BusReF: Infrared-Visible images registration and fusion focus on reconstructible area using one set of features

Zeyang Zhang,Hui Li,Tianyang Xu,Xiaojun Wu
Jiangnan University
No.1800 Lihu Avenue, Wuxi City, Jiangsu Province, China
zzy_jnu_cv@163.com

Josef Kittler
University of Surry
Guildford, Surrey GU2 7XH, United Kingdom
j.kittler@surrey.ac.uk

Abstract

In a scenario where multi-modal cameras are operating together, the problem of working with non-aligned images cannot be avoided. Yet, existing image fusion algorithms rely heavily on strictly registered input image pairs to produce more precise fusion results, as a way to improve the performance of downstream high-level vision tasks. In order to relax this assumption, one can attempt to register images first. However, the existing methods for registering multiple modalities have limitations, such as complex structures and reliance on significant semantic information. This paper aims to address the problem of image registration and fusion in a single framework, called BusRef. We focus on Infrared-Visible image registration and fusion task (IVRF). In this framework, the input unaligned image pairs will pass through three stages: Coarse registration, Fine registration and Fusion. It will be shown that the unified approach enables more robust IVRF. We also propose a novel training and evaluation strategy, involving the use of masks to reduce the influence of non-reconstructible regions on the loss functions, which greatly improves the accuracy and robustness of the fusion task. Last but not least, a gradient-aware fusion network is designed to preserve the complementary information. The advanced performance of this algorithm is demonstrated by comparing it with different registration/fusion algorithms.

1 Introduction

Image fusion is an important technique in the field of computer vision. The main purpose of image fusion is to generate fused images by integrating complementary information from multiple source images of the same scene, with the aim to improve the performance of downstream high-level semantic vision tasks [14, 24, 36].

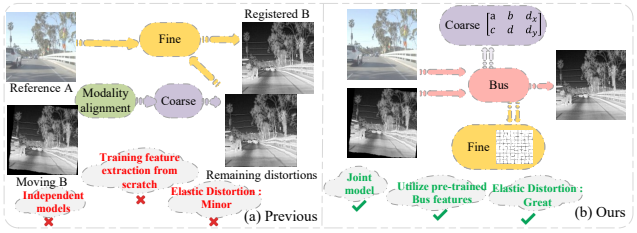

Figure 1. Comparison of frameworks. Existing serial training methods and our proposed Bus like training.

Typically, image fusion involves multiple sensors to capture scene data and the use of fusion algorithms to integrate the complementary information [23]. The normal prerequisite for multi-modal image fusion is that the input image pairs are strictly aligned. The impact of misalignment is severe ghosting [34]. However, in most scenarios, due to the difference between the internal and external parameters of infrared cameras and digital RGB cameras, it is very hard to obtain strictly aligned multi-modal image pairs. To remedy this situation, it is crucial to perform the registration of multiple images before the fusion task.

Due to the difference in imaging principles between infrared and digital cameras, feature-point matching-based registration algorithms often achieve unsatisfactory results. In most cases, the gradients of the salient target edges have a negative correlation between the two modalities, and sometimes the salient targets are even lost due to factors such as

smoke occlusion. In these cases, most similarity operators lose effectiveness and even mislead the image registration.

In general, there are three existing solutions for Infrared - Visible image registration:

- Specially designed similarity loss [13]
- Search for identical information between modalities [30, 34, 35]
- Semantic based methods [33]

Most of the existing algorithms contain multiple stages, with no interdependence between them. For example, a method may include a modality alignment module, a coarse registration module, and a fine registration module, which are trained [30, 33–35] independently. However, this one-at-a-time training approach from scratch is very unsatisfactory, as it limits the feature extraction capabilities of each model. It ignores the fact that the image registration task often requires fine-grained structural descriptions [35].

In this paper, we design a unified framework for multi-modal image fusion, linking coarse and fine registration modules through a backbone. The unique features of our framework are shown in the comparison with conventional approaches in Figure 1. The representation extracted by the backbone network provides the fine-grain features to sub-modules for registration. Specifically, an Auto-Encoder (AE) framework [15, 17] is trained as the bus of our proposed algorithm. Once the training is completed, the parameters of AE are fixed, and the design proceeds with training the coarse registration module [19], fine registration module [1], and multi-modal image fusion module [4, 5, 16, 18] mounted on the bus (BusReF).

In our approach, we also address the problem of fusing unaligned multi-modal images subject to content inconsistencies. In many existing algorithms, one assumes that the information from one modality is used as a reference and the other as a moving image [13]. Under this assumption, the registration methods tend to resample the moving image on a grid to match the reference image. However, due to the difference in the perspective projections of multimodal cameras, resulting in incongruent magnifications, some of the information in the reference images, compared to the moving images, during the un-warping process may be missing. It is worth noting that most of the "missing information" continuously affects the reference image, constituting "non-reconstructible" regions. Existing algorithms often ignore the impact of these "non-reconstructible" regions on the registration problem, and their model will forcibly attempt to register these parts regardless. To address this problem, we propose a reconstructible mask and apply it to the loss function during the training process, which greatly improves the registration ability of the model and reduces the risk of false matches.

Finally, in order to improve the performance further, the task of image fusion and registration should be in a mutually reinforcing relationship [35]. We propose a gradient-aware fusion network(GAF) and use it as guidance during the training phase of image registration.

