electronics

MDPI

Article

Lightweight Infrared and Visible Image Fusion via Adaptive DenseNet with Knowledge Distillation

Zongqing Zhao, Shaojing Su, Junyu Wei *, Xiaozhong Tong and Weijia Gao

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; zhaozongqing17@nudt.edu.cn (Z.Z.); ssjing@nudt.edu.cn (S.S.); tongxiaozhong@nudt.edu.cn (X.T.); gaoweijia21@nudt.edu.cn (W.G.)
* Correspondence: yujy@nudt.edu.cn

Abstract: The fusion of infrared and visible images produces a complementary image that captures both infrared radiation information and visible texture structure details using the respective sensors. However, the current deep-learning-based fusion approaches mainly tend to prioritize visual quality and statistical metrics, leading to an increased model complexity and weight parameter sizes. To address these challenges, we propose a novel dual-light fusion approach using adaptive DenseNet with knowledge distillation to learn and compress from pre-existing fusion models, which achieves the goals of model compression through the use of hyperparameters such as the width and depth of the model network. The effectiveness of our proposed approach is evaluated on a new dataset comprising three public datasets (MSRS, M3FD, and LLVIP), and both qualitative and quantitative experimental results show that the distillated adaptive DenseNet model effectively matches the original fusion models' performance with smaller model weight parameters and shorter inference times.

Keywords: infrared image; visible image; image fusion; adaptive DenseNet; knowledge distillation

Citation: Zhao, Z.; Su, S.; Wei, J.; Tong, X.; Gao, W. Lightweight Infrared and Visible Image Fusion via Adaptive DenseNet with Knowledge Distillation. *Electronics* **2023**, *12*, 2773. https://doi.org/10.3390/electronics12132773

Academic Editors: Peter Sarcevic, Sašo Tomažič, Akos Odry, Sara Stančin and Gábor Kertész

Received: 10 May 2023
Revised: 4 June 2023
Accepted: 6 June 2023
Published: 22 June 2023

1. Introduction

The fusion of infrared and visible images presents an opportunity to leverage the unique advantages of each imaging sensor. Infrared imaging sensors are capable of detecting infrared radiation emitted by objects and temperature variations with less susceptibility to harsh weather conditions and illumination [1–3]. Nonetheless, the resolution of infrared sensors is comparatively lower, and the texture details are deficient. Conversely, visible imaging sensors can capture texture structure details and color information, but their performance is compromised under varying light conditions and weather situations. The fusion of infrared and visible images yields a synergistic outcome in the form of a complementary image that encompasses both infrared radiation information and visible texture structure details. Therefore, fused images have immense potential for deployment in low-light and adverse weather environments. With the advantages of infrared and visible image fusion, there are diverse measurement and detection applications across various fields. In the realm of medical image measurement tasks, by fusing images captured by multiple sensors with different spectral ranges, we can obtain more comprehensive and accurate measurements of biological tissues and structures [4]. Additionally, the fusion of infrared and visible images can provide detailed and accurate information on environmental condition measurement, enabling land cover mapping, vegetation analysis, and detection of environmental changes [5]. Furthermore, robotics measurement and control also can utilize the superior quality of the fused images to enhance navigation, object recognition, and obstacle avoidance in low-light and adverse weather conditions [6]. Additionally, security systems can exploit the benefits of fused images for facial recognition, access control, and intrusion detection, leading to a more reliable and robust representation of the scene [7,8]. Overall, infrared and visible image fusion has significant potential in a wide range of measurement and detection applications, offering enhanced accuracy, efficiency, and reliability. Moreover,

current optical vision stereo measurements based on deep learning require efficient feature extraction methods to support them [9].

With the huge demand for fusion measurement applications, many image fusion methods have been developed and proposed to address the challenge of fusing infrared and visible images. Based on the different feature extraction and fusion strategies, these methods can be classified into conventional fusion methods and end-to-end deep learning methods. According to the hand-crafted feature decomposition and generation rules, conventional fusion methods mainly consist of multiscale transform-based [10], sparse representation-based [11–13], saliency-based [14–17], fuzzy set-based [18–20], and hybrid-based [21–23] methods. To summarize, conventional methods for image fusion typically comprise three primary stages. Initially, a specific transformation is utilized to extract features from the source images, commonly referred to as the feature extraction stage. Subsequently, fusion strategies are applied to combine these features in the feature fusion phase. Lastly, the merged features are used to reconstruct the fused image by applying the corresponding inverse transformation in the feature reconstruction stage.

Recently, conventional methods have been integrated with deep learning techniques due to their ability to effectively leverage multi-modal features and facilitate efficient fusion. In general, deep-learning-based end-to-end fusion methods can be categorized into four distinct groups, namely those founded on auto-encoders (AEs) [3,24–26], convolutional neural networks (CNNs) [27–29], generative adversarial networks (GANs) [30–33], and transformer architectures [34–37]. AE-based techniques utilize an auto-encoder to execute both feature extraction and feature reconstruction, while feature fusion is carried out via specific fusion strategies. In contrast, CNN-based approaches can achieve feature extraction, feature fusion, and feature reconstruction, obtaining unique fusion results through well-designed network architectures and loss functions. Diverging from CNN-based methods, GAN-based techniques introduce generative adversarial mechanisms to the realm of infrared and visible image fusion. Such fusion methods typically incorporate generators and discriminators. On the other hand, transformer-based methods harness self-attention mechanisms to extract global features for the completion of fusion tasks.

While existing fusion approaches have shown promising performance, most of them tend to prioritize the visual quality and statistical metrics of the fused images by altering the deep network structure and enhancing the layers. Thus, the fusion training network becomes increasingly convoluted and the model's reasoning parameters undergo a surge, without duly considering the requirement of real-time performance when computational resources on edge devices are scarce. Drawing inspiration from the successful application of knowledge distillation (KD) technology in recommender systems and natural language processing, we have developed an adaptive distillation paradigm for real-time infrared and visible image fusion tasks. KD is a model agnostic technique that facilitates the transfer of knowledge from a pre-trained large teacher model to a smaller, yet powerful, student model. Therefore, we utilize previously trained large models based on deep learning, such as MCnet, MCTNet, and TCPMFNet [38–42], which have demonstrated effectiveness in image fusion work, as teacher models. This paper combines the strengths of previous large models and KD and proposes a lightweight infrared and visible image fusion method that integrates an adaptive mechanism and knowledge distillation. By collecting visible and infrared image pairs from multiple public dual-light datasets, including MSRS, M3FD, and LLVIP, we innovatively constructed a typical dual-light dataset covering multiple complex scenes, and the proposed approach yields satisfactory fusion results on this created dataset. The primary contributions of this study are as follows.

(1) We introduce an adaptive knowledge distillation network for infrared and visible image fusion tasks, which can be trained by leveraging the knowledge from pre-existing large fusion models and achieves a comparable performance with smaller model parameters and a more simplified model structure.
(2) The devised knowledge distillation network exhibits the capability of adaptively tuning hyperparameters and can be trained by various categories of pre-existing

fusion models, encompassing CNN-based, transformer-based, and high-vision-level-based models.

(3) The newly created dataset comprises 3288 pairs of infrared and visible images captured from multiple typical scenes, such as city roads, pedestrian crossings, parking lots, mountain forests, and buildings in the background. The images were meticulously selected to offer a broad representation of real-world scenarios, making the dataset suitable for training and evaluating models for infrared and visible image fusion tasks.

The paper exhibits meticulous organization, structured as follows. Section 2 presents an exposition of related works concerning knowledge distillation and adaptive mechanisms for infrared and visible image fusion. A detailed account of our approach and methodology is provided in Section 3. Section 4 comprises an explanation of the constructed dataset, accompanied by the results and analysis of the experiments conducted offline. In Section 5, the real-world applications of the trained adaptive DenseNet are described in detail, including its deployment on mobile platforms. Finally, the paper draws conclusions in Section 6.

2. Related Works

In this section, we first present some related works on knowledge distillation and adaptive mechanisms. Subsequently, we provide an overview of various state-of-the-art large models for infrared and visible image fusion based on different algorithms.

2.1. Knowledge Distillation

Knowledge distillation is a procedure for model compression where a small (student) model is trained to match a large pre-trained (teacher) model [43]. Knowledge is transferred from the teacher model to the student by minimizing a loss function aimed at matching softened teacher logits as well as ground-truth labels [8]. Chen et al. [44] proposed cross-stage connection paths in knowledge distillation that use low-level features in the teacher network to supervise deeper features for the student, resulting in a much improved overall performance. This modification reveals the great importance of designing connection paths in knowledge distillation. Xiao et al. [45] proposed a heterogeneous knowledge distillation network with multilayer attention embedding to address the issue of low-resolution fusion results in infrared–visible image fusion. This method has a high-resolution fusion network (teacher) and a low-resolution fusion and super-resolution network (student), where the teacher guides the student's ability to implement joint fusion and super-resolution. Liu et al. [46] proposed a perceptual distillation method to train image fusion networks without ground truths using a main autoencoder as the student network, a teacher network with well-trained network representations, and a multi-autoencoder architecture trained with self-supervision. To leverage depth maps, Zhao et al. [47] proposed a new depth-distilled multi-focus image fusion (MFIF) framework called D2MFIF, featuring a depth-distilled model (DDM) to transfer depth knowledge and a multi-level fusion mechanism to improve the final predictions. Mi et al. [48] proposed a medical image fusion model addressing the challenges of limited publicly accessible medical image datasets by using knowledge distillation and an explainable AI-module-based generative adversarial network with dual discriminators. This model reduces the required dataset size for training and generates clear fused images using small-scale datasets while helping to reduce overfitting. While existing image fusion methods utilizing knowledge distillation can enhance the performance of the fusion model, increase the predictive accuracy, and decrease the dataset size necessary for training, there exists potential for further optimization of the deep neural network through knowledge distillation combined with adaptive mechanisms. Additionally, compressed models ought to be tailored to facilitate portability to edge computing devices and sustain high-level visual measurement tasks that necessitate real-time image fusion. The advantages and disadvantages of knowledge distillation related works are summarized in the Table 1.

