Multi-interactive Feature Learning and a Full-time Multi-modality Benchmark for Image Fusion and Segmentation

Jinyuan Liu[†], Zhu Liu[‡], Guanyao Wu[‡], Long Ma[‡], Risheng Liu[‡§], Wei Zhong[‡], Zhongxuan Luo[‡], Xin Fan[‡*]
[†]School of Mechanical Engineering, Dalian University of Technology
[‡]International School of Information Science Engineering, Dalian University of Technology
[§]Peng Cheng Laboratory
atlantis918@hotmail.com, liuzhu@mail.dlut.edu.cn, xin.fan@dlut.edu.cn

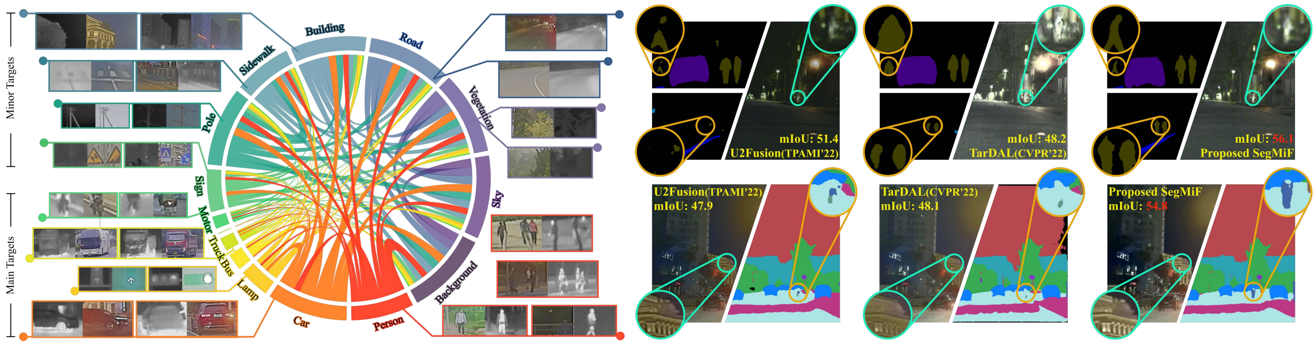


Figure 1: The chord diagram on the left shows the association and relative number of various labels in the proposed FMB dataset. Its branches display part of the well registered targets, which are extraordinarily affluent and manifold. The zoomed-in regions on the right show the segmentation comparison in cyan circles and fusion comparison in orange circles. It is obvious that the proposed method is superior to SOTA methods on both visual effects and mIoU.

Abstract

Multi-modality image fusion and segmentation play a vital role in autonomous driving and robotic operation. Early efforts focus on boosting the performance for only one task, e.g., *fusion or segmentation, making it hard to reach 'Best of Both Worlds'. To overcome this issue, in this paper, we propose a* ***M**ulti-**i**nteractive **F**eature learning architecture for image fusion and **Seg**mentation, namely SegMiF, and exploit dual-task correlation to promote the performance of both tasks. The SegMiF is of a cascade structure, containing a fusion sub-network and a commonly used segmentation sub-network. By slickly bridging intermediate features between two components, the knowledge learned from the segmentation task can effectively assist the fusion task. Also, the benefited fusion network supports the segmentation one to perform more pretentiously. Besides, a hierarchical interactive attention block is established to ensure fine-grained mapping of all the vital information between two tasks, so that the modality/semantic features can be fully mutual-interactive. In addition, a dynamic weight factor is introduced to automatically adjust the corresponding weights of each task, which can balance the interactive feature correspondence and break through the limitation of laborious tuning. Furthermore, we construct a smart multi-wave binocular imaging system and collect a full-time multi-modality benchmark with 15 annotated pixel-level categories for image fusion and segmentation. Extensive experiments on several public datasets and our benchmark demonstrate that the proposed method outputs visually appealing fused images and perform averagely* 7.66% *higher segmentation mIoU in the real-world scene than the state-of-the-art approaches. The source code and benchmark are available at* `https://github.com/JinyuanLiu-CV/SegMiF`.

1. Introduction

Accurate and robust scene parsing[1, 2] is a fundamental technology for autonomous driving. However, in complex environments, *e.g.,* inclement weather, only using visible sensors may fail to accurately recognize targets. On the contrary, infrared sensors are free from the aforementioned issues but limited in low spatial resolution. Consequently, fusing the infrared and visible image [3, 4, 5, 6, 7] has become a mainstream solution for better scene understanding.

Multi-modality fusion for scene parsing needs to provide: (i). *robust visual appealing image*: they require continually generating high-quality images in dynamic scenes. (ii). *accurate semantic segmentation*: they demand to assign category labels to each pixel. Towards these goals, jointly solving multi-modality image fusion and segmentation becomes an urgent issue.

Numerous learning-based multi-modality image fusion methods have been fast development [8, 9, 10, 11, 12]. However, most of them concentrate on developing various networks for generating visual-appealing images rather than considering the follow-up high-level vision tasks, posing an obstacle to better scene parsing. Recently, few studies[13, 14, 15, 16] have attempted to design multi-task learning-based loss functions by cascading the fusion network and high-level tasks. Unfortunately, seeking unified appropriate features for either task simultaneously is still a tough issue.

Moreover, exploring multi-modality fusion and segmentation demands a comprehensive collection of well-alignment image pairs with pixel-level annotated labels. Also, as for one image, the annotated needs to cover a wide range of pixels. Unfortunately, existing multi-modality data collections either focus on image fusion or lack whole image annotated segmentation labels, placing an obstacle to exploring the correlation of the fusion and segmentation.

