arXiv:2301.11274v1 [cs.CV] 26 Jan 2023

基于跨输入一致性的自监督 RGB-T 跟踪

Xingchen Zhang*，*IEEE* 会员，Yiannis Demiris，*IEEE* 高级会员

摘要—本文提出一种自监督 RGB-T 跟踪方法。不同于现有使用大量带标注 RGB-T 图像对进行训练的深度 RGB-T 跟踪器，我们的 RGB-T 跟踪器以自监督方式使用无标注RGB-T视频对进行训练。我们提出了一种新颖的基于跨输入一致性的自监督训练策略，其依据是跟踪可以通过不同输入来完成这一理念。具体而言，我们使用无标注RGB-T视频对构建两种不同的输入。随后，我们分别使用这两种输入跟踪目标并生成结果，并据此构建跨输入一致性损失。同时，我们提出一种重加权策略，使损失函数能够适应低质量训练样本。我们的跟踪器基于孪生相关滤波网络构建。据我们所知，这是首个自监督 RGB-T 跟踪器。在两个公开的 RGB-T 跟踪基准上进行的大量实验表明，所提出的训练策略行之有效。值得注意的是，尽管仅使用无标注 RGB-T 视频对语料进行训练，我们的跟踪器在 GTOT 数据集上的表现仍优于七种有监督 RGB-T 跟踪器。

索引词—RGBT 跟踪、目标跟踪、热成像图像、图像融合、信息融合

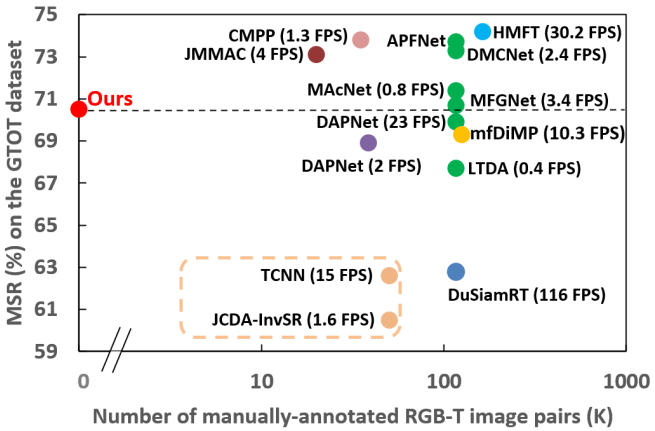


图 1. 性能与训练所需标注 RGB-T 图像对数量的关系。尽管我们的 RGB-T 跟踪器不需要任何标注训练数据，但在 GTOT 数据集 [12] 上的表现优于七种有监督 RGB-T 跟踪器。请注意，TCNN [13] 和 JCDA-InvSR [14]（均为有监督跟踪器）所用标注数据的确切数量并未在其论文中提及。

I. 引言

OBJECT 标跟踪是一项重要任务，在机器人和监控等领域有着广泛应用。近年来，人们提出了许多跟踪算法，跟踪性能也得到了显著提升。然而，大多数视觉跟踪器都在 RGB图像上运行。当 RGB图像不可靠时（例如处于较差的光照条件下），这些跟踪器的性能会显著下降，从而限制了其实际应用。

为提升跟踪性能，研究人员将热成像图像与 RGB图像结合起来执行 RGB-T 跟踪 [1–7]。这是基于以下事实：热成像图像对光照变化不敏感，而 RGB图像包含更多纹理细节 [8]。尽管人们已投入大量精力开发基于深度学习的 RGB-T 跟踪器，且 RGB-T 跟踪性能得到了显著提升，但现有的深度 RGB-T 跟踪器 [6, 9–11]需要大量经过标注的 RGB-T 图像对，如图1所示。

众所周知，标注既耗时又昂贵。为避免对标注的需求，研究者提出了一些无监督单目标跟踪器。例如，Wang等人 [15, 16] 和 Zhu 等人 [17]基于前向-后向跟踪的循环一致性来训练跟踪器。还有一些跟踪器在训练期间只使用极少量的标注，例如，仅标注初始帧 [18–20]。然而，所有无监督跟踪器都只使用单模态图像，即只使用 RGB图像 [15, 16, 18–24] 或热成像图像[25]。

本文提出了一种自监督 RGB-T 跟踪器，在训练过程中无需任何人工标注。为此，我们提出了一种跨输入一致性训练策略，以利用无标注 RGB-T 视频中的时间信息。我们的思路源于这样一种观察：目标跟踪可以使用不同的输入来完成。如图2所示，给定帧 t中的一个目标，我们可以跟踪它，从而获得其在帧$(t+1)$ 中的位置，所用输入可以不同（例如 RGB图像、热成像图像，或二者的组合）。理想情况下，如果所有跟踪均成功，那么帧$(t+1)$中的跟踪结果应当一致。

我们将自监督训练策略融入基于孪生网络的判别式相关滤波器（DCF）框架。在实现中，我们为跟踪构建了两个不同的输入，以建立跨输入一致性。这种跨输入一致性基于无标注 *RGB-T* 视频对中的时间信息，可用于指导 RGB-T 跟踪器的训练。此外，由于我们不想使用任何人工标注，因此在训练数据中随机初始化一个边界框。因此，训练样本通常带有噪声或质量较差。我们提出一种重新加权策略，对损失函数进行重新加权，使训练更加轻松、有效。总之，主要贡献

X. Zhang 和 Y. Demiris 就职于伦敦帝国理工学院电气与电子工程系个人机器人实验室，地址为英国伦敦 SW7 2AZ（电子邮箱：xingchen.zhang@imperial.ac.uk，y.demiris@imperial.ac.uk）

* 通讯作者：Xingchen Zhang

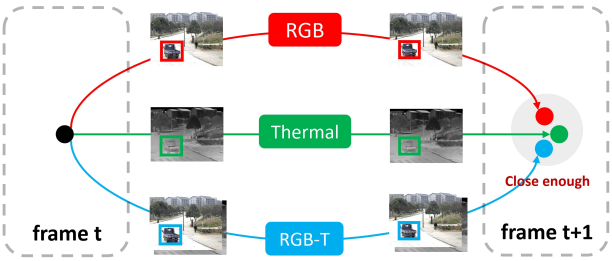


图 2. 我们的跨输入一致性基于这样一种观察：使用不同输入均可进行跟踪。理想情况下，利用不同输入获得的帧（$t+1$）跟踪结果应足够接近。

本文的内容包括：

- 我们提出了一种利用 RGB-T 视频对进行训练的自监督 RGB-T 跟踪器，无需人工标注。
- 我们提出了一种基于跨输入一致性的策略，以实现自监督训练。我们使用 RGB图像和热成像图像构造不同的跟踪输入，并据此构建跨输入一致性损失来指导训练。
- 我们提出了一种重加权方案，对损失函数进行重新加权，以提高训练效果。
- 在两个 RGB-T 跟踪基准上的大量实验表明，所提出的方法表现出色，也展现了自监督 RGB-T 跟踪的潜力。

本文其余部分安排如下。第II 节介绍相关工作。随后，第III 节详细介绍所提方法，接着在第IV节介绍训练数据处理。然后，第V节展示实验结果，第VI 节进行讨论。最后，第VII节对本文进行总结。

II. 相关工作

A. 单目标跟踪

单目标跟踪方法主要包括基于深度学习的方法 [16, 26, 27] 和基于判别相关滤波器（DCF）的方法 [28, 29]。大多数跟踪器以RGB图像作为输入，并且对良好的光照条件要求较高。为了使跟踪器不受光照条件影响，一些研究人员使用热成像图像进行跟踪 [30, 31]。然而，热成像图像缺乏足够的纹理细节，因此在光照条件良好时，其性能不如基于 RGB 的跟踪器。

B. *RGB-T* 跟踪

为缓解基于 RGB 和热成像的跟踪器所面临的问题，研究人员开展了 RGB-T 跟踪 [1, 2, 9,32]。例如，Zhang 等人 [32] 提出了一种基于像素级融合的 RGB-T 跟踪器。相比之下，一些 RGB-T 跟踪器采用特征级融合 [7, 33, 34]、决策级融合[35] 或结合多个融合层级 [9]。RGB-T 跟踪器的性能已得到显著提升。然而，现有的深度 RGB-T 跟踪器需要大量 RGB-T 图像对进行训练。

C. 无监督目标跟踪

研究人员提出了无监督跟踪器，以减轻对标注的需求。例如，Vondrick 等人 [21]提出通过视频着色来训练 RGB 跟踪器。Wang 等人 [15, 16] 和 Shen 等人 [23] 提出利用循环一致性训练 RGB 跟踪器。Yuan 等人 [18] 和 Shen 等人 [23] 进一步在循环一致性框架中使用区域提议网络。还有一些基于不同思路的无监督 RGB 跟踪器，例如循环记忆学习 [20], 、裁剪-变换-粘贴操作[22], ， 以及利用图像及其裁剪区域进行训练[36]。此外，也有研究提出了基于循环一致性的无监督热成像跟踪器 [25]。然而，现有的无监督跟踪器仅限于单一模态，即仅基于 RGB 图像或仅基于热成像图像。

