Equivariant Multi-Modality Image Fusion

Zixiang Zhao[1,2] Haowen Bai[1] Jiangshe Zhang[1] Yulun Zhang[3*] Kai Zhang[4]
Shuang Xu[5] Dongdong Chen[6*] Radu Timofte[2,7] Luc Van Gool[2,8]
[1]Xi'an Jiaotong University [2]ETH Zürich [3]Shanghai Jiao Tong University
[4]Nanjing University [5]Northwestern Polytechnical University [6]Heriot-Watt University
[7]University of Würzburg [8]INSAIT
zixiangzhao@stu.xjtu.edu.cn

Abstract

Multi-modality image fusion is a technique that combines information from different sensors or modalities, enabling the fused image to retain complementary features from each modality, such as functional highlights and texture details. However, effective training of such fusion models is challenging due to the scarcity of ground truth fusion data. To tackle this issue, we propose the ***E****quivariant* ***M****ulti-****M****odality im****A****ge fusion (****EMMA****) paradigm for end-to-end self-supervised learning. Our approach is rooted in the prior knowledge that natural imaging responses are equivariant to certain transformations. Consequently, we introduce a novel training paradigm that encompasses a fusion module, a pseudo-sensing module, and an equivariant fusion module. These components enable the net training to follow the principles of the natural sensing-imaging process while satisfying the equivariant imaging prior. Extensive experiments confirm that EMMA yields high-quality fusion results for infrared-visible and medical images, concurrently facilitating downstream multi-modal segmentation and detection tasks. The code is available at* https://github.com/Zhaozixiang1228/MMIF-EMMA.

1. Introduction

Multi-modality image fusion serves as an image restoration method that synthesizes information from multiple sensors and modalities to generate a comprehensive representation of scenes and objects [29, 38, 51, 54]. It finds widespread application in tasks such as image registration [13, 36, 39], scene information enhancement or restoration [7, 20, 42, 43, 53], and downstream tasks such as object detection [1, 21] and semantic segmentation [24, 33] in scenes with multiple sensors. Notable tasks include infrared-visible image fusion (IVF) and medical image fusion (MIF). IVF focuses on merging thermal radiation information from input infrared images and intricate texture details from input visible images, resulting in fusion images that mitigate the limitations of visible images affected by illumination variations and infrared images susceptible to low resolution and noise [48, 56]. The primary goal of MIF is to provide a comprehensive representation of any abnormalities in a patient's medical condition. This is accomplished by integrating multiple imaging techniques, thereby enabling an intelligent decision-making system that supports both diagnostic and therapeutic processes [12].

We assume that the underlying ground truth fused image is information-rich, but in practice we can only measure the same ground truth through different sensing processes which are typically nonlinear and difficult to model, thus obtaining observations in different modalities. Therefore, the multi-modality image fusion problem can be regarded as a challenging *nonlinear and blind* inverse problem, which can be regarded as the following negative log-likelihood minimization problem:

$$\min_{\boldsymbol{f}}\{-\log p\left(\boldsymbol{f} \mid \boldsymbol{i}_1, \boldsymbol{i}_2\right)\} \tag{1a}$$

$$\propto \min_{\boldsymbol{f}}\{-\log p\left(\boldsymbol{i}_1, \boldsymbol{i}_2 \mid \boldsymbol{f}\right) - \log p\left(\boldsymbol{f}\right)\} \tag{1b}$$

$$\propto \min_{\boldsymbol{f}}\{\mathcal{L}(\boldsymbol{f}, \boldsymbol{i}_1, \boldsymbol{i}_2) + \mathcal{R}(\boldsymbol{f})\} \tag{1c}$$

where $\boldsymbol{i}_1$, $\boldsymbol{i}_2$, and $\boldsymbol{f}$ represent two input source images and the output fusion image, respectively. Eq. (1b) originates from Bayes' theorem. In Eq. (1c), the first term is the *data fidelity term*, indicating that $\boldsymbol{i}_1$ and $\boldsymbol{i}_2$ are sensed from $\boldsymbol{f}$; the second term is the *prior term*, indicating that $\boldsymbol{f}$ needs to satisfy certain fusion image prior or empirical characteristics.

In the era of deep learning, numerous advanced methods strive to better model this problem. However, several pressing issues remain unaddressed in this task. For the first term of Eq. (1c), it is evident that individual sensors are limited to capturing modality-specific features; no singular "super" sensor exists that can perceive all modal information simultaneously in reality. Consequently, the absence of a

*Corresponding authors.

definitive ground truth hampers the effective application of deep learning's supervised learning paradigm to image fusion tasks. While generative model-based methods [21, 26] attempt to achieve fusion by making the source image and the fused image belong to a similar distribution, they suffer from a lack of interpretability, controllability, and present training challenges. On the other hand, methods based on manually crafted loss functions [18, 38, 51] often push the fusion image to resemble the source images by minimizing the ℓ_1 or ℓ_2 distance. However, such direct computation of $\|\boldsymbol{f}-\boldsymbol{i}_1\|+\|\boldsymbol{f}-\boldsymbol{i}_2\|$ to determine $\boldsymbol{f}$ neglects the potential domain differences between the fused images and the source images, failing to consider that $\boldsymbol{f}$ may not reside on the same feature manifold as $\boldsymbol{i}_1$ and $\boldsymbol{i}_2$. Meanwhile, for the second term of Eq. (1c), researchers often presuppose that the fused image exhibits certain structures, such as low-rank [17, 19], sparsity [6, 8], multi-scale decomposition [51, 54], *etc*., and impose priors to restrict the solution space. Nonetheless, due to that ground truth fused images are inaccessible, these priors typically depend on speculative assumptions about the fused images or extrapolations from natural image priors, thereby overly relying on domain knowledge and exhibiting limited adaptability to unseen scenarios.

In response to the challenges mentioned above, we plan to address them from two aspects. First, since aligning distributions and manually crafted loss functions are challenging tasks, we propose to start with the sensing and imaging processes. We aim to learn the sensing, or say, the inverse mapping from the fusion image back to images of various modalities. This approach is intuitively simpler than mastering the process of fusion itself. By doing so, we can measure the loss between the input source images and the (pseudo) sensing results, which are obtained by applying the fusion images to different sensing functions. This strategy overcomes the problem of not having ground truth images for fusion. Furthermore, as image fusion is an inherently ill-posed problem, merely optimizing the aforementioned sensing loss may not yield the optimum fused image. Consequently, we introduce a conceptually simple yet effective prior, which is based on the inherent priors of the imaging systems and does not rely on domain-specific knowledge of fusion images. This non-domain-specific prior is predicated on the understanding that natural imaging responses are equivariant to transformations such as shifts, rotations, and reflections. In other words, the transformed fused image, after sensing and re-fused, should yield the same outcome as before sensing. Leveraging the equivariance prior of the natural imaging system offers stronger constraints and guidance for the learning process within the fusion network. In summary, regarding the common learning paradigms for image fusion, we have made the following improvements:

$$\begin{aligned}&\|\boldsymbol{f}-\boldsymbol{i}_1\|+\|\boldsymbol{f}-\boldsymbol{i}_2\|+\text{fusion image prior}(\boldsymbol{f})\\ \Longrightarrow&\|\hat{\boldsymbol{i}}_1-\boldsymbol{i}_1\|+\|\hat{\boldsymbol{i}}_2-\boldsymbol{i}_2\|+\text{equivariance prior}(\mathcal{F}\circ\mathcal{A})\end{aligned}\tag{2}$$

where $\mathcal{F}$ represents the fusion model and $\mathcal{A}$ is the sensing model. $\hat{\boldsymbol{i}}_1=\mathcal{A}_1(\boldsymbol{f})$ and $\hat{\boldsymbol{i}}_2=\mathcal{A}_2(\boldsymbol{f})$ denote the respective sensing results for $\boldsymbol{i}_1$ and $\boldsymbol{i}_2$, as determined by their corresponding sensing models $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. $\mathcal{A}_1$ and $\mathcal{A}_2$ together comprise the sensing model $\mathcal{A}$.

