MetaFusion: Infrared and Visible Image Fusion via Meta-Feature Embedding from Object Detection

Wenda Zhao[1], Shigeng Xie[1], Fan Zhao[2]*, You He[3], Huchuan Lu[1]
[1] Dalian University of Technology, China
[2] Liaoning Normal University, China
[3] Tsinghua University, China
zhaowenda@dlut.edu.cn; xieshigeng@mail.dlut.edu.cn
Fan_Zhao20@163.com; youhe_nau@163.com; lhchuan@dlut.edu.cn

Abstract

Fusing infrared and visible images can provide more texture details for subsequent object detection task. Conversely, detection task furnishes object semantic information to improve the infrared and visible image fusion. Thus, a joint fusion and detection learning to use their mutual promotion is attracting more attention. However, the feature gap between these two different-level tasks hinders the progress. Addressing this issue, this paper proposes an infrared and visible image fusion via meta-feature embedding from object detection. The core idea is that meta-feature embedding model is designed to generate object semantic features according to fusion network ability, and thus the semantic features are naturally compatible with fusion features. It is optimized by simulating a meta learning. Moreover, we further implement a mutual promotion learning between fusion and detection tasks to improve their performances. Comprehensive experiments on three public datasets demonstrate the effectiveness of our method. Code and model are available at: https://github.com/wdzhao123/MetaFusion.

1. Introduction

Multi-modality sensor technology has promoted the application of multi-modality images in different areas. Among them, infrared images and visible images have been utilized commonly, as the information contained in these two modalities is complementary. Specifically, infrared images can supply object thermal structures without being affected by illumination. But they are short of texture details. On the contrary, visible images can catch the texture information for the scene. But they are severely affected by light. Thus, many methods [15,25,35,43–45,47] focus on studying pixel-level infrared and visible image fusion (IVIF), thereby helping high-level tasks improve performance, e.g., object detection (OD) [20, 34].

IVIF and OD can greatly benefit form each other. IVIF

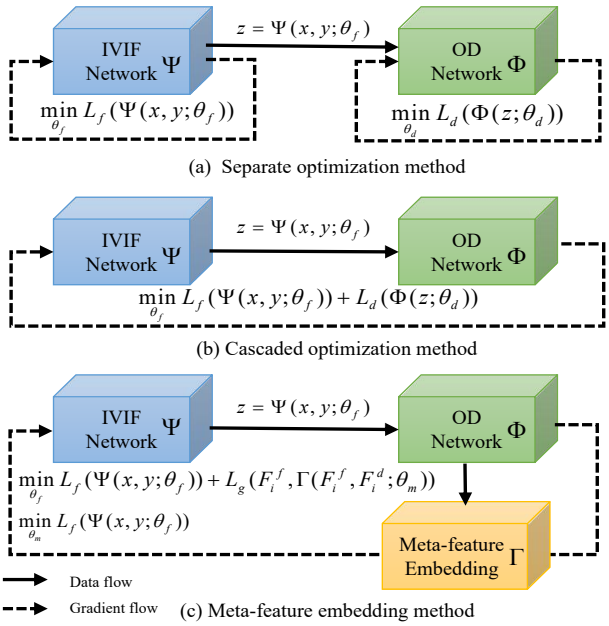

Figure 1. Different joint learning methods of infrared and visible image fusion (IVIF) and object detection (OD). (a) Separate optimization method: IVIF network Ψ is firstly optimized by fusion loss L_f. Then, fusion result z is generated by Ψ from the input infrared and visible image pair x, y. Finally, OD network Φ is optimized by detection loss L_d using z. (b) Cascaded optimization method: OD network Φ is treated as a constraint to optimize IVIF network Ψ by the loss L_f and L_d. (c) Meta-feature embedding method: Meta-feature embedding Γ is optimized to learn how to guide Ψ to have a low L_f. Then, Γ generates meta feature from detection feature F_i^d and fusion feature F_i^d. Finally, the meta feature is used to guide Ψ by the loss L_d.

*Corresponding author

generates the fused image that contains more information than any single modality image to improve OD. However, IVIF mainly focuses on the pixel relationship between an image pair, and there is little consideration for object semantic. In contrast, OD can provide rich object semantic information to IVIF, as its aim is locating the objects. Therefore, this paper studies a joint learning framework between IVIF and OD to improve their performances.

Existing joint learning methods of IVIF and OD can be divided into two categories: separate optimization and cascaded optimization. Separate optimization firstly trains IVIF network, and then trains OD network using IVIF results, as shown in Figure 1(a). Thus, most methods focus on improving the fusion effect, e.g., designing networks [13, 23, 35, 36] and introducing constraints [10, 26, 45]. Obviously, separate optimization neglects the help of OD. Cascaded optimization adopts OD network as a constraint to train IVIF network, and thus forces the IVIF network to generate fusion images with easily detected objects [20], as shown in Figure 1(b). However, directly utilizing the high-level OD constraint to guide the pixel-level IVIF will result in limited effect. Therefore, we leverage OD feature maps that guide IVIF feature maps to obtain more semantic information. Unfortunately, OD features are mismatched with IVIF features due to their task-level difference. Addressing this issue, we propose a meta-feature embedding network (MFE), as shown in Figure 1(c). The idea is that if MFE generates OD features according to the IVIF network ability, the OD features are naturally compatible with the IVIF network, and the optimization can be achieved by simulating a meta learning.

