DIVIDUAL: A Disentangled Visible And Infrared Image Fusion Contrastive Learning Method

Shaoqi Yang[1] and Dan He[2*]

[1] School of Safety Engineering, Shenyang Aerospace University, No.37 Daoyi South Avenue, Shenbei New Area, Shenyang, 110136, China
[2] Dalian University of Finance and Economics, No. 80 Renwen Street, Jinzhou New Area, Dalian, 116622, China
[*] Corresponding author. E-mail: hedanc@outlook.com

Received: Mar. 15, 2024; Accepted: Apr. 14, 2024

Visible and Infrared Image Fusion (VIIF), as a vital fundamental component in vehicle applications, has been attracting plenty of attention from the academic and industrial communities over past few years. Various deep learning based methods have been proposed to effectively fuse visible and infrared images for improving the comprehensiveness of vehicle sensing and monitoring capabilities. However, due to the complex coupling property of the vehicle observational environment, it is still a challenging problem to effectively decouple and fusion visible and infrared images. To address this problem, we propose a DIsentangled Visible and InfrareD image fUsion contrAstive Learning method (DIVIDUAL). For capturing common and complementary information between the two domains, DIVIDUAL contains a self-supervised decoupling framework to separate domain-invariant and domain-specific representations. Meanwhile, for removing the noise in domain-specific representations and extracting clean domain-invariant representations, DIVIDUAL deploys a decoupling contrastive loss to effectively separate noise information and retain critical information in domains. Finally, DIVIDUAL generates fused images in an end-to-end manner. Extensive performance and generalization experiments on TNO and RoadScene datasets demonstrate that DIVIDUAL has superior visual results.

Keywords: Disentangled Learning; Image Fusion; Contrastive Learning; Self-unsupervised Learning
© The Author('s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

http://dx.doi.org/10.6180/jase.202505_28(5).0005

1. Introduction

As aircraft technology rapidly advances, the wide application of drones, airships and other aircraft in military, civil and scientific research fields has become a reality. In the process of aircraft mission execution, the comprehensive perception of the environment and the accurate identification of targets have become key technical challenges [1]. The application of VIIF in the vision system of aircraft provides a strong support to improve the perception ability and mission execution effect of the aircraft in the complex environment [2].

VIIF is aimed at extracting critical, complementary information and features from the source image to acquire a more comprehensive fusion image with superior perceptions and higher saliency of the target, thereby improving image recognition, target detection and scene analysis. Specifically, the visible image is intuitive to the human eye and can show the color and shape of an object, but the performance of the visible image may be limited in some specific situations, such as at night or under adverse weather conditions. In contrast, infrared images capture information about an object's thermal radiation and therefore perform better at nighttime or inclement weather conditions. By fusing visible and infrared images, it is possible to overcome shortcomings associated with each, resulting in a more comprehensive and robust perception system that provide strong support for the mission execution of aircraft

in complex environments.

VIIF methods are dedicated to designing different means for the extraction of features and rules for their fusion, so as to obtain high-quality information from the source image and achieve well-quality image fusion. Existing VIIF methods fall into two general categories. That is, traditional methods and deep learning-based methods. The former primarily centers on extracting features from source images and amalgamating them. Specifically, multiscale transform [3, 4] is the more common method, which often employs Laplace pyramid, wavelet transform, Laplace pyramid, wavelet transform, and multiscale geometric analysis to extract high-frequency sub-images and low-frequency sub-images, which will then enable the fusion of images by analyzing the sub-images. In contrast, sparse representation-based methods [5, 6] utilize sparse bases in an over-complete dictionary to represent the source image which enables feature extraction. Although such methods can improve the under-information capture problem in the multi-scale transform, the over-complete dictionary still has limited capability for signal representation, is prone to lose texture details in source images, and faces challenges related to low computational efficiency. The subspace-based fusion method [7] is generally designed to capture potential structures within source images. This is achieved by mapping high-dimensional images to low-dimensional space, which can effectively align images from different domains within both semantic and visual spaces. Despite the improved computational efficiency, it still has limitations in representing critical information within source images. In recent times, numerous deep learning-based approaches have been proposed [8, 9], leveraging auto-encoders, convolutional neural networks, and generative adversarial networks as the underlying architecture, yielding fusion results that can highlight salient targets and reveal the inherent styles and attributes of source images. Despite existing methods achieve superior results, most of them usually ignore the complementarity of different domains and suffer from the confusion of key texture and contextual information during feature extraction.

To solve the above problem, in combination with the idea of decoupling learning, some methods not only capture the domain-invariant information between different modalities [10], but also isolate the domain-specific information, thus achieving the decoupling of critical texture and contextual information. However, most of the methods rely on generative adversarial mechanisms, which are unable to obtain clean and complete domain-invariant information. In addition, during the fusion process, most of the methods usually incorporate information specific to two domains, thus introduce noise. Based on this observation, it is possible solution to enhance the effectiveness and robustness of VIIF by explicitly capturing complete and clean domain-invariant representations, and preserving complementarity and eliminating noise from domain-specific representations. To achieve this goal, there are two critical problems in VIIF that need to be resolved, i.e., (i) There is the requirement to achieve excellent separation between domain-invariant and domain-specific information, as these aspects are frequently intertwined and confused. (ii) Both domain-invariant and domain-specific information contain noise, necessitating a need to further distinguish the noise and preserve the complementary information.

In this paper, to tackle the aforementioned issues, we depart from the conventional decoupled learning methods employed in prior research, and investigate a new self-supervised learning framework, i.e., disentangled contrastive learning, to perform effective and robust VIIF. To this end, we initially devise a new disentangled framework for VIIF to extract complete and clean domain-invariant representations as well as domain-specific representations, addressing issue (i). Furthermore, we design the disentangled contrastive loss to further distinguish the noise and preserve the complementary information to combat issue (ii). As a result, the domain-specific and domain-inconvenient representations separated by the proposed disentangled contrastive learning method can be provably decoupled and contain more critical contextual and texture information, yielding fused images that are favorable for downstream tasks.

The contributions are summarized as follows:

- We introduce a novel disentangled method for VIIF out of generative adversarial mechanisms. To this end, we devise a disentangled architecture based on a self-supervised mechanism which is able to obtain clean and complete domain-invariant and domain-specific representations.
- We introduce a disentangled contrastive constraint in the disentangled framework that preserves complementary texture information and common contextual information in domain-invariant and domain-specific representations. It helps eliminate noise in domain-specific representations.
- Our approach has been extensively tested through a series of experiments, demonstrating superior performance in qualitative and quantitative analyses.

The organizational structure of this article is as follows: Section 1, i.e., this section, introduces the VIIF task and

the problem of existing VIIF methods, showcases how the proposed method works and summarizes the contributions. Section 2 furnishes a review of pertinent literature concerning VIIF, as well as disentangled representation learning, and states the motivation behind the proposed method. Section 3 delves into the specifics of the proposed method, while Section 4 assesses its performance and analyzes the experimental results. The paper is concluded in Section 5.

2. Related works

This section provides a brief overview of topics relevant to this work including traditional VIIF methods, deep learning-based VIIF methods, and disentangled representation learning for machine learning, to shed light on the motivation behind the proposed approach.

2.1. Traditional VIIF methods

Over the past few years, due to the highly complementary nature of infrared and visible images, fusing both for downstream tasks has attracted widespread attention and achieved remarkable performance. Traditional fusion methods are composed by two phases of feature extraction and feature fusion, including the multi-scale transform (MST)-based method [3, 4], the sparse representation (SR)-based method [5, 6], and the subspace-based fusion method [7].

