论文

DCEvo：基于判别式跨维度进化学习的红外与可见光图像融合方法

刘晋源[1]，张博为[2]，梅青云[2]，李星源[2]，
邹阳[3]，姜智颖[4]，马龙[2]，刘日升[2]，樊鑫[2*]

1. 大连理工大学机械工程学院
2. 大连理工大学软件学院与大连理工大学-立命馆大学国际信息与软件学院
3. 西北工业大学计算机科学学院
4. 大连海事大学信息科学技术学院
* 通信作者. E-mail: xin.fan@dlut.edu.cn

摘要 红外与可见光图像融合通过整合不同光谱波段的信息，在保留各自模态优势的同时弥补其局限性，从而提升图像质量。现有方法通常将图像融合与后续高级任务视为独立过程，导致融合图像对任务性能提升有限，且无法为融合过程提供优化。为突破这些限制，本文提出一种判别式跨维度进化学习框架（DCEvo），可同时提升视觉质量与感知精度。本方法利用进化学习的强大搜索能力，将双任务优化建模为多目标问题，采用进化算法（EA）动态平衡损失函数参数。受视觉神经科学启发，我们在编码器与解码器中集成判别增强器（DE），有效学习不同模态的互补特征。此外，我们提出的跨维度嵌入（CDE）模块实现了高级任务特征与低维融合特征的互增强，确保特征融合过程的协同高效。在三个基准数据集上的实验表明，本方法显著优于现有技术，视觉质量平均提升 9.32%，同时有效增强后续高级任务性能。代码已开源至 https://github.com/Beate-Suy-Zhang/DCEvo。

关键词 判别式跨维度进化学习，红外与可见光图像融合，多目标协同优化，跨模态特征互补

1 引言

红外图像能够捕捉热辐射信息，在黑暗环境以及烟雾等条件下表现良好，但存在分辨率较低和纹理细节有限的问题 [1,4]。相反，可见光图像具有高分辨率、丰富的色彩和清晰的边缘细节，但在低光照或恶劣天气条件下性能会大打折扣。因此，将红外图像和可见光图像进行融合，可以提高整体图像质量，显著提升在自动驾驶 [7]、遥感 [11]、安全监控 [10]、医学成像 [13] 和军事侦察 [5] 等复杂环境中的信息获取能力。

为满足多样化应用需求，红外与可见光图像融合（IVIF）技术致力于通过像素级增强生成高质量视觉图像，以提升人类观察者的视觉清晰度与细节感知 [9,10]。此外，该技术还可增强目标检测、场景分析等下游感知任务的精度，实现任务级优化 [2,6]。通过同时达成像素级与任务级的双重提升，IVIF 技术为智能系统提供更精准的环境理解与决策支持。

近年来，深度学习的快速发展极大地推动了 IVIF 领域的进步。基于深度学习的方法在融合性能上显著超越了传统方法 [14]。尽管如此，大多数 IVIF 方法仍然主要关注提升视觉美感，而不是增强后续高级视觉任务的效果，然而这些任务对于众多实际应用来说至关重要 [7]。例如，现有的策略通常将图像融合和目标检测视为不同的问题，仅将检测作为后续处理环节。因此，融合后的图像在检测精度上的提升微乎其微，并且检测结果无法为进一步优化融合过程提供有价值的信息 [5,8,12]。

若摒弃传统的分治策略而尝试协同优化双任务，则需解决以下挑战：**多维特征协调难题**：不同任务在模型推理过程中依赖信息不同，导致适用于图像融合的特征可能不适合后续高级任务，反之亦然。**损失函数调参困境**：现有方案常通过加权组合损失函数联合优化双目标，但其效果高度依赖精细的参数调整。人工调参方法往往以牺牲某一目标为代价提升另一指标，难以实现双任务的最优平衡。

为了解决 IVIF 中**多维特征协调难题**和**损失函数调参困境**，本文提出了一种判别式跨维度进化学习框架，用于实现视觉增强和精确感知。受到进化学习强大搜索能力的启发，我们引入了一种新的方法来应对双任务学习过程中协同优化的挑战。我们将其构建为一个多目标优化问题，利用进化算法（EA）来更新跨任务约束的系数，并学习每个任务的最优参数。为了支撑该框架，我们首先基于视觉神经机制与物理特性设计了一种判别增强器，并将其嵌入模型的编码器和解码器架构中。随后，我们建立了一个跨维度嵌入模块，将高维任务生成的目标感知特征嵌入到低维融合特征中。我们的贡献主要体现在以下四个方面：