Following this methodology, we devise a self-supervised learning paradigm named ***E**quivariant **M**ulti-**M**odality im**A**ge fusion* (**EMMA**). This framework consists of a fusion module, a pseudo-sensing module, and an equivariant fusion module. The fusion module, named U-Fuser, is a U-Net-like [30] structure that incorporates Restormer [45]-CNN blocks, and is employed to model both global and local features, thereby effectively aggregating information. The pseudo-sensing module, based on U-Net [30], is a learnable construct that maps the fused image back to the source images, simulating the natural process of perception imaging. Lastly, the equivariant fusion module is designed to ensure that fused images adhere to the established prior of equivariant imaging. Our contributions are as follows:

- We propose a novel self-supervised learning paradigm named EMMA, designed to address the absence of ground truth in image fusion. EMMA leverages the natural sensing-imaging process with the non-domain-specific prior that imaging responses are equivariant to transformations such as shift, rotation, and reflection.
- We refine the inappropriate handling of domain differences between fused images and source inputs in the conventional fusion loss by simulating the perceptual imaging process via pseudo-sensing module and the sensing loss component effectively.
- The U-Fuser fusion module proposed in EMMA proficiently models long- and short-range dependencies across multiple scales to integrate the source information.
- Our approach demonstrates excellent performance in infrared-visible image fusion and medical image fusion, which is also proved to facilitate downstream multi-modal object detection and semantic segmentation tasks.

2. Related Work

Multi-modality image fusion. In the deep learning era, multi-modality image fusion methods can be classified into four primary groups: generative models [26–28], autoencoder-based models [16, 18, 22, 24, 46], algorithm unrolling models [6, 8, 41, 52], and unified models [15, 37, 38, 47, 49]. Generative models represent the distribution of fused images and source images in the latent space through generative adversarial networks [26–28] or denoising diffusion model [55]. Autoencoder-based models use the encoder/decoder with CNN or Transformer block as the basic unit to model the mapping/inverse mapping between the image domain and the feature domain [20, 35, 54]. Algorithm unrolling models shift the algorithm focus from data-driven learning to model-driven learning, which replace

complex operators with CNN/Transformer blocks while retaining the original computational graph structure, achieving lightweight and interpretable learning [19, 52]. Unified models identify meta-knowledge between different tasks through cross-task learning, enabling rapid adaptation to new tasks and improved performance with fewer examples [38, 46]. Moreover, the multi-modality image fusion task is often integrated into coupled systems with upstream (pre-processing) image registration [13, 36, 39] and downstream object detection and semantic segmentation tasks [21, 23, 31, 33]. Image registration can effectively eliminate image artifacts and unaligned areas, enhance edge clarity and expand the perception field [11, 39, 40]. Furthermore, gradient of the recognition loss in downstream tasks can effectively guide the production of the fused image [21, 23, 33, 50].

Equivariant Imaging. Equivariant imaging (EI) [2–4] is an emerging fully unsupervised imaging framework that exploits the group invariance property in natural signals to learn a reconstruction function from partial measurement data alone. The main idea behind EI is to use the fact that natural signals often have certain symmetries. For example, images are often translation invariant, meaning that they look the same if they are shifted around. With this invariance prior, the whole imaging system (from sensing to reconstruction) is transformation equivariant. Under certain sensing conditions [32], the reconstruction function will be able to correctly reconstruct images that have been transformed around, even if it has never seen those images before. As a promising new approach to imaging and a new way to acquire and process images, EI has been shown to be effective for a variety of linear inverse problems [4]. This paper devotes to exploring the potential of EI on a more challenging task, *i.e.*, non-linear and blind inverse problems in multi-modality image fusion.

Comparison with existing approaches. a) Compared to the regular fusion loss, *i.e.* $\|\boldsymbol{f}-\boldsymbol{i}_1\|+\|\boldsymbol{f}-\boldsymbol{i}_2\|$ in the image or feature domains [18, 33, 51], the pseudo-sensing loss item in Eq. (2) from EMMA mitigates the irrationality in traditional loss caused by the manifold difference between $\boldsymbol{f}$ and $\{\boldsymbol{i}_1,\boldsymbol{i}_2\}$, ensuring that the distances calculated between $\{\hat{\boldsymbol{i}}_1,\boldsymbol{i}_1\}$ and $\{\hat{\boldsymbol{i}}_2,\boldsymbol{i}_2\}$ are within the same domain. **b)** Similar fusion-to-source mapping concepts [44, 46] aim to make $\boldsymbol{f}$ decomposable into $\{\boldsymbol{i}_1,\boldsymbol{i}_2\}$ to ensure it containing the source image information. However, their decomposition module, as an integral part of the fusion algorithm, undergoes updates during training, and the fusion output is considered as a feature for source reconstruction. Thus, proficiency in decomposition learning does not invariably correlate with enhanced information in fusion. In contrast, within the EMMA paradigm, the learning of the pseudo-sensing module is decoupled from that of the fusion network, and it remains frozen during EMMA training, thus ensuring that the mapping from the fused image back to the source image is explicit and determinate. This enhances the rationality and interpretability of the sensing module. **c)** Furthermore, other prior-based optimizations [19, 51] often necessitate domain knowledge of fusion images. However, in EMMA , we only need to use the imaging system prior rather than the fusion image prior to accomplish self-supervised learning.

3. Method

In this section, we first provide the model formalization, including the sensing module and the fusion module, and give the model hypotheses for establishing the equivariant image fusion paradigm. Then, we take the IVF task as an example and present the implementation details of EMMA. Other image fusion tasks can be analogously derived.

3.1. Problem Overview

Let $\boldsymbol{i}$, $\boldsymbol{v}$, and $\boldsymbol{f}$ refer to infrared, visible, and fused images, respectively, with $\boldsymbol{i}\in\mathbb{R}^{HW}$, $\boldsymbol{v}\in\mathbb{R}^{3HW}$, and $\boldsymbol{f}\in\mathbb{R}^{3HW}$. We assume the existence of an information-rich $\boldsymbol{f}$ that contains multi-sensory and multi-modal information and needs to be predicted. However, there is no perception device in real life that can fully sense $\boldsymbol{f}$ up to now. Thus, as an unsupervised task, there is no ground truth for $\boldsymbol{f}$. Therefore, we model the fusion process and the sensing process as follows:

$$\boldsymbol{f}=\mathcal{F}\left(\boldsymbol{i},\boldsymbol{v}\right)+\boldsymbol{n}_f \Leftrightarrow \boldsymbol{i}=\mathcal{A}_i\left(\boldsymbol{f}\right)+\boldsymbol{n}_i, \boldsymbol{v}=\mathcal{A}_v\left(\boldsymbol{f}\right)+\boldsymbol{n}_v, \quad (3)$$

where $\mathcal{F}\left(\cdot,\cdot\right)$ represents the fusion model, $\mathcal{A}_i\left(\cdot\right)$ and $\mathcal{A}_v\left(\cdot\right)$ represent the sensing model of $\boldsymbol{i}$ and $\boldsymbol{v}$, *i.e.*, the infrared and RGB cameras, respectively. In the traditional image inverse problem $\boldsymbol{y}=\mathcal{A}(\boldsymbol{x})+\boldsymbol{n}$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the ground truth image and the measurement, the degradation operator $\mathcal{A}(\cdot)$ is known (such as the noise distribution in denoising tasks and the blur kernel in super-resolution tasks). However, in image fusion, we cannot explicitly obtain $\mathcal{A}_i$ and $\mathcal{A}_v$. Nevertheless, we can set them learnable, in order to simulate the perceptual process and assist the network in self-supervised learning.