Specifically, an infrared and visible image fusion via meta-feature embedding from object detection is proposed, which is named as MetaFusion. MetaFusion includes IVIF network (F), OD network (D), and MFE. In particular, MFE is expected to generate meta features to bridge the gap between F and D, which is optimized by two alternate steps: inner update and outer update. In the inner update process, we firstly optimize F using meta training set S_{mtr} to obtain its updated network F'. Then, F' calculates the fusion loss on meta testing set S_{mts} to optimize MFE. The motivation is that if MFE successfully generates meta features which are compatible with F, F' will produce better fused images, i.e., the fusion loss should be lower. In the outer update process, F is optimized with the guide of the meta features generated by the fixed MFE on S_{mtr} and S_{mts}. In this way, F can learn how to extract semantic information to improve fusion quality. In the above two alternate steps, D is fixed to offer detection semantic information. Thus, we further implement a mutual promotion learning, where we use the improved F to generate fusion results to finetune D, and then the improved D offers better semantic information to optimize F.

In summary, our contributions are as follows. (1) We explore the joint learning framework of IVIF and OD, and propose MetaFuison to obtain superior performance on these two tasks. (2) Meta-feature embedding network is designed to generate meta features that bridge the gap between F and D. (3) Sequentially, mutual promotion learning between F and D is introduced to improve their performances. (4) Extensive experiments on image fusion and object detection validate the effectiveness of the proposed method.

2. Related Work

Image Fusion. Traditional image fusion methods use hand-crafted features, such as sparse representation [22], spectral variation [53] and low-rank representation [14], which can not handle the complex scenes well. Nowadays, deep learning-based image fusion methods [16, 24, 27, 39, 49, 50, 54, 56] are proposed. Ma $et\ al.$ [23] propose a swin transformer based image fusion method with cross-domain long-range learning. Xu $et\ al.$ [44] use feature extraction and measurement to estimate the degree of information preservation in image fusion. Zhao $et\ al.$ [51] design a self-supervised feature extraction model. Besides, feature diversity enhancement and fusion is received attention [52, 55]. However, most of them neglect the help of high-level tasks. Recently, Liu $et\ al.$ [20] propose a joint learning of IVIF and OD. They treat OD network as an additional constraint to help IVIF network generate fusion result with more clear objects.

However, directly using the high-level OD network to guide pixel-level IVIF may not obtain an effective constraint. Moreover, they use the fusion results as the connection between IVIF network and OD network, ignoring the pixel-level semantic information in OD features. In contrast, we adopt semantic information in OD features to help IVIF, and conduct a meta-feature embedding to generate meta features from OD features. Then, those meta features are used to guide IVIF network to learn pixel-level semantic information.

Object Detection. Deep learning-based object detection methods have made great progress, e.g., network structure-designed methods [2, 8, 30, 31] and loss constraint methods [18, 32, 42]. Carion $et\ al.$ [3] introduce a transformer encoder-decoder architecture for object detection. Xu $et\ al.$ [42] rank positive and negative sample pairs and improve the ranking effect by a clustering algorithm to built object detection loss. On the other hand, many weakly supervised methods [40, 46, 48] are proposed. Wu $et\ al.$ [40] design a model to suppress the background activation and obtain object region, where image-level labels are used. Zhang $et\ al.$ [48] use instance proposal grouping to learn object detection from a single point annotation.

Most of those object detection methods leverage single modality image, e.g., visible images. However, the visible

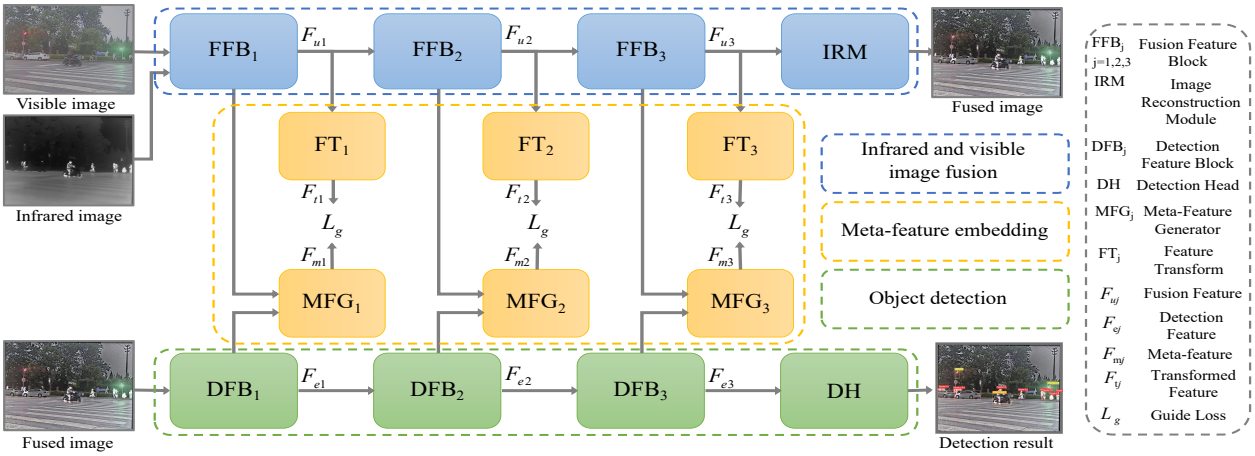

Figure 2. Framework illustration of the proposed MetaFusion. MetaFusion includes three parts: infrared and visible image fusion (IVIF), object detection (OD), and meta-feature embedding (MFE). IFIV network includes four blocks: three feature fusion blocks FFB_1, FFB_2, FFB_3 used to extract and fuse features and one image reconstruction module IRM used to reconstruct fusion result. MFE contains three meta-feature generators MFG_1, MFG_2, MFG_3 conducted to generate muti-level meta features from OD features F_{ej}, and three feature transform networks FT_1, FT_2, FT_3 used to transfer the meta features to fusion features. OD network is divided as three feature extraction blocks DFB_1, DFB_2, DFB_3 and one OD detection head DH.

images are affected by light conditions, thereby influencing the performance of object detection. On the contrary, we leverage IVIF results to provide infrared information without be influenced by illumination, therefore can offer more stable detection performance.

