

利用可学习的融合损失实现任务驱动的图像融合

白浩闻¹ 赵子祥^{1,2} 张讲社¹ 吴一尘³ 邓李伦¹ 崔玉昆¹ 姜柏松¹ 徐爽⁴

¹ 西安交通大学 ² ETH Zürich ³ 香港城市大学 ⁴ 西北工业大学

hwbai@stu.xjtu.edu.cn

Abstract

多模态图像融合通过聚合来自多个传感器源的信息，能够比任何单一来源实现更优的视觉质量和感知特性，通常能提升下游任务的表现。然而，目前用于下游任务的融合方法仍然使用预定义的融合目标，这些目标可能与下游任务不匹配，从而限制了适应性指导并降低了模型的灵活性。为了解决这个问题，我们提出了基于任务驱动的图像融合框架——**TDFusion**，该框架结合了由任务损失指导的可学习融合损失。具体而言，我们的融合损失包括由神经网络建模的可学习参数，该网络被称为损失生成模块。该模块通过元学习的方式，由下游任务的损失进行监督。学习目标是通过优化融合损失来最小化融合图像的任务损失。融合模块与损失模块之间的迭代更新确保融合网络朝着最小化任务损失的方向演化，从而引导融合过程朝着任务目标进行。**TDFusion**的训练仅依赖于下游任务的损失，使其能够适应任何特定任务。它可以应用于任何融合与任务网络架构。实验结果展示了**TDFusion**在融合和任务相关应用中的表现，包括四个公开的融合数据集、语义分割和目标检测任务。代码发布在 <https://github.com/HaowenBai/TDFusion>。

1. 引言

多模态图像融合 [20, 24, 37, 45, 56, 63, 66] 结合了来自多个传感器的信息，以产生更全面和详细的表示。红外图像捕捉热辐射，其不受光照条件的影响，而可

见光图像则提供更多细节和纹理。融合图像的优势，如信息丰富度的增加和鲁棒性的提升 [14, 30, 31, 54, 55]，使得图像融合在下游任务中具有重要的应用价值。它通常在语义分割 [9, 26, 42]、目标检测 [4, 25] 等任务中优于单模态输入 [15]。大多数方法将融合任务视为一个简单的图像重建问题，依赖于预定义的无监督损失函数 [60, 67, 71–73] 或感知损失 [16, 18, 61]。这些方法侧重于信息聚合的预定义目标，主要在视觉层面上实现融合。因此，它们通常未能在特征提取过程中捕获关键的语义信息，妨碍了场景理解并降低了任务性能 [10, 19, 38]。最近，一些融合方法进一步探索了融合与下游任务之间的相互增强关系。通过将融合网络与下游任务网络级联 [28, 35]，任务损失约束了融合学习，确保融合图像满足任务要求 [25, 42]。另外，一些方法则嵌入了高级视觉任务特征 [26, 60, 62]，或者学习了最优的初始化 [29]。

然而，尽管下游任务已被集成，现有框架仍然依赖于预定义的融合损失项，这些损失项无法动态适应这些任务。通过特定组合的下游任务影响依然有限。手动定义的损失本质上涉及预定义的指导，往往忽视了任务特定的需求。这种指导对融合过程施加了手动设计的先验约束，限制了下游任务对特定图像对的动态和自适应影响。这些方法，无论是嵌入有用的下游任务特征 [26, 62]，还是直接使用下游损失来指导融合网络 [25, 42]，仍然面临固定融合损失项的限制。专门为下游任务量身定制的网络 [60] 引入了融合与任务之间的依赖，限制了灵活性并限制了其在各种高级视觉任务中的适用性。为了解决这些问题，我们提出了一个下游任务驱动的学习框架，其中包含一个可学习的损失。融合损失包含可学习的参数，由损失生成模块生成，旨在为特定的下游任务保留源图像的强度信息。

¹张讲社为本文的通讯作者。

²本文为 *Task-driven Image Fusion with Learnable Fusion Loss* (CVPR 2025) 的中文翻译版。

更新融合损失的目的是引导融合网络生成融合图像，从而最小化下游任务损失，增强适应性。此外，融合损失的更新仅仅依赖于下游任务损失，使其独立于任何特定的任务或网络架构。

损失生成模块产生融合损失，并用于随后更新融合模块。这一复杂性给标准的端到端训练带来了挑战。幸运的是，元学习技术作为学习如何学习的策略，可以有效地实现损失生成模块的学习目标。这涉及通过优化的融合损失最小化融合图像的任务损失。元学习解决了深度学习中的常见挑战，如数据有限、计算成本高以及需要更好的泛化能力。关键优化领域包括参数调优 [7]、优化策略 [21] 和网络架构设计 [23]。在本文中，我们从模型无关元学习 (MAML) 方法 [7] 中汲取灵感，来训练我们的损失生成模块。具体来说，训练损失生成模块包括两个阶段：内部更新和外部更新。在内部更新阶段，损失生成模块的输出创建融合模块的一步更新替身，而不融合模块改变原始参数。在外部更新阶段，替身融合模块生成的融合图像被输入到任务网络中。由此产生的任务损失通过反向传播更新损失生成模块。这种交替训练确保了损失生成模块持续产生最小化融合图像下游任务损失的融合损失。

本文提出了TDFusion，一个由下游任务驱动的任务导向融合框架。该框架包括一个融合模块、一个任务模块和一个损失生成模块，后者学习优化融合损失。融合损失结合了源图像的强度偏好和梯度保留，并通过下游任务损失来指导强度偏好的优化。该模型遵循先进方法中使用的融合损失的通用形式 [60, 67, 72, 73]，确保了其对各种任务的适应性。损失生成模块的更新使用元学习方法，在每次更新融合模块后，基于融合图像的任务损失优化损失生成模块参数。这个过程确保损失函数持续引导融合模块，优化特征聚合并最小化下游任务损失。损失生成模块通过与融合模块和任务模块的交替更新动态调整，在每个模型状态下生成最优的融合损失。我们的贡献可以总结如下：

- 我们提出了TDFusion，这是一种基于元学习的融合框架，利用下游任务的损失函数进行训练。该方法促进了任务驱动的融合，并缓解了融合图像中缺乏真实值带来的挑战。此外，该框架与特定的下游任务或网络架构无关，从而增强了其适应性和灵活性。
- 我们的框架包括一个动态更新的、可学习的融合损失生成模块。它有选择地提取源图像信息，最小化

下游任务中的损失。这确保了最佳的融合性能，同时最大限度地适应下游任务。

- 我们分析了下游任务的信息偏好，如语义分割和目标检测，提供了对多模态高级任务的更深入见解。
- TDFusion在融合和高级视觉任务中取得了卓越的性能，在四个融合数据集上通过语义分割和目标检测进行了验证。