The MST-based feature fusion methods are more typical ones, which utilize Laplacian pyramid [11] and multiscale geometric analysis [4] to extract multi-modal features, while preserving maximum information in the fusion result. Such methods are able to design specific strategies based on different frequencies of sub-bands. However, in most cases, they do not have translation invariance and have information redundancy, which makes them prone to lose structural information. DT-DWT [3] combines the weighted average method of normalized Shannon entropy to enhance the wavelet coefficients, which improves the fusion results. NSST-NMF [12] introduces fast nonnegative matrix decomposition into multiscale geometric analysis to adaptively fuse multi-modal critical features, which alleviates the problem of information redundancy. Although such methods have achieved good performance, they usually require the setting of predefined functions, which not only lacks universality but also suffers from the problem of insufficient information extraction.

To address this drawback, several SR-based methods have been introduced in recent times, including joint sparse representation (JSR) and latent low-rank representation (LRR). JSR [5] utilizes the sliding window to alleviate the issue of insufficient retention within the image edge information. LatLRR [6] decomposes source images into low-pass and high-pass sub-images to achieve adaptive weighted fusion of different source images. They are able to obtain content-rich fused images, but tend to possess high computational complexity. Subspace-based fusion methods can map high-dimensional data to low-dimensional subspaces, reducing computational cost while preserving the underlying structure of the image. HybridHDR-ICA [7] independent component analysis to extract uncorrelated variables, reducing the dimensionality while preserving information of the data. They reduce computational cost but lose critical information in multi-source images.

2.2. Deep learning-based VIIF methods

Recently, deep learning methods have experienced substantial advancements, leveraging their potent representation learning capabilities. They have been successfully applied to tasks involving VIIF.

With the first introduction of CNN to VIIF, MST-CNN [13] utilizes twin CNNs to obtain weight maps and joint image pyramids to achieve multiscale fusion, which leads the advancement in this field. Subsequently, UNFusion [14] and MMF [15] utilize CNNs with different sized convolutional kernels to extract multi-scale information during fusion process. These methods facilitate the integration of low-frequency and high-frequency information. FusionGAN [16], DDcGAN [17], and PMGI [18], which utilize adversarial relationships between generators and discriminators to enable the models to generate information-rich fused images under unsupervised conditions. VDFEFuse [19] extracts visual differentiation features based on encoder-decoder architectures, whose fusion results exhibit highlighted salient targets and provide sufficient texture information.

Moreover, in order to extract the long range dependence of the two modalities, some methods utilize Transformers to accomplish the task of VIIF. MFST [20] fuses multiscale features from multiscale encoders based on the focal self-attention mechanism, which preserves important information of the source images. DATFuse [21] employs the dual attention residual module to model long-range dependencies across different modalities, preserving global complementary information, and the fusion results have well generalization power. Despite their success, existing methods still fail to capture the complete common information between source images, and it is difficult to effectively utilize the private information from different sources to complement the fusion results, which is of great importance for downstream tasks.

2.3. Network architecture and notation

Disentangled representation learning is crafted to acquire decoupled representations capable of modeling the underlying factors within the data, and the theory has been successfully used in computer vision tasks.

Based on the generative manner, DRIT [22] and DRIT++ [23] capture cross-domain shared content information and domain-specific attribute information by introducing cyclic consistency loss in image-to-image translation tasks. DR-GAN [24] utilizes an encoder-decoder architecture to achieve a clear separation of facial pose from other facial variations for face recognition tasks. FDMER [25] learns public and private features of multimodal emotion data by defining multiple subspaces to enhance the performance of emotion recognition. WTI [26] operates on a decoupled framework capturing sequential and hierarchical representations to enable cross-modal interactions in text-video retrieval tasks. DMACCN [27] obtains complementary information of multimodal data by an adversarial cycle-consistent encoding-decoding architecture including modality-specific and modality-common sub-networks in unsupervised multimodal pattern mining.

Inspired by the prosperity in other fields, decoupled representation learning has attracted dramatic interest in infrared-visible fusion. DRF [28] utilizes dual encoders to decompose the image into representations related to scene and sensor modalities. DCDR-GAN [29] utilizes a densely connected generative adversarial architecture to decouple the content and modal features of the two modalities, whose obtained fused images have well visual quality. While the above methods have achieved superior results in VIIF, they are typically based on generating adversarial networks and rely on additional dense connectivity, failing to account for the potentially complex structure between the different modalities and neglecting to account for the complementarity and noise of each modality, which results in sub-optimal performance.

2.4. Motivation

Current methods for VIIF strive to reconstruct well-fused images using common information from different modalities, which can reveal the inherent style and properties of the image. However, they usually neglect the complementary nature of domain-specific information between different modalities and may confuse the critical attribute and content. e.g., if the infrared image is affected by equipment faults and may be noisy and blurred, with part of the content being present only in the visible image, it proves advantageous to acquire critical content by integrating complementary information from the visible image.

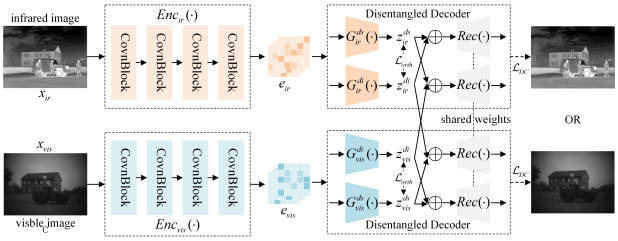

Fig. 1. The architecture of DIVIDUAL.

Therefore, it is necessary to consider both domain-invariant information for each modality and domain-specific information with complementary relationships. Since the public information helps to reconstruct the common scene in the fused image, it becomes crucial to capture domain-invariant information between different modalities. Such domain invariant information should contain all and only common information, i.e., clean and complete domain invariant information. Therefore, how to acquire clean and complete domain invariant information is the first problem to be addressed. Subsequently, due to the image acquisition process or mode, the visible and infrared images may contain noise, which is mixed with the complementary information. Merging the domain-specific representation with noise can easily impact the quality of the fused images. Therefore, how to obtain clean complementary domain-specific information is the second problem that needs to be solved.

Recently, people motivated by disentangled learning have been able to obtain domain-invariant information and domain-special information. It is frequently based on the coarse decoupling of generative adversarial networks, which fails to obtain clean and complete domain-invariant information. In addition, the presence of complex noise in visible and infrared images has limited the exploration of the complementary relationship between domain-invariance spaces. Addressing the aforementioned challenges through generative disentangled learning remains problematic. Considering this, this paper introduces disentangled multi-modal fusion method designed to tackle these issues.

3. Methodology

In this section, a comprehensive overview of DIVIDUAL is presented. To begin, we introduce the notations together with the generic network architecture of the method. Then, we describe the disentangled contrast mechanism, which utilizes domain-invariant information extraction with domain-specific information constraints to solve the above two problems. Finally, we outline the overall object.