Meta Learning. Meta learning aims to use the learned knowledge to quickly adapt to new tasks, which has been used to many fields: model pre-training [5, 11, 29], image classification [17, 19], few-shot learning [1, 4, 6, 41], *etc*. For example, Raghu *et al*. [29] propose an implicit differentiation and backpropagation meta-learning method and improve the efficiency of learning pre-training hyper-parameters. Xu *et al*. [41] design a dynamic alignment in meta learning, which can emphasize information in query according to the support. Li *et al*. [12] use the meta learning to handle different input resolutions, thereby generating fusion results with arbitrary resolutions.

Inspired by the above meta learning methods, we design a meta-feature embedding network to generate the meta features from object detection features, thereby helping IVIF network fuse more object semantic information.

3. Proposed Method

Our MetaFusion framework is shown in Figure 2, which includes three subnetworks: IVIF network (F) generates fusion results, OD network (D) offers semantic features, and meta-feature embedding network (MFE) uses OD features to guide IVIF to fuse more object semantic information. The challenge is that OD features can hardly be directly used to guide IVIF due to their task gap. Thus, our focus is to design MFE and optimize it. Details are as follows.

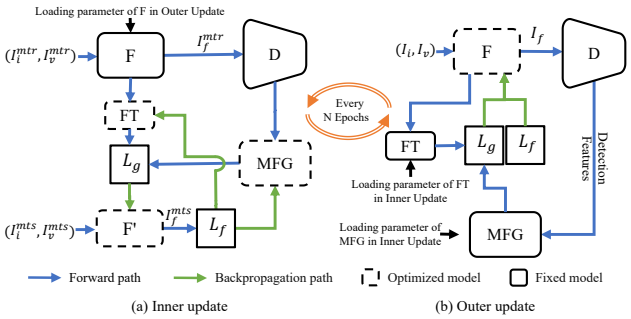

Figure 3. Illustration of the meta-feature embedding learning. There are two stages in the meta-feature embedding: (a) inner update and (b) outer update. In the inner update, MFG and FT are trained according to current F' semantic extracting ability to generate meta features. In the outer update, F is trained which is guided by the meta features, thereby learning to extract object semantic. The two stages alternate every N epoches.

3.1. Meta-feature Embedding

Meta-feature embedding includes meta-feature generator (MFG) and feature transform (FT). MFG generates meta feature F_{mj} according to IVIF feature F_{uj} from OD feature F_{ej}, i.e., $F_{mj} = MFG(F_{uj}, F_{ej})$, where j is the feature level index. FT transfers the meta feature F_{mj} to fusion feature F_{uj} by producing feature bridge F_{tj}. Specifically, meta-feature embedding learning is divided two stages: inner update stage and outer update stage, as shown in Figure 3.

In the inner update stage, we optimize MFG, FT and F (see Figure 3(a)). Given meta training set S_{mtr} and meta testing set S_{mts}, we firstly update the parameter θ_F of F on

S_{mtr} through meta feature guidance:

$$\theta_{F'} = \theta_F - \beta_{F'} \nabla_{\theta_F} L_g(F_{mj}, F_{tj}) = \theta_F - \beta_{F'} \frac{\partial L_g(F_{mj}, F_{tj})}{\partial \theta_F}, \tag{1}$$

where the guide loss L_g is L_2 distance, $\beta_{F'}$ is the learning rate of F', $\theta_{F'}$ is the parameter of F' that is updated from F. Then, we use F' with parameter $\theta_{F'}$ to calculate the fusion loss L_f on meta testing set S_{mts}. Here, L_f can measure the effect of using the meta features to guide F, i.e., when L_f is lower, the semantic extracting ability of F will be improved. Therefore, we use L_f to update parameter θ_{MFG} of MFG and parameter θ_{FT} of FT by

$$\theta_{MFG} = \theta_{MFG} - \beta_{MFG} \nabla_{\theta_{MFG}} L_f(I_f^{mts}, I_i^{mts}, I_v^{mts}), \tag{2}$$

$$\theta_{FT} = \theta_{FT} - \beta_{FT} \nabla_{\theta_{FT}} L_f(I_f^{mts}, I_i^{mts}, I_v^{mts}), \tag{3}$$

where the fusion loss L_f is SSIM loss [38], I_i^{mts} and I_v^{mts} are infrared and visible images from S_{mts}, $I_f^{mts} = F'(I_i^{mts}, I_v^{mts})$ is the meta testing fusion result, and β_{MFG} and β_{FT} are the learning rate of MFG and FT. $\nabla_{\theta_{MFG}} L_f$ can be calculated by

$$\begin{aligned} &\nabla_{\theta_{MFG}} L_f(F'(I_i^{mts}, I_v^{mts}), I_i^{mts}, I_v^{mts}) \\ &= \frac{\partial L_f(F'(I_i^{mts}, I_v^{mts}), I_i^{mts}, I_v^{mts})}{\partial \theta_{F'}} * (-\frac{\partial^2 L_g(F_{mj}, F_{tj})}{\partial \theta_F \partial \theta_{MFG}}). \end{aligned} \tag{4}$$

Similarly, $\nabla_{\theta_{FT}} L_f$ can be calculated by

$$\begin{aligned} &\nabla_{\theta_{FT}} L_f(F'(I_i^{mts}, I_v^{mts}), I_i^{mts}, I_v^{mts}) \\ &= \frac{\partial L_f(F'(I_i^{mts}, I_v^{mts}), I_i^{mts}, I_v^{mts})}{\partial \theta_{F'}} * (-\frac{\partial^2 L_g(F_{mj}, F_{tj})}{\partial \theta_F \partial \theta_{FT}}). \end{aligned} \tag{5}$$

Thus, MFG can learn how to generate meta features according to current F' semantic extracting ability, i.e., makes meta features be compatible with F.

