Beyond Strict Pairing: Arbitrarily Paired Training for High-Performance Infrared and Visible Image Fusion

Supplementary Material

7. Method

7.1. Baseline models

As shown in Fig. 8, three baseline IVIF models based on CNN, Transformer, and GAN are constructed to support the subsequent validation experiments. It is worth noting that our aim is to explore APTP and UPTP and then compare all training paradigms; thus, the three proposed baseline models are applicable to every one of them.

7.1.1. U-shaped Hierarchical Multi-scale Architecture

To achieve high-quality fusion results at a low cost with minimal information loss, we construct all baselines upon a concise and efficient U-shaped hierarchical multi-scale architecture, which is highly scalable and can be adjusted according to different task requirements. As shown in Fig. 8 (a) and (b), the model consists of three layers, each following a pipeline of feature extraction, fusion, and reconstruction, which can be respectively represented by $E_i(\cdot)$, $F_i(\cdot)$ and $R_i(\cdot), (i = 1, 2, 3)$.

E_1 and R_1 are used to map the input data to high-dimensional feature maps (sequences) and reconstruct them into the fused image, respectively. $E_i, R_i (i = 2, 3)$ consists of cascaded dense-blocks (transformer-encoder blocks) to deal with multi-scale features. Each layer of $E_i (i = 1, 2, 3)$ is followed by a global(local) fusion layer $F_i(\cdot), (i = 1, 2, 3)$ to obtain global-local fusion features. For the CNN baseline, we achieve global feature fusion by modeling global interactions between serialised local features. Conversely, for the Transformer baseline, we enhance local features by compressing and refining extracted global information. The transformer baseline employs channel attention, surviving from incorrect positional priors. The global-local fusion features are then obtained by adding them to the original features. The forward propagation process of networks can be explicitly seen from the Fig. 8.

7.1.2. Generative Adversarial process

As shown in Fig. 8 (c), two discriminators are designed for each modality, engaging in an adversarial game with the generator $\mathcal{G}$. In addition, the saliency discriminator $\mathcal{D}_s$ primarily focuses on distinguishing the thermal radiation targets between I_{ir} and O after masking the background details by detail mask m_s, while the detail discriminator $\mathcal{D}_d$ mainly evaluates the texture information of I_{vis} and O after masking the thermal radiation information by saliency mask m_s. Specifically, m_s and m_d satisfy the relation $m_s + m_d = 1$. $\mathcal{D}_s$ and $\mathcal{D}_d$ share the same network structure with independent parameters.

7.2. Loss Function

The adversarial loss of $\mathcal{D}_s$ and $\mathcal{D}_d$ are:

$$\mathcal{L}_{\mathcal{D}_s} = \mathbb{E}_{\tilde{x}\sim p(\mathcal{M}_s(O))}\mathcal{D}_s(\tilde{x}) - \mathbb{E}_{x\sim p(\mathcal{M}_s(I_{ir}))}\mathcal{D}_s(x), \tag{16}$$

$$\mathcal{L}_{\mathcal{D}_d} = \mathbb{E}_{\tilde{x}\sim p(\mathcal{M}_d(\nabla O))}\mathcal{D}_d(\tilde{x}) - \mathbb{E}_{x\sim p(\mathcal{M}_d(\nabla I_{vis}))}\mathcal{D}_d(x), \tag{17}$$

where $\mathcal{M}_s(\cdot)$ and $\mathcal{M}_d(\cdot)$ represent the mask operation. The training of the generator is jointly driven by the adversarial loss $\mathcal{L}_{adv}$ and the total loss ($\mathcal{L}_G$):

$$\mathcal{L}_{adv} = \mathbb{E}_{\hat{x}\sim p(\mathcal{M}_s(O))}\mathcal{D}_s(\hat{x}) + \mathbb{E}_{\hat{x}\sim p(\mathcal{M}_d(\nabla O))}\mathcal{D}_d(\hat{x}), \tag{18}$$

$$\mathcal{L}_G = \alpha\mathcal{L}_{int} + \beta\mathcal{L}_{grad} + \mathcal{L}_{ssim} - \lambda\mathcal{L}_{adv}. \tag{19}$$

8. Experiments

8.1. Challenging with Unbalanced Source Inputs

In real-world scenarios, the two modalities may not provide the same number of images. Therefore, we design two controlled experiments, *i.e.*, one with 75 infrared and 150 visible images, and another with 150 infrared and 75 visible images. They are all expanded to 7,500 image pairs. Even when one modality is scarcer, its images can still be paired with the richer modality in new combinations, preserving the optimal fusion performance.

