VIFA: An Efficient Visible and Infrared Image Fusion Architecture for Multi-task Applications via Continual Learning

Jiaxing Shi[1], Ao Ren*[1], Wei Zhuang*[2], Yang Hua[2], ZhiYong Qin[2], Zhenyu Wang[1], Yang Song[1], Yujuan Tan[1], and Duo Liu[1]

[1] College of Computer Science, Chongqing University, China
sjx@stu.cqu.edu.cn, ren.ao@cqu.edu.cn, zhenyuwang@cqu.edu.cn, sy@stu.cqu.edu.cn, tanyujuan@gmail.com, liuduo@cqu.edu.cn
[2] Beijing Microelectronics Technology Institute, China
{zhuangw, huay, qinzhy}@mxtronics.com

Abstract. Visible-infrared image fusion has attracted great attention in a range of computer vision applications. Aiming at improving task-specific performance, recent studies have employed a cascading approach, where the fusion network is trained using feedback from the specific downstream task network. However, this training strategy will result in the overfitting of the fusion network, and deploying a different fusion network for each downstream task is inefficient for multi-task scenarios. To address this challenge, we propose VIFA, a visible-infrared image fusion architecture for multi-task applications. This architecture effectively mitigates the catastrophic forgetting problem by partitioning the fusion network into a knowledge-sharing backbone and task-specific components. To facilitate knowledge sharing, we introduce a key channel-constrained distillation strategy, which identifies and retains informative features, while allowing non-critical channels to learn new knowledge. In addition, we propose a reference model-guided distillation to compress the task-specific components while maintaining model performance. Evaluations on multiple representative fusion networks show that VIFA can significantly improve task performance and speed.

Keywords: Visible-infrared Image Fusion · Knowledge Distillation · Continual Learning · Multi-task Inference

1 INTRODUCTION

Visible-infrared image fusion aims to combine the advantages of the two types of signals. Infrared images can capture the thermal information of objects but lack texture information and are susceptible to noise. On the contrary, visible images

* Corresponding author
This work was supported by the Natural Science Foundation of China No. 62102051 and No. 62072059. We would like to thank the anonymous reviewers for their valuable comments and improvements to this paper.

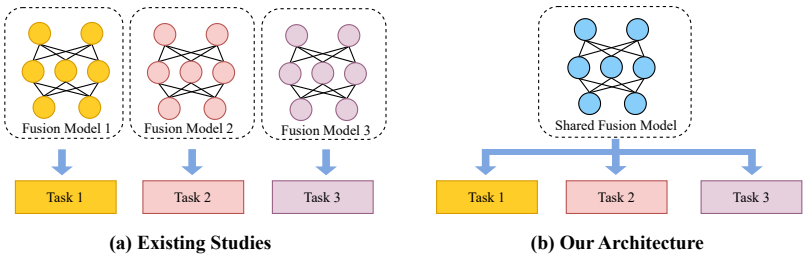

Fig. 1. Comparison of our architecture with prior studies.

typically contain rich texture details but are sensitive to illumination changes, which result in poorly imaged objects in low-light environments. Visible-infrared image fusion has been widely deployed in computer vision tasks, such as image segmentation[21], object detection[25][19], object tracking [30], and biometric[1].

Toward accommodating the fusion images to downstream tasks, recent studies utilize specific downstream tasks to facilitate the training of fusion models. Tardal[10] employs an object detection network to perform a bi-level optimization process. DetFusion[20] uses object-related information learned from the object detection network to guide the multimodal image fusion process, thus motivating the fusion network to learn object-specific information. SeAFusion[21] and SegMif[11] employ feedback from the semantic segmentation networks to facilitate the training of fusion networks. Nonetheless, these methods overfit in a specific task and suffer in other downstream tasks. In multi-task scenarios, this approach requires multiple fusion models that are jointly optimized with the downstream tasks, incurring significant storage and computational overheads, and are inefficient in resource-constrained edge devices.

Inspired by continual learning, we explore the architecture as illustrated in Fig. 1, where one fusion network is shared by all the downstream tasks, thereby improving multi-task inference efficiency. Nevertheless, direct training on this architecture results in catastrophic forgetting issues within the fusion network. Conventional continual learning studies have proposed various solutions to overcome the catastrophic forgetting issues. However, they primarily focus on the supervised learning field, such as image classification [7][18], segmentation[5][34], and object detection [4][12]. They utilize ground-truth labels to alleviate catastrophic forgetting, while visible-infrared image fusion falls into the unsupervised learning field, therefore, traditional continual learning methods cannot be directly applied to fusion networks.

