	4,052	74.60 K	6.02 G	50.61	1.00	12.60	51.40	90.70	97.30
ISTA-Net+ Zhang and Ghanem (2018)	5,504	0.38 M	7.70 G	51.02	1.00	13.70	52.70	90.40	93.70
ISTA-Net++ You et al. (2021)	1,751	0.76 M	16.54 G	50.50	0.70	10.40	49.20	92.8	99.40
LISTA Gregor and LeCun (2010)	490	21.10 M	0.42 G	9.39	0.10	1.70	6.90	15.40	22.70
USRNet Zhang et al. (2020)	622	1.07 M	11.26 G	49.25	0.70	9.80	46.60	91.20	98.90
TiLISTA Liu and Chen (2019)	4,716	2.22 M	86.97 M	13.52	0.20	2.10	9.50	22.60	33.30
RPCANet Wu et al. (2024b)	2,601	0.68 M	14.81 G	47.17	0.70	10.20	44.50	84.60	95.90
⋆ **DeRefNet (我们的方法)**	367	0.89 M	15.70 G	**51.55**	1.00	14.40	54.90	90.40	97.10

2) 与超分辨率方法相比（例如 SRCNN、SRGAN 和 GMFN），DeRefNet 展现出明显的优势。虽然传统方法在序列空间邻近红外小目标解混中的性能有限，但 DeRefNet 通过融合多帧信息和时间处理，实现了显著更高的精度。通过利用多帧数据和时间序列中的动态变化，该模型能够捕捉到更精细的细节和时间关系，从而提高了精度。这凸显了 DeRefNet 在处理动态和复杂目标场景方面的优越性。此外，DeRefNet 受益于深度展开范式，并在网络的早期而非末端执行超分辨率。这些设计使 DeRefNet 能够更有效地处理复杂的目标场景。

3) 与深度展开方法相比， DeRefNet 通过增强的信息处理进一步展示了其优越性。多帧可变形对齐的使用在这一改进中起着至关重要的作用。在复杂的空间邻近红外小目标场景中，多帧可变形对齐捕捉了动态的帧间变化，使 DeRefNet 能够超越其他深度展开方法。这使得 DeRefNet 能够在各种阈值下保持强大的性能，展示了其在挑战性环境中进行细粒度目标解混的鲁棒性和能力。

DeRefNet 有效地平衡了效率和性能。仅用 $0.89M$ 的可学习参数，它证明了通过优化相对较少数量的参数可以实现最佳的解混性能。然而，其 FLOPs 达到 $15.70G$，相对于其轻量级的参数规模来说相对较高。这一增长是由于使用了可变形卷积，它增强了空间适应性和更深层次的特征提取。尽管其 FLOPs 中等，DeRefNet 实现了 367 FPS 的平均处理速度，这远超出典型的红外视频序列采集速率（通常为 25-30 FPS）。这表明在实际场景中具有很强的实时部署潜力。此外，与其他具有相似参数或 FLOPs 的模型相比，DeRefNet 的 51.55 的 mAP 展示了其卓越的效率和非凡的目标解混能力。

总之，DeRefNet 模型在处理序列空间邻近红外小目标解混方面表现出色，特别是在高精度识别和管理复杂场景方面。它在各种模型中实现了最佳的 mAP 性能，凸显了深度展开范式和可变形对齐在序列空间邻近红外小目标解混中的优越性。这些结果将 DeRefNet 定位为类似应用的领先模型，并为序列空间邻近红外小目标解混的未来发展提供了宝贵的见解，展示了显著的实用价值和广泛的应用潜力。