Table 1. Summary of knowledge-distillation-related works.

Related Works	Advantages	Disadvantages
Original knowledge distillation [43]	Compresses a large model into a small one	May lose some information or performance
Cross-stage connection path in knowledge distillation [44]	Uses low-level features to supervise deeper features	Increase the complexity of the distillation process
Heterogeneous knowledge distillation network [45]	Implements joint fusion and super-resolution	Dependent on the quality of the teacher network
Perceptual distillation [46]	Trains image fusion networks without ground truths	Requires a high-quality pre-trained teacher network
Depth-distilled multi-focus image fusion (MFIF) [47]	Transfers depth knowledge to improve fusion accuracy	Transfers depth knowledge to improve fusion accuracy
Medical image fusion model [48]	Reduces the dataset size and overfitting risk	Limited generalization and sensitivity to hyperparameters and loss functions

In summary, the related works utilizing knowledge distillation have been shown to improve the performance of image fusion models by compressing large pre-trained models and transferring knowledge from teacher to student models. However, these methods have their limitations, including the dependence on high-quality pre-trained networks, the increased computational cost, and the need for proper tuning of the architecture. These limitations highlight the potential for further optimization of deep neural networks through knowledge distillation combined with adaptive mechanisms, including the development of compressed models that facilitate portability to edge computing devices and sustaining high-level visual measurement tasks that require real-time image fusion.

2.2. Adaptive Mechanisms

Adaptive mechanisms refer to techniques or algorithms that enable a system to adjust or modify its behavior in response to changes in its environment or operating conditions. Adaptive mechanisms for image fusion have been proposed by various researchers over the years. For example, Xia et al. [4] proposed a parameter-adaptive pulse-coupled neural network method to obtain a better fusion effect. Lu et al. [49] proposed an image retrieval strategy using adaptive features and information entropy theory to extract features, calculate similarity, and obtain the initial results. Kong et al. [50] proposed an adaptive normalization-mechanism-based fusion method, which injects detailed features into the structure feature. This adaptive normalization mechanism significantly improves the fusion performance. In brief, adaptive mechanisms for image fusion exhibit various benefits compared to other algorithmic fusion methods. For instance, they can blend the effectiveness of multiple features in an unsupervised way and outperform single-feature retrieval in terms of accuracy and generalization. Moreover, they can enhance image fusion precision, curtail noise interference, and refine the algorithm's real-time performance. Conversely, various adaptive mechanisms have been proposed to optimize deep neural network structures. These techniques include adaptive selection of the loss function, the adaptive activation function, and adaptive sampling [51]. Additionally, global group sparse coding has been utilized to automatically learn inter-layer connections and determine the depth of a neural network [52]. In recent years, novel network structures have been proposed that demonstrate superior performance compared to traditional feedforward neural networks [53]. To summarize, the utilization of adaptive mechanisms provides a means to optimize the hyperparameters and structure of the designed knowledge distillation network. The advantages and disadvantages of adaptive mechanism related works are summarized in the Table 2.

Table 2. Summary of adaptive-mechanism-related works.

Related Works	Advantages	Disadvantages
Parameter-adaptive pulse-coupled neural network [4]	Obtains a better fusion effect	May be sensitive to parameter settings
Image retrieval using adaptive features and information entropy [49]	Extracts features and calculates similarity effectively	May not handle complex scenes well
Adaptive normalization-mechanism-based fusion [50]	Injects detailed features into structure feature	May introduce artifacts or distortions
Adaptive selection of the loss function, activation function, and sampling [51]	Optimizes the performance of deep neural networks	May require more computational resources or tuning
Global group sparse coding [52]	Learns inter-layer connections and determines network depth automatically	May suffer from sparsity or redundancy issues
Novel network structures [53]	Outperforms traditional feedforward neural networks	May be difficult to design or interpret

In summary, adaptive mechanisms for image fusion have various benefits compared to other algorithmic fusion methods. They are able to adapt to changing operating conditions, blend the effectiveness of multiple features in an unsupervised way, and outperform single-feature retrieval in terms of accuracy and generalization. They are also able to enhance image fusion precision, curtail noise interference, and refine the algorithm's real-time performance. However, proper selection of a mechanism and optimization of parameters can be challenging. Adaptive mechanisms for optimizing deep neural network structures are also available, but they may increase the computational cost and require an appropriate selection of the network structure, the loss function, the activation function, sampling, or the coding method. In general, the utilization of adaptive mechanisms provides a means to optimize the hyperparameters and structure of the designed knowledge distillation network.

2.3. Typical Image Fusion Model

The application of deep-learning-based approaches to infrared and visible image fusion has yielded significant advancements in the domain of visual detection and measurements owing to the implementation of multi-layered and intricately structured deep neural networks. Li et al. [54] introduced DenseFuse, a deep learning architecture that uses a combination of convolutional layers, a fusion layer, and dense blocks to extract more useful features from the source infrared and visible images, resulting in a fused image reconstructed by a decoder that outperforms existing fusion methods in both objective and subjective assessments. Tang et al. [55] proposed SeAFusion, a real-time image fusion network that combines image fusion and semantic segmentation to guide high-level semantic information and uses gradient residual dense blocks to enhance the fusion network's description ability. Xu et al. [56] proposed U2Fusion, an unsupervised and unified image fusion network that adapts to different fusion tasks by estimating source image importance, preserving adaptive similarity, and avoiding loss of previous fusion capabilities, thus mitigating the requirements of a ground truth and specific metrics. Wei et al. [57] utilized a dynamic transformer module to extract local features and context information and a Y-shaped network to maintain both thermal radiation information and scene details in infrared and visible image fusion. Wu et al. [58] proposed an end-to-end fusion network architecture (RFN-Nest) to tackle the challenging problem of designing an appropriate strategy for generating fused images, incorporating a residual fusion network, detail-preserving and feature-enhancing loss functions, and a two-stage training strategy including an auto-encoder and the RFN. Z-R. Jin et al. [59] proposed a simple but effective bilateral activation mechanism (BAM) which can be applied to the activation function to offer an efficient feature extraction model. It also introduces a Bilateral ReLU Residual Block

(BRRB) and a BRResNet architecture to achieve state-of-the-art performance in two-image fusion tasks, i.e., pansharpening and hyperspectral image super-resolution (HISR). Despite the significant progress made by these fusion models in infrared and visible light fusion, the utilization of intricate hyperparameters and significant amounts of memory render these models inadequate for real-time processing and lightweight applications, thereby necessitating model compression and parameter optimization to achieve the optimal performance and strike a balance between model efficacy and real-time constraints. The advantages and disadvantages of image fusion model related works are summarized in the Table 3.

Table 3. Summary of image-fusion-model-related works.

Related Works	Advantages	Disadvantages
DenseFuse [54]	Uses dense blocks to extract more useful features	May not preserve the contrast and brightness of the source images
SeAFusion [55]	Combines image fusion and semantic segmentation	May not handle complex scenes or occlusions well
U2Fusion [56]	Adapts to different fusion tasks by estimating source image importance	May not be stable or robust to noise or distortion
Dynamic transformer module and Y-shaped network [57]	Extracts local features and context information	May introduce artifacts or blur in the fused image
RFN-Nest [58]	Incorporates a residual fusion network and a two-stage training strategy	May require a large amount of training data and time

In summary, various deep-learning-based approaches have been proposed for infrared and visible image fusion, utilizing multi-layered and intricately structured deep neural networks. These models are able to extract useful features from source images, guide high-level semantic information, adapt to different fusion tasks, and tackle the challenging problem of generating fused images. However, they require the appropriate implementation of complex modules, the selection of hyperparameters, and significant memory utilization. Additionally, achieving real-time processing and lightweight applications necessitates model compression and parameter optimization to strike a balance between model efficacy and real-time constraints.

3. Framework and Methodology

In this section, we present our proposed methodology and framework. Firstly, we describe the framework of our method, followed by a comprehensive description of the novel adaptive DenseNet and the corresponding loss function for the knowledge distillation process. Lastly, we introduce the optimization and lightweight strategy employed in the constructed fusion network.

3.1. The Framework of Fusion Network

The diagram in Figure 1 depicts the proposed methodology's framework. Our approach involves treating the amalgamation of visible and infrared images as a knowledge distillation of the neural network fitting problem. Specifically, we leverage pre-existing fusion models and feed them with pairs of visible and infrared images to obtain fused images. As a result, we construct a dataset comprising visible–infrared image pairs as inputs and fused images as labels. Subsequently, we perform supervised learning on our fusion model using the Adaptive DenseNet and Huber Loss network models for convergence. To this end, we construct an Adaptive DenseNet set that facilitates the selection of the optimal DenseNet network architecture for our fitting task. Importantly, our network model is characterized by a reduced complexity, a smaller storage size, and fewer parameters

when compared to existing trained models, while delivering equivalent or superior image fusion performance.

