Searching a Hierarchically Aggregated Fusion Architecture for Fast Multi-Modality Image Fusion

Risheng Liu*
Dalian University of Technology
rsliu@dlut.edu.cn

Zhu Liu
Dalian University of Technology
liuzhu_dlut@mail.dlut.edu.cn

Jinyuan Liu
Dalian University of Technology
atlantis918@hotmail.com

Xin Fan
Dalian University of Technology
xin.fan@dlut.edu.cn

ABSTRACT

Multi-modality image fusion refers to generating a complementary image that integrates typical characteristics from source images. In recent years, we have witnessed the remarkable progress of deep learning models for multi-modality fusion. Existing CNN-based approaches strain every nerve to design various architectures for realizing these tasks in an end-to-end manner. However, these handcrafted designs are unable to cope with the high demanding fusion tasks, resulting in blurred targets and lost textural details. To alleviate these issues, in this paper, we propose a novel approach, aiming at searching effective architectures according to various modality principles and fusion mechanisms. Specifically, we construct a hierarchically aggregated fusion architecture to extract and refine fused features from feature-level and object-level fusion perspectives, which is responsible for obtaining complementary target/detail representations. Then by investigating diverse effective practices, we composite a more flexible fusion-specific search space. Motivated by the collaborative principle, we employ a new search strategy with different principled losses and hardware constraints for sufficient discovery of components. As a result, we can obtain a task-specific architecture with fast inference time. Extensive quantitative and qualitative results demonstrate the superiority and versatility of our method against state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Hierachically aggregated fusion architecture, fusion-oriented search space, collaborative architecture search, multi-modality fusion

ACM Reference Format:
Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan. 2021. Searching a Hierarchically Aggregated Fusion Architecture for Fast Multi-Modality Image Fusion. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475299

*Corresponding author: Risheng Liu.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475299

1 INTRODUCTION

Multi-modality image fusion, a hot topic in the multimedia and computer vision community, targets to provide a fuse image keeping the complementary advantages of different modality images. The fused images have strong ability to describe various scenes with comprehensive information, which presents better visual understanding, and also can be applied to subsequent high-level vision tasks, *e.g.*, object detection [4], salient detection [31, 42] and tracking, to name a few. Typically, Infrared and Visible Image Fusion (IVIF) and Medical Image Fusion (MIF) play an irreplaceable role in the field of image fusion. To solve these tasks, a large number of approaches have been developed in the past few decades. These methods can be roughly divided into conventional framework-based [12, 25, 41] and deep learning-based [15, 36, 37]. In methods based on the conventional fusion framework, multi-scale transform [14, 41], sparse representation [22, 33, 34], subspace [12], optimization model [8, 27, 28, 35] have been proposed for realizing multi-modality image fusion tasks. For instance, a mass of methods [5, 14, 39] are based on multi-scale transform technique. Although these methods applied well, it still remains limitations. The handcrafted feature extraction of inputs and manual fusion rule designing make the algorithms more complex and time-consuming.

In recent years, widespread attention has been drawn on deep learning schemes, due to its strong ability to extract deep salient features. To this end, some multi-modality image fusion methods [15, 17, 24] take advantage of deep learning network to represent features more comprehensive and more robust. In literature, a shallow CNN [32] is used to estimate a weight map for MIF fusion. Li *et. al.* proposed a IVIF fusion method [15] based on several dense blocks which are employed to extract features and $\ell 1$ norm to fuse the intermediate features. Apart from that, Ma *et. al.* [36, 37] introduced various Generative Adversarial Networks (GAN) for IVIF task. Lately, some methods (*e.g.,* PMGI [50], FusionDN [46] and U2Fusion [45]) aimed to construct unified architectures by investigating the proportional maintenance of latent information based on dense architectures. Recently, high/low frequency decomposition technique [18, 26, 51] was leveraged, that can exact structural edges and base background exactly for IVIF task.

Lately, Neural Architecture Search (NAS) methodology has made significant progress, which can be divided into three categories,

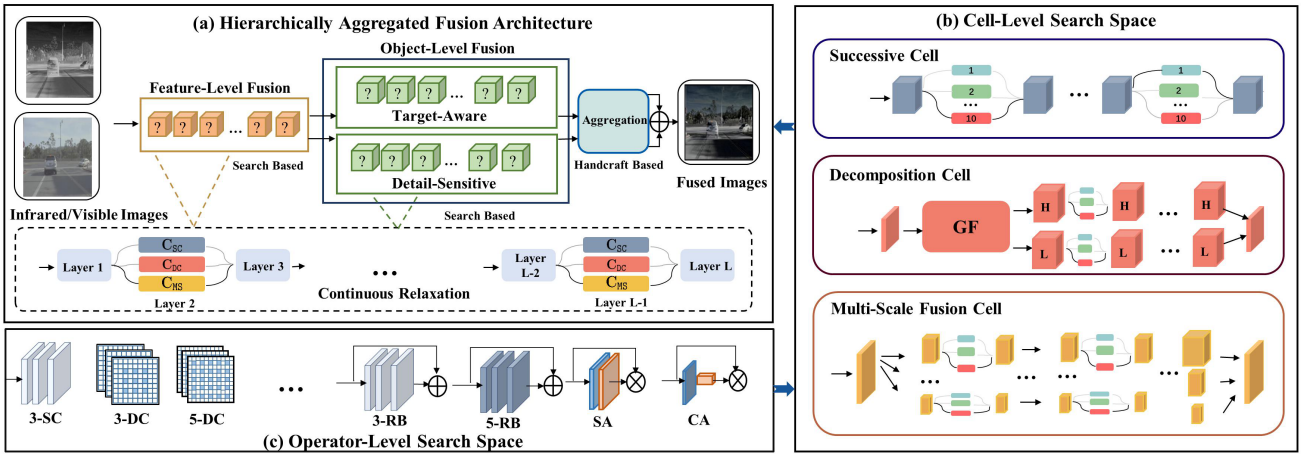

Figure 1: Schematic of the main components of our proposed architecture and search space. In subfigure (a), we illustrate the hierarchically aggregated fusion architecture, including fusion-level fusion, target/detail-specific object-level fusion and the aggregation module. We plot the search space based on three principled cells and several fundamental operators in (b) and (c). As indicated by two blue arrows, these cells are composed of operator-level search space and the whole architecture can be considered as the integration of cells. "GF" denotes the guided filters. "H" and "L" represent the high/low frequency features.

early evolutionary algorithms [44], reinforcement learning based [6] and differentiable gradient-based search [23]. Specifically, differentiable schemes [23, 30] have been widely used for various vision tasks, such as image restoration [26, 29, 49] and classification [23]. Recently, NAS schemes provide the tools for multi-modality fusion [1, 40]. These methods pursuit the accuracy of final prediction and ignore the fusion procedure. Primitive search spaces (*e.g.*, separable convolutions and pooling) neglect of task-specific domain knowledges that make NAS stay at an initial stage. Thus, designing a proper search scheme requires a good awareness of exploring the characteristics of fusion images.