2 Related Work

Multi-modal image registration has been widely discussed mainly in the field of medical images [6, 11, 41]. Due to the differences in imaging principles between modalities, the appearance of a target rendered by different modalities may vary greatly. In general, solving the image registration problem involves two major approaches: i) finding similar features between modalities and computing a spatial transformation based on them, ii) learning a spatial transformation to maximise the similarity between the modalities.

2.1 Multi-Modality Registration

Nowadays there are many methods based on feature matching. SIFT [21] is used as a local feature description operator. LoFTR [29] innovatively combines CNN and Transformer for feature matching [7, 20, 26, 27]. MatchFormer [31] leverages the powerful global feature interaction capabilities of the Transformer to perform multi-scale feature matching through a hierarchical structure. However, the aforementioned algorithms are designed for single modality image pairs and the performance will be seriously degraded when directly applied to processing multi-modal unaligned image pairs. Due to the different imaging principles of different sensors, the corresponding regions of different images may exhibit only a weak correlation. When it comes to Thermal Infrared–Visible (TIR) image pairs, the gradients of salient target regions at the edges are likely to show a negative correlation. Sometimes the target will be completely invisible. Semantics led all (SemLA) [33] achieves robust feature matching in simple scenes by using the results of semantic segmentation and restricting the feature matching to significant semantic target regions.

Many works seek a representation that is shared by multiple modalities. Cross-modality Perceptual Style Transfer Network (CPSTN) [30] was the first method to apply CycleGAN [42] to perform style transformation and reduce inter-modality differences. Mutually Reinforcing Multi-modal Image Registration and Fusion (MURF) [25, 35] has a similar idea. The only difference is that it uses contrastive learning to learn a robust and unified description of multi-modality images.

The quest for a unified description of multi-modal data is challenging. In most cases the intermediate features are less detailed, making it difficult to achieve high accuracy in learning multi-modal registration. Another promising approach has been suggested by Li et al. [19] who directly and

dynamically align the features of the convolutional network and depends on these features for image reconstruction.

However, all existing methods suffer from low coupling between the modules, task fragmentation and poor feature reusability. Finding intermediate features, performing a coarse registration, refining it and finally conducting image fusion are all formulated as unrelated multiple tasks, with multiple models to solve (rather than sub-modules within a model), and trained separately. This serial training approach makes it necessary for each of the models to learn feature descriptions from scratch, and it is difficult to ensure that the features learned by each model are optimal for different scenarios.

2.2 Multi-Modality Fusion

Multi-modal image fusion requires a high-precision alignment of image pairs. Otherwise, artefacts will appear in the fused image, hindering subsequent high-level semantic tasks [40]. The fusion algorithm should be able to preserve complementary information between multi-modal images and present it clearly in the fused image. For infrared-visible fusion, the widely accepted definition of complementary information is that the fused image should reflect the texture details of the visible image and the highlighted salient targets in the infrared image [39].

In deep learning-based image fusion algorithms, DenseFuse [14] pioneered the application of Auto-Encoder, while encouraging feature reuse and specially designed feature fusion rules to improve the performance of image fusion dramatically. In recent years, end-to-end fusion also has gained wide attention due to its limited dependence on human heuristics, and its high robustness.

In this paper, a multi-modal image registration and fusion network that imitates a bus structure is proposed. Specifically, our multi-modality registration network has an Auto-Encoder as a bus, and the sub-modules required for registration are mounted on the bus one by one for cooperative training. Since the Auto-Encoder requires features that can reconstruct the original image, they are inevitably rich in texture details, and therefore well suited for the registration task. Other innovations of our approach are the proposed reconstructible region mask, a gradient-aware fusion network, and a fusion strategy for registration training, designed to provide more robust registration and better visual results for unaligned multi-modal image fusion.

3 Methods

3.1 Reconstructible Mask

Due to the differences in intrinsic and extrinsic parameters of different sensors, it is normal that captured image pairs

Figure 2. (a) Original image. The red mask represents the unreconstructible area, the green represents our reconstructible content mask. (b) Artificially generated unaligned image. (c) Ground truth registration results. (d) Possible artifacts of not applying the reconstructible mask.

are not fully registered. As shown in Figure 2, the red portion represents the area of missing information in the moving image.

In this work, we propose the use of a reconstructible mask, which is used to guide the network to focus on image information that can be registered. Let us assume that I_x and I_y are a strictly registered multi-modal image pair. θ is a randomly generated 6-dof affine transformation parameters [12] to simulate the rigid deformation between the images, while the deformation field ϕ is to simulate the elastic deformation [10]. See the appendix for details of the generation process.

$$\theta = \begin{bmatrix} a & b & d_x \\ c & d & d_y \end{bmatrix}, \theta^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} \begin{bmatrix} -d_x \\ -d_y \end{bmatrix} \tag{1}$$

$$\phi^{-1} = -\phi \tag{2}$$

Then the spatial mapping between the images can be defined by an affine transformation $S()$. The spatially transformed image I_y is $I_y^R = S(I_y, \theta)$. The deformation field can be modelled by elastic deformation, denoted as $E()$, the elastic transformed I_y^R is $I_y^{RF} = E(I_y^R, \phi)$. Here $S()$ and $E()$ are grid sample functions. By this formulation, we have obtained an artificially unaligned image pair and forward and backward transforms. For strictly aligned images, the reconstructible region would be the full image. Hence, the mask M can be initialised as a matrix of ones, of size I_x, I_y. The forward transformation is applied to M:

$$\begin{aligned} M^{RF} &= E(S(M, \phi), \theta), \\ M' &= E(S(M^{RF}, \phi^{-1}), \theta^{-1}), \\ \bar{M} &= \begin{cases} 1, & M' > 0, \\ 0, & else. \end{cases} \end{aligned} \tag{3}$$

where $\bar{M}$ represents the reconstructible area between the reference images and moving images.