To address this challenge, we further analyze the catastrophic forgetting issue in visible-infrared image fusion networks. The results lead to the following findings: i) Fusion models trained with different fusion datasets exhibit relatively mild catastrophic forgetting. ii) When the fusion model is jointly trained with different downstream tasks, the catastrophic forgetting becomes much more severe. The results imply that the fusion knowledge learned from different fusion

datasets can be shared, while the knowledge learned from downstream tasks is difficult to share and incurs forgetting.

Based on our findings, we propose VIFA, a task-specific parameter-separated architecture. VIFA partitions the fusion network into a knowledge-sharing component and task-specific components. To effectively preserve learned knowledge in the knowledge-sharing component, we analyze the feature map characteristics of the fusion network. We reveal that the channels of the feature maps hold high similarities and are inclined to cluster together. Accordingly, we propose the key channel-constrained distillation strategy. For each cluster, the channel that holds the highest entropy value is defined as the key channel. The key channels are constrained when training on new tasks, aiming at preserving the knowledge learned from the previous task. The task-specific components are jointly trained with the specific downstream task models. However, the sizes of the task-specific components vary with downstream models and can be too large to deploy efficiently. Consequently, we propose a reference model-guided distillation strategy for the task-specific components. This strategy can produce high-quality fusion images with only one layer, thereby improving the multi-task inference performance while maintaining the accuracy of downstream tasks. It is important to note that this work focuses on designing an efficient multi-task architecture, and the proposed methods are orthogonal to other fine-grained model compression methods, such as pruning and quantization.

Our contributions are summarized as follows:

- We figure out the cause of the catastrophic forgetting issues in fusion networks, and we propose VIFA that partitions the fusion network into a knowledge-sharing component and task-specific components to address the issues.
- We reveal that the intermediate features in the fusion model hold high similarities, based on which, we propose the key channel-constrained distillation to preserve the learned knowledge in the knowledge-sharing backbone.
- We propose a reference model-guided distillation strategy that compresses the task-specific component to only one layer, leading to high efficiency for multi-task inference.

VIFA is evaluated with multiple representative networks, and the results show that our approach achieves 1.85×-1.98× speedups compared to prior studies.

2 RELATED WORK

2.1 Visible-infrared Image Fusion

In recent years, multimodal image fusion algorithms based on deep learning have made significant progress. By innovating the network architecture, researchers have significantly enhanced the capacity of the fusion network to extract features, which generates more informative fusion images. DIDFuse[32] decomposes the input images into background and detail features through an encoder, while a decoder is used to reconstruct the fusion image. FusionDN[24] develops an

adaptive fusion framework that dynamically evaluates the significance of different image sources. PIAFusion[22] introduces an illumination-aware module that guides the fusion process through an illumination-aware loss function. To address the issue of long-range dependencies in image fusion, SwinFusion[15] employs a combination of CNN and Swin Transformer architectures to extract features containing both local and global information. CDDFuse[31] uses correlation-driven feature decomposition fusion for multimodal feature decomposition and image fusion.

To adapt to the needs of downstream visual tasks, some studies have used specific downstream tasks to facilitate the training process of the fusion network. SeAFusion[21] provides the fusion network with semantic feedback by cascading the semantic segmentation network. SegMif[11] integrates the solutions to the fusion and segmentation tasks into a single optimization objective. Similarly, Tardal[10] constrains the fusion network to retain comprehensive target information from the perspective of target detection. DetFusion[20] independently trains object detection models based on infrared and visible images and utilizes these models to guide the training of the fusion network jointly. BDLFusion[14] has further investigated the potential of utilizing feedback from specific downstream task networks to optimize the fusion network. However, these approaches guide the fusion models through task-specific models, which may result in a reduction in the performance of the generated fusion images on other tasks. If the fusion model is tailored to each downstream task, it results in additional costs on computational and storage resources, which constrains the model's feasibility for multi-task applications in constrained edge environments. Consequently, there is a necessity to develop architectures that can simultaneously fulfill the requirements of multiple downstream tasks efficiently.

