Where Elegance Meets Precision: Towards a Compact, Automatic, and Flexible Framework for Multi-modality Image Fusion and Applications

Jinyuan Liu[1] , **Guanyao Wu**[2] , **Zhu Liu**[2] , **Long Ma**[2] , **Risheng Liu**[2] , **Xin Fan**[2*]

[1]School of Mechanical Engineering, Dalian University of Technology
[2]School of Software Technology, Dalian University of Technology
atlantis918@hotmail.com, rollingplainko@gmail.com, {rsliu, xin.fan}@dlut.edu.cn

Abstract

Multi-modality image fusion aims to integrate images from multiple sensors, producing an image that is visually appealing and offers more comprehensive information than any single one. To ensure high visual quality and facilitate accurate subsequent perception tasks, previous methods have often cascaded networks using weighted loss functions. However, such simplistic strategies struggle to truly achieve the "Best of Both Worlds", and the adjustment of numerous hand-crafted parameters becomes burdensome. To address these challenges, this paper introduces a Compact, Automatic and Flexible framework, dubbed CAF, designed for infrared and visible image fusion, along with subsequent tasks. Concretely, we recast the combined problem of fusion and perception into a single objective, allowing mutual optimization of information from both tasks. Then we also utilize the perception task to inform the design of fusion loss functions, facilitating the automatic identification of optimal fusion objectives tailored to the task. Furthermore, CAF can support seamless integration with existing approaches easily, offering flexibility in adapting to various tasks and network structures. Extensive experiments demonstrate the superiority of CAF, which not only produces visually admirable fused results but also realizes 1.7 higher detection mAP@.5 and 2.0 higher segmentation mIoU than the state-of-the-art methods. The code is available at https://github.com/RollingPlain/CAF_IVIF.

1 Introduction

With the increasing complexity of application scenarios in intelligent systems, the drawbacks of single-type sensors have become apparent. They are incapable of providing a precise and comprehensive description of scenes or targets, leading to difficulties in accomplishing tasks in real-world challenging environments [Ma *et al.*, 2019a]. To address this, image fusion technology has emerged, which aims to conduct comprehensive analysis and processing of information from different sources, achieving a consistent description of the observed scenes or targets. Among these, infrared and visible sensors play a crucial role in the perception process of intelligent systems. However, both infrared and visible images have their own limitations in diverse scenes [Zhang and Demiris, 2023]. Therefore, fusing the infrared and visible images can intelligent systems maintain stable perception in dynamically changing environments.

Drawing on strong non-linear fitting ability of Convolutional Neural Networks (CNNs) [Liu *et al.*, 2023c; Liu *et al.*, 2023b], extensive learning-based Infrared and Visible Image Fusion (IVIF) approaches have been proposed and applied well [Ma *et al.*, 2019b; Liu *et al.*, 2022a; Xu *et al.*, 2020a]. A naive way to solve IVIF and its downstream tasks is to cascade dual networks with weighted mean functions, with achieving certain performance improvement [Liu *et al.*, 2022a; Sun *et al.*, 2022; Cao *et al.*, 2023]. However, such a direct alternative optimization process make the dual networks incline to focus on only one task, thus bringing imbalanced performance across the dual tasks. Besides, the joint loss function still relies on empirical designing, and the tuning hypter-parameter is laborious, placing an obstacle to achieve high perception accuracy while generating favorable fused images for human inspection.

In the realm of IVIF, there is no ground truth for generating expected fused images. Consequently, the design of the loss function plays a pivotal role in determining the final outcome. That is to say, devising a formulation that seamlessly integrates the two tasks is a fundamental methodology for achieving superior fusion and perception results. Following this observation, this study addresses these critical issues by consolidating the fusion and perception tasks into a single objective, culminating in a Compact, Automatic and Flexible Framework, dubbed CAF. Specifically, we formulate the fusion and perception tasks from a coupling constraint optimization prospective, wherein each component is tailored to its respective loss function. Then, we optimize the dual tasks by introducing a new loss searching scheme. By allowing the perception task to guide the entire search process, the fusion loss function can be automatically derived, and the key features from both tasks are mutually reinforced and optimized.

Furthermore, the proposed CAF can be seamlessly integrated into existing fusion network architectures, outperforming them in terms of results, thereby showcasing the flexibility and generalization capability of our CAF.

The main contribution can be distilled as three main aspects as follows:

- **Compact Formulation.** From a new aspect, we embrace the image fusion and perception in a coupling constraint optimization. By constructing corresponding loss functions at different levels, the dual tasks mutually constrain each other, thereby achieving a tightly integration.
- **Automatic Optimization.** Considering the need to reduce the burden of laborious parameter tuning and optimize the dual tasks, we employ the guidance of higher-level tasks to automatically optimize and search for suitable fusion loss functions. To the best of our knowledge, this is the first time that loss function searching has been introduced to the IVIF community.
- **Flexible Framework.** It is worth mentioning that CAF is a "ready-to-use" framework. We have extended it to several existing state-of-the-art fusion network architectures, achieving remarkable progress in both fusion and perception performance compared to their original versions. This demonstrates the flexibility and generalizability of our proposed CAF.

Our method achieves superior fused results on par with existing state-of-the-art methods and yields significant improvements for perception tasks (1.7 higher mAP@.5 in detection and 2.0 higher mIoU in segmentation).

2 Related Works

Deep learning has achieved significant results in the field of IVIF [Liu *et al.*, 2023a; Zhao *et al.*, 2023a; Zhao *et al.*, 2023b; Liu *et al.*, 2023d; Liu *et al.*, 2022b] due to its powerful feature extraction capability based on neural networks. As a task without ground truth, numerous loss functions have been proposed and widely used. In early methods focused on visual effects, Li et al. [Li and Wu, 2018] made a preliminary attempt at deep learning, with a loss function composed of two parts to constrain at both the pixel and structural levels. Ma et al. [Ma *et al.*, 2019b] introduced adversarial constraint loss between the fused image and the visible image; furthermore, they [Ma *et al.*, 2020] adjusted the strategy of adversarial loss to estimate the distribution of the visible and infrared domains simultaneously, achieving significant contrast and rich texture details. However, these methods all involve laborious manual design of losses and adjustment of weights. Xu et al. [Xu *et al.*, 2020b] introduced an image quality detection mechanism to dynamically control the weights between different parts of the loss, creating a versatile fusion framework. Besides, Xu et al. [Xu *et al.*, 2020a] introduced information fidelity to dynamically balance the impact of the two source images on the loss function, which somewhat reduced the burden of parameter tuning. However, these attempts could not eliminate the need for designing the loss function itself.