3.2 Bus Like Training Strategy

Multi-modality image registration requires the inputs to contain detailed information, which is difficult to achieve by training each module from scratch. As shown in Figure 3, Auto-Encoder is used as the bus in our framework, the features are extracted by Encoder from source images which are also reconstructed by Decoder with these features. This processing can ensure that all the detailed information is retained in the extracted feature maps. The Auto-Encoder as a reconstructor contains three down-sampling modules, four residual feature extraction modules (RFE) and three up-sampling modules. Four RFEs are denoted as $\{R_1 \sim R_4\}$, and their outputs are $\{F_1 \sim F_4\}$, representing the features from shallow to deep, respectively. The loss function for the reconstructor is composed of the l_2 norm based pixel loss and structural similarity index measure(SSIM) loss [32],

$$\mathcal{L}_{re} = \|I - I_{re}\|_2 + \lambda(1 - SSIM(I, I_{re})) \tag{4}$$

where λ is a balancing hyperparameter, making the two loss functions to be of similar magnitude, I_{re} is the reconstructed image.

To make the registration server better for multi-modal image fusion, the fusion module should be trained in coordination with the registration modules to enhance the performance of each module. When edges in the multi-modal images perfectly overlap in the fused image, it represents a highly accurate alignment.

Gradient information tends to reflect edges in the images. The greater the consistency between the edges, the better the registration. Therefore the fusion network should be able to gauge the gradient. As shown in Figure 4, the spatial attention is utilized to establish the gradient map, based on the existing gradient-weighted fusion network [35]. The loss function of the fusion network includes a weighted SSIM $\mathcal{L}_{wsim}$ and gradient loss $\mathcal{L}_{grad}$.

$$\mathcal{L}_{wsim} = 1 - \frac{SSIM(I_f, I_v) + SSIM(I_f, I_r)}{2} \tag{5}$$

where I_f is the fused image, I_v is the visible light image and I_r is the infrared light image.

$\mathcal{L}_{grad}$ requires I_f and I_t to have similar gradient information:

$$\mathcal{L}_{grad} = \|\nabla I_f - \nabla I_t\|_2 \tag{6}$$

where ∇ is the laplacian operator. The target gradient we want to retain is ∇I_t:

$$\begin{aligned} \nabla I_t &= w\left(\frac{\nabla I_v \cdot |\nabla I_v|^{\gamma}}{|\nabla I_v|}\right) + (1-w)\left(\frac{\nabla I_r \cdot |\nabla I_r|^{\gamma}}{|\nabla I_r|}\right) \\ s.t. \quad w[i,j] &= \begin{cases} 1, & |\nabla I_x[i,j]| > |\nabla I_y[i,j]| \\ 0, & else. \end{cases} \end{aligned} \tag{7}$$

where γ is the enhancement factor, which is set to 0.7.

The loss function of the fusion network is the weighted summation of $\mathcal{L}_{wsim}$ and $\mathcal{L}_{grad}$

$$\mathcal{L}_{fuse} = \sigma\mathcal{L}_{wsim} + \mathcal{L}_{grad} \tag{8}$$

Once the reconstruction and fusion network is designed, we treat it as a bus for the framework and freeze the parameters in it. Then we mount Affine Net for learning coarse registration. The input unaligned image pairs are first processed $\{F_1 \sim F_4\}$ by the reconstructor and used as the input to Affine Net. Multiple down-sampling processes are included in Affine Net and before each down-sampling we use a large scale 7×7 convolution and a dynamic convolution to achieve a sufficiently large receptive field [3]. Finally, the global pooling and MLP are used to obtain the desired affine transformation parameter θ.

At the same time, we mount up Deformable Net for elastic deformation learning. The input of Deformable Net is $\{F_1 \sim F_4\}$. It is worth noting that the output layer of Deformable Net is a simple 3×3 convolution with 8 input channels and 2 output channels. We perform group convolution [37] on $\{F_1 \sim F_4\}$ to ensure that each F_i only outputs 2 channels. This is equivalent to each layer of the RFE giving an elastic deformation proposal. Finally, we use a convolutional layer to linearly weight these 4 deformation proposals to output the final Deformation Field.

$$\begin{aligned} I_y^{Rf} &= E(I_y^{RF}, A(\{F_1 \sim F_4\}_x, \{F_1 \sim F_4\}_y)), \\ I_y^{rf} &= S(I_y^{Rf}, D(\{F_1 \sim F_4\}_x, \{F_1 \sim F_4\}_y)) \end{aligned} \tag{9}$$

where $A()$ is Affine Net, $D()$ is Deformable Net and I_y^{rf} is the reversed transformation result. During the training of the registration module, $A()$ and $D()$ are simultaneously mounted on the bus, and the loss function at this stage is constituted by the masked NCC loss L_{MNCC} and the masked gradient loss L_{MG} [38].

$$\mathcal{L}_{MNCC} = -MNCC(I_y^{rf}, I_y, \bar{M}) \tag{10}$$

$$MNCC(x, y, \bar{M}) = \frac{\sum_{i}^{i\in\bar{M}=1}\sum_{j}^{j\in\bar{M}=1}(x_{i,j} - \bar{x})(y_{i,j} - \bar{y})}{\sqrt{\sum_{i}^{i\in\bar{M}=1}\sum_{j}^{j\in\bar{M}=1}(x_{i,j} - \bar{x})^2}\sqrt{\sum_{i}^{i\in\bar{M}=1}\sum_{j}^{j\in\bar{M}=1}(y_{i,j} - \bar{y})^2}} \tag{11}$$