- 我们提出了一种用于红外与可见光图像融合的判别式跨维度进化学习方法（DCEvo），所得到的融合图像不仅视觉效果好，而且在任务性能上保持了较高的准确性。据我们所知，这是首次将进化学习集成到 IVIF 领域的研究。
- 为了有效地学习互补特征并增强不同模态的独特特性，我们引入了一种判别增强器，它被很好地集成到编码器和解码器中。
- 我们建立了一个跨维度嵌入模块，在这个模块中，高维任务特征可以监督低维融合特征，反之亦然。这种设计促进了不同维度特征之间的全面交互，从而实现了融合任务和下游任务的相互增强。
- 我们提出了一种进化学习策略，逐步确定损失函数中平衡两个任务的超参数。这种方法克服了经验性超参数调整的挑战，通过系统地优化参数，在相互竞争的目标之间实现平衡。

2 本文方法

本节详细介绍我们的 DCEvo 框架，包括进化学习超参数、判别增强器和跨维度特征嵌入，如图 1 所示。

2.1 进化学习超参数

选择模型超参数仍然是一个重大挑战，传统方法依赖先验知识。然而，基于不同的先验分别优化两个网络，无法同时满足融合和下游任务的要求。此外，总损失函数由多个部分组成，形成了一个多目标问题（MOP），不同部分之间可能存在冲突。进化学习可以在一次运行中同时探索解空间的不同区域，并找到多个非支配解，非常适合解决 MOP 问题。因此，我们采用**进化算法（EA）**来优化损失函数的系数，以改进训练过程，如图 2（c）所示。我们的 EA 使用遗传算法，根据每个解

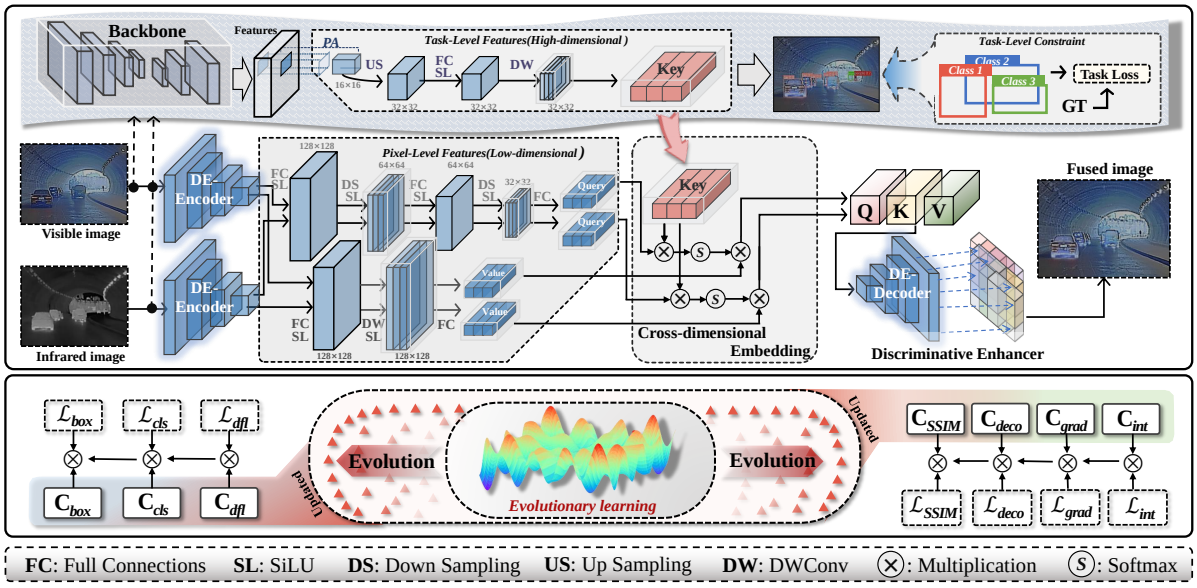
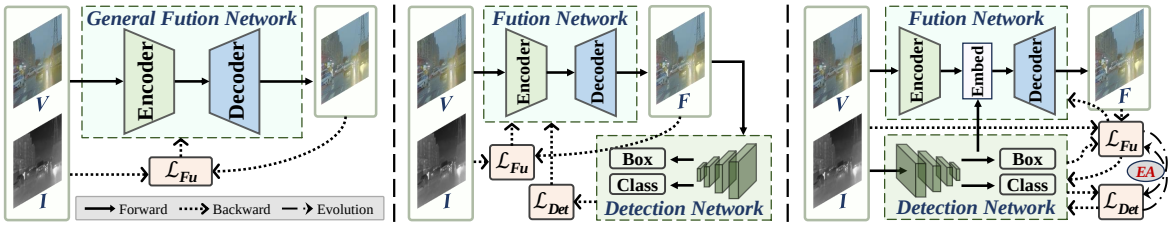