D. 自监督训练

一些自监督学习方法，例如 BYOL [37]和 SimCLR [38], ，会先使用不同的数据增强生成两个相关视图，然后最大化相似度，以学习用于下游任务的表征。我们的想法受到这些自监督学习方法的启发。然而，在我们的工作中，我们使用来自不同模态的图像替代传统的数据增强。此外，我们采用输入不同的目标跟踪框架，利用 RGB-T 视频对中的时间信息，并构建跨输入一致性来指导训练。此外，与许多研究不同，我们不使用前置任务和下游任务，而是仅执行一项任务（RGB-T 跟踪）。我们直接使用所提出的训练方法来获得一个 RGB-T 跟踪器。

E. 跨输入一致性

跨输入一致性很少用于跟踪。Bastani 等人 [39]将跨输入一致性应用于开发一种自监督多目标跟踪器。我们的工作受到 [39]的启发，旨在训练一个 RGB-T 单目标跟踪器。我们利用 RGB-T 视频对作为不同输入来构建跨输入一致性。

III. 所提方法

本文的基本思想（见图2）是，使用不同输入进行目标跟踪，以生成一致的结果。具体而言，本文构建了两种不同的输入，即 RGB 图像和 RGB-T 图像对，以实现跨输入一致性，如图 3(a)所示。RGB 输入由 RGB跟踪器处理，RGB-T 输入则由我们提出的 RGB-T跟踪器处理。我们在孪生网络式DCF跟踪框架中实现了跨输入一致性自监督训练策略。

A. 背景：孪生网络式 *DCF*跟踪器

给定来自无标签视频的连续两帧，我们首先裁剪出模板块 **T** 和搜索块 **S**。在基于孪生网络的 DCF 跟踪器中， [15, 40], 首先使用卷积神经网络从以下位置提取特征： **T** 和 **S**。然后，学习一个滤波器**W** 。

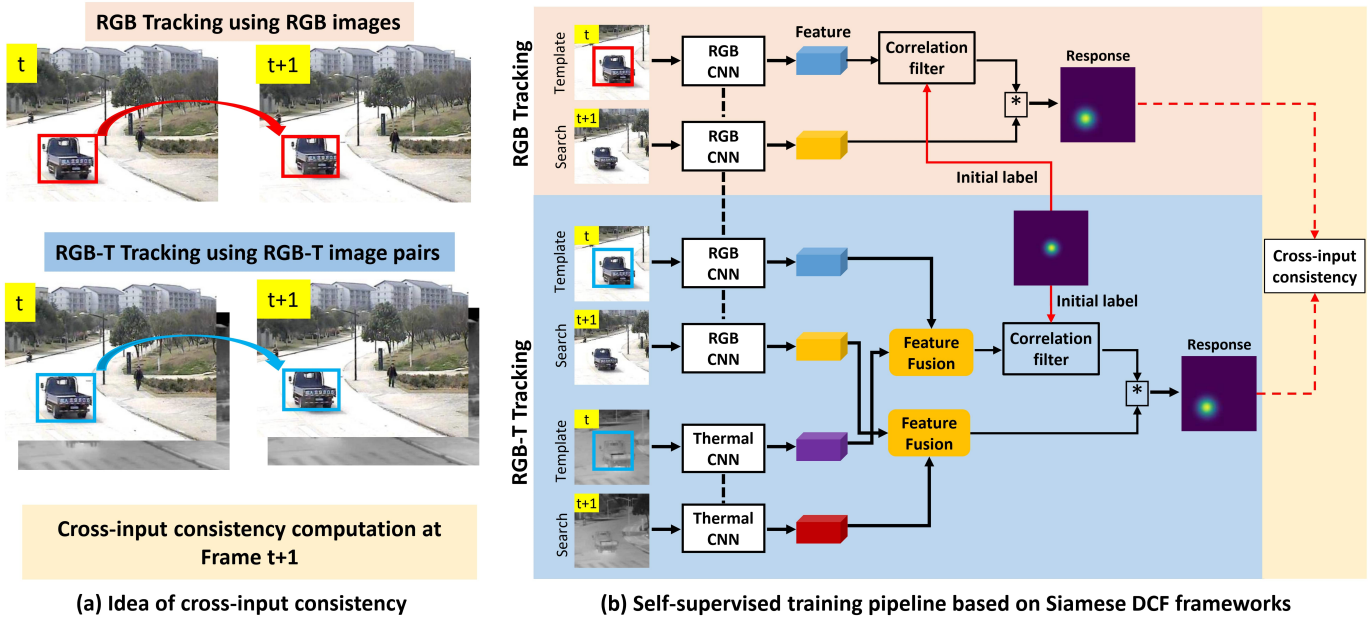


图 3. 我们方法的概览。(a) 使用 RGB 和 RGB-T 图像作为输入的跨输入一致性的基本思想。(b) 基于孪生 DCF 跟踪框架的训练流程。损失函数根据响应图计算。通过特征级融合获得融合后的模板特征和搜索特征。**(b) 中的蓝色部分用于训练后的目标跟踪**。

通过将$\mathbf{W}$ 与搜索块 $\mathbf{S}$的特征进行卷积，即可生成响应图。响应图用于目标定位。具体而言，RGB图像的滤波器$\mathbf{W}$可表示为

$$\mathbf{W}_{\mathbf{T}_{\mathrm{RGB}}}=\mathcal{F}^{-1}\left(\frac{\mathcal{F}(\phi_{\mathrm{RGB}}(\mathbf{T}_{\mathrm{RGB}}))\odot\mathcal{F}^{\star}(\mathbf{Y}_{\mathbf{T}_{\mathrm{RGB}}})}{\mathcal{F}^{\star}(\phi_{\mathrm{RGB}}(\mathbf{T}_{\mathrm{RGB}}))\odot\mathcal{F}(\phi_{\mathrm{RGB}}(\mathbf{T}_{\mathrm{RGB}}))+\lambda}\right), \tag{1}$$

其中，$\odot$ 表示逐元素乘积，$\mathcal{F}$ 表示离散傅里叶变换（DFT），$\mathcal{F}^{-1}$ 表示逆离散傅里叶变换，$\star$ 表示复共轭操作，$\phi_{\mathrm{RGB}}()$ 是用于提取 RGB 特征的卷积神经网络，$\mathbf{Y}_{\mathbf{T}_{\mathrm{RGB}}}$ 是 RGB 模板块的标签，它是一个以边界框区域为中心的高斯响应图。一旦获得滤波器 $\mathbf{W}_{\mathbf{T}_{\mathrm{RGB}}}$，RGB 搜索块 $\mathbf{S}_{\mathrm{RGB}}$ 的响应图便为

$$\mathbf{R}_{\mathbf{S}_{\mathrm{RGB}}}=\mathcal{F}^{-1}(\mathcal{F}^{\star}(\mathbf{W}_{\mathbf{T}_{\mathrm{RGB}}})\odot\mathcal{F}(\phi_{\mathrm{RGB}}(\mathbf{S}_{\mathrm{RGB}}))). \tag{2}$$

在基于 DCF 的跟踪器中使用卷积神经网络的主要优势在于，卷积神经网络与相关滤波层集成在一个端到端框架中。因此，卷积神经网络能够学习提取更适合跟踪的特征。我们的 RGB跟踪器和 RGB-T跟踪器都采用基于孪生网络的 DCF 框架，但RGB卷积神经网络和热红外卷积神经网络的权重不同。

B. 跨输入一致性

如图3所示，我们的框架采用两种不同的输入来构建跨输入一致性。第一输入是 RGB图像，第二输入是 RGB-T 图像对。我们的自监督训练策略的核心思想是：如果 RGB跟踪器和 RGB-T跟踪器均能良好运行，那么无论使用哪种输入进行跟踪，我们都能确定目标在帧$(t+1)$ 中的位置，t即目标在帧中的位置。

在训练过程中，以边界框区域为中心的高斯响应图被用作 RGB跟踪器和 RGB-T跟踪器的初始标签。我们使用跨输入一致性损失来指导 RGB跟踪器和 RGB-T跟踪器的联合训练。主要目标是训练跟踪器中的卷积神经网络模型，使其学习适用于跟踪的特征。**在推理阶段，我们仅使用 RGB-T跟踪器，以 RGB 和热成像图像作为输入执行跟踪，如图3(b)**所示。

我们的跨输入一致性具有通用性。在本研究中，我们使用不同模态的视频来构建跨输入一致性。也可以采用其他方案来构建跨输入一致性，并取得相当或更好的性能。此外，正如我们将在实验中展示的那样，我们还可以在热成像图像与 RGB-T 图像对之间，或在 RGB 图像、热成像图像与 RGB-T 图像对之间构建跨输入一致性。

C. 我们的 *RGB-T*跟踪器

我们的 RGB-T 跟踪器架构如图 3(b)的蓝色部分所示。如图所示，我们的 RGB-T 跟踪器由两个 RGB 卷积神经网络和两个热红外卷积神经网络组成。两个 RGB 卷积神经网络用于提取 RGB 模板特征和搜索特征，两个热红外卷积神经网络用于提取热红外模板特征和搜索特征。RGB 模板特征和热红外模板特征融合后得到融合模板特征，而 RGB 搜索特征和热红外搜索特征融合后得到融合搜索特征。然后，按照 [41],利用融合模板特征和融合搜索特征，通过相关滤波器和循环卷积运算生成响应图，即

$$\mathbf{R}_{\mathbf{S}_{\mathrm{RGBT}}}=\mathcal{F}^{-1}(\mathcal{F}^{\star}(\mathbf{W}_{\mathbf{T}_{\mathrm{RGBT}}})\odot\mathcal{F}(\phi_{\mathrm{RGB}}(\mathbf{S}_{\mathrm{RGB}})\oplus\phi_{\mathrm{T}}(\mathbf{S}_{\mathrm{T}}))), \tag{3}$$