2.2 Continual Learning

Continual learning aims to overcome catastrophic forgetting in neural networks by enabling a single model to integrate knowledge from multiple datasets. This approach is promising to realize a multi-task architecture for fusion networks. The Elastic Weight Consolidation (EWC)[8] algorithm is initially introduced into the image fusion field by FusionDN[24] so that the fusion network can support the fusion of multiple source images, including multi-exposure images, visible infrared images, and multi-focus images. During training, an additional L2 loss is appended to the original weights to maintain previous knowledge. However, subsequent studies have shown that the EWC algorithm is suboptimal. The majority of contemporary research in continual learning focuses on supervised learning tasks, such as classification, segmentation, and detection tasks. Replay-based approaches, such as PCR[9] and CO2L[2], propose to store partial data from prior datasets and train the model with a mixture of old and new data. However, accessing prior data is usually restricted due to privacy and security considerations. Some studies maintain knowledge of old classes through class similarity. CL-DETR[13] maintains model consistency of known data by merging the class predictions of the old model and the new class labels, thereby ensuring

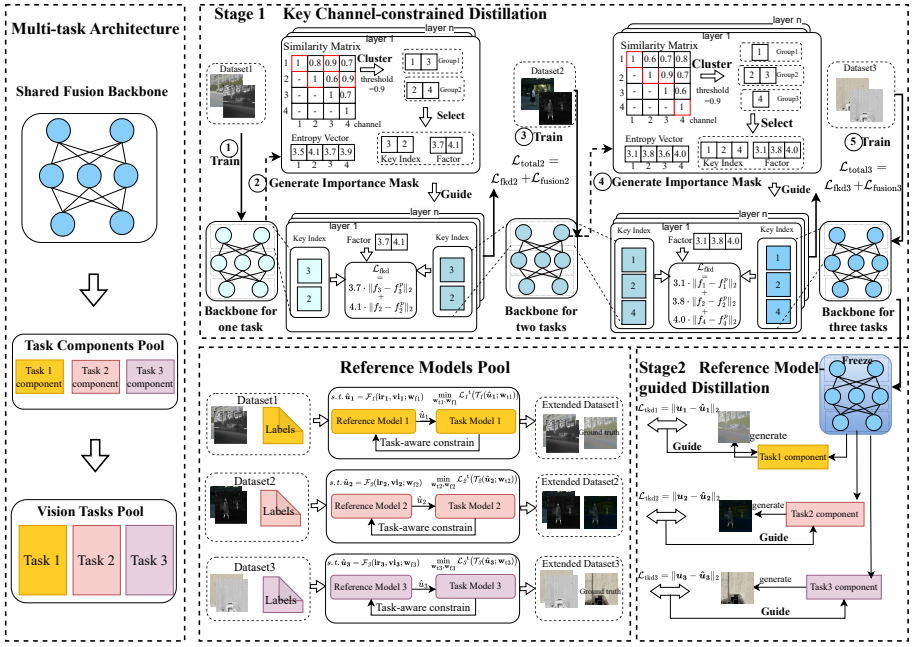

Fig. 2. Overview Framework of VIFA. Left shows the multi-task architecture, and right shows the two-stage training process.

the prediction performance for known classes. LGKD[26] employs a label-guided knowledge distillation loss to form semantically appropriate class correspondences with new model outputs. These methods typically rely on information such as class labels to facilitate the retention of old knowledge in new models. However, fusion models, as an unsupervised learning task, lack direct labeling information, which makes it difficult to apply existing continual learning strategies directly to fusion networks. Besides, some studies have investigated continual learning for self-supervised tasks. However, these studies[28][33][27] focus on the network's ability to work across multiple datasets, and their approaches are hard to address the challenge of catastrophic forgetting and conflict caused by downstream tasks.

To address this challenge, this paper proposes a new solution aimed at overcoming the catastrophic forgetting of fusion networks and thus enabling their applications in multi-task learning architectures.

3 METHODOLOGY

In this section, we first investigate the catastrophic forgetting issues in the visible-infrared image fusion networks, and the results indicate that task-related knowledge is harder to share than fusion knowledge. Inspired by this finding, we propose VIFA, which partitions the fusion network into a shared backbone and

task-specific components. Then we study the feature characteristics of the fusion network and propose a key channel-constrained distillation method, aiming at tackling the catastrophic forgetting issues. Furthermore, we propose a reference model-guided distillation strategy to compress the task-specific components and improve the multi-task inference efficiency.

3.1 Task-specific Parameter Separation Architecture

Table 1. Fusion and Task Performance Results. Seq 1 indicates that only the M3FD dataset was trained. Seq2 indicates that the model is trained again with the MSRS dataset.