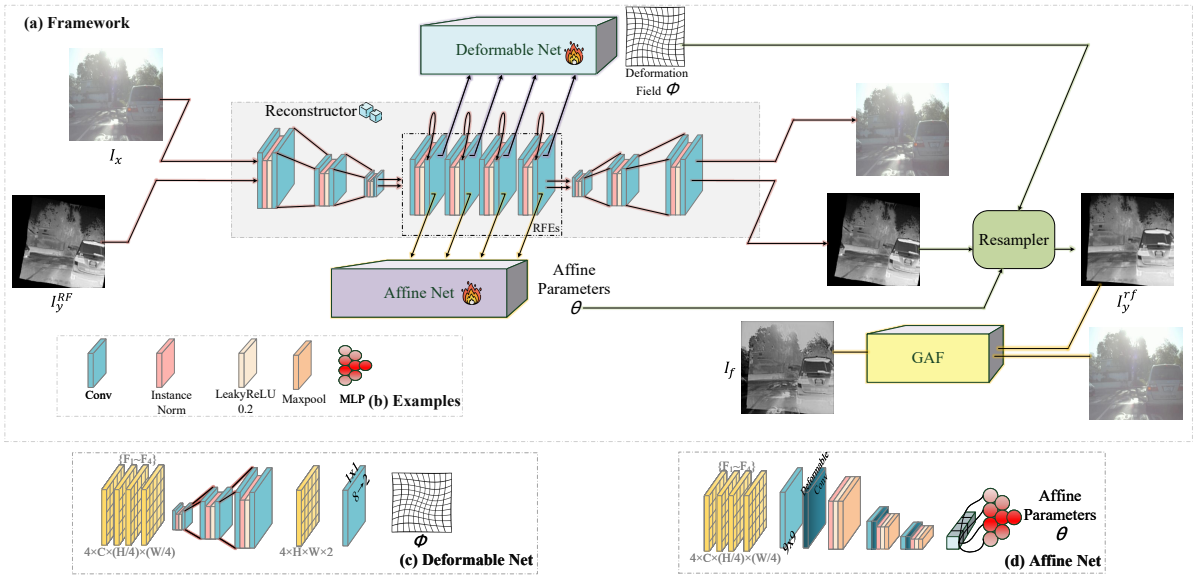

Figure 3. The Reconstructor is an Auto-Encoder framework that is trained by simultaneously inputting Infrared-Visible images and requiring the reconstruction of the input images to ensure the ability to extract multi-modal features. The registration module is mounted on this pre-trained framework to ensure the acquisition of detailed features. Finally, the affine transformation parameter θ and the deformation field ϕ corresponding to the rigid and elastic transforms are learnt. Finally, the mesh resampling is performed to achieve the image registration.

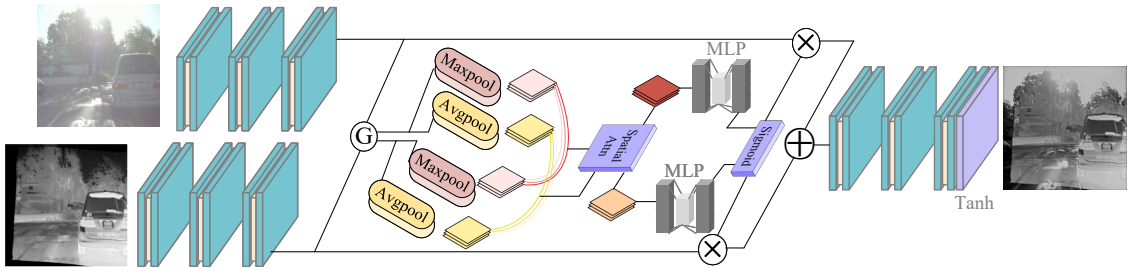

Figure 4. The architecture of GAF. The registered image pairs are inputted to the feature extractor, and gradient sensing is performed on the extracted features. G is the Laplacian operator. After separating the high-frequency part and low-frequency part by Maxpool and Avgpool respectively, two MLPs are used to learn inter-modality weighting. Finally, the fused image is obtained by output convolutional layers and Tanh activation.

where x, y indicate two images. $\bar{x}, \bar{y}$ denote their mean values.

$$\mathcal{L}_{MG} = \|(\nabla F(I_x, I_y^{rf}) - \nabla F(I_x, I_y)) \cdot \bar{M}\|_2 \quad (12)$$

where $F()$ is the GAF. The loss function to train registration is $\mathcal{L}_{reg}$

$$\mathcal{L}_{reg} = \epsilon \mathcal{L}_{MNCC} + \mathcal{L}_{MG} \quad (13)$$

4 Experiments

Unaligned multi-modality image registration and fusion is a very challenging task and there are a few competitive works to compare with BusReF. In this section, some single modality registration algorithms such as LoFTR and SIFT operator are chosen for comparison, using GAF to perform the fusion. The algorithms compared include also some representative state-of-the-art multi-modality registration and fusion algorithms such as SemLA and MURF.

4.1 Qualitative Comparison

The TIR registration and fusion results are given in Figure 5. The red and green edge salient maps are the superimposed gradient maps of the registered IR images and the human manually registered images.