其中，$\oplus$ 表示特征融合。随后可根据响应图获得跟踪结果。

1）特征融合：特征融合可以通过多种方式实现。在本研究中，为了使我们的 RGB-T 跟踪器更加轻量，从而能够快速运行，我们没有采用复杂的特征融合模块。相反，我们将RGB特征和热特征拼接起来，生成融合特征。正如我们将在第 V-B 节所示，这种方法简单而有效。

2）在线目标跟踪：我们首先进行离线训练，以训练卷积神经网络。然后，使用 RGB-T 跟踪器执行在线跟踪。在跟踪过程中，所有卷积神经网络均保持固定。遵循先前的研究 [15, 16, 41]，我们更新 RGB-T 跟踪器中的判别相关滤波器参数，以提高跟踪器的鲁棒性，即

$$\mathbf{W}^t_{\mathrm{RGBT}} = (1-\alpha_t)\mathbf{W}^{t-1}_{\mathrm{RGBT}} + \alpha_t \mathbf{W}_{\mathrm{RGBT}}, \tag{4}$$

其中 α_t 是控制更新速度的参数。

D. 跨输入一致性损失函数

理想情况下，如果所有跟踪器都工作良好，那么不同输入所得到的跟踪结果应当相同。我们通过损失函数最小化不同输入下得到的响应图之间的差异。具体而言，我们的跨输入一致性损失为

$$\mathcal{L} = ||\mathbf{R}_{\mathbf{S}_{\mathrm{RGB}}} - \mathbf{R}_{\mathbf{S}_{\mathrm{RGBT}}}||, \tag{5}$$

其中，$\mathbf{R}_{\mathbf{S}_{\mathrm{RGB}}}$ 是由 RGB 跟踪器生成的响应图，而 $\mathbf{R}_{\mathbf{S}_{\mathrm{RGBT}}}$ 是由 RGB-T 跟踪器生成的响应图。

IV. 训练数据处理与损失函数重加权

A. 训练数据处理

我们不希望在训练中使用任何人工标签。因此，在自监督训练中获取良好的初始边界框（伪标签）至关重要。在这项工作中，我们从 RGB-T 视频对中裁剪出中心图像块来生成训练数据，这与 Wang 等人所做的方法相同。 [15]通过这种方式，我们跟踪出现在裁剪区域中心的目标。请注意，中心的目标可能只是目标的一部分。图中展示了一些裁剪图像的示例。4。可以看到，一些裁剪图像包含有用的运动目标，而另一些图像仅包含背景信息。本研究受 Wang 等人的启发，提出了几种方法来提高这些训练数据的利用率。 [15]。

噪声样本丢弃。裁剪得到的中心图块包含噪声样本，会产生很大的损失值。这些噪声样本会使训练过程不稳定，效果也较差。我们为每个训练对分配一个权重值$\mathbf{D}^i_{\mathrm{noisy}}$，以排除损失值极高的 10% 训练对。根据我们的观察，这些样本通常包含摄像机突然运动或外观剧烈变化。不同于 Wang 等人 [15] 利用响应图的方法，我们使用 RGB 模板块与 RGB 搜索块之间的差异，即

$$\mathbf{D}^i = \frac{||\mathbf{T}^i_{\mathrm{RGB}} - \mathbf{S}^i_{\mathrm{RGB}}||_2^2}{\mathrm{H}\times\mathrm{W}}, \tag{6}$$

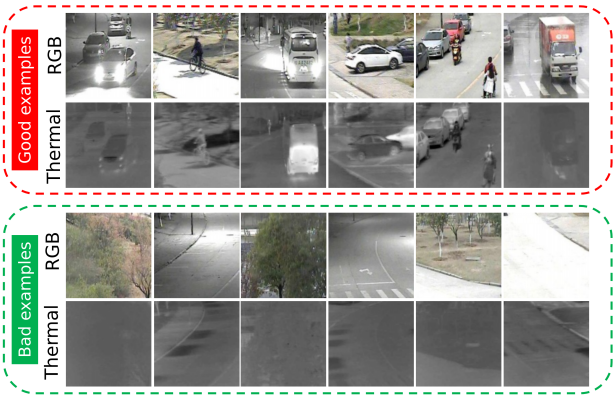


图 4. RGBT234 [1]中的裁剪图像块示例。上：良好示例。下：不良示例。

其中，H 和 W 分别表示训练样本的高度和宽度。我们对 $\mathbf{D}$中的元素进行排序。然后，为每个训练对分配一个权重值$\mathbf{D}^i_{\mathrm{noisy}}$。权重向量$\mathbf{D}^{\mathrm{noisy}}$ 中对应噪声样本的 10% 元素设为 0。这样，我们便排除了产生较大差异值的 10% 训练对。

背景样本丢弃。如图 4所示，一些裁剪的中心图块仅包含背景或静止物体。这些背景样本对模型训练的贡献很小。为排除这些背景训练样本，我们为每个训练对设置一个值 $\mathbf{D}^i_{\mathrm{background}}$。权重向量 25% 的元素 $\mathbf{D}^{\mathrm{background}}$ 中对应于 25% 个最低值的元素 $\mathbf{D}$ 被设为零。结合 $\mathbf{D}^{\mathrm{noisy}}$ 和 $\mathbf{D}^{\mathrm{background}}$，我们可以对每个训练对的权重进行归一化，以确保一个小批次中有用权重之和为 1，即

$$\mathbf{D}^i_{\mathrm{norm}} = \frac{\mathbf{D}^i_{\mathrm{noisy}} \cdot \mathbf{D}^i_{\mathrm{background}}}{\sum_{i=1}^{n} \mathbf{D}^i_{\mathrm{noisy}} \cdot \mathbf{D}^i_{\mathrm{background}}}, \tag{7}$$

其中 n 是一个小批次中的训练对数量。

B. 损失函数重加权

使用公式（7）计算重加权权重向量后，我们使用它对从不同质量训练样本中获得的损失进行重加权，即

$$\mathcal{L}_{\mathrm{final}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{D}^i_{\mathrm{norm}} \cdot \mathcal{L}^i, \tag{8}$$

其中 $\mathcal{L}^i$ 根据公式（5）计算。使用该损失可以使训练更加有效，并避免过拟合。

V. 实验

实现细节。随后 [15, 16]，，我们在跟踪器中使用轻量级卷积神经网络。具体而言，我们的卷积神经网络中两个卷积层的滤波器尺寸分别为 $3\times3\times3\times32$和 $3\times3\times32\times32$。所有实验均使用一台配备两块英伟达 RTX3090 GPU 和一颗 i9-10900X 中央处理器的台式机完成。批次大小为 32。我们将学习率从 10^{-4} 调整为 10^{-6}，调整过程从第 0 个训练轮次持续到第 30 个训练轮次。权重衰减为 5×10^{-5}。

测试集。我们在 GTOT 数据集 [12]和 RGBT234 数据集上报告结果，这两个数据集已被广泛用于 RGB-T 跟踪研究 [9, 10, 42]。GTOT 包含 50 个 RGB-T 视频（15.8K 帧）。此外，每个序列都标注了七种属性，包括遮挡（OCC）、大尺度变化（LSV）、快速运动（FM）、低照度（LI）、热交叉（TC）、小目标（SO）和形变（DEF）。RGBT234 包含 234 个 RGB-T 视频对（约 233.8K 帧），并标注了 12 种属性。与 GTOT 相比，RGBT234 更具挑战性，因为其视频包含更长的帧序列以及更具挑战性的属性。

训练数据。在 GTOT 数据集上进行测试时，我们使用 RGBT234 数据集 [1] 作为训练数据。训练过程中随机选取 10,000 对 RGB-T 作为验证集。在 RGBT234 数据集上进行测试时，我们使用 GTOT 数据集作为训练数据，并在训练过程中随机选取 1000 对 RGB-T 作为验证集。

评估指标。在本文中，我们采用 RGB-T 跟踪中两种常用的评估指标——最大精确率（MPR）和最大成功率（MSR）[1, 34], 来评估跟踪器的性能。遵循以往研究 [1, 6, 12], , 对于 GTOT，MPR 的阈值设为 5 像素（因为 GTOT 中的目标相对较小），对于 RGBT234 则设为 20 像素。

A. 自监督训练与监督训练

为展示自监督训练策略的有效性，我们使用 RGBT234 数据集的真实标注来训练有监督 RGB-T 跟踪器。具体而言，我们仅训练图3(b)蓝色部分所示的 RGB-T 跟踪器。有监督 RGB-T 跟踪器与我们的自监督 RGB-T 跟踪器之间的比较结果见表I。可以看出，在 GTOT 上，我们的自监督 RGB-T 跟踪器性能优于有监督跟踪器。这一点颇为有趣，也令人意外，因为使用真实标注标签进行训练通常更有效。一个可能的原因是，将 RGBT234（包含 110K 对 RGB-T 图像）中的中心裁剪区域用作训练数据后，训练集包含的目标类别比真实标注标签更多。一些无监督 RGB 跟踪研究也观察到了类似现象 [43], , 其中无监督 DCFNet 的表现略优于有监督 DCFNet。我们还使用 GTOT 数据集的真实标注训练了一个有监督 RGB-T 跟踪器，并在 RGBT234 上对其进行测试。从表I可以看出，我们的自监督 RGB-T 跟踪器表现略逊于有监督版本。这可能是因为 GTOT 数据集规模较小（包含 7.9K 对 RGB-T 图像），未能为我们的自监督训练提供足够的高质量训练数据。