3.1. Network architecture and notations

DIVIDUAL is consists of two key components: feature extractor and disentangled decoder, illustrated in Fig. 1. Given a VIS images domain $X_{vis} = \left\{x_{vis}^i\right\}_i^N \in \mathbb{R}^{N\times F}$ and an IR images domain $X_{ir} = \left\{x_{ir}^i\right\}_i^N \in \mathbb{R}^{N\times F}$, where N and F stand for the number and dimension of images, respectively. Firstly, the proposed DIVIDUAL obtains different modal features $\mathcal{E}_{\text{domain}}$ through two feature extractor $Enc_{\text{domain}}(\cdot)$. Then, learning the disentangled domain-invariant representations $Z_{\text{domain}}^{\text{di}} \in \mathbb{R}^{N\times d}$ and the domain-specific representations $Z_{\text{domain}}^{ds} \in \mathbb{R}^{N\times d}$ via disentangled networks Dis $(\cdot)$ with disentangled contrastive mechanism, where domain $= \{vis, ir\}$. Next, the fused representations $\mathcal{F} \in \mathbb{R}^{N\times(d+d)}$ are obtained by concatenating the optimized Z_{domain}^{di} and Z_{domain}^{ds} representations, where d are dimensions of the representation space. Finally, a reconstruction network Rec$(\cdot)$ is employed to map the fusion representation $\mathcal{F}$ to fusion images domain $x_f = \left\{x_f^i\right\}_i^N \in \mathbb{R}^{N\times F}$. For the sake of readability, Table 1 presents the notations employed throughout the paper.

We utilize a ConvBlock-based dual-encoder architecture for the feature extraction in the infrared image and visible image domains, formalized as $e_{vis} = Enc_{vi}\left(x_{vi}^i\right)$ and $e_{ir} = Enc_{irr}\left(x_{ir}^i\right)$. For both feature extractors, residual mapping is used to bridge different convolutional blocks, which alleviates both gradient vanishing and gradient explosion problems at the same time. Specifically, as in Fig. 2, two convolutional layers with kernel sizes of 1×1 and 3×3 form a convolutional block, each of which contains a residual operation, and four convolutional blocks form a feature extractor.

For each image domain, we used two disentangled networks to separate the features extracted by the feature extractor into domain-invariant and domain-specific representations, respectively (in fact, the two features have the same dimension). Subsequently, a 6-layer block, comprising an up-sampling layer and a convolutional layer, was employed as a reconstruction network. This network serves to merge and transform the domain-invariant and domain-specific representations into fusion images.

Formally, taking the visible light image domain as an example, the two decoupled branches $G_{vis}^{ds}(\cdot)$ and $G_{vis}^{di}(\cdot)$ are defined to be used for extracting the domain-invariant representation $z_{v,s}^{di,i}$ and the domain-specific representation $z_{vis}^{cb,i}$ for the i-th image. Similarly, the domain-invariant representation $z_{ir}^{d,i,i}$ and the domain-specific representation $z_{ir}^{d\tau,i}$ of the i-th image in the infrared optical image domain are obtained from the decoupled branches $G_{ir}^{ds}(\cdot)$ and $G_{ir}^{di}(\cdot)$. To solve the two problems in motivation, $z_{v,\delta}^{d,i,i}$ and $z_{ir}^{a,i,i}$

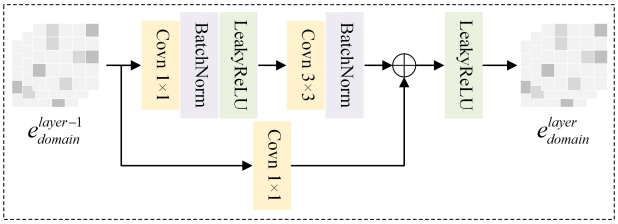

Fig. 2. Architecture of the ConvBlock.

are supposed to describe the same scene features, while $z_{v,s}^{ai,i}$ and $z_{ir}^{\text{ci,i,i}}$ are supposed to contain different complementary features. In this regard, we directly splice the domain-invariant and domain-specific representations as well as define the reconstruction network Rec$(\cdot)$ to generate the fused images. Specifically, we argue that DIVIDUAL should realize the following four types of fusion images:

$$\begin{aligned} x_{f\_1}^i &= \text{Rec}\left(z_{vis}^{ds,i} \oplus z_{vis}^{di,i}\right) \\ x_{f\_2}^i &= \text{Rec}\left(z_{ir}^{ds,i} \oplus z_{ir}^{di,i}\right) \\ x_{f\_3}^i &= \text{Rec}\left(z_{vis}^{ds,i} \oplus z_{ir}^{di,i}\right) \\ x_{f\_4}^i &= \text{Rec}\left(z_{irr}^{ds,i} \oplus z_{vis}^{di,i}\right) \end{aligned} \tag{1}$$

3.2. Disentangled contrastive learning

General disentangled learning tends to coarsely decoupled representations. Therefore, we design a disentangled contrastive learning that more easily captures complete and clean domain-invariant as well as domain-specific representations. This achieves effective separation of scene features and texture details in a self-supervised manner and suppresses the side effects of domain-specific representations. Since it is difficult to determine whether the public information in different domains is consistent or not, directly constructing positive and negative samples by data augmentation is prone to introduce many additional noises. In this regard, we design a patch-level disentangled contrastive loss in combination with the generated fused images.

Specifically, given i-th fusion image $x_{f\_m}^i m$, infrared image x_{ir}^i and visible image x_{vis}^i, where $m \in \{1,2,3,4\}$. First, to extract image features, we reuse dyadic feature extractor Enc domain $(\cdot)$ of the original domain and access a weight-sharing network mapping head $H(\cdot)$, which is MLP. As a result, following features can be obtained:

Table 1. Notations.

Notations	Description
X_{vis}, X_{ir}, and X_f	VIS, IR, and fusion images domain
Z_{vis}^{di}, Z_{ir}^{di}, Z_{vis}^{ds}, and Z_{ir}^{ds}	Domain-invariant and domain-specific representation of X_{vis} and X_{ir}
F	The fused representations
E_{is} and E_{ir}	The feature of X_{vis} and X_{ir}
L	Loss function
$Enc_{vis}(\cdot)$ and $Enc_{ir}(\cdot)$	The feature extractor of X_{vis} and X_{ir}
$\mathrm{Rec}(\cdot)$	The reconstruction network
$G_{vis}^{di}(\cdot)$, $G_{vis}^{ds}(\cdot)$, $G_{ij}^{di}(\cdot)$, and $G_{ij}^{ds}(\cdot)$	The decoupled branch for extracting the domain invariant and domain-specific representations of X_{vis} and X_{ir}
β_1, β_2, and β_3	The balancing factors
λ_1 and λ_2	The moderating factors

$$\begin{aligned} s_{m\_vis}^{i} &= H_{vis}\left(Enc_{vis}\left(x_{f\_m}^{i}\right)\right) \\ s_{m\_ir}^{i} &= H_{ir}\left(Enc_{ir}\left(x_{f\_m}^{i}\right)\right) \\ s_{ir}^{i} &= H_{ir}\left(Enc_{ir}\left(x_{ir}^{i}\right)\right) \\ s_{vis}^{i} &= H_{ir}\left(Enc_{ir}\left(x_{vis}^{i}\right)\right) \end{aligned} \tag{2}$$