Model		M3FD				MSRS				M3FD	MSRS
		EN	SF	VIF	SSIM	EN	SF	VIF	SSIM	mAP@0.5	mIOU
PIAFusion	Seq.1	7.4	24.16	0.78	0.88	-	-	-	-	0.794	-
	Seq.2	7.34	22.84	0.65	0.79	5.98	10.23	0.65	0.79	0.723	0.658
Tardal	Seq.1	6.43	18.10	0.83	0.86	-	-	-	-	0.826	-
	Seq.2	6.41	19.68	0.67	0.83	5.47	11.97	0.65	0.83	0.711	0.652

Traditional visible-infrared image fusion models are jointly trained with one specific downstream task and suffer from catastrophic forgetting issues. Aiming at proposing a visible-infrared image fusion architecture that supports multiple downstream tasks, we first analyze the cause of catastrophic forgetting. Specifically, We first train only the fusion network on M3FD[10] and MSRS[22] datasets in sequence, to analyze how much catastrophic forgetting effect the fusion datasets have applied. As shown in Table 1, the fusion images of Seq2 on the M3FD dataset exhibit stability in terms of entropy (EN) and structural fidelity (SF) compared to Seq1. However, Seq2 shows decreased visual information fidelity (VIF) and structural similarity index (SSIM) metrics. It reflects that training on new datasets has reduced the similarity between the fusion image and the source images though the amount of information in the fusion images does not change too much. This is due to the differences in the distribution of various datasets or biases in the collection devices. Furthermore, we jointly train the fusion network with the object detection network on the M3FD dataset, and subsequently, we jointly train the same fusion network with the semantic segmentation network on the MSRS dataset. The results in Table 1 show that the network jointly trained with the semantic segmentation network causes a significant accuracy loss in object detection, which indicates severe catastrophic forgetting.

Inspired by the results, we propose to partition the fusion model into a cross-task knowledge-sharing component and task-specific components, and the architecture is illustrated in Fig.2. The task-specific components are responsible for generating the fusion images that are adapted to the respective downstream tasks.

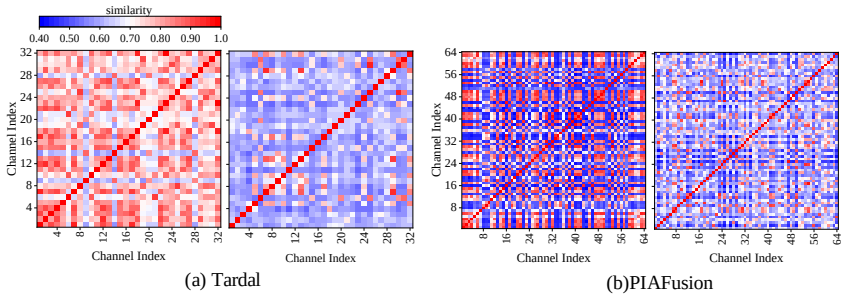

Fig. 3. Inter-channel similarity visualization on Tardal(a) and PIAFusion(b). The left part shows results from the M3FD dataset, while the right part shows results on the two datasets with distillation.

Constructing a fusion network backbone that facilitates cross-task sharing requires retaining prior knowledge while ensuring sufficient capacity for learning new tasks. Therefore, to further address the catastrophic forgetting that still resides in the knowledge-sharing component, we analyze the feature characteristics of the fusion models. Specifically, we perform a similarity analysis among the channels in Tardal[10] and PIAFusion[22], respectively. The similarity is measured with the FSIM[29], a metric for representing the low-level features of the image, such as color, texture, and shape. The channels from the same layer are compared to generate a two-dimensional similarity matrix. As illustrated in Fig. 3, the fusion network exhibits a high similarity among the channels. This indicates that multiple channels focus on similar features of the image and there exists some redundancy. Therefore, it is promising to solve the catastrophic forgetting by retaining some key channels to keep previous knowledge, meanwhile letting other channels learn from new tasks.

Based on our findings, we propose a two-phase training approach for VIFA. In the phase for training the knowledge-sharing component, we propose a key channel-constrained distillation, which selects the key channels from the intermediate feature maps and allows only minor changes when the component is trained for another task. The details for the key channel-constrained distillation are introduced in Section 3.2. Besides, we propose a reference model-guided distillation, which employs a pool of reference models to guide the training of the task-specific components, leading to highly compressed task-specific components. This architecture effectively mitigates the catastrophic forgetting problem and enhances performance in multi-task scenarios. The details for the reference model-guided distillation are introduced in Section 3.3.

3.2 Key Channel-constrained Distillation

Leveraging the characteristic that the fusion network exhibits a high similarity among the channels, we propose the key channel-constrained distillation, as

illustrated in Fig. 2. We cluster similar channels and the channels with the highest entropy in each cluster are identified as the key channels, as they extract unique features of input images. During the training on new tasks, the key channels are constrained such that only minor adjustments can be applied, thus effectively preserving previously learned knowledge. In addition, no constraints are applied to the non-critical channels, enabling them to adapt to new tasks.