In the outer update stage, F is trained with the initial parameter θ_F, as shown in Figure 3(b). Given the training data $S = \{S_{mtr}, S_{mts}\}$, F is optimized using the fusion loss L_f and guide loss L_g by

$$\begin{aligned} \theta_F &= \theta_F - \beta_F \nabla_{\theta_F} (L_f(I_f, I_i, I_v) + \lambda_g \sum_{j=1}^{3} L_g(F_{mj}, F_{tj})) \\ &= \theta_F - \beta_F (\frac{\partial L_f(I_f, I_i, I_v)}{\partial \theta_F} + \lambda_g \sum_{j=1}^{3} \frac{\partial L_g(F_{mj}, F_{tj})}{\partial \theta_F}), \end{aligned} \tag{6}$$

where I_i and I_v are infrared and visible images from S, λ_g is a hyper-parameter to balance L_f and L_g, and β_F is the learning rate of F.

Finally, the inner update stage and outer update stage are carried out alternately for every N epochs, which improves the optimization of MFG and F.

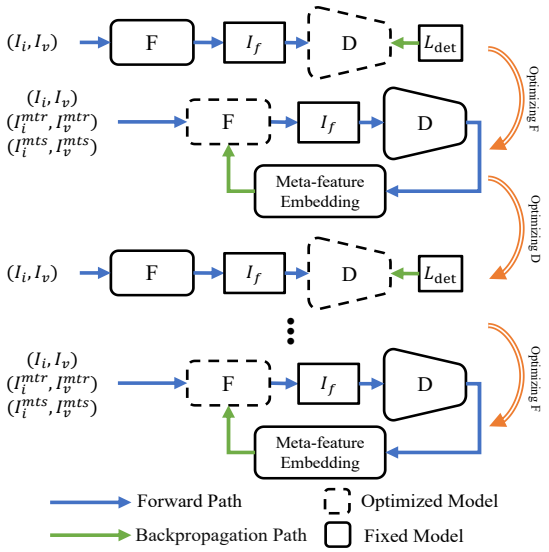

Figure 4. Illustration of the mutual promotion learning between F and D. Firstly, F is used to generate fusion results to optimize D. Thus, a fine-tuned D with higher semantic extracting ability is obtained. Then, D is adopted to optimize F. This process is alternated.

3.2. Mutual Promotion Learning

In Sec. 3.1, we fix the object detection network D to offer stable object semantic features. In this way, MFG does not need to handle variable object semantic features, thereby reducing the difficulty of training MFG. However, the fixed D will restrict the object semantic extracting ability of F. Addressing this problem, we propose the mutual promotion learning that promotes not only F but also D, as shown in Figure 4. Specifically, we firstly train F individually, and then use the fusion results to train D. Sequentially, D offers object semantic features to the meta-feature embedding, thereby improving F. After that, the improved fusion results generated by F are used to fine-tune D. In this way, both F and D are improved.

3.3. Architecture

IVIF network F aims to generate fusion results according to the input infrared and visible images, which is flexible. Here, we adopt the similar structure with [50]. F consists of three feature fusion blocks ($FFB_j, j = 1, 2, 3$) and one image reconstruction module (IRM), as shown in Figure 2. FFB_j is used to extract and fuse features from input infrared image I_i and visible image I_v. It is designed based on the idea of [50] that if FFB_j can well fuse the features of I_i and I_v, then the two kinds of features can be recovered from the fused feature. Thus, FFB_j is described as

$$FFB_j = \begin{cases} C_2T_1 \uplus C_2T_1, j = 1 \\ C_2T_{j-1} \uplus C_2T_{j-1}, j > 1 \end{cases} \tag{7}$$

where j is the block index, C_k represents the number of "Convolution with 3×3 kernel+ReLU" layers is k, $F_{u(j-1)}$ is the output feature of the $j-1$th block, $T_1 = C_2(I_i, I_v)$ and $T_{j-1} = C_2(F_{u(j-1)})$ represent feature integration operations, and $\biguplus$ means feature concatenation. IRM reconstructs the fusion result, which contains six "Convolution with 3×3 kernel +ReLU" layers.

Object detection network D offers object semantic features. In our framework, we choose yolov5s[1] as our D. According to the size of the feature maps, the backbone of D is divided into three feature extraction blocks ($\{DFB_i, i = 1, 2, 3\}$). In addition, we represent the neck and detection head in yolov5s into one block (DH) to simplify the expression.

Meta-feature embedding network MFE includes meta-feature generator MFG_j and feature transform network FT_j. Since OD feature F_{ej} generated by DFB_j offers object semantic information and F_{uj} provides scene details, they are fed into the MFG_j. Then, MFG_j is built as

$$MFG_j = C_6(C_4(F_{uj}) \biguplus C_2(Up(F_{ej}))), \tag{8}$$

where $Up()$ is the feature up-sampling. FT_j transfers the meta feature to fusion feature, which includes three "Convolution with 3×3 kernel +ReLU" layers.

3.4. Training

The training details of our framework is shown in Algorithm 1, which contains four steps. Generally, we firstly pretrain F, and then use the fusion results to finetune D. Sequentially, we train FT and MFG, where MFG and FT are updated $n = 200$ times every $N = 8$ epochs. Lastly, the mutual promotion between F and D is trained for $R = 2$ rounds, thereby improving their performances.