Nowadays, adjusting the loss functions between different tasks has become an main means of adapting fusion methods [Wu *et al.*, 2024] to perception tasks. For object detection, Liu et al. [Liu *et al.*, 2022a] proposed a dual-level adversarial learning network, while Sun et al. [Sun *et al.*, 2022] proposed a detection-driven network; both of which cascade networks for different tasks to merge their respective loss functions. For semantic segmentation, Tang et al. [Tang *et al.*, 2022] also followed a similar approach, using semantic loss to guide the fusion network to acquire more high-level semantic information. However, this kind of weighting of task losses is difficult to achieve a perfect balance to 'Best of Both Worlds'.

3 The Proposed Method

In this section, we first given a compact formulation to model the relationship of the dual tasks. Then we employ the high-level tasks to automatically optimize model in a loss-free manner. In the end, the hierarchical training is developed to achieve "Best of Both Worlds".

3.1 Problem Formulation

Given a pairs of aligned infrared and visible images, our goal is to generate a visual-appealing fused image, along with high perception accuracy. The infrared, visible and fused images are all gray-scale with the size of $m \times n$, denoted as column vectors $\mathbf{I}_{\mathtt{ir}}$, $\mathbf{I}_{\mathtt{vis}}$, and $\mathbf{I}_{\mathtt{f}} \in \mathbf{R}^{mn\times 1}$, respectively. We formulate the dual tasks into one goal, in which the respected parameters are mutually interacted. The coupling constraint formulation can be written as:

$$\min_{\omega_{\mathtt{t}}} \mathcal{L}^{\mathtt{t}}\big(\mathcal{T}(\mathbf{I}_{\mathrm{f}}, \boldsymbol{\omega}_{\mathtt{t}})\big), s.t. \begin{cases} \mathbf{I}_{\mathrm{f}} = \mathcal{F}(\mathbf{I}_{\mathrm{vis}}, \mathbf{I}_{\mathrm{ir}}; \boldsymbol{\omega}_{\mathtt{f}}^{*}), \\ \boldsymbol{\omega}_{\mathtt{f}}^{*} \in \arg\min_{\boldsymbol{\omega}_{\mathtt{f}}} \mathcal{L}^{\mathtt{f}}\big(\mathcal{F}(\mathbf{I}_{\mathrm{vis}}, \mathbf{I}_{\mathrm{ir}}; \boldsymbol{\omega}_{\mathtt{f}})\big). \end{cases} \tag{1}$$

where $\mathcal{L}^{\mathtt{t}}$ is the task-specific training loss, and $\mathcal{T}$ denotes a perception task network with learnable parameters $\boldsymbol{\omega}_{\mathtt{t}}$. $\mathcal{L}^{\mathtt{f}}$ is the fusion loss, and $\mathcal{F}$ represents fusion network with learnable parameters $\boldsymbol{\omega}_{\mathtt{f}}$. The fusion process can be symbolized as $\mathbf{I}_{\mathtt{f}} = \mathcal{F}(\mathbf{I}_{\mathtt{vis}}, \mathbf{I}_{\mathtt{ir}}; \boldsymbol{\omega}_{\mathtt{f}})$. For the perception task, the objective is to generate an output based on $\mathbf{I}_{\mathtt{f}}$, incorporating its own loss function to guide the training process.

Compared to previous methods focused primarily on visual effects, our proposed model delineates a distinct relationship, capturing the inherent correspondence between the two tasks. The performance of $\mathbf{I}_{\mathtt{f}}$ hinges on defining the training loss, shaped by meticulously crafted hyper-parameters. However, selecting the appropriate loss type and adjusting weight hyper-parameters pose considerable challenges. To address this,e present a pioneering strategy to navigate these intricacies associated with loss selection and labor-intensive tuning.

3.2 Automatic Loss Search

As aforementioned, $\mathcal{L}^{\mathtt{f}}$ plays a significant role to investigate the complementary modal characteristics for diverse perception tasks. To overcome the core obstacle, we present the automatic design scheme to identify task-specific weights that guide the fusion process. Denoting these weighted parameters as $\boldsymbol{\beta}$, the aggregated losses based on $\mathcal{L}^{\mathtt{f}}$ can be expressed as $\mathcal{L}^{\mathtt{f}} = \sum_{i=1}^{n} \beta_i \mathcal{L}_i$. Specially, we utilize the bi-level formulation to elucidate the automatic search process, which can be written as following:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \mathcal{L}^{\mathtt{t}}_{\mathtt{val}}\big(\boldsymbol{\beta}; \mathcal{T}(\mathcal{F}(\mathbf{I}_{\mathtt{vis}}, \mathbf{I}_{\mathtt{ir}}; \boldsymbol{\omega}_{\mathtt{f}}^{*}); \boldsymbol{\omega}_{\mathtt{t}})\big), \\ &s.t.\ \boldsymbol{\omega}_{\mathtt{f}}^{*} \in \arg\min_{\boldsymbol{\omega}_{\mathtt{f}}} \mathcal{L}^{\mathtt{f}}_{\mathtt{train}}\big(\mathcal{F}(\mathbf{I}_{\mathtt{vis}}, \mathbf{I}_{\mathtt{ir}}; \boldsymbol{\omega}_{\mathtt{f}}); \boldsymbol{\beta}\big), \end{aligned} \tag{2}$$

where "train" and "val" denote the losses on the training and validation datasets, respectively. The upper-level objective seeks to minimize the weighted parameters, gauging them against the measurements of perception tasks derived from the fused images. Conversely, the lower-level optimization aims to ascertain the optimal fusion parameters in the context of the provided $\boldsymbol{\beta}$. We contend that this formulation aptly achieves task-specific image fusion. On the one hand, the image fusion constraint aids the perception task in precisely assessing the impact of $\boldsymbol{\beta}$. On the other hand, feedback from downstream perception tasks can encourage adjustments in $\boldsymbol{\beta}$ to further refine the fusion process.