3.2. Model hypothesis

To provide comprehensive sensing and fusion models and further support the subsequent introduction of EMMA framework, we first need to establish some necessary hypotheses. **a) Measurement consistency.** We assume that the fusion function $\mathcal{F}\left(\cdot,\cdot\right)$ maintains consistency within the measurement domain, that is,

$$\mathcal{A}_i\left(\mathcal{F}\left(\boldsymbol{i},\boldsymbol{v}\right)\right)=\boldsymbol{i},\ \mathcal{A}_v\left(\mathcal{F}\left(\boldsymbol{i},\boldsymbol{v}\right)\right)=\boldsymbol{v}. \quad (4)$$

However, due to the underdetermined nature of the sensing process, the estimation of $\mathcal{F}\left(\boldsymbol{i},\boldsymbol{v}\right)$ cannot be achieved by estimating the inverse of $\mathcal{A}_i$ or $\mathcal{A}_v$, and we have to learn more information beyond the range space of their inverse.
b) Invariant set consistency. We first give two definitions in the equivariant imaging [4]:

Definition 1 (Invariant set)**.** *For a set of transformations* $\mathcal{G}=\left\{g_1,\ldots,g_{|\mathcal{G}|}\right\}$ *composed of unitary matrices* $T_g\in$

$\mathbb{R}^{n\times n}$, *$\mathcal{X}$ is the invariant set with respect to transformations $\mathcal{G}$, if $T_g x \in \mathcal{X}$ holds for $\forall x \in \mathcal{X}$ and $\forall g \in \mathcal{G}$, i.e., $T_g\mathcal{X}$ and $\mathcal{X}$ are identical.*

Definition 2 (Equivariant function)**.** *If function $\mathcal{I}$ satisfies $\mathcal{I}\left(T_g x\right) = T_g \mathcal{I}(x)$ for $\forall x \in \mathcal{X}$ and $\forall g \in \mathcal{G}$, we call $\mathcal{I}$ is an equivariant function with respect to the transformation $\mathcal{G}$.*

Regarding the corollary of Definition 1, if $\mathcal{X}$ represents a set of natural images, it is evident that the result remains natural images after transformations that include translations, rotations, and reflections. Hence, $\mathcal{X}$ is an invariant set for transformation group $\mathcal{G}$. Furthermore, the set composed of fused images $\boldsymbol{f}$, being a subset of $\mathcal{X}$, is also an invariant set to $\mathcal{G}$. Moreover, in Definitions 1 and 2, "invariance" pertains to the properties of the dataset, while "equivariance" characterizes the properties of the imaging system, meaning that the imaging system (denoted as $\mathcal{F}\circ\mathcal{A}$ in our paper) is the equivariant function with respect to $\mathcal{G}$. Consequently, we propose the following theorem:

Theorem 1 (Equivariant image fusion theorem)**.** *If we regard $\mathcal{I}$ in Definition 2 to be the composite function $\mathcal{F}\circ\mathcal{A}$, where $\mathcal{F}$ is the fusion model and $\mathcal{A}$ (including $\mathcal{A}_i$ and $\mathcal{A}_v$) is the sensing model, the equivariant image fusion theorem is:*

$$\mathcal{F}\left(\mathcal{A}_i\left(T_g\boldsymbol{f}\right), \mathcal{A}_v\left(T_g\boldsymbol{f}\right)\right) = T_g\mathcal{F}\left(\mathcal{A}_i\left(\boldsymbol{f}\right), \mathcal{A}_v\left(\boldsymbol{f}\right)\right). \quad (5)$$

Proof. Consider a set of natural images $\mathcal{X}$ satisfying the invariance property, by Definition 2 the imaging system $\mathcal{F}\circ\mathcal{A}$ should be equivariant to the group actions $\{T_g\}$. Hence, for $\forall \boldsymbol{f} \in \mathcal{X}$, we have $\mathcal{F}\circ\mathcal{A}(T_g\boldsymbol{f}) = T_g\mathcal{F}\circ\mathcal{A}(\boldsymbol{f})$. Furthermore, by separating $\mathcal{A}$ into $\mathcal{A}_i$ and $\mathcal{A}_v$, we can get Eq. (5). ■

Remark 1. *For Eq. (5), it does not necessitate $\mathcal{F}$ or $\mathcal{A}$ to be equivariant to T_g, instead, $\mathcal{F}\circ\mathcal{A}$ is required to be equivariant. Thus, $\mathcal{F}$ and $\mathcal{A}$ here can be set to any form of mapping without restriction.*

In the following , we will demonstrate how to establish our equivariant image fusion paradigm based on Theorem 1.

3.3. Equivariant image fusion paradigm

The main focus of this paper is to present EMMA, a self-supervised image fusion framework based on the equivariant imaging prior, with the specific workflow shown in Fig. 1.

Overall paradigm. Firstly, we establish a U-Net-like fusion module $\mathcal{F}(\cdot)$ named *U-Fuser*, which combines a Restormer [45] with CNN blocks as the basic unit to generate the fused image $\boldsymbol{f}$ from inputs $\boldsymbol{i}$ and $\boldsymbol{v}$. Subsequently, based on the equivariant image fusion theorem in Theorem 1, an equivariant prior-based self-supervised framework, comprising *U-Fuser module* and learnable *(pseudo) sensing modules* $\mathcal{A}_i$ and $\mathcal{A}_v$, is employed to better preserve the source image information in the absence of the fusion ground truth. Specifically, we transform $\boldsymbol{f}$, estimated by U-Fuser, through a series of transformations T_g to obtain $\boldsymbol{f}_t$, then pass $\boldsymbol{f}_t$ through pseudo sensing modules $\{\mathcal{A}_i, \mathcal{A}_v\}$ to obtain pseudo-images $\{\boldsymbol{i}_t, \boldsymbol{v}_t\}$. Finally, we fuse $\{\boldsymbol{i}_t, \boldsymbol{v}_t\}$ with U-Fuser again to obtain $\hat{\boldsymbol{f}}_t$.

Unlike other methods that require a well-designed loss function to minimize the distance between $\boldsymbol{f}$ and $\{\boldsymbol{i}, \boldsymbol{v}\}$, EMMA's loss focuses on making the pseudo-images $\{\mathcal{A}_i(\boldsymbol{f}), \mathcal{A}_v(\boldsymbol{f})\}$ generated by the sensing module from $\boldsymbol{f}$ as close to the original $\{\boldsymbol{i}, \boldsymbol{v}\}$ as possible, while making $\boldsymbol{f}_t$ close to $\hat{\boldsymbol{f}}_t$ simultaneously. Thus, from a natural imaging perspective, the optimal fusion image $\boldsymbol{f}$ is found.