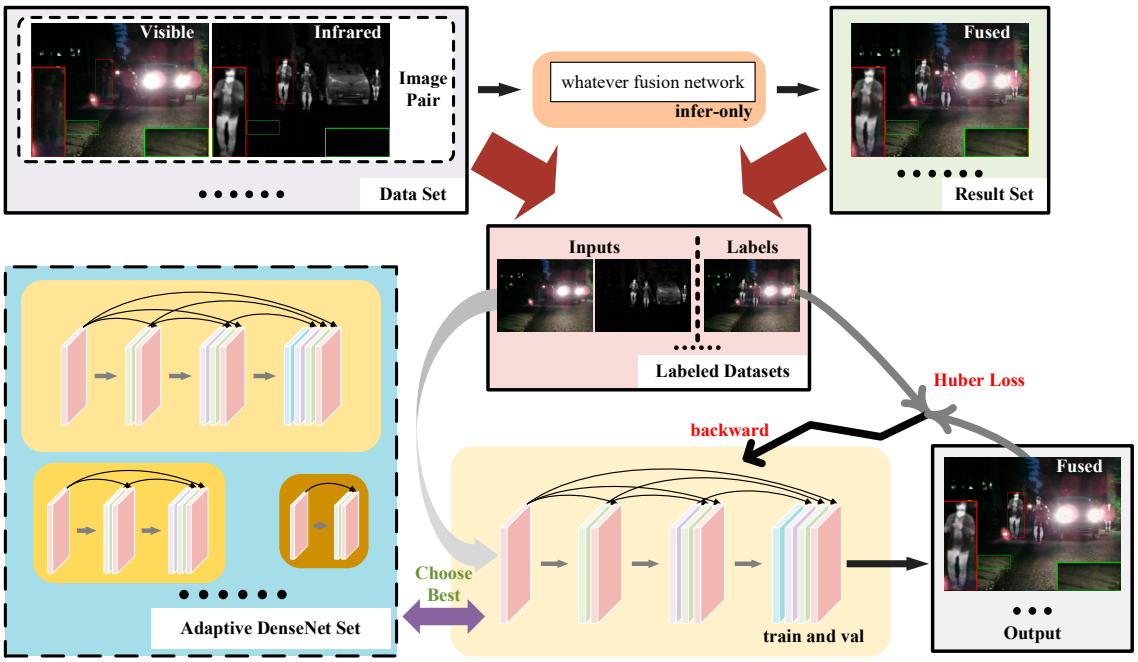

Figure 1. The framework of the proposed approach using the trained infrared and visible images fusion models and the novel adaptive DenseNet Set for knowledge distillation. The red box in the image zooms in on the local area where the pedestrian is located, while the green box zooms in on the local area where the ground object is located.

The specific steps can be divided into the following five steps:

(1) Construct a paired dataset of visible light and infrared images.
(2) Select a teacher model (SeAFusion in this case) and use its pre-trained weights to infer the visible light and infrared images in the dataset to obtain a collection of fused images.
(3) Combine the fused images obtained in step 2 with the original dataset to form a new dataset with visible light and infrared images as inputs and corresponding fusion images as labels.
(4) Use the labeled dataset from step 3 to train a student model, which is described in detail in Section 3.3 of the manuscript. This step is the actual knowledge distillation process, as the student model is trained to capture the information in the labeled dataset using the fusion images obtained from the teacher model as soft labels.
(5) Obtain the student model, which is a self-adaptive DenseNet with weights trained in step 4. The student model has similar fusion effects to the teacher model SeAFusion, but is far superior in terms of inference speed and model size.

3.2. Adaptive DenseNet and Loss Function

The structure of the proposed adaptive DenseNet is presented in Figure 2, and it integrates two variables that function to regulate the network structure and hyperparameters of the adaptive DenseNet for implementing knowledge distillation. Input visible and infrared image pairs are concatenated and subjected to ConvBlock processing within dense layers (n), where n represents the number of dense layers, akin to depth in the YOLO network. A dense layer is composed of ConvBlock and concatenation modules, with the latter facilitating the integration of input and ConvBlock features at the next level, where m represents the number of each dense layer output channel produced by the ConvBlock,

similar to the width in the YOLO network. Finally, convolution and the Tanh activation function are used to obtain the fusion output (comprising three channels).

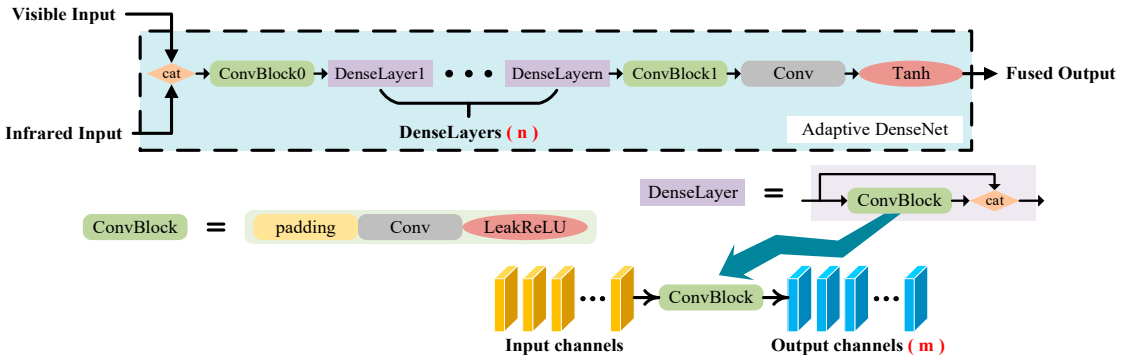

Figure 2. The structure of the proposed adaptive DenseNet.

Importantly, the network does not have upsampling or downsampling layers, and the convolutional layers use the following parameters: padding = 1, stride = 1, and kernel_size = 3. This ensures that the input and output features of each dense layer differ only in the number of channels, while the feature size remains the same.

Table 4 depicts the hierarchical structure of the network, as well as the input and output channel counts in each layer, with examples provided for n = 2, m = 8 and n = 5, m = 16. It is noteworthy that only the number of layers and channels are selected to adaptively tune the performance of the designed DenseNet.

Table 4. Examples of the hierarchical structure and corresponding parameters of the network.

Layer Name	n = 2, m = 8		n = 5, m = 16	
	Input Channels	Output Channels	Input Channels	Output Channels
ConvBlock0	4	8	4	16
DenseLayer1	8	16	16	32
DenseLayer2	16	24	32	48
DenseLayer3			48	64
DenseLayer4			64	80
DenseLayer5			80	96
ConvBlock1	24	32	96	32
Conv	32	3	32	3
Activation Function	Tanh			

Our model is designed to accommodate the results produced by other fusion models; thus, the loss function can be specified as either the mean absolute error loss (*MAELoss*) or the mean square error loss (*MSELoss*). The *MAELoss* function is formulated as follows:

$$MAELoss_i = |x_i - y_i| \tag{1}$$

where x_i is the output fusion image sequence of the proposed adaptive model and y_i is the original model fusion output sequence as the label. *MSELoss* function is formulated as follows:

$$MSELoss_i = (x_i - y_i)^2 \tag{2}$$

Generally, the *MAELoss* exhibits greater resilience towards outliers; however, it may experience oscillations in proximity to the global minimum of the loss function, leading to difficulty in achieving convergence to the optimal value. While *MSELoss* has a faster

convergence rate, it is more sensitive to outliers compared to *MAELoss*. To balance between the two, we adopt HuberLoss in our model, which is defined by the following equation:

$$l_i = \begin{cases} \frac{1}{2} MSELoss_i{}^2, \text{ if } \left| x_i - y_i \right| \leq \delta \\ \delta \cdot MAELoss_i - \frac{1}{2}\delta^2, \text{ if } \left| x_i - y_i \right| > \delta \end{cases} \tag{3}$$

where δ is typically set to a threshold value of 1. Thus, the proposed approach combines the advantages of both *MAELoss* and *MSELoss* by using HuberLoss, which behaves similarly to *MAELoss* when the loss is large and resembles *MSELoss* when the loss is small, thereby reducing the impact of outliers while also promoting faster convergence. In relation to the sample set, the mean loss is utilized to denote the loss, which can be defined as follows:

$$L(x, y) = mean\{l_1, l_2 \cdots l_i\} \tag{4}$$

3.3. Adaptive Optimal Strategy for DenseNet

The objective of this study is to achieve knowledge distillation by constructing an optimal network structure to fit existing fusion algorithms. Based on DenseNet, we simplify the problem to an adaptive optimization problem under discrete conditions, which involves finding the optimal values of (n, m) to achieve good fitting performance while minimizing the inference time. The mathematical model of the adaptive optimization is expressed as follows:

$$\begin{aligned} &\text{Variables} : (n, m) \\ &\text{Destination} : \min\{ \text{ inference time for each pair images} | (n, m)\} \\ &\text{Constraints} : |f(n, m) - f(n_{best}, m_{best})| < \sigma, \ (n, m) \in N \end{aligned} \tag{5}$$

where m is the number of dense layer output channels, n is the number of dense layers, $f(n, m)$ represents the fitting performance, $f(n_{best}, m_{best})$ denotes the theoretically optimal fitting performance, σ represents the designed threshold, and N is a natural number. The adaptive optimal strategy for DenseNet is made up of three steps:

Step 1 (Pre-Processing Stage): To begin with, the time required to perform inference for each pair of images is calculated for a fixed (n, m) value and given typical input image pixels in order to adaptively select the model structure with the shortest inference time in subsequent operations. Figure 3 demonstrates the adaptive DenseNet's inference time per pair of images, while processing the fusion of 640 × 480 size infrared and visible light images with varying $n \in$ [2,5] and $m \in$ [4,16]. In this figure, the horizontal axis represents the number of output channels in each dense layer (i.e., m value), while the four curves represent the number of dense layers (i.e., n value) ranging from 2 to 5. The vertical axis shows the inference time, measured in milliseconds, required for each pair of infrared and visible light images, and represents the average of 10,000 tests. Upon analyzing Figure 3, it is evident that the inference time does not necessarily increase with an increase in m, while n is constant. Moreover, when m is a power of 2, the model complexity increases without an increase in inference time, indicating a higher cost-effectiveness. Therefore, we eliminated the low cost-effective DenseNet network structures constructed with certain (n, m) values during the pre-processing stage. For instance, the DenseNet network structure with n = 2 and m = 11 has a significantly lower actual complexity than that of n = 2 and m = 12, but its inference time is significantly higher. Table 5 presents the optional adaptive DenseNet network architectures sorted by inference time from low to high after removing the low cost-effective (n, m) values, where each (n, m) pair corresponds to a sequence number s.