However, solving the bi-level optimization in Eq. (2) has been demonstrated as a challenging task, which acquires the massive computation resource to address the nested optimization [Liu *et al.*, 2021c; Liu *et al.*, 2021d]. In order to effectively solve Eq. (2), inspired by the remarkably success of differentiable architecture search [Liu *et al.*, 2019], we adopt the one-step truncated approximation strategy. To simplify the representation, we utilize the $\mathcal{L}^{\mathtt{t}}_{\mathtt{val}}(\boldsymbol{\beta};\boldsymbol{\omega}^*_{\mathtt{f}})$ and $\mathcal{L}^{\mathtt{f}}_{\mathtt{train}}(\boldsymbol{\beta};\boldsymbol{\omega}_{\mathtt{f}})$ to concisely describe the upper and lower objectives in Eq. (2). In detail, the gradient of upper-level objective can be approximated as $\nabla_{\boldsymbol{\beta}}\mathcal{L}^{\mathtt{t}}_{\mathtt{val}}(\boldsymbol{\beta};\boldsymbol{\omega}^*_{\mathtt{f}}) \approx \nabla_{\boldsymbol{\beta}}\mathcal{L}^{\mathtt{t}}_{\mathtt{val}}(\boldsymbol{\beta};\boldsymbol{\omega}_{\mathtt{f}} - \eta\nabla_{\boldsymbol{\omega}_{\mathtt{f}}}\mathcal{L}^{\mathtt{f}}_{\mathtt{train}}(\boldsymbol{\beta};\boldsymbol{\omega}_{\mathtt{f}}))$, where η denotes the learning rate of fusion. In this way, $\boldsymbol{\omega}^*_{\mathtt{f}}$ can be approximated by the one-step gradient of image fusion. Moreover, in order to address the hessian matrix, we further utilize the finite difference approximation to solve this second-order gradient, which can describe the intrinsic latent relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\omega}_{\mathtt{f}}$. Denoted $\boldsymbol{\omega}_{\mathtt{f}} - \eta\nabla_{\boldsymbol{\omega}_{\mathtt{f}}}\mathcal{L}^{\mathtt{f}}_{\mathtt{train}}(\boldsymbol{\beta};\boldsymbol{\omega}_{\mathtt{f}})$ as $\boldsymbol{\omega}^{'}_{\mathtt{f}}$, the finite difference approximation can be written as

$$\nabla^2_{\boldsymbol{\beta},\boldsymbol{\omega}_{\mathtt{f}}}\mathcal{L}^{\mathtt{t}}_{\mathtt{val}}\nabla_{\boldsymbol{\omega}^{'}_{\mathtt{f}}}\mathcal{L}^{\mathtt{f}}_{\mathtt{train}} \approx \frac{\nabla_{\boldsymbol{\beta}}\mathcal{L}^{\mathtt{f}}_{\mathtt{train}}(\boldsymbol{\beta};\boldsymbol{\omega}^{+}_{\mathtt{f}}) - \nabla_{\boldsymbol{\beta}}\mathcal{L}^{\mathtt{f}}_{\mathtt{train}}(\boldsymbol{\beta};\boldsymbol{\omega}^{-}_{\mathtt{f}})}{2\delta}, \tag{3}$$

where δ is a constant and we define $\boldsymbol{\omega}^{\pm}_{\mathtt{f}} = \boldsymbol{\omega}_{\mathtt{f}} \pm \delta\nabla_{\boldsymbol{\omega}^{'}_{\mathtt{f}}}\mathcal{L}^{\mathtt{t}}_{\mathtt{val}}(\boldsymbol{\beta};\boldsymbol{\omega}^{'}_{\mathtt{f}})$. Thus, we obtain the efficiency approximated gradient update of upper-level objective to perform the automatic search for $\mathcal{L}^{\mathtt{f}}$. We also elaborately illustrate the search procedure in Algorithm 1.

Algorithm 1 Automatic Loss Function Search

Require: The loss functions $\mathcal{L}^{\mathtt{t}}$, $\mathcal{L}^{\mathtt{f}}$, and other necessary hyper-parameters.
Ensure: The optimal parameters $\boldsymbol{\beta}^*$.

```
while not converged do
    % Optimizing the perception task weights ω_t;
    ω^t ← ω^t − ∇L^t_train(I_f; ω^t);
    % Optimizing the image fusion network;
    ω^f ← ω^f − ∇L^f_train(I_vis, I_ir; ω^f, β);
    % Optimizing the hyper-parameter β using first-order approximation;
    β ← β − ∇_β L^t_val(I_vis, I_ir; β, ω_f − η∇_{ω_f} L^f_train(β; ω_f)));
end while
return β*.
```