After a fusion model has been trained on the first dataset as in step ① of Fig.2, we sample N images to generate intermediate feature maps, whose channel features are then analyzed. For instance, for the l-th layer, its feature map is represented as $F_l \in R^{N\times C\times H\times W}$, where C denotes the number of channels, and H and W are the height and width of the feature map. Subsequently, the similarities among the channels and the entropy of each channel are computed. The channels are clustered according to their similarities by a predefined threshold, and the channels with the highest entropy in each cluster are selected as the key channels. Consequently, an importance mask M_l for the key channels is generated. Detailed steps can be found in Step ②. This mask records the index of the channel to be retained in the l-th layer, and then the entropy corresponding to each channel is utilized as the importance factor. This procedure is repeated for all intermediate layers, and we can generate importance masks for each dataset and use them to guide the training for the next dataset.

After the masks have been generated, the previous backbone is used as the teacher model and also a starting point for the training on new datasets as shown in step ③. Specifically, utilizing the importance mask, the intermediate feature maps are distilled as defined in Eq. (1).

$$\mathcal{L}_{\text{fkd}} = \sum_{l=1}^{L} \sum_{c\in M_l^p} M_{l,c}^p \cdot \|\boldsymbol{f}_{l,c} - \boldsymbol{f}_{l,c}^p\|_2 \tag{1}$$

where $M_{l,c}^p$ denotes the importance factor of c-th channel in the l-th layer, $\boldsymbol{f}_{l,c}^p$ denotes the c-th channel feature map of the l-th layer of the previous model, $\boldsymbol{f}_{l,c}$ denotes the c-th channel feature map of the l-th layer of the newly trained model.

The Fig.3 illustrate the similarities among the channels for one intermediate layer after the model has been trained with the key-channel constrained distillation on two datasets. Compared to left figure of each network, we can observe a significant decline in the similarity among channels. This indicates that the proposed method enhances channel expressibility and facilitates the fusion network in learning new knowledge.

3.3 Reference Model-guided Distillation

The proposed parameter-isolated architecture can address the catastrophic forgetting issues and share the fusion backbone, thus improving the multi-task inference. Nonetheless, the task-specific components have a varied number of layers and the size could be large, diminishing the benefits of the architecture. We evaluated Tardal and PIAFusion as the fusion backbones, and each attached

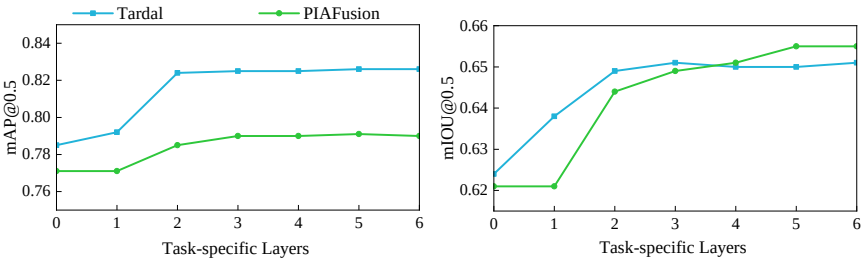

Fig. 4. Performance of Tardal and PIAFusion with YoloV5s (left) and the DeeplabV3+ (right) with different numbers of task-specific layers.

with two downstream tasks, the M3FD for object detection with YoloV5s[23] and MSRS for semantic segmentation with DeeplabV3+[3]. The models are trained until the downstream task performance converges. As shown in Fig.4, to reach the saturated performance, Tardal needs two task-specific layers for YoloV5s and three task-specific layers for DeeplabV3+, and PIAFusion needs three and five task-specific layers, respectively. This is because the fusion model is trained unsupervised, and the feedback from the downstream task model is weak, making it difficult for the downstream task to provide explicit image generation guidance to the fusion network. As a result, larger task-specific components are needed to generate fusion images for the downstream tasks.