2. 相关工作

2.1. 基于深度学习的图像融合

基于深度学习的图像融合方法通过利用神经网络强大的特征提取能力，彻底改变了这一领域 [49, 61]。这些方法大致分为判别方法和生成方法。判别方法 [6, 65, 69, 72] 利用神经网络的强重建能力，直接学习源图像和融合图像之间的映射 [16, 18, 64, 66]。另一方面，生成方法使用生成方法对图像生成过程进行建模，从分布的角度整合源图像。这些方法包括基于生成对抗网络 [32–34, 71] 和扩散模型 [57, 68] 的方法。统一的融合方法 [59, 73] 弥合了不同融合子任务之间的差距，结合了持续学习 [51] 和自监督分解技术 [22] 等策略。配准模块的引入有助于缓解源图像中的未对齐问题 [12, 47, 52]。最近的研究进一步探索了融合任务与高级视觉任务之间的协同作用。这些研究包括利用下游任务损失来优化融合网络 [25, 42]，嵌入高级任务特征 [26, 60, 62]，以及采用初始化技术 [29]。

2.2. 视觉中的元学习

元学习开发算法以自动微调特定任务的超参数，展示了其在各个领域的多功能性和有效性。MAML [7] 及其变体 [8, 36, 39] 专注于学习高效的初始化参数，以便使用最少的数据快速适应新任务。Meta-SGD [21] 通过学习最佳更新方向和速率扩展了 MAML，特别是在少样本学习场景中具有优势。其他方法如 MW-Net [41] 和 L2RW [40] 优先选择相关的样本权重，以通过使用紧凑的验证集来处理噪声数据。此外，一些研究专注于通过学习损失函数来提高模型的适应性 [1, 3, 11]。

在图像融合中，[17] 学习滤波器以生成任意分辨率的融合图像。MetaFusion [62] 引入了一种机制，通过将语义与融合特定特征对齐来改善图像融合和目标检测。元学习还支持神经架构搜索 [27, 29]，以识别图像融合的最佳网络架构，并为各种任务定制网络初始化 [29]。

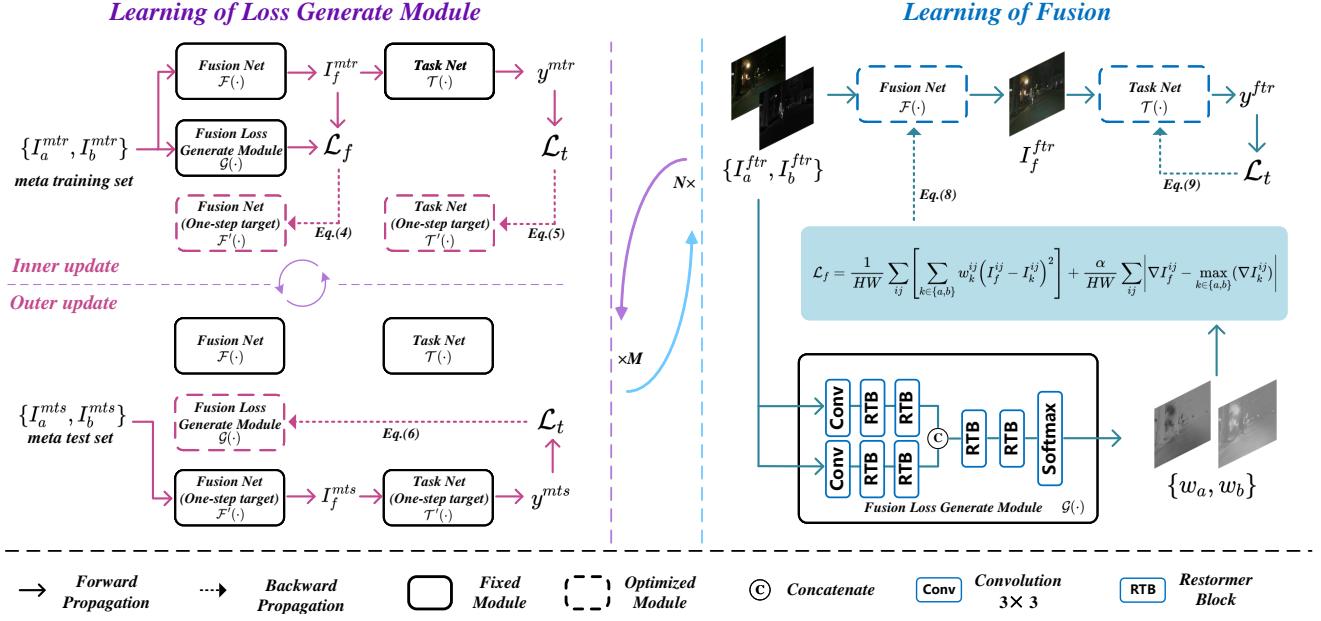


图 1. TD-Fusion的工作流程在训练损失生成模块和融合模块之间交替进行。损失生成模块的训练包括通过元学习进行的内部和外部更新。

此外，持续学习 [50]集成多任务融合也符合元学习原则。

2.3. 与现有方法的区别

我们提出了一种为下游任务量身定制的新颖图像融合方法，利用由任务特定目标驱动的可学习融合损失。我们的方法采用元学习算法，在内部和外部更新之间交替进行，使下游任务损失能够指导可学习融合参数的优化。这导致生成的融合图像能够最小化下游任务损失，从而增强其在各种任务中的适应性。与以往的方法不同，我们的方法开发了任务特定的融合损失，将重点从传统因素如分辨率和网络结构转移，并避免依赖预定义的融合损失项。这使得我们的融合框架更加灵活，适用于各种场景。

3. 方法

3.1. 总览

我们的 TD-Fusion 框架如图 1 所示，由融合网络 $\mathcal{F}(\cdot)$ 、下游任务网络 $\mathcal{T}(\cdot)$ 和用于生成可学习损失函数的参数的融合损失生成模块 $\mathcal{G}(\cdot)$ 组成。这些模块的参数分别表示为 $\theta_{\mathcal{F}}$ 、 $\theta_{\mathcal{T}}$ 和 $\theta_{\mathcal{G}}$ 。 \mathcal{F} 和 \mathcal{T} 的一步更新版本分别表示为 \mathcal{F}' 和 \mathcal{T}' ，其参数为 $\theta_{\mathcal{F}'}$ 和 $\theta_{\mathcal{T}'}$ 。在更新过程中，融合网络和损失生成模块交替学习，如图 1 中

蓝色和紫色部分所示。损失生成模块的更新包括内部更新和外部更新，将在接下来的小节中介绍。 \mathcal{L}_f 和 \mathcal{L}_t 分别表示可学习融合损失和特定任务损失，其具体公式将在下一节给出。