表 I 所提出的自监督训练与监督训练的比较。较好的结果以粗体标示。

变体	GTOT		RGBT234	
	MPR(↑)	MSR(↑)	MPR(↑)	MSR(↑)
有监督	76.7	66.1	**59.5**	**43.7**
Ours	**85.6**	**70.5**	56.2	41.6

表 I 第一分支输入的影响。

第一输入	第二输入	MPR(↑)	MSR(↑)
RGB	RGB-T	**85.6**	**70.5**
热成像	RGB-T	75.4	64.9
四通道可见光-热红外	RGB-T	81.6	67.8

B. 消融研究与分析

除非另有说明，消融研究均使用 GTOT 数据集。

不同的跨输入组合。我们可以使用不同的输入组合来构建跨输入一致性。在本研究中，我们固定第二输入（RGB-T 图像对），并将第一输入替换为不同的变体。具体而言，我们训练了两种变体。在第一种变体中，我们使用热成像图像作为第一输入。在第二种变体中，我们使用四通道可见光-热红外图像作为第一输入。使用四通道可见光-热红外图像时，我们调整第一卷积层的维度，以适配这些四通道图像。如表II所示，所有这些组合都可用于指导基于跨输入一致性的训练，并得到有效的 RGB-T 跟踪器。在这些组合中，当第一输入为 RGB 图像、第二输入为 RGB-T 图像对时，我们的 RGB-T 跟踪器表现出最佳的跟踪性能。

在跨输入一致性中使用更多分支。此前，我们使用两个分支（不同的输入）来构建跨输入一致性，如图3所示。我们可以将这一思路扩展到更多分支。例如，可以使用三种输入，即 RGB 图像、热成像图像和 RGB-T 图像对。在这种情况下，我们可以再增加两个损失项，即 $\mathcal{L}_{\mathrm{RGB-T}}$ 和$\mathcal{L}_{\mathrm{T-RGBT}}$，以计算每两种输出之间的差异。采用这一思路，我们得到一个 MPR 为 83.7、MSR 为 70.3 的 RGB-T 跟踪器，其性能优于使用四通道 RGBT 图像和 RGB-T 图像对的方案。不过，其性能略逊于将 RGB 图像和 RGB-T 图像对作为不同输入的方案。

损失函数的影响。均方误差损失（MSE损失）通常用于训练无监督跟踪器 [15, 16, 25]。在本研究中，我们使用L1损失代替。表III 展示了在基于交叉输入一致性的自监督训练中，使用均方误差损失和L1损失的性能对比。可以看出，L1损失在所有情况下都能取得更好的性能。这可能是因为我们的训练样本含有噪声（尽管我们采用了重新加权策略），均方误差损失会放大噪声训练样本带来的误差，从而降低训练效果。

损失重加权的影响。我们提出了两个基于裁剪图像块生成权重向量的组件，即噪声样本丢弃和背景样本丢弃。本节报告了在表IV中移除其中一个组件后的结果。可以看出，移除任一组件后，跟踪性能都会略有下降，这表明我们的损失重加权策略是有效的。

无标签训练数据规模的影响。我们使用 RGBT234 的不同比例作为训练集。可以看出，总体而言，所提出的自监督训练策略能够从使用更多

表 III 损失函数的影响。在这些变体中，L1损失的性能优于均方误差损失。

变体		MPR(↑)	MSR(↑)
三分支	均方误差损失	72.5	63.1
	L1损失	75.3	65.6
RGB，RGB-T	均方误差损失	78.8	64.9
	L1损失	**85.6**	**70.5**
热成像，RGB-T	均方误差损失	65.7	58.2
	L1损失	75.4	64.9

表 IV 损失重加权方案的影响。对损失函数进行重加权后，结果更佳。

损失重加权		MPR(↑)	MSR(↑)
噪声样本丢弃	背景样本丢弃		
		82.6	69.1
	✓	83.7	69.5
✓	✓	**85.6**	**70.5**

无标注RGB-T视频对进行训练中获益。由于无标注RGB-T 视频对比有标注的视频对更容易获取，我们的方法表明，无监督RGB-T跟踪具有巨大的潜力。

表 V 训练数据规模的消融研究。使用更多无标注RGB-T视频进行训练时，所提出的RGB-T跟踪器在GTOT 数据集上取得了更好的结果。

Size	MPR(↑)	MSR(↑)	Size	MPR(↑)	MSR(↑)
RGBT234（90%）	**85.6**	**70.5**	RGBT234（50%）	81.5	67.5
RGBT234（70%）	81.2	67.3	RGBT234（20%）	73.6	61.3

特征融合方法的影响。我们使用三种特征级融合方法训练了三种RGB-T跟踪器变体，分别是逐元素平均、拼接以及Zhang 等人提出的DFF融合模块。 [9]。具体而言，我们保留图3(b) 中的RGB跟踪器，并更改RGB-T跟踪器中的特征融合方法。我们在表VI中比较了不同的融合方法，结果表明，融合层级对RGB-T跟踪器的性能有显著影响。具体而言，拼接RGB特征和热特征可获得最佳跟踪性能。

表 VI 特征融合方法的影响。拼接方法简单，却能带来最佳性能。

变体	MPR(↑)	MSR(↑)
特征级（平均）	83.7	69.1
特征级（拼接）	**85.6**	**70.5**
特征级（DFF [9]）	79.1	66.0

RGB卷积神经网络与热红外卷积神经网络之间共享权重的影响。在实验中，我们发现，RGB卷积神经网络与热红外卷积神经网络是否共享权重（请参见论文中的图3(b) ）会影响跟踪器的性能。我们设计了所提方法的一个变体，即让RGB卷积神经网络与热红外卷积神经网络共享权重。结果

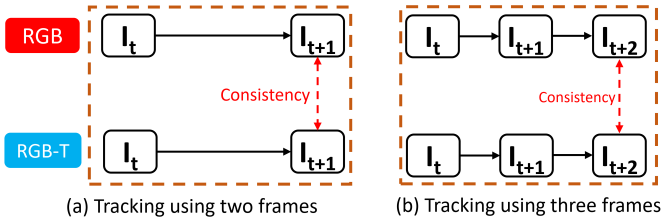


图5. 使用两帧（a）和三帧（b）进行训练。 红色箭头：跨输入一致性。

如表VII所示，我们的RGB-T跟踪器性能优于RGB卷积神经网络与热红外卷积神经网络共享权重的变体。

表VII RGB卷积神经网络与热红外卷积神经网络之间共享权重对跟踪性能的影响。使用GTOT 数据集。

变体	MPR(↑)	MSR(↑)
共享	82.2	68.2
不共享	**85.6**	**70.5**

跟踪序列长度的影响。本文中，我们使用两帧构建基于跨输入一致性的训练策略，即帧 t 和帧 $(t+1)$，如图5(a) 所示。实际上，我们的基于跨输入一致性的训练策略可以扩展到更多帧，例如帧 t、$(t+1)$和$(t+2)$，如图 5所示。在这种情况下，在帧 $(t+1)$ 生成的响应图将用作帧 $(t+1)$ 的伪标签。基于该伪标签，我们的跟踪器从 $(t+1)$ 跟踪目标至帧 $(t+2)$。

使用更多帧时，跨输入一致性损失在最后一帧计算，如图 5(b)中的红色虚线箭头所示。表 VIII展示了跟踪序列长度对跟踪性能的影响。从结果可以看出，使用三帧同样可以训练出良好的 RGB-T 跟踪器。然而，其性能略逊于训练时使用两帧的情况。这些结果表明，我们的自监督训练策略只需在每个训练对中使用两帧，就能有效训练 RGB-T 跟踪器。

表 VIII 跟踪序列长度对跟踪性能的影响。使用 GTOT 数据集。

变体	MPR(↑)	MSR(↑)
使用三帧	84.3	69.5
使用两帧	**85.6**	**70.5**

C. 与循环一致性的比较

循环一致性通常用于训练无监督的基于 RGB 的跟踪器 [15–19, 23] 或热红外跟踪器 [25]。本节将我们的跨输入一致性与循环一致性进行比较。具体而言，我们使用循环一致性训练 RGB-T 跟踪器（图3(b)中的蓝色部分）。我们训练了两种变体，一种使用两帧，另一种使用三帧，如图 6所示。其他所有