We address this problem by converting the joint training process to a supervised learning process. Specifically, we generate ground truth labels for image fusion, thereby forcing the image generation of the fusion network to respond to the ground truth. As illustrated in stage two in Fig.2, for each downstream task, we train a reference model, which is obtained by jointly training the downstream model with the whole fusion model rather than merely the task-specific component. The reference model is trained as Eq. (2).

$$\min_{\mathbf{w}_\mathrm{t},\mathbf{w}_\mathrm{f}} \mathcal{L}^\mathrm{t}\big(\mathcal{T}(\hat{\boldsymbol{u}};\mathbf{w}_\mathrm{t})\big), s.t. \hat{\boldsymbol{u}} = \mathcal{F}(\mathbf{ir},\mathbf{vi};\mathbf{w}_\mathrm{f}). \tag{2}$$

where $\mathcal{L}^\mathrm{t}$ denotes current downstream task-oriented loss, $\mathcal{T}$ is the current downstream task network, $\hat{u}$ is fusion images generated by the reference model, $\mathcal{F}$ is the reference fusion network and $\mathcal{T}$ is the task network, and $\mathbf{w}$ denotes the paramters of the network. This strategy produces a dedicated fusion model for that downstream task and can generate high-quality fusion images. Then, the generated images are utilized to create an extended dataset, comprising paired source images ($\mathbf{ir}$, $\mathbf{vi}$) and their fusion images $\hat{u}$ as the ground truth label. Next, the dataset is utilized to train the corresponding task-specific component of VIFA. Specifically, the shared knowledge component obtained from the first training stage is frozen. Only the task-specific component is trained to separate the two knowledge components. To align the model output with the ground truth of the

Table 2. Fusion Performance Results

Model		M3FD dataset				MSRS dataset			
		EN	SF	VIF	SSIM	EN	SF	VIF	SSIM
PIAFusion	*Base.*	7.40	24.16	0.78	0.88	5.93	9.56	0.64	0.76
	Seq.	7.34	22.84	0.65	0.79	5.98	10.23	0.65	0.79
	VIFA	7.63	23.11	0.73	0.85	6.07	10.19	0.64	0.77
Tardal	*Base.*	6.43	18.10	0.83	0.86	5.37	13.36	0.69	0.83
	Seq.	6.41	19.68	0.67	0.83	5.47	11.97	0.65	0.83
	VIFA	6.53	20.90	0.77	0.78	5.95	10.45	0.68	0.85
DIDFuse	*Base.*	6.55	19.43	0.56	0.68	6.41	14.76	0.57	0.87
	Seq.	7.02	22.28	0.42	0.49	6.72	14.29	0.51	0.89
	VIFA	7.40	27.90	0.52	0.62	6.81	14.15	0.55	0.87
SwinFusion	*Base.*	7.38	18.25	0.68	0.92	6.37	14.65	0.72	0.88
	Seq.	7.25	19.32	0.57	0.81	6.39	14.86	0.74	0.86
	VIFA	7.32	19.52	0.67	0.88	6.41	14.82	0.74	0.85

reference model at the output layer, we introduce a distillation loss, as shown in Eq. (3), which facilitates the learning of the task-specific parameters.

$$\mathcal{L}_{\text{tkd}} = \|\boldsymbol{u} - \hat{\boldsymbol{u}}\|_2 \tag{3}$$

where u denotes the fusion images generated by the task-specific component, and $\hat{u}$ denotes the images generated by the reference model, which are used as the ground truth. We use ℓ_2-norm to constrain the output.

4 Evalution

4.1 Experimental Setup

Datasets We evaluated our VIFA on three representative benchmark datasets. The M3FD[10] and MSRS[22] datasets were employed to evaluate the impact of visible-infrared image fusion on the performance of downstream tasks, which are about streetscapes. The M3FD is for object detection and the MSRS is for semantic segmentation. Furthermore, to investigate the scalability of the architecture, we used the VEDAI[17]. The VEDAI is designed for detecting small targets in remote sensing images. The training, validation, and test sets were divided among the datasets involved in training in the ratio of 8:1:1.

Implementation details Four representative fusion networks were tested: PIAFusion[22], which is a CNN-based model; Tardal[10], which is a GAN-based model; DIDFuse[32], which is an AutoEncoder-based model; and SwinFusion[15], which is a Transformer-based model. We sampled 128 images to generate the importance masks, and the similarity threshold was set to 0.9. For downstream models, we used YoloV5s[23] for the object detection task and DeepLabV3+[3] for the semantic segmentation task. Both networks are widely utilized in edge computing deployments due to their exceptional performance and efficiency. The downstream

Table 3. Performance Evaluation of the Fusion Networks on Downstream Tasks. We focus on the average accuracy of our method (bolded) versus the sequential training method (underlined) on the two datasets.