To partially overcome the above limitations, this work proposes a hierarchically aggregated architecture rather than using simple networks to construct a macrostructure for various fusion tasks. Concretely, we first present a feature-level fusion module to obtain initial fused result. Then we investigate the typical property of different modality images and design an object-level fusion module to refine the results based on the inhere information of target/detail object. By introducing the aggregation module, we can obtain a complementary and comprehensive fused result with distinct targets and abundant textual details. Furthermore, different from current handcrafted CNN schemes, we construct the structure based on the proposed efficient and flexible search space. Taking fusion-specific knowledge into consideration, we integrate high/low frequency decomposition and multi-scale fusion mechanisms and efficient blocks (*e.g.*, residual, dense and attention modules) into the search space. Then, we provide a collaborative learning strategy, a new solution based on differentiable search, to discover the whole architecture from different modules progressively that guarantees the sufficient search of each principled component. In different stages, various kinds of losses are used to constrain the basic role of these modules. The main contributions of this manuscript can be summarized as three folds:

- Targeting to address the major stumbling blocks in CNN-based methods, we first construct a hierarchically aggregated architecture to obtain complementary fused images, integrating by the feature-level fusion and inhere modality-oriented properties.
- We propose a fusion-oriented search space that leverages the fusion principles including high/low frequency representation and multi-scale mechanisms with introducing effective fundamental operations to facilitate image fusion.
- A collaborative search strategy is designed to search the whole architecture from modular level progressively with the assistance of different principled losses and hardware-constraints. The proposed search strategy ensures the design principle of structure can be fully characterized.

2 THE PROPOSED METHOD

2.1 Hierarchically Aggregated Fusion

As aforementioned in Section. 1, current heuristic deep learning-based methods come across a few common problems. (i) most proposed architectures are unsophisticated and fail to take full use of the different statistics of modality information, and hence, the blurred edges or unclear targets may emerge on the fusion result. (ii) as a consequence of computational complexity and large parameter quantities, which leads to most existing fusion approaches are often less competitive in terms of time.

To address these issues, we first propose a hierarchically aggregated fusion architecture, decoupled into three vital principled hierarchies, *i.e.,* feature-level fusion, object-level fusion and aggregation mechanism, shown in the subfigure (a) of Fig. 1. Noting that this hierarchical architecture can be constructed by autonomic search, considered as a super-network (*i.e.,* the integration of various latent architectures). We propose a dominated module, named

Feature-level Fusion Module (FFM), aiming to design a general fusion module to extract, reconstruct features and generate fused images. In detail, FFM consists of a series of candidate cells, besides candidate operations, which provides a more general but flexible architecture compared with the manually design of dense or residual blocks, shown in the last row of subfigure (a). The cell-level continuous relaxation at l-th layer can be implemented with weighted sum of outputs, that makes the whole architecture differentiable and is computed with

$$\mathbf{F}^{l+1} = \sum_k \beta_l^k \mathbf{C}^k(\mathbf{F}^l, \boldsymbol{\alpha}^k), \tag{1}$$

where we denote $\mathbf{F}^l$, $\mathbf{F}^{l+1}$ and $\mathbf{C}^k$ as the input, output and candidate cell respectively. $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ denote the continuous weights of cells and inner operations. More details about search space (*e.g.,* cells and operators) and strategy will discussed in following subsections.

Investigating different characteristics and complementary information from source images based on architecture design is a primary difference compared with existing CNN methods. The target information and visible/structural details from multi-modality images also need a scheme to extract object-specific information in different ways. Therefore, we introduce the Object-level Fusion Module (OFM) to refine fused images.

In OFM, we decouple the fusion object into two parts, *i.e.*, target-aware object and detail-sensitive object. In detail, as for target information, thermal radiation often represents the targets in infrared images with obvious distinction of pixel intensity. On the other hand, as for detail information, visible observations with more texture details can improve scenario awareness. Supervised by different principled losses, designed in following subsection, the goal of designing target/detail-object module can be achieved.

Lastly, we also construct an aggregation module to confirm retention degrees of different features. It is implemented by spatial attention mechanism with two layers of convolutions and sigmoid function to generate a weight mask $\mathbf{M}$ for aggregating two target/detail-specific outputs. Noting that, this aggregation module is pre-defined without architecture search.

2.2 Fusion-Oriented Search Space

In this part, we design a flexible fusion-oriented search space elaborately from two degrees of external cells and internal operators, as shown in subfigure (b) and (c) in Fig. 1. It sets the tone for discovering a desired fusion structure based on whole architecture.

2.2.1 Fusion-Specific Cell Architectures. We compose three types of cells into cell-level search space, which provides the external latent structure construction of this hierarchical architecture, taking consideration of effective practices for fusion tasks, *i.e.*, high/low frequency representation and multi-scale fusion.

Successive Cell. Dealing with hardware-specific scenario, many NAS methods [11, 13] adopt successive cell (also known as "choice blocks") to reduce search cost. $\mathbf{N}_c$ choice blocks are connected in series to composite a cell (denoted as $\mathbf{C}_{\mathsf{SC}}$). Naturally, several operators (defined in search space) are embedded in each block with continuous weights. Finally, operator with the largest weight is selected in each block. The whole structure is shown in the first row of subfigure (b), where boxes with different colours denotes the operators.

Decomposition Cell. The basic cell (*i.e.*, $\mathbf{C}_{\mathsf{DC}}$) is composited by three parts, *i.e.*, feature decomposition, parallel streaming and feature fusion. Deep image-guided filters [20, 21] are performed to decouple high/low frequency of features. Focusing on two kinds of feature, multiple neural blocks are connected at each row. Every block indicates one possible operation. We define this cell has $\mathbf{N}_d$ blocks. Note that blocks at two rows of cells share the same architecture due to the hardware computation limitation. Finally, *conv* 1×1 layer is performed to fuse high/row frequency components. The whole architecture is shown the second row of subfigure (b).

Multi-Scale Fusion Cell. Based on multi-scale fusion principles, we construct the following cell. Detailed structure is shown in the last row in subfigure (b). Multi-scale fusion cell (*i.e.*, $\mathbf{C}_{\mathsf{MS}}$) comprises four key elements, *i.e.*, downsampling, parallel streaming, upsampling and fusion. We utilize strided convolution to reduce the resolution of features. In this paper, we define the resolutions of multi-scale features are 1x, 0.5x and 0.25x. Then same with decomposition cell, each row of cell contains $\mathbf{N}_m$ blocks. After upsampling resized features, we concatenate and fuse them by *conv* 1×1 layer.