3.2. 损失函数

可学习的融合损失 \mathcal{L}_f 由强度项和梯度项组成。强度项由损失生成模块的输出定义 $\{w_a, w_b\} = \mathcal{G}(I_a, I_b)$ ，其中 w_a 和 w_b 控制融合损失中的强度偏好。损失生成模块中的 $Softmax(\cdot)$ 函数确保每个像素 $w_a^{ij} + w_b^{ij} = 1$ ，从而有选择地保留源图像中的强度信息。梯度项强调来自输入图像的较高梯度值 [60, 67, 73]，旨在最大化保留来自源图像的信息。可学习融合损失的详细公式如下：

$$\mathcal{L}_f = \mathcal{L}_f^{int} + \alpha \mathcal{L}_f^{grad}, \quad (1)$$

$$\mathcal{L}_f^{int} = \frac{1}{HW} \sum_{ij} \left[\sum_{k \in \{a,b\}} w_k^{ij} (I_f^{ij} - I_k^{ij})^2 \right], \quad (2)$$

$$\mathcal{L}_f^{grad} = \frac{1}{HW} \sum_{ij} \left| \nabla I_f^{ij} - \max_{k \in \{a,b\}} (\nabla I_k^{ij}) \right|, \quad (3)$$

其中 ∇ 表示 Sobel 算子，通常用于图像融合中的梯度提取 [25, 43, 67]。参数 α 用作缩放因子，而 \mathcal{L}_f^{int} 和 \mathcal{L}_f^{grad} 分别表示强度损失和梯度损失。权重 $\{w_a, w_b\}$ 控制损

失函数对每个源图像的强度信息的强调。这些参数使得融合过程能够有选择地聚合和整合源图像中的相关信息。 θ_G 的变化导致 w_a 和 w_b 的不同配置，从而影响融合损失的特性。在步骤更新过程中， θ_G 在高级视觉任务损失的驱动下进行更新，详细信息见 3.4 节。

损失函数 \mathcal{L}_t 取决于具体任务。在本研究中，我们采用 SegFormer [5] 和 YOLOv8 [46] 进行下游任务，分别使用交叉熵损失和 YOLO 损失 [46]。

3.3. 数据集划分

为了增强损失生成模块 \mathcal{G} 在指导融合任务中的有效性，我们在每个训练周期创建大小为 M 的不重叠子集。元训练集 $\{I_a^{mtr}, I_b^{mtr}\}$ 和元测试集 $\{I_a^{mts}, I_b^{mts}\}$ 是从融合训练集 $\{I_a^{ftr}, I_b^{ftr}\}$ 中随机抽取的。在损失生成模块的训练过程中，这些子集会按顺序输入到模型中，覆盖了内部更新和外部更新的过程。整个融合训练集 $\{I_a^{ftr}, I_b^{ftr}\}$ 在融合网络的训练过程中被充分利用。

3.4. 损失生成模块的学习

图 1 展示了损失生成模块和融合模块交替训练，确保在不同阶段和不同状态下优化融合损失。优化融合损失的训练过程包含两个关键步骤：内部更新和外部更新。在内部更新中，分别生成融合网络和任务网络的克隆，每个网络使用融合损失和任务损失进行一次训练迭代。这个过程旨在通过融合损失获取网络的状态。在外部更新过程中，计算由更新后的克隆生成的融合图像的任务损失。然后，这个损失将反向传播到损失生成网络。该步骤的目标是指导融合网络生成能够降低下游任务损失的融合图像，一旦得到融合损失的指导。内部更新和外部更新的交替步骤构成了损失生成模块的学习过程。

3.4.1. 内部更新

在内部更新阶段，融合网络 \mathcal{F} 在融合损失的指导下进行一次更新，该损失依赖于网络 \mathcal{G} 的当前状态。此更新主要旨在计算中间参数 $\theta_{\mathcal{F}'}$ 和 $\theta_{\mathcal{T}'}$ ，这些参数对于在后续阶段更新 θ_G 至关重要。紫色区域的上半部分在图 1 中说明了这一过程。在此阶段，来自元训练集 $\{I_a^{mtr}, I_b^{mtr}\}$ 的图像被输入到模型中，其中 \mathcal{F} 进行一次梯度下降更新：

$$\theta_{\mathcal{F}'} = \theta_{\mathcal{F}} - \eta_{\mathcal{F}'} \frac{\partial \mathcal{L}_f(I_a^{mtr}, I_b^{mtr}, I_f^{mtr}; \theta_G)}{\partial \theta_{\mathcal{F}}}, \quad (4)$$

θ_G 表示损失生成模块 \mathcal{G} 的参数，这些参数决定了可学习融合损失的参数。在更新方程中，以步长 $\eta_{\mathcal{F}'}$ 更新融合模块。模块 \mathcal{F}' 是 \mathcal{F} 的临时替身，其参数经历了一次更新步骤的调整。同时， \mathcal{F} 的参数，记为 $\theta_{\mathcal{F}}$ ，保持不变。类似地， \mathcal{T}' 使用当前任务网络 \mathcal{T} 的参数 $\theta_{\mathcal{T}}$ 进行进一步更新：

$$\theta_{\mathcal{T}'} = \theta_{\mathcal{T}} - \eta_{\mathcal{T}'} \frac{\partial \mathcal{L}_t(I_f^{mtr})}{\partial \theta_{\mathcal{T}}}. \quad (5)$$

在内部更新过程中更新的参数 $\theta_{\mathcal{F}'}$ 和 $\theta_{\mathcal{T}'}$ 也确保了 $\theta_{\mathcal{F}'}$ 相对于 θ_G 的计算图被保留。这个保留的图对于在外部更新中优化 θ_G 至关重要。

3.4.2. 外部更新

外部更新的主要目标是评估和改进 \mathcal{G} 的融合指导能力，特别是通过加强损失函数 \mathcal{L}_f 在引导融合模块 \mathcal{F} 中的影响。在框架图中，这一阶段在紫色区域的下半部分表示。从内部更新中派生的模块 \mathcal{F}' 和 \mathcal{T}' 代表了 \mathcal{G} 的当前指导能力。在理想情况下，最佳融合损失应提高下游任务在融合图像上的性能。在此阶段，元测试集 $\{I_a^{mts}, I_b^{mts}\}$ 被使用。随后使用任务损失 \mathcal{L}_t 更新参数 θ_G ，该损失由 \mathcal{F}' 和 \mathcal{T}' 计算：

$$\theta_G = \theta_G - \eta_G \frac{\partial \mathcal{L}_t(I_f^{mts})}{\partial \theta_G}, \quad (6)$$