This paper proposes a multi-interactive feature learning architecture for the joint problem of multi-modality fusion and segmentation, namely SegMiF. SegMiF is constructed by a fusion network and a segmentation network, in which the intrinsic features of either one interact via a new proposed hierarchical interactive attention (HIA). HIA fully integrates semantic-/modality-oriented features by fine-grained mapping. We also derive a dynamic weighting factor and seamless it in the interactive training scheme, to automatically learn the optimal parameters for either task. Figure 1 demonstrates that our SegMiF assigns the category to each pixel from the visual-friendly fused result more accurately than the state-of-the-arts (SOTAs). Our contributions can be distilled into four main aspects as follows:

- We formulate both image fusion and segmentation in a joint manner, in which the semantic and pixel-based features can mutually interact. To this end, two tasks can achieve the 'Best of Both Worlds', generating visual-appealing fused images along with accurate scene parsing.
- A hierarchical interactive attention is introduced to bridge the feature gap between the fusion network and the segmentation one. Establishing the semantic/modality multi-head attention mechanism in HIA simultaneously preserves intrinsic modality features and brings more attention to semantic features.
- An interactive feature training scheme is proposed to overcome the shortcoming of insufficient feature interaction between fusion and segmentation. Seamlessly integrating a dynamic weighting factor allows the exploration of the optimal parameters of each task in an automatic manner.
- We construct a smart multi-wave binocular imaging system, and introduce a full-time multi-modality benchmark, namely FMB, to promote the research of both image fusion and segmentation. FMB contains 1500 well-registered infrared and visible image pairs with 15 annotated pixel-level categories (see the left part of Figure 1). Also, it covers a wide range of pixel variations and various severe environments, *e.g.,* dense fog, heavy rain, and low-light condition.

2. Related Works

Multi-modality image fusion In recent years, deep learning based multi-modality image fusion approaches achieved significant progress[11, 12, 17, 18, 19]. Early efforts [8, 9, 10, 20, 21] tend to achieve excellent fusion effects by adjusting the network structure or loss functions. However, a minority pay attention to whether the downstream tasks can be well adapted to fusion. DIDFuse [10] was the first to apply deep decomposition for high and low-frequency features, while AUIF [22] proposed a decomposition and fusion framework based on algorithm unfolding from an optimization perspective. Subsequently, CDDFuse [6] upgraded the decomposition network to a Transformer-CNN dual-stream structure. DDFM [7], for the first time, utilized denoising diffusion models for image fusion tasks. Recently, some methods [23, 13, 24] cascaded fusion and downstream tasks, focusing on improving task performance by achieving oriented fusion. Nevertheless, this kind of native gradient back propagation hinders the fusion network to adapt to subsequent tasks heuristicly from the feature level.

Multi-modality Segmentation Recently, two-stream-based feature fusion models are proposed to perform the segmentation directly. Most of existing methods mostly develop various simple fusion strategies, such as the weighted average [25, 26, 27], summation [28, 29] and concatenation [30, 31]. Nonetheless, direct feature fusion lacks explicit fusion principle to preserve the typical modality feature and pay no attention to the pixel-level visual effects.

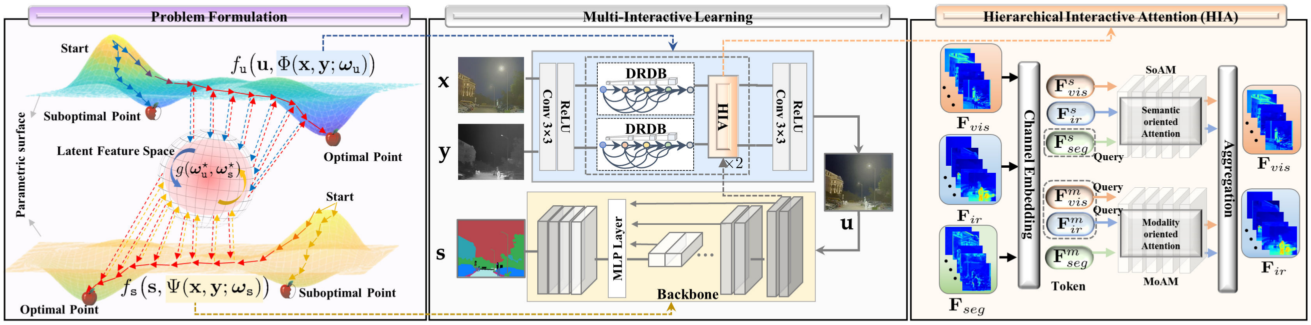


Figure 2: Workflow of our proposed SegMiF. The left part depicts the latent interactive relationship between image fusion and segmentation. The middle part plots the concrete architecture of the SegMiF. The right part details the components of proposed hierarchical interactive attention.

3. The Proposed Method

3.1. Problem formulation

As for image fusion or segmentation tasks, one of the most commonly used ways is to design a neural network, and fully utilize it to find a set of optimal parameters. For this purpose, we suppose that the visible, infrared, and fused image are all gray-scale with the size of $m \times n$, denoted as column vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{u} \in \mathbb{R}^{mn \times 1}$, respectively. The optimization model is formulated as:

$$\min_{\boldsymbol{\omega}_{\mathrm{k}}} f\left(\mathbf{k}, \mathcal{N}(\mathbf{x}, \mathbf{y}; \boldsymbol{\omega}_{\mathrm{k}})\right), \tag{1}$$

where $\mathbf{k}$ denotes the output of the task-related network $\mathcal{N}$ with the learnable parameters $\boldsymbol{\omega}_{\mathrm{k}}$. $f(\cdot)$ is a fidelity term.

Previous approaches solely design image fusion or segmentation networks, which only can achieve outstanding results for one task. To generate visual-appealing fused images along with accurate scene segmentation results, we jointly formulate the two tasks into one goal [32, 33, 34], which can be rewritten as:

$$\min_{\boldsymbol{\omega}_{\mathrm{u}}, \boldsymbol{\omega}_{\mathrm{s}}} f_{\mathrm{u}}\left(\mathbf{u}, \Phi(\mathbf{x}, \mathbf{y}; \boldsymbol{\omega}_{\mathrm{u}})\right) + f_{\mathrm{s}}\left(\mathbf{s}, \Psi(\mathbf{x}, \mathbf{y}; \boldsymbol{\omega}_{\mathrm{s}})\right) + g(\boldsymbol{\omega}^{\star}), \tag{2}$$

where $\boldsymbol{\omega}^{\star} = [\boldsymbol{\omega}_{\mathrm{u}}, \boldsymbol{\omega}_{\mathrm{s}}]$. $\mathbf{u}$ and $\mathbf{s}$ denote the fused image and segmentation map, which are produced by the fusion network Φ and segmentation network Ψ with the learnable parameters $\boldsymbol{\omega}_{\mathrm{u}}$ and $\boldsymbol{\omega}_{\mathrm{s}}$. $g\left(\cdot\right)$ is a constrained term to joint optimize the two tasks. In this paper, we regard the $g\left(\cdot\right)$ as a feature learning constrained manner, and achieve this goal by designing a hierarchical attention along with the interactive training scheme. The visualized illustration is plotted in the left part of Figure 2.