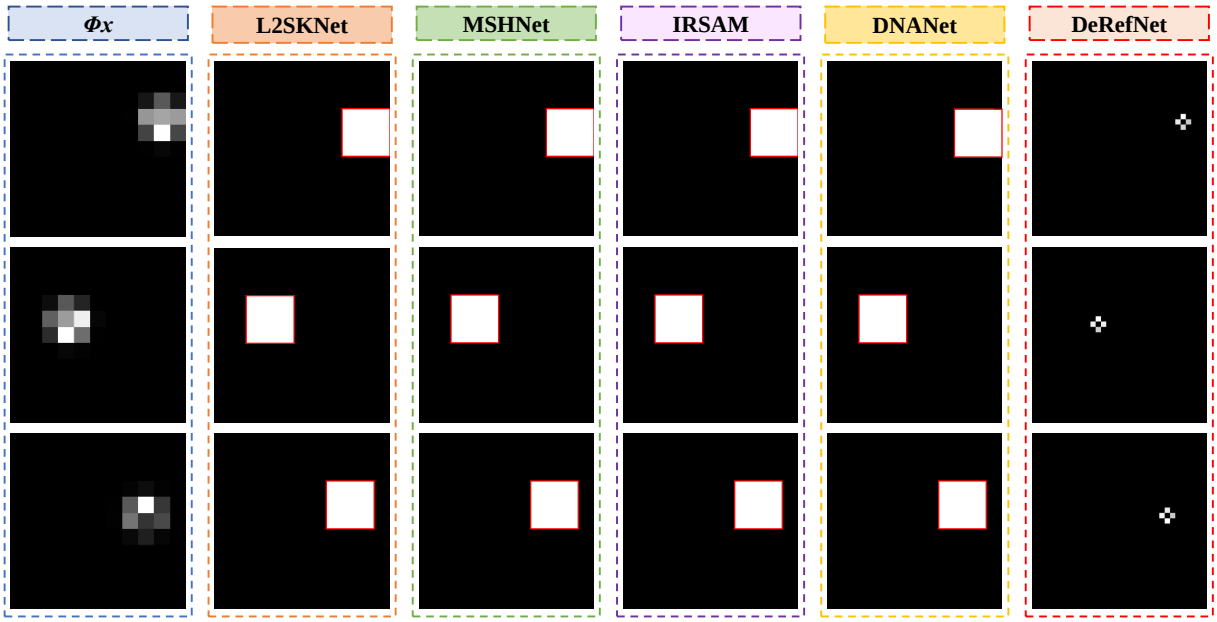

图 5 不同轨迹子序列的部分帧解混效果。该可视化展示了 DeRefNet 在解析邻近红外小目标方面的卓越能力，成功地从传统检测方法只能提供粗略二值掩码结果的密集聚类区域中，提取出单个目标特征和精确的亚像素位置。