Algorithm 1 MetaFusion Training Algorithm

Input: Training dataset $S = \{S_{mtr}, S_{mts}\}$, IVIF model $F(\theta_F)$, meta-feature generator $MFG(\theta_F)$, feature transform network $FT(\theta_{FT})$, OD network $D(\theta_D)$
Output: IVIF model $F(\theta_F^*)$

Initialize $F(\theta_F)$, $MFG(\theta_F)$, $FT(\theta_{FT})$, $D(\theta_D)$
/* Pretrain F */
while not converged **do**
  Sample image pair (I_i, I_v) from S
  Forward pass on F to generate fusion result I_f
  Optimize F using fusion loss L_f
end while
/* Pretrain D */
while not converged **do**
  Sample image I_f from the fused images generated by F
  Forward pass on D to get detection result
  Optimize D using detection loss L_d
end while
/* Meta-feature embedding outer update */
while not converged **do**
  Sample a batch image pair (I_i, I_v) from S
  Forward pass on F to generate fusion result I_f
  Forward pass on MFG to generate meta feature
  Optimize F using fusion loss L_f and guide loss L_g by Eq. 6
  /* Meta-feature embedding inner update */
  if epoch%N==0 **then**
    for t=1 to n **do**
      Sample an image pair (I_i^{mtr}, I_v^{mtr}) from S_{mtr}
      Compute $\theta_{F'}$ by Eq. 1
      Sample an image pair (I_i^{mts}, I_v^{mts}) from S_{mts}
      Calculate L_f using $\theta_{F'}$
      Update θ_{MFG} and θ_{FT} using the gradient of L_f by Eqs. 2 and 3
    end for
  end if
end while
/* Mutual promotion*/
if Number of round $\leq$ R **then**
  Go back to line 9
end if

4. Experiments

4.1. Setup

Dataset. We conduct experiment on three widely-used datasets: M^{3}FD [20], RoadScene [44] and TNO [37]. M^{3}FD is divided as training set (2940 image pairs) and testing set (1260 image pairs). RoadScene with 221 image pairs and TNO with 40 image pairs are only used for testing. Besides, M^{3}FD is adopted to evaluate OD performance.

Implementation. Our framework is implemented with PyTorch on a NVIDIA GeForce RTX 3090 GPU. F and MFE are trained using optimizer Adam with the learning rate β_F, $\beta_{F'}$, β_{FT} and β_{MFG} of 1×10^{-3}, respectively. D is trained followed by yolov5s that adopts the optimizer SGD using the learning rate of 1×10^{-2} with decaying rate of 0.1 every round. We firstly train F for 100 epochs and D for 150 epochs. Then, we conduct MFE for 50 epochs. After that, D is finetuned for 150 epochs. Following the same strategy, the mutual promotion is carried out R rounds. The images are resized to 512×384 with batchsize of 1. The hyperparameter λ_g is set to 0.1.

Metric. Three metrics are used for IVIF evaluation: entropy (EN) [33], mutual information (MI) [28] and visual information fidelity (VIF) [7]. EN evaluates the information richness in an image, and the higher EN means more information. MI evaluates the information similarity between

[1] https://github.com/ultralytics/YOLOv5

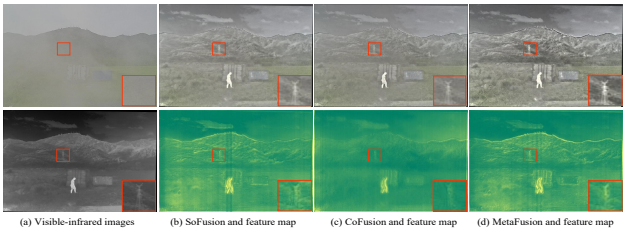

Figure 5. Visual comparison of different optimization methods.

Table 1. Effect study of meta-feature embedding by comparing with different optimization methods on M^3FD. The best result is in red.

Method	Metric		
	MI	EN	VIF
SoFusion	14.164	7.078	1.190
CoFusion	14.187	7.089	1.345
MetaFusion	14.511	7.249	1.515

the input images and fused image. The higher MI illustrates more information of the input images is fused. VIF measures the ability to extract visible information from the input image, and a larger VIF represents less visible distortion in the fused result. Here, we use the V channel in HSV space of the fusion results to calculate these metrics. Moreover, we use $mAP_{50\to95}$ [9] to comprehensively evaluate OD performance, where the average of mAPs sampling every 5 from AP_{50} to AP_{95} is calculated. A higher $mAP_{50\to95}$ means better OD effect.

4.2. Ablation Study

Effect of meta-feature embedding. In Sec. 3.1, we introduce the meta-feature embedding that generates the meta features from object detection features to help IVIF network learn more object semantic information. To verify its effect, separate optimization method (SoFusion) and cascaded optimization method (CoFusion) are compared, as shown in Figure 1(a)-(b). Specifically, SoFusion firstly optimizes F and then uses the generated fusion results to optimize D. CoFusion treats D as a constraint to optimize F with the loss L_f and L_d. The results are shown in Table 1. Our MetaFusion achieves the best results on IVIF. The reason is that SoFusion neglects the help of D, and directly utilizing the high-level D to guide the pixel-level F has feature mismatch in CoFusion. In contrast, we implement the meta-feature embedding that generates meta features from D according to the ability of F. Thus, the meta features are naturally compatible with F, thereby providing effective object semantic features to F. Visual comparison is shown in Figure 5. MetaFusion highlights the object feature and generates the clearest result.

Influence of mutual promotion learning. As described in Sec. 3.2, we propose the mutual promotion learning promote F and D alternately. Here, we study the influence of mutual promotion learning by evaluating round $R = 0, 1, 2$. In particular, $R = 0$ represents F and D are trained respectively. As shown in Table 2, the results of F are better with the R increases. Comprehensively considering training efficiency and performance, we take $R = 2$.

Table 2. Influence study of mutual promotion learning by evaluating the performances of round R on M^3FD. The best result is in red.

Method	Metric			
	MI	EN	VIF	$mAP_{50\to95}$ (%)
$R = 0$	14.164	7.078	1.190	55.6
$R = 1$	14.464	7.226	1.474	55.8
$R = 2$	14.511	7.249	1.515	56.5

Table 3. Study of multi-level meta-feature embedding by comparing different configurations on M^3FD. The best result is in red.