Tab. 5 shows consistent improvement in multiple metrics across three datasets, despite EN fluctuations, indicating that the model can learn more robust capabilities from the enriched trainable pixel relationships. These include processing inter-modal pixel relationships, fusing modal details and structures while suppressing noise previously caused by limited data. The clearly sharper infrared architectural textures in Fig. 9 provide further evidence.

8.2. Compatibility with State-of-the-Art Methods

Existing methods are all trained under the SPTP. However, considering that their supervisions and model architectures do not explicitly enforce content consistency, they can be directly generalised to UPTP and APTP. To verify that the robustness gains offered by our paradigms also hold for some current SOTAs, we select one representative for each of the CNN, Transformer, and GAN variants, to conduct contrastive experiments.

Using only 150 pairs of paired data and by altering pairing relationships, we expand the trainable dataset to 15,000 pairs, and retrain the three SOTA methods using UPTP. The

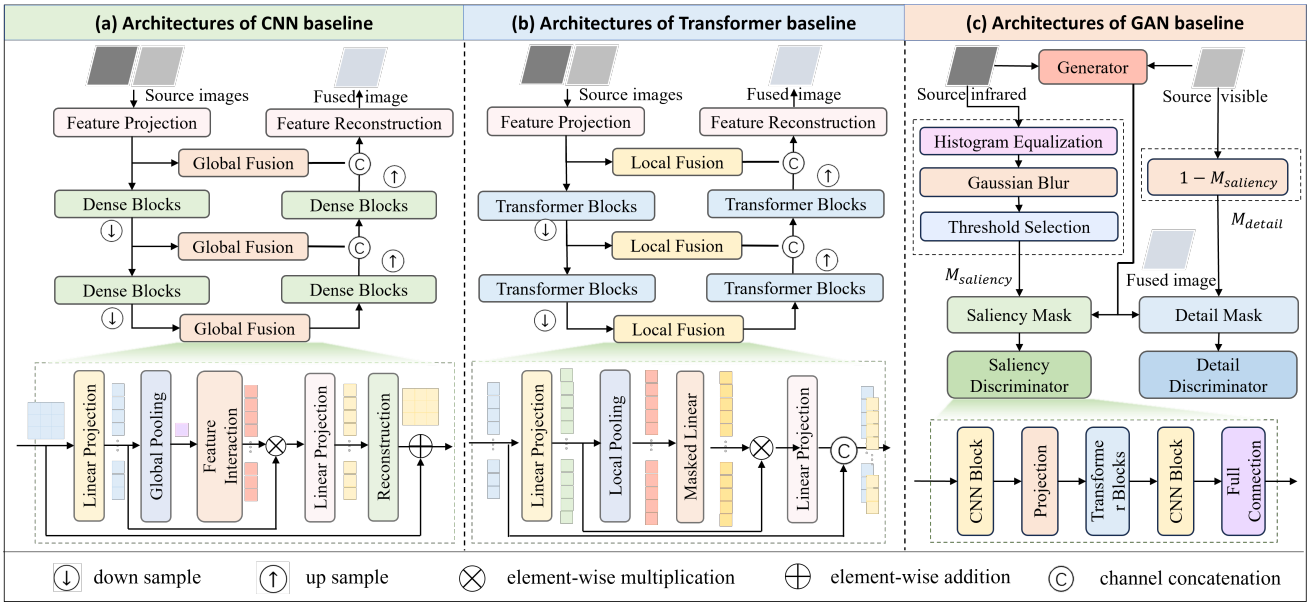

Figure 8. The architectures of the three baselines.

Figure 9. Visualisation of fused images: unpaired training with missing modalities vs paired training with complete modalities.