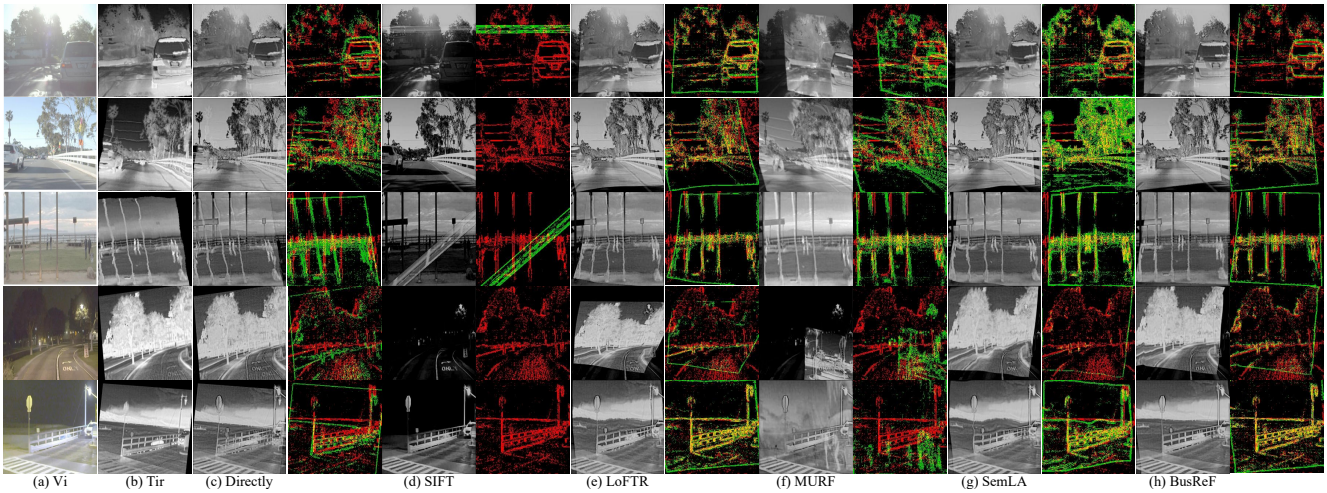

Figure 5. A qualitative comparison of fusion and registration on the TIR task (RoadScene) dataset. (a), (b) the original visible image and thermal Infrared image, respectively. (c) the directly fused results obtained by GAF. (d) SIFT+GAF. (e) LoFTR+GAF. (f), (g) the results obtained by MURF and SemLA. (h) the results of BusReF.

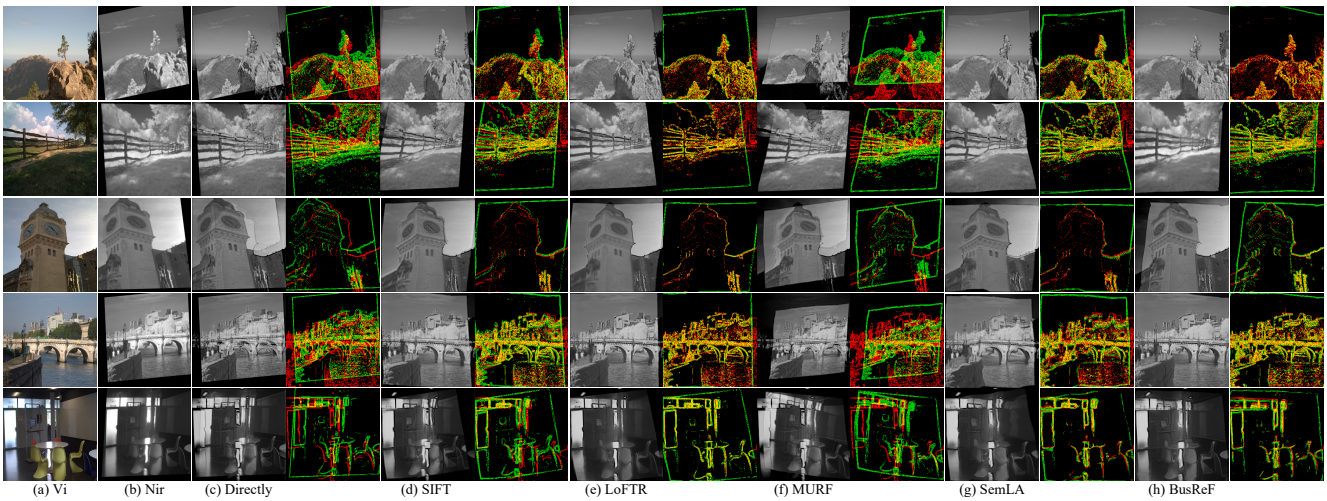

Figure 6. A qualitative comparison of fusion and registration on the NIR task. The setup of the experiment was kept consistent with Figure 5.

As seen in Figure 5 (c) and its edge salient maps, a direct fusion of unaligned multi-modality images produces many artefacts, resulting in the injection of additional noise instead. The SIFT operator and LoFTR are designed for unimodal unaligned images, where SIFT is completely incapable of solving the TIR-VI task, whereas LoFTR, thanks to the deep features extracted by Convolution and Transformer networks, is able to perform the registration in some of the cases.

Observing the results of the fourth row in Figure 5, the woods in the image are highlighted in the infrared modality and almost completely black and invisible in the visible modality, which shows a negative correlation trend. In contrast, the traffic lines and symbols on the road belong to the same highlighted information, and their gradients show a positive correlation.