Method	Metric		
	MI	EN	VIF
MetaFusion-L1	14.450	7.220	1.259
MetaFusion-L2	14.452	7.220	1.377
MetaFusion-L3	14.464	7.226	1.474

Study of multi-level meta-feature embedding. We implement multi-level meta-feature embedding to build MetaFusion framework, as introduced in Sec. 3.3. Here, we study the number of meta-feature embedding levels with the following configurations. One-level meta-feature embedding (MetaFusion-L1) with $MFE = \{MFG_1, FT_1\}$, $F = \{FFB_1, IRM\}$ and $D = \{DFB_1, DH\}$. Two-level meta-feature embedding (MetaFusion-L2) with $MFE = \{MFG_1, MFG_2, FT_1, FT_2\}$, $F = \{FFB_1, FFB_2, IRM\}$ and $D = \{DFB_1, DFB_2, DH\}$. Three-level meta-feature embedding (MetaFusion-L3) with $MFE = \{MFG_1, MFG_2, MFG_3, FT_1, FT_2, FT_3\}$, $F = \{FFB_1, FFB_2, FFB_3, IRM\}$ and $D = \{DFB_1, DFB_2, DFB_3, DH\}$. The results are shown in Table 3. With the number of multi-level meta-feature embedding increases, F achieves higher image fusion quality. Since meta-feature embedding provide multi-level object semantic features to F.

4.3. Comparison with State-of-the-art Methods

We compare the proposed MetaFusion with eight SOTA fusion methods to verify the superiority: FusionGAN [25], GANMcC [26], MFEIF [21], YDTR [36], PIAFusion [35], SwinFusion [23], Tardal [20], and U2Fusion [44]. Their available codes and recommended parameter settings are

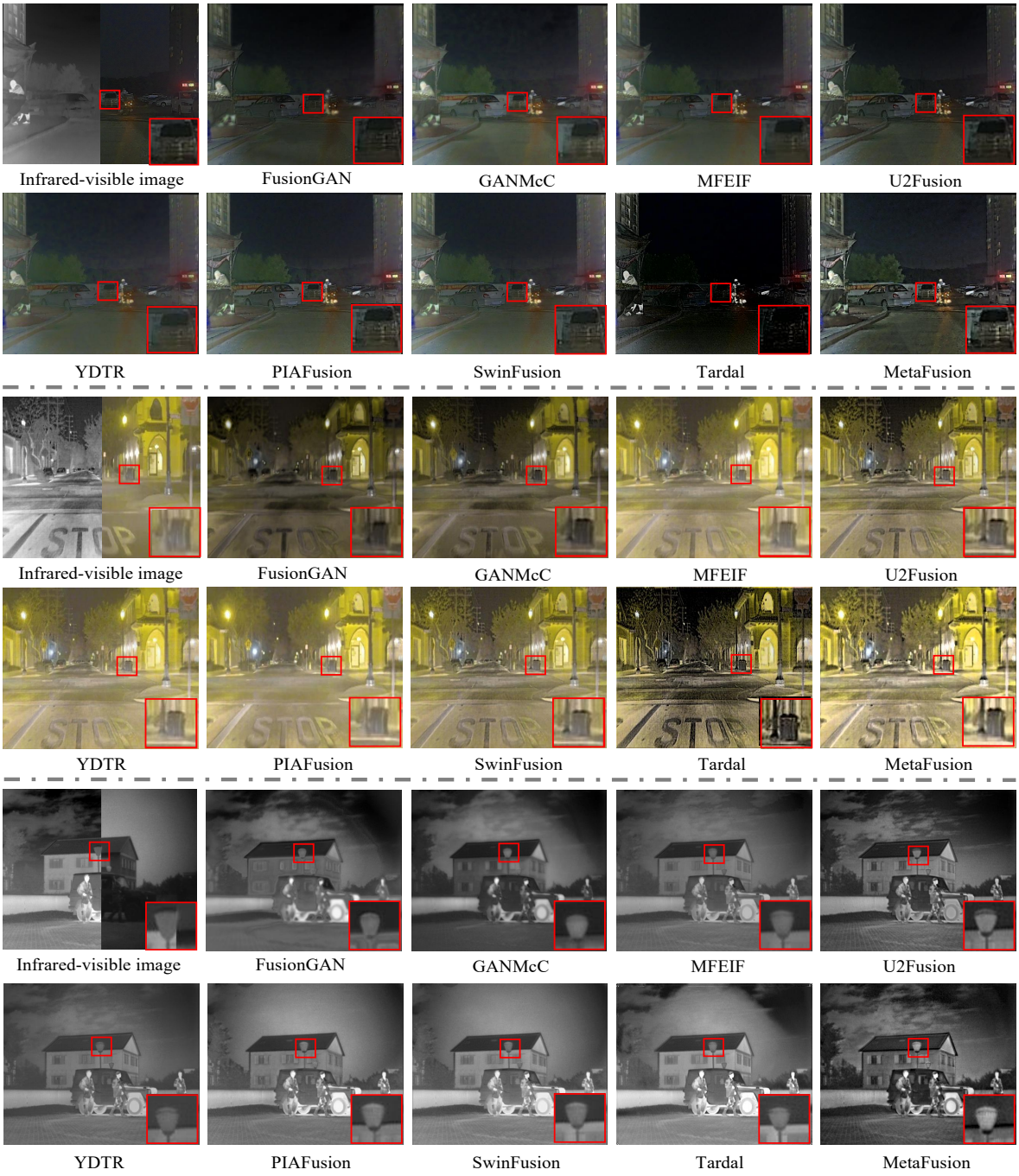

Figure 6. Qualitative results of different fusion methods on M^3FD (first group), RoadScene (second group) and TNO (third group).

adopted to generate fusion results for the fair comparison.