2.2.2 Effective Operators. Constructing an effective operation-level search space (denoted as $\mathbf{A}$) is one of vital aspects for architecture search. The current search spaces for low-level vision [44, 49] are mainly designed with primitive operators (*e.g.,* separable *conv* 3×3), where such basic operators maybe not sufficient or suitable for guiding to optimal architectures. Therefore, we investigate more effective operators in existing fusion tasks for the construction of above three principled cells. We introduce the following ten candidate operators:

- 3×3 Separable *Conv* (3-SC)
- 3×3 Dilated *Conv* (3-DC)
- 3×3 Residual Blocks (3-RB)
- 3×3 Dense Blocks (3-DB)
- Spatial Attention (SA)
- 5×5 Separable *Conv* (5-SC)
- 5×5 Dilated *Conv* (5-DC)
- 3×3 Residual Blocks (5-RB)
- 5×5 Dense Blocks (5-DB)
- Channel Attention (CA)

The main concrete illustration of operators are shown in subfigure (c). In detail, we discard skip connection, pooling and *zero/none* operations, which is not suitable for fusion tasks.

2.3 Collaborative Search Strategy

In this part, we introduce three key denotations to perform continuous relaxation, *i.e.*, vectorized form $\boldsymbol{\alpha} := \{\boldsymbol{\alpha}_F, \boldsymbol{\alpha}_O\}$ as internal architecture for FFM and OFM, $\boldsymbol{\beta} := \{\boldsymbol{\beta}_F, \boldsymbol{\beta}_O\}$ as external architecture and $\boldsymbol{\omega} := \{\boldsymbol{\omega}_F, \boldsymbol{\omega}_O\}$ as weight parameters. Noting that we define $\boldsymbol{\omega}_O$ denotes the weights of OFM and aggregation module. After the definition of search space, we propose a collaborative search strategy to discover the whole architecture progressively, which is a new solution compared with naive search strategy [23]. This is because current gradient-based search paradigms only optimize $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ and $\boldsymbol{\omega}$ straightforwardly, neglecting our proposed fusion principles and having difficulty to utilize this complex search space.

2.3.1 Bilevel Learning with Hardware-Aware Constraint. In order to constrain the computational cost of searched modules, we leverage inference latency as the hardware-aware constraint. We introduce

this latency-constrained term as a regularization loss to composite the validation loss in search phase. The search process that discovers low-latency structures from super-network can be formulated as:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \mathcal{L}_{\mathtt{val}}(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{\omega}^{\star}) + \lambda(\mathrm{LAT}(\boldsymbol{\alpha},\boldsymbol{\beta})) \\ \text{s.t.} \quad & \boldsymbol{\omega}^{\star} = \arg\min_{\boldsymbol{\omega}} \mathcal{L}_{\mathtt{train}}(\boldsymbol{\omega};\boldsymbol{\alpha},\boldsymbol{\beta}), \end{aligned} \tag{2}$$

where this differentiable bilevel optimization can be utilized to search our proposed modules with task-specific losses. $\mathcal{L}_{\mathtt{val}}$ and $\mathcal{L}_{\mathtt{train}}$ are denoted as validation and training losses. We utilize * to denote the optimal weights. Specifically speaking, the function LAT can be calculated by weighted linear sum of operations:

$$\mathrm{LAT}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{l}\sum_{k}\sum_{i} \beta_l^k \alpha_i^k \mathrm{LAT}(\mathrm{op}_i), \mathrm{op}_i \in \mathbf{A}, \tag{3}$$

where we denote β_l^k as the relaxation weight of k-th cell in l-th layer. α_i^k is denoted as the i-th operation weight.

2.3.2 Collaborative Architecture Search. The Alg. 1 illustrates the whole collaborative search strategy performed for the hierarchical architecture. We detail the original strategy with two key considerations. Firstly, the well-behaved FFM can provide a fused image with sufficient features to help the search process of OFM from global to the local. In other words, The performance of OFM benefits greatly from a favourable FFM. Secondly, the search for external architecture can guide the construction of inner operators from macro to micro. That indicates the relationship of search of FFM and OFM is collaborative. Thus, we present a new progressive architecture search strategy to solve Eq. (2).

In detail, the search strategy of whole architecture can be decoupled into two stages, the discovery of FFM and OFM respectively. We first perform the strategy on FFM using first-order approximation [23]. $\mathcal{L}_{\mathtt{val}}^{F'}$ denotes the validation loss with latency constraint. After obtaining the optimal structure of FFM with fixed architecture, we perform the learning process of OFM. Note that, network parameters $\boldsymbol{\omega_F}$ are still optimized to preserve the consistency of whole architecture. Specifically, as for the search for each module, the structure is searched from external to internal, ***i.e.,*** updating $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ alternatively, as shown in the steps (4-13) in Alg. 1.

2.3.3 Loss Functions. Four types of losses are used to search/train our candidate networks.

The intensity loss, aiming to capture the contrast information (*e.g.*, thermal radiation highlighted in pixel intensity) is introduced, which is defined as

$$\mathcal{L}_{int} = \frac{1}{HW}\|\mathbf{I}_A - \mathbf{I}_B\|_2^2, \tag{4}$$

where the height and width are denoted as H and W.

Recently, perceptual loss is leveraged to measure the discrepancy in feature domain (*e.g.*, VGG network ϕ), which can represent the difference of pixel distinct and global targets. We define the perceptual loss as:

$$\mathcal{L}_{per} = \frac{1}{C_i H_i W_i}\|\phi_i(\mathbf{I}_A) - \phi_i(\mathbf{I}_B)\|_2^2, \tag{5}$$

where C is the channel number and i represents the layer index.

Structural similarity between source images and fused images is also considered by SSIM metric. SSIM loss $\mathcal{L}_{ssim}$ can be constructed as $1 - \mathrm{SSIM}(\mathbf{I}_A, \mathbf{I}_B)$.