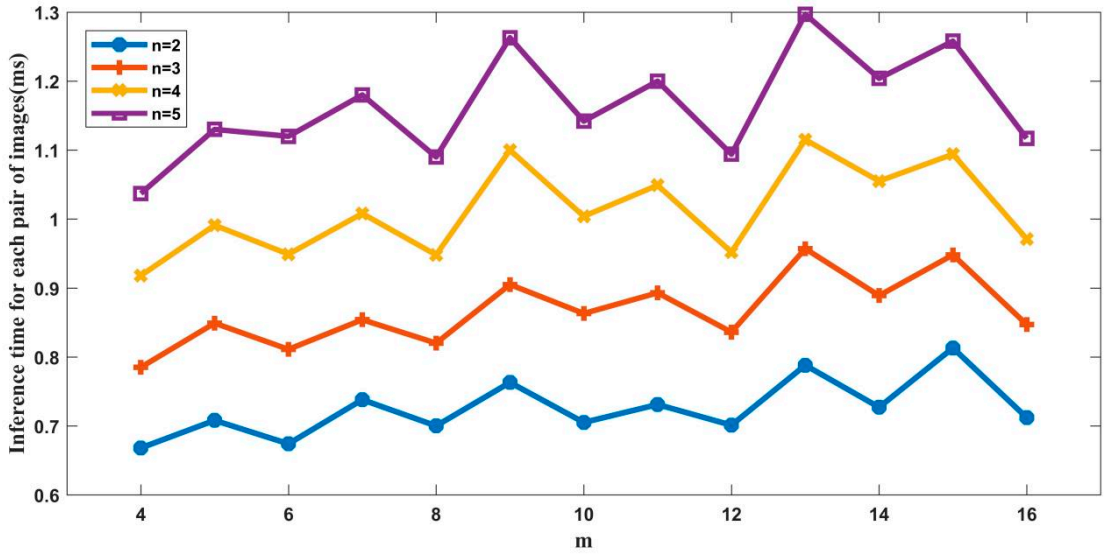

Figure 3. The inference time required for each pair of images using varying numbers of dense layers and output channels per dense layer.

Table 5. The optional adaptive DenseNet network architecture.

s	1	2	3	4	5	6	7	8
n	2	2	2	2	3	3	3	3
m	4	8	12	16	4	8	12	16
inference time (ms)	0.668	0.700	0.701	0.712	0.785	0.82	0.836	0.847
s	9	10	11	12	13	14	15	16
n	4	4	4	4	5	5	5	5
m	4	8	12	16	4	8	12	16
inference time (ms)	0.918	0.948	0.952	0.971	1.037	1.09	1.094	1.117

Step 2 (Assumption Stage): The proposed method utilizes the following assumptions based on common knowledge in machine learning before performing an adaptive search for optimal solutions:

A. After removing the low cost-effective (n, m) pairs, the complexity of the model is positively correlated with the inference time, i.e., the larger the sequence number s, the more complex the model.

B. The complexity of the model is positively correlated with its generalization fitting capability. Specifically, given two models, A and B, if A is more complex than B and both models converge during function fitting with adequate training and validation data, A's fitting performance will be at least as good as that of B. Conversely, if there are insufficient training and validation data, neither model A nor model B may converge.

Step 3 (Solution Stage): Let $f(n, m)$ be defined as $1 - HuberLoss_{stable}(n, m)$, where $HuberLoss_{stable}(n, m)$ represents the mean value of the HuberLoss on the validation set when the model converges at the given values of (n, m). It is evident that as $HuberLoss_{stable}(n, m)$ becomes smaller, the generated images by the model are closer to the expected images, indicating a better fitting effect. The pseudo code of the adaptive optimal search strategy is shown in Algorithm 1. To obtain the optimal adaptive DenseNet architecture and dual-light image fusion effect, a binary search strategy is employed to adaptively obtain the best hyperparameters (n, m). The hyperparameters $(n, m)_{s_{low}}$, $(n, m)_{s_{mid}}$, and $(n, m)_{s_{high}}$ correspond to the network structure at s equal to 1, 8, and 16, respectively. After conducting multiple tests, the hyperparameters for initializing the model architecture are set to $(n_0, m_0) = (5, 16)$ and $(n_1, m_1) = (2, 4)$, while the initial threshold is set to $\sigma = 0.01 * [f(n_0, m_0) - f(n_1, m_1)]$.

The value of s is considered the optimal situation when it is minimized. The adaptive DenseNet constructed based on the (n, m) values obtained from the solution is referred to as the optimal network.

Algorithm 1: Adaptive Optimal Search

Input: train set T$\{visible_i, infrared_i, label_i\}$, val set V$\left\{visible_j, infrared_j, label_j\right\}$
Output: s$\rightarrow(n, m)_s$
Begin
1. Calculate f(n0, m0) and σ
2. $s_{low} = 1 \rightarrow (n, m)_{s_{low}}$
3. $s_{high} = 16 \rightarrow (n, m)_{s_{high}}$
4. $s_{mid} = (s_{low} + s_{high})//2 = 8 \rightarrow (n, m)_{s_{mid}}$
5. **While** $s_{low} \neq s_{high}$ **do**
6. **if** $\left|\mathrm{f}(n, m)_{s_{mid}} - \mathrm{f}(n_0, m_0)\right| < \sigma$ **do**
7. $s_{high} = s_{mid}$
8. $s_{mid} = (s_{low} + s_{high})//2$
9. **else do**
10. $s_{low} = s_{mid} + 1$
11. $s_{mid} = (s_{low} + s_{high})//2$
12. return $s_{mid} \rightarrow (n, m)_{s_{mid}}$
end

4. Offline Experimental Setup and Comparative Analysis

In this section, we first provide the experimental settings and dataset. Then, we introduce six existing infrared and visible light fusion models, including DenseFuse, U2Fusion, RFN-NEST, YDTR, SwinFusion, and SeAFusion. Next, we propose the adaptive DenseNet structure obtained through the proposed adaptive optimization learning method. Then, we compare the performance of the six fusion algorithms and the corresponding adaptive DenseNet in terms of the model weight and size and the inference time. Furthermore, to comprehensively compare the fitting performance of the adaptive Optimal DenseNet and the original model network, we use 21 typical pairs of dual-light images in the VIFB (visible and infrared image fusion benchmark) dataset as training and testing data and employ five categories of 17 popular evaluation metrics in the infrared and visible light fusion field for quantitative comparison. Finally, we qualitatively compare the fusion effect of the original algorithm with that of the adaptive DenseNet.

All the networks were trained, validated, and tested on a high-performance workstation equipped with an Nvidia Tesla A100 GPU with 80 GB memory and an AMD Ryzen Threadripper PRO 5995WX 64-Core CPU. The deep learning framework was PyTorch and the CUDA version used is 11.7.

In describing the algorithm model of this paper, we distinguish between the original and our models. For each original algorithm, we have a corresponding adaptive DenseNet structure. For example, YDTR_ori is the original YDTR algorithm, and its model weights were trained by its authors. YDTR_our refers to our use of adaptive DenseNet to fit its fusion effect.

4.1. Dataset Preparation

The datasets used for training and validation were selected from MSRS, M3FD_fusion, and LLVIP, comprising a total of 3288 pairs of visible and infrared images [55,60,61]. These were divided into a training set of 2562 pairs and a validation set of 726 pairs. The images used for training and validation were cropped to a size of 640 × 480 and strictly aligned spatially. During testing, the VIFB dataset was selected and used to comprehensively compare the original fusion algorithms with the proposed adaptive DenseNet [62,63].

Table 6 shows the selected quantitative evaluation metrics, which include entropy (EN), mutual information (MI), pixel feature mutual information (FMI_pixel), wavelet feature

mutual information (FMI_w), discrete cosine feature mutual information (FMI_dct), the peak signal-to-noise ratio (PSNR), edge-information-based indicators (Qabf), artifact-based indicators (Nabf), the structural similarity index measure (SSIM), the multi-scale structural similarity index measure (MS-SSIM), the mean square error (MSE), spatial frequency (SF), the standard deviation (SD), the average gradient (AG), visual information fidelity (VIF), the correlation coefficient (CC), and the sum of correlation differences (SCD).

Table 6. The selected quantitative evaluation metrics [62].

Theory	Evaluation Metrics
Information Theory	EN, MI, FMI_pixel, FMI_w, FMI_dct, PSNR, Qabf, Nabf
Structural Similarity	SSIM, MS_SSIM, MSE
Image Features	SF, SD, AG
Human Visual Perception	VIF
Correlation	CC, SCD

If the evaluation indicators are for the fused image itself, then visible and infrared images are not needed, such as for entropy (*EN*):

$$EN = -\sum_{l=0}^{L-1} p_l \log_2 p_l$$

In the formula, L represents the number of gray levels and p_l represents the normalized histogram of the corresponding gray level in the fused image.

In the case of evaluation indicators that are based on both the input and output images, such as the correlation coefficient (*CC*), the average value is taken:

$$CC = \frac{r(A,F) + r(B,F)}{2}$$

In the formula for *CC*, A represents the visible light image, B represents the infrared image, and F represents the fused image. In addition,

$$r(X,F) = \frac{\sum_{i=1}^{M}\sum_{i=1}^{N}(X(i,j)-\overline{X})(F(i,j)-\overline{F})}{\sqrt{\sum_{i=1}^{M}\sum_{i=1}^{N}(X(i,j)-\overline{X})^2\sum_{i=1}^{M}\sum_{i=1}^{N}(F(i,j)-\overline{F})^2}}$$

where $\overline{X}$ means the mean value of the source image.

In the qualitative comparison, the test results of man.jpg in the VIFB dataset were selected for analysis, with a focus on comparing the details, brightness, and saliency of the fused image.

4.2. Adaptive DenseNet Knowledge Distillation

For the current six popular dual-light fusion models, our proposed method for searching the optimal structure of their corresponding adaptive DenseNet through knowledge distillation can converge to a stable value of the loss function within 20 epochs. Moreover, when s ∈ [1,16], at most log2(16) + 1 ≈ 5 searches are needed to find each optimal (n, m) combination, i.e., the corresponding adaptive DenseNet structure. Table 7 presents the model weight sizes and inference times of the corresponding adaptive DenseNet obtained by our proposed method for the six different fusion models. The images used for testing were all 640 × 480 resolution and consisted of pairs of three-channel infrared and visible light images. Table 7 demonstrates that the generated adaptive DenseNet through knowledge distillation significantly reduced both the inference time and the model parameter size. SwinFusion achieved the greatest reduction, with the inference time compressed

by 0.00057 times and the model parameters compressed by 0.002 times that of the original algorithm. Although DenseFuse and U2Fusion both have network models based on DenseNet, their inference times can still be reduced by less than 0.4 times. U2Fusion had the smallest reduction in inference time, at 0.305 times lower, but its network structure is more reasonable, with less redundancy compared to other algorithm models.