In the following text, we will first introduce the fusion module U-Fuser $\mathcal{F}(\cdot)$ and the pseudo sensing modules $\{\mathcal{A}_i, \mathcal{A}_v\}$, then illustrate the entire self-supervised learning framework, and finally provide the training loss function.

U-Fuser module. We adopt a U-Net-like structure for fusing $\boldsymbol{i}$ and $\boldsymbol{v}$ and generating the fused image $\boldsymbol{f}$. At each scale, since the input cross-modal features contain both global features such as environment and background information, as well as local features like the highlighting and detailed texture object features, we design a Transformer-CNN structure to better model the cross-modal features by leveraging their respective inductive biases. For the selection of Transformer block, we adopt Restormer block [45], which implements self-attention in channel dimension to model global features without too much computation load. In the CNN block, we use Res-block [10]. The input features of the Restormer-CNN block are embedded and then parallelly processed by the Restormer block and the Res-block, followed by embedding interaction and a CNN layer, and finally input to the next scale. Features of $\boldsymbol{i}$ and $\boldsymbol{v}$ at the same scale are fused in the fusion layer, and are passed to the reconstruction branch at the previous scale via skip connections. Design of blocks for feature fusion and reconstruction is the same as Restormer-CNN block used in the feature extraction branch.

Pseudo sensing module. In contrast to other works in this field where their algorithm mainly focuses on the design of the fusion function $\mathcal{F}$, in this paper, we propose a self-supervised learning framework based on equivariant imaging prior to address the issue of lacking ground truth for fused images. According to the equivariant image fusion theorem stated in Theorem 1, we need to obtain pseudo imaging results from $\mathcal{A}_i(\boldsymbol{f})$ and $\mathcal{A}_v(\boldsymbol{f})$. To achieve this goal, we need to simulate the process of sensing infrared and visible images from the (imagined) fused image, as described in Eq. (4). Since it is not feasible to explicitly give the structures of $\mathcal{A}_i$ and $\mathcal{A}_v$, we adopt a data-driven learning approach to obtain them. Recently, many deep learning-based methods have shown promising results in image fusion. Therefore, we selected fifteen state-of-the-art (SOTA) methods that have recently appeared in top venues. They are DIDFuse [51], U2Fusion [38], SDNet [46], RFN-Nest [18], AUIF [52], RFNet [39], TarDAL [21], DeFusion [20], ReCoNet [11],

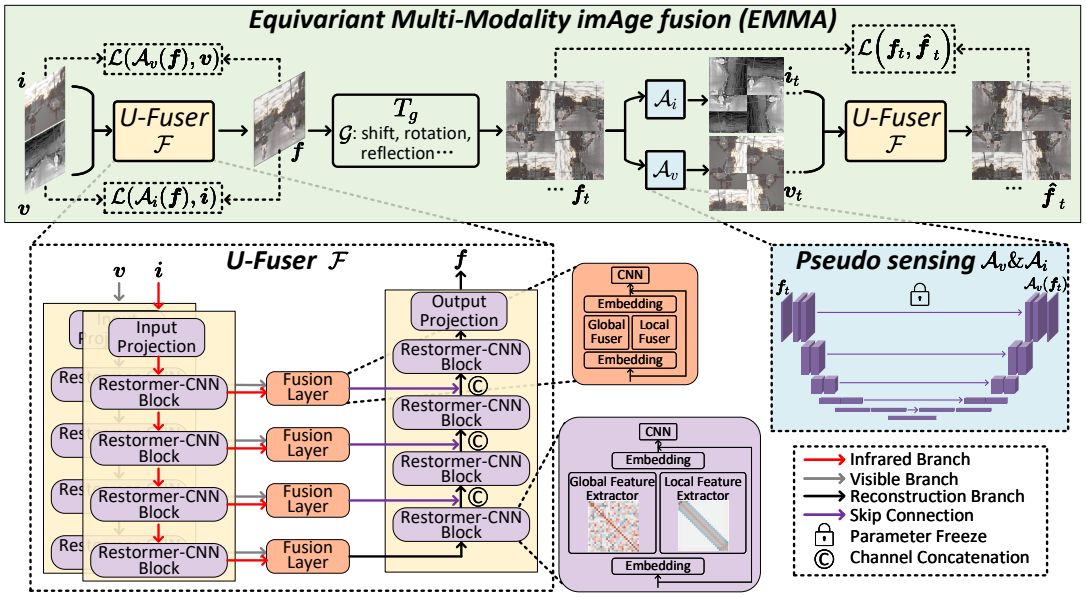

Figure 1. Workflow for EMMA. The image pair $\{\boldsymbol{i}, \boldsymbol{v}\}$ are initially input into U-Fuser $\mathcal{F}$, resulting in the fused image $\boldsymbol{f}$. Next, a series of transformations T_g containing shift, rotation, reflection, *etc.*, are applied to $\boldsymbol{f}$ to produce $\boldsymbol{f}_t$. $\boldsymbol{f}_t$ is then fed into the parameter-frozen $\{\mathcal{A}_i, \mathcal{A}_v\}$ to generate the pseudo-sensing images $\{\boldsymbol{i}_t, \boldsymbol{v}_t\}$, which are finally input into $\mathcal{F}$ to obtain the re-fused image $\hat{\boldsymbol{f}}_t$.

MetaFusion [50], CDDFuse [54], LRRNet [19], MURF [40], DDFM [55] and SegMIF [23]. We use their fusion results as the (pseudo) ground truth for the fused images and then learn the mappings from the fused images to $\boldsymbol{i}$ and $\boldsymbol{v}$, which can be regarded as $\mathcal{A}_i$ and $\mathcal{A}_v$, respectively. Considering that both the input and output of the mapping have the same image size, we choose U-Net [30] as the backbone of $\mathcal{A}_i$ and $\mathcal{A}_v$ and conduct the end-to-end training paradigm. The specific network details are in the supplementary material.

Equivariant image fusion. After obtaining the U-Fuser $\mathcal{F}$ and pseudo-sensing functions $\{\mathcal{A}_i, \mathcal{A}_v\}$, we introduce our self-supervised learning framework based on image equivariant prior. As shown in Fig. 1, we first input the image pairs $\{\boldsymbol{i}, \boldsymbol{v}\}$ into $\mathcal{F}$, and obtain fused image $\boldsymbol{f}$ (which is the entire operation of conventional fusion algorithms). Then, we apply a series of transformations T_g to $\boldsymbol{f}$, including shift, rotation, reflection, *etc.*, to obtain $\boldsymbol{f}_t$. Subsequently, $\boldsymbol{f}_t$ is input into the well-trained $\{\mathcal{A}_i, \mathcal{A}_v\}$ to obtain the pseudo-sensing images $\{\boldsymbol{i}_t, \boldsymbol{v}_t\}$, which contain the information from $\boldsymbol{f}_t$ and satisfy the imaging characteristics of infrared and visible images, respectively. Finally, paired $\{\boldsymbol{i}_t, \boldsymbol{v}_t\}$ are fed into $\mathcal{F}$ to obtain the re-fused image $\hat{\boldsymbol{f}}_t$. Throughout the framework, we aim to aggregate information from $\{\boldsymbol{i}, \boldsymbol{v}\}$ into $\boldsymbol{f}$, and according to the equivariant image fusion theorem (Theorem 1), $\boldsymbol{f}_t$ and $\hat{\boldsymbol{f}}_t$ should be sufficiently close. These will be guaranteed through the designed loss function.