Then, the image blocks in the same position of fused image domain and original domain image are used as positive pairs, and image blocks in N different positions are randomly sampled as negative pairs. The block features in the positive pair are treated as $s^p \in \mathbb{R}^C$, and the other block features at the same feature level are treated as $s^{P/p} \in \mathbb{R}^{(P-1)\times C}$, where C is channel number and $s \in \{1, \ldots, S\}$ is spatial location number. Motivated by patchNCE [30], we define the disentangled contrastive loss as:

$$\mathrm{L}_{DC} = \lambda_1 E_{x\sim X} l_{DC}^{ir} + (1-\lambda_1)\, E_{x\sim X} l_{DC}^{is} \tag{3}$$

where λ_1 is the moderating factor:

$$\mathrm{L}_{DC} = \lambda_1 E_{x\sim X} l_{DC}^{ir} + (1-\lambda_1)\, E_{x\sim X} l_{DC}^{is} \tag{4}$$

And

$$\begin{aligned} l_{DC}^{ir} &= \sum_{p=1}^{P} \\ &- \log\left[\frac{\exp\left(s_m^i p_{-ir}^{p} \cdot s_{irr}^i p/\tau\right)}{\exp\left(s_m^{m_- - p} \cdot s_{ir}^{i-p}/\tau\right) + \sum_{p=1}^{P-1}\exp\left(s_m^{i-p} \cdot \cdot s_{ir}^{i-P/p}/\tau\right)}\right] \\ l_{DC}^{vis} &= \sum_{p=1}^{P} \\ &- \log\left[\frac{\exp\left(s_{m\_p}^{\mathrm{vis}} \cdot s_{vis}^i p/\tau\right)}{\exp\left(s_{m\_-p}^{i_{\mathrm{vis}}} \cdot s_{\mathrm{vis}}^{i-p}/\tau\right) + \sum_{p=1}^{P-1}\exp\left(s_{m\_-p}^{v_{\mathrm{vis}}} \cdot s_{\mathrm{vis}}^{i-P/p}/\tau\right)}\right] \end{aligned} \tag{5}$$

where τ denotes the temperature parameter. The objective is to encourage the fusion image to faithfully restore distinct source images by evaluating the similarity of the output features. This loss constrains how closely the fused image generated by domain-invariant and domain-specific representations approximate source images. As much as possible, the domain-invariant representation and domain-specific representation contain key texture and background information in source images.

In order to capture purer domain-invariant and domain-specific information, DIVIDUAL is concerned with decoupled the two, aiming to render them independent of each other. To achieve this, the orthogonalization rule is further introduced and we constrain that F and F should be orthogonal to each other. The orthogonalization rule $\mathrm{L}_{\mathrm{orth}}$ is defined as follows:

$$\mathrm{L}_{\mathrm{orth}} = \left\|Z_{vis}^{di}, Z_{uis}^{ds}\right\|_2^2 + \left\|Z_{ir}^{di}, Z_{ir}^{ds}\right\|_2^2 \tag{6}$$

During the learning process, minimizing $\mathrm{L}_{\mathrm{orth}}$ can help force the DOT product of Z_{comain}^{ci} and Z_{comain}^{ck} to approach zero so that they are orthogonal to each other, thus con-

straining domain-invariant and domain-specific information within each modal image to be dissimilar.

3.3. Objective function

In addition to the use of disentangled contrastive loss and orthogonality constraints, we additionally introduce SSIM loss to learn the structural features of an image. SSIM can efficiently evaluate the quality of an image through brightness, contrast and structure [30]. Given P_1 and P_2, SSIM can be defined as:

$$\mathrm{SSIM}\left(P_1, P_2\right) = \frac{\left(2\rho_{P_1P_2} + \sigma\right)\left(2v_{P_1}v_{P_2} + \mu\right)}{\left({\rho_{P_1}}^2 + {\rho_{P_2}}^2 + \sigma\right)\left({v_{P_1}}^2 + {v_{P_2}}^2 + \mu\right)} \tag{7}$$

where ρ and v stand for the variance/covariance and mean of corresponding P_1 and P_2, respectively. μ and σ are two floating point numbers that tend to zero. From this, the SSIM loss can be formalized as:

$$E(I \mid W) = \frac{1}{mn}\sum_{i=1}^{mn} P_i \tag{8}$$

where W represents the sliding window. Referring to disentangled contrastive loss, for different approximations, we impose moderators for the SSIM loss. Thus, SSIM is defined in the following form:

$$\begin{aligned} L_{SSIM} = 1 - \lambda_2 \tfrac{1}{N}\textstyle\sum_{w=1}^{W} \mathrm{Score}\left(X_{f\_m}, X_{ir} \mid W\right) \\ - \left(1-\lambda_2\right) \tfrac{1}{N}\textstyle\sum_{w=1}^{W} \mathrm{Score}\left(X_{f\_m}, X_{vis} \mid W\right) \end{aligned} \tag{9}$$

where λ_2 is the moderating factor:

$$\begin{cases} \lambda_2 = 1, \text{ if } \left(X_{ir} \mid W\right) > \left(X_{vis} \mid W\right) \\ \lambda_2 = 0, \text{ if } \left(X_{vis} \mid W\right) > \left(X_{ir} \mid W\right) \end{cases} \tag{10}$$

In this regard, SSIM can guarantee the structural consistency by constraining the shallow features of fused images and source images. This loss function is employed with the aim of encouraging the retention of almost as much complementary information between different source domains. Ultimately, DIVIDUAL is trained in an end-to-end manner by combining disentangled contrastive loss, orthogonality constraint loss, and SSIM loss.

$$\mathrm{L}_{\text{overall}} = \beta_1 \mathrm{L}_{DC} + \beta_2 \mathrm{L}_{\text{orth}} + \beta_3 \mathrm{L}_{SSIM} \tag{11}$$

where $\beta_1, \beta_2, and \beta_3$ are the balancing factors.

4. Experiments

This section begins by introducing the datasets and evaluation metrics. Subsequently, experimental setup, selected baselines, and the configuration of DIVIDUAL are detailed. This section presents qualitative and quantitative comparisons to demonstrate that DIVIDUAL is effective.

4.1. Datasets

We were used two real-world benchmark datasets for evaluation: TNO dataset [31] and RoadScene dataset. Specifically, TNO dataset encompasses nighttime images depicting various military-related scenarios, extensively utilized for VIIF. It includes 60 pairs of visible and infrared images along with screenshots in three video sequences. RoadScene dataset comprises FLIR video [32] image of pedestrian, roadway, and vehicular scenes, and it includes 221 aligned image pairs.

We formed a training set comprising 45 pairs of visible and infrared images from TNO dataset. To augment this dataset, we employed cropping techniques, in which the sliding window is configured at 128×128, with a step size of 32, and 8745 image patch pairs were finally obtained. During the testing phase, we chose 20 pairs of images from the TNO dataset to conduct a comparative evaluation. To assess the generalizability of DIVIDUAL, we conducted experiments using 20 image pairs from the RoadScene dataset. The visible images in RGB format from RoadScene dataset were converted to a single channel (Y channel in the YCbCr color space).

4.2. Evaluation metrics

Six communal metrics are used as the evaluation metrics, as defined below:

- **EN**. Information Entropy [33], which assesses the information of fusion images by analyzing pixel distribution, with increased values corresponding to more information within images.
- **MI**. Mutual Information [34], which gauges the information linkage strength between fusion and source images, with increased values signifying superior image fusion performance.
- **VIF**. Visual Information Fidelity [35], which assesses fusion images by simulating the characteristics of the human visual system. Elevated values indicate greater consistency with human vision, i.e., high fidelity.
- **SD**. Standard Deviation [36], which characterizes the discrete gray values, with increased values corresponding to heightened image contrast.
- **SF**. Spatial Frequency [37], which quantifies the percentage change in pixel grayscale by computing the disparity between the center pixel and its adjacent pixels. Elevated values indicate sharper images.
- **AG**. Average Gradient [38], which mirrors the fused image's capacity to portray details and textures.