3.2. Feature interaction architecture

Overview of the whole network. Our proposed SegMiF is designed with cascade principle, composited by image fusion and segmentation sub-network. Details of the whole architecture is shown at Figure 2. In specific, we utilize two parallel dilated residual dense blocks (DRDB) [35, 36] to extract features from visual and infrared images. SegFormer [37] is leveraged as the baseline segmentation network to provide semantic parsing. Two scales of semantic features from the backbone, interpolated with original resolutions are embedded into fusion network. In order to sufficiently realize the semantic information sharing, we propose the hierarchical interactive attention (HIA) to transfer high-level knowledge.

Hierarchical interactive attention. After obtaining modality feature $\mathbf{F}_{ir}, \mathbf{F}_{vis}$ from fusion network and segmentation feature $\mathbf{F}_{seg}$, we build the HIA to construct the fine-grained mapping of these features and strengthen the mutually beneficial representation. Features including $\mathbf{F}_{ir}, \mathbf{F}_{vis}$ and $\mathbf{F}_{seg}$ as inputs, two attention mechanisms are leveraged to globally exchange intermediate features. Concatenating features from attentions, fresh modality features are generated based on a residual connection.

In detail, channel embedding is utilize to decompose modality/semantic features. We can denote the outputs from the linear embedding as $\{\mathbf{F}^{s}_{x}, \mathbf{F}^{m}_{x}\}$ with size $\mathbb{R}^{mn \times C}$, where $x \in \{ir, vis, seg\}$ is under a vector formulation. Instead of utilizing the original self-attention mechanisms directly, we bridge these features with complementary interaction from different representation subspaces by MoAM and SoAM.

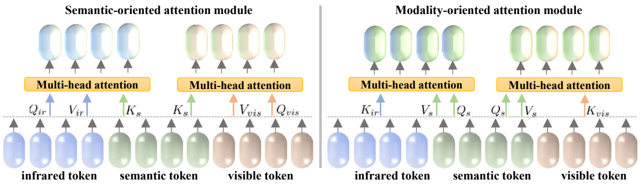


Figure 3: Detailed architectures of SoAM and MoAM.

Both multi-head attention modules are plotted in Figure 3. Semantic-oriented attention module targets to provide more semantic attention for the modality feature. SoAM utilizes the token $\mathbf{F}^{s}_{seg}$ to generate the Query Q_s, which represents the inhere semantic information that needs to be enhanced. The global context representation of each can be calculated by as ${K_{ir}}^{T}V_{ir}$ and ${K_{vis}}^{T}V_{vis}$, where the corresponding Key and Value of each head are from the modality tokens $\{\mathbf{F}^{m}_{ir}, \mathbf{F}^{m}_{vis}\}$. Denoted the modality context representation as $\mathbf{G}_{ir}$, $\mathbf{G}_{vis}$, we can calculate the attention as $S_{ir} = Q_s\mathbf{G}_{ir}$ and $S_{vis} = Q_s\mathbf{G}_{vis}$. On the other hand, MoAM is with the complementary principle to investigate the significant feature from semantic contexts. Specifically, MoAM is to introduce two modality queries Q_{ir}, Q_{vis} to represent the intrinsic modality feature (*e.g.,* targets and details) from $\{\mathbf{F}^{m}_{ir}, \mathbf{F}^{m}_{vis}\}$. Similarly, we can obtain the global context of segmentation $\mathbf{G}_s$ by ${K_s}^{T}V_s$. The cross attention can be formulated as $M_{vis} = Q_s\mathbf{G}_{vis}$ and $M_{vis} = Q_s\mathbf{G}_{vis}$. By concatenating the groups $\{S_{vis}, M_{vis}\}$ and $\{S_{ir}, M_{ir}\}$, we can obtain the comprehensive features with two parallel MLPs with residual connection to aggregate features from SoAM and MoAM.

3.3. Loss function

The total loss function is combined of an image fusion loss function $\mathcal{L}_{\mathtt{f}}$ and segmentation loss $\mathcal{L}_{\mathtt{s}}$. $\mathcal{L}_{\mathtt{f}}$ consists of three types of losses, *i.e.,* structure loss $\mathcal{L}_{\mathtt{SSIM}}$, pixel loss $\mathcal{L}_{\mathtt{MSE}}$ and gradient loss $\mathcal{L}_{\mathtt{grad}}$. For one fused image, it should preserve overall structures from source images. To this end, the structural similarity index (SSIM) [38, 39, 40, 41] is introduced in function:

$$\mathcal{L}_{\mathtt{SSIM}} = (1 - \mathtt{SSIM}_{\mathbf{u},\mathbf{x}})/2 + (1 - \mathtt{SSIM}_{\mathbf{u},\mathbf{y}})/2, \quad (3)$$

where $\mathcal{L}_{\mathtt{SSIM}}$ denotes structure similarity loss. To maintains the vital intensity in the fused image, we employ the saliency-based pixel loss, it formulated as :

$$\mathcal{L}_{\mathtt{MSE}} = \|\mathbf{u} - \boldsymbol{m_1}\mathbf{x}\|_2^2 + \|\mathbf{u} - \boldsymbol{m_2}\mathbf{y}\|_2^2, \quad (4)$$

where $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ are saliency weight maps calculated by VSM [42]. Besides, gradient information of images always characterizes texture details, thus, we used $\mathcal{L}_{\mathtt{grad}}$ to constrain these textual factors to a multi-scale manner:

$$\mathcal{L}_{\mathtt{grad}} = \sum_{k=3,5,7} \|\nabla^k\mathbf{u} - \mathbf{max}(\nabla^k\mathbf{x}, \nabla^k\mathbf{y})\|_2^2 \quad (5)$$

where ∇ denotes gradient operators that calculate by $\nabla = \mathbf{u} - \mathcal{G}(\mathbf{u})$ with combination of different Gauss ($\mathcal{G}$) kernel size k. Totally, we obtained $\mathcal{L}_{\mathtt{f}} = \mathcal{L}_{\mathtt{SSIM}} + \mathcal{L}_{\mathtt{MSE}} + \eta\mathcal{L}_{\mathtt{grad}}$.