Name	Expression	X	Variable
$\mathcal{L}_1$	$\|\mathbf{F}-\mathbf{X}\|_1$	$\mathbf{I}_{ir}$	λ_a
		$\mathbf{I}_{vi}$	λ_b
		$\mathrm{Max}(\mathbf{I}_{ir},\mathbf{I}_{vi})$	λ_c
	$\|\mathcal{M}\odot\mathbf{F}-\mathbf{X}_1\|_1+\left\|\tilde{\mathcal{M}}\odot\mathbf{F}-\mathbf{X}_2\right\|_1$	$\mathbf{X}_1:\mathcal{M}\odot\mathbf{I}_{ir}$	λ_d
		$\mathbf{X}_2:\tilde{\mathcal{M}}\odot\mathbf{I}_{vi}$	λ_e
$\mathcal{L}_{MSE}$	$\|\mathbf{F}-\mathbf{X}\|_2$	$\mathbf{I}_{ir}$	λ_f
		$\mathbf{I}_{vi}$	λ_g
		$\mathrm{Max}(\mathbf{I}_{ir},\mathbf{I}_{vi})$	λ_h
	$\|\mathcal{M}\odot\mathbf{F}-\mathbf{X}_1\|_2+\left\|\tilde{\mathcal{M}}\odot\mathbf{F}-\mathbf{X}_2\right\|_2$	$\mathbf{X}_1:\mathcal{M}\odot\mathbf{I}_{ir}$	
		$\mathbf{X}_2:\tilde{\mathcal{M}}\odot\mathbf{I}_{vi}$	
$\mathcal{L}_{SSIM}$	$1-SSIM(\mathbf{F},\mathbf{X})$	$\mathbf{I}_{ir}$	λ_i
		$\mathbf{I}_{vi}$	λ_j
$\mathcal{L}_{grad}$	$\frac{1}{HW}\|\nabla\mathbf{F}-\mathbf{X}\|_1$	$\nabla\mathbf{I}_{ir}$	λ_k
		$\nabla\mathbf{I}_{vi}$	λ_l
		$\mathrm{Max}(\nabla\mathbf{I}_{ir},\nabla\mathbf{I}_{vi})$	λ_m
$\mathcal{L}_{per}$	$\sum_{l=1}^{L}\frac{1}{C_lH_lW_l}\|\varphi_l(\mathbf{F})-\varphi_l(\mathbf{X})\|_1^2$	$\mathbf{I}_{ir}$	λ_n
		$\mathbf{I}_{vi}$	λ_o

Figure 1: The search space of fusion loss functions.

3.3 Loss Search Space

We introduce 4 distinct loss functions: pixel-level loss (comprising $\mathcal{L}_1$ and $\mathcal{L}_{\mathtt{MSE}}$), structural similarity loss $\mathcal{L}_{\mathtt{SSIM}}$, gradient loss $\mathcal{L}_{\mathtt{grad}}$, and perceptual loss $\mathcal{L}_{\mathtt{per}}$. Each function is tailored to measure specific image characteristics. Detailed formulations of loss functions are presented in Figure 1.

- Pixel-wise similarity loss, including $\mathcal{L}_1$ and $\mathcal{L}_{\mathtt{MSE}}$, are introduced, which aim to ensure that the pixel intensity of the fused image is consistent with reference images with $\mathtt{p} = \{1, 2\}$. $\mathcal{L}_1$ has been proven to be effective in previous fusion work[Huang *et al.*, 2022; Liu *et al.*, 2022a; Sun *et al.*, 2022; Ma *et al.*, 2021], while $\mathcal{L}_{\mathtt{MSE}}$ has also widely applied[Xu *et al.*, 2020a; Liu *et al.*, 2021a]. Specifically, we consider there candidate references, including source images ($\mathbf{I}_{\mathtt{vis}}$ and $\mathbf{I}_{\mathtt{ir}}$), pixel intensity maximum ($\max(\mathbf{I}_{\mathtt{vis}}, \mathbf{I}_{\mathtt{ir}})$), and visual saliency guidance ($\mathcal{M}(\mathbf{I}_{\mathtt{vis}}, \mathbf{I}_{\mathtt{ir}})$), calculated by [Liu *et al.*, 2022a]. $\mathcal{M}$ denotes the aggregation operation based on saliency mask, which can be written as $\mathtt{Mask} \otimes \mathbf{I}_{\mathtt{ir}} + (1 - \mathtt{Mask}) \otimes \mathbf{I}_{\mathtt{vis}}$.
- Structural similarity (SSIM) [Wang *et al.*, 2004] has been widely used in the field of image fusion[Zhao *et al.*, 2020; Xu *et al.*, 2020a; Liu *et al.*, 2022a; Liu *et al.*, 2021b]. It considers the similarity in brightness, contrast, and structure, which can better reflect human perception of image quality.
- Gradient loss is added into the search space, which focuses on the details and local edge information of the image, widely used in [Zhang and Ma, 2021; Wang *et al.*, 2022; Sun *et al.*, 2022; Ma *et al.*, 2021].
- Perceptual loss, compared with others, focuses more on the perceptual quality and visual effects of images[Ledig *et al.*, 2017], which is widely used in the fields of image fusion, generation, and restoration[Wang *et al.*, 2022;

Han *et al.*, 2022]. In the formulation, φ_l denotes the l-th layer in VGG-16 and $L = \{1, 3, 5, 9, 13\}$.

3.4 Hierarchical Training

After obtaining the optimal combination β for the $\mathcal{L}^{\mathrm{F}}$, we can further solve the major objective Eq. (1) with the hierarchical training. Due to the huge gradient computation of parameters for this nested formulation, instead of utilizing the exact solutions and above approximation directly, we introduce the stage-wise gradual learning to hierarchically address this objective. In details, we first optimize the constraint to obtain the optimal parameters of fusion (*i.e.,* $\boldsymbol{\omega}_{\mathrm{f}}^*$). Then we conduct the optimization for the upper-level objective to obtain the desired parameters of perception. Owing to the link and guidance that the proposed CAF established between the task and the fusion network during the searching, we can employ the training for fusion and perception separately. This strategy not only circumvents the potential negative influence of task loss on the fusion network parameters but also obviates the rigid demand for labeled data intrinsic to cascaded networks. Importantly, this approach realizes "Best of Both Worlds", fulfilling the inherent requirements of the image fusion and perception tasks.

Discussion. We argue that our framework has two significant characteristics, including model-irrelevant generalization and task-related flexibility. Firstly, our framework is generic enough to replace the arbitrary fusion networks and discover suitable loss combination for improve perception and visual quality. Secondly, due to the flexible formulation (Eq. (1)) and effective task-guided loss search (Eq. (2)), our framework is effective to address various multi-modality semantic perception tasks, without introducing redundant auxiliary learning mechanisms (*i.e.,* feature interaction). The related experimental details are reported below.

4 Experiments

4.1 Implementation Details

Search configurations. We employed labeled datasets M^{3}FD [Liu *et al.*, 2022a] and MFNet [Ha *et al.*, 2017] for searching, with the former being randomly partitioned and the latter using the standard division. The fusion network $\mathcal{F}$ is simply composed of three dense blocks, without any specific design. We selected state-of-the-art perceptual task methods (YOLOv5[1] for detection, and Segformer [Xie *et al.*, 2021] for segmentation) as the task network $\mathcal{T}$ to integrate into the framework, adopting the provided *s/mb1* models and maintaining their original loss functions. During the search for the fusion hyper-parameters matrix β, we used the Adam optimizer with an initial learning rate of 5e-2 for 10 epochs of iterations, and adopted a step learning rate decay strategy at a rate of 0.998.