图 1　这是我们 DCEvo 架构的整体示意图。中间部分展示了我们的红外与可见光图像融合网络，该网络通过结合像素级特征和任务级特征来生成图像。上半部分表示检测网络，它嵌入任务级特征以进行融合监督，使融合后的图像包含目标信息。在检测网络和融合网络的协同训练过程中，我们提出了一种进化学习策略来搜索优化目标的系数，如底部部分所示。

的损失值评估其适应度——损失越低，适应度越高。选择适应度高的个体进行交叉操作，增加种群中优质解的比例。变异操作引入随机变化，探索更广阔的解空间，每次迭代通过轮盘赌选择适应度高的个体进入下一代。具体过程如算法 1 所示。



(a) IVIF with only general fusion network.　(b) Target-aware cascaded IVIF network.　(c) Our EA-based training approach.

图 2　面向高级任务的 IVIF 网络工作流程与学习策略对比，策略（a）仅通过低级约束训练融合网络，策略（b）级联检测网络以引导具有额外高级约束的融合网络。我们的策略（c）采用进化算法来协同有效地优化两个任务。

下面我们给出模型公式的相关符号表示。红外图像、可见光图像和融合图像分别记为 $\mathbf{I}_{ir}, \mathbf{I}_{vis}, \mathbf{I}_f$。在图像融合的优化过程中，为了获得具有高视觉质量的融合图像，我们应用了多种约束条件。损失函数为 $\mathcal{L}_{Fu} = \mathcal{L}_{SSIM} + \mathcal{L}_{deco} + \mathcal{L}_{grad} + \mathcal{L}_{int}$，其中 $\mathcal{L}_{SSIM} = \mathcal{L}_{SSIM}(\mathbf{I}_{ir}, \mathbf{I}_f) + \mathcal{L}_{SSIM}(\mathbf{I}_{vis}, \mathbf{I}_f)$，$\mathcal{L}_{int} = \frac{1}{HW} \|\mathbf{I}_f - max(\mathbf{I}_{ir}, \mathbf{I}_{vis})\|$ ，$\mathcal{L}_{grad} = \frac{1}{HW} \||\nabla \mathbf{I}_f| - max(|\nabla \mathbf{I}_{ir}|, |\nabla \mathbf{I}_{vis}|)\|$. $\mathcal{L}_{deco}$ 是在文献 [13] 中提出的，用于关联跨模态输入的基本特征。

目标检测部分的损失函数由三部分组成，即 $\mathcal{L}_{Det} = \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{dfl}$ 第一部分是分类损失，计

算法 1 进化学习超参数优化

输入: Number of population N, Number of iterations I_n, Probability of population individual mutating P_m

输出: Global best solution G_b

Start with a random initial population P_1
Calculate initial population fitness F_1
for t = 1 to I_n **do**
    Sort (population, descending, F_t)
    Select individuals with high fitness for crossover
    NewIndividuals ⟵ **Crossover**()
        Q_t ⟵ NewIndividuals
    for k = 1 to N **do**
        $x \sim U(0, 100)$
        if x < P_m **then**
            Mutation(k-th individual)
        end if
    end for
    NewIndividuals ⟵ **Mutate**()
    Q_t ⟵ Q_t∪ NewIndividuals
    Calculate population M_t fitness
    Calculate **selected** probability of each individual
        $p\left(x_i\right)=\frac{\text{fitness }\left(x_i\right)}{\sum_{j=1}^{\left|M_t\right|}\text{fitness}\left(x_j\right)}, \quad \forall x_i \in M_t$
    Selecting a new generation of population through
        the **roulette wheel**, $r \sim U(0, 1)$
        $P_{t+1}=\left\{x_i \mid \sum_{i=1}^{k-1} p(x_i) \leqslant r \leqslant \sum_{i=1}^{k} p(x_i)\right\}$
end for