Training detail and loss function. During the entire training process of EMMA, we first trained $\mathcal{A}_i$ and $\mathcal{A}_v$ using ℓ_2 loss as the loss function, *i.e.*, $\mathcal{L}_I^{Rec} = \ell_2\left(\boldsymbol{i}, \mathcal{A}_i(\tilde{\boldsymbol{f}})\right)$ and $\mathcal{L}_V^{Rec} = \ell_2\left(\boldsymbol{v}, \mathcal{A}_v(\tilde{\boldsymbol{f}})\right)$, where $\tilde{\boldsymbol{f}}$ are the fusion results from the SOTA methods in Sec. 3.3. Then, we freeze the parameters of $\mathcal{A}_i$ and $\mathcal{A}_v$, which means that parameters of the pseudo-sensing module will no longer be updated. Afterwards, we train U-Fuser module with the total loss function:

$$\mathcal{L}_{total} = \mathcal{L}\left(\mathcal{A}_i(\boldsymbol{f}), \boldsymbol{i}\right) + \alpha_1 \mathcal{L}\left(\mathcal{A}_v(\boldsymbol{f}), \boldsymbol{v}\right) + \alpha_2 \mathcal{L}\left(\boldsymbol{f}_t, \hat{\boldsymbol{f}}_t\right), \quad (6)$$

where $\mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \ell_1(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \ell_1(\nabla \boldsymbol{x}, \nabla \hat{\boldsymbol{x}})$. α_1 and α_2 are the tuning parameters, and ∇ indicates the Sobel operator. In particular, the first and second terms of Eq. (6) ensure our paradigm satisfies the *measurement consistency* of model hypothesis in Sec. 3.2, while the third term ensures it satisfies the *invariant set consistency* of model hypothesis.

3.4. Explanations

Here we will explain why the unsupervised fusion of EMMA works. By the fact that image set $\{\boldsymbol{f}\}$ is invariant to a group of invertible transformations $\{T_g\}$, give any image $\boldsymbol{f}$ from the invariant set $\{\boldsymbol{f}\}$, then $T_g\boldsymbol{f}$ also belongs to the set for all $g = 1, \cdots, |G|$. Under the equivariant theorem in Theorem 1, we have $\{\boldsymbol{i}, \boldsymbol{v}\} = \mathcal{A}\boldsymbol{f} = \mathcal{A}T_g T_g^{-1}\boldsymbol{f} = \mathcal{A}_g\boldsymbol{f}'$ for $g = 1, \cdots, |G|$, where $\mathcal{A}_g = \mathcal{A}T_g$ and $\boldsymbol{f}' = T_g^{-1}\boldsymbol{f}$ belongs to $\{\boldsymbol{f}\}$. That is to say, applying transformations is equal to generating multiple virtual sensing operators $\{\mathcal{A}_g\}_{g=1,\cdots,|G|}$. Since those virtual operators $\mathcal{A}_g$ are with potentially different nullspaces, this allows us to learn beyond the range space of inverse $\mathcal{A}$ (see [32]).

The lack of ground truth leads to potential inaccuracies in modeling $\mathcal{A}_i$ and $\mathcal{A}_v$, making the reconstruction of $\boldsymbol{f}$ potentially unsatisfactory in the first few training epochs. Fortunately, the combination of transformation for $\boldsymbol{f}_t$ and learning via equivariant imaging prior allows the completion

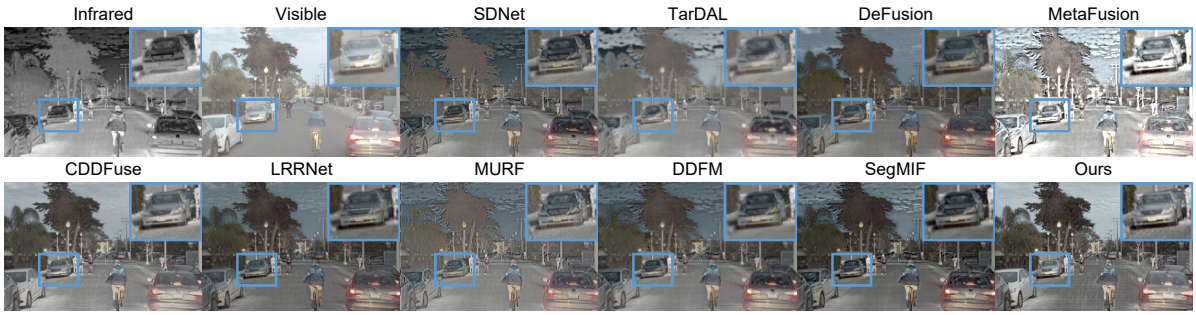

Figure 2. Visual comparison of "06832" from RoadScene [37] IVF dataset.

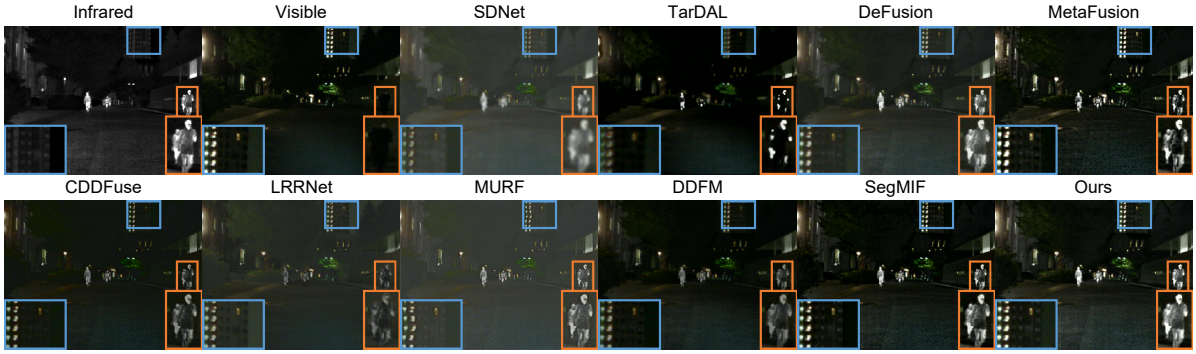

Figure 3. Visual comparison of "00782N" from MSRS [34] IVF dataset.

of the originally missing knowledge to calibrate and refine the fusion results, *i.e.*, achieving the recovering of the missed null space component. Notably, in the final algorithm deployment phase, only the fine-tuned U-Fuser $\mathcal{F}$ is needed, and all other modules will be disregarded, such as $\mathcal{A}_i$ and $\mathcal{A}_v$. Finally, the proposed equivariant fusion module differs from data augmentation (DA), which mainly extends data based on the ground truth. However, ground truth is firmly inaccessible in the image fusion task and DA cannot provide extra information gains when learning to image without ground truth [2, 4]. Fortunately, as we have shown, with the equivariance prior the proposed EMMA can provide extra information and figure out principle-plausible fusion results.

4. Experiment

4.1. Infrared and visible image fusion

Setup. We conduct experiments on three fashion benchmarks: MSRS [34], RoadScene [37] and M^{3}FD [21]. The network is trained on the MSRS training set and tested on its test set to evaluate the performance. In addition, the trained model is implemented to RoadScene and M^{3}FD without fine-tuning to verify the generalization performance. Our experiments are performed using PyTorch on a computer equipped with two NVIDIA GeForce RTX 3090 GPUs. The training image pairs are cropped into 128×128 patches randomly and with a batchsize of 8 before being fed into the network. α_1 and α_2 in Eq. (6) are set to 1 and 0.1, to ensure comparable magnitudes among the terms in the loss function. We train the network for 100 epochs using the Adam optimizer, with an initial learning rate of 1e-4 and decreasing by a factor of 0.5 every 20 epochs. U-Fuser is set to contain a four-layer structure. $\mathcal{A}_i$ and $\mathcal{A}_v$ are set as five-layer U-Nets [30]. They are pre-trained and parameter-frozen prior to the U-Fuser training. As for the transformation set $\mathcal{G}$, we will discuss it in our supplementary material.