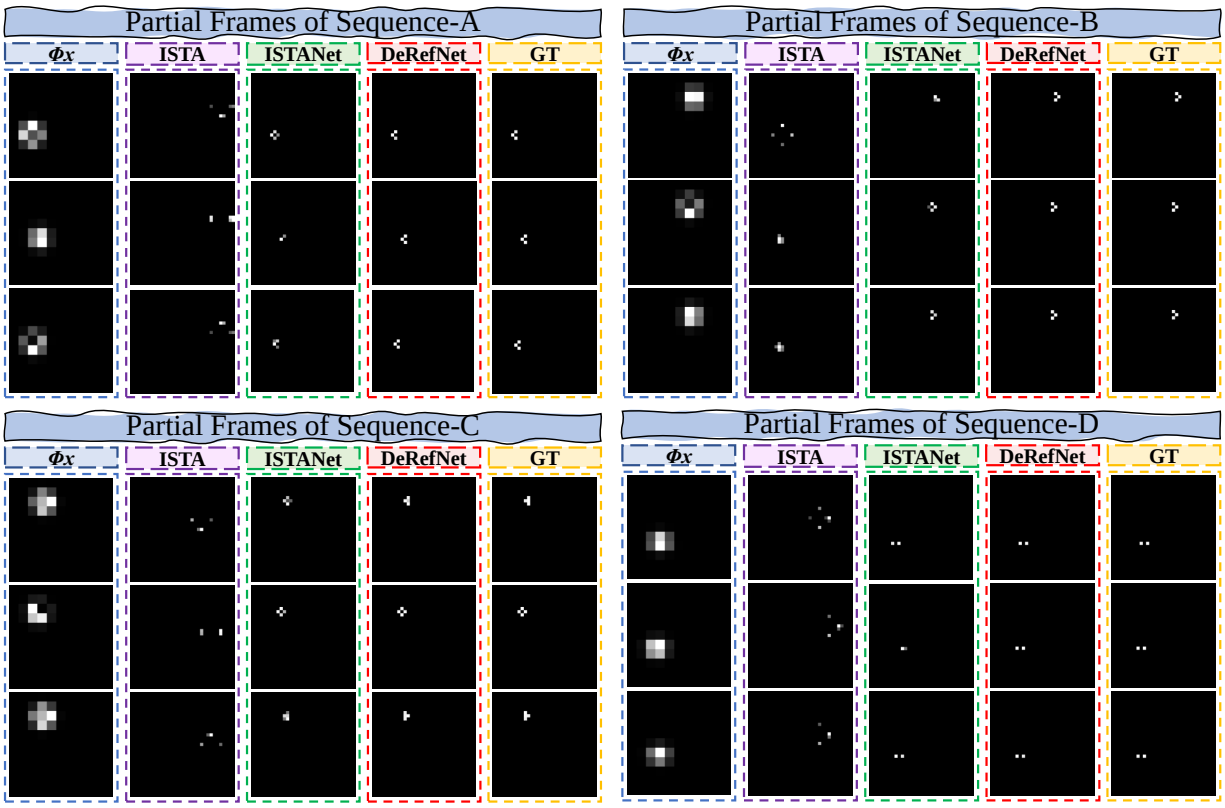

图 6 不同方法在邻近红外小目标序列上的解混性能视觉对比。从左至右：输入的混叠观测值 (Φx)、传统基于迭代的方法 ISTA、基于深度展开的方法 ISTA-Net、空间邻近红外小目标解混方法 DeRefNet 以及真实目标分布。DeRefNet 在准确分离单个目标同时保持清晰边界方面表现出卓越能力，尤其是在传统方法难以处理模糊和混叠特征的空间邻近红外小目标场景中。

5.4 可视化分析

为了更直观地凸显我们模型的优越性，我们提供了一些实验结果的可视化。

如表 8所示，虽然这些红外小目标检测方法实现了很高的检测率（一些模型的 Pd 甚至高达 100%），但它们从根本上解决的是一个不同范围的问题。所有红外小目标检测方法，无论其检测精度如何，都只能提供边界框定位，而无法确定聚集的红外信号内单个目标的数量。这在处理空间邻近红外小目标时是一个根本性的限制，因为在这种情况下，由于大气效应、传感器限制或目标间距离过近，多个目标可能显示为单个模糊的光斑。

表 8 传统检测方法性能概览：用于对比可视化研究的补充评估结果

模型	IoU (%)	Pd (%)	Fa
MSHNet Liu et al. (2024)	99.27	16.67	0.0315
L2SKNet Wu et al. (2024a)	99.96	100.00	0.0001
IRSAM Zhang et al. (2024a)	99.67	100.00	0.0017
DNANet Li et al. (2023a)	99.22	99.36	0.0004

如图 5所示，红外小目标检测方法将聚集的目标视为单个实体，仅提供二值的检测掩码。相比之下，我们的时间可变形对齐机制能够在每个检测到的区域内进行亚像素解混，揭示目标的实际数量及其各自的特征——这些信息对于传统的红外小目标检测方法来说是完全无法获取的。此外，我们还考察了其他代表性的红外小目标检测方法，包括 ISNet Zhang et al. (2022b)、RKFormer Zhang et al. (2022a)、IRPruneDet Zhang et al. (2024b)，以及最近的工作 Unleashing the Power of Generic Segmentation Model Zhang et al. (2024c)。尽管这些方法具有良好的检测性能，但它们同样局限于生成二值检测掩码，并且天生无法揭示密集聚集区域内目标的精确数量或亚像素位置。这进一步凸显了我们的亚像素解混方法在解决空间邻近红外小目标案例中的独特性和必要性。

如图 6所示，该图展示了多个序列的子序列的解混效果。第一列 Φx 是通过光学镜头获得的邻近红外小目标图像，而最后一列 ground truth 显示了实际的目标分布。中间三列分别展示了 ISTA、ISTA-Net 和 DeRefNet 的解混性能。

在不同数量的子目标下，这些模型在解混质量上表现出显著差异。ISTA 难以恢复混叠小目标的位置、轮廓和数量，特别是当它们密集排列时，会导致特征模糊和重叠，难以区分每个子目标。与 ISTA 相比，ISTA-Net 具有更强的特征提取能力，目标定位更准确。这凸显了在有充足数据的情况下，基于模型的深度学习相对于传统纯基于模型的迭代算法在特征提取方面的优势。然而，随着目标数量的增加，ISTA-Net 难以实现清晰的分离，导致恢复结果模糊或重叠。总体而言，ISTA-Net 在解混方面优于 ISTA，但两者都未能完全分离密集的子目标。