其中 $I_f^{mts} = \mathcal{F}'(I_a^{mts}, I_b^{mts})$ ，梯度 $\partial \mathcal{L}_t / \partial \theta_G$ 可以计算为：

$$\frac{\partial \mathcal{L}_t}{\partial \theta_G} = \frac{\partial \mathcal{L}_t}{\partial \theta_{\mathcal{F}'}} * \left(-\eta_{\mathcal{F}'} \frac{\partial^2 \mathcal{L}_f(I_a^{mtr}, I_b^{mtr}, I_f^{mtr}; \theta_G)}{\partial \theta_{\mathcal{F}} \partial \theta_G} \right). \quad (7)$$

公式 6 成立是因为任务损失 \mathcal{L}_t 由 I_f^{mts} 决定，而 I_f^{mts} 又依赖于 $\theta_{\mathcal{F}'}$ 。通过在内更新过程中保留 $\theta_{\mathcal{F}'}$ 和 θ_G 之间的计算关系，实现了通过 \mathcal{L}_t 优化 θ_G 。更新后的 \mathcal{G} 模块获得了生成增强融合损失函数的能力，使融合模块能够更有效地将源图像中的相关信息整合到用于下游任务的融合输出中。

3.5. 融合网络的学习

交替的内外更新迭代形成了一种灵活且有效的机制，以响应 \mathcal{F} 的不断变化状态来优化 \mathcal{G} 。在优化 \mathcal{G} 之后，它被用于进一步改进 \mathcal{F} 的训练。在图中以蓝色表示的这个阶段涉及处理融合训练集图像 $\{I_a^{ftr}, I_b^{ftr}\}$ 。通过应用融合损失 \mathcal{L}_f 和任务损失 \mathcal{L}_t ，同时更新 \mathcal{F} 和 \mathcal{T} ：

$$\theta_{\mathcal{F}} = \theta_{\mathcal{F}} - \eta_{\mathcal{F}} \frac{\partial \mathcal{L}_f(I_a^{ftr}, I_b^{ftr}, I_f^{ftr}; \theta_G)}{\partial \theta_{\mathcal{F}}}, \quad (8)$$

Algorithm 1 TDFusion 训练算法

Require: 融合训练集 $\{I_a^{ftr}, I_b^{ftr}\}$, 大小为 N 。
Ensure: 经过充分训练的 $\theta_{\mathcal{F}}$, $\theta_{\mathcal{T}}$, $\theta_{\mathcal{G}}$ 。

- 1: 初始化 $\theta_{\mathcal{F}}$, $\theta_{\mathcal{T}}$, $\theta_{\mathcal{G}}$ 。
- 2: **for** $epoch = 1$ **to** L **do**
- 3: 采样 $\{I_a^{mtr}, I_b^{mtr}\}$ 和 $\{I_a^{mts}, I_b^{mts}\}$ 。
- 4: **for** $step = 1$ **to** M **do**
- 5: % 内部更新: 应用 \mathcal{G} 。
 采样 (I_a^{mtr}, I_b^{mtr}) 并获取 (I_f^{mtr}, y^{mtr}) 。
 通过公式 4 和公式 5 计算 $\theta_{\mathcal{F}'}$ 和 $\theta_{\mathcal{T}'}$ 。
 % 外部更新: 优化 \mathcal{G} 。
 采样 (I_a^{mts}, I_b^{mts}) 并获取 $((I_f^{mts}, y^{mts})$ 。
 通过公式 6 更新 $\theta_{\mathcal{G}}$ 。
- 9: **end for**
- 10: **for** $step = 1$ **to** N **do**
- 11: % 融合更新: 优化 \mathcal{F} 和 \mathcal{T} 。
 采样 (I_a^{ftr}, I_b^{ftr}) 并获取 (I_f^{ftr}, y^{ftr}) 。
 通过公式 8 和公式 9 更新 $\theta_{\mathcal{F}}$ 和 $\theta_{\mathcal{T}}$ 。
- 13: **end for**
- 14: **end for**

$$\theta_{\mathcal{T}} = \theta_{\mathcal{T}} - \eta_{\mathcal{T}} \frac{\partial \mathcal{L}_t(I_f^{ftr})}{\partial \theta_{\mathcal{T}}} \quad (9)$$

在完成对融合网络的几次训练后, 重点转回到融合生成模块的学习阶段。融合框架通过一系列交替阶段演变, 每个阶段根据融合网络的当前状态微调融合损失。这种交替阶段确保融合网络在其进程中始终应用最合适的融合损失。最终, 这导致了一个高度有效的融合网络的发展, 优化以达到最佳性能。完整的训练过程详见算法 1。

3.6. 网络架构

TDFusion 由三个模块组成: 融合网络、下游任务网络和损失生成网络。融合网络与 [2] 中的架构相同, 是一个基于 Restormer Block (RTB) [58] 构建的轻量级模型。该网络中集成了自适应融合模块以促进特征整合。表 1 展示了损失生成模块的结构。它同样采用 Restormer Block (RTB) [58] 作为核心单元, 接收输入 $\{I_a, I_b\}$ 。在应用 $Softmax(\cdot)$ 后, 最终输出保证每个像素满足 $w_a^{ij} + w_b^{ij} = 1$ 。这种设计确保了融合图像满足相似性约束, 并消除了对损失生成模块中初

始化的依赖。下游任务网络 $\mathcal{T}(\cdot)$ 的架构取决于具体任务。对于损失生成模块的学习, 我们选择了 SegFormer [5] 和 YOLOv8 [46] 的最轻量级模型, 分别用于语义分割和目标检测。

3.7. 理论分析

为了更好地理解损失生成模块 \mathcal{G} 的加权机制, 我们研究了生成权重 $\{w_a, w_b\}$ 的优化过程, 记为 $\theta_{\mathcal{G}}$ 。为清晰起见, 我们将公式 2 重写如下:

$$\begin{aligned} \mathcal{L}_f^{int} &= [w_a \odot (I_a - I_f) \odot (I_a - I_f) \\ &\quad + w_b \odot (I_b - I_f) \odot (I_b - I_f)] \times \frac{1}{HW} \\ &= [\mathcal{G}(I_a, I_b; \theta_{\mathcal{G}}) \odot (I_a - \mathcal{F}_{\theta_{\mathcal{F}}}(I_a, I_b)) \\ &\quad \odot (I_a - \mathcal{F}_{\theta_{\mathcal{F}}}(I_a, I_b)) \\ &\quad + (1 - \mathcal{G}(I_a, I_b; \theta_{\mathcal{G}})) \odot (I_b - \mathcal{F}_{\theta_{\mathcal{F}}}(I_a, I_b)) \\ &\quad \odot (I_b - \mathcal{F}_{\theta_{\mathcal{F}}}(I_a, I_b))] \times \frac{1}{HW}. \end{aligned} \quad (10)$$