Algorithm 1 Collaborative Architecture Search

Require: Search space **A**, initial $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$, hyper-parameter *n, m, t, s* and necessary parameters.
Ensure: The Searched optimal architecture.
1: Initialize $\boldsymbol{\alpha}_u$ by the normal random function.
2: %% Stage 1: Search for FFM.
3: **for** *m* epochs **do**
4: %% Search from cell-level space (updating $\boldsymbol{\beta}_F$ and $\boldsymbol{\omega}_F$).
5: **for** *t* epochs **do**
6: $\boldsymbol{\omega}_F \leftarrow \boldsymbol{\omega}_F - \nabla_{\boldsymbol{\omega}_F}\mathcal{L}_{\mathtt{train}}^{F}(\boldsymbol{\omega}_F, \boldsymbol{\alpha}_F, \boldsymbol{\beta}_F)$;
7: $\boldsymbol{\beta}_F \leftarrow \boldsymbol{\beta}_F - \nabla_{\boldsymbol{\beta}_F}\mathcal{L}_{\mathtt{val}}^{F'}(\boldsymbol{\beta}_F; \boldsymbol{\omega}_F - \nabla_{\boldsymbol{\beta}_F}\mathcal{L}_{\mathtt{train}}^{F}, \boldsymbol{\alpha}_F)$;
8: **end for**
9: %% Search based on operators (updating $\boldsymbol{\alpha}_F$ and $\boldsymbol{\omega}_F$).
10: **for** *s* epochs **do**
11: $\boldsymbol{\omega}_F \leftarrow \boldsymbol{\omega}_F - \nabla_{\boldsymbol{\omega}_F}\mathcal{L}_{\mathtt{train}}^{F}(\boldsymbol{\omega}_F, \boldsymbol{\alpha}_F, \boldsymbol{\beta}_F)$;
12: $\boldsymbol{\alpha}_F \leftarrow \boldsymbol{\alpha}_F - \nabla_{\boldsymbol{\alpha}_F}\mathcal{L}_{\mathtt{val}}^{F'}(\boldsymbol{\alpha}_F; \boldsymbol{\omega}_F - \nabla_{\boldsymbol{\omega}_F}\mathcal{L}_{\mathtt{train}}^{F}, \boldsymbol{\beta}_F)$;
13: **end for**
14: **end for**
15: %% Stage 2: Search for OFM, fixed weights of $\{\boldsymbol{\beta}_F, \boldsymbol{\alpha}_F\}$.
16: **for** *n-m* epochs **do**
17: %% Omitted. Similar with the search process of FFM.
18: **end for**
19: **return** Derive architecture based on $\boldsymbol{\alpha}_F^{\star}, \boldsymbol{\beta}_F^{\star}, \boldsymbol{\alpha}_R^{\star}, \boldsymbol{\beta}_R^{\star}$.

Moreover, gradient information of images always characterizes texture details and scenario structure. Therefore, we use the gradient loss to constrain these textual factors, *i.e.*,

$$\mathcal{L}_{grad} = \frac{1}{HW}\|\nabla\mathbf{I}_A - \nabla\mathbf{I}_B\|_2^2, \tag{6}$$

where we denote ∇ as the gradient operator.

2.3.4 Configuration for Infrared-Visible Image Fusion. It is obvious that the purpose of this task is to preserve contrast intensities of targets from infrared images and textural details (shown obviously in gradient domain) from visible ones. We introduce this principle in the search of OFM. Denoted $\mathbf{I}_{in}, \mathbf{I}_{vis}, \mathbf{I}_f^{'}$ as the inputs of infrared, visible images and the output of FFM, we utilize $\mathcal{L}_{int} + \mathcal{L}_{ssim}$ with source image pairs as $\mathcal{L}_{\mathtt{val}}^{F}$ and $\mathcal{L}_{\mathtt{train}}^{F}$. We follow the feature adaption weights [45] to control the preservation degrees. Then we denote $\mathbf{I}_f, \mathbf{I}_t, \mathbf{I}_d$ as the fused images, outputs of target-object and detail-object modules. As for OFM, the definition of $\mathcal{L}_{\mathtt{val}}^{O}$ and $\mathcal{L}_{\mathtt{train}}^{O}$ is $\mathcal{L}_{per}(\mathbf{I}_t, \mathbf{I}_{in}) + \mathcal{L}_{grad}(\mathbf{I}_d, \mathbf{I}_{vis}) + \sigma_1\mathcal{L}_{int}(\mathbf{I}_f, \mathbf{I}_{in}) + \sigma_2\mathcal{L}_{ssim}(\mathbf{I}_f, \mathbf{I}_{vis})$. $\{\sigma_1, \sigma_2\}$ are also adaptive weights generated by VGG features.

3 EXPERIMENTAL RESULTS

3.1 Infrared-Visible Image Fusion

3.1.1 Datasets. As for IVIF task, we leveraged part images from two public released datasets (*i.e.*, *TNO* and *RoadScene*) to search, train and test our networks. Data augmentation strategy was adopted in our schemes to enlarge our training datasets. We used substantial 64×64 image patches by cropping and clipping source images randomly in our training process of both tasks. Finally, we selected 37 images form *TNO*, 26 images from *RoadScene* as testing pairs to make a comparison with other methods respectively.

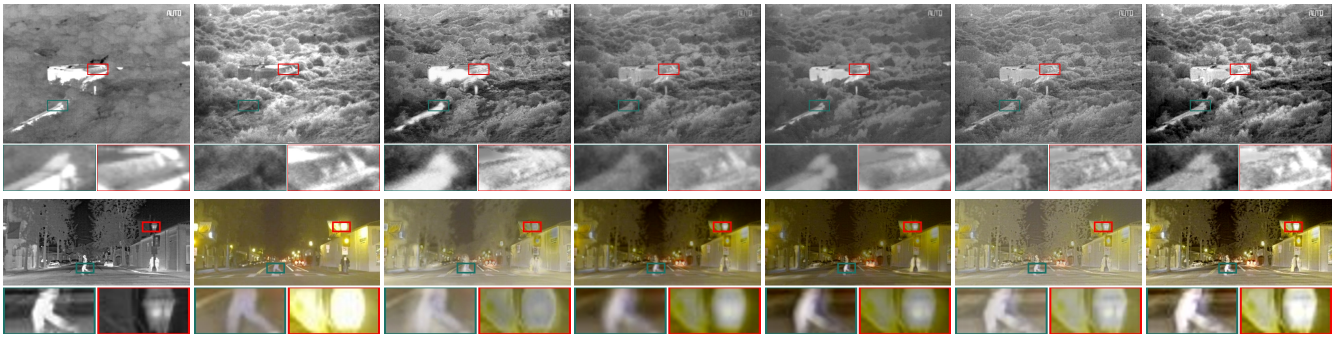

Figure 2: Qualitative comparison of our scheme with four state-of-the-arts on two groups typical infrared-visible images from *TNO* and *RoadScene* datasets respectively.

Table 1: Quantitative comparison with a series of completive CNN-based methods on *TNO* and *RoadScene* datasets.

Dataset	Metrics	PMGI	FGAN	Dense	DDcGAN	FusionDN	RFN	GANMcC	Nest	DID	U2Fusion	Ours$_F$	Ours$_{F+O}$
TNO	SD	9.715	8.659	9.487	**10.295**	9.926	9.618	9.263	9.528	9.837	9.731	9.817	10.036
	VIF	0.857	0.628	0.798	0.691	0.850	0.803	0.681	0.862	0.828	0.789	0.855	**0.869**
	CC	0.555	0.478	**0.583**	0.529	0.547	0.575	0.573	0.537	0.551	0.573	0.581	0.582
	SCD	1.614	1.194	1.761	1.565	1.770	1.772	1.662	1.678	1.794	1.752	1.813	**1.832**
RoadScene	SD	10.111	10.185	10.105	10.339	10.299	10.145	10.223	10.468	**11.004**	10.249	10.295	10.682
	VIF	0.841	0.614	0.797	0.612	0.797	0.771	0.736	0.832	0.821	0.771	0.811	**0.843**
	CC	0.597	0.574	0.673	0.620	0.649	0.659	0.656	0.640	0.653	0.647	0.667	**0.679**
	SCD	1.291	1.027	1.698	1.559	1.782	1.713	1.553	1.703	1.782	1.638	1.828	**1.889**

Table 2: The parameters, FLOPs and inference time compared with recent competitive CNN-based methods.