相比之下，DeRefNet 始终表现出色，即使在密集排列的场景中，也能准确地解开每个子目标并保持清晰度。这种卓越的解混能力凸显了我们的工作在不同密度下的鲁棒性和优越性能，实现了最精确的目标分离结果。

6 总结

在本研究中，我们提出了一项名为序列化邻近红外小目标解混（Sequential CSIST Unmixing）的新任务，并设计了 DeRefNet 框架。该框架由三个关键部分组成：一个稀疏性驱动的特征提取模块、一个位置编码模块以及一个时间可变形特征对齐（TDFA）模块。本研究首次将深度展开范式引入到序列化邻近红外小目标解混的设计中。通过对比研究和大量实验，我们证明了所提出的 DeRefNet 框架能够有效解决红外图像中的邻近小目标能量混叠问题，实现了目标的解混与亚像素定位。此外，我们还构建了一个用于红外目标解混的开源生态系统，其中包含序列化基准数据集和一个工具包，为相关研究提供了宝贵的资源。

References

CL Philip Chen, Hong Li, Yantao Wei, Tian Xia, and Yuan Yan Tang. A local contrast method for small infrared target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):574–581, 2013. 2

Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10337–10346, 2020. 2

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017. 7

Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Attentional local contrast networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11):9813–9824, 2021. 1

Yimian Dai, Peiwen Pan, Yulei Qian, Yuxuan Li, Xiang Li, Jian Yang, and Huan Wang. Pick of the bunch: Detecting infrared small targets beyond hit-miss trade-offs via selective rank-aware attention. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004. 11

He Deng, Yonglei Zhang, Yuqing Li, Kai Cheng, and Zhong Chen. BEmST: Multi-frame infrared small-dim target detection using probabilistic estimation of sequential backgrounds. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2015. 11

Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015. 7

Shou Feng, Rui Feng, Dan Wu, Chunhui Zhao, Wei Li, and Ran Tao. A coarse-to-fine hyperspectral target detection method based on low-rank tensor decomposition. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 2

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *27th International Conference on Machine Learning*, pages 399–406, 2010. 10, 11

Jinhui Han, Yong Ma, Bo Zhou, Fan Fan, Kun Liang, and Yu Fang. A robust infrared small target detection algorithm based on human visual system. *IEEE Geoscience and Remote Sensing Letters*, 11(12):2168–2172, 2014. 2

Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2018. 11

Yuanxin Huang, Xiyang Zhi, Jianming Hu, Lijian Yu, Qichao Han, Wenbin Chen, and Wei Zhang. LMAFormer: Local motion aware transformer for small moving infrared target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 11

Anat Levin, Yair Weiss, Frédo Durand, and William T Freeman. Blind motion deblurring using image statistics. In *Advances in Neural Information Processing Systems*, pages 841–848, 2007. 11

Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32:1745–1758, 2022. 1

Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32:1745–1758, 2023a. 13

Qilei Li, Zhen Li, Lu Lu, Gwanggil Jeon, Kai Liu, and Xiaomin Yang. Gated multiple feedback network for image super-resolution. In *The British Machine Vision Conference (BMVC)*, 2019. 11

Xin Li, Tao Ma, Yuenan Hou, Botian Shi, Yuchen Yang, Youquan Liu, Xingjiao Wu, Qin Chen, Yikang Li, Yu Qiao, et al. LoGoNet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17524–17534, 2023b. 1

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 11

Liangkui Lin, Hui Xu, Dan Xu, and W An. Resolution of closely spaced objects via infrared focal plane using reversible jump Markov chain Monte-Carlo method. *Acta Optica Sinica*, 31(5):0510004, 2011. 1

Liangkui Lin, Weidong Sheng, and Dan Xu. Bayesian approach to joint super-resolution and trajectory estimation for midcourse closely spaced objects via space-based infrared sensor. *Optical Engineering*, 51(11): 117003–117003, 2012. 4

Jialin Liu and Xiaohan Chen. ALISTA: Analytic weights are as good as learned weights in lista. In *International Conference on Learning Representations*, 2019. 11

Qiankun Liu, Rui Liu, Bolun Zheng, Hongkui Wang, and Ying Fu. Infrared small target detection with scale and location sensitivity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17490–17499, 2024. 4, 13