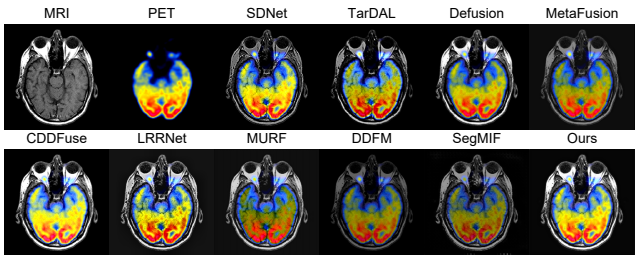

Figure 4. Visual comparison for MIF task.

SOTA methods and metrics. We compare EMMA with SOTA methods of IVF, including SDNet [46], TarDAL [21], DeFusion [20], MetaFusion [50], CDDFuse [54], LRRNet [19], MURF [40], DDFM [55] and SegMIF [23]. Six metrics are used to objectively compare fusion performance, including entropy (EN), standard deviation (SD), spatial frequency (SF), average gradient (AG), structure content dissimilarity (SCD) and visual information fidelity (VIF). Higher values indicate superior fusion effects and the calculation details are in [25].

Qualitative comparison. We compare the fusion outcomes of EMMA with SOTAs in Figs. 2 and 3. Our results successfully integrate thermal radiation information derived from infrared images with detailed texture features extracted from visible images. Fig. 2 shows that the fused image accurately

Infrared-Visible Image Fusion on MSRS Dataset [34]							Infrared-Visible Image Fusion on RoadScene Dataset [37]						
	EN ↑	SD ↑	SF ↑	AG ↑	SCD ↑	VIF ↑		EN ↑	SD ↑	SF ↑	AG ↑	SCD ↑	VIF ↑
SDN [46]	5.25	17.35	8.67	2.67	0.99	0.50	SDN [46]	7.30	44.06	14.58	5.80	1.37	0.61
TarD [21]	5.28	25.22	5.98	1.83	0.71	0.42	TarD [21]	7.26	47.44	11.11	4.14	1.40	0.56
DeF [20]	6.46	37.63	8.60	2.80	1.35	0.77	DeF [20]	7.36	47.03	10.99	4.38	1.62	0.63
Meta [50]	5.65	24.97	9.99	3.40	1.14	0.31	Meta [50]	6.88	31.97	14.38	5.57	0.92	0.55
CDDF [54]	6.70	43.38	11.56	3.73	1.62	**1.05**	CDDF [54]	7.52	54.42	14.97	5.81	1.65	0.66
LRR [19]	6.19	31.78	8.46	2.63	0.79	0.54	LRR [19]	7.12	39.16	11.41	4.37	1.46	0.45
MURF [40]	5.04	16.37	8.31	2.67	0.86	0.40	MURF [40]	6.91	33.34	13.88	5.37	1.04	0.52
DDFM [55]	6.19	29.26	7.44	2.51	1.45	0.73	DDFM [55]	7.24	42.43	10.68	4.15	1.64	0.62
SegM [23]	5.95	37.28	11.10	3.47	1.57	0.88	SegM [23]	7.29	46.14	14.47	5.57	1.61	0.65
Ours	**6.71**	**44.13**	**11.56**	**3.76**	**1.63**	0.97	Ours	**7.52**	**54.81**	**15.21**	**5.83**	**1.69**	**0.66**
Infrared-Visible Image Fusion on M^{3}FD Dataset [21]							**Medical Image Fusion on Harvard Dataset** [9]						
	EN ↑	SD ↑	SF ↑	AG ↑	SCD ↑	VIF ↑		EN ↑	SD ↑	SF ↑	AG ↑	SCD ↑	VIF ↑
SDN [46]	6.87	36.22	15.32	5.61	1.41	0.55	SDN [46]	3.79	52.53	21.91	5.51	0.87	0.52
TarD [21]	6.80	41.77	8.65	3.17	1.35	0.51	TarD [21]	4.74	55.73	18.02	5.35	0.86	0.31
DeF [20]	6.90	36.81	9.85	3.65	1.42	0.58	DeF [20]	4.00	57.48	17.09	4.19	0.84	0.59
Meta [50]	6.73	30.56	16.48	6.02	1.31	0.65	Meta [50]	3.90	65.18	**28.69**	**6.29**	1.33	0.54
CDDF [54]	7.04	42.02	16.56	5.84	1.41	0.65	CDDF [54]	4.13	68.46	21.58	5.83	1.61	0.66
LRR [19]	6.58	30.28	11.83	4.21	1.34	0.54	LRR [19]	4.15	45.71	17.39	4.47	0.23	0.51
MURF [40]	6.59	28.89	11.82	4.81	1.21	0.39	MURF [40]	4.42	36.35	27.18	5.98	0.35	0.37
DDFM [55]	6.82	32.68	10.07	3.71	1.35	0.60	DDFM [55]	3.97	59.81	16.43	4.11	1.49	0.63
SegM [23]	6.88	36.20	16.19	5.83	1.38	**0.75**	SegM [23]	3.67	57.79	21.91	5.56	1.05	0.66
Ours	**7.12**	**44.01**	**16.92**	**6.23**	**1.48**	0.66	Ours	**4.81**	**69.42**	22.15	6.02	**1.64**	**0.66**

Table 1. Quantitative results of IVF and MIF task. Best and second-best values are **highlighted** and underlined.

	Configurations	EN	SD	SF	AG	SCD	VIF
I	w/o Equivariant Loss	6.36	39.22	9.09	3.01	1.18	0.72
II	w/o Sensing Loss	6.42	40.12	9.12	3.29	1.24	0.79
III	w/ $\ell_1(\boldsymbol{f},\boldsymbol{i})+\ell_1(\boldsymbol{f},\boldsymbol{v})$ Loss	6.21	38.96	8.85	2.99	0.85	0.76
IV	Exp. III w/ augmentations	6.26	39.11	8.73	3.02	0.96	0.77
V	w/o Global Extractor	6.45	39.37	9.44	3.24	1.42	0.81
VI	w/o Local Extractor	6.52	40.49	9.79	2.82	1.46	0.82
	Ours	**6.71**	**44.13**	**11.56**	**3.76**	**1.63**	**0.97**

Table 2. Ablation experiment results. **Bold** indicates the best value.

captures the advantages of each modality while eliminating redundant information. The fusion process enhances object visibility, sharpens textures, and reduces artifacts. In Fig. 3, objects situated in inadequately illuminated surroundings are prominently highlighted with well-defined edges and abundant contours. This distinctiveness facilitates the differentiation between foreground objects and the background, thereby enhancing our comprehension of the depicted scene.

Quantitative comparison. The fusion outcomes are quantitatively compared using six metrics, as shown in Tab. 1. Our method exhibits remarkable performance across nearly all metrics, affirming its suitability for various environmental conditions and object categories. They indicate the capability of EMMA to produce images that align with human visual perception while preserving the integrity of the source image features and producing informative fused images.