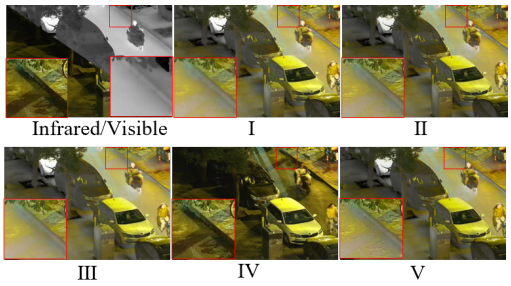

Figure 10. Visualisation of fused images supervised by differernt loss function groups.

results are compared with those obtained from models re-trained on strictly paired datasets of 150 and 15,000 pairs, respectively. As evidenced by the metric values in metric columns and the average metric improvement percentage in the last column, our proposed novel training paradigm not only generalises directly to existing SOTA but also significantly enhances model performance even with limited data, even surpassing the effect achieved with 100 times the amount of strictly-paired data.

8.3. Ablation Study of Loss Items

To validate the effectiveness of the proposed adaptive weighted loss function, we designed several different combinations of loss functions for comparison. The intensity loss function balances cross-modal pixel distribution, helping establish a fundamental image foundation. The gradient loss function further supplements texture details based on this foundation, while the structural similarity loss function contributes additional information such as contrast, structure, and luminance.

The first row in Tab. 7 corresponds to using only the intensity loss, aligning with the third column in Fig. 10. Clearly, the fused images in this case retain only basic content information, lacking gradient texture details compared to the experimental setup in the second row of Tab. 7, and demonstrating deficiencies in overall image structure and visual fidelity compared to the third row. The fourth row in the table represents the configuration using only gradient loss and structural similarity loss. While it effectively preserves visible texture and content information, it loses a significant amount of infrared thermal radiation information (fifth column in Fig. 10). In summary, the combination of all three loss functions achieves the best qualitative and quantitative results.

8.4. Comparison with State-of-the-Art Methods

Here, we provide the training settings for all methods in Tab. 4, we also present more visualisation results in addition to the quantitative results in Tab. 4. In terms of the training settings, different methods employ distinct datasets and pre-processing operations such as cropping or resizing on their training data. Notably, methods including DCINN, CTHIE, FreeFusion, and GIFNet use two or more datasets, while others like DCINN, SAGE, and GIFNet were trained on the specific datasets tested in Table 4. It is noteworthy

Table 5. Quantitative comparison of models trained on imbalanced source modality data from the MSRS dataset, evaluated on the MSRS, LLVIP, and M3FD test sets. Bold indicates that the model performs better than the one trained on its original paired data.

Dataset	Quantity						EN	MI	VIF	Q_{abf}	SSIM
	Infrared			Visible							
	PB	w/o E	T	PB	w/o E	T					
MSRS	150	×	150	150	×	150	6.54	2.57	0.88	0.60	0.96
	75	✓	7500	150	✓	7500	**6.56**	**2.70**	**0.92**	**0.65**	**0.98**
	150	✓	7500	75	✓	7500	**6.58**	**2.71**	**0.93**	**0.66**	**0.99**
LLVIP	150	×	150	150	×	150	7.37	2.27	0.78	0.57	0.81
	75	✓	7500	150	✓	7500	7.29	**2.53**	**0.84**	**0.63**	**0.91**
	150	✓	7500	75	✓	7500	7.31	**2.59**	**.85**	**0.62**	**0.91**
M3FD	150	×	150	150	×	150	6.95	2.18	0.71	0.56	0.91
	75	✓	7500	150	✓	7500	6.81	**2.42**	**0.73**	**0.58**	**0.98**
	150	✓	7500	75	✓	7500	6.83	**2.49**	**0.74**	**0.57**	**0.98**

Table 6. Quantitative results on LLVIP of applying new paradigm to SOTA methods with the same training numbers.