Daniel Macumber, Sabino Gadaleta, Allison Floyd, and Aubrey Poore. Hierarchical closely spaced object (CSO) resolution for ir sensor surveillance. In *Signal and Data Processing of Small Targets 2005*, pages 32–46. SPIE, 2005. 4

Chris Metzler, Ali Mousavi, and Richard Baraniuk. Learned d-amp: Principled neural network based compressive image recovery. *Advances in Neural Information Processing Systems*, 30, 2017. 11

John T Reagan and Theagenis J Abatzoglou. Model-based superresolution CSO processing. In *Signal and Data Processing of Small Targets 1993*, pages 204–218. SPIE, 1993. 4

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 11

Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. TDAN: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020. 9, 11

Xiaozhong Tong, Zhen Zuo, Shaojing Su, Junyu Wei, Xiaoyong Sun, Peng Wu, and Zongqing Zhao. ST-Trans: Spatial-temporal transformer for infrared small target detection in sequential images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–19, 2024. 2

Di Wang, Jinyuan Liu, Long Ma, Risheng Liu, and Xin Fan. Improving misaligned multi-modality image fusion with one-stage progressive dense registration. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 11

Zhangyang Wang, Qing Ling, and Thomas Huang. Learning deep ℓ_0 encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016. 11

Fengyi Wu, Anran Liu, Tianfang Zhang, Luping Zhang, Junhai Luo, and Zhenming Peng. Saliency at the helm: Steering infrared small target detection with learnable kernels. *IEEE Transactions on Geoscience and Remote Sensing*, 2024a. 13

Fengyi Wu, Tianfang Zhang, Lei Li, Yian Huang, and Zhenming Peng. RPCANet: Deep unfolding RPCA based infrared small target detection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 4809–4818, 2024b. 11

Xin Wu, Danfeng Hong, and Jocelyn Chanussot. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions on Image Processing*, 32:364–376, 2022. 1

Jinxi Xiang, Yonggui Dong, and Yunjie Yang. FISTA-Net: Learning a fast iterative shrinkage thresholding network for inverse problems in imaging. *IEEE Transactions on Medical Imaging*, 40(5):1329–1339, 2021. 11

Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. ADMM-CSNet: A deep learning approach for image compressive sensing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):521–538, 2018. 3

Di You, Jingfen Xie, and Jian Zhang. ISTA-Net++: Flexible deep unfolding network for compressive sensing. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 3, 11

Jian Zhang and Bernard Ghanem. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1828–1837, 2018. 3, 5, 6, 7, 10, 11

Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3217–3226, 2020. 3, 11

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4791–4800, 2021. 11

Mingjin Zhang, Haichen Bai, Jing Zhang, Rui Zhang, Chaoyue Wang, Jie Guo, and Xinbo Gao. Rkformer: Runge-kutta transformer with random-connection attention for infrared small target detection. In *Pro-*

ceedings of the 30th ACM International Conference on Multimedia, pages 1730–1738, 2022a. 13

Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. ISNet: Shape matters for infrared small target detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 877–886, 2022b. 1, 13

Mingjin Zhang, Yuchun Wang, Jie Guo, Yunsong Li, Xinbo Gao, and Jing Zhang. Irsam: Advancing segment anything model for infrared small target detection. In *European Conference on Computer Vision*, pages 233–249. Springer, 2024a. 13

Mingjin Zhang, Handi Yang, Jie Guo, Yunsong Li, Xinbo Gao, and Jing Zhang. IRPruneDet: Efficient infrared small target detection via wavelet structure-regularized soft channel pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7224–7232, 2024b. 13

Mingjin Zhang, Chi Zhang, Qiming Zhang, Yunsong Li, Xinbo Gao, and Jing Zhang. Unleashing the power of generic segmentation model: A simple baseline for infrared small target detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10392–10401, 2024c. 13

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 11

补充材料

1 多目标成像

当目标位于较远距离时，可将其近似为点光源，其在传感器焦平面上的能量分布由 PSF 描述，该函数近似于二维高斯函数。当多个目标在近距离存在时，探测器接收到的能量响应是各 PSF 的线性叠加。这一叠加原理是空间邻近目标解混过程的基础，其中叠加响应必须被分解以识别和表征单个目标，如图 7所示。