这里, $w_a, w_b \in \mathbb{R}^{H \times W}$, $I_a, I_b \in \mathbb{R}^{H \times W}$, \odot 表示元素级乘法操作。令 Ω' 为集合 $\{\theta_{\mathcal{F}'}, \theta_{\mathcal{T}'}\}$, 可以得到:

$$\begin{aligned} \theta_{\mathcal{G}} &= \theta_{\mathcal{G}} - \eta_{\mathcal{G}} \frac{\partial \mathcal{L}_t^{mts}(\Omega'(\theta_{\mathcal{G}}))}{\partial \theta_{\mathcal{G}}} \\ &= \theta_{\mathcal{G}} - \eta_{\mathcal{G}} \frac{\partial \mathcal{L}_t^{mts}(\Omega'(\theta_{\mathcal{G}}))}{\partial \Omega'} \frac{\partial \Omega'(\theta_{\mathcal{G}})}{\partial \theta_{\mathcal{G}}} \\ &= \theta_{\mathcal{G}} - \eta_{\mathcal{G}} \eta_{\mathcal{F}'} \underbrace{\frac{\partial \mathcal{L}_t^{mts}(\Omega'(\theta_{\mathcal{G}}))}{\partial \Omega'}}_{(a)} \times \frac{\partial \mathcal{G}(I_a, I_b; \theta_{\mathcal{G}})}{\theta_{\mathcal{G}}} \\ &\quad \times [(I_a - \underbrace{\frac{\partial \mathcal{F}_{\theta_{\mathcal{F}}}}{\partial \theta_{\mathcal{F}}}}_{(b)}) \odot (I_a - \frac{\partial \mathcal{F}_{\theta_{\mathcal{F}}}}{\partial \theta_{\mathcal{F}}}) \\ &\quad - (I_b - \underbrace{\frac{\partial \mathcal{F}_{\theta_{\mathcal{F}}}}{\partial \theta_{\mathcal{F}}}}_{(b)}) \odot (I_b - \frac{\partial \mathcal{F}_{\theta_{\mathcal{F}}}}{\partial \theta_{\mathcal{F}}})] \\ &= \theta_{\mathcal{G}} - \eta_{\mathcal{G}} \eta_{\mathcal{F}'} \mathbf{G} \times \frac{\partial \mathcal{G}(I_a, I_b; \theta_{\mathcal{G}})}{\theta_{\mathcal{G}}}. \end{aligned} \quad (11)$$

这里, \mathbf{G} 表示两个梯度之间的内积: 第一个是从任务损失中导出的, 使用元测试集, 第二个是从融合损失中计算的, 基于元训练集。因此, 模块 $\theta_{\mathcal{G}}$ 的优化由任务损失驱动, 目标是在融合过程中保留任务特定的信息。

4. 实验

4.1. 设置

实验设置. 在我们的实验中, epoch 数 L 和损失生成模块的训练迭代次数 M 被设置为 50 和 200。融合网



图 2. 融合结果的视觉比较。案例包括 MSRS 数据集中的“01258N”、FMB 数据集中的“00122”、M3FD 数据集中的“00449”和 LLVIP 数据集中的“200304”。

络的学习迭代次数 N 取决于数据集的大小。我们使用 Adam 优化器，学习率为 $1e-4$ ，批量大小为 2，超参数 α 设置为 1。所有实验均在配备单个 NVIDIA RTX 3090 GPU 的 PC 上进行。

评估指标和比较方法. 我们研究中比较的先进融合方法包括 TarDAL [25]、SegMIF [26]、MURF [53]、EMMA [70]、DCINN [48]、MRFS [60] 和 TIMFusion [29]。融合性能的评估指标包括熵 (EN)、空间频率 (SF)、相关差异 (SCD)、视觉信息保真度 (VIF)、 $Q^{AB/F}$ 和结构相似性指数 (SSIM)。

数据集划分. 我们使用四个包含下游任务标签的数据集，包括 MSRS [44]、FMB [26]、M3FD [25] 和 LLVIP [13]。MSRS 包含 1083/361 对图像用于训练/测试，FMB 包含 1220/280 对图像用于训练/测试。我们遵循原始论文的划分方法。M3FD 数据集包含 4200 对用于检测的图像，其中 300 对指定用于融合评估。4200 张检测图像被分为 3150 张用于训练和 1050 张用于测试，确保 300 对用于融合评估的图像包含在检测测试集中。这 300 张融合图像随后用于评估融合性能。原始 LLVIP 数据集包含 12025/3463 对图像作为训练/测试集。由于其较大规模，我们每隔 10 张图像选择一张来

形成我们的训练和测试集，最终得到 1203/347 对图像用于训练和测试。我们对 M3FD 和 LLVIP 的划分将公开可用。

4.2. 融合实验

图 2 展示了不同方法的视觉比较。TDFusion 生成的融合图像在保留关键细节、实现亮度平衡和保持清晰的物体轮廓方面表现出色。它有效地保留了红外图像中的目标特征和可见光图像中的背景细节，生成的融合图像在不同环境中更加自然且清晰。这些结果突显了 TDFusion 在细节保留和视觉表现方面的优势。更多结果在补充材料中提供。

表 1 展示了四个数据集上的融合定量比较。TDFusion 在大多数指标上优于其他方法。这表明 TDFusion 不仅能够增强图像细节，还能在各种场景中提供一致的融合结果。与其他方法相比，TDFusion 显示出了更强的适应性和鲁棒性，特别是在处理不同的图像特征和具有挑战性的融合任务时。