Methods	FGAN	Dense	DDcGAN	FusionDN	RFN	GANMcC	Nest	U2Fusion	Ours$_F$	Ours$_{F+O}$
SIZE(M)	0.925	**0.074**	1.098	1.163	10.93	1.864	10.934	0.659	0.941	1.131
FLOPs (G)	497.76	**48.96**	896.84	933.57	-	1002.56	-	366.34	125.94	181.86
TIME (S)	0.124	0.251	0.211	1.162	0.239	0.246	11.762	0.123	**0.075**	0.121

3.1.2 Searching Details. To alleviate the redundant calculation, we only defined basic hyper-parameters empirically to initialize our search process. The FFM of collaborative architecture consists of 2 layers for candidate cells. Each type of cells has 2 nodes (*i.e.*, $N_c = N_d = N_m = 2$). As for OFM, several stem (*i.e.*, target/detail object sub-module) has one layer and two candidate blocks are defined in each layer. Then we performed our search strategy (*i.e.*, Alg. 1) on hybrid *TNO-RoadScene* datasets. Specifically, we divided evenly the dataset into two parts. One groups are leveraged for training ω, applying on training loss $\mathcal{L}_{train}$. Another part of images is used to update architectures α and β. As for local search, we optimize α and β with five epochs (*i.e.*, $t = s = 5$) alternatively. As for global search, we employed this strategy with 12 epochs (*i.e.*, m =12 and n = 24) respectively. With batch size of 8, we searched the whole super-net with 200 epochs, introducing cosine the annealing strategy with SGD optimizer to decay learning rate from 0.001.

3.1.3 Training Details. Our training strategy also follows with collaborative principle, rather than straightforward end-to-end training. At the first stage, we trained the derived FFM with 100 epochs, in order to guarantee its module to generate coarse fused images. At second stage, we trained the whole network (including OFM) end-to-end with 1000 epochs, using Adam optimizer with initial learning rate of 0.0005. We utilized the same training losses of search process to guide our training phase. Note that, grayscale images are as inputs of our network. As for color images, such as visible images in *RoadScene* and PET images, we transformed them from RGB channels to YCbCr channels firstly. Our method is based on the PyTorch framework and runs on a NVIDIA GTX1070 GPU. We evaluated the effectiveness and efficiency of our scheme by comparing with a number of recent CNN-based methods (*i.e.*, PMGI [50], FGAN [37], DDcGAN [36], FusionDN [46], U2Fusion [45], RFN [19], GANMcC [38], Nest [16], DID [51] and Dense [15]).

3.1.4 Qualitative Comparisons. As for qualitative comparisons, we carried out experiments on two typical images and show the visual

results produced by various methods in Fig. 2. That demonstrates our competitive performance and superiority of our paradigm. We can conclude three main advantages compared with other methods. First, the apparent target can be preserved effectively. The thermal radiation information of targets is highlighted, such as the house in the first row and person in the second row. However the results produced by RFN and DDcGAN remain rich textures, the targets are not clear enough. Moreover, our method contains more texture details from visible images. For instance, the structure of street lamp in the second row is restored exactly. Furthermore, less artifacts are produced by our methods. Obviously, the boundary of person (the second row) are burred in other results.

3.1.5 Quantitative Comparisons. In addition to visual evaluation, we also conducted quantitative comparisons, using four objective metrics, including Standard Deviation (SD), Visual Information Fidelity (VIF) [10], Correlation Coefficient (CC) and Sum of Correlation of difference (SCD) [2]. Table. 1 reports the numerical comparisons with these CNN-based methods. According to these statistical results, our scheme achieves comparable performances on both datasets *TNO* and *RoadScene*. The largest results of VIF and SCD indicate that our method can preserve texture details and fuse visual sissified results. Furthermore, our method is also correlated high with source images, measured by CC. Higher SD points to significant contrast of our methods.

3.1.6 Hardware-Constrained Experiment. To verify the improvement of inference time constrained by hardware latency, we carried out the computation efficiency experiment, calculated on ten images from TNO with size 620×448. Table. 2 compares the parameters, FLOPs and inference time with recent methods. The strategy-oriented methods (*e.g.*, Dense, RFN and Nest) consider more about the fusion rule and strategy (*e.g.*, ℓ_1 and nuclear strategy). Though these methods have much smaller model or FLOPs, the inference time of them is not faster. Our method considers the GPU latency of each operators and performs these constrain into the construction of architectures. Thus, our method can achieve fast inference compared with other methods. In the following, we also verify the control effects of λ.

Table 3: Ablation study of hyper-parameter λ on *TNO*.

λ	Cells	FLOPs	Paramters	Time	SD
0	$\{C_{DC}, C_{MS}, C_{SC}, C_{MS}\}$	195.62	1.223	0.362	10.01
0.02	$\{C_{MS}, C_{MS}, C_{SC}, C_{SC}\}$	280.90	1.981	0.198	9.97
0.05	$\{C_{SC}, C_{MS}, C_{SC}, C_{SC}\}$	181.86	1.131	0.121	10.04
0.08	$\{C_{SC}, C_{MS}, C_{SC}, C_{SC}\}$	202.27	1.583	0.118	9.59
0.2	$\{C_{SC}, C_{SC}, C_{SC}, C_{SC}\}$	78.95	0.279	0.077	9.81

3.2 Ablation Study

Firstly, we verified the impacts of hardware-constraint parameter λ and the reuslts are listed in Table. 3. One could observe that the larger λ will guide to the faster inference time. However, the FLOPs and model size cannot establish strong correlation with λ. The possible reason is that we introduce the GPU latency of each operators (*i.e.*, the running time of operators) as the constraint without considering the influence of FLOPs. Furthermore, larger λ will lead to selecting lightweight external cells (*e.g.*, C_{SC}) [1].

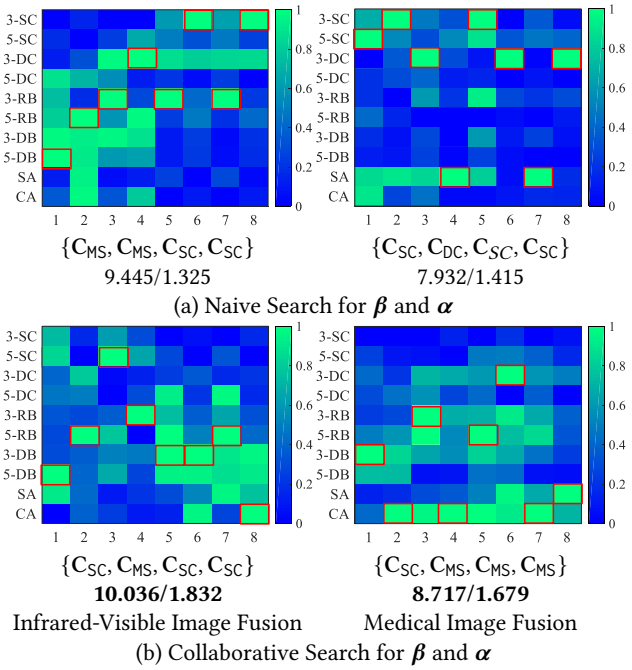

Figure 3: Heatmaps of searched architectures based on naive search (searching α and β simultaneously) and proposed collaborative search strategy for two fusion tasks. Red boxes indicate the final architecture. The SD/SCD metrics are listed in the below of heatmaps.