2 CSO-mAP 指标

一旦预测结果被分类为真正例（True Positive，TP）或假正例（False Positive，FP），将按照 COCO 数据集标准的常见做法构建一个二进制列表，其中真正例预测用 1 表示，假正例预测用 0 表示。该二进制列表是生成精确率-召回率（Precision Recall，PR）曲线的基础。通过动态调整检测亮度阈值，可以获得一系列精确度和召回率值，从而形成 PR 曲线。平均精确度（Average Precision，AP）作为 PR 曲线下的面积，能够全面评估模型在不同灰度阈值下的性能，准确总结精确度与召回率之间的权衡关系。最后，我们引入 CSO-mAP，其通过对不同距离阈值 δ_k 下的 AP 取平均值，为 CSIST 解混任务中模型性能的比较提供了一个标准化指标。

3 GrokCSO 工具箱

鉴于该领域缺乏专业且易于使用的工具，我们推出了 GrokCSO——一个专为提升空间邻近红外小目标重建效果而设计的综合性开源工具包。尽管通用计算机视觉领域已拥有 MMDetection 和 GluonCV 等丰富的目标检测工具包，但由于红外小目标解混任务的特殊性，迫切需要一个专属平台。缺乏此类平台导致的研究工作碎片化，阻碍了实验的可重复性和算法的比较分析。

基于强大的 PyTorch 框架开发，GrokCSO 经过精心设计，旨在解决 CSIST 解混过程中固有的独特挑战。其独特之处在于：

1. **预训练模型与可复现性**：GrokCSO 为研究者提供了一套完整的预训练模型库，包含前沿算法的训练脚本和日志记录。这些资源不仅增强了研究结果的可复现性，还支持对不同算法方案进行细致入微的对比分析。
2. **定制化灵活性与评估严谨性**：GrokCSO 通过大量可适配的主干网络和颈部结构，支持更广泛的计算策略。该工具包整合了专用数据集加载器、尖端注意力机制以及多功能数据增强流程。更重要的是，GrokCSO 针对空间邻近红外小目标解混任务的特殊性，专门设计了精细化评估指标，这些指标充分尊重了 CSO 问题独有的挑战特性。

4 超参数分析

阶段数量。我们使用阶段数 $K = 2, 4, 6, 8,$ 和 10 对网络进行了训练。性能曲线如图 8 所示。随着阶段数从 2 增加到 6，CSO-mAP 分数有所提升，并在 6 个阶段时达到峰值（46.74%）。此后，增加阶段数带来的收益递减，这可能由于模型复杂度的提升。因此，选择 6 个阶段可在准确性和效率之间实现最优平衡。

动态变换权重。我们调整了动态变换分支相对于两个分支总贡献的比例，并观察性能变化。从 0% 系数开始，随着动态分支影响力的增加，性能逐渐提升，并在 30% 系数时达到峰值。模型在 50% 至 90% 的系数设置下保持稳定性能，展现出鲁棒性。然而，当系数达到 100% 时，性能急剧下降至 30.62%，这证实了合适的动态变换分支系数对实现最优特征表示至关重要。

5 初始化以及学习目标

初始化。 给定数据集 $(z_i, s_i)_{i=1}^M$，其中 z_i 表示混叠目标的图像，s_i 是对应的超分辨图像。我们利用目标信息 (x_i, y_i, g_i) 和缩放因子 c 来计算 s_i 中的高分辨率目标坐标。例如，高分辨率图像中位置 $(c \cdot x_i + \frac{c-1}{2}, c \cdot y_i + \frac{c-1}{2})$ 处的灰度值 g_i 被用于生成 s_i。

令 $Z = [z_1, \ldots, z_M]$ 、$S = [s_1, \ldots, s_M]$。用于初始化 $\tilde{s}^{(0)}$ 的矩阵 Q_{init} 计算如下:

$$Q_{\text{init}} = \arg\min_Q \|QZ - S\|_F^2 = SZ^T(ZZ^T)^{-1}.$$

因此, 初始解混图像 $\tilde{s}^{(0)}$ 的计算表达式为:

$$\tilde{s}^{(0)} = Q_{\text{init}} Z.$$

损失函数。 为确保解混后的图像 $\tilde{s}$ 在保持结构约束 $\tilde{\mathcal{F}}(\cdot) \circ \mathcal{F}_d(\cdot) = \mathbf{I}$ 的同时尽可能接近真实图像 s，我们为 DISTA-Net 设计了如下端到端训练损失函数，其中训练数据集规模为 M，网络包含 N 个阶段，图像尺寸为 N_s:

$$\mathcal{L} = \mathcal{L}_{\text{discrepancy}} + \gamma \mathcal{L}_{\text{constraint}},$$