Method		M3FD dataset							MSRS dataset								
		Peo	Car	Bus	Lam	Mot	Tru	mAP@0.5	Unl	Car	Per	Bik	Cur	CS	GD	CC	BU	mIOU
VI		0.701	0.894	0.921	0.741	0.668	0.736	0.777	0.972	0.792	0.677	0.616	0.431	0.477	0.58	0.433	0.515	0.61
IR		0.669	0.88	0.903	0.725	0.654	0.742	0.762	0.97	0.768	0.674	0.59	0.413	0.456	0.538	0.411	0.422	0.582
PIAFusion	*Base.*	0.687	0.881	0.901	0.737	0.671	0.75	0.771	0.974	0.838	0.646	0.647	0.372	0.536	0.549	0.535	0.49	0.621
	Ref.	0.725	0.904	0.947	0.7	0.675	0.81	0.794	0.976	0.849	0.661	0.68	0.434	0.559	0.679	0.557	0.51	0.656
	Seq.	0.638	0.866	0.874	0.575	0.661	0.722	0.723	0.976	0.852	0.66	0.675	0.444	0.562	0.7	0.563	0.494	0.658
	VIFA	0.722	0.912	0.946	0.714	0.667	0.783	**0.791**	0.976	0.842	0.653	0.668	0.431	0.552	0.672	0.563	0.547	**0.656**
Tardal	*Base.*	0.714	0.905	0.931	0.742	0.668	0.753	0.785	0.974	0.826	0.687	0.638	0.418	0.506	0.583	0.491	0.491	0.624
	Ref.	0.745	0.925	0.956	0.793	0.712	0.825	0.826	0.975	0.837	0.673	0.654	0.42	0.569	0.667	0.536	0.546	0.653
	Seq.	0.692	0.875	0.869	0.554	0.67	0.604	0.711	0.975	0.837	0.675	0.654	0.421	0.559	0.664	0.537	0.544	0.652
	VIFA	0.743	0.922	0.953	0.787	0.711	0.813	**0.822**	0.975	0.835	0.673	0.65	0.423	0.564	0.651	0.532	0.543	**0.65**
DIDFuse	*Base.*	0.709	0.901	0.922	0.741	0.672	0.738	0.781	0.97	0.808	0.586	0.608	0.281	0.496	0.613	0.503	0.438	0.589
	Ref.	0.721	0.907	0.926	0.793	0.656	0.787	0.798	0.975	0.846	0.647	0.657	0.418	0.561	0.682	0.528	0.452	0.641
	Seq.	0.668	0.885	0.887	0.732	0.602	0.712	0.748	0.974	0.845	0.643	0.654	0.421	0.565	0.69	0.535	0.45	0.642
	VIFA	0.718	0.903	0.919	0.793	0.661	0.781	**0.796**	0.974	0.842	0.649	0.654	0.435	0.546	0.667	0.541	0.424	**0.637**
SwinFusion	*Base.*	0.683	0.902	0.918	0.752	0.663	0.792	0.785	0.975	0.843	0.652	0.652	0.387	0.539	0.583	0.547	0.505	0.631
	Ref.	0.705	0.915	0.931	0.777	0.679	0.812	0.803	0.975	0.848	0.658	0.673	0.414	0.546	0.599	0.561	0.529	0.645
	Seq.	0.676	0.853	0.898	0.624	0.628	0.745	0.737	0.976	0.848	0.659	0.674	0.413	0.541	0.595	0.561	0.532	0.644
	VIFA	0.712	0.911	0.928	0.771	0.668	0.803	**0.799**	0.976	0.845	0.66	0.67	0.414	0.549	0.598	0.566	0.508	**0.643**
RecoNet[6]		0.717	0.872	0.965	0.753	0.614	0.805	0.788	0.972	0.825	0.608	0.598	0.35	0.555	0.538	0.431	0.521	0.6
CDDFuse[31]		0.69	0.905	0.917	0.764	0.671	0.803	0.792	0.975	0.845	0.64	0.647	0.439	0.491	0.665	0.521	0.521	0.638
BDLFusion		0.72	0.925	0.947	0.747	0.667	0.813	0.803	-	-	-	-	-	-	-	-	-	-
SegMif		-	-	-	-	-	-	-	0.976	0.846	0.65	0.67	0.412	0.536	0.664	0.562	0.483	0.644

models were connected with each of the fusion models. For each pair of fusion backbones and downstream models, we implemented four versions. *Base.* represents that the fusion backbone and the downstream models are separately trained and then connected. *Ref.* represents that the whole fusion backbone is fine-tuned by a single downstream model, which reaches the upper bound for downstream task performance. *Seq.* indicates that the whole fusion backbone is fine-tuned by the downstream models sequentially, which may cause catastrophic forgetting issues. Finally, VIFA represents our approach, that is, a fusion backbone is trained by multiple fusion datasets with key channel-constrained distillation, and the task-specific components are trained with reference model-guided distillation.