4.3. 下游任务应用

在这一部分，为了公平比较，我们使用 SegFormer [5] 和 YOLOv8 [46] 作为主干网络，并重新训练每种融

表 1. 红外-可见光图像融合的定量比较。红色和蓝色标记分别代表最佳和次佳值。

MSRS [44] 数据集上的红外-可见光图像融合						FMB [26] 数据集上的红外-可见光图像融合							
	EN ↑	SF ↑	SCD ↑	VIF ↑	Q_{abf} ↑	SSIM ↑		EN ↑	SF ↑	SCD ↑	VIF ↑	Q_{abf} ↑	SSIM ↑
TarDAL [25]	5.28	5.98	0.71	0.21	0.18	0.47	TarDAL [25]	6.63	6.94	1.03	0.28	0.29	0.74
SegMIF [26]	5.95	11.10	1.57	0.44	0.63	0.55	SegMIF [26]	6.83	13.69	1.72	0.39	0.65	0.60
MURF [53]	5.04	10.49	1.02	0.22	0.37	0.60	MURF [53]	6.37	13.88	1.34	0.22	0.37	0.68
EMMA [70]	6.73	11.56	1.62	0.49	0.64	0.70	EMMA [70]	6.77	15.00	1.50	0.42	0.65	0.72
DCINN [48]	6.00	10.51	1.49	0.41	0.57	0.52	DCINN [48]	6.47	11.47	1.39	0.38	0.59	0.74
MRFS [60]	7.00	8.86	1.42	0.37	0.49	0.55	MRFS [60]	6.78	12.42	1.24	0.38	0.62	0.73
TIMFusion [29]	6.27	9.67	1.34	0.32	0.48	0.68	TIMFusion [29]	6.51	12.23	1.24	0.35	0.59	0.73
TDFusion (Ours)	6.74	11.30	1.86	0.50	0.67	0.70	TDFusion (Ours)	6.86	14.16	1.76	0.43	0.68	0.75
M3FD [25] 数据集上的红外-可见光图像融合						LLVIP [13] 数据集上的红外-可见光图像融合							
	EN ↑	SF ↑	SCD ↑	VIF ↑	Q_{abf} ↑	SSIM ↑		EN ↑	SF ↑	SCD ↑	VIF ↑	Q_{abf} ↑	SSIM ↑
TarDAL [25]	6.87	7.63	1.29	0.27	0.30	0.71	TarDAL [25]	6.32	7.42	1.04	0.27	0.22	0.58
SegMIF [26]	6.85	14.14	1.72	0.37	0.60	0.59	SegMIF [26]	6.68	15.46	1.38	0.40	0.66	0.57
MURF [53]	6.50	12.55	1.46	0.21	0.32	0.64	MURF [53]	6.13	15.08	0.96	0.21	0.31	0.57
EMMA [70]	6.92	15.23	1.49	0.38	0.59	0.69	EMMA [70]	7.35	15.37	1.57	0.41	0.64	0.66
DCINN [48]	6.59	11.21	1.46	0.34	0.51	0.72	DCINN [48]	6.98	13.34	1.43	0.38	0.52	0.64
MRFS [60]	6.94	12.07	1.26	0.34	0.55	0.70	MRFS [60]	6.83	11.04	1.23	0.31	0.42	0.64
TIMFusion [29]	6.75	12.31	1.37	0.35	0.53	0.70	TIMFusion [29]	6.58	13.52	1.14	0.33	0.46	0.64
TDFusion (Ours)	6.99	14.49	1.83	0.41	0.65	0.72	TDFusion (Ours)	7.36	16.38	1.75	0.46	0.70	0.67

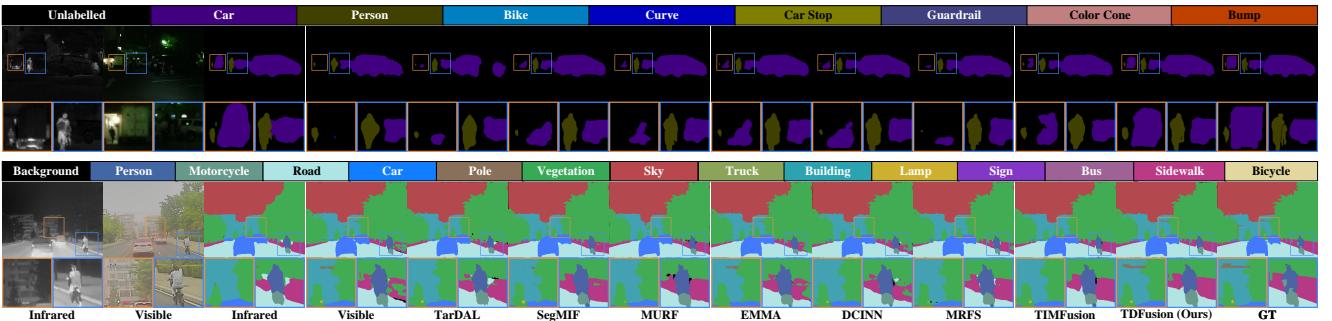


图 3. 语义分割的视觉比较。案例是 MSRS 数据集中的“00726N”和 FMB 数据集中的“01438”。

合方法的任务网络，训练300个epoch以评估它们对下游任务的适应性。图 3 和图 4 分别展示了语义分割和目标检测的视觉比较。TDFusion 在图像细节保留、边缘清晰度和目标识别方面表现优异，能够有效地识别和分割物体。在语义分割中，生成的图像清晰地区分了不同的类别区域，并且与真实标签高度匹配。在目标检测中，融合图像展现了更精确的关键物体边界定位。这表明 TDFusion 在细节和整体背景之间保持了更好的平衡。更多结果请参考补充材料。

表 2 展示了不同方法在语义分割和目标检测中的性能比较。TDFusion 在大多数指标上优于其他方法，特别是在 mIoU 和 mAP 等关键指标上。这表明 TDFu-

表 2. 下游任务应用性能比较。红色和蓝色标记分别代表最佳和次佳。

方法	语义分割				目标检测			
	MSRS		FMB		M3FD		LLVIP	
	mAcc	mIoU	mAcc	mIoU	mAP50	mAP75	AP50	AP75
Infrared	83.23	69.49	58.85	51.98	79.12	53.05	96.03	72.07
Visible	83.44	73.76	65.12	57.96	82.21	54.82	91.78	48.66
TarDAL	81.93	71.35	62.86	55.33	83.16	56.39	93.79	62.71
SegMIF	85.73	74.25	65.97	58.41	83.61	58.23	93.95	66.45
MURF	85.03	74.08	64.10	56.96	80.58	54.22	94.24	68.04
EMMA	85.99	74.48	62.45	56.28	83.71	56.91	94.00	66.21
DCINN	84.11	74.35	61.09	54.81	82.69	57.37	94.92	68.34
MRFS	84.76	74.50	61.93	55.71	83.28	57.74	93.03	67.21
TIMFusion	83.67	73.58	63.70	57.24	83.22	56.08	93.76	61.33
TDFusion (Ours)	86.04	75.09	67.17	60.50	86.27	59.71	95.00	69.18