Methods	Venue	Mode	Quantity			EN	MI	VIF	Qabf	SSIM	Avg.Imp.
			PB	w/o E	T						
CDD	23' CVPR	SPTP	150	×	150	6.62	2.58	0.93	0.63	1.00	–
		UPTP	150	✓	15000	**6.73**	**3.77**	**1.03**	**0.70**	**1.01**	**15.93%**
		SPTP	15000	×	15000	6.69	3.50	1.04	0.72	0.98	–
		UPTP	150	✓	15000	**6.73**	**3.77**	1.03	0.70	**1.01**	**7.52%**
MMDR	24' MM	SPTP	150	×	150	6.11	2.40	0.51	0.31	0.57	–
		UPTP	150	✓	15000	**6.69**	**2.86**	**0.97**	**0.65**	**1.02**	**62.38%**
		SPTP	15000	×	15000	6.68	2.86	0.93	0.63	0.99	–
		UPTP	150	✓	15000	**6.69**	2.86	**0.97**	**0.65**	**1.02**	**2.11%**
TGF	23' TIP	SPTP	150	×	150	6.52	2.00	0.83	0.59	0.88	–
		UPTP	150	✓	15000	**6.67**	**2.26**	**0.95**	**0.66**	**0.99**	**10.49%**
		SPTP	15000	×	15000	6.66	2.24	0.94	0.66	1.00	–
		UPTP	150	✓	15000	**6.67**	**2.26**	**0.95**	0.66	0.99	**0.96%**

Table 7. Qualitatively ablation results on LLVIP of CNN baseline supervised by different loss function combinations. Bold shows the best value.

Settings	Items			EN	MI	VIF	Qabf	SSIM
	$\mathcal{L}_{int}$	$\mathcal{L}_{grad}$	$\mathcal{L}_{ssim}$					
I	✓	×	×	7.33	2.86	0.82	0.52	0.83
II	✓	✓	×	7.31	2.65	0.85	0.62	0.85
III	✓	×	✓	7.34	2.77	0.85	0.52	0.89
IV	×	✓	✓	6.43	**3.24**	**0.95**	0.61	0.81
V	✓	✓	✓	**7.34**	2.54	0.88	**0.63**	**0.91**

that among the compared methods, only DCINN and FreeFusion are trained on fewer than 1,000 samples (231 and 914). In contrast, our three baseline models use only 10 image pairs each. By cutting these into 150 patches (128 × 128) and applying our proposed UPTP, we effectively yield 15,000 trainable image pairs. Despite this significantly smaller training set, our proposed training paradigm enables the model to learn from rich, relational information within the data, ultimately achieving superior fusion performance compared to all the aforementioned methods. It is supported by the quantitative results in Tab. 4 and the qualitative results in Fig. 11 . Fig. 11 presents visual comparisons of fusion results on the LLVIP, MSRS, RoadScene datasets, including SOTA methods and three baseline approaches.

For rows 1 to 2, the red box highlights tree roots, while the orange box shows a car behind green leaves—both cases demonstrate each method's capability to retain infrared thermal radiation information. Only DeFusion and our methods successfully preserve thermal radiation in these two areas, whereas many methods(LRRNet, DDFM, CrossFuse, DCINN, and GIFNet) fail to even maintain the thermal radiation of the car in the upper left corner. However, as shown in the green box, the ground texture appears significantly clearer using our methods compared to DeFusion. For rows 3 to 4, the red boxes highlight eaves under strong illumination, and the yellow boxes indicate leaf edges. Seven methods—DeFusion, MetaFusion, DDFM, CrossFuse, FreeFusion, SAGE, and GIFNet—all exhibit artifacts, particularly around the leaf edges. Meanwhile, both

Table 8. Training settings of ours and SOTA methods in Tab. 4. C: COCO, K: KAIST, M^3: M^3FD, MS: MSRS, I: ImageNet, F: FILR, R: RoadScene, T: TNO, L: LLVIP, P: Postdam, W: WHU, M: MFNet.

Method	DeF	LRR	Meta	DDF	Cross	DCINN	CTHI	FreeFusion	SAGE	GIFNet	**Ours**
Dataset	C	K	M^3	I	K	R T	K F	P W M	R T M^3	L	**MS M^3**
Num(k)	50	20	2.940	1280	40	0.231	12	0.914	3.611	12.025	**0.01**
Process	resize	resize	resize	resize	resize	crop	crop	crop	crop resize	crop	**crop**