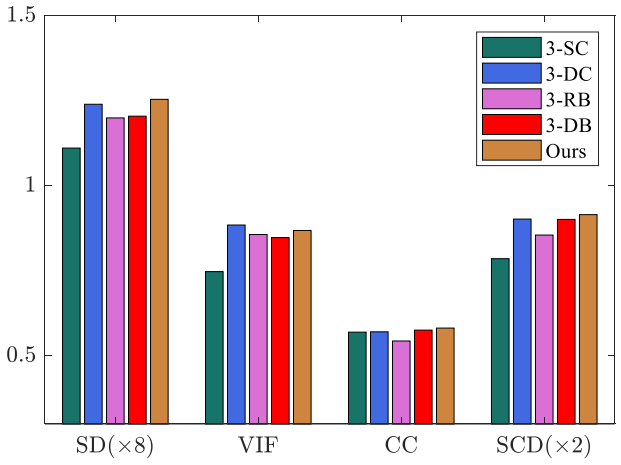

Figure 4: Comparison of heuristic single operator composited architectures with same external structure. Four metrics of these networks are plotted. $(\times\cdot)$ means the zoom scale.

[1] C_{SC}, C_{DC} and C_{MS} denotes the successive, decomposition and multi-scale cells.

Furthermore, we explored the impacts of proposed search/train strategy, compared with naive search strategy and different losses. Naive strategy considers the architecture as an entire network without the principle guidance, *i.e.*, updating architecture $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ and weight parameters $\boldsymbol{\omega}$ alternatively. Obviously, we can observe that numerical results of naive search are much lower than results under our proposed strategy in Fig. 3. Moreover, the architectures searched by naive search tend to use basic convolutions. Our final architecture for IVIF contains several dense blocks, which is utilized widely in previous works. We extend our search scheme to deal with MIF task. Architecture for MIF contains more attentions, in order to extract obvious structural/functional information. More details are reported in the following subsection. We also verified

Table 4: Ablation study of different losses on *TNO*.

losses	SD	VIF	CC	SCD
w/o $\mathcal{L}^F_{train}$	9.112	0.676	0.578	1.612
w/o object-level losses	9.769	**0.887**	0.561	1.828
ours	**10.036**	0.869	**0.582**	**1.832**

the effectiveness of internal operator search, which we keep the external architecture and replace the inner operators using single 3-SC, 3-DC, 3-RB, 3-DB respectively. As shown in Fig. 4, the heuristic operator-based architectures cannot obtain the best performance. Obviously, we can observe that the searched architecture can improve these numerical results. The performances of training strategy ablation are reported in Table. 4, discarding the pre-training of FFM will damage the final performance. Object-level losses can help the whole architecture to consider more visible details and target information.

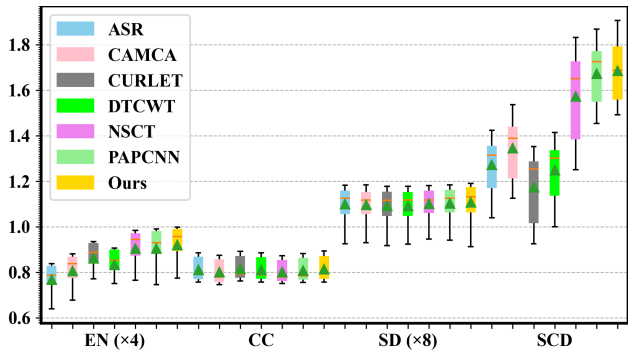

Figure 5: Qualitative comparison on one groups of MRI-PET images with various medical fusion algorithms. $(\times\cdot)$ means the zoom scale.

3.3 Medical Image Fusion

In this part, we extended our search strategy with hierarchical aggregated architecture to address MIF task. As for medical image fusion, we selected one representative scenario (*i.e.*, MRI-PET fusion) to conduct our experiment. We collected 136 MRI-PET pairs of medical images from Harvard medical websites and used 16 images as testing pairs. Six remarkable methods (*i.e.*, ASR [34], CSMCA [33], CURVELET [47], DTCWT [5], NSCT [3] and PAPCNN [48]) are employed to make a comparison.

3.3.1 Search Configuration for MRI-PET Image Fusion. As for medical image fusion, MRI images (*i.e.*, $\mathbf{I}_{mri}$) contain structural texture information, such as the details of organ. PET images (*i.e.*, $\mathbf{I}_{pet}$) are one type of functional images that show metabolic representation. The major difference with infrared-visible images is the pixel intensity of PET images is distinct with MRI images, that would damage the fused structural information coming from MRI images. Therefore, we need to preserve the pixel intensity of MRI images and PET ones simultaneously. We set $\mathcal{L}^O_{\texttt{val}}$ and $\mathcal{L}^O_{\texttt{train}}$ as $\mathcal{L}_{per}(\mathbf{I}_t, \mathbf{I}_{mri}) + \mathcal{L}_{per}(\mathbf{I}_t, \mathbf{I}_{pet}) + \mathcal{L}_{grad}(\mathbf{I}_d, \mathbf{I}_{mri}) + \sigma_1 \mathcal{L}_{int}(\mathbf{I}_f, \mathbf{I}_{pet}) + \sigma_2 \mathcal{L}_{int}(\mathbf{I}_f, \mathbf{I}_{mri})$. Noting that, other searching and training settings are similar with the experiments performed in IVIF task.

3.3.2 Qualitative Comparisons. Fig. 6 shows the fused results on two kinds of brain-hemispheric transaxial sections. Among them, the results of ASR and CSMCA maintain the structural information (mainly from MRI) but have color distortion of PER images, which cannot extract sufficient functional information from PET images accurately. On the contrary, NSCT, PAPCNN and our method preserve clear edges, generate vivid texture details and have proper contrast. That is also reflected in the blow metrics of Fig. 6.

3.3.3 Quantitative Comparisons. Due to the particular characteristic of medical images, it requires us to measure the information remaining in fused images. The particularity comes from the clear representation with high contrast, such as the shapes, edges of organ and physiologically functional information. So we utilized entropy (EN) [43] additionally to measure the amount of information. We plotted the Fig. 5 to report mean and variance values in terms of these four metrics. Obviously, the results produced by our method have the largest amount of information, clear texture details and proper feature representation from source images.