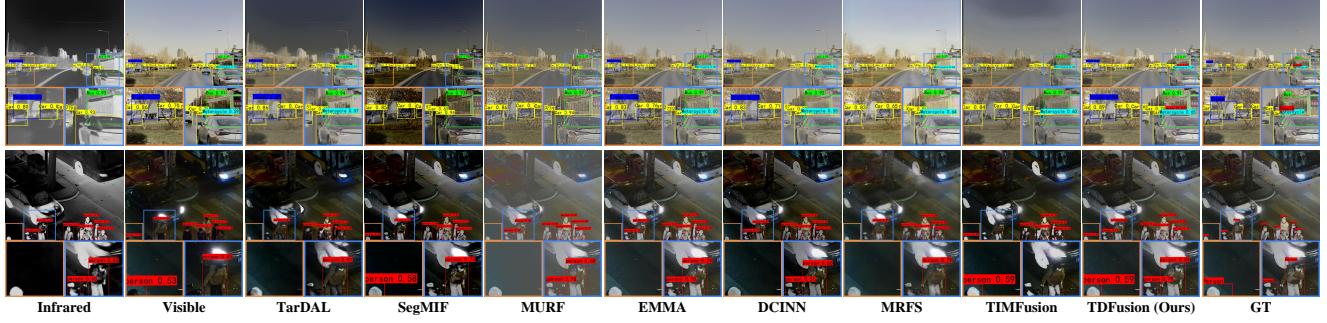


图 4. 物体检测的视觉比较。案例为 M3FD 数据集中的“02236”和 LLVIP 数据集中的“210145”。

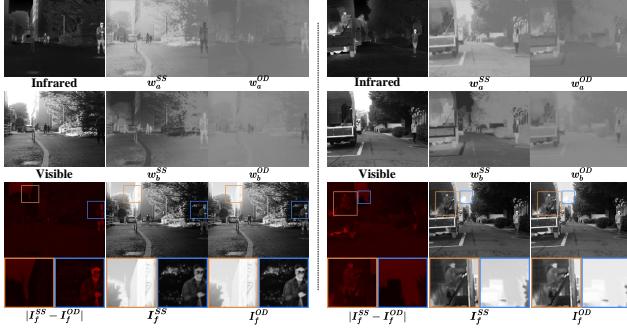


图 5. 不同下游任务的可学习损失的视觉比较。

sion 有效地增强了融合图像的细节。它还提高了下游任务的性能，特别是在复杂场景中的准确性和鲁棒性方面。类别结果在补充材料中提供。

4.4. 任务驱动的可学习损失

我们的框架包含一个可学习的融合损失，它对下游任务从源图像中获取信息的偏好进行建模。在 FMB 数据集和 LLVIP 数据集上训练的模型分别标记为 *SS* 和 *OD*，并在 MSRS 数据集上进行评估，以模拟在未知场景中的性能，如图 5 所示。结果表明，融合模型自适应地从红外和可见光图像中选择信息以满足任务需求。在语义分割中，模型结合了场景结构和纹理，优先考虑边界。这在不同光照条件下提高了分割效果。融合权重 w_a^{SS} 和 w_b^{SS} 表明了对可见光细节和低光条件下红外优势的偏好。而在目标检测中，模型侧重于边缘和对比度信息，特别是行人和车辆等目标。更高的融合权重分配给红外图像中的亮区，从而增强了低光条件下的目标检测。融合权重 w_a^{OD} 和 w_b^{OD} 反映了这种偏好。不同任务的融合损失比较揭示了显著差异，特别是在高亮区域，确认了模型通过选择多模态图像中最相关的信息来适应任务特定的要求。更多结果在补充材料中提供。

表 3. 红外-可见光融合的消融实验。红色表示最佳。

配置	FMB数据集上的融合消融研究					
	EN	SF	SCD	VIF	Q_{abf}	SSIM
I 固定 w_a 和 w_b 为 1/2	6.60	13.73	1.58	0.39	0.60	0.72
II 取消 \mathcal{L}_f^{grad}	6.77	11.65	1.63	0.37	0.64	0.73
III 令 θ_f 受 \mathcal{L}_t 影响	6.80	13.85	1.70	0.41	0.66	0.73
IV 取消融合学习阶段	6.82	14.07	1.72	0.41	0.67	0.72
V $I_f = w_a * I_a + w_b * I_b$	6.75	11.49	1.65	0.38	0.62	0.73
Ours	6.86	14.16	1.76	0.43	0.68	0.75

4.5. 消融实验

为了全面评估我们提出的算法的性能，我们在 FMB 数据集上进行了系列消融实验，详细结果如表 3 所示。在实验 I 中，我们通过将 w_a 和 w_b 固定为 1/2 来排除可学习的融合损失。在实验 II 中，我们从损失函数中省略了梯度损失。在实验 III 中，我们还允许融合模块参数由任务损失 \mathcal{L}_t 和融合损失 \mathcal{L}_f 共同优化。在实验 IV 中，我们排除了融合模块的专用学习阶段，并在损失模块的外部更新期间对其进行更新，以测试融合模块训练计划的影响。在实验 V 中，我们用判别方法替换了我们的融合方法。在不同配置下观察到的性能下降证实了我们提出的方法的合理性和有效性。消融实验的视觉对比展示在补充材料中。

5. 结论

为了克服预定义融合损失的局限性，这些损失通常无法有效地指导下游任务的融合过程，我们提出了一种基于元学习的任务引导融合框架。该框架包括一个损失生成模块，该模块输出可学习融合损失的参数。该模块使用元学习方法进行更新，通过交替进行内循环和外循环步骤来增强其指导融合网络的能力。在不同的融合条件下，该模块为下游任务生成最优的融合损失。这使得融合网络能够生成最小化任务特定损失

的融合图像。理论分析解释了下游任务损失如何在我们的框架中指导融合损失。在四个公开可用的融合数据集和下游任务（包括语义分割和目标检测）上的实验，证明了我们方法的有效性。

参考文献

- [1] Antreas Antoniou and Amos J. Storkey. Learning to learn by self-critique. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 9936–9946, 2019. [2](#)
- [2] Haowen Bai, Zixiang Zhao, Jiangshe Zhang, Yichen Wu, Lilun Deng, Yukun Cui, Shuang Xu, and Baisong Jiang. Refusion: Learning image fusion from reconstruction with learnable loss via meta-learning, 2024. [5](#)
- [3] Sungyong Baik, Janghoon Choi, Heewon Kim, Dohee Cho, Jaesik Min, and Kyoung Mu Lee. Meta-learning with task-adaptive loss function for few-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9445–9454. IEEE, 2021. [2](#)
- [4] Yanpeng Cao, Dayan Guan, Weilin Huang, Jiangxin Yang, Yanlong Cao, and Yu Qiao. Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. *Information Fusion*, 46:206–217, 2019. [1](#)
- [5] Bo Cheng, Xiang Li, Yujie Wei, Cheng Huang, Xiaoyong Zhang, Yandong Jiang, Tianyu Zhang, Na Xu, Shuai Yu, Xinxin Zhan, et al. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12022–12032, 2021. [4, 5, 6](#)
- [6] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3333–3348, 2021. [2](#)
- [7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International conference on machine learning (ICML)*, pages 1126–1135, 2017. [2](#)
- [8] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *Proceedings of the International conference on machine learning (ICML)*, pages 1920–1930, 2019. [2](#)
- [9] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IROS*, pages 5108–5115. IEEE, 2017. [1](#)
- [10] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. In *ICONIP (5)*, pages 387–395. Springer, 2021. [1](#)
- [11] Rein Houthooft, Yuhua Chen, Phillip Isola, Bradly C. Stadie, Filip Wolski, Jonathan Ho, and Pieter Abbeel. Evolved policy gradients. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 5405–5414, 2018. [2](#)
- [12] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 539–555. Springer, 2022. [2](#)
- [13] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021. [6, 7](#)
- [14] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, pages 4369–4376, 2020. [1](#)
- [15] Chenglong Li, Chengli Zhu, Yan Huang, Jin Tang, and Liang Wang. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In *ECCV (13)*, pages 831–847. Springer, 2018. [1](#)
- [16] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. [1, 2](#)
- [17] Huafeng Li, Yueliang Cen, Yu Liu, Xun Chen, and Zhengtao Yu. Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion. *IEEE Trans. Image Process.*, 30:4070–4083, 2021. [2](#)
- [18] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021. [1, 2](#)
- [19] Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K. Tokuda, Roberto Hirata Junior, Roberto Marcondes Cesar Junior, Jiawan Zhang, Xiaojie Guo, and Xiachun Cao. Single image deraining: A comprehensive