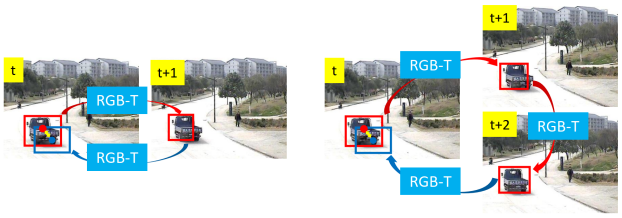


图 6. 使用两帧（左）和三帧（右）的循环一致性。

设置均与基于跨输入一致性的训练保持相同。从表IX可以看出，我们基于跨输入一致性的自监督训练策略比循环一致性更有效。表 XI 中给出的属性性能也表明，我们基于跨输入一致性的训练比基于循环一致性的训练更有效。此外，在基于循环一致性的训练中使用三帧比使用两帧更有效。除了 RGB-T 跟踪器之外，我们还使用训练好的卷积神经网络，基于图3(b)中的红色部分运行 RGB 跟踪器和热成像跟踪器。可以看出，我们的训练策略在 RGB 跟踪器、热成像跟踪器和 RGB-T 跟踪器上均带来了稳定的性能提升。原因在于，基于前向-后向跟踪的循环一致性在某些具有挑战性的真实世界跟踪场景中可能并不成立 [22]。此外，值得一提的是，基于循环一致性的策略（如 UDT [15],）仅应用于单模态目标跟踪，而我们提出的基于跨输入一致性的自监督训练策略则应用于多模态跟踪。

表IX 循环一致性性能比较。

变体	RGB跟踪器		热跟踪器		RGB-T跟踪器	
	PR	SR	PR	SR	MPR	MSR
循环一致性（2帧）	67.6	56.7	62.4	56.0	74.1	64.5
循环一致性（3帧）	64.5	55.6	62.8	56.2	75.4	65.6
Ours	**68.1**	**57.8**	**64.8**	**57.3**	**85.6**	**70.5**

D. 与最先进水平的 *RGB-T* 跟踪器进行比较

对比方法。目前尚无现有的无监督深度RGB-T跟踪器。我们选取了以下有监督 RGB-T 跟踪器进行比较，即 HMFT [9],CMPP [10], DMCNet [11], JMMAC [6], CBPNet [46],MaCNet[2], MANet [47], CMP [48], MFGNet [49],DAFNet [50], DAPNet [51], mfDiMP [52],LTDA [43], DuSiamRT [53], TCNN [13], JCDA-InvSR [14], CFNet [40]+RGBT,SiamDW [54]+RGBT。此外，我们还选取了一种基于Transformer的方法APFNet[45]，进行比较。我们还选取了一些非深度 RGB-T 跟踪器，包括 CMCF [55],NRCMR[44], LGMG [42], CMR [56], SGT [57], Li等人的方法 [58], CSR [12], [59],MEEF[60]+RGBT 和 KCF[61]+RGBT。几乎涵盖了所有类别的 RGB-T 方法。

GTOT 上的结果。GTOT 数据集上的跟踪结果如表X所示。如图所示，我们的 RGB-T跟踪器在两项指标上均表现更佳。此外，我们的自监督 RGB-T 跟踪器优于七种有监督 RGB-T 跟踪器（DAFNet、DAPNet、mfDiMP、LTDA、DuSiamRT、TCNN、JCDA-InvSR）以及 CFNet 的 RGB-T 版本。我们的跟踪器与若干有监督跟踪器（如 MFGNet、CMP 和 MANet）的性能也不相上下。尽管我们的性能与最先进的有监督 RGB-T 跟踪器仍有差距，但这是可以理解的，因为这些有监督 RGB-T 跟踪器使用大规模标注的 RGB-T 图像对进行训练。相比之下，我们的方法没有使用任何标注。此外，这些跟踪器采用了更复杂的模型，这一点从其跟踪速度也可看出。例如，CMPP 和 DMCNet 的每秒帧数分别为 1.3 和 2.4。

基于属性的结果。表 XI 展示了 GTOT 数据集上的基于属性的性能。如表所示，作为一种无需真实标注标签进行训练的轻量级 RGB-T 跟踪器，我们的跟踪器在 LSV、LI、TC 和 SO 方面取得了极具竞争力的性能。尤其是，我们的 RGB-T 跟踪器在 LSV 方面的性能优于大多数有监督 RGB-T 跟踪器，这表明我们的跟踪器能够很好地应对目标的尺度变化。

RGBT234 上的结果。表 X 展示了 RGBT234 数据集上的结果。从表中可以看出，所提出的 RGB-T 跟踪器在 MSR 方面优于四种有监督跟踪器和三种非学习型 RGB-T 跟踪器。然而，我们的 RGB-T 跟踪器在 RGBT234 上的性能不如在 GTOT 上。这是因为 RGBT234 比 GTOT 更具挑战性：其图像数量更多（223.8K 帧，而 GTOT 为 15.8K 帧），且具有挑战性的场景更多（12 种具有挑战性的属性，而 GTOT 为 7 种属性）。大多数深度 RGB-T 跟踪器通过采用复杂的模型架构并使用大量带标注的 RGB-T 图像对进行训练，在 RGBT234 上取得了良好的性能。相比之下，我们的跟踪器使用 6,900 对无标注 RGB-T 图像对进行训练。在未来的工作中，我们将尝试通过采用更好的骨干跟踪器或更大的训练集来缩小这一差距。不过，值得一提的是，我们的自监督 RGB-T 跟踪器与最先进的有监督 RGB-T 跟踪器 HMFT 之间的性能差距是可以接受的：MPR 相差 22.6%，MSR 相差 15.2%。最先进的无监督 RGB 跟踪器与有监督 RGB 跟踪器之间也存在类似的性能差距。例如，ULAST [23] 在 TrackingNet 上的精确率为 59.2%，成功率为 65.4%；而 SwinTrack [62] 分别达到 82.8%和 84.0%。

E. 定性结果

1) 与最先进水平的跟踪器相比：我们首先在图中展示了我们的 RGB-T跟踪器与其他 RGB-T 跟踪器的定性比较。7。我们选取了若干 RGB-T 跟踪器，包括 SGT[57],LGMG [42], DAPNet [51],MaCNet [2]。可以看出，我们的 RGB-T跟踪器在以下四个序列上表现更佳，分别是*BlackCar*、*Exposure2*、*LightOcc*和*carNig*。

2) 困难案例：本节展示了图8中的一些困难案例的定性结果。在这些案例中，基于 RGB 图像和热成像图像进行跟踪都会失败，而我们的 RGB-T 跟踪器能够正确跟踪目标。这些

表X 与现有RGB-T 跟踪器在GTOT 数据集和RGBT234 数据集上的比较。这些带有 † 标记的结果由我们使用原始跟踪结果计算得出。这些带有 § 标记的结果复制自 [1, 12, 44]。其他结果摘自相应论文。“-”表示相应论文中未提及。比我们方法差的结果以绿色标记。

跟踪器	GTOT		RGBT234		每秒帧数 (↑)	类别	有监督	发表场所
	MPR (↑)	MSR (↑)	MPR (↑)	MSR (↑)				
HMFT† [9]	90.6	74.2	78.8	56.8	30.2	基于深度学习	Yes	CVPR 2022年
CMPP [10]	92.6	73.8	82.3	57.5	1.3	基于深度学习	Yes	CVPR 2020年
APFNet [45]	90.5	73.7	82.7	57.9	-	基于深度学习	Yes	AAAI 2022
DMCNet [11]	90.9	73.3	83.9	59.3	2.4	基于深度学习	Yes	IEEE 神经网络与学习系统汇刊 2022年
JMMAC† [6]	89.3	73.1	79.0	57.3	4	基于深度学习	Yes	IEEE 图像处理汇刊 2021年
CBPNet [46]	88.5	71.6	79.4	54.1	3.7	基于深度学习	Yes	IEEE 多媒体汇刊 2021年
MaCNet [2]	88.0	71.4	79.0	55.4	0.8	基于深度学习	Yes	Sensors 2020
MANet† [47]	88.9	71.1	77.7	53.9	3.1	基于深度学习	Yes	ICCVW 2019
CMP [48]	86.9	71.1	75.1	49.1	35	基于深度学习	Yes	Neurocomputing 2021
MFGNet [49]	88.9	70.7	78.3	53.5	3.4	基于深度学习	Yes	IEEE 多媒体汇刊 2022年
DAFNet† [50]	88.6	69.9	79.6	54.4	23	基于深度学习	Yes	ICCVW 2019
DAPNet† [51]	87.4	68.9	76.6	53.7	2	基于深度学习	Yes	ACM 多媒体大会 2019年
mfDiMP§ [52]	84.1	69.3	78.5	55.9	10.3	基于深度学习	Yes	ICCVW 2019
LTDA [43]	84.3	67.7	78.7	54.5	0.4	基于深度学习	Yes	ICIP 2019
DuSiamRT [53]	76.6	62.8	56.7	38.4	116	基于深度学习	Yes	《The Visual Computer》2022年
TCNN [13]	85.2	62.6	-	-	15	基于深度学习	Yes	神经计算 2018年
JCDA-InvSR [14]	-	60.5	60.6	41.4	1.6	基于机器学习	Yes	IEEE 图像处理汇刊 2019年
CFNet [40] + RGBT§	-	-	55.1	39.0	-	基于深度学习	Yes	CVPR 2017
SiamDW [54] + RGBT§	68.0	56.5	60.4	39.7	-	基于深度学习	Yes	CVPR 2019
CMCF [55]	77.0	63.2	-	-	227	非深度学习（基于相关滤波）		神经计算 2019年
NRCMR [44]	83.7	66.4	72.9	50.2	7	非深度学习（基于图）		IEEE TNNLS 2021
LGMG [42]	83.7	65.8	-	-	7	非深度学习（基于图）		IEEE TCSVT 2019
CMR [56]	82.7	64.3	-	-	8	非深度学习（基于图）		ECCV 2018
SGT§ [57]	85.1	62.8	72.0	47.2	5	非深度学习（基于图）		ACM MM 2017
[58]	84.2	62.2	-	-	7	非深度学习（基于图）		SPIC 2018
CSR [12]	74.5	61.5	46.3	32.8	1.6	非深度学习（基于超分辨率）		IEEE TIP 2016
[59]	77.3	61.2	72.9	48.6	1.2	非深度学习（基于图）		Neurocomputing 2022
MEEF[60] + RGBT§	-	52.0	63.6	40.5	4.9	非深度学习		ECCV 2014
KCF [61] + RGBT§	-	42.0	46.3	30.5	124.1	非深度学习（基于相关滤波）		IEEE TPAMI 2014
Ours	85.6	70.5	56.2	41.6	179.3	基于深度学习	No	