4.1.1 Ablation studies

We conduct ablation studies on the MSRS testset to prove the rationality of EMMA, with the results shown in Tab. 2.

Terms in loss function. In Exp. I, we eliminate the last term in Eq. (6), which is the equivariant term. Even though the fusion module is capable of completing image fusion, it is unable to constrain the solution space through the equivariant prior. Thus, the network yields weaker results. In Exp. II, we modified the first two terms of Eq. (6) to be $\ell_1(\boldsymbol{f},\boldsymbol{i})+\ell_1(\boldsymbol{f},\boldsymbol{v})$, which is the traditional loss in other fusion tasks. The first two terms of Eq. (6), *i.e.*, sensing loss, guarantee that the fused image needs to inherit enough information from source images, so that the output pseudo-perceptual imaging result can be closer to the source images. While the traditional loss function purely forces the fused image to closely resemble the source images. Results in Exp. II demonstrate the necessity of sensing loss term. In Exp. III, we replace the loss in Eq. (6) with that in Exp. II. The results indicate that without equivariant loss and sensing loss, relying on $\ell_1(\boldsymbol{f},\boldsymbol{i})+\ell_1(\boldsymbol{f},\boldsymbol{v})$ loss makes it difficult to achieve an ideal fusion network. In Exp. IV, to further demonstrate our claim, we employ the same transformation as EMMA for conducting data augmentation (DA) on input images $\boldsymbol{i}$ and $\boldsymbol{v}$, expanding upon the ablation experiment Exp. III. That is, we employ the same transformation group as EMMA on the original network input, and the fusion training framework follows traditional approaches. Specifically, the loss function becomes: $\|\boldsymbol{f}-\boldsymbol{i}\|+\|\boldsymbol{f}-\boldsymbol{v}\|+\|\boldsymbol{f}_t-\boldsymbol{i}_t\|+\|\boldsymbol{f}_t-\boldsymbol{v}_t\|$, where $\boldsymbol{f}_t=T_g\boldsymbol{f}$. Experimental results demonstrate that under the same transformation, there is only a slight improvement for DA on $\boldsymbol{i}$ and $\boldsymbol{v}$. Conversely, in comparison to EMMA, substantial differences in effectiveness are observed. Thus, our equivariant fusion module fundamentally differs from traditional DA, as DA cannot provide additional information gains when learning to image without ground truth.

U-Fuser. Then, in Exp. V and Exp. VI, we separately eliminated the Restormer-block or the Res-block, ensuring a consistent number of parameters by increasing the remaining

	Bus	Car	Lam	Mot	Peo	Tru	mAP@0.5
IR	78.8	88.7	70.2	63.4	80.9	65.8	74.6
VI	78.3	90.7	86.4	69.3	70.5	70.9	77.7
SDN [46]	81.4	92.3	84.1	67.4	79.4	69.3	79.0
TarD [21]	81.3	**94.8**	87.1	69.3	81.5	68.7	80.5
DeF [20]	82.9	92.5	**87.8**	69.5	80.8	71.4	80.8
Meta [50]	83.0	93.4	87.3	74.8	81.6	68.8	81.5
CDDF [54]	81.8	92.9	87.6	72.8	81.8	72.9	81.6
LRR [19]	80.1	92.3	86.2	73.6	78.3	68.6	79.9
MURF [40]	81.3	92.6	86.5	70.8	80.2	69.9	80.2
DDFM [55]	82.2	93.2	87.6	68.4	81.0	71.3	80.6
SegM [23]	81.8	93.1	86.8	72.3	79.9	70.9	80.8
Ours	**83.2**	93.5	87.7	**77.7**	**82.0**	**73.5**	**82.9**

Table 3. AP@0.5(%) for MM detection.

	Unl	Car	Per	Bik	Cur	CS	GD	CC	Bu	mIOU
IR	90.5	75.6	45.4	59.4	37.2	51.0	46.4	43.5	50.2	55.4
VI	84.7	67.8	56.4	51.8	34.6	39.3	42.2	40.2	48.4	51.7
SDN [46]	97.3	78.4	62.5	61.7	35.7	49.3	52.4	42.2	52.9	59.2
TarD [21]	97.1	79.1	55.4	59.0	33.6	49.4	54.9	42.6	53.5	58.3
DeF [20]	97.5	82.6	61.1	62.6	40.4	51.5	48.1	47.9	54.8	60.7
Meta [50]	97.3	81.6	61.2	62.1	37.2	52.9	59.8	46.2	56.2	61.6
CDDF [54]	**97.8**	82.5	63.2	62.2	40.8	52.7	56.2	45.3	58.7	62.2
LRR [19]	97.4	81.2	62.4	61.9	40.3	50.7	48.1	45.3	47.3	59.4
MURF [40]	97.2	81.4	62.0	60.9	39.7	52.3	55.5	46.8	56.1	61.3
DDFM [55]	97.4	82.5	60.4	62.0	41.7	52.9	56.2	46.3	53.7	61.2
SegM [23]	97.6	**84.6**	64.8	**63.6**	40.2	52.9	59.9	49.4	56.2	63.2
Ours	97.6	84.0	**65.2**	63.1	**42.4**	**53.6**	**60.2**	**50.5**	**56.3**	**63.7**

Table 4. IoU(%) for MM segmentation.

blocks number. The results demonstrate that an incomplete feature extraction module leads to deficiencies in modeling local texture details or capturing long-range dependencies, thereby resulting in a degradation of performance.

4.2. Downstream IVF applications

This section aims to examine the impact of image fusion on downstream vision tasks. We assess the performance of fusion results in both *multi-modal semantic segmentation* (MMSS) tasks and *multi-modal object detection* (MMOD) tasks. To ensure fairness, we individually re-train the network for each task using fusion results obtained from their own methods. Due to space limitations, the visual comparisons are placed in the supplementary material.

Infrared-visible object detection. MMOD task is conducted on the M^3FD dataset [21], which comprises 4200 images encompassing six categories of labels: people, cars, buses, motorcycles, trucks, and lamps. We partition M^3FD dataset into training/validation/test sets in an 8:1:1 ratio. YOLOv5 detector [14] is trained using the SGD optimizer for 400 epochs. Batch size is 8 and the initial learning is 0.01. We evaluate the detection performance by comparing the mAP@0.5. Tab. 3 indicates that EMMA exhibits the most superior detection capabilities, enhancing the detection accuracy by merging thermal radiation and RGB information and emphasizing hard-to-detect objects.

Infrared-visible semantic segmentation. MSRS dataset [34] is designed for MMSS task and encompasses nine categories of pixel-level labels: background, bump, color cone, guardrail, curve, bike, person, car stop, and car. We select DeeplabV3+ [5] as the segmentation network and value the performances via Intersection over Union (IoU). The division of training and test sets adheres to the protocol in the original dataset paper [34]. We employ the cross-entropy loss along with the SGD optimizer. The total number of epochs is 340 while the backbone is frozen for the first 100 epochs. The batch size and the initial learning rate are set to 8 and 7e-3, and the learning rate follows cosine annealing delayed as the epoch number increases. Segmentation outcomes are displayed in Tab. 4. EMMA effectively combines the edge and contour details present in the source images, thereby improving the model's capability to recognize the object's boundary, and leading to more precise segmentation.

4.3. Medical image fusion

Setup. We conducted MIF experiments via the Harvard Medical dataset [9], which included 50 pairs of MRI-CT/MRI-PET/MRI-SPECT images. We directly generalize the models trained on the IVF task to the MIF task without fine-tuning. The quantitative metrics used are the same as those employed in the IVF task.