- benchmark analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3838–3847. Computer Vision Foundation / IEEE, 2019. 1
- [20] Xiaoling Li, Yanfeng Li, Houjin Chen, Yahui Peng, and Pan Pan. Ccafusion: cross-modal coordinate attention network for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1
- [21] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Metasgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017. 2
- [22] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 719–735. Springer, 2022. 2
- [23] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018. 2
- [24] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):105–119, 2021. 1
- [25] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5792–5801, 2022. 1, 2, 3, 6, 7
- [26] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8115–8124, 2023. 1, 2, 6, 7
- [27] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 1600–1608. ACM, 2021. 2
- [28] Risheng Liu, Long Ma, Tengyu Ma, Xin Fan, and Zhongxuan Luo. Learning with nested scene modeling and cooperative architecture search for low-light vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5953–5969, 2022. 1
- [29] Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and metainitialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2, 6, 7
- [30] ShuMin Liu, Jiajia Chen, and Susanto Rahardja. A new multi-focus image fusion algorithm and its efficient implementation. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(5):1374–1384, 2020. 1
- [31] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019. 1
- [32] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019. 2
- [33] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020.
- [34] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiaoping (Steven) Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020. 2
- [35] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5637–5646, 2022. 1
- [36] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999, 2018. 2
- [37] Seonghyun Park, An Gia Vien, and Chul Lee. Cross-modal transformers for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2):770–785, 2023. 1
- [38] Yanting Pei, Yaping Huang, Qi Zou, Yuhang Lu, and Song Wang. Does haze removal help cnn-based image classification? In *ECCV (10)*, pages 697–712. Springer, 2018. 1
- [39] Xinran Qin, Yuhui Quan, Tongyao Pang, and Hui Ji. Ground-truth free meta-learning for deep compressive sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vi-*

- sion and Pattern Recognition (CVPR)*, pages 9947–9956, 2023. [2](#)
- [40] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the International conference on machine learning (ICML)*, pages 4334–4343, 2018. [2](#)
- [41] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. [2](#)
- [42] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022. [1, 2](#)
- [43] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022. [3](#)
- [44] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Infromation Fusion*, 83-84:79–92, 2022. [6, 7](#)
- [45] Wei Tang, Fazhi He, Yu Liu, Yansong Duan, and Tongzhen Si. Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3159–3172, 2023. [1](#)
- [46] Ultralytics. Yolov8: A family of models for fast, accurate object detection, 2023. [4, 5, 6](#)
- [47] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3508–3515. ijcai.org, 2022. [2](#)
- [48] Wu Wang, Liang-Jian Deng, Ran Ran, and Gemine Vivone. A general paradigm with detail-preserving conditional invertible network for image fusion. *International Journal of Computer Vision*, 132(4):1029–1054, 2024. [6, 7](#)
- [49] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, pages 12484–12491. Proceedings of the AAAI conference on artificial intelligence (AAAI) Press, 2020. [2](#)
- [50] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. [3](#)
- [51] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. [2](#)
- [52] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19647–19656. Computer Vision Foundation / IEEE, 2022. [2](#)
- [53] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. [6, 7](#)
- [54] Shuang Xu, Ouafa Amira, Junmin Liu, Chun-Xia Zhang, Jiangshe Zhang, and Guanghai Li. HAM-MFN: hyperspectral and multispectral image multiscale fusion network with RAP loss. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7):4618–4628, 2020. [1](#)
- [55] Shuang Xu, Lizhen Ji, Zhe Wang, Pengfei Li, Kai Sun, Chunxia Zhang, and Jiangshe Zhang. Towards reducing severe defocus spread effects for multi-focus image fusion via an optimization based strategy. *IEEE Transactions Computational Imaging*, 6:1561–1570, 2020. [1](#)
- [56] Yong Yang, Jiaxiang Liu, Shuying Huang, Weiguo Wan, Wenying Wen, and Juwei Guan. Infrared and visible image fusion via texture conditional generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(12):4771–4783, 2021. [1](#)
- [57] Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024. [2](#)
- [58] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5728–5739, 2022. [5](#)
- [59] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *Inter-*

- national Journal of Computer Vision*, 129(10):2761–2785, 2021. [2](#)
- [60] Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26974–26983, 2024. [1](#), [2](#), [3](#), [6](#), [7](#)
- [61] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54: 99–118, 2020. [1](#), [2](#)
- [62] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13955–13965, 2023. [1](#), [2](#)
- [63] Yangyang Zhao, Qingchun Zheng, Peihao Zhu, Xu Zhang, and Wenpeng Ma. Tufusion: A transformer-based universal fusion algorithm for multimodal images. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. [1](#)
- [64] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. Didfuse: Deep image decomposition for infrared and visible image fusion. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 970–976. ijcai.org, 2020. [2](#)
- [65] Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1186–1196, 2022. [2](#)
- [66] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5906–5916. Computer Vision Foundation / IEEE, 2023. [1](#), [2](#)
- [67] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023. [1](#), [2](#), [3](#)
- [68] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. DDFM: denoising diffusion model for multi-modality image fusion. *CoRR*, abs/2303.06840, 2023. [2](#)
- [69] Zixiang Zhao, Jiang-She Zhang, Haowen Bai, Yicheng Wang, Yukun Cui, Lilun Deng, Kai Sun, Chunxia Zhang, Junmin Liu, and Shuang Xu. Deep convolutional sparse coding networks for interpretable image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2369–2377. Computer Vision Foundation / IEEE, 2023. [2](#)
- [70] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25912–25921, 2024. [6](#), [7](#)
- [71] Huabing Zhou, Wei Wu, Yanduo Zhang, Jiayi Ma, and Haibin Ling. Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network. *IEEE Transactions on Multimedia*, 25:635–648, 2023. [1](#), [2](#)
- [72] Man Zhou, Naishan Zheng, Xuanhua He, Danfeng Hong, and Jocelyn Chanussot. Probing synergistic high-order interaction for multi-modal image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [2](#)
- [73] Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7099–7108, 2024. [1](#), [2](#), [3](#)