Qualitative results of different fusion methods are shown in Figure 6. All the fusion methods can fuse the main features of the infrared and visible images to some extent. However, FusionGAN, GANMcC, MFEIF and YDTR produce blurred edge details, and U2Fusion, PIAFusion, SwinFusion and Tardal generate low contrast objects, as shown in the red rectangular boxes. In contrast, the images generated by the proposed MetaFusion contain clear edge details and high contrast objects.

Sequentially, we provide quantitative results of different fusion methods in Table 4. Our MetaFusion generally achieves the largest or the second-largest metric values. In detail, the high EN and MI illustrate the fusion images by our MetaFusion contain high contrast object and clear edge details. The large VIF shows our fusion results have high-

Table 4. Quantitative results of different fusion methods on TNO, RoadScene and M^3FD datasets. The model inference time is counted on a NVIDIA GeForce RTX 2080 Ti. The best result is in red and the second best one is in violet.

Method	M3FD			RoadScene			TNO40			Time (s)
	MI	Ent	VIF	MI	Ent	VIF	MI	Ent	VIF	
FusionGAN [25]	13.445	6.722	0.303	14.203	7.101	0.251	13.068	6.534	0.252	0.040
GANMcC [26]	13.731	6.865	0.453	14.017	7.008	0.422	13.485	6.742	0.424	0.081
MFEIF [21]	12.957	6.478	0.401	13.489	6.742	0.260	13.360	6.680	0.395	0.029
U2Fusion [44]	13.816	6.908	0.545	13.888	6.944	0.456	13.889	6.944	0.636	0.043
YDTR [36]	12.694	6.347	0.361	13.166	6.583	0.236	12.862	6.431	0.280	0.086
PIAFusion [35]	13.326	6.663	0.437	13.262	6.533	0.205	13.931	6.960	0.506	0.055
SwinFusion [23]	13.051	6.525	0.460	13.267	6.633	0.313	13.879	6.933	0.479	1.081
Tardal [20]	13.636	6.818	0.650	15.009	7.504	0.483	13.030	6.515	0.910	0.030
MetaFusion	14.511	7.249	1.515	14.218	7.022	0.969	14.657	7.323	1.462	0.015

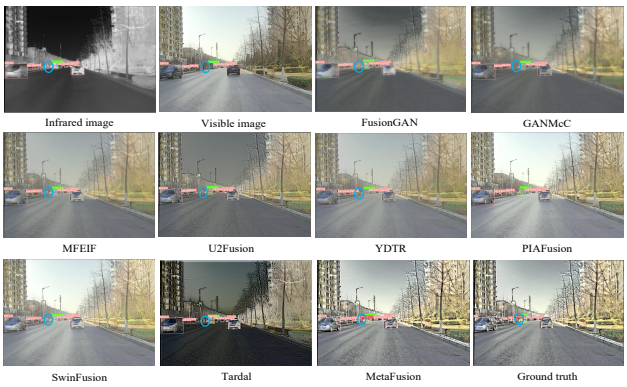

Figure 7. Visual results of object detection based on different fusion methods on M^3FD.

quality visual effect and small distortion compared with the source images. Moreover, our method achieves the fastest inference time that only needs 0.015s to generate one fusion result.

4.4. Evaluation on Infrared-visible Object Detection

A better fusion image can provide more information to improve object detection. Here, we verify the effectiveness of our method by evaluating object detection accuracy with the infrared and visible fusion images. To make a fair comparison, we firstly generate fusion results using our method and SOTA methods, and then use the results to retrain the object detection baseline YOLOv5s, respectively.

Figure 7 shows the object detection results. In general, the SOTA methods improve the performance of object detection. In contrast, our MetaFusion achieves a better performance, e.g., the person is accurately detected (see the blue ellipse). Table 5 shows the quantitative results. Our fusion results help the detection network achieve the highest object detection accuracy. This further proves that our method can generate high-quality fusion results especially for the objects.

Table 5. Quantitative results of object detection based on different fusion methods on M^3FD dataset. The best result is in red and the second best one is in violet.

Method	$mAP_{50\rightarrow95}$ (%)
FusionGAN [25]	54.2
GANMcC [26]	55.2
MFEIF [21]	55.4
U2Fusion [44]	55.7
YDTR [36]	55.4
PIAFusion [35]	55.6
SwinFusion [23]	55.4
Tardal [20]	54.4
MetaFusion	56.5

5. Conclusion

This paper presents a joint fusion and detection learning framework through introducing the meta-feature embedding model. Based on the meta learning idea, the meta-feature embedding model can generate object semantic features according to the fusion network ability, thereby bridging the feature gap between these two different-level tasks. Moreover, a mutual promotion learning between fusion and detection tasks is further implemented to improve their performances. Both quantitative and qualitative results demonstrate the superior performance of our method compared with the state-of-the-art methods.

Acknowledgement. This work is supported by National Natural Science Foundation of China under Grant Nos. 62176038, 62001450 and U1903215.

References

[1] Sungyong Baik, Janghoon Choi, Heewon Kim, Dohee Cho, Jaesik Min, and Kyoung Mu Lee. Meta-learning with task-adaptive loss function for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9465–9474, October 2021. 3

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing. 2

[4] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14166–14175, June 2022. 3

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 3

[6] Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. Meta faster R-CNN: towards accurate few-shot object detection with attentive feature alignment. In *AAAI*, pages 780–789. AAAI Press, 2022. 3

[7] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14:–, 04 2013. 5

[8] Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2

[9] Liqiang He and Sinisa Todorovic. Destr: Object detection with split transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9367–9376, 2022. 6

[10] Ruichao Hou, Dongming Zhou, Rencan Nie, Dong Liu, Lei Xiong, Yanbu Guo, and Chuanbo Yu. Vif-net: An unsupervised framework for infrared and visible image fusion. *IEEE Transactions on Computational Imaging*, 6:640–651, 2020. 2

[11] Muhammad Abdullah Jamal, Liqiang Wang, and Boqing Gong. A lazy approach to long-horizon gradient-based meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6577–6586, October 2021. 3