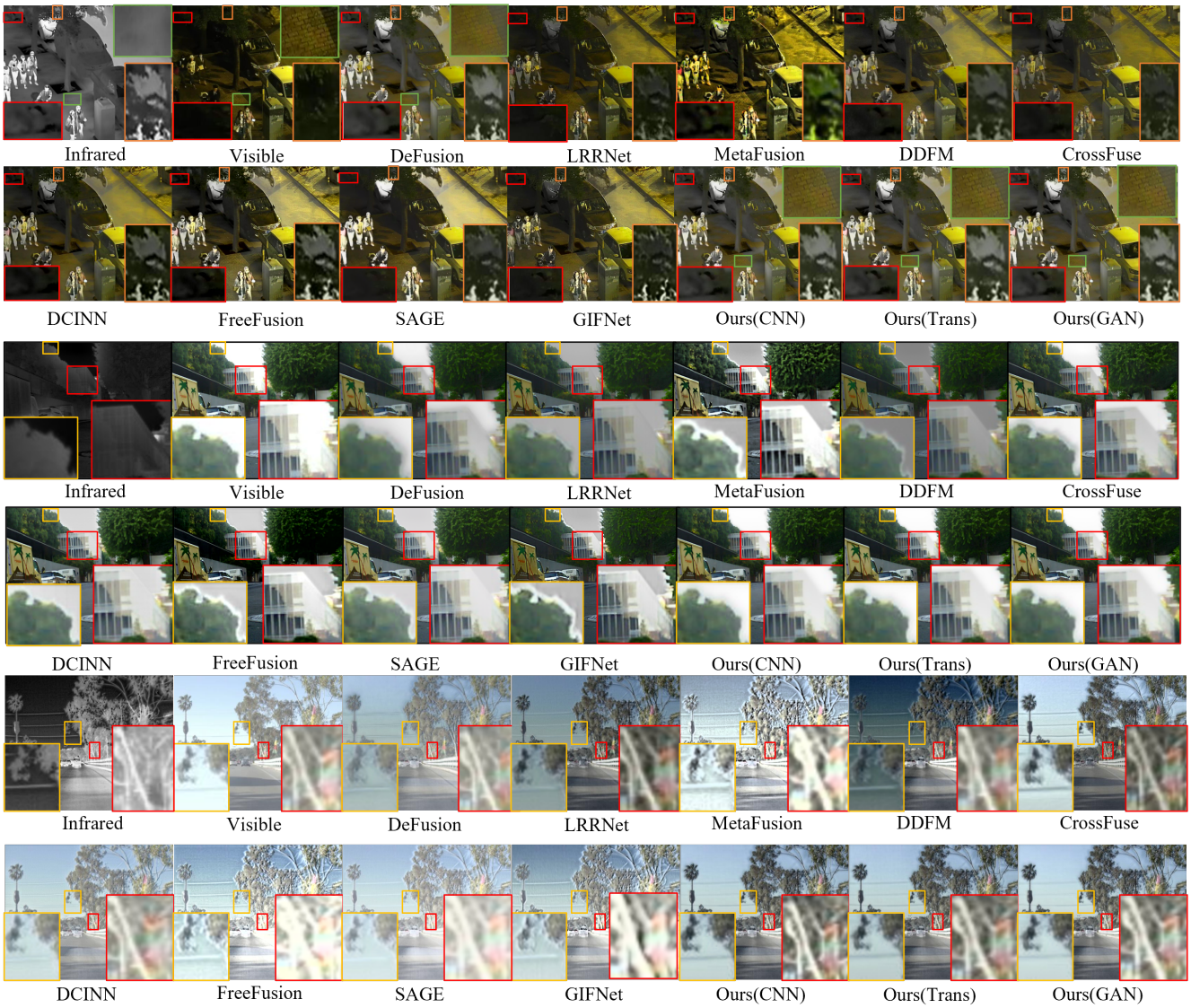

Figure 11. Visualisation of SOTA fused images on the LLVIP, MSRS, and RoadScene datasets (from top to bottom).

LRRNet and DCINN suffer from color distortion and fail to render distinct infrared information at the edges of the house. In contrast, our baselines clearly preserve the infrared details of the eaves and successfully integrate the two modalities without introducing any artifacts.

For rows 5 to 6, the red boxes highlight roadside tree trunks and markers, while the yellow boxes enclose power lines and leaves. Eight methods, except CrossFuse, fail to achieve a satisfactory balance between texture color information and infrared thermal radiation. This failure results in either distorted power lines or blurred edges. Furthermore, the CrossFuse method does not render the infrared information of the tree trunks in the red box with sufficient distinctness. In contrast, our methods excel in both of the aforementioned aspects. In summary, by comparison, our baselines trained by proposed training paradigm demonstrate superior performance in preserving both visible texture and infrared thermal radiation information.