Training configurations. With the searched loss function $\mathcal{L}_{\mathrm{f}}$, we optimize $\boldsymbol{\omega}_{\mathrm{f}}$ on 1k image blocks collected from M^{3}FD, TNO [Toet, 2017], and RoadScene [Xu *et al.*, 2020a], using the Adam optimizer with a learning rate of 1e-4 for 100 epochs. After training the fusion network $\mathcal{F}$, the generated fusion images are fed into the task network for fine-tuning to test task performance. Except for the training rounds (100 epochs for detection, 4k iterations for segmentation), all other settings remain unchanged as per the original source code. All search and training experiments are performed on two NVIDIA Tesla V100 GPUs within PyTorch framework.

In the following, we compared with 7 SOTA competitors, including DIDFuse [Zhao *et al.*, 2020], SDNet [Zhang and Ma, 2021], UMFusion [Wang *et al.*, 2022], DeFusion [Liang *et al.*, 2022], ReCoNet [Huang *et al.*, 2022], U2Fusion [Xu *et al.*, 2020a], and TarDAL [Liu *et al.*, 2022a][2].

4.2 Evaluation in Object Detection

Qualitative detection comparisons. We visualize some results of object detection in Figure 2. The first set of images are daytime road scenes, in which pedestrians are basically invisible in the visible images, while the infrared images have significant target information. The methods in the first row of red area fail to fully retain useful information from both modalities, and their low-contrast fusion results do not support high-precision detection. Benefiting from the detection-favorable loss obtained through searching, our method significantly outperforms the SOTA method TarDAL, in the detection performance of this scene. The second set of images features small targets in overcast, bright night scenes. Similarly, our method achieves the highest detection accuracy. However, TarDAL misses the car in the lower right corner due to excessive contrast.

Quantitative detection comparisons. Table 1 presents the object detection results of various fusion methods on the M^{3}FD and Multispectral [Takumi *et al.*, 2017] datasets, which are measured by average precision (AP) and mean average precision (mAP). The former includes a variety of challenging road scenes, while the latter mostly comprises low-resolution campus scenes. The two source image modalities show particularly notable results in the categories of people and motorcycle, where other methods focusing on visual effects cannot maintain consistent superior detection performance. The SOTA method TarDAL, achieves second best in multiple categories, but its overall performance still has a considerable gap compared to our method on both datasets. That also verifies that our search framework can effectively provide a task-appropriate loss to boost performance.

4.3 Evaluation in Semantic Segmentation

Qualitative segmentation comparisons. Figure 3 displays the visual effects of segmentation, where the labels are embedded in the source images. In both daytime and nighttime scenarios, infrared and visible images exhibit complementary segmentation expressions characteristic of their own modalities. Notably, the proposed method effectively withstands the interference from infrared to segment thermally insensitive objects (*e.g.,* the car in the night scene). On the other hand, other fusion-based methods cannot correctly predict thermally sensitive objects (*e.g.,* the pedestrian in the

[1]https://github.com/ultralytics/yolov5

[2]The version guided by detection is selected for comparision

Method	Source	M^3FD dataset								Multispectral dataset						
		People	Car	Bus	Lamp	Motor	Truck	mAP@.5:.95	mAP@.5	Person	Car	Bike	Cone	CStop	mAP@0.5:.95	mAP@.5
Infrared	-	**31.6**	50.5	58.6	17.1	25.6	43.3	37.8	65.3	**35.3**	40.2	**28.5**	12.8	27.6	28.9	57.0
Visible	-	21.1	50.3	55.2	25.0	**29.0**	42.8	37.2	65.4	27.5	41.9	26.3	20.0	30.8	29.3	58.9
DIDFuse	IJCAI'20	28.1	50.9	57.8	24.1	25.3	45.1	38.6	66.3	32.8	42.1	26.7	19.9	29.6	30.2	61.3
SDNet	IJCV'21	29.8	52.3	57.3	24.3	24.8	43.6	38.7	67.2	34.3	41.6	26.5	18.6	28.9	30.0	61.7
UMFusion	IJCAI'22	29.9	52.2	58.4	23.5	26.1	42.8	38.8	64.9	31.9	41.3	26.8	19.3	30.2	39.9	61.4
DeFusion	ECCV'22	30.3	52.3	59.3	**26.1**	25.7	45.5	39.9	67.8	34.6	41.5	25.4	19.4	30.8	30.3	61.1
ReCoNet	ECCV'22	28.6	**52.4**	58.1	25.8	24.1	43.6	38.8	67.3	34.2	42.0	24.8	18.5	31.2	30.1	61.4
U2Fusion	TPAMI'22	29.3	51.3	58.7	25.7	26.1	44.4	39.3	65.9	31.5	41.7	26.5	20.1	30.4	30.0	61.3
TarDAL	CVPR'22	29.9	52.4	**60.1**	23.2	**28.1**	**45.6**	**39.9**	**67.9**	34.8	**42.9**	27.6	**20.4**	**32.1**	**31.6**	**61.7**
Ours	Proposed	**31.5**	**53.6**	**60.6**	**26.4**	24.6	**46.1**	**40.5**	**69.6**	**35.3**	**43.7**	**27.4**	**20.8**	**31.6**	**31.8**	**62.4**

Table 1: Quantitative results of object detection. **Red**: best; **blue**: 2nd best.

Figure 2: Qualitative demonstrations of different approaches on the M^3FD dataset.