Higher values signify that the image encompasses more abundant information and exhibits superior definition.

4.3. Experimental setup

4.3.1. Baselines

The proposed method's performance is accentuated through a comparative analysis with several baselines, outlined as follows:

- DenseFuse [39]: It is a deep learning architecture that utilizes intensive blocks to capture more beneficial features.
- IFSepR [40]: It utilizes multi-branch encoder learning for de-entanglement learning of different characteristic information.
- RFN-Nest [41]: It constructs residual fusion networks based on residual architectures, which is a two-stage fusion method.
- U2Fusion [42]: It automatically assigns weights to the source image and proposes adaptive information preservation degrees, whose are unsupervised fusion methods.
- UMF-CMGR [43]: It allows vector displacement between different kinds of infrared images and feature fusion through image alignment and interaction mechanisms.
- PIAFusion [44]: It maximizes the preservation of highlight patches preserving the background while also preserving the texture information of the image.
- SeAFusion [45]: It preserves the semantic information in the fused image through supervision.
- CLF-Net [46]: It constructs an unsupervised contrastive learning framework that utilizes structural similarity loss and contrast loss to effectively guide feature extraction and fusion.
- PSFusion [47]: The fusion is achieved by progressively injecting semantics and constraining scene fidelity.

4.3.2. Implementation details

The original image for each viewpoint is normalized to the range $[-1, 1]$. The MLP includes a fully connected operation and a normalization operation, a LeakyReLU activation function, and a fully-connected operation for final output. The input and output channels of the inverse convolutional layer are 128 and 64 respectively. During the training process, the batch size, learning rate and epoch are set to 4, 0.002, and 20, respectively.

4.4. Performance experiment

To conduct a thorough analytical evaluation, DIVIDUAL was compared with nine baselines by TNO dataset.

4.4.1. Qualitative results

To intuitively identify the performance discrepancies among different algorithms, we opted to showcase results of various algorithms on four image pairs. In Fig. 3a, for comparison, we zoomed in on the smaller character information patches and displayed them in the bottom corner of each fusion image. It can be clearly seen from the zoomed-in image that DIVIDUAL has preserved the maximum amount of character texture information, including the light and dark contrast information of the arm, clothes, etc. On the contrary, DenseFuse [39], IFSepR [40], RFN-Nest [41], U2Fusion [42], UMF-CMGR [43], and PIAFusion [44] lost a lot of texture features of the characters. In addition, the yellow box in each image demonstrates that although SeAFusion [45], CLF-Net [46], PSFusion [47]captured relatively rich texture information, their fused images were interfered with by infrared light. The brightness of the sky portion of the SeAFusion [45], CLF-Net [46], PSFusion [47] fusion result is more closely aligned to infrared light than to visible light. In contrast, benefiting from the de-entanglement learning of public and private information, DIVIDUAL preserves both the detailed character texture and reasonable sky background in the fused image.

Fig. 3b demonstrates the difference in image fusion performance of different algorithms in the case of visible images with occlusion. For comparison, we have likewise labeled the character patches and background patches in each fused image with red and yellow boxes, respectively. As depicted in Fig. 3b, it is evident that DIVIDUAL excels in producing a distinct silhouette of the figure with minimal interference from thermal infrared information. Additionally, DIVIDUAL successfully captures the brightest sky, aligning with the corresponding region. Noteworthy, while PSFusion [47] implements clear character outlines similar to DIVIDUAL, its sky is gray and only highlights the silhouettes of the background trees. The reason for this PSFusion [47] is highly disturbed by thermal IR information, leading to the dropping of most features within visible images. The loss function only imposes constraints on retaining contour constraints on retaining contour features in the fused image. DenseFuse [39], IFSepR [40], RFN-Nest [41], U2Fusion [42] are affected by visible light occlusion, and does not produce the silhouette of images very clearly.

Fig. 4a demonstrates the fused image generation results of different algorithms fusing thermal infrared light when the visible image brightness is dim. Our proposed DIVID-

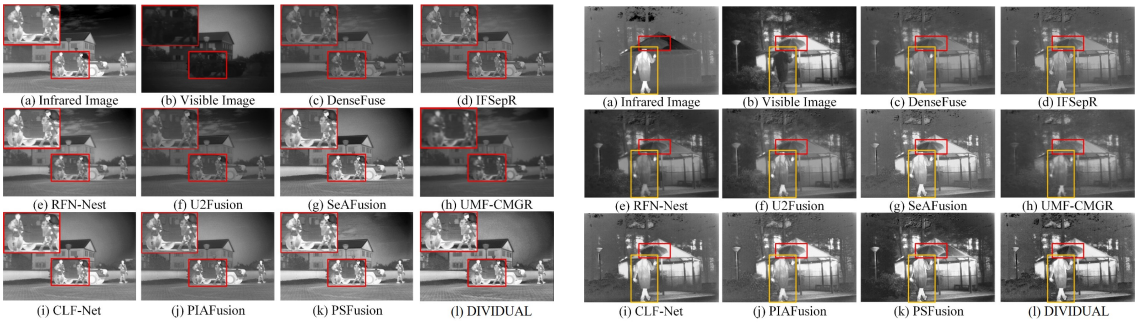

Fig. 3. For comparison purposes, we have identified the highlighted objects using red boxes and will place them at the bottom right corner. In addition, we have used orange boxes to display the figure. (a) Qualitative comparison results in dimly lit images of people and cars. (b) Qualitative comparison results in images where the object and background temperatures are similar (IR image chromaticity is similar

UAL captures richer character texture features compared to existing methods [39–41]. And although PSFusion [47] can extract richer texture features, its character edges are too sharpened and do not resemble visible images. In contrast, DIVIDUAL is more in line with the actual semantics of images.

Fig. 4b illustrates the baseline and DIVIDUAL fusion performance when the figure and the background have similar chromaticity in the infrared image. Ideally, when the object and the background have similar chromaticity in IR light, the fused image should highlight the difference in visible light color more. The red box in Fig. 4b demonstrates that DIVIDUAL can show the visible light color differences between the umbrella and the gazebo more clearly when they are similar in IR light. On the contrary, PSFusion [46] fails to preserve the visible chromaticity of the gazebo. PIAFusion [44] and SeAFusion [45] are able to capture the visible color of the gazebo, but these methods sacrifice the texture features of the pedestrians for sub-optimal fusion results.

4.4.2. *Quantitative results*

The outcomes for the 6 general quantitative metrics are summarized in Table 2. Notably, DIVIDUAL achieved the best performance in terms of VIF, SD, AG, and SF.

Regarding the MI and EN metrics, DIVIDUAL performs optimally, except for CLF-Net [46] and PSFusion. In the case of the EN metric, DIVIDUAL closely aligns with PSFusion, whose superior EN is due to its ability to capture image texture features, but it is not able to capture texture features and image structure at the same time, and thus the VIF value of PSFusion is much lower than that of DIVIDUAL. Whereas, for metric, DIVIDUAL is only slightly behind CLF-Net [46], which is due to the fact that CLF-Net [46] maximizes the mutual information between different domains using contrastive contrastive learning method of PatchNCE, but the limitation of this approach is that it is unable to decouple the public and private information within the viewpoints, which leads to the fact that the CLF-Net [46] can only unidirectionally resolve the visible image or the infrared image without being able to simultaneously preserve their respective features, and this defect is also reflected in the lower AG value of CLF-Net [46]. On the contrary, the DIVIDUAL proposed in this paper effectively remedies this defect and achieves superior AG values.