Common to previous works, $\mathcal{L}_{\mathtt{s}}$ is defined as:

$$\mathcal{L}_{\mathtt{s}}(\mathbf{s}, \mathbf{s}^*) = -\sum_{class} \mathbf{s}^* \log(\mathbf{s}), \quad (6)$$

where $\mathbf{s}^*$ represents the segmentation label. We adopt the effective semantic segmentation method SegFormer network Ψ [37]. The total loss function is:

$$\mathcal{L}_{\mathtt{total}} = \lambda_1\mathcal{L}_{\mathtt{f}} + \lambda_2\mathcal{L}_{\mathtt{s}}, \quad (7)$$

where λ_1 and λ_2 are dynamic weighting factor which will be discussed below.

3.4. Dynamic factor for interactive learning

By exploiting an alternatively recurrent iterations, we can progressively introduce the task-preferred features into the framework optimization. As mentioned above, at the phase of image fusion, we introduce the dynamic weighting factors λ_1 and λ_2 to rapidly measure the importance of task-related losses. We observe that the task-specific balance can be derived from the convergent rate. The intuition is that if the value of losses cannot be further descended, the network may obtain the corresponding optimal weights. If the convergent rate descends fast, we should pay more attention to this task. Learning from the dynamic weight average [43], we further present the factor of task preference $\boldsymbol{\eta}$ to emphasize the primary goal. Denoted r_i as the convergent rate of i-th task with loss $\mathcal{L}_i$, we can compute the rate as:

$$r_{\mathtt{i}}(n-1) = \frac{\mathcal{L}_i(n-1)}{\mathcal{L}_i(n-2)}. \quad (8)$$

Then the procedure of dynamic weight factor of i-th task is formulated as:

$$\lambda_{\mathtt{i}}(n) = \frac{\eta_{\mathtt{i}}\exp(r_{\mathtt{i}}(n-1)/T)}{\sum_k \exp(r_{\mathtt{k}}(n-1)/T)}, \quad (9)$$

where T is a temperature to control the sensitiveness of two tasks. Different from widely sued GDN [44], this dynamic strategy can actually avoid the complicated computation of various task gradients.

Then based on the fusion image generated with semantic feature, we can train the segmentation network end-to-end, using the gradient decending $\boldsymbol{\omega}_s \leftarrow \boldsymbol{\omega}_s - \nabla_{\boldsymbol{\omega}_s}\mathcal{L}_{\mathtt{s}}(\mathbf{u}; \boldsymbol{\omega}_s)$. The two learning processes are trained alternately until full convergence. Noting that, this training strategy is actually task-agnostic, we also can introduce other different high-level vision tasks into the unified consideration, rather than designing for segmentation unilaterally.

4. Full-time Multi-modality Benchmark

Existing two multi-modality segmentation datasets suffer from few label categories, sparse annotation and monotonous scene, as shown in Figure 6. The proposed FMB dataset aims to overcome these difficulties and promote the development of whole field. A glimpse of FMB is given in Figure 4 .

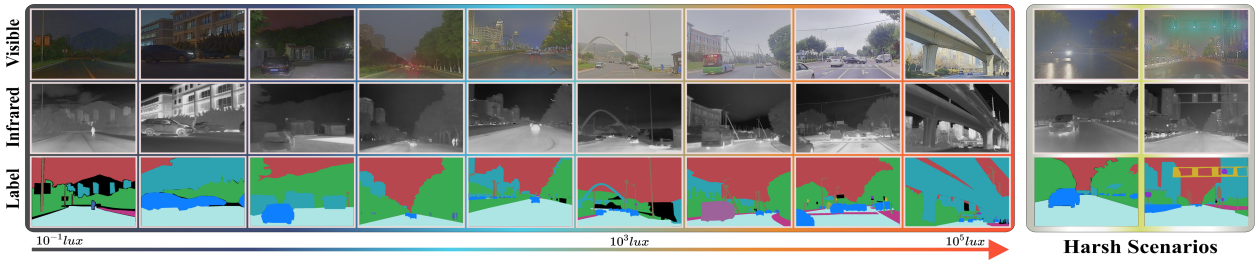


Figure 4: Visualization of visible/infrared/segmentation images in the proposed FMB dataset. The dataset contains a wide range of real driving scenes under different lighting conditions, and also includes special scenarios with rain, fog, strong light, and even Tyndall Effect.

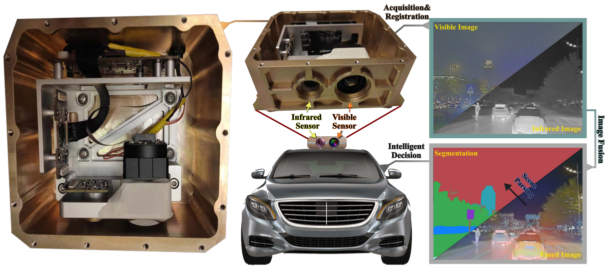


Figure 5: Illustration of the binocular imaging system.

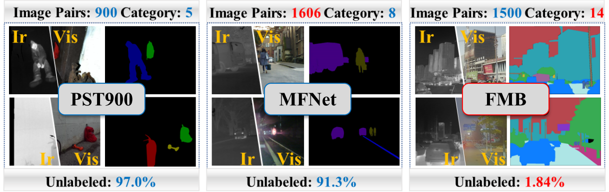


Image Pairs: 900 Category: 5
Image Pairs: 1606 Category: 8
Image Pairs: 1500 Category: 14
Ir
Vis
PST900
MFNet
FMB
Unlabeled: 97.0%
Unlabeled: 91.3%
Unlabeled: 1.84%

Figure 6: Comparison of FMB with existing multi-modality segmentation datasets (*i.e.,* PST900 [31] and MFNet [30]).

We built a binocular imaging system that can be placed on the car roof, including a visible camera and an infrared sensor with a wavelength range of 8-14μm(as shown in the Figure 5). We finally obtained 1500 aligned image pairs with a resolution of 800×600. In details, two sensors are individually calibrated using their respected calibration board to obtain internal parameters, and their relative pose relationship is obtained through joint calibration. We calculate the homography matrix H by employing RANSAC[1]. The infrared images are projected onto visible coordinates using H and cropped, ultimately resulting in pixel-level registered image pairs with a size of 800×600.