3.4 Applications in Related Tasks

In this subsection, we investigate our searched network of IVIF task to support a series of related upstream applications, especially in severe dark night scenarios, collected from *RoadScene* dataset. **Salient Object Detection.** We adopted the remarkable BASNet [42] to carry out this experiment. Taking an example from Fig. 7, the hard light from lamp damages the detection of car region obviously. Furthermore, we confirm that the infrared information can obtain a better detection result with low-light condition. However, the reflection and blurs captured by infrared devices would introduce much artifacts. Moreover, recent methods focus more on infrared information (*e.g.*, DDcGAN) or visible details (*e.g.*, RFN) that cannot estimate the main natural object. Compared with U2Fusion, our method estimated the whole region without artifacts generated by distribution of traffic lights.

Depth Estimation. As a non-trivial byproduct, we also illustrated the effectiveness of our method to support depth estimation task, based on Monodepth2 [4] framework. Actually, recent depth estimation methods are trained with daytime road data set (*e.g.*, KITTI [9] and CityScapes [7]), that exists the gap between daytime and night

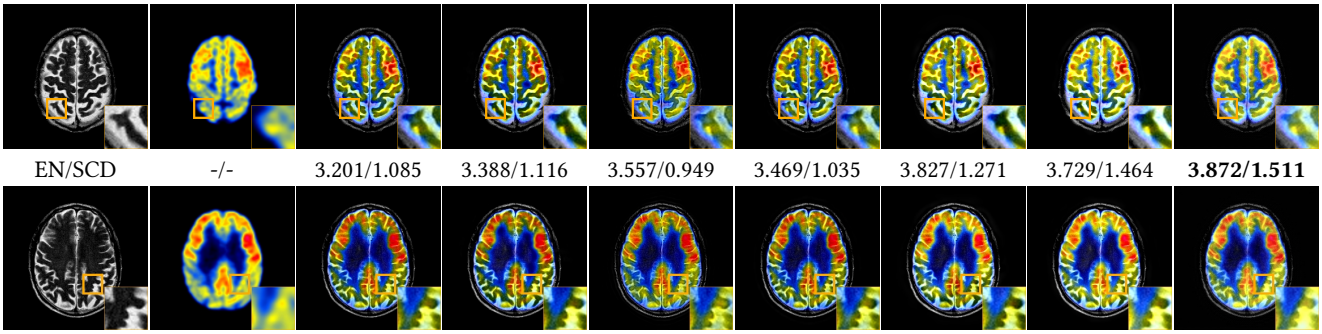

Figure 6: Qualitative comparison on two groups of MRI-PET images with various medical fusion algorithms.

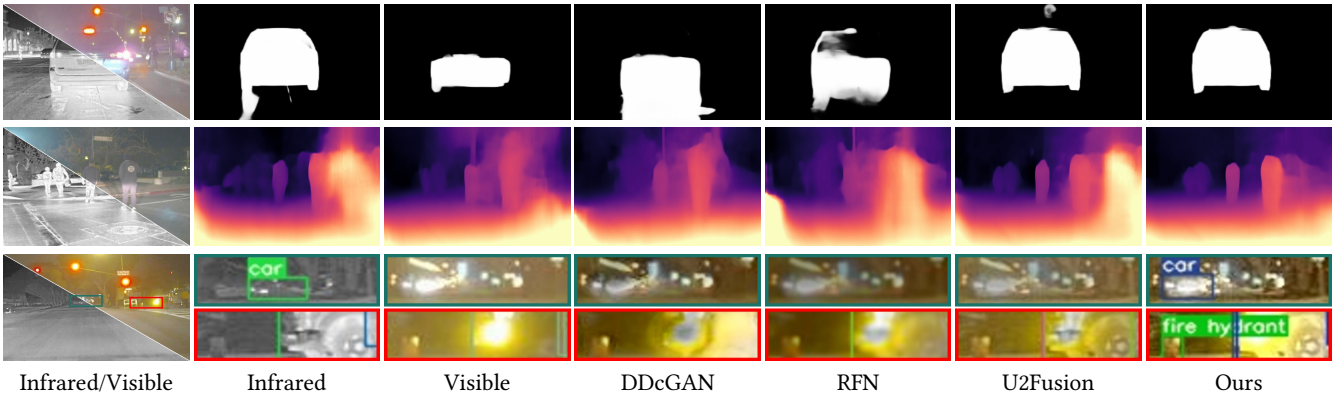

Figure 7: Visual results on salient object detection, depth estimation and object detection applications.

scenarios. The second row of Fig. 7 visualizes the depth maps estimated by various methods. We can observe that the depths of pedestrians have been estimated fail in the visible images and most of CNN methods. The redundant artifacts are shown in the results of infrared image and U2Fusion, degraded by the light reflection. On the contrary, our scheme can estimate a depth map with obvious shapes of pedestrians clearly, which provide a new assistance scheme for real-world depth estimation.

Object Detection. Subsequently, we demonstrated the significant improvement of our method on object detection, based on well-known YOLO-v4 network [4]. The last row of Fig. 7 shows the final detected results based on different modality images and recent state-of-the-arts. The clear differences are blown out by different boxes. Obviously, a car in the distance can be detected successfully based on infrared image and our result, where the fused result of our method can estimate the position exactly. However, detection based on other methods lost this object. Moreover, a fire hydrant is detected by the fused image of our method only, where the fused results of U2Fusion and ours preserve thermal radiation structure and visible details properly (*e.g.*, carriage wheels).

4 CONCLUSIONS

In this paper, we proposed a novel architecture search scheme for multi-modality fusion by integrating fusion task characteristics and efficient search strategy. Firstly, we constructed one hierarchically aggregated fusion architecture, composed of three principled modules, dedicated to feature-level fusion, object-level refinement and aggregation. Then we established a flexible task-oriented search space to construct latent architectures based on various external cells and internal operators. With the collaborative search strategy, restrained by principled losses and hardware constraints, we can obtain a fast and efficient task-specific architecture. The experiments performed on various benchmarks and multiply related applications corroborated the significance of our method.

ACKNOWLEDGEMENTS

This work is partially supported by the National Key R&D Program of China (2020YFB1313503), the National Natural Science Foundation of China (Nos. 61922019, 61733002, and 61672125), LiaoNing Revitalization Talents Program (XLYC1807088), and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Stefano Alletto, Shenyang Huang, Vincent Francois-Lavet, Yohei Nakata, and Guillaume Rabusseau. 2020. RandomNet: Towards Fully Automatic Neural Architecture Design for Multimodal Learning. *AAAI* (2020).

[2] V Aslantas and Emre Bendes. 2015. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-international Journal of electronics and communications* 69, 12 (2015), 1890–1896.

[3] Gaurav Bhatnagar, QM Jonathan Wu, and Zheng Liu. 2013. Directive contrast based multimodal medical image fusion in NSCT domain. *IEEE transactions on multimedia* 15, 5 (2013), 1014–1024.