Comparison with SOTA methods. In both visual perception and quantitative measures in Fig. 4 and Tab. 1, EMMA demonstrates superior accuracy in extracting structural highlights and detailed texture features, and effectively integrates characteristic features within the fused image. Consequently, it achieves remarkable fusion results.

5. Conclusion

This paper tackles the lack of ground truth in image fusion by employing a conceptually straightforward yet potent prior that natural imaging responses exhibit equivariance to translations like shifts, rotations, and reflections. Upon this foundation, we propose a self-supervised paradigm called equivariant image fusion, which adjusts the inherent patterns of the loss function by taking into account the principles of natural imaging, making it simulate the sensing-imaging process. We also introduce a U-Net-like fusion module using the Restormer-CNN block as its basic unit, facilitating global-local feature extraction and efficient information fusion. Experimental results corroborate the effectiveness of our proposed paradigm in multi-modality image fusion, and its propensity to facilitate downstream tasks like multi-modality segmentation and detection.

Acknowledgement

This work has been supported by the National Natural Science Foundation of China under Grant 12371512 and 12201497, the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011358, and partly supported by the Alexander von Humboldt Foundation.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *CoRR*, abs/2004.10934, 2020. 1

[2] Dongdong Chen, Julián Tachella, and Mike E. Davies. Equivariant imaging: Learning beyond the range space. In *ICCV*, pages 4359–4368. IEEE, 2021. 3, 6

[3] Dongdong Chen, Julián Tachella, and Mike E. Davies. Robust equivariant imaging: a fully unsupervised framework for learning to image from noisy and partial measurements. In *CVPR*, pages 5637–5646. IEEE, 2022.

[4] Dongdong Chen, Mike E. Davies, Matthias J. Ehrhardt, Carola-Bibiane Schönlieb, Ferdia Sherry, and Julián Tachella. Imaging with equivariant deep learning: From unrolled network design to fully unsupervised learning. *IEEE Signal Process. Mag.*, 40(1):134–147, 2023. 3, 6

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 833–851. Springer, 2018. 8

[6] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3333–3348, 2021. 2

[7] Li Fang, Qian Wang, and Long Ye. Glgnet: light field angular superresolution with arbitrary interpolation rates. *Visual Intelligence*, 2(1):6, 2024. 1

[8] Fangyuan Gao, Xin Deng, Mai Xu, Jingyi Xu, and Pier Luigi Dragotti. Multi-modal convolutional dictionary learning. *IEEE Trans. Image Process.*, 31:1325–1339, 2022. 2

[9] Harvard Medical website. http://www.med.harvard.edu/AANLIB/home.html. 7, 8

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 770–778, 2016. 4

[11] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 4

[12] Alex Pappachen James and Belur V. Dasarathy. Medical image fusion: A survey of the state of the art. *Inf. Fusion*, 19: 4–19, 2014. 1

[13] Zhiying Jiang, Zengxi Zhang, Xin Fan, and Risheng Liu. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *ACM Multimedia*, pages 3783–3791, 2022. 1, 3

[14] Glenn Jocher. ultralytics/yolov5. https://github.com/ultralytics/yolov5, 2020. 8

[15] Hyungjoo Jung, Youngjung Kim, Hyunsung Jang, Namkoo Ha, and Kwanghoon Sohn. Unsupervised deep image fusion with structure tensor representations. *IEEE Trans. Image Process.*, 29:3845–3858, 2020. 2

[16] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018. 2

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

[18] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Inf. Fusion*, 73:72–86, 2021. 2, 3, 4

[19] Hui Li, Tianyang Xu, Xiaojun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):11040–11052, 2023. 2, 3, 5, 6, 7, 8

[20] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 4, 6, 7, 8

[21] 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 *CVPR*, pages 5792–5801. IEEE, 2022. 1, 2, 3, 4, 6, 7, 8

[22] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, pages 1–28, 2023. 2

[23] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8115–8124, 2023. 3, 5, 6, 7, 8

[24] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *ACM Multimedia*, pages 1600–1608. ACM, 2021. 1, 2

[25] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019. 6

[26] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019. 2

[27] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020.

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

[29] Bikash Meher, Sanjay Agrawal, Rutuparna Panda, and Ajith Abraham. A survey on region based image fusion methods. *Information Fusion*, 48:119–132, 2019. 1

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2, 5, 6

[31] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *ACM Multimedia*, pages 4003–4011, 2022. 3

[32] Julián Tachella, Dongdong Chen, and Mike Davies. Sensing theorems for unsupervised learning in linear inverse problems. *Journal of Machine Learning Research*, 24(39):1–45, 2023. 3, 5

[33] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion*, 82: 28–42, 2022. 1, 3

[34] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion*, 83-84:79–92, 2022. 6, 7, 8

[35] Vibashan Vs, Jeya Maria Jose Valanarasu, Poojan Oza, and Vishal M Patel. Image fusion transformer. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3566–3570. IEEE, 2022. 2

[36] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *IJCAI*, pages 3508–3515. ijcai.org, 2022. 1, 3

[37] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *AAAI Conference on Artificial Intelligence, AAAI*, pages 12484–12491, 2020. 2, 6, 7

[38] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):502–518, 2022. 1, 2, 3, 4

[39] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *CVPR*, pages 19647–19656. IEEE, 2022. 1, 3, 4

[40] Han Xu, Jiteng Yuan, and Jiayi Ma. MURF: mutually reinforcing multi-modal image registration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):12148–12166, 2023. 3, 5, 6, 7, 8

[41] Shuang Xu, Zixiang Zhao, Yicheng Wang, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Deep convolutional sparse coding networks for image fusion. *CoRR*, abs/2005.08448, 2020. 2

[42] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Guangyu Li, Jun Li, and Jian Yang. Learning complementary correlations for depth super-resolution with incomplete data in real world. *IEEE transactions on neural networks and learning systems*, 2022. 1

[43] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Rignet: Repetitive image guided network for depth completion. In *European Conference on Computer Vision*, pages 214–230. Springer, 2022. 1

[44] Wuyang Ye, Tao Yan, Jiahui Gao, and Yang Yang. Lfienet: Light field image enhancement network by fusing exposures of lf-dslr image pairs. *IEEE Transactions on Computational Imaging*, 2023. 3

[45] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5718–5729. IEEE, 2022. 2, 4

[46] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.*, 129(10):2761–2785, 2021. 2, 3, 4, 6, 7, 8

[47] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *AAAI*, pages 12797–12804. AAAI Press, 2020. 2

[48] Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2023. 1

[49] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion*, 54:99–118, 2020. 2

[50] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *CVPR*, pages 13955–13965. IEEE, 2023. 3, 5, 6, 7, 8

[51] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. DIDFuse: Deep image decomposition for infrared and visible image fusion. In *International Joint Conference on Artificial Intelligence, IJCAI*, pages 970–976, 2020. 1, 2, 3, 4

[52] Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Trans. Circuits Syst. Video Technol.*, 32(3):1186–1196, 2022. 2, 3, 4

[53] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5697–5707, 2022. 1

[54] 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. 1, 2, 5, 6, 7, 8

[55] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: Denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8082–8093, 2023. 2, 5, 6, 7, 8

[56] Zixiang Zhao, Jiangshe Zhang, Xiang Gu, Chengli Tan, Shuang Xu, Yulun Zhang, Radu Timofte, and Luc Van Gool. Spherical space feature decomposition for guided depth map super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12547–12558, 2023. 1