Table 7. The model size and inference time of the original fusion models and their corresponding adaptive DenseNet.

Model	Original		Ours (Adaptive DenseNet)				Ratio	
	Inference Time (ms)	Model Size (KB)	*n*	*m*	Inference Time (ms)	Model Size (KB)	Inference Time Ratio	Model Size Ratio
YDTR	63	873	2	8	0.7	42	0.011	0.048
DenseFuse	3	296	2	8	0.7	42	0.233	0.14
SeAFusion	4	667	4	8	0.95	77	0.238	0.12
U2Fusion	2.2	2590	2	4	0.67	23	0.305	0.009
RFN-NEST	55	18,730	3	16	0.85	136	0.015	0.007
SwinFusion	1920	54,025	5	8	1.1	98	0.00057	0.002

4.3. Qualitative Analysis

During the qualitative analysis, the "man.jpg" image from VIFB was selected to demonstrate the contrast and texture details of the infrared and visible light fusion images among the six original fusion models and their corresponding distilled models. Figure 4 illustrates the test results of all models on "man.jpg", and typical regions of the character and ground mark were selected for a magnified comparison. The results reveal that the original DenseNet and U2Fusion models do not highlight the character region significantly, whereas the distilled U2Fusion model using our proposed method shows a higher overall contrast, which is more consistent with human visual perception. In the ground mark region, both RFN-NEST and its corresponding distilled model lack significant texture features and fail to exhibit clear ground markings. Overall, although the defects in the original fusion models also appear in the distilled models, our proposed knowledge distillation adaptive DenseNet method effectively fits the original fusion models' performance and achieves the goals of model compression with a shorter inference time.

4.4. Quantitative Analysis

To test the effectiveness of the proposed distillation model and the efficacy of dual-light image fusion, we evaluated the performance using six popular fusion models and their corresponding distilled models on 21 pairs of VIFB images. A set of 17 commonly used metrics were employed to assess the performance. Figure 5 presents the results of the quantitative analysis using 17 metrics on the 21 pairs of images. It can be observed that the distribution of fusion results obtained by our proposed method using the distilled models is similar to that of the original models with respect to the corresponding evaluation metrics. Specifically, based on the information-theory-based EN metric, the algorithm fitted to the U2Fusion method achieves the best performance on most image pairs, indicating that the proposed distillation method can still enrich the information in the images when fitting the U2Fusion fusion algorithm. This is mainly because the obtained corresponding adaptive DenseNet not only regresses to the result of the U2Fusion algorithm, but also can obtain information from the input visible–infrared images.

Regarding the structural similarity (SSIM) and multi-scale structural similarity (MS_SSIM) metrics, the distilled model obtained for the DenseFuse model is extremely similar to the original model and overall performs better. On the one hand, our network structure is similar to DenseFuse, so the fitting degree is higher. On the other hand, DenseFuse's fusion strategy is a conventional strategy, which is more biased towards specific scene applications. In contrast, the knowledge distilled model we proposed has no application restrictions, and

therefore has stronger universality. Moreover, the original RFN-NEST algorithm performed poorly on the three SSIM metrics. Similarly, the distilled model for RFN-NEST did not show a relatively strong performance.

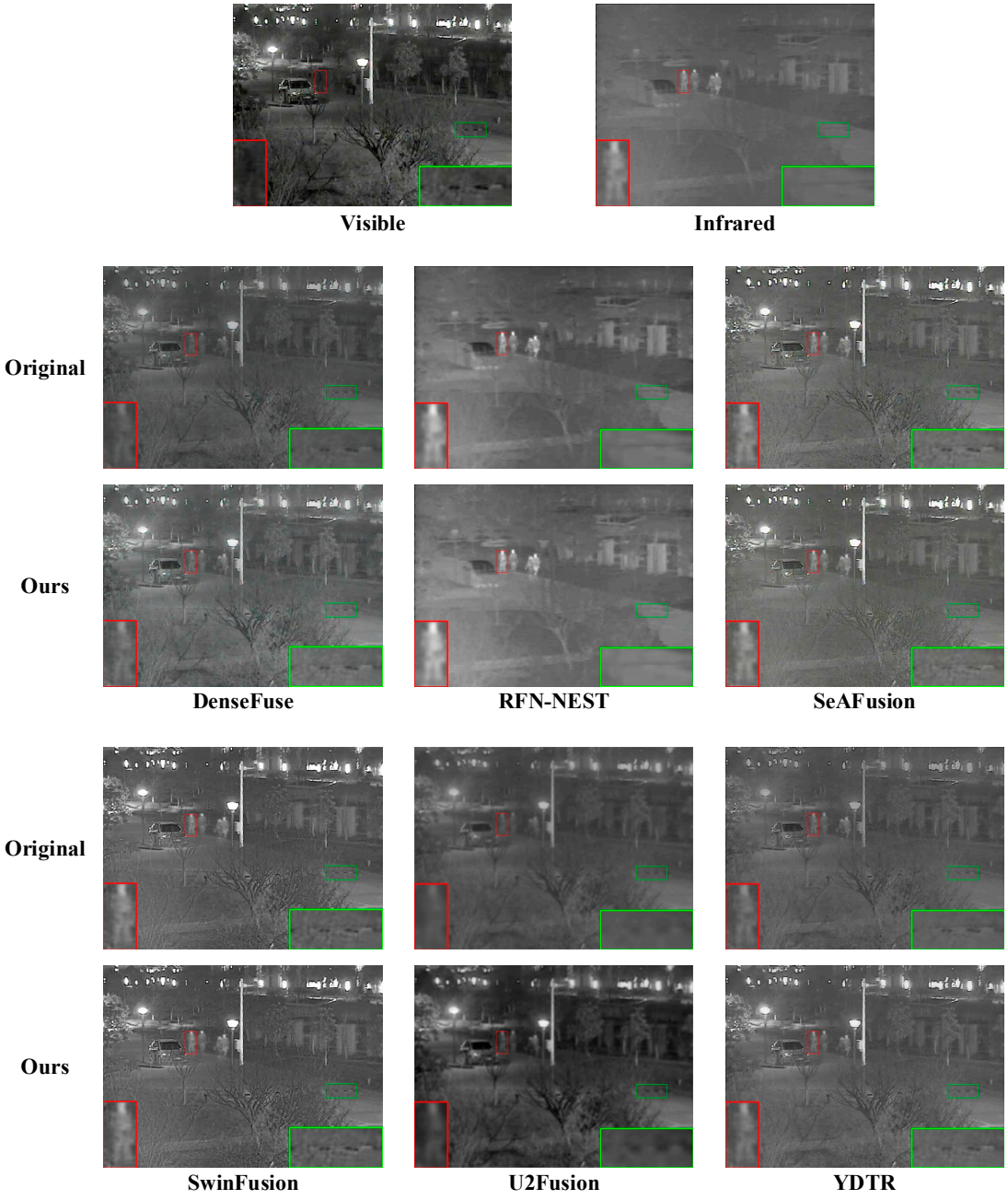

Figure 4. Examples of qualitative analyses on the VIFB dataset. The red box in the image zooms in on the local area where the pedestrian is located, while the green box zooms in on the local area where the ground object is located.

In terms of the VIF metric based on human visual perception, the original RFN-NEST algorithm achieves the best performance, while the results obtained by our distilled algorithm come in second place, but both outperform the other algorithms by a significant margin. Regarding the VIF metric based on human visual perception, the original RFN-NEST algorithm shows the best performance, followed by the distilled algorithm we propose, both of which show significant differences from the other algorithms. The quality of the input image pairs also has a significant impact on the fusion results. For example, for

image pair number 17, the RFN-NEST algorithm obtains a very low score in the MS_SSIM metric, which directly leads to poor performance of the corresponding distilled adaptive DenseNet on image pair number 17.