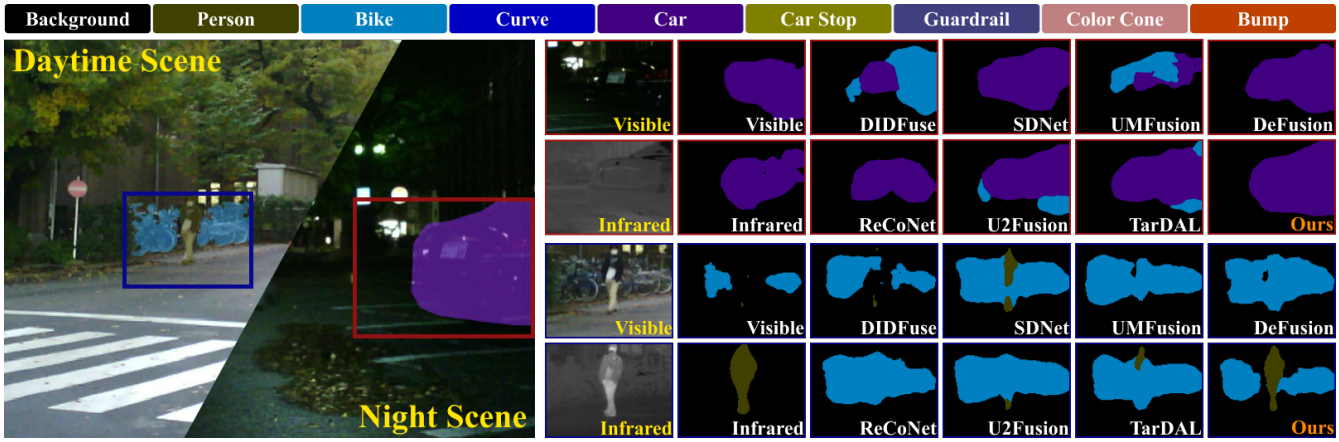

Figure 3: Qualitative demonstrations of different approaches on the MFNet dataset.

daytime scene), demonstrating that our result is capable of making full use of infrared information.

Quantitative segmentation comparisons. Table 2 presents the quantitative results of various fusion methods on the MFNet dataset, measured by Intersection-over-Union (IoU) and Accuracy (ACC). The proposed method achieves the highest numerical performance in multiple categories with the fusion loss obtained from search. Furthermore, fusion schemes with pleasant visual effects struggle to produce excellent segmentation performance, which also validates the effectiveness of the proposed framework that compactly integrates the two tasks.

Method	Source	Unlabel		Car		Person		Bike		Curve		Guardrail		Cone		Bump		mAcc	mIoU
		Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
Infrared	-	98.7	96.3	74.8	67.0	65.6	56.1	53.5	44.2	38.2	30.2	7.2	1.3	23.8	18.9	**58.3**	**43.9**	49.7	41.3
Visible	-	**99.4**	97.2	86.1	**81.0**	53.2	46.6	64.1	56.4	28.5	22.9	**33.9**	**7.4**	40.1	36.8	31.7	30.5	50.5	43.9
DIDFuse	IJCAI'20	**99.3**	97.2	85.1	79.0	66.5	57.1	65.4	55.3	23.4	19.9	1.6	0.8	38.3	34.6	25.8	25.0	48.3	43.7
SDNet	IJCV'21	99.2	**97.5**	**88.7**	80.9	**75.0**	**62.8**	65.0	57.5	35.3	28.9	17.0	3.3	**46.0**	**40.7**	18.6	18.3	51.6	45.3
UMFusion	IJCAI'22	99.1	97.3	**88.3**	80.6	69.6	58.4	66.2	56.4	24.1	20.2	0.4	0.1	43.3	37.9	40.3	35.8	50.6	45.3
DeFusion	ECCV'22	99.2	**97.4**	87.8	80.5	71.2	59.5	64.3	56.1	**41.5**	**32.8**	13.0	3.0	44.3	36.7	33.9	31.4	**52.6**	**45.9**
ReCoNet	ECCV'22	**99.4**	97.2	84.5	78.6	62.2	54.5	63.5	55.3	17.6	15.8	18.4	**7.5**	40.0	36.5	30.7	29.4	48.8	43.9
U2Fusion	TPAMI'22	**99.3**	97.3	84.6	79.2	65.9	56.0	**67.8**	**57.8**	26.1	22.5	0.6	0.2	44.9	39.6	**45.2**	39.3	50.1	45.2
TarDAL	CVPR'22	99.1	**97.4**	**88.3**	80.4	**72.0**	60.5	**68.8**	57.1	38.2	29.8	0	0	40.6	37.1	29.0	28.4	50.6	45.3
Ours	Proposed	99.2	**97.5**	86.1	**81.2**	**75.0**	**62.5**	67.4	**57.9**	**44.0**	**32.5**	**23.1**	4.1	**46.4**	**40.6**	41.5	**39.7**	**55.6**	**47.9**

Table 2: Quantitative results of semantic segmentation. **Red**: best; **blue**: 2nd best.

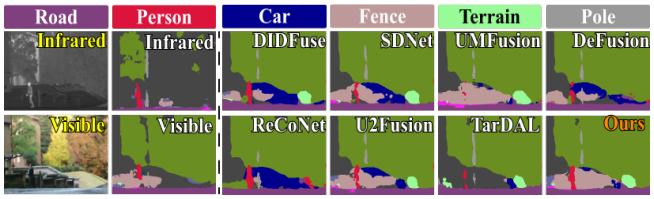

Figure 4: Qualitative demonstrations of different approaches from pre-trained Deeplab-V3+.

Figure 5: Results of different application tasks set on CAF.

Validation through a pretained basedline. Besides the fine-tuned model, Deeplab-V3+ [Chen *et al.*, 2018] trained on the Cityscapes dataset [Cordts *et al.*, 2016] is also leveraged to measure various approaches. Source and fused images from MFNet are directly input into Deeplabv3+, respectively, and their visualized results are reported in Figure 4. In the complex lower region, our method nearly flawlessly segments the terrain, people, and the black fence. The results demonstrate that the loss function searched from the proposed framework are task-informative and accommodate various scenarios of the perception task.