It can be seen that LoFTR based on feature matching is able to register positively correlated information but not negatively correlated information. SemLA belongs to the family of algorithms based on feature matching, but some regions of the registered images are too distorted. The proposed method, BusReF, predicts the affine transformations in the TIR-VI task more accurately and can correct most of the elastic deformations. Moreover, BusReF is also able to

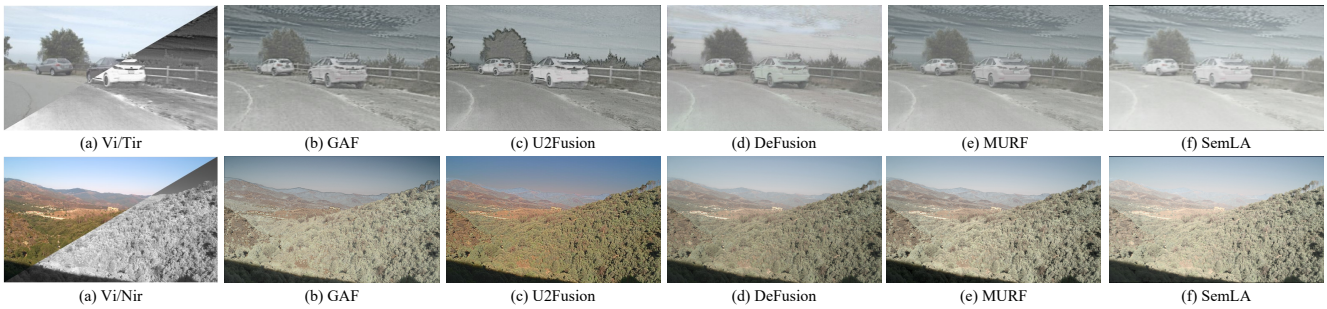

Figure 7. A qualitative comparison of the GAF with state-of-the-art multi-modality fusion algorithms. The left set is from the RoadScene dataset, showing a car parked by a railing, and the right set is from the NIRScene dataset showing a mountain scene. (a) the original images from the datasets (b) the GAF fusion result. (c),(d) the output of state-of-the-art unified image fusion algorithms. The (e),(f) The two works include both registration and fusion modules, we only show the results of their fusion modules here.

perform high-precision registration in dark scenes, such as the penultimate and penultimate rows of Figure 5.

Figure 6 shows the registration and fusion results of NIR-VI. Five representative scenes are selected for a qualitative comparison. Starting from the first row of Figure 6 shows "Mountain", "Country", "Old-Building", "Water", "Indoor" scenes. The differences between NIR-VI modalities are small, and there are fewer areas of negative correlation between the NIR-VI image pairs. Therefore most of the algorithms are able to complete the alignment to some extent. However, as seen by visualising the gradient salient maps, BusReF results in fewer artefacts and better edge overlap.

When evaluating the fusion performance of GAF for the thermal infrared modality, as can be seen from the image of the car parked by the guardrail in Figure 7, the fusion result of GAF has sharper edges and the clouds in the sky are well preserved. The thermal infrared information of the TIR modality is also well represented in the fused image.

We also give the NIR-VI fusion results of the mountains in Figure 7. Among all the compared algorithms, the GAF's fusion image has the clearest distant mountains and the infrared radiation information of the near mountains is well represented. The result obtained by U2Fusion produces a better view, but the distant mountains become transparent and invisible after fusion. The SemLA fusion was the brightest, but too close to the TIR image and the distant mountains were also less visible.

4.2 Quantitative Comparison

For objective evaluation of the BusReF registration capabilities, we give the results of quantitative experiments in Table 1.

The first two rows of these two tables show the metrics computed on a full-map scale, with the higher NCC representing a higher correlation between the registration results and those manual registration results in the dataset. BusReF is marginally better in NCC and MSE metrics (around 0.01) than SOTA.

Furthermore, observing the third and fourth rows of the two tables, MNCC and MMSE demonstrate that the calculation of these two metrics occurs only within the theoretically reconstructible area. In the TIR-VI task, the MNCC of BusReF is 0.1 higher than the SOTA with less variance, and BusReF leads in the NIR-VI task. Meanwhile, from Figure 2(c) and (d), a higher full-map NCC does not necessarily represent higher registration accuracy, and accordingly, the

Table 1. A comparison of the registrations of 221 image pairs on the RoadScene dataset and 469 image pairs on the NIRScene dataset. **Red** bold is the best, **blue** bold is the second.

RoadScene (TIR)	SIFT	LoFTR	SemLA (Matchformer)	MURF	BusReF
NCC	0.113±0.277	0.801±0.034	**0.820±0.023**	0.442±0.111	**0.876±0.028**
MSE↓	0.256±0.009	0.059±0.001	**0.040±0.004**	0.106±0.010	**0.042±0.005**
MNCC	0.073±0.024	0.755±0.019	**0.813±0.007**	0.373±0.071	**0.916±0.002**
MMSE↓	0.228±0.007	0.035±0.006	**0.025±0.000**	0.110±0.007	**0.011±0.005**
NIRScene (NIR)	SIFT	LoFTR	SemLA (Matchformer)	MURF	BusReF
NCC	0.345±0.117	0.832±0.036	**0.851±0.033**	0.796±0.042	**0.869±0.040**
MSE↓	0.125±0.018	**0.029±0.000**	**0.028±0.000**	0.097±0.002	0.030±0.000
MNCC	0.361±0.098	**0.647±0.034**	0.619±0.032	0.450±0.027	**0.897±0.005**
MMSE↓	0.117±0.142	0.353±0.006	**0.036±0.001**	0.078±0.001	**0.009±0.004**

Table 2. A comparison of the fusion of 221 image pairs on the RoadScene dataset and 469 image pairs on the NIRScene dataset.