The FMB dataset includes rich scenes under different illumination conditions, so that it enables fusion/segmentation model to improve the generalization ability greatly. We labeled 98.16% of all pixels into 14 different categories including *Road, Sidewalk, Building, Traffic Light, Traffic Sign, Vegetation, Sky, Person, Car, Truck, Bus, Motorcycle, Bicycle and Pole*, which often appear in real-world automatic driving and semantic understanding tasks.

5. Experiments

Two representative datasets including MFNet and proposed FMB are utilized for the training and evaluation. The details of these datasets are reported in the above section. Several data augmentation techniques are utilized for the whole training procedure: random resizing with a ratio of 0.5-2.0, random cropping to 360×360, brightness distortion and normalization. The Adam optimizer with poly learning rate adjustment is utilized to optimize both networks. As for the fusion, the initial learning rate is $1e^{-4}$ and decayed to $1e^{-8}$ progressively. As for segmentation, we first utilized $1e^{-6}$ to warm start the training with 3k iterations, then we conducted the training with intilial learning rate $8e^{-5}$. With batch size of 8, we trained the framework for 8 rounds. For each round, we set 10k iterations for training segmentation and 5k iterations for fusion. Noting that, a similar configuration of segmentation is also utilized for the training of fusion-based methods. All experiments are performed on an NVIDIA Tesla V100 GPU with PyTorch framework.

5.1. Results of multi-modality image fusion

We demonstrate our fusion quality based on qualitative and quantitative analyses with six state-of-the-art competitors, including DIDFuse [10], DenseFuse [45], ReCoNet[11], UMFusion[12], TarDAL[13] and U2Fusion [9].

Qualitative Comparisons. The qualitative results on MFNet and FMB datasets are depicted in Figure 7, in which we can clearly observe two remarkable advantages of our method. First, the significant characteristics of infrared images can be effectively highlighted. For instance, as shown on the green rectangle of the first group, DIDFuse, ReCoNet and TarDAL are susceptible to strong illumination. In contrast, our method can remarkably preserve this information from infrared images, *e.g.,* the structure of cars and pedestrians. Furthermore, benefiting from the guidance of semantic

[1] Efficient RANSAC for point-cloud shape detection

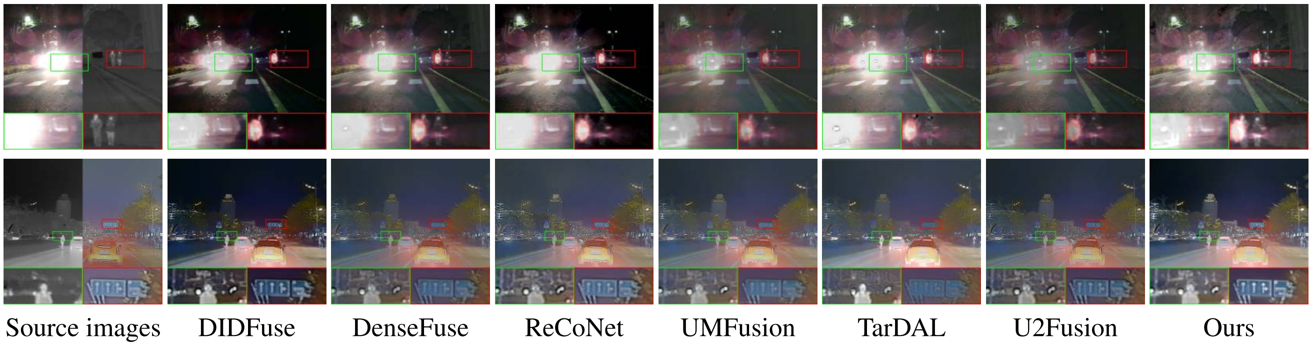


Figure 7: Visual comparison of different fusion approaches on the MFNet and FMB dataset, respectively.

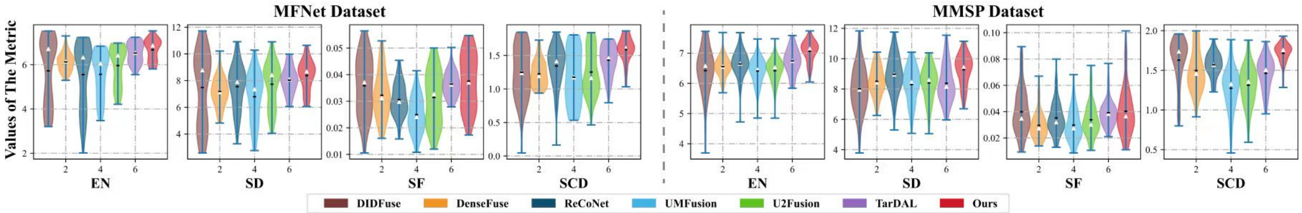


Figure 8: Quantitative comparisons of image fusion with six SOTA methods on two datasets. Violin plots illustrating the distribution of the four metrics, in which the white triangles and the black lines indicate mean values and medium values.

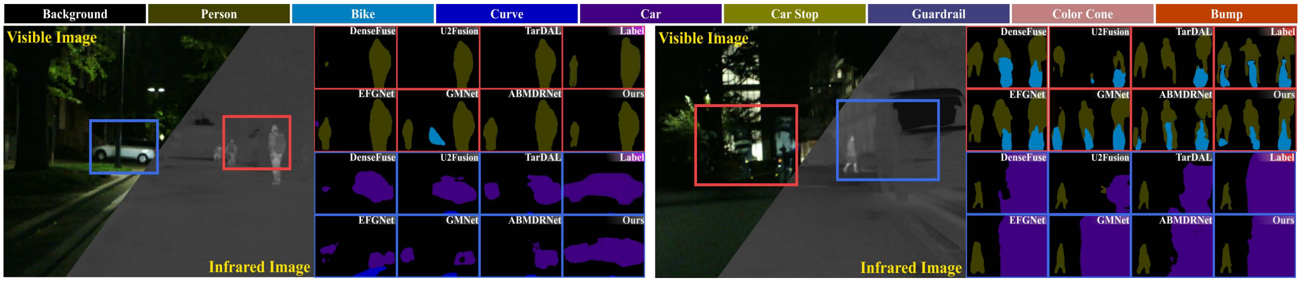


Figure 9: Qualitative demonstrations of different approaches on the MFNet dataset.

information, our method can enhance the texture details of the given scene from either dataset. Compared with other competitors in the second row of Figure 7, our results exhibit a sharper appearance.