其中:

$$\mathcal{L}_{\text{discrepancy}} = \frac{1}{MN_s} \sum_{i=1}^{M} \|\tilde{s}_i^{(N)} - s_i\|_2^2,$$

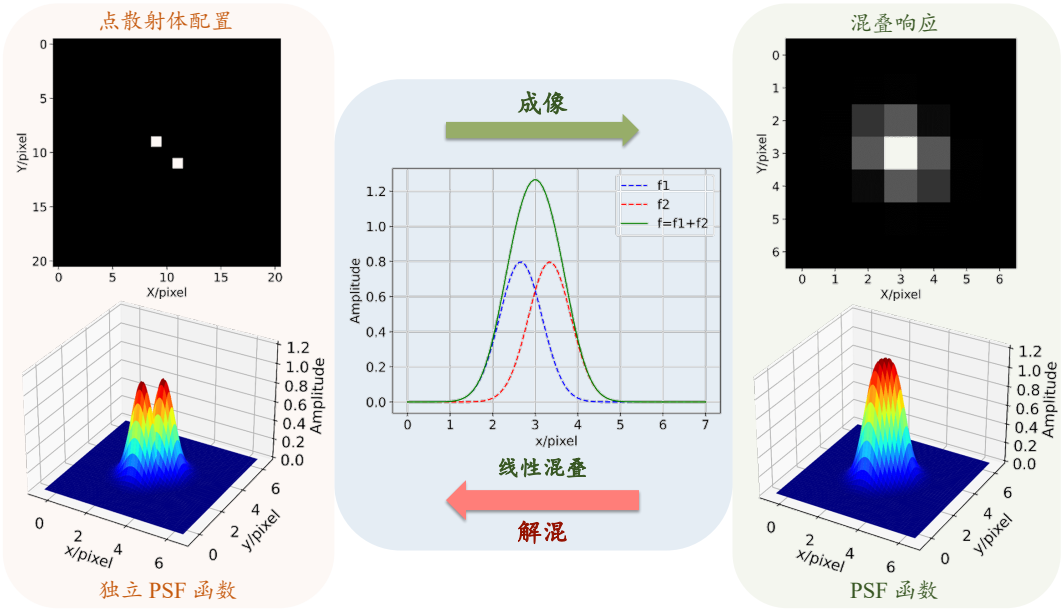

图 7 多目标成像中，远距离目标在成像平面上的成像可视为点源目标通过点扩散函数（Point Spread Function，PSF）进行的能量扩散过程。多目标成像本质上是多个叠加点源累积响应的结果。

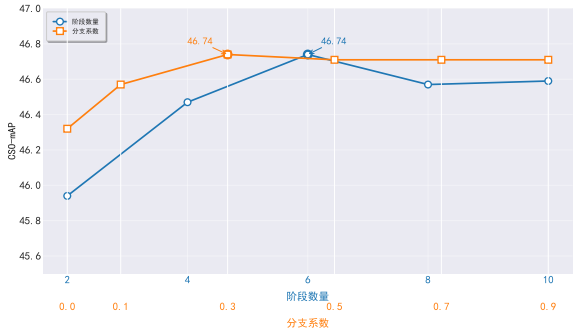

图 8 不同阶段数和分支系数下的 CSO-mAP 性能表现

$$\mathcal{L}_{\text{constraint}} = \frac{1}{MN_s} \sum_{i=1}^{M} \sum_{k=1}^{N} \|\tilde{\mathcal{F}}^{(k)}(\mathcal{F}_d^{(k)}(s_i)) - s_i\|_2^2.$$

其中，$L_{\text{discrepancy}}$ 用于衡量超分辨率图像 $\tilde{s}_i^{(N)}$ 与真实图像 s_i 之间的均方误差。$L_{\text{constraint}}$ 通过确保每阶段 k 中 $\tilde{\mathcal{F}}^{(k)}$ 与 $\mathcal{F}_d^{(k)}$ 的复合变换近似恒等变换，来强制执行结构约束。γ 是用于平衡差异项与约束项的参数。该损失函数旨在平衡解混图像的精度与变换函数的结构完整性。