RoadScene (TIR)	DeFusion	SemLA	U2Fusion	MURF	GAF
EI	35.06±5.7	48.98±7.6	50.49±10.9	**50.59±11.2**	**63.98±7.0**
CE↓	0.91±0.1	0.93±0.1	0.99±0.2	**0.81±0.3**	**0.79±0.2**
SF	9.16±1.1	**18.76±2.5**	12.37±1.8	15.36±3.6	**17.45±2.4**
FMI_w	0.27±0.0	**0.36±0.1**	**0.37±0.0**	0.26±0.0	0.24±0.0
Q_{cv}	172.31±64.1	**687.53±53.2**	255.47±25.5	497.67±43.6	**556.34±35.8**
NIRScene (NIR)	DeFusion	SemLA	U2Fusion	MURF	GAF
EI	57.97±12.3	**69.04±14.5**	64.77±8.2	67.37±15.6	**70.01±16.6**
CE↓	**0.75±0.1**	0.98±0.2	0.90±0.1	0.88±0.3	**0.72±0.3**
SF	18.05±5.7	**25.02±6.8**	21.40±8.5	20.43±6.2	**22.50±4.6**
FMI_w	**0.54±0.2**	0.41±0.1	0.34±0.0	0.37±0.0	**0.48±0.1**
Q_{cv}	659.47±160.3	845.23±210.5	**1192.17±126.4**	**946.89±266.2**	832.02±136.7

improvement of MNCC is more in line with human viewing.

The current SOTA fusion algorithms do not have multi-modality registration algorithms that can be used directly. In order to compare the performance of the fusion module GAF and the SOTA fusion algorithms, we input the strictly aligned image pairs from the original dataset into GAF and other fusion algorithms. The results are presented in Table 2.

To evaluate the performance of multi-modal image fusion, we targeted EI [22], SF [8], Q_{cv} [2], FMI_w [9] and CE [28] as evaluation metrics. EI and SF reflect the density and sharpness of the edges in the fused image, and Q_{cv} reflects the image quality of the local region, calculated as the mean square error of the weighted difference between the fused image and the source image. FMI_w denotes the degree of correlation between the two images, and CE analyses the degree of information retention between the fused image and the source image from an information theoretic perspective.

From Table 2, GAF achieves two best values and two second-best values in TIR-VI image fusion. The EI and SF show that the images fused by our algorithm have richer and sharper edges. CE achieving the first place represents that the fusion result of GAF is able to retain more information of the source image and no additional noise is introduced. However, the FMI_w metric appears not to be good in the fusion task of TIR-VI, but this is not supported by the results of the visualization. This is due to the fact that the fused image is able to reflect the information of both modalities at the same time. In the TIR-VI task, the value decreases when correlation calculations are performed with a single modality.

Referring to NIR-VI, GAF achieves second place in the FMI_w metric, which is not consistent with TIR-VI. This inconsistency arises from the magnitude of variability in the multi-modality data, with less variability between NIR-VI modes and more variability between TIR-VI modes. Thus the correlation between the fused images of the TIR-VI task and any single modality is small, whereas the correlation between the fused images of the NIR-VI task is considerable. This discrepancy reflects the ability of the GAF to embody multi-modal information simultaneously. The remaining metrics achieved two firsts and one second in the NIR-VI task, and the Q_{cv} metric exhibited a small gap compared to the remaining algorithms.

4.3 Ablation Study

We conducted ablation experiments investigating the three proposed improvement measures. In order to capture the impact of the method on the combined TIR and NIR tasks, the metrics we adopted were averaged over the two datasets.

Table 3. A quantitative comparison of full BusReF without the reconstructible region masks, no Bus training, and lack of fusion task guidance. The data in the table are averaged over the TIR and NIR datasets for the registration results.

Avg TIR & NIR	Full	Mask	Bus Training	Fusion Guide
NCC	**0.872**	0.852	0.574	0.829
MSE ↓	0.036	**0.014**	0.124	0.071
MNCC	**0.906**	0.638	0.418	0.839
MMSE ↓	**0.010**	0.087	0.201	0.036

First, we use the same training strategy, but without using the Reconstructible mask in the loss function for training. As seen in the third column of Table 3, NCC and MSE are almost indistinguishable from the full method, and the MSE metric is even better than the full method, but the MNCC and MMSE metrics decrease substantially. This means that networks trained without Reconstructible masks are likely to forcefully register the regions that cannot be reconstructed to optimise the loss functions. This can lead to regions that should be interested not being handled well.

In the second ablation study, the registration modules Affine Net and Deformable Net were removed from the bus for serial individual training. Specifically, the unaligned multi-modal image pairs are fed to the registration modules, which output the transformation parameters to register infrared images. As seen in the fourth column of Table 3, all four metrics show a substantial improvement, but the network is unable to complete registration. This is because the registration modules require a lot of capacity for learning an adaptive feature representation. This implies that the difficult task of multi-modality registration can not be learned.

Finally, we removed the guidance signal from the fusion module during the registration training process. As can be seen in the fourth column of Table 3 all the four metrics have undergone only a slight decrease and the network is still able to perform most of the registration. This suggests that the guidance signals, given by the fusion module, are mainly local details of the registration and are able to refine the BusReF performance.