Quantitative Comparisons. We also plot the numerical results with other six fusion competitors on 50 pairs from MFNet and 50 pairs from FMB in Figure 8. Four objective metrics are leveraged for the comparison, including entropy (EN)[46], standard deviation (SD)[47], spatial frequency (SF)[48] and the sum of the correlations of differences (SCD)[49]. Note that our results achieve consistent superiority in terms of these statistical metrics. Specifically, the highest EN and SCD indicate that our method can significantly preserve the largest amount of considerable information transferred from source images. Moreover, the immense average value on SD reflects the high pixel contrast for visual observations. Furthermore, higher SF reflects our method has rich texture details and contrasts. In summary, our method enhances the texture details for precise observation and stably preserves abundant typical information to support semantic parsing tasks.

5.2. Results of multi-modality segmentation

We provide another comprehensive analysis for image segmentation. Besides comparing with the newest fusion-based methods, we also conduct the evaluations with competitive dual-stream methods: GMNet [50], FEANet [51], EGFNet [52], ABMDRNet [53] and LASNet [54].[2]

Qualitative Comparisons. Visualized results of segmentation on MFNet are depicted in Figure 9. We also compare various competitive methods in Figure 10 under the newly proposed dataset, which is more challenging with rich categories, complex imaging conditions and complicated scene

[2] We also retrained dual-stream methods on the FMB dataset.

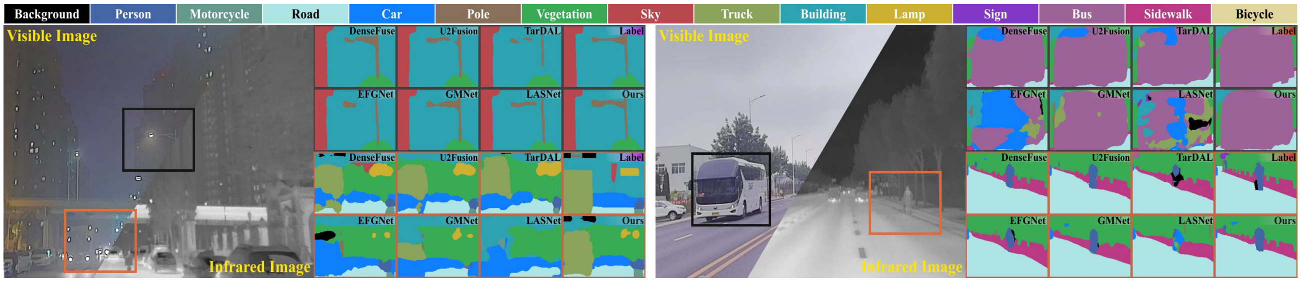


Figure 10: Qualitative demonstrations for the different methods in daytime and nighttime scenarios on FMB benchmark.

Methods	Unlabel		Car		Person		Bike		Curve		Car Stop		Cone		Bump		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
Visible	98.2	97.5	94.6	82.2	81.7	54.7	71.2	61.0	54.9	21.6	61.9	16.6	82.1	43.7	77.3	30.8	69.7	45.9
Infrared	98.6	97.4	90.3	81.8	85.3	67.3	71.6	54.8	51.3	37.3	53.8	24.3	64.7	33.4	70.2	48.0	65.6	49.9
LASNet	99.2	97.4	94.9	84.2	81.7	67.1	82.1	56.9	70.7	41.1	56.8	39.6	58.1	48.8	77.2	40.1	75.4	54.9
EGFNet	99.3	97.7	95.8	87.6	89.0	69.8	80.6	58.8	71.5	42.8	48.7	33.8	65.3	48.3	71.1	47.1	72.7	54.8
FEANet	99.1	97.8	93.9	87.8	82.7	71.1	76.7	61.1	65.5	46.5	26.6	22.1	66.6	55.3	77.3	48.9	73.2	55.3
ABMDR	99.3	98.4	94.3	84.8	90.0	69.6	75.7	60.3	64.0	45.1	44.1	33.1	61.7	47.4	66.2	50.0	69.5	54.8
DIDFuse	98.0	97.3	95.0	79.1	79.5	64.0	80.2	58.5	44.1	19.9	64.0	23.3	77.5	37.8	69.0	20.4	67.5	44.5
ReCoNet	98.1	97.3	95.8	80.4	88.9	60.0	65.0	55.4	47.0	20.7	69.0	25.8	77.8	39.8	46.6	17.4	65.9	44.5
U2Fusion	98.3	97.7	95.2	82.8	85.4	64.8	77.7	61.0	62.7	32.3	66.7	20.9	75.5	45.2	82.3	50.2	71.9	50.8
TarDAL	98.3	97.6	93.5	80.7	86.2	67.1	76.5	60.1	53.8	34.9	55.3	10.5	88.6	38.7	90.6	45.5	71.7	48.6
Ours	98.7	98.1	96.3	87.8	89.6	71.4	81.2	63.2	63.5	47.5	66.7	31.1	85.3	48.9	84.8	50.3	74.8	56.1

Table 1: Quantitative semantic segmentation results of different methods on the MFNet dataset.

details. As discussed above, existing fusion methods cannot highlight the dimness of infrared targets, and the distant pedestrian can not be recognized. As for dual-stream methods, which utilize the modality feature directly, they are easy to introduce conflicts and weaken the accuracy without a clear feature fusion principle. The results, such as the car occluded by barriers (the first column of Figure 9) and the shape of the human (the second column of Figure 10) can not be precisely classified. It is worth mentioning that interaction feature learning from segmentation can drastically transfer the complementary characteristics for image fusion and further improve the segmentation performance. Thus, our method can continuously classify the objects of diverse scenes with high accuracy.

Quantitative Comparisons. Table 1 and Table 2 reported the qualitative results among different categories of competitors on MFNet and FMB datasets. These results illustrate our method is ahead of other state-of-the-art methods by a large margin on both segmentation datasets. Noting that our numerical results outperform other methods concerning mIOU and rank second in terms of mACC. Compared with the second one, our method improves 7.66% and 1.45% of mIOU on FMB and MFNet respectively. More specifically, the classification of Car and Person is important for the current intelligent perception system. The top two results in these two categories indicate the high performance of our method to employ for real-world perception. On the other hand, for thermal-insensitive categories, such as traffic sign, building, bump, due to the effective visual quality preservation and enhancement, our method achieves significant superiority.