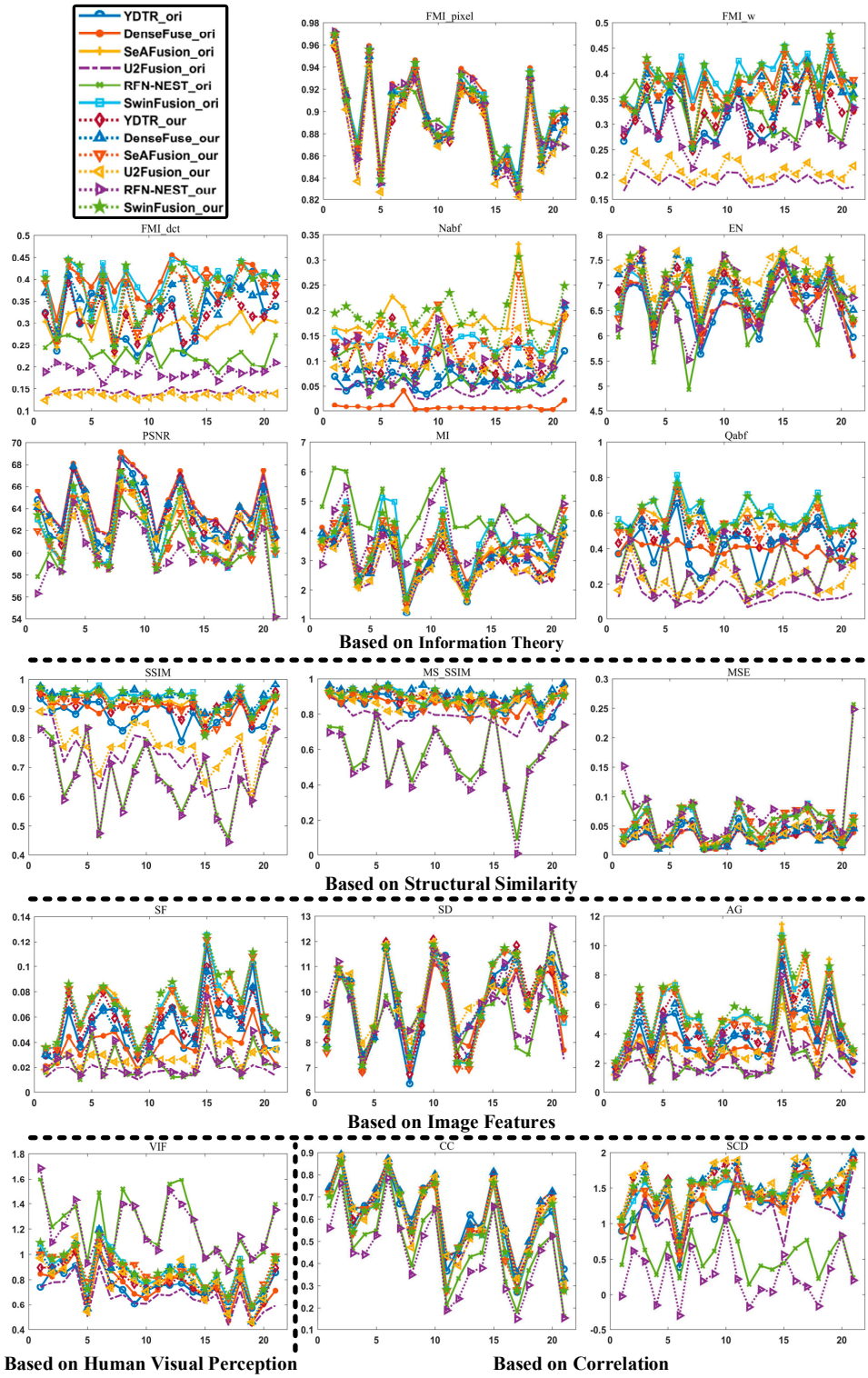

Figure 5. Quantitative analysis results, where the original model is denoted with a suffix '_ori', while the distilled model is denoted with a suffix '_our'.

Table 8 presents the average scores for each model corresponding to each metric. According to Table 8, the following conclusions can be obtained. Firstly, the proposed

knowledge distillation method effectively fits different dual-light fusion models based on different network structures, with similar or even better performance on various evaluation metrics compared to the original models. For example, for the YDTR algorithm, the distilled model achieves an SSIM value of 0.915, which exceeds the original model's 0.879. This is because the distilled model's network structure is based on DenseNet, which can maximize the retention of features extracted from dual-light images at each convolutional layer. Although the knowledge distillation process is aimed at fitting the original model, the adaptive DenseNet distilled after the process is better at feature extraction and retention for fusion images, resulting in a higher SSIM value than the original YDTR model. Secondly, among the 17 evaluation metrics, the original six fusion models achieved 10 first places, 8 second places, and 7 third places, while our proposed method achieved 7 first places, 9 second places, and 10 third places. Specifically, the distilled fusion model performs poorly in the Nabf metric, and it is found that all other original models exhibited the same problem. Therefore, the corresponding distilled model's poor performance in this metric is mainly due to the original model's low performance. In contrast, in the three image-feature-based metrics, including SF, SD, and AG, the distilled models proposed in this study ranked first. This is because the adaptive DenseNet network structure obtained by our proposed method can better reflect the shallow image features and original pixel features in the fusion image compared to other model architectures, resulting in a better performance in contrast and texture features of the image.

Table 8. The average values for each model corresponding to each metric [62]. In the table, the number in red with an asterisk (*) indicates that it ranks first in the index, the bold black number indicates that it ranks second, and the bold blue number indicates that it ranks third.

	Information Theory							
	FMI_pixel	**FMI_w**	**FMI_dct**	**Nabf**	**EN**	**PSNR**	**MI**	**Qabf**
YDTR_ori	0.893	0.334	0.322	*0.062*	6.633	63.37	3.126	0.420
DenseFuse_ori	0.901 *	*0.381*	0.399 *	0.009 *	6.680	64.53 *	3.214	0.396
SeAFusion_ori	0.897	0.356	0.293	0.184	6.974	61.64	3.172	*0.560*
U2Fusion_ori	0.885	0.187	0.143	**0.044**	6.578	**64.35**	2.763	0.140
RFN-NEST_ori	0.896	0.333	0.231	0.081	6.606	60.58	4.602 *	0.277
SwinFusion_ori	**0.900**	0.394 *	**0.395**	0.139	6.921	61.81	*3.616*	0.593 *
YDTR_our	0.891	0.326	0.316	0.110	6.923	63.41	2.971	0.479
DenseFuse_our	0.895	0.358	0.351	0.088	**7.032**	*63.82*	3.097	0.506
SeAFusion_our	0.895	0.369	0.367	0.144	6.844	61.34	3.426	0.532
U2Fusion_our	0.885	0.208	0.135	0.107	7.269 *	62.88	2.857	0.210
RFN-NEST_our	0.895	0.284	0.190	0.101	6.734	60.00	**3.836**	0.271
SwinFusion_our	*0.899*	**0.383**	*0.376*	0.187	*7.023*	62.02	3.472	**0.575**

	Structural Similarity			**Image Features**			**Human Visual Perception**	**Correlation**	
	SSIM	**MS_SSIM**	**MSE**	**SF**	**SD**	**AG**	**VIF**	**CC**	**SCD**
YDTR_ori	0.879	0.855	0.035	0.053	9.535	3.854	0.747	0.607	1.294
DenseFuse_ori	0.906	0.880	0.026 *	0.039	9.452	3.237	0.746	0.637 *	1.300
SeAFusion_ori	0.925	0.889	0.051	**0.069**	*9.668*	**5.604**	0.850	0.587	1.446
U2Fusion_ori	0.733	0.786	**0.027**	0.018	9.311	1.839	0.661	0.608	1.169
RFN-NEST_ori	0.666	0.567	0.069	0.025	9.375	2.416	1.223 *	0.484	0.566
SwinFusion_ori	**0.938**	*0.899*	0.051	0.067	9.559	*5.373*	0.874	0.585	1.455
YDTR_our	0.915	**0.903**	0.034	0.057	9.591	4.514	0.773	*0.614*	*1.515*
DenseFuse_our	0.938 *	0.925 *	*0.030*	0.053	**9.805**	4.335	0.823	**0.636**	**1.530**
SeAFusion_our	0.909	0.867	0.054	*0.067*	9.554	5.133	0.867	0.586	1.412
U2Fusion_our	0.778	0.873	0.036	0.028	9.984 *	2.954	0.796	0.607	1.539 *
RFN-NEST_our	0.658	0.539	0.076	0.027	9.600	2.662	**1.187**	0.443	0.267
SwinFusion_our	*0.932*	0.898	0.049	0.071 *	9.637	5.678 *	*0.880*	0.585	1.476

5. Real-World Applications and Results Analysis

This section presents the practical implementation and analysis of our algorithm deployed on mobile platforms. In order to demonstrate the efficacy of our approach, we design and implement a real-time fusion detection system as per the framework illustrated in Figure 6. The system utilizes a DJI M300 RTK unmanned aerial vehicle as the airborne payload platform and a Zenmuse H20T as the visible light and infrared sensor payload. In our real-time fusion and detection system scheme A, we employes an NVIDIA Jetson Xavier NX as the computing platform and directly installed it on the drone to perform data acquisition, visible light and infrared image fusion, and target detection of the fused image. Subsequently, the fusion detection results were transmitted to a laptop for display. In contrast, in the real-time fusion and detection system scheme B, a laptop functions as the computing platform. The laptop reads the required visible and infrared images from the DJ remote control and conducts visible and infrared image fusion, followed by detecting targets in the fused images. The fusion detection results are displayed simultaneously.

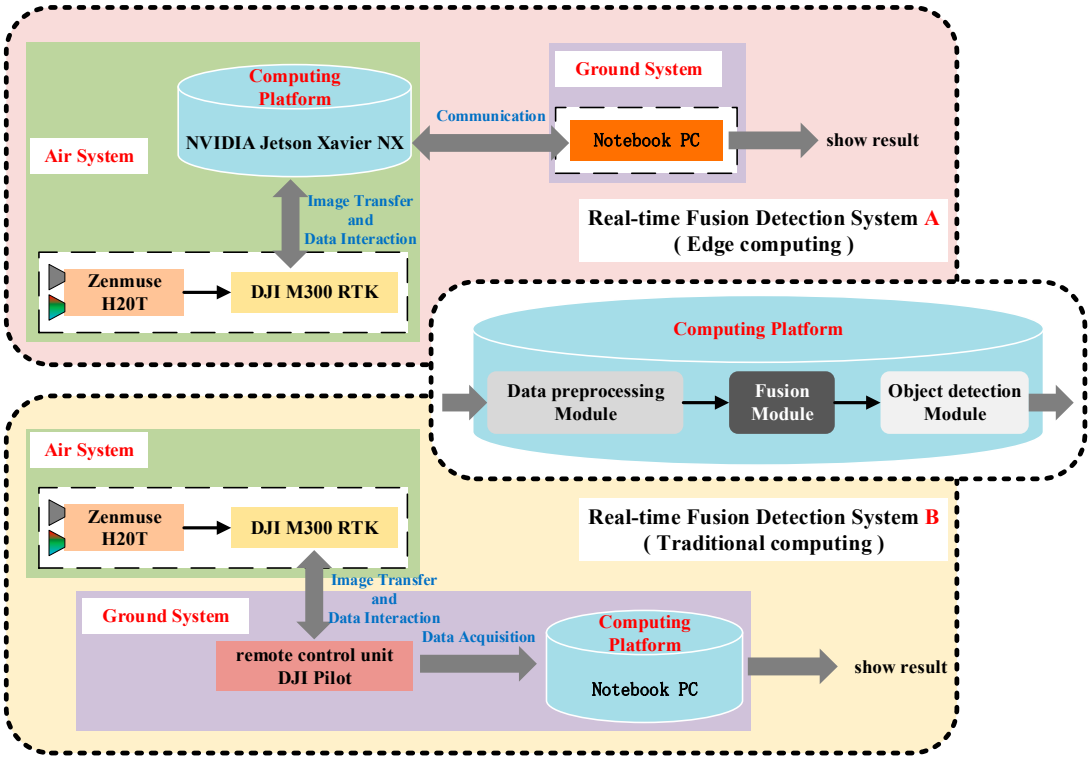

Figure 6. The practical implementation and framework of our algorithm deployed on mobile platforms.