表 XI 基于属性的性能（成功率）在 GTOT 上的表现。低于我们方法的结果以绿色标记。

方法	OCC	LSV	FM	LI	TC	SO	DEF
HMFT	72.1	74.0	75.2	75.6	72.3	70.8	72.3
MANet	68.9	69.4	69.2	71.9	69.1	68.6	73.4
CBPNet	68.6	68.8	67.5	73.1	69.3	69.0	75.4
MaCNet	68.7	67.3	65.9	73.1	69.7	69.5	76.5
DAPNet	67.4	64.8	61.9	72.2	69.0	69.2	77.1
DuSiamRT	57.7	64.5	58.0	62.3	61.4	64.2	62.9
TCNN	55.6	64.6	51.7	64.2	59.5	59.1	73.4
SGT	56.7	54.7	55.9	65.1	61.5	61.8	73.3
Cycle-2	59.1	67.4	60.3	63.9	65.7	61.9	61.3
Cycle-3	62.0	67.1	62.9	67.9	65.6	61.9	65.0
Ours	65.9	72.2	65.2	73.0	68.8	66.9	67.7

示例表明，我们的自监督 RGB-T 跟踪器能够有效融合 RGB 和热信息，从而提升跟踪性能。例如，在 *LightOcc* 案例中，当汽车被树木严重遮挡时，RGB 跟踪器和热成像跟踪器都难以正常工作。相比之下，我们的自监督 RGB-T 跟踪器能够通过融合 RGB 特征和热特征，成功跟踪汽车。请注意，RGB 跟踪器使用的是我们的 RGB 卷积神经网络，而热成像跟踪器使用的是我们的热红外卷积神经网络。

F. 失败案例

我们的自监督 RGB-T 跟踪器在一些极具挑战性的案例中会跟踪失败。图 9 展示了我们方法的失败案例。在这些案例中，RGB跟踪器、热成像跟踪器和 RGB-T跟踪器均跟踪失败。在 *Pool* 案例中，树木造成的遮挡，以及树木与行人衣物颜色相近，是导致失败的主要原因。在 *RainyCar2* 案例中，树木造成的遮挡，以及白色汽车与道路颜色相近，是导致失败的主要原因。

VI. 讨论

我们自监督训练策略的优势。首先，所提出的训练策略使我们能够在没有任何人工标注的情况下训练 RGB-T 跟踪器。一些自监督跟踪方法 [18, 19, 22] 仍然需要一些标注（非常稀疏）。一些自监督跟踪方法使用无监督光流模型 [20, 23] 来生成伪标签，或使用 EdgeBox [24] 来生成候选目标。我们的方法完全不需要任何标注或光流模型，同时仍能取得合理的跟踪性能，这一点非常有益。其次，我们的跨输入一致性具有灵活性。具体而言，我们可以使用不同模态作为不同输入，这非常适合不同模态的视觉数据，例如 RGB图像和热成像图像。这一理念有望应用于其他模态，例如 RGB-D跟踪。我们也可以轻松地添加或移除分支

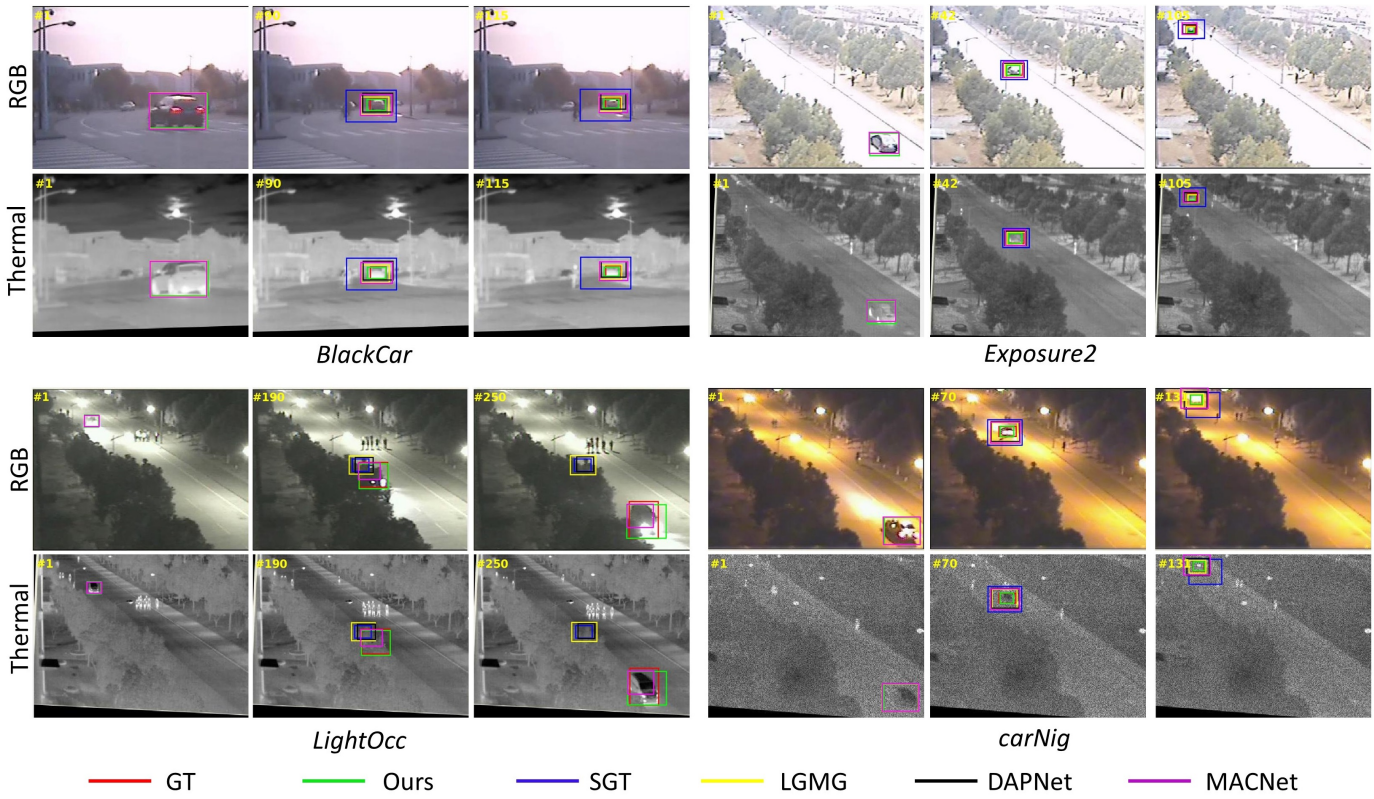
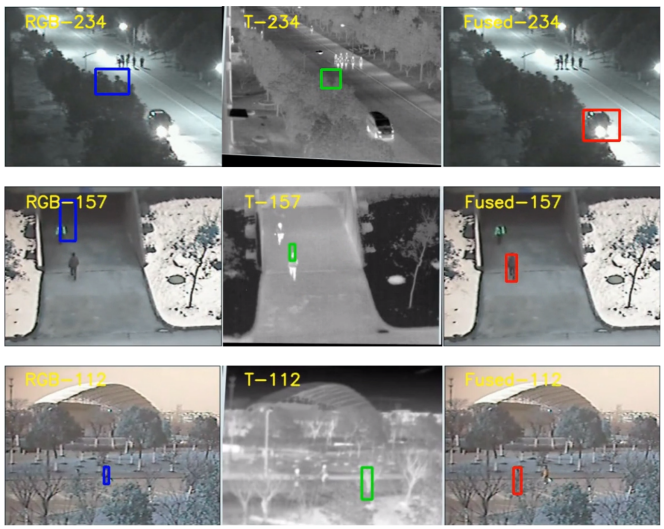


图7. 我们的 RGB-T 跟踪器与其他 RGB-T 跟踪器的定性比较。四个序列分别为来自 GTOT 数据集的 *BlackCar*、*Exposure2*、*LightOcc* 和 *carNig* 。我们展示了 RGB 图像（第1行和第3行）以及热成像图像（第2行和第4行）上的结果。



使用 RGB 图像进行跟踪和使用热成像图像进行跟踪均会失败。我们的 RGB-T 跟踪器能够成功跟踪目标。从上到下：*LightOcc*、*Quarreling*、*WalkingOcc*。左侧：RGB 跟踪器。中间：热成像跟踪器。右侧：我们的自监督 RGB-T 跟踪器。