5 Conclusion

In this paper, we proposed a bus-like architecture for training a multi-modal image registration and fusion network. The method, for the first time, unifies the features of multiple registration modules. The commonly used registration training and evaluation metrics are improved so that the network focuses on learning the registration and fusion in reconstructible regions. The experiments show that BusReF registered multi-modal images are more in line with the apriori knowledge based on human vision and are able

to achieve higher registration accuracy.

However, the proposed multi-modal image registration still has a lot of room for development. Although BusReF achieves feature reuse by mounting multiple registration modules to a bus, the gradient during training cannot be carried through the whole network. Perhaps end-to-end multi-modality registration algorithms will emerge in the future, which will further simplify the process and improve accuracy. Furthermore, is it possible for the results of multi-modality registration and fusion to be solved completely by end-to-end training? Although it is difficult, it is a very meaningful and promising topic. If a fully end-to-end output of registration and fusion can be achieved, it will be possible to find the correlation between the features for registration and the features for fusion.

References

[1] Moab Arar, Yiftach Ginger, Dov Danon, Amit H. Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[2] Hao Chen and Pramod K. Varshney. A human perception inspired quality metric for image fusion based on regional information. *Information Fusion*, 8(2):193–207, 2007. Special Issue on Image Fusion: Advances in the State of the Art. 8

[3] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[4] Chunyang Cheng, Xiao-Jun Wu, Tianyang Xu, and Guoyang Chen. Unifusion: A lightweight unified image fusion network. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2021. 2

[5] Chunyang Cheng, Tianyang Xu, and Xiao-Jun Wu. Mufusion: A general unsupervised image fusion network based on memory unit. *Information Fusion*, 92:80–92, 2023. 2

[6] Phu-Hung Dinh. Medical image fusion based on enhanced three-layer image decomposition and chameleon swarm algorithm. *Biomedical Signal Processing and Control*, 84: 104740, 2023. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] A.M. Eskicioglu and P.S. Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995. 8

[9] Mohammad Haghighat and Masoud Amirkabiri Razian. Fast-fmi: Non-reference image fusion metric. In *2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–3, 2014. 8

[10] Pierre Hellier, Christian Barillot, Etienne Mémin, and Patrick Pérez. Hierarchical estimation of a dense deformation field for 3-d robust registration. *IEEE transactions on medical imaging*, 20(5):388–402, 2001. 3

[11] Derek LG Hill, Philipp G Batchelor, Mark Holden, and David J Hawkes. Medical image registration. *Physics in medicine & biology*, 46(3):R1, 2001. 2

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 3

[13] Lingke Kong, X. Sharon Qi, Qijin Shen, Jiacheng Wang, Jingyi Zhang, Yanle Hu, and Qichao Zhou. Indescribable multi-modal spatial evaluator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9853–9862, 2023. 2

[14] 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. 1, 3

[15] Hui Li, Xiao-Jun Wu, and Tariq Durrani. Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Transactions on Instrumentation and Measurement*, 69(12): 9645–9656, 2020. 2

[16] 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

[17] 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. 2

[18] 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 Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11040–11052, 2023. 2

[19] Huafeng Li, Junzhi Zhao, Jinxing Li, Zhengtao Yu, and Guangming Lu. Feature dynamic alignment and refinement for infrared–visible image fusion: Translation robust fusion. *Information Fusion*, 95:26–41, 2023. 2

[20] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. In *Advances in Neural Information Processing Systems*, pages 17346–17357. Curran Associates, Inc., 2020. 2

[21] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[22] Xiaoqing Luo, Zhancheng Zhang, Cuiying Zhang, and Xiaojun Wu. Multi-focus image fusion using hosvd and edge intensity. *Journal of Visual Communication and Image Representation*, 45:46–61, 2017. 8

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

[24] 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. 1

[25] Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Natasa Sladoje. Comir: Contrastive multimodal image representation for registration. In *Advances in Neural Information Processing Systems*, pages 18433–18444. Curran Associates, Inc., 2020. 2

[26] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 2

[27] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *Computer Vision – ECCV 2020*, pages 605–621, Cham, 2020. Springer International Publishing. 2

[28] BK Shreyamsha Kumar. Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform. *Signal, Image and Video Processing*, 7:1125–1143, 2013. 8

[29] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8922–8931, 2021. 2

[30] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876*, 2022. 2

[31] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 2746–2762, 2022. 2

[32] 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

[33] Housheng Xie, Yukuan Zhang, Junhui Qiu, Xiangshuai Zhai, Xuedong Liu, Yang Yang, Shan Zhao, Yongfang Luo, and Jianbo Zhong. Semantics lead all: Towards unified image registration and fusion from a semantic perspective. *Information Fusion*, 98:101835, 2023. 2

[34] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19679–19688, 2022. 1, 2

[35] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023. 2, 4

[36] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021. 1

[37] Ting Zhang, Guo-Jun Qi, Bin Xiao, and Jingdong Wang. Interleaved group convolutions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 4

[38] Feng Zhao, Qingming Huang, and Wen Gao. Image matching by normalized cross-correlation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, pages II–II, 2006. 4

[39] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*, 2020. 3

[40] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5906–5916, 2023. 3

[41] Tao Zhou, Qi Li, Huiling Lu, Qianru Cheng, and Xiangxiang Zhang. Gan review: Models and medical image fusion applications. *Information Fusion*, 91:134–148, 2023. 2

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2