Drawing on the aforementioned quantitative analysis, DIVIDUAL utilizes the disentangled learning to adequately weigh the importance of information between different domains, and achieves good performance on the TNO dataset.

4.5. Generalization experiment

We conduct experiments on the RoadScene dataset to investigate the generalizability of DIVIDUAL.

4.5.1. *Qualitative results*

As with the TNO dataset, we selected four image pairs from the RoadScene dataset for analysis. To thoroughly assess the generalization ability of the fusion methods, we choose two images taken during the daytime and two taken at night. Fig. 5a shows the fusion performance of DIVIDUAL and the other nine methods for a single car during daytime. We used a similar method to the TNO dataset to zoom in on the details of the fused images. Fig. 5a shows that DIVIDUAL can clearly preserve the texture features of the car, such as the license plate number, during daytime. While the images fused by other methods lead to blurring of the license plate number. In addition, DIVIDUAL also preserves the sky chromaticity in daylight.

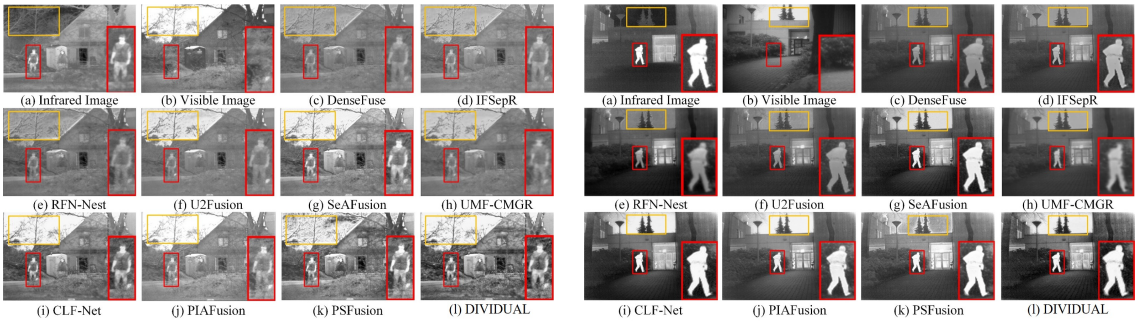

Fig. 4. For comparison purposes, we have identified the highlighted objects using red boxes and will place them at the bottom right corner. In addition, we have used orange boxes to display the figure. (a) Qualitative comparison results in dimly lit images of people and cars. (b) Qualitative comparison results in images where the object and background temperatures are similar (IR image chromaticity is similar).

Table 2. Notations.

	TNO						RoadScene					
	EN	MI	VIF	SD	AG	SF	EN	MI	VIF	SD	AG	SF
DenseFuse	6.398	2.104	0.634	8.505	2.701	0.026	6.654	2.886	0.629	9.478	3.114	0.031
IFSepR	6.559	2.090	0.683	8.722	3.579	0.037	6.708	2.683	0.610	9.758	4.081	0.049
RFN-Nest	6.965	2.005	0.761	9.155	2.870	0.024	7.260	2.720	0.704	10.031	3.315	0.030
U2Fusion	6.990	1.769	0.740	9.438	5.313	0.047	7.154	2.760	0.680	9.980	5.925	0.058
SeAFusion	7.124	2.765	0.951	9.445	5.867	0.052	7.595	3.239	0.918	10.665	5.468	0.064
UMF-CMGR	6.724	1.924	1.016	9.381	5.465	0.0545	6.016	1.207	0.895	9.901	4.821	0.049
CLF-Net	7.226	3.635	1.048	9.613	4.989	0.049	7.454	5.164	1.083	10.461	6.076	0.068
PIAFusion	6.501	3.431	0.877	9.765	5.331	0.092	6.211	4.057	0.922	8.803	4.021	0.062
PSFusion	7.525	3.281	0.734	9.322	6.611	0.132	7.509	3.110	0.664	9.021	6.085	0.104
DIVIDUAL	7.506	3.360	1.126	9.771	7.974	0.132	7.610	3.332	1.096	9.760	7.974	0.129

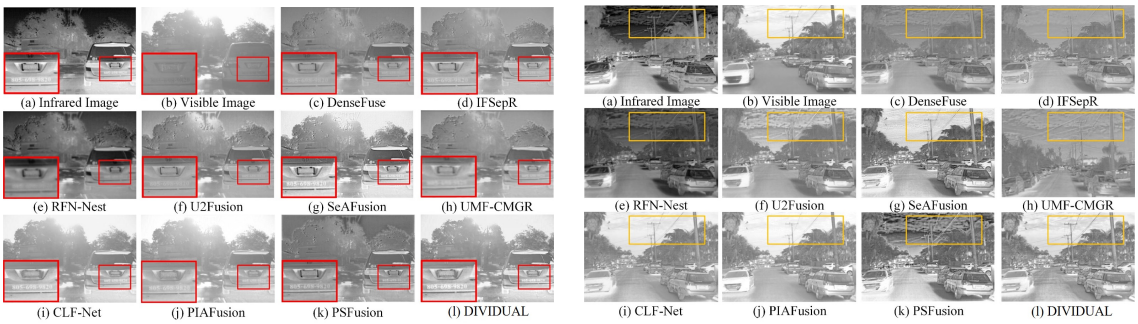

Fig. 5. Qualitative comparison of DIVIDUAL results with baseline methods. We define highlighted objects with orange boxes and place them at the bottom right corner. (a) Qualitative comparison results in individual car images in daylight. (b) Qualitative comparison results in mixed images of poles and vehicles.

Fig. 5b demonstrates the fusion performance of DIVIDUAL and nine other methods on a hybrid image of utility poles and vehicles. The orange boxes in Fig. 5b label the texture effect of the utility poles. The experimental results show that DIVIDUAL can adequately capture the wire texture among the utility poles. Since the wires are thin in the image, DenseFuse [39], IFSepR [40], RFN-Nest [41], U2Fusion [42], UMF-CMGR [43] lost the texture features of the wires during fusion. In addition, we can observe that PSFusion [47] retains clearer cloud textures in the sky compared to DIVIDUAL, which is due to the fact that PSFusion [47] captures all the textures in the image indiscriminately.

And thanks to the de-entanglement learning, DIVIDUAL treats the sky as the private information of the lightable image, which is reasonable in practical applications because the sky tends to have less semantic texture information and more chromaticity information.

Fig. 6a shows the fusion performance of DIVIDUAL and nine other methods for mixed images of buildings and vehicles at night. We observed the fusion effect on the window areas of the building and zoomed in on them at the bottom of the image. It was observed that the other 9 methods struggled to preserve the texture of the windows at night, while DIVIDUAL can outline the window borders more clearly. We note that DenseFuse [39], IFSepR [40], RFN-Nest [41], U2Fusion [42], UMF-CMGR [43], PSFusion [47] fused building images have darker window chromaticity compared to the walls, which suggests that they prefer to capture the features of building patches, while PIAFusion [44], SeAFusion [45], and CLF-Net [46] prefer to capture the features of building patches. In contrast, DIVIDUAL retains the most informative parts of the images from both modalities, which leads DIVIDUAL to exacerbate the chromatic contrast between windows and walls, allowing for clearer outlines. In addition, DIVIDUAL loses the light speckle features in tower buildings. The reason for this phenomenon could be that the speckle features are learned as public information by DIVIDUAL, resulting in that they are not preserved in the private information of the visible light image. Other possible reasons are that the speckles are recognized by DIVIDUAL as a noisy part. However, the generalization performance of DIVIDUAL is still optimally beneficial in terms of overall fusion results.