算公式如下:

$$\mathcal{L}_{cls}=-\sum_{i=1}^{N_{cls}}[\mathbf{y}_i \log(1+e^{-\hat{\mathbf{y}}_i})+(1-\mathbf{y}_i)\log(1+e^{\hat{\mathbf{y}}_i})], \tag{1}$$

其中 $\hat{\mathbf{y}}$ 是预测的类别标签，$\mathbf{y}$ 是真实标签，N_{cls} 是类别数量。第二部分是完全交并比损失，计算公式如下:

$$\mathcal{L}_{box}=1-(IOU-\frac{d_2^2}{d_C^2}-\frac{v^2}{(1-IOU)+v}), \tag{2}$$

其中，$v=\frac{4}{\pi^2}(\arctan\frac{w_g}{h_g}-\arctan\frac{w_p}{h_p})^2$, d_2 是两个边界框中心点之间的欧几里得距离，d_C 是最小外接矩形的对角线距离。w_g 和 h_g 是真实边界框的宽度和高度，而 w_p 和 h_p 是预测边界框的宽度和高度。第三部分是分布焦点损失 $\mathcal{L}_{dfl}$，用于快速聚焦于标签框 [3]。

2.2 判别增强器

在 IVIF 中，特征提取网络至关重要，但特征图的固有属性却很少得到关注。为此，我们引入**判别增强器（DE）**来提升特征表示能力，进而提高网络的学习效率。

在特征表示中，每个特征图都包含物体的特定语义信息。为了区分物体和背景特征，对于单个特征图 $\mathbf{X} \in \mathbb{R}^{H\times W}$，假设 $\mathbf{X}=\mathbf{X}_o+\mathbf{X}_b$，其中 $\mathbf{X}_o$ 和 $\mathbf{X}_b$ 分别是仅包含物体像素和背景像素的特征图。$\mathbf{X}$、$\mathbf{X}_o$、$\mathbf{X}_b$ 中的像素值都在 $[0,1]$ 范围内。$\mathbf{X}$ 的均值为 $\mu=\frac{1}{M}\sum_{i=1}^{M}x_i$，其中 x_i 表示特征图 $\mathbf{X}$ 中的一个像素，$M=H\times W$。因此，增大 $||\mathbf{X}_o-\mathbf{X}_b||_1$ 的值是很有必要的。$\mathbf{X}_o$ 的均值为 $\mu_o=\frac{1}{M_o}\sum_{i=1}^{M_o}x_{o,i}\in(\mu,1]$，而 $\mathbf{X}_b$ 的均值为 $\mu_b=\frac{1}{M_b}\sum_{i=1}^{M_b}x_{b,i}\in[0,\mu)$。由于在 IVIF 中，有温度的

物体是突出显示的，我们假设 $\mu_0 > \mu > \mu_b$。因此，为了对物体的周围环境进行调制，可以增大差异 $D(\mathbf{X}) = \sum_{i=1}^{M} |x_i - \mu|$。由于 $\mu_o > \mu > \mu_b$，很容易看出 $D(\mathbf{X}) < \sum_{i=1}^{M} \left| \frac{x_i^2}{\mu} - \mu \right|$，这表明 $\frac{x_i^2}{\mu}$ 的值比 x_i 更能起到调制作用。

需要注意的是，这个公式需要满足 $\mathbf{X}$、$\mathbf{X}_o$、$\mathbf{X}_b > 0$ 的条件。由于 Sigmoid 激活函数 $S(x)$ 具有单调性和有界性，我们使用它来进行重要性值映射。因此，我们使用 Sigmoid 激活函数将像素值映射到 $y_i = S(x_i) \in (0, 1)$，其均值为 μ_y。我们为像素 x_i 赋予重要性权重以增强特征。根据 $\mathbf{X}$ 的周围环境，增强特征图 $\widetilde{\mathbf{X}}$ 中的每个增强像素为：