5.3. Ablation studies

Study on HIA. HIA plays a key role in preserving intrinsic modality features from the semantic feature guidance. Firstly, we visualized the representative feature to discuss the effectiveness of HIA in Figure 11. Clearly, HIA can remarkably preserve the salient infrared features with abundant semantic information, avoiding the interference of harsh weather and strong light. Then we plot different variants of HIA to illustrate the inner mechanisms of HIA. As shown in Figure 12. Obviously, the version w/o SoAM loses the classify ability to distinguish confused objectives, e.g., the orange circle in the second row. Meanwhile, "w/o MoAM" cannot protect the details at nighttime with color distortion, e.g, the building in the distance. It is worth pointing out that our full model not only provides clear visual observation but also has high sensitiveness to segmentation. Similarly, the quantitative results reported in Table 3 also demonstrate the effectiveness of full HIA for both segmentation benchmarks compared with direct aggregation and other model variants. In brief, HIA is capable enough to

Methods	Car		Person		Truck		T- Lamp		T-Sign		Building		Vegetation		Pole		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
Visible	84.5	78.3	78.1	46.6	73.2	43.4	82.5	23.7	83.6	64.0	89.0	77.8	88.5	82.1	71.6	41.8	73.5	50.5
Infrared	75.6	69.1	89.9	63.3	66.6	12.6	63.1	24.7	80.5	52.9	87.4	78.0	83.7	75.5	62.1	23.0	69.6	43.9
GMNet	87.6	79.3	77.3	60.1	49.0	22.2	54.1	21.6	78.5	69.0	89.1	79.1	88.9	83.8	59.7	39.8	64.4	49.2
LASNet	81.3	72.6	76.4	48.6	29.6	14.8	20.7	2.9	79.3	59.0	86.9	75.4	87.6	81.6	56.6	36.7	56.9	42.5
EGFNet	83.6	77.4	79.5	63.0	33.5	17.1	58.6	25.2	82.6	66.6	88.5	77.2	89.3	83.5	63.8	41.5	63.0	47.3
FEANet	82.3	73.9	78.8	60.7	44.7	32.3	53.6	13.5	73.3	55.6	87.6	79.4	89.0	81.2	66.2	36.8	64.5	46.8
DIDFuse	86.3	77.7	87.4	64.4	66.3	28.8	75.9	29.2	81.1	64.4	87.1	78.4	89.5	82.4	79.2	41.8	73.0	50.6
ReCoNet	83.7	75.9	87.7	65.8	34.7	14.9	83.3	34.7	85.6	66.6	89.0	79.2	88.2	81.3	73.3	44.9	71.4	50.9
U2Fusion	85.0	76.6	87.7	61.9	84.6	14.4	75.1	28.3	81.3	68.9	89.5	78.8	92.5	82.2	74.5	42.2	70.1	47.9
TarDAL	81.8	74.2	93.3	56.0	66.3	18.8	75.0	29.6	81.2	66.5	88.1	79.1	87.9	81.7	65.9	41.9	74.8	48.1
Ours	85.3	78.3	78.3	65.4	74.4	47.3	86.4	43.1	86.1	74.8	90.0	82.0	91.6	85.0	72.5	49.8	74.5	54.8

Table 2: Quantitative semantic segmentation results of different methods on the FMB dataset.

bridge the fusion and segmentation tasks.

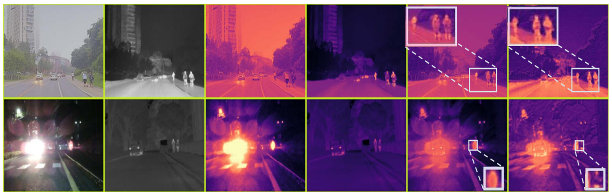
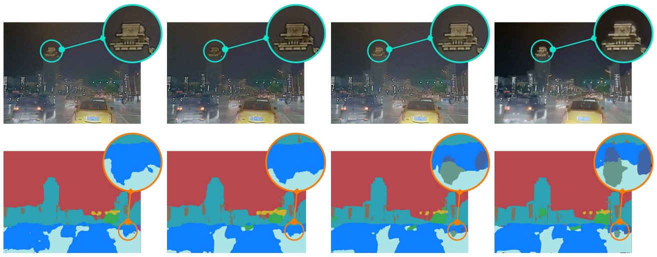


Figure 11: Feature visualization of different stages. From left to right: visible image, infrared image, their features, w/o HIA, and w/ HIA.

Model	HIA		MF Dataset		FMB Dataset	
	SoAM	MoAM	mAcc	mIoU	mAcc	mIoU
"Concatenate"	✗	✗	72.6	52.7	72.3	51.3
"Summation"	✗	✗	71.5	52.6	73.7	52.1
"Average"	✗	✗	72.9	52.1	72.4	50.7
M1	✗	✗	67.2	51.9	72.4	50.5
M2	✗	✔	73.6	55.0	72.7	52.3
M3	✔	✗	72.0	53.4	73.1	54.1
M4	✔	✔	74.8	56.1	74.5	54.8

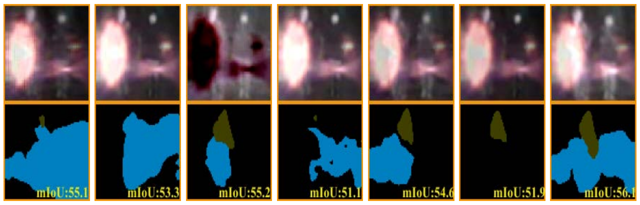
Table 3: Numerical results about the effectiveness of HIA. The first three are the results of direct feature aggregation. Latter are the results of model variants with HIA.

Analyzing the dynamic factor. We discussed the impact of the proposed dynamic factor for interactive learning compared with existing multi-task optimization methods, as shown in Figure 13. Manual adjustment requires plenty of prior knowledge and labor consumption. But concurrently, it achieves decent segmentation results. The other five training strategies hardly coordinate the relationship between the two tasks and fail to achieve good visualization and segmentation performance. The dynamic factor enables to better introduce task preferences into optimization, thus achieving excellent results on both tasks.



w/o HIA w/o SoAM w/o MoAM Full Model

Figure 12: Visual comparisons of different models.