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

[5] Liu Cao, Longxu Jin, Hongjiang Tao, Guoning Li, Zhuang Zhuang, and Yanfu Zhang. 2014. Multi-focus image fusion based on spatial frequency in discrete cosine transform domain. *IEEE signal processing letters* (2014), 220–224.

[6] Yun-Chun Chen, Chen Gao, Esther Robb, and Jia-Bin Huang. 2020. Nas-dip: Learning deep image prior with neural architecture search. *arXiv preprint arXiv:2008.11713* (2020).

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.

[8] Qinglei Du, Han Xu, Yong Ma, Jun Huang, and Fan Fan. 2018. Fusing infrared and visible images of different resolutions via total variation model. *Sensors* (2018), 3827.

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.

[10] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. 2013. A new image fusion performance metric based on visual information fidelity. *Information fusion* 14, 2 (2013), 127–135.

[11] Cong Hao, Xiaofan Zhang, Yuhong Li, Sitao Huang, Jinjun Xiong, Kyle Rupnow, Wen-mei Hwu, and Deming Chen. 2019. Fpga/dnn co-design: An efficient design methodology for 1ot intelligence on the edge. In *2019 56th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.

[12] Changtao He, Quanxi Liu, Hongliang Li, and Haixu Wang. 2010. Multimodal medical image fusion based on IHS and PCA. *Procedia Engineering* 7 (2010).

[13] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1314–1324.

[14] John J Lewis, Robert J O'Callaghan, Stavri G Nikolov, David R Bull, and Nishan Canagarajah. 2007. Pixel-and region-based image fusion with complex wavelets. *Information fusion* (2007), 119–130.

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

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

[17] Hui Li, Xiao-Jun Wu, and Josef Kittler. 2018. Infrared and visible image fusion using a deep learning framework. In *ICPR*. IEEE, 2705–2710.

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

[19] Hui Li, Xiao-Jun Wu, and Josef Kittler. 2021. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion* (2021).

[20] Ruoteng Li, Loong-Fah Cheong, and Robby T. Tan. 2019. Heavy Rain Image Restoration: Integrating Physics Model and Conditional Adversarial Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Shutao Li, Xudong Kang, and Jianwen Hu. 2013. Image fusion with guided filtering. *IEEE Transactions on Image processing* 22, 7 (2013), 2864–2875.

[22] CH Liu, Y Qi, and WR Ding. 2017. Infrared and visible image fusion method based on saliency detection in sparse domain. *Infrared Physics & Technology* 83 (2017), 94–102.

[23] Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2018. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055* (2018).

[24] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. 2021. Learning a Deep Multi-scale Feature Ensemble and an Edge-attention Guidance for Image Fusion. *IEEE Transactions on Circuits and Systems for Video Technology* (2021).

[25] Risheng Liu, Shichao Cheng, Yi He, Xin Fan, Zhouchen Lin, and Zhongxuan Luo. 2019. On the convergence of learning-based iterative methods for nonconvex inverse problems. *IEEE transactions on pattern analysis and machine intelligence* 42, 12 (2019), 3027–3039.

[26] Risheng Liu, Jinyuan Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. 2020. A Bilevel Integrated Model With Data-Driven Layer Ensemble for Multi-Modality Image Fusion. *IEEE Transactions on Image Processing* 30 (2020), 1261–1274.

[27] Risheng Liu, Xuan Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. 2021. A Hessian-free Interior-point Method for Non-convex Bilevel Optimization. In *International Conference on Machine Learning*.

[28] Risheng Liu, Long Ma, Yiyang Wang, and Lei Zhang. 2018. Learning converged propagations with deep prior ensemble for image enhancement. *IEEE TIP* 28 (2018), 1528–1543.

[29] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. 2021. Retinex-inspired Unrolling with Cooperative Prior Architecture Search for Low-light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[30] Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. 2020. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *International Conference on Machine Learning*.

[31] Risheng Liu, Guangyu Zhong, Junjie Cao, Zhouchen Lin, Shiguang Shan, and Zhongxuan Luo. 2016. Learning to diffuse: A new perspective to design pdes for visual analysis. *IEEE transactions on pattern analysis and machine intelligence* 38, 12 (2016), 2457–2471.

[32] Yu Liu, Xun Chen, Juan Cheng, and Hu Peng. 2017. A medical image fusion method based on convolutional neural networks. In *2017 20th international conference on information fusion (Fusion)*. IEEE, 1–7.

[33] Yu Liu, Xun Chen, Rabab K Ward, and Z Jane Wang. 2019. Medical image fusion via convolutional sparsity based morphological component analysis. *IEEE Signal Processing Letters* (2019), 485–489.

[34] Yu Liu and Zengfu Wang. 2014. Simultaneous image fusion and denoising with adaptive sparse representation. *IET Image Processing* (2014), 347–357.

[35] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. 2016. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion* 31 (2016), 100–109.

[36] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. 2020. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing* 29 (2020).

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

[38] Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. 2020. GANMcC: A Generative Adversarial Network With Multiclassification Constraints for Infrared and Visible Image Fusion. *IEEE Transactions on Instrumentation and Measurement* 70 (2020), 1–14.

[39] Jinlei Ma, Zhiqiang Zhou, Bo Wang, and Hua Zong. 2017. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Physics & Technology* 82 (2017), 8–17.

[40] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. 2019. Mfas: Multimodal fusion architecture search. In *CVPR*. 6966–6975.

[41] Vladimir S Petrovic and Costas S Xydeas. 2004. Gradient-based multiresolution image fusion. *IEEE Transactions on Image processing* (2004), 228–237.

[42] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7479–7489.

[43] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. 2008. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing* 2, 1 (2008), 023522.

[44] Masanori Suganuma, Mete Ozay, and Takayuki Okatani. 2018. Exploiting the potential of standard convolutional autoencoders for image restoration by evolutionary search. In *International Conference on Machine Learning*. 4771–4780.

[45] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. 2020. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[46] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. 2020. Fusiondn: A unified densely connected network for image fusion. In *AAAI*, Vol. 34. 12484–12491.

[47] L Yang, BL Guo, and W Ni. 2008. Multimodality medical image fusion based on multiscale geometric analysis of contourlet transform. *Neurocomputing* 72, 1-3 (2008), 203–211.

[48] Ming Yin, Xiaoning Liu, Yu Liu, and Xun Chen. 2018. Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. *IEEE Transactions on Instrumentation and Measurement* 68, 1 (2018), 49–64.

[49] Haokui Zhang, Ying Li, Hao Chen, and Chunhua Shen. 2020. Memory-efficient hierarchical neural architecture search for image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3657–3666.

[50] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. 2020. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *AAAI*, Vol. 34. 12797–12804.

[51] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. 2020. DIDFuse: Deep Image Decomposition for Infrared and Visible Image Fusion. In *IJCAI*. 970–976.