Finally, we chose a multi-pedestrian image to evaluate the fusion performance of DIVIDUAL, and the results are shown in Fig. 6b. First, the orange color of Fig. 6b indicates that DIVIDUAL retains clear tree edge textures under darkness, while the other methods blend the trees and the night sky into one. It is worth noting that the moonlight spots in the original image pair are recognized as noise by DIVIDUAL and removed from the private information. The red boxes in Fig. 6b indicate that all 10 methods, including DIVIDUAL, were able to capture relatively clear character textures. However, DIVIDUAL, PIAFusion [44], and PSFusion [47] can generate texture details of the pedestrian clothing and limbs in greater detail. The blue box in Fig. 6b highlights the texture details of the crowd in the distance, through which it can be seen that DIVIDUAL is slightly less effective than PSFusion [47] in fusing the blurred textures in the distance. This is due to the loss function for filtering private information noise recognizes the blue patches as private information noise of the thermal infrared image when comparing the visible and infrared images.

4.5.2. *Quantitative results*

To quantitatively assess the difference in generalizability between DIVIDUAL and the other 9 algorithms, similar to the quantitative analysis, we followed the same experimental steps to evaluate 6 quantitative metrics on the Roadscene dataset, as shown in Table 2. Among the six metrics, DIVIDUAL obtains the best quantitative results on all the other five metrics except SD. It is worth noting that DIVIDUAL results in four metrics, MI, VIF, AG, SF, are basically equal to the TNO dataset, and DIVIDUAL EN metrics under the road scenario dataset even outperform surpass the TNO dataset. This fully demonstrates that DIVIDUAL has superior generalization ability to out-of-distribution datasets. The SD results of DIVIDUAL are lower than those of SeAFusion [44], but they are close to those of the TNO dataset. We believe that DIVIDUAL images trade-off image texture contrast while preserving public information. The texture contrast of DIVIDUAL fused images is closer to that of visible light images. DIVIDUAL trades an acceptable weakening of texture contrast for closer to real image information.

Based on the above quantitative analysis, the fusion effect of DIVIDUAL on the outer distribution dataset is slightly inferior to the existing methods in terms of SD metrics, but it still has a great generalization ability when analyzing all the metrics together.

5. Conclusion

In this work, we suggest a VIIF framework based on disentangled learning and self-supervised mechanisms, namely DIVIDUAL, which combines the advantages of encoding-decoding models. Specifically, to achieve effective decoupling of source images, we propose to separate the complex information within the source image using a disentangled sub-network for obtaining domain-invariant and domain-specific representations. Next, the utilization of disentangled contrastive constraint assures that the separated information adequately contains common and critical information in source images. Then, to remove the noisy information embedded in the domain-specific representation, we introduce orthogonality constraints to optimize the domain-specific representation. Finally, we utilize additional structural losses to aid in the retention of information in the source image. On two generalized real datasets, performance experiments and generalization experiments are conducted, which proves DIVIDUAL is superior.

In addition, lightweight and portable versions of models are often required for better integration of the proposed

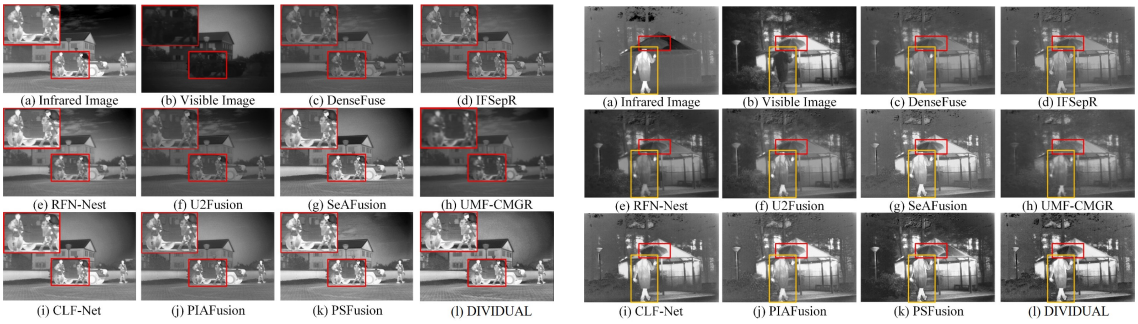

Fig. 6. For comparison purposes, we used the red box to select salient objects and place them at the bottom right corner. In addition, blue boxes enlarge the pedestrian vignettes in the distance, and the orange box labels patches of the night sky and trees. (a) Qualitative comparison results in mixed images of buildings and vehicles at night. (b) Qualitative comparison results in ulti-pedestrian images.

methods with real-world applications in image fusion tasks. Especially in the field of aircraft sensing, lightweight models will facilitate the execution of downstream tasks such as target identification and environment monitoring in aircraft. However, in DIVIDUAL, utilizing disentangled networks based on encoding-decoding architectures to obtain some pre-trained features tends to increase the size of the model. There must be a more lightweight model that can combine disentangled learning with encoding-decoding architecture for image fusion. According to this demand, solving the model size problem of image fusion methods based on disentangled learning and better adapting to practical application scenarios are future research directions in VIIF.

References

[1] J. Pisane, S. Azarian, M. Lesturgie, and J. Verly, (2014) *"Automatic target recognition for passive radar"* **IEEE Transactions on Aerospace and Electronic Systems *50***(1): 371–392. DOI: 10.1109/TAES.2013.120486.

[2] C. Sun, C. Zhang, and N. Xiong, (2020) *"Infrared and visible image fusion techniques based on deep learning: A review"* **Electronics *9***(12): 2162. DOI: 10.3390/electronics9122162.

[3] Y. Zou, X. Liang, and T. Wang, (2013) *"Visible and infrared image fusion using the lifting wavelet"* **TELKOMNIKA Indonesian Journal of Electrical Engineering *11***(11): 6290–6295.

[4] Q. Zhang and X. Maldague, (2016) *"An adaptive fusion approach for infrared and visible images based on NSCT and compressed sensing"* **Infrared Physics & Technology *74***: 11–20. DOI: 10.1016/j.infrared.2015.11.003.

[5] Y. Liu, X. Yang, R. Zhang, M. K. Albertini, T. Celik, and G. Jeon, (2020) *"Entropy-based image fusion with joint sparse representation and rolling guidance filter"* **Entropy *22***(1): 118. DOI: 10.3390/e22010118.

[6] B. Cheng, L. Jin, and G. Li, (2018) *"General fusion method for infrared and visual images via latent low-rank representation and local non-subsampled shearlet transform"* **Infrared Physics & Technology *92***: 68–77. DOI: 10.1016/j.infrared.2018.05.006.

[7] Y. Huang and K. Yao. "Multi-Exposure Image Fusion Method Based on Independent Component Analysis". In: *Proceedings of the 2020 International Conference on Pattern Recognition and Intelligent Systems*. 2020, 1–6.

[8] P. Li, J. Gao, J. Zhang, S. Jin, and Z. Chen, (2023) *"Deep Reinforcement Clustering"* **IEEE Transactions on Multimedia *25***: 8183–8193.