Manual DWA[43] GLS[55] GDN[44] RLM[56] UW[57] Ours

Figure 13: Comparisons of different strategies of adjusting dynamic factors.

6. Conclusion

In this paper, a multi-interactive architecture was proposed to formulate fusion and segmentation in a harmonious manner. We introduced a hierarchical interactive attention with dynamic factors, which bridges gaps of cross-task features from architecture and learning perspectives. In addition, we proposed a comprehensive full-time multi-modality benchmark, with well-registered targets, abundant scenes and affluent labels.

Acknowledgments. This work is partially supported by the National Key R&D Program of China (No. 2022YFA1004101), the National Natural Science Foundation of China (No. U22B2052, 62027826).

References

[1] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE/CVF CVPR*, pages 2881–2890, 2017.

[2] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE/CVF CVPR*, pages 633–641, 2017.

[3] Wujie Zhou, Xinyang Lin, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Mffenet: Multiscale feature fusion and enhancement network for rgb–thermal urban road scene parsing. *IEEE TMM*, 24:2526–2538, 2021.

[4] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb–thermal scene parsing. In *AAAI*, volume 36, pages 3571–3579, 2022.

[5] Zhiying Jiang, Zengxi Zhang, Xin Fan, and Risheng Liu. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *ACM MM*, pages 3783–3791, 2022.

[6] 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. *IEEE/CVF CVPR*, 2023.

[7] 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. *IEEE/CVF ICCV*, 2023.

[8] 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 TCSVT*, 2021.

[9] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE TPAMI*, 2020.

[10] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *IJCAI*, 2020.

[11] 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 *ECCV*, pages 539–555. Springer, 2022.

[12] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *IJCAI*, 2022.

[13] 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 *IEEE/CVF CVPR*, pages 5802–5811, 2022.

[14] 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.

[15] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. *IEEE/CVF CVPR*, pages 5637–5646, 2022.

[16] 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 TPAMI*, 2022.

[17] Jinghao Zhang, Jie Huang, Mingde Yao, Zizheng Yang, Hu Yu, Man Zhou, and Feng Zhao. Ingredient-oriented multi-degradation learning for image restoration. In *IEEE/CVF CVPR*, pages 5825–5835, 2023.

[18] Jie Huang, Feng Zhao, Man Zhou, Jie Xiao, Naishan Zheng, Kaiwen Zheng, and Zhiwei Xiong. Learning sample relationship for exposure correction. In *IEEE/CVF CVPR*, pages 9904–9913, 2023.

[19] 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.

[20] 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. *IJCAI*, 2023.

[21] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *ACM MM*, pages 1600–1608, 2021.

[22] 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 TCSVT*, 2021.

[23] Di Wang, Jinyuan Liu, Risheng Liu, and Xin Fan. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Information Fusion*, 98:101828, 2023.

[24] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE TPAMI*, 2023.

[25] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019.

[26] Qiang Zhang, Tonglin Xiao, Nianchang Huang, Dingwen Zhang, and Jungong Han. Revisiting feature fusion for rgb-t salient object detection. *IEEE TCSVT*, 2020.

[27] Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *arXiv preprint arXiv:2305.15862*, 2023.

[28] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE RAL*, 4(3):2576–2583, 2019.

[29] Man Zhou, Jie Huang, Xueyang Fu, Feng Zhao, and Danfeng Hong. Effective pan-sharpening by multiscale invertible neural network and heterogeneous task distilling. *IEEE TGRS*, 60:1–14, 2022.

[30] 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.

[31] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *ICRA*, pages 9441–9447. IEEE, 2020.

[32] 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 TPAMI*, 2021.

[33] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *ACM MM*, pages 4003–4011, 2022.

[34] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE TCSVT*, 32(10):6700–6713, 2022.

[35] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *IEEE/CVF CVPR*, pages 5687–5697. IEEE, 2022.

[36] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *IEEE/CVF CVPR*, pages 1751–1760, 2019.

[37] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34:12077–12090, 2021.

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

[39] Long Ma, Tianjiao Ma, Xinwei Xue, Xin Fan, Zhongxuan Luo, and Risheng Liu. Practical exposure correction: Great truths are always simple. 2022.

[40] Yuhui Wu, Zhu Liu, Jinyuan Liu, Xin Fan, and Risheng Liu. Breaking free from fusion rule: A fully semantic-driven infrared and visible image fusion. *arXiv preprint arXiv:2211.12286*, 2022.

[41] Risheng Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. Knowledge-driven deep unrolling for robust image layer separation. *IEEE TNNLS*, 2019.

[42] Jinlei Ma, Zhiqiang Zhou, Bo Wang, and Hua Zong. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Physics & Technology*, 82:8–17, 2017.

[43] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *IEEE/CVF CVPR*, pages 1871–1880, 2019.

[44] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, pages 794–803. PMLR, 2018.

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

[46] Wesley J. Roberts, Jan A. Aardt Van, and Fethi Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):1–28, 2008.

[47] V Aslantas and E 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.

[48] Guangmang Cui, Huajun Feng, Zhihai Xu, Li Qi, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341(341):199–209, 2015.

[49] Kaiwen Zheng, Jie Huang, Man Zhou, Danfeng Hong, and Feng Zhao. Deep adaptive pansharpening via uncertainty-aware image fusion. *IEEE TGRS*, 2023.

[50] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE TIP*, 30:7790–7802, 2021.

[51] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *IROS*, pages 4467–4473. IEEE, 2021.

[52] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb thermal scene parsing. *AAAI*, 2022.

[53] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *IEEE/CVF CVPR*, pages 2633–2642, 2021.

[54] Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. Rgb-t semantic segmentation with location, activation, and sharpening. *IEEE TCSVT*, 2022.

[55] Sumanth Chennupati, Ganesh Sistu, Senthil Yogamani, and Samir A Rawashdeh. Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning. In *IEEE/CVF CVPR*, 2019.

[56] Baijiong Lin, Feiyang Ye, and Yu Zhang. A closer look at loss weighting in multi-task learning. *arXiv preprint arXiv:2111.10603*, 2021.

[57] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE/CVF CVPR*, pages 7482–7491, 2018.