$$\widetilde{x_i} = x_i \times S(\frac{y_i^2}{\mu_y}), \tag{3}$$

其中 $S(\cdot)$ 表示 Sigmoid 激活函数。

我们在融合编码器和解码器中都使用了 DE，它对物体的周围进行调制，以生成具有突出物体信息的融合图像。

2.3 跨维度特征嵌入

除了突出的物体表示，我们提出一种**跨维度特征嵌入（CDE）**方法，用于监督融合网络整合跨任务特征。

编码器 - 解码器网络能够有效地表示多模态特征。然而，它们忽略了不同任务的特征表示之间的相互影响。任务级模型可以提取高维信息。例如，在目标检测中，特征金字塔网络能够将物体与周围环境区分开来，为检测头提供语义信息。因此，我们的 DCEvo 方法通过使用 CDE 模块利用检测特征，以整合来自两个任务的特征。

如图 1 所示，融合编码器将红外图像 $\mathbf{I}_{ir}$ 和可见光图像 $\mathbf{I}_{vis}$ 分别转换为特征 $\mathbf{F}_{ir}$ 和 $\mathbf{F}_{vis}$，而检测网络将 $\mathbf{I}_{vis}$ 转换为 $\mathbf{F}_{det}$。与其他任务中的特征嵌入不同，图像特征融合要求两个输入的特征图包含相同的场景实体。为了正确匹配所表示的特征，我们对检测特征图 $\mathbf{F}_{det}$ 和切片的输入融合块进行对齐操作。切片后的 $\mathbf{F}_{det}$ 特征标记为 $\mathbf{F}_{pdet}$。然后，$\mathbf{F}_{pdet}$、$\mathbf{F}_{ir}$ 和 $\mathbf{F}_{vis}$ 被输入到不同的 CNN 模块中，以获得 $\widehat{\mathbf{F}}_{pdet}$、$\widehat{\mathbf{F}}_{ir}$ 和 $\widehat{\mathbf{F}}_{vis}$ 用于特征嵌入。令 $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} = \{\widehat{\mathbf{F}}_{pdet}, \widehat{\mathbf{F}}_{ir}, \widehat{\mathbf{F}}_{vis}\}$，那么我们的跨维度嵌入操作输出 $\widehat{\mathbf{F}}_{ca}$ 为：

$$\widehat{\mathbf{F}}_{ca} = \mathcal{CDE}(\mathbf{Q}, \mathbf{K}, \mathbf{V}), \tag{4}$$

其中 $\mathcal{CDE}(\cdot)$ 表示跨纬度嵌入。

通过上述特征嵌入，检测特征被集成到图像融合网络中，使融合图像具有目标感知能力。随后，特征图 $\widehat{\mathbf{F}}_{ca}$ 被输入到自注意力模块中。最后，图像解码器生成融合图像。

参考文献

1 Wassim A. El Ahmar, Dhanvin Kolhatkar, Farzan Erlik Nowruzi, Hamzah AlGhamdi, Jonathan Hou, and Robert Laganiere. Multiple object detection and tracking in the thermal spectrum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 277–285, June 2022.

2 Mengyue Geng, Lin Zhu, Lizhi Wang, Wei Zhang, Ruiqin Xiong, and Yonghong Tian. Event-based visible and infrared fusion via multi-task collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26929–26939, June 2024.

3 Xiang Li, Chengqi Lv, Wenhai Wang, Gang Li, Lingfeng Yang, and Jian Yang. Generalized focal loss: Towards efficient representation learning for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3139–3153, 2023.

4 Xingyuan Li, Jinyuan Liu, Zhixin Chen, Yang Zou, Long Ma, Xin Fan, and Risheng Liu. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *European Conference on Computer Vision*, pages 270–288. Springer, 2024.

5 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*, pages 5802–5811, June 2022.

6 Jinyuan Liu, Xingyuan Li, Zirui Wang, Zhiying Jiang, Wei Zhong, Wei Fan, and Bin Xu. Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*, 2024.

7 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*, pages 8115–8124, October 2023.

8 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.

9 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.

10 Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27026–27035, June 2024.

11 Zhenjie Yu, Shuang Li, Yirui Shen, Chi Harold Liu, and Shuigen Wang. On the difficulty of unpaired infrared-to-visible video translation: Fine-grained content-rich patches transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1631–1640, June 2023.

12 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, June 2023.

13 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, June 2023.

14 Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, and Feng Zhao. Probing synergistic high-order interaction in infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26384–26395, June 2024.