In Figure 7a, the real-life application setup of the deployment experiment is depicted. The left image displays the unmanned aerial platform with a payload section, while the right image showcases the ground dual light image data processing section. Specifically, the Zenmuse H20T, with infrared images on the left and visible light images on the right in each frame, can generate a 30 fps and 1080 p video output. The laptop used in this experiment features an Intel Core i7-12800HX CPU and an NVIDIA GeForce RTX 3070Ti Laptop GPU with 8 GB memory. Moreover, the NVIDIA Jetson Xavier NX comprises a six-core NVIDIA Carmel ARM processor and an NVIDIA Volta architecture graphics card with 384 NVIDIA CUDA cores. Figure 7b displays the outcomes of the real-time fusion detection algorithm utilizing the adaptive DenseNet fitting SeAFusion. For the detection algorithm, yolov5s was utilized as the network structure adopting YOLOv5. Notably, our algorithm produces comparable results to the original SeAFusion. The fusion results demonstrate that individuals can be distinctly differentiated and the detection network can correctly identify targets in the fused image.

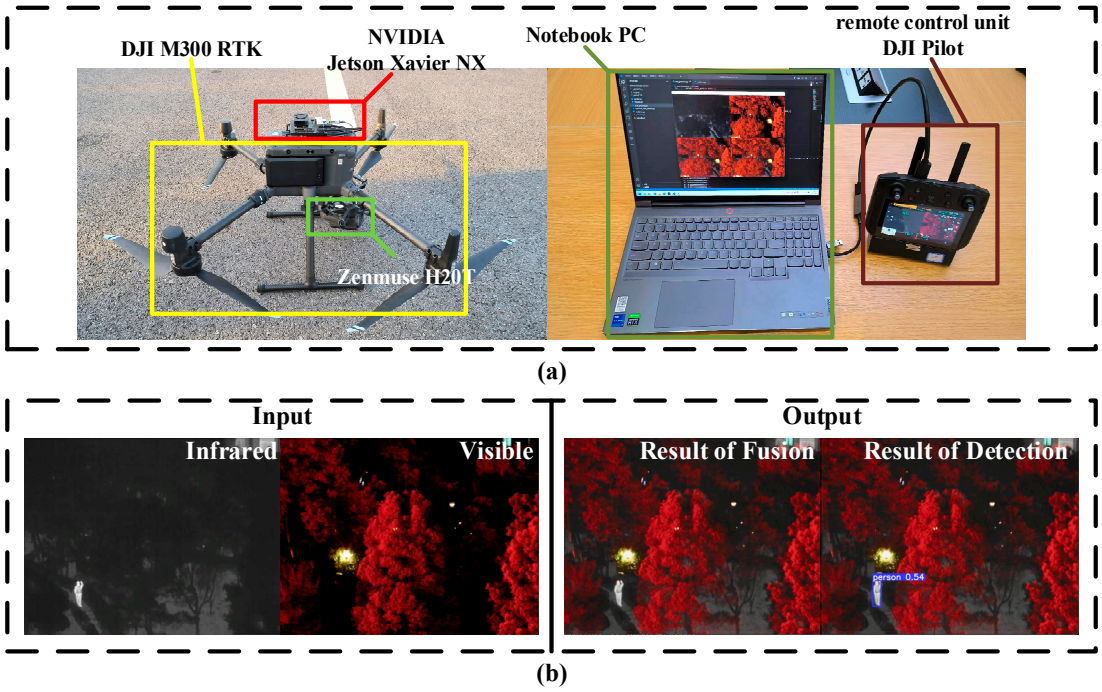

Figure 7. The real-life application setup of the deployment experiment and the outcomes of dual light fusion and detection.

Table 9 provides an analysis of time consumption for two real-time fusion systems. The original fusion algorithm consumed more time than the image preprocessing time, and the time consumed relative to detection was also significant and cannot be ignored. This indicates that the original algorithm may not be efficient enough for practical applications. In contrast, our proposed fusion process consumes a significantly lower proportion of time compared to the detection process, accounting for less than one-sixth of the detection time. Following a multi-process optimization, scheme A yields a frame rate of 10.1 fps, while scheme B approaches the upper limit of the original video's frame rate, achieving 28 fps. Therefore, our algorithm may be more suitable for practical applications where efficiency is a concern.

Table 9. Analysis of time consumption for two real-time fusion systems.

Computing Platform	Fusion Module	Average Time Consumption (ms)				
		Preprocessing	Fusion	Detection	NMS	Results Presentation
Jetson Xavier NX	Original	23.4	29.4	75.8	14.4	92.0
	Ours		7.6			
Notebook PC	Original	3.5	6.6	9.6	2.0	17.3
	Ours		1.6			

6. Conclusions

This paper presents a novel method for infrared and visible image fusion that employs adaptive DenseNet and knowledge distillation to compress six existing dual-light image fusion models based on different algorithms, including DenseFuse, U2Fusion, RFN-NEST, YDTR, SwinFusion, and SeAFusion. The proposed method involves utilizing the designed adaptive DenseNet to learn from pre-existing fusion models and implementing the fusion of visible and infrared images as a solution to the neural network fitting problem by using the hyperparameters of the model's structure, including the inference time and dense layer number. Then, the best adaptive DenseNet networks are generated corresponding to the respective existing fusion models. In addition, we have carefully selected a number of

dual-light image pairs in public datasets for alignment and markinge and produced a novel representative infrared and visible light dataset in multiple scenes to verify the performance of various fusion models after knowledge distillation by 17 popular evaluation metrics of dual-light fusion. The proposed method has undergone qualitative and quantitative experimental evaluations, which have demonstrated its effectiveness in matching the performance of the original fusion models and achieving model compression with a shorter inference time. Furthermore, the deployment of the adaptive DenseNet on edge computing devices has exhibited the proposed method's efficacy in real-world applications. Despite the exceptional outcomes, our approach presents potential avenues for further research and enhancement. It may be imperative to explore the influence of other hyperparameters on the knowledge distillation performance of fusion models.

Author Contributions: Z.Z.: conceptualization, methodology, investigation, and software. S.S.: supervision and validation. J.W.: writing—original draft, visualization, and software. X.T.: conception, design, and resources. W.G.: software and data curation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the National Natural Science Youth Foundation of China under grant no. 62201598.

Data Availability Statement: The data for this study are openly available at the following links: https://github.com/Linfeng-Tang/MSRS for MSRS [55]; https://github.com/JinyuanLiu-CV/TarDAL for M3FD [60]; https://github.com/bupt-ai-cz/LLVIP for LLVIP [61]; https://github.com/xingchenzhang/VIFB for VIFB [63].

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationship that could have appeared to influence the work reported in this paper.

References

1. Ma, W.; Wang, K.; Li, J.; Yang, S.X.; Li, J.; Song, L.; Li, Q. Infrared and Visible Image Fusion Technology and Application: A Review. *Sensors* **2023**, *23*, 599. [CrossRef]
2. Guo, X.; Yang, F.; Ji, L. MLF: A mimic layered fusion method for infrared and visible video. *Infrared Phys. Technol.* **2022**, *126*, 104349. [CrossRef]
3. Zhao, Z.; Xu, S.; Zhang, C.; Liu, J.; Zhang, J.; Li, P. DIDFuse: Deep Image Decomposition for Infrared and Visible Image Fusion. *arXiv* **2020**, arXiv:2003.09210.
4. Xia, J.; Lu, Y.; Tan, L. Research of Multimodal Medical Image Fusion Based on Parameter-Adaptive Pulse-Coupled Neural Network and Convolutional Sparse Representation. *Comput. Math. Methods Med.* **2020**, *2020*, 3290136. [CrossRef]
5. Nencini, F.; Garzelli, A.; Baronti, S.; Alparone, L. Remote sensing image fusion using the curvelet transform. *Inf. Fusion* **2007**, *8*, 143–156. [CrossRef]
6. Bin Peng, X.; Coumans, E.; Zhang, T.; Lee, T.-W.; Tan, J.; Levine, S. Learning Agile Robotic Locomotion Skills by Imitating Animals. *arXiv* **2020**, arXiv:2004.00784.
7. Rai, A.K.; Senthilkumar, R.; Kumar, A. Combining pixel selection with covariance similarity approach in hyperspectral face recognition based on convolution neural network. *Microprocess. Microsyst.* **2020**, *76*, 103096. [CrossRef]
8. Wang, M.; Liu, R.; Hajime, N.; Narishige, A.; Uchida, H.; Matsunami, T. Improved knowledge distillation for training fast low resolution face recognition model. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Montreal, BC, Canada, 11–17 October 2021.
9. Ju, Y.; Lam, K.M.; Xiao, J.; Zhang, C.; Yang, C.; Dong, J. Efficient Feature Fusion for Learning-Based Photometric Stereo. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
10. Lewis, J.J.; O'Callaghan, R.J.; Nikolov, S.G.; Bull, D.R.; Canagarajah, N. Pixel-and region-based image fusion with complex wavelets. *Inf. Fusion* **2007**, *8*, 119–130. [CrossRef]
11. Zhu, Z.; Yin, H.; Chai, Y.; Li, Y.; Qi, G. A novel multi-modality image fusion method based on image decomposition and sparse representation. *Inf. Sci.* **2018**, *432*, 516–529. [CrossRef]
12. Zhang, Q.; Liu, Y.; Blum, R.S.; Han, J.; Tao, D. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Inf. Fusion* **2018**, *40*, 57–75. [CrossRef]
13. Zhang, Q.; Li, G.; Cao, Y.; Han, J. Multi-focus image fusion based on non-negative sparse representation and patch-level consistency rectification. *Pattern Recognit.* **2020**, *104*, 107325. [CrossRef]
14. Zhang, S.; Li, X.; Zhang, X.; Zhang, S. Infrared and visible image fusion based on saliency detection and two-scale transform decomposition. *Infrared Phys. Technol.* **2021**, *114*, 103626. [CrossRef]