4.4 Analysis of Image Fusion

Influence of different tasks. We first analyze the impact of different tasks on the search results when they are integrated into the framework. Figure 5 shows the final search results, with the visualization on the left and main parts of the final searched loss function on the right. It is quite apparent that different tasks generate different interests in the components within the search space: compared to the segmentation task, the detection loss is more abstract, so it pays more attention to the loss at semantic feature level. In contrast, the segmentation task corresponds to the pixel level and prefers to capture the most interesting parts of the image through constraints on the maximum pixel of the source image. Reflected in the visualization effect, the result guided by segmentation tends to be closer to the visible image (as shown in the green boxes), with lower contrast and limited preservation of scene details; whereas, the result guided by detection incorporates more infrared features, and the content within the salient pedestrians have richer contours (as shown in the red box).

Qualitative comparisons. Figure 6 presents visual comparison of various methods. In the smoggy scene, our approach perfectly retains the rich details from the visible image (leaves shown in the red box) and also preserves non-prominent information from the infrared image (the bushes on the right side of the person in green box). In the tunnel scene, we retain information from darker areas that is closest to the intensity of infrared (the area in the red box), a feat that most other methods struggle to achieve; moreover, we effectively preserve the visible content at the exit. All in all, our method achieves a fusion effect that surpasses the current state-of-the-art in terms of visual appeal.

Quantitative comparisons. Besides, we make comparisons on three datasets using four objective metrics, including Mutual Information (MI) [Qu *et al.*, 2002], Standard Deviation (SD), Visual Information Fidelity (VIF) [Han *et al.*, 2013], and Sum of Correlation Differences (SCD) [Aslantas and Bendes, 2015]. As shown in Figure 7, our results demonstrate consistent advantages in these statistical metrics. Specifically, higher MI and SCD indicate that we are able to significantly preserve the correlated information between source images. Excellent SD and VIF suggest that our results possess higher contrast and are in line with visual perception.

4.5 Analysis of the Proposed Framework

Model-irrelevant generality. In essence, CAF represents a generalized learning paradigm that can potentially be directly applied to existing works. It effectively introduce the responses/guidance from percetion tasks by the bi-level optimization, enabling the adaptive search for loss functions that are tailored to the current network. This approach enhances

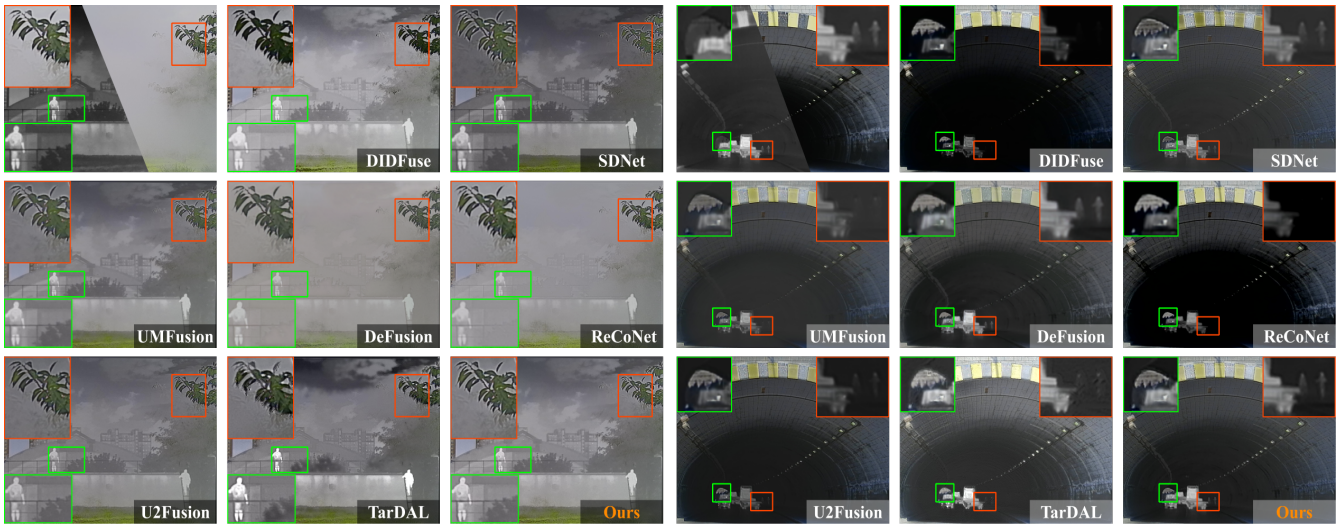

Figure 6: Qualitative demonstrations of different approaches on the M^3FD dataset.

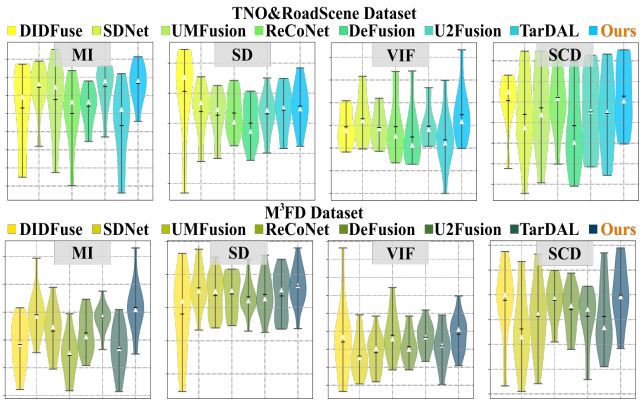

Figure 7: Qualitative demonstrations of different approaches, where the triangles represent the mean, while the lines indicate the median.

Model	Fusion		Detection		Segmentation		
	MI	VIF	People	Car	Person	Car	mIoU
UMFusion	1.263	0.881	29.4	51.7	58.4	80.6	45.3
UMFusion$_{\mathrm{P}}$	1.242	0.923	30.8	52.1	59.1	80.7	47.5
ReCoNet	1.089	0.978	28.3	52.0	54.5	78.6	43.9
ReCoNet$_{\mathrm{P}}$	1.275	0.955	29.9	52.5	56.6	79.2	45.2
U2Fusion	1.139	0.918	29.8	51.7	57.9	80.7	46.6
U2Fusion$_{\mathrm{P}}$	1.315	0.931	30.7	52.9	58.7	81.0	47.0
Ours$_{\mathrm{H}}$	1.244	0.933	26.3	51.6	56.3	77.2	43.9
Ours$_{\mathrm{R}}$	1.075	0.815	22.4	50.5	52.1	72.9	42.0
Ours$_{\mathrm{A}}$	1.091	0.867	25.8	51.1	53.4	75.5	42.8
Ours	1.432	0.983	31.5	53.6	62.5	81.2	47.9

Table 3: Evaluating of generality and effectiveness of CAF on M^3FD & MFNet datasets.

the understanding of model characteristics at the level of loss functions, facilitating better alignment with downstream perception tasks. In this regard, we explore the application of our CAF using representative works including U2Fusion, ReCoNet, and UMFusion as the fusion network $\mathcal{F}$.