我们的 RGB-T 跟踪器的失败案例。展示了 *Pool* （上）和 *Rainy- Car2* （下）的结果。左侧：RGB 跟踪器。中间：热成像跟踪器。右侧：我们的自监督 RGB-T 跟踪器。

（输入）。第三，所提出的自监督训练策略具有通用性。理论上，可以使用不同的骨干跟踪器。我们将在未来对此进行探索。

模型坍塌。在我们基于跨输入一致性的方法中，可能存在一种坍塌解，即不同的分支总是生成相同的错误结果。然而，我们的实验中从未出现过这种坍塌解。我们的跟踪器能够有效避免这一问题。一个可能的原因是，我们的输入来自不同的模态，因此不同分支不太可能生成相同的特征。

局限性：我们的方法与最先进的有监督 RGB-T 跟踪器之间存在明显的性能差距。未来应投入更多努力来缩小这一差距。一种可行的解决方案是将我们的自监督训练策略应用于性能更好的骨干跟踪器，例如 SiamAtt [63],SiamDW [64], 和 ATOM [65]。另一种有前景的方法是使用规模更大的 RGB-T 跟踪数据集进行训练。此外，

我们没有学习用于边界框回归的 IoU 网络。这是无监督跟踪中的一个常见问题 [15–17, 19]。我们也将在未来的研究中解决这一问题。不过，值得一提的是，尽管所提出的自监督方法与最先进的有监督方法之间存在性能差距，但所提方法具有很大的潜力，因为它不需要任何人工标注。

七、结论

本文提出了一种基于跨输入一致性的自监督训练策略，用于训练 RGB-T 跟踪器。我们利用 RGB图像和热红外图像构造两种不同的输入，然后根据这两种输入得到的跟踪结果计算跨输入一致性损失。我们还提出了一种损失重加权方案，以改进训练。所提方法的主要优点是训练只需使用无标注RGB-T视频对。实验表明，所提方法取得了良好的性能。具体而言，采用非常简单的卷积神经网络和简单的特征融合方法时，我们的 RGB-T 跟踪器在 GTOT 数据集上的表现优于若干有监督 RGB-T 跟踪器。我们还表明，使用更多无标注RGB-T视频可以进一步提升跟踪性能。未来，我们将采用更复杂的网络，并在框架中加入边界框回归模块。

致谢

本研究获得了欧盟“地平线2020”研究与创新计划的资助，资助依据为玛丽·斯克沃多夫斯卡-居里资助协议，编号为101025274。本工作还获得了英国皇家工程院新兴技术讲席项目的资助，受资助人为YD。

参考文献

[1] C. Zhao、J. Tang、Y. Luo、“RGB-T目标跟踪：基准与基线”，《模式识别》，第106977页，2019年。[2] H. Zhang、L. Zhang、L. Zhuo和J. Zhang，“利用模态感知注意力网络和竞争学习进行RGB-T视频目标跟踪”，《传感器》，第20卷，第2期，第393页，2020年。[3] X. Zhang、P. Ye、S. Peng、J. Liu、K. Gong和G. Xiao，“SiamFT：一种基于全卷积孪生网络的RGB-红外融合跟踪方法”，《IEEE Access》，第7卷，第122 122–122 133页，2019年。[4] X. Zhang、P. Ye、S. Peng等，“DSiamMFT：一种利用多层特征融合的动态孪生网络实现的RGB-T融合跟踪方法”，《信号处理：图像通信》，第115756页，2020年。[5] X. Zhang、P. Ye、H. Leung等，“基于可见光和红外图像的目标融合跟踪：全面综述”，《信息融合》，第63卷，第166–187页，2020年。[6] P. Zhang、J. Zhao、C. Bo、D. Wang、H. Lu和X. Yang，“联合建模运动与外观线索以实现鲁棒的RGB-T跟踪”，《IEEE图像处理汇刊》，第30卷，第3335–3347页，2021年。[7] A. Lu、C. Li、Y. Yan、J. Tang和B. Luo，“基于多适配器网络与层次散度损失的RGBT跟踪”，《IEEE图像处理汇刊》，2021年。[8] X. Zhang、P. Ye和G. Xiao，“VIFB：可见光与红外图像融合基准”，载于《IEEE/CVF计算机视觉与模式识别会议研讨会论文集》，2020年，第104–105页。

[9] P. Zhang、J. Zhao、D. Wang、H. Lu 和 X. Ruan，《可见光-热成像无人机跟踪：大规模基准与新基线》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2022年，第8886–8895 页。[10] C. Wang、C. Xu、Z. Cui、L. Zhou、T. Zhang、X. Zhang 和 J. Yang，《用于 RGB-T 跟踪的跨模态模式传播》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第 7064–7073 页。[11] A. Lu、C. Qian、C. Li、J. Tang 和 L. Wang，《用于 RGBT 跟踪的双重门控互条件网络》，IEEE神经网络与学习系统汇刊，2022年。[12] C. Li、H. Cheng、S. Hu、X. Liu、J. Tang 和 L. Lin，《灰度-热红外跟踪中的协同稀疏表示学习》，IEEE图像处理汇刊，第 25 卷，第 12 期，第5743–5756 页，2016年。[13] C. Li、X. Wu、N. Zhao、X. Cao 和 J. Tang，《融合双流卷积神经网络进行 RGB-T目标跟踪》，神经计算，第 281 卷，第 78–85 页，2018年。[14] B. Kang、D. Liang、W. Ding、H. Zhou 和 W.-P. Zhu，《通过基于逆稀疏表示的协同编码进行灰度-热红外跟踪》，IEEE图像处理汇刊，第 29 卷，第 3401–3415 页，2019年。[15] N. Wang、Y. Song、C. Ma、W. Zhou、W. Liu 和 H. Li，《无监督深度跟踪》，载于 IEEE计算机视觉与模式识别会议论文集，2019年，第 1308–1317 页。[16] N. Wang、W. Zhou、Y. Song、C. Ma、W. Liu 和 H. Li，《用于实时跟踪的无监督深度表征学习》，国际计算机视觉杂志，第 129 卷，第 2 期，第 400–418 页，2021年。[17] J. Zhu、C. Ma、S. Jia 和 S. Xu，《用于无监督视觉跟踪的对比循环一致性学习》，载于中国模式识别与计算机视觉会议（PRCV）。Springer，2021年，第 564–576 页。[18] W. Yuan、M. Y. Wang 和 Q. Chen，《利用循环一致孪生网络进行自监督目标跟踪》，载于 2020年 IEEE/RSJ智能机器人与系统国际会议（IROS）。IEEE，2020年，第 10 351–10 358 页。[19] D. Yuan、X. Chang、P.-Y. Huang、Q. Liu 和 Z. He，《自监督深度相关跟踪》，IEEE图像处理汇刊，第 30 卷，第 976–985 页，2021年。[20] J. Zheng、C. Ma、H. Peng 和 X. Yang，《学习从无标签视频中跟踪目标》，载于 IEEE/CVF国际计算机视觉会议论文集，2021年，第 13 546–13 555 页。[21] C. Vondrick、A. Shrivastava、A. Fathi、S. Guadarrama 和 K. Murphy，《通过视频着色实现跟踪》，载于欧洲计算机视觉会议（ECCV）论文集，2018年，第 391–408 页。[22] X. Li、W. Pei、Z. Zhou、Z. He、H. Lu 和 M.-H. Yang，《通过目标感知数据合成进行自监督跟踪》，arXiv 预印本 arXiv:2106.10900，2021年。[23] Q. Shen、L. Qiao、J. Guo、P. Li、X. Li、B. Li、W. Feng、W. Gan、W. Wu 和 W. Ouyang，《无监督精准孪生跟踪学习》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2022年，第8101–8110 页。[24] Q. Wu、J. Wan 和 A. B. Chan，《面向视觉目标跟踪的渐进式无监督学习》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第 2993–3002 页。[25] Y. Huang、Y. He、R. Lu、X. Li 和 X. Yang，《利用无监督深度相关滤波器进行热红外目标跟踪》，数字信号处理，第 123 卷，第103432 页，2022 年。[26] G. Bhat 、 M. Danelljan 、 L. V. Gool 和 R. Timofte，《学习用于跟踪的判别式模型预测》，载于 IEEE/CVF国际计算机视觉会议论文集，2019年，第 6182–6191 页。