[9] J. Gao, P. Li, A. A. Laghari, G. Srivastava, T. R. Gadekallu, S. Abbas, and J. Zhang, (2024) *"Incomplete Multiview Clustering via Semidiscrete Optimal Transport for Multimedia Data Mining in IoT"* **ACM Transactions on Multimedia Computing, Communications and Applications *20***(6): 158:1–158:20. DOI: 10.1145/3625548.

[10] J. Gao, M. Liu, P. Li, J. Zhang, and Z. Chen, (2023) *"Deep Multiview Adaptive Clustering With Semantic Invariance"* **IEEE Transactions on Neural Networks and Learning Systems**: 1–14. DOI: 10.1109/TNNLS.2023.3265699.

[11] P. J. Burt and E. H. Adelson. "The Laplacian pyramid as a compact image code". In: *Readings in computer vision*. Elsevier, 1987, 671–679.

[12] W. Kong, Y. Lei, and H. Zhao, (2014) *"Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization"* **Infrared Physics & Technology *67***: 161–172. DOI: 10.1016/j.infrared.2014.07.019.

[13] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, (2018) *"Infrared and visible image fusion with convolutional neural networks"* **International Journal of Wavelets, Multiresolution and Information Processing *16***(03): 1850018.

[14] Z. Wang, J. Wang, Y. Wu, J. Xu, and X. Zhang, (2021) *"UNFusion: A unified multi-scale densely connected network for infrared and visible image fusion"* **IEEE Transactions on Circuits and Systems for Video Technology *32***(6): 3360–3374.

[15] Y. Liu, C. Miao, J. Ji, and X. Li, (2021) *"MMF: A Multi-scale MobileNet based fusion method for infrared and visible image"* **Infrared Physics & Technology *119***: 103894.

[16] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, (2019) *"FusionGAN: A generative adversarial network for infrared and visible image fusion"* **Information fusion *48***: 11–26. DOI: 10.1016/j.inffus.2018.09.004.

[17] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, (2020) *"DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion"* **IEEE Transactions on Image Processing *29***: 4980–4995. DOI: 10.1109/TIP.2020.2977573.

[18] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma. "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity". In: *Proceedings of the AAAI conference on artificial intelligence*. ***34***. 07. 2020, 12797–12804.

[19] S. Hao, T. He, B. An, X. Ma, H. Wen, and F. Wang, (2022) *"VDFEFuse: A novel fusion approach to infrared and visible images"* **Infrared Physics & Technology *121***: 104048. DOI: 10.1016/j.infrared.2022.104048.

[20] X. Liu, H. Gao, Q. Miao, Y. Xi, Y. Ai, and D. Gao, (2022) *"MFST: Multi-modal feature self-adaptive transformer for infrared and visible image fusion"* **Remote Sensing *14***(13): 3233.

[21] W. Tang, F. He, Y. Liu, Y. Duan, and T. Si, (2023) *"Datfuse: Infrared and visible image fusion via dual attention transformer"* **IEEE Transactions on Circuits and Systems for Video Technology**:

[22] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. "Diverse image-to-image translation via disentangled representations". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, 35–51.

[23] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, (2020) *"Drit++: Diverse image-to-image translation via disentangled representations"* **International Journal of Computer Vision *128***: 2402–2417.

[24] L. Tran, X. Yin, and X. Liu. "Disentangled representation learning gan for pose-invariant face recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, 1415–1424.

[25] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang. "Disentangled representation learning for multimodal emotion recognition". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, 1642–1651.

[26] Q. Wang, Y. Zhang, Y. Zheng, P. Pan, and X.-S. Hua, (2022) *"Disentangled representation learning for text-video retrieval"* **arXiv preprint arXiv:2203.07111**:

[27] P. Li, A. A. Laghari, M. Rashid, J. Gao, T. R. Gadekallu, A. R. Javed, and S. Yin, (2023) *"A Deep Multimodal Adversarial Cycle-Consistent Network for Smart Enterprise System"* **IEEE Transactions on Industrial Informatics *19***(1): 693–702. DOI: 10.1109/TII.2022.3197201.

[28] H. Xu, X. Wang, and J. Ma, (2021) *"DRF: Disentangled representation for visible and infrared image fusion"* **IEEE Transactions on Instrumentation and Measurement *70***: 1–13.

[29] Y. Gao, S. Ma, and J. Liu, (2022) *"DCDR-GAN: A densely connected disentangled representation generative adversarial network for infrared and visible image fusion"* **IEEE Transactions on Circuits and Systems for Video Technology *33***(2): 549–561.

[30] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, (2020) *"Unsupervised learning of visual features by contrasting cluster assignments"* **Advances in neural information processing systems *33***: 9912–9924.

[31] URL: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029.

[32] URL: https://www.flir.com/oem/adas/adas-dataset-form/.

[33] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed, (2008) *"Assessment of image fusion procedures using entropy, image quality, and multispectral classification"* **Journal of Applied Remote Sensing** *2*(1): 023522.

[34] G. Qu, D. Zhang, and P. Yan, (2002) *"Information measure for performance of image fusion"* **Electronics letters** *38*(7): 1.

[35] Y. Han, Y. Cai, Y. Cao, and X. Xu, (2013) *"A new image fusion performance metric based on visual information fidelity"* **Information fusion** *14*(2): 127–135.

[36] Y. J. Rao, (1997) *"In-fibre Bragg grating sensors"* **Measurement Science & Technology** *8*(4): 355–375.

[37] G. Cui, H. Feng, Z. Xu, Q. Li, and Y. Chen, (2015) *"Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition"* **Optics Communications** *341*: 199–209.

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

[39] H. Li and X.-J. Wu, (2018) *"DenseFuse: A fusion approach to infrared and visible images"* **IEEE Transactions on Image Processing** *28*(5): 2614–2623.

[40] X. Luo, Y. Gao, A. Wang, Z. Zhang, and X.-J. Wu, (2021) *"IFSepR: A general framework for image fusion based on separate representation learning"* **IEEE Transactions on Multimedia**:

[41] H. Li, X.-J. Wu, and J. Kittler, (2021) *"RFN-Nest: An end-to-end residual fusion network for infrared and visible images"* **Information Fusion** *73*: 72–86.

[42] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, (2020) *"U2Fusion: A unified unsupervised image fusion network"* **IEEE Transactions on Pattern Analysis and Machine Intelligence** *44*(1): 502–518.

[43] D. Wang, J. Liu, X. Fan, and R. Liu, (2022) *"Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration"* **arXiv preprint arXiv:2205.11876**:

[44] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, (2022) *"PIAFusion: A progressive infrared and visible image fusion network based on illumination aware"* **Information Fusion** *83*: 79–92.

[45] L. Tang, J. Yuan, and J. Ma, (2022) *"Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network"* **Information Fusion** *82*: 28–42. DOI: 10.1016/j.inffus.2021.12.004.

[46] Z. Zhu, X. Yang, R. Lu, T. Shen, X. Xie, and T. Zhang, (2022) *"Clf-Net: Contrastive learning for infrared and visible image fusion network"* **IEEE Transactions on Instrumentation and Measurement** *71*: 1–15. DOI: 10.1109/TIM.2022.3203000.

[47] L. Tang, H. Zhang, H. Xu, and J. Ma, (2023) *"Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity"* **Information Fusion**: 101870. DOI: 10.1016/j.inffus.2023.101870.