Table 3 presents the quantitative results of fusion visual metrics and task-specific metrics before and after (with $_{\mathrm{P}}$) employing our CAF for adaptive search. They all achieved consistent improvements, both in terms of fusion quality and performance on perception tasks, which highlight the flexibility and strong model-irrelevant generality of our framework.

Effectiveness of automatic loss function search. To verify the effectiveness of the proposed framework, we constructed several variants that discard the search process ($_{\mathrm{H}}/_{\mathrm{R}}/_{\mathrm{A}}$ represent handcrafting/random/average initialization of β). The last three rows of Table 3 present the related results. Specifically, the manually designed variant introducing prior knowledge exhibited some effectiveness than policy-free initialization. It is evident that results of CAF are far superior to any other variants in both fusion and task performance.

5 Conclusion and Remark

In this paper, we established a compact, automatic, and flexible framework for infrared and visible image fusion and applications. We not only conducted an in-depth exploration to demonstrate the exceptional characteristics of CAF, but also performed comprehensive experiments to affirm our superior performance in both image fusion and its follow-up high-level tasks.

Broader impacts. From the task perspective, CAF provides a solution paradigm to better understand the relationship between fusion and subsequent tasks. In terms of method design, as a highly universal learning search framework, CAF is capable of being inspiringly applied to existing methods and providing reference and guidance for the manual design of loss functions.

Acknowledgments

This work is partially supported by the National Key R&D Program of China (No. 2022YFA1004101), the China Postdoctoral Science Foundation (2023M730741), the National Natural Science Foundation of China (Nos. U22B2052, 62302078, 61936002) and the Liaoning Revitalization Talents Program (No. 2022RG04).

References

[Aslantas and Bendes, 2015] V Aslantas and Emre Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-international Journal of electronics and communications*, 69(12):1890–1896, 2015.

[Cao *et al.*, 2023] Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23555–23564, 2023.

[Chen *et al.*, 2018] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[Ha *et al.*, 2017] 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 *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.

[Han *et al.*, 2013] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information fusion*, 14(2):127–135, 2013.

[Han *et al.*, 2022] Dong Han, Liang Li, Xiaojie Guo, and Jiayi Ma. Multi-exposure image fusion via deep perceptual enhancement. *Information Fusion*, 79:248–262, 2022.

[Huang *et al.*, 2022] 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.

[Ledig *et al.*, 2017] 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.

[Li and Wu, 2018] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.

[Liang *et al.*, 2022] 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.

[Liu *et al.*, 2019] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. 2019.

[Liu *et al.*, 2021a] 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.

[Liu *et al.*, 2021b] Jinyuan Liu, Yuhui Wu, Zhanbo Huang, Risheng Liu, and Xin Fan. Smoa: Searching a modality-oriented architecture for infrared and visible image fusion. *IEEE Signal Processing Letters*, 28:1818–1822, 2021.

[Liu *et al.*, 2021c] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10045–10067, 2021.

[Liu *et al.*, 2021d] Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34:8662–8675, 2021.

[Liu *et al.*, 2022a] 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, 2022.

[Liu *et al.*, 2022b] Jinyuan Liu, Yuhui Wu, Guanyao Wu, Risheng Liu, and Xin Fan. Learn to search a lightweight architecture for target-aware infrared and visible image fusion. *IEEE Signal Processing Letters*, 29:1614–1618, 2022.

[Liu *et al.*, 2023a] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, pages 1–28, 2023.

[Liu *et al.*, 2023b] 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, 2023.

[Liu *et al.*, 2023c] Jinyuan Liu, Guanyao Wu, Junsheng Luan, Zhiying Jiang, Risheng Liu, and Xin Fan. Holoco: Holistic and local contrastive learning network for multi-exposure image fusion. *Information Fusion*, 95:237–249, 2023.

[Liu *et al.*, 2023d] Zhu Liu, Jinyuan Liu, Guanyao Wu, Long Ma, Xin Fan, and Risheng Liu. Bi-level dynamic learning for jointly multi-modality image fusion and beyond. *International Joint Conference on Artificial Intelligence*, 2023.

[Ma *et al.*, 2019a] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019.

[Ma *et al.*, 2019b] 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.

[Ma *et al.*, 2020] Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2020.

[Ma *et al.*, 2021] Jiayi Ma, Linfeng Tang, Meilong Xu, Hao Zhang, and Guobao Xiao. Stdfusionnet: An infrared and visible image fusion network based on salient target detection. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021.

[Qu *et al.*, 2002] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics letters*, 38(7):1, 2002.

[Sun *et al.*, 2022] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4003–4011, 2022.

[Takumi *et al.*, 2017] Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 35–43, 2017.

[Tang *et al.*, 2022] 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.

[Toet, 2017] Alexander Toet. The tno multiband image data collection. *Data in brief*, 15:249–251, 2017.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2022] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *International Joint Conference on Artificial Intelligence*, 2022.

[Wu *et al.*, 2024] Guanyao Wu, Hongming Fu, Jinyuan Liu, Long Ma, Xin Fan, and Risheng Liu. Hybrid-supervised dual-search: Leveraging automatic learning for loss-free multi-exposure image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5985–5993, 2024.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[Xu *et al.*, 2020a] 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, 2020.

[Xu *et al.*, 2020b] 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*, volume 34, pages 12484–12491, 2020.

[Zhang and Demiris, 2023] Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Zhang and Ma, 2021] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129:2761–2785, 2021.

[Zhao *et al.*, 2020] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *International Joint Conference on Artificial Intelligence*, 2020.

[Zhao *et al.*, 2023a] 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.

[Zhao *et al.*, 2023b] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8082–8093, 2023.