15. Chen, J.; Wu, K.; Cheng, Z.; Luo, L. A saliency-based multiscale approach for infrared and visible image fusion. *Signal Process.* **2021**, *182*, 107936. [CrossRef]
16. Liu, C.; Qi, Y.; Ding, W. Infrared and visible image fusion method based on saliency detection in sparse domain. *Infrared Phys. Technol.* **2017**, *83*, 94–102. [CrossRef]
17. Ma, J.; Tang, L.; Xu, M.; Zhang, H.; Xiao, G. STDFusionNet: An Infrared and Visible Image Fusion Network Based on Salient Target Detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13. [CrossRef]
18. Alghamdi, R.S.; Alshehri, N.O. Fusion of infrared and visible images using neutrosophic fuzzy sets. *Multimedia Tools Appl.* **2021**, *80*, 25927–25941. [CrossRef]
19. Cai, H.; Zhuo, L.; Chen, X.; Zhang, W. Infrared and visible image fusion based on BEMSD and improved fuzzy set. *Infrared Phys. Technol.* **2019**, *98*, 201–211. [CrossRef]
20. Cai, H.; Zhuo, L.; Zhu, P.; Huang, Z.; Wu, X. Fusion of infrared and visible images based on non-subsampled contourlet transform and intuitionistic fuzzy set. *Acta Photonica Sin.* **2018**, *47*, 125479664.
21. Yin, W.; He, K.; Xu, D.; Luo, Y.; Gong, J. Adaptive enhanced infrared and visible image fusion using hybrid decomposition and coupled dictionary. *Neural Comput. Appl.* **2022**, *34*, 20831–20849. [CrossRef]
22. Luo, Y.; He, K.; Xu, D.; Yin, W.; Liu, W. Infrared and visible image fusion based on visibility enhancement and hybrid multiscale decomposition. *Optik* **2022**, *258*, 168914. [CrossRef]
23. Zhang, Y.; Li, D.; Zhu, W. Infrared and Visible Image Fusion with Hybrid Image Filtering. *Math. Probl. Eng.* **2020**, *2020*, 1757214. [CrossRef]
24. Ren, L.; Pan, Z.; Cao, J.; Liao, J. Infrared and visible image fusion based on variational auto-encoder and infrared feature compensation. *Infrared Phys. Technol.* **2021**, *117*, 103839. [CrossRef]
25. Xu, H.; Gong, M.; Tian, X.; Huang, J.; Ma, J. CUFD: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition. *Comput. Vis. Image Underst.* **2022**, *218*, 103407. [CrossRef]
26. Su, W.; Huang, Y.; Li, Q.; Zuo, F.; Liu, L. Infrared and Visible Image Fusion Based on Adversarial Feature Extraction and Stable Image Reconstruction. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–14. [CrossRef]
27. An, W.-B.; Wang, H.-M. Infrared and visible image fusion with supervised convolutional neural network. *Optik* **2020**, *219*, 165120. [CrossRef]
28. Liu, Y.; Chen, X.; Cheng, J.; Peng, H.; Wang, Z. Infrared and visible image fusion with convolutional neural networks. *Int. J. Wavelets Multiresolut. Inf. Process.* **2018**, *16*, 1850018. [CrossRef]
29. Liu, N.; Zhou, D.; Nie, R.; Hou, R. Infrared and visible image fusion based on convolutional neural network model and saliency detection via hybrid l0-l1 layer decomposition. *J. Electron. Imaging* **2018**, *27*, 063036. [CrossRef]
30. Hou, J.; Zhang, D.; Wu, W.; Ma, J.; Zhou, H. A Generative Adversarial Network for Infrared and Visible Image Fusion Based on Semantic Segmentation. *Entropy* **2021**, *23*, 376. [CrossRef]
31. Li, J.; Huo, H.; Liu, K.; Li, C. Infrared and visible image fusion using dual discriminators generative adversarial networks with Wasserstein distance. *Inf. Sci.* **2020**, *529*, 28–41. [CrossRef]
32. Li, J.; Huo, H.; Li, C.; Wang, R.; Feng, Q. AttentionFGAN: Infrared and Visible Image Fusion Using Attention-Based Generative Adversarial Networks. *IEEE Trans. Multimed.* **2021**, *23*, 1383–1396. [CrossRef]
33. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2018**, *48*, 11–26. [CrossRef]
34. Wang, Z.; Chen, Y.; Shao, W.; Li, H.; Zhang, L. SwinFuse: A Residual Swin Transformer Fusion Network for Infrared and Visible Images. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [CrossRef]
35. Rao, D.; Wu, X.; Xu, T. TGFuse: An Infrared and Visible Image Fusion Approach Based on Transformer and Generative Ad-versarial Network. *arXiv* **2022**, arXiv:2201.10147. [CrossRef] [PubMed]
36. Li, J.; Zhu, J.; Li, C.; Chen, X.; Yang, B. CGTF: Convolution-Guided Transformer for Infrared and Visible Image Fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–14. [CrossRef]
37. Tang, W.; He, F.; Liu, Y. TCCFusion: An infrared and visible image fusion method based on transformer and cross correlation. *Pattern Recognit.* **2023**, *137*, 109295. [CrossRef]
38. Yi, S.; Jiang, G.; Liu, X.; Li, J.; Chen, L. TCPMFNet: An infrared and visible image fusion network with composite auto encoder and transformer–convolutional parallel mixed fusion strategy. *Infrared Phys. Technol.* **2022**, *127*, 104405. [CrossRef]
39. Xiao, Z.; Xie, P.; Wang, G. Multi-scale Cross-Modal Transformer Network for RGB-D Object Detection. In Proceedings of the MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, 6–10 June 2022; pp. 352–363.
40. Wang, X.; Wang, X.; Song, R.; Zhao, X.; Zhao, K. MCT-Net: Multi-hierarchical cross transformer for hyperspectral and multi-spectral image fusion. *Knowl. -Based Syst.* **2023**, *264*, 110362. [CrossRef]
41. Zhou, D.; Jin, X.; Jiang, Q.; Cai, L.; Lee, S.; Yao, S. MCRD-Net: An unsupervised dense network with multi-scale convolutional block attention for multi-focus image fusion. *IET Image Process.* **2022**, *16*, 1558–1574. [CrossRef]
42. Zhang, D.; Song, K.; Xu, J.; He, Y.; Niu, M.; Yan, Y. MCnet: Multiple Context Information Segmentation Network of No-Service Rail Surface Defects. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 5004309. [CrossRef]
43. Niyaz, U.; Bathula, D.R. Augmenting Knowledge Distillation with Peer-to-Peer Mutual Learning for Model Compression. In Proceedings of the 19th International Symposium on Biomedical Imaging (ISBI), Kolkata, India, 28–31 March 2022.

44. Chen, P.; Liu, S.; Zhao, H.; Jia, J. Distilling Knowledge via Knowledge Review. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5006–5015.
45. Xiao, W.; Zhang, Y.; Wang, H.; Li, F.; Jin, H. Heterogeneous Knowledge Distillation for Simultaneous Infrared-Visible Image Fusion and Super-Resolution. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–15. [CrossRef]
46. Liu, X.; Hirota, K.; Jia, Z.; Dai, Y. A multi-autoencoder fusion network guided by perceptual distillation. *Inf. Sci.* **2022**, *606*, 1–20. [CrossRef]
47. Zhao, F.; Zhao, W.; Lu, H.; Liu, Y.; Yao, L.; Liu, Y. Depth-Distilled Multi-Focus Image Fusion. *IEEE Trans. Multimed.* **2021**, *25*, 966–978. [CrossRef]
48. Mi, J.; Wang, L.; Liu, Y.; Zhang, J. KDE-GAN: A multimodal medical image-fusion model based on knowledge distillation and explainable AI modules. *Comput. Biol. Med.* **2022**, *151*, 106273. [CrossRef] [PubMed]
49. Lu, X.; Zhang, L.; Niu, L.; Chen, Q.; Wang, J. A Novel Adaptive Feature Fusion Strategy for Image Retrieval. *Entropy* **2021**, *23*, 1670. [CrossRef] [PubMed]
50. Wang, L.; Hu, Z.; Kong, Q.; Qi, Q.; Liao, Q. Infrared and Visible Image Fusion via Attention-Based Adaptive Feature Fusion. *Entropy* **2023**, *25*, 407. [CrossRef] [PubMed]
51. Zeng, S.; Zhang, Z.; Zou, Q. Adaptive deep neural networks methods for high-dimensional partial differential equations. *J. Comput. Phys.* **2022**, *463*, 111232. [CrossRef]
52. Yuan, J.; Pan, F.; Zhou, C.; Qin, T.; Liu, T.Y. Learning Structures for Deep Neural Networks. *arXiv* **2021**, arXiv:2105.13905.
53. Li, H.; Yang, Y.; Chen, D.; Lin, Z. Optimization Algorithm Inspired Deep Neural Network Structure Design. *arXiv* **2018**, arXiv:1810.01638.
54. Li, H.; Wu, X.-J. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [CrossRef]
55. Tang, L.; Yuan, J.; Ma, J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion* **2022**, *82*, 28–42. [CrossRef]
56. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef] [PubMed]
57. Tang, W.; He, F.; Liu, Y. YDTR: Infrared and Visible Image Fusion via Y-shape Dynamic Transformer. *IEEE Trans. Multimedia* **2022**, 1–16. [CrossRef]
58. Hui, L.; Xjw, A.; Jk, B. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Inf. Fusion* **2021**, *73*, 72–86.
59. Jin, Z.R.; Deng, L.J.; Zhang, T.J.; Jin, X.X. BAM: Bilateral activation mechanism for image fusion. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 4315–4323.
60. Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; Luo, Z. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 5802–5811.
61. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. *LLVIP: A Visible-Infrared Paired Dataset for Low-Light Vision*; Beijing Laboratory of Advanced Information Networks, Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications: Beijing, China, 2021.
62. Tang, L.; Zang, H.; Xu, H.; Ma, J.Y. Deep learning-based image fusion: A survey. *J. Image Graph.* **2023**, *28*, 3–36.
63. Zhang, X.; Ye, P.; Xiao, G. VIFB: A Visible and Infrared Image Fusion Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; Shanghai Jiao Tong University, School of Aeronautics and Astronautics: Shanghai, China, 2020.