[27] X. Li、C. Ma、B. Wu、Z. He 和 M.-H. Yang，《目标感知深度跟踪》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2019年，第 1369–1378 页。[28]Y. Li、C. Fu、F. Ding、Z. Huang 和 G. Lu，《Auto-track：借助自动时空正则化实现高性能无人机视觉跟踪》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第 11 923–11 932 页。[29] G. Zheng、C. Fu、J. Ye、F. Lin 和 F. Ding，《用于自适应混合标签实时无人机跟踪的突变敏感相关滤波器》，载于 2022年 IEEE机器人与自动化国际会议（ICRA），2021年。[30] M. Felsberg、A. Berg、G. Hager、J. Ahlberg、M. Kristan、J. Matas、A. Leonardis、L. Cehovin、G. Fernandez、T. Vojir 等，《热红外视觉目标跟踪 VOT-TIR2015 挑战赛结果》，载于 IEEE国际计算机视觉会议研讨会论文集，2015年，第 76–88 页。[31] Q. Liu、X. Li、Z. He、N. Fan、D. Yuan 和 H. Wang，《学习用于热红外目标跟踪的深度多层相似性》，IEEE多媒体汇刊，第 23 卷，2126，2020年 2114页。– 。[32] X. Zhang、P. Ye、D. Qiao、J. Zhao、S. Peng 和 G. Xiao，《基于全卷积孪生网络的可见光与红外图像目标融合跟踪》，载于 2019年第三十二届信息融合国际会议。IEEE，2019年，第 1–8 页。[33] C. Li、L. Liu、A. Lu、Q. Ji 和 J. Tang，《挑战感知 RGBT 跟踪》，载于欧洲计算机视觉会议。Springer，2020年，第 222–237 页。[34] P. Zhang、D. Wang、H. Lu 和 X. Yang，《学习用于实时 RGB-T 跟踪的自适应属性驱动表征》，国际计算机视觉杂志，第 1–16 页，2021年。[35] X. Zhang、P. Ye、J. Liu、K. Gonge 和 G. Xiao，《通过孪生网络进行决策级可见光与红外融合跟踪》，载于第九届中国信息融合大会，2019年。[36] C. H. Sio、Y.-J. Ma、H.-H. Shuai、J.-C. Chen 和 W.-H. Cheng，《S2siamfc：用于视觉跟踪的自监督全卷积孪生网络》，载于第 28 届 ACM 多媒体国际会议论文集，2020年，第 1948–1957 页。[37] J.-B. Grill、F. Strub、F. Altché、C. Tallec、P. Richemond、E. Buchatskaya、C. Doersch、B. Avila Pires、Z. Guo、M. Ghesh-laghi Azar 等，《自举潜在表征：一种自监督学习的新方法》，神经信息处理系统进展，第 33 卷，第 21 271–21 284 页，2020年。[38] T. Chen、S. Kornblith、M. Norouzi 和 G. Hinton，《视觉表征对比学习的简单框架》，载于机器学习国际会议。PMLR，2020年，第 1597–1607 页。[39] F. Bastani、S. He 和 S. Madden，《利用跨输入一致性进行自监督多目标跟踪》，神经信息处理系统进展，第 34 卷，2021年。[40] J. Valmadre、L. Bertinetto、J. Henriques、A. Vedaldi 和 P. H. Torr，《用于基于相关滤波器跟踪的端到端表征学习》，载于 IEEE计算机视觉与模式识别会议论文集，2017年，第 2805–2813 页。[41] Q. Wang、J. Gao、J. Xing、M. Zhang 和 W. Hu，《DCFNet：用于视觉跟踪的判别式相关滤波器网络》，arXiv 预印本 arXiv:1704.04057，2017年。[42] C. Li、C. Zhu、J. Zhang、B. Luo、X. Wu 和 J. Tang，《学习用于 RGB-T目标跟踪的局部-全局多图描述符》，IEEE视频技术电路与系统汇刊，第 29 卷，第 10 期，第 2913–2926 页，2019年。[43] R. Yang、Y. Zhu、X. Wang、C. Li 和 J. Tang，《学习面向目标的双重注意力，实现鲁棒RGB-T跟踪》，载于 2019年 IEEE图像处理国际会议（ICIP）。IEEE，2019年，第 3975–3979 页。[44] C. Li、Z. Xiang、J. Tang、B. Luo 和 F. Wang，《通过抗噪跨模态排序进行 RGBT 跟踪》，IEEE汇刊神经网络与学习系统，2021年。[45] Y. Xiao、M. Yang、C. Li、L. Liu 和 J. Tang，“用于 RGBT 跟踪的基于属性的渐进式融合网络”，载于 AAAI人工智能会议论文集，2022年。[46] Q. Xu、Y. Mei、J. Liu 和 C. Li，“用于 RGBT 跟踪的多模态跨层双线性池化”，IEEE多媒体汇刊，第24卷，567–580页，2021年。[47] C. Li、A. Lu、A. Hua Zheng、Z. Tu 和 J. Tang，“多适配器RGBT跟踪”，载于 IEEE/CVF国际计算机视觉会议研讨会论文集，2019年，页码 0–0。[48] R. Yang、X. Wang、C. Li、J. Hu 和 J. Tang，“通过跨模态消息传递进行RGBT跟踪”，神经计算，第462卷，365–375页，2021年。[49] X. Wang、X. Shu、S. Zhang、B. Jiang、Y. Wang、Y. Tian 和 F. Wu，“MFGNet：用于 RGB-T 跟踪的动态模态感知滤波器生成”，IEEE多媒体汇刊，2022年。[50] Y. Gao、C. Li、Y. Zhu、J. Tang、T. He 和 F. Wang，“用于高性能RGBT跟踪的深度自适应融合网络”，载于 IEEE/CVF国际计算机视觉会议研讨会论文集，2019年，页码 0–0。[51] Y. Zhu、C. Li、B. Luo、J. Tang 和 X. Wang，“用于 RGBT 跟踪的密集特征聚合与剪枝”，载于第27届ACM国际多媒体会议论文集，2019年，465–472页。[52] L. Zhang、M. Danelljan、A. Gonzalez-Garcia、J. van de Weijer 和 F. Shahbaz Khan，“用于端到端 RGB-T 跟踪的多模态融合”，载于 IEEE/CVF国际计算机视觉会议研讨会论文集，2019年，页码 0–0。[53] C. Guo、D. Yang、C. Li 和 P. Song，“通过融合预测位置图进行RGBT跟踪的双孪生网络”，视觉计算机，第38卷，第7期，2555–2567页，2022年。[54] Z. Zhang 和 H. Peng，“ 用于实时视觉跟踪的更深、更宽孪生网络 ”，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2019年，4591–4600页。[55] S. Zhai、P. Shao、X. Liang 和 X. Wang，“通过跨模态相关滤波器实现快速 RGB-T 跟踪”，神经计算，第334卷，172–181页，2019年。[56] C. Li、C. Zhu、Y. Huang、J. Tang 和 L. Wang，“利用软一致性和噪声标签进行跨模态排序，实现鲁棒 RGB-T 跟踪”，载于欧洲计算机视觉会议（ECCV）论文集，2018年，808–823页。[57] C. Li、N. Zhao、Y. Lu、C. Zhu 和 J. Tang，“用于 RGB-T 目标跟踪的加权稀疏表示正则化图学习”，载于第25届ACM国际多媒体会议论文集，2017年，1856–1864 页。[58] C. Li、C. Zhu、S. Zheng、B. Luo 和 J. Tang，“用于 RGB-T 跟踪的两阶段模态图正则化流形排序”，信号处理：图像通信，第68卷，207–217页，2018年。[59] L. Shen、X. Wang、L. Liu、B. Hou、Y. Jian、J. Tang 和 B. Luo，“基于协同低秩图模型的RGBT跟踪”，神经计算，第492卷，370–381页，2022年。[60] J. Zhang、S. Ma 和 S. Sclaroff，“MEEM：通过熵最小化实现多专家鲁棒跟踪”，载于欧洲计算机视觉会议。Springer，2014年，188–203页。[61] J. F. Henriques、R. Caseiro、P. Martins 和 J. Batista，“使用核化相关滤波器进行高速跟踪”，IEEE模式分析与机器智能汇刊，第37卷，第3期，583–596页，2014年。[62] L. Lin、H. Fan、Y. Xu 和 H. Ling，“SwinTrack：一种简单而强大的Transformer跟踪基线”，载于神经信息处理系统进展，2022年。[63] K. Yang、Z. He、Z. Zhou 和 N. Fan，“Siamatt：用于视觉跟踪的孪生注意力网络”，知识型系统，第203卷，106079页，2020年。[64] Z. Zhang 和 H. Peng，“用于实时视觉跟踪的更深、更宽孪生网络”，载于 IEEE计算机视觉与模式识别会议

计算机视觉与模式识别（CVPR），2019年6月。[65] M. Danelljan、G. Bhat、F. S. Khan 和 M. Felsberg，“ATOM：通过最大化重叠率实现精准跟踪”，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2019年，4660–4669页。



Xingchen Zhang（M’21）于2012年获得华中科技大学理学学士学位，并于2018年获得伦敦玛丽女王大学博士学位。他目前是伦敦帝国理工学院电气与电子工程系个人机器人实验室的玛丽·斯克沃多夫斯卡-居里个人研究员。此前，他曾在该系担任教学研究员和研究助理。他的主要研究兴趣包括人类意图预测、图像融合和目标跟踪。他是第九届中国信息融合大会最佳论文荣誉提名奖获得者。他是《图像融合》一书的合著者，该书获得了国家科学技术学术著作出版基金（2019年）资助。他担任英国研究与创新署未来领袖奖学金、英国工程与自然科学研究委员会新研究员奖和英国工程与自然科学研究委员会开放奖学金的评审人。



Yiannis Demiris (SM’03) 于1994年和1999年分别获得英国爱丁堡大学人工智能与计算机科学理学学士（荣誉）学位，以及智能机器人学博士学位。他现任英国伦敦帝国理工学院电气与电子工程系教授，同时担任英国皇家工程院新兴技术讲席教授及个人机器人实验室主任。他目前的研究兴趣包括人机交互、机器学习、用户建模和辅助机器人技术。Demiris教授于2012年获得校长卓越教学奖，并于同年获得工程学院工程教育卓越奖。他是工程技术学会（IET）和英国计算机学会（BCS）的会士。