[12] Huafeng Li, Yueliang Cen, Yu Liu, Xun Chen, and Zhengtao Yu. Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 30:4070–4083, 2021. 3

[13] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. 2

[14] Hui Li, Xiao-Jun Wu, and Josef Kittler. Mdlatlrr: A novel decomposition method for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 29:4733–4746, 2020. 2

[15] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021. 1

[16] Jing Li, Jianming Zhu, Chang Li, Xun Chen, and Bin Yang. Cgtf: Convolution-guided transformer for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022. 2

[17] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5212–5221, 2021. 3

[18] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 07 2018. 2

[19] Benlin Liu, Yongming Rao, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Metadistiller: Network self-boosting via meta-learned top-down distillation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 694–709. Springer, 2020. 3

[20] 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. 1, 2, 5, 6, 8

[21] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. 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*, pages 1–1, 2021. 6, 8

[22] Yu Liu, Shuping Liu, and Zengfu Wang. A general framework for image fusion based on multi-scale transform and sparse representation. *Information Fusion*, 24:147–164, 2015. 2

[23] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. 2, 6, 8

[24] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020. 2

[25] 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, 08 2019. 1, 6, 8

[26] 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, 2021. 2, 6, 8

[27] Yan Mo, Xudong Kang, Puhong Duan, Bin Sun, and Shutao Li. Attribute filter based infrared and visible image fusion. *Information Fusion*, 75:41–54, 2021. 2

[28] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics Letters*, 38:313 – 315, 04 2002. 5

[29] Aniruddh Raghu, Jonathan Lorraine, Simon Kornblith, Matthew McDermott, and David K Duvenaud. Meta-learning to improve pre-training. *Advances in Neural Information Processing Systems*, 34:23231–23244, 2021. 3

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 2

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 2

[32] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019. 2

[33] Wesley Roberts, Jan van Aardt, and Fethi Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2:1–28, 05 2008. 5

[34] 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*, Thematic Workshops '17, page 3543, New York, NY, USA, 2017. Association for Computing Machinery. 1

[35] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84:79–92, 2022. 1, 2, 6, 8

[36] Wei Tang, Fazhi He, and Yu Liu. Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. *IEEE Transactions on Multimedia*, pages 1–16, 2022. 2, 6, 8

[37] Alexander Toet. The tno multiband image data collection. *Data in Brief*, 15:249–251, 2017. 5

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

[39] Zhishe Wang, Yuanyuan Wu, Junyao Wang, Jiawei Xu, and Wenyu Shao. Res2fusion: Infrared and visible image fusion based on dense res2net and double nonlocal attention models. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022. 2

[40] Pingyu Wu, Wei Zhai, and Yang Cao. Background activation suppression for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14248–14257, June 2022. 2

[41] Chengming Xu, Yanwei Fu, Chen Liu, Chengjie Wang, Jilin Li, Feiyue Huang, Li Zhang, and Xiangyang Xue. Learning dynamic alignment via meta-filter for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5182–5191, June 2021. 3

[42] Dongli Xu, Jinghong Deng, and Wen Li. Revisiting ap loss for dense object detection: Adaptive ranking pair selection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 14187–14196, 2022. 2

[43] Han Xu, Meiqi Gong, Xin Tian, Jun Huang, and Jiayi Ma. Cufd: An encoderdecoder network for visible and infrared image fusion based on common and unique feature decomposition. *Computer Vision and Image Understanding*, 218:103407, 03 2022. 1

[44] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. 1, 2, 5, 6, 8

[45] Han Xu, Xinya Wang, and Jiayi Ma. Drf: Disentangled representation for visible and infrared image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021. 1, 2

[46] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Rui-Wei Zhao, Tao Zhang, Xuequan Lu, and Shang Gao. Cream: Weakly supervised object localization via class re-activation mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9437–9446, June 2022. 2

[47] Meilong Xu, Linfeng Tang, Hao Zhang, and Jiayi Ma. Infrared and visible image fusion via parallel scene and texture learning. *Pattern Recognition*, 132:108929, 2022. 1

[48] Shilong Zhang, Zhuoran Yu, Liyang Liu, Xinjiang Wang, Aojun Zhou, and Kai Chen. Group r-cnn for weakly semi-supervised object detection with points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9417–9426, June 2022. 2

[49] Fan Zhao and Wenda Zhao. Learning specific and general realm feature representations for image fusion. *IEEE Transactions on Multimedia*, 23:2745–2756, 2020. 2

[50] Fan Zhao, Wenda Zhao, and Huchuan Lu. Interactive feature embedding for infrared and visible image fusion. *arXiv preprint arXiv:2211.04877*, 2022. 2, 4

[51] Fan Zhao, Wenda Zhao, Libo Yao, and Yu Liu. Self-supervised feature adaption for infrared and visible image fusion. *Information Fusion*, 76:189–203, 2021. 2

[52] Wenda Zhao, Xueqing Hou, You He, and Huchuan Lu. Defocus blur detection via boosting diversity of deep ensemble networks. *IEEE Transactions on Image Processing*, 30:5426–5438, 2021. 2

[53] Wenda Zhao, Huimin Lu, and Dong Wang. Multisensor image fusion and enhancement in spectral total variation domain. *IEEE Transactions on Multimedia*, 20(4):866–879, 2018. 2

[54] Wenda Zhao, Dong Wang, and Huchuan Lu. Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(4):1102–1115, 2018. 2

[55] Wenda Zhao, Bowen Zheng, Qiuhua Lin, and Huchuan Lu. Enhancing diversity of defocus blur detectors via cross-ensemble network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8905–8913, 2019. 2

[56] Zhengjie Zhu, Xiaogang Yang, Ruitao Lu, Tong Shen, Xueli Xie, and Tao Zhang. Clf-net: Contrastive learning for infrared and visible image fusion network. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022. 2