4.2 Performance of Image Fusion

To evaluate the performance of the fusion model obtained by stage 1, four metrics were employed. Entropy (EN) and structural fidelity(SF) were employed to measure the information richness of the fusion image, and visual information fidelity (VIF) and structural similarity index (SSIM) were employed to measure the similarity between the fusion image and the source images. The higher values of these metrics suggest higher quality of the fusion images. As shown in Table 2, the *Seq.* models on the M3FD dataset exhibit a slight decline in the VIF and SSIM metrics in comparison to *Ref.*, which suggests a small degree of catastrophic forgetting. VIFA recovers the similarity metrics, indicating that the knowledge acquired from the previous dataset was effectively retained during the training of the second dataset. Nevertheless, since no constraints were imposed on the output layer in this stage, the similarity remains distinct from that of *Ref.*

Table 4. Scalability Experimental Results.

	PIAFusion				Tardal			
	Base.	*Ref.*	*Seq.*	**VIFA**	*Base.*	*Ref.*	*Seq.*	**VIFA**
M3FD	0.771	0.794	0.559	**0.788**	0.785	0.826	0.535	**0.819**
MSRS	0.621	0.656	0.383	**0.653**	0.624	0.653	0.428	**0.648**
VEDAI	0.533	0.551	0.55	**0.548**	0.565	0.578	0.578	**0.577**

4.3 Performance of Vision Tasks

Table 3 shows the performance comparison on downstream tasks, specifically on object detection and segmentation. Compared to the *Base.* models, the *Ref.* models show a notable enhancement, demonstrating that training the fusion network with feedback from the downstream models can significantly improve the downstream task performance. For the *Base.* models, we first trained them with the M3FD dataset for object detection, and then we trained them with the MSRS dataset for segmentation without any constraints. The results show a decline of approximately 0.05-0.07 in the mAP metric, indicating that the fusion networks are overfitting to the latest task and exhibit drastic catastrophic forgetting.

In comparison, for all four sets of networks, our method can simultaneously support object detection and segmentation tasks, achieving an accuracy close to the Ref. models, which are the performance upper bound. Furthermore, we compare with other state-of-the-art fusion networks, such as BDLfusion[14], which focuses on object detection, and SegMif[11], which is specifically designed for semantic segmentation. The results show that our method is comparable with these SOTA methods, and more importantly, our method can simultaneously support multiple downstream tasks.

4.4 Evaluation on Scalability

Table 4 presents the findings of supplementary scalability experiments conducted on the VEDAI dataset. We perform stage 1 for the VEDAI dataset based on the existing shared backbone and then perform stage 2 to add the task component of VEDAI. In particular, the *Base.* result, after being trained on the VEDAI dataset with the guidance of the object detection task, exhibits a severe catastrophic forgetting phenomenon on the previous tasks. This phenomenon is primarily attributable to the fact that the VEDAI dataset is specifically designed for the task of detecting small targets in remote sensing images, which is significantly distinct from the nature of the tasks in the previous two datasets. Consequently, there are significant task conflicts. In contrast, our proposed method effectively isolates the effects of the downstream tasks, resulting in only a slight performance degradation on the previous tasks. The experimental results demonstrate the good scalability of the proposed approach.

Table 5. Comparative Analysis of Deployment Overhead Between VIFA and Ref.

Model	*Ref.*			**VIFA**			
	Size(MB)	Flops(G)	Time(ms)	Size(MB)	Flops(G)	Time(ms)	Speedup
PIAfusion	8.98	573.78	1677.38	4.56	291.73	866.03	**1.9x**
Tardal	2.26	144.86	526.02	1.202	77.34	283.89	**1.85x**
DIDFuse	1.98	182.56	2594.54	1.27	110.58	1403.66	**1.85x**
SwinFusion	7.2	474.84	19255.92	3.616	238.52	9711.77	**1.98x**

Table 6. Task-specific Component Overhead Proportions in the Entire Network

Model	Size(Perecent)	Flops(Perecent)	Time(Perecent)
PIAfusion	1.56%	1.68%	3.26%
Tardal	6.3%	6.78%	7.94%
DIDFuse	28.9%	21.1%	8.2%
SwinFusion	0.45%	0.46%	0.87%

4.5 Efficiency Analysis

We deployed our proposed VIFA on embedded devices and tested the inference efficiency with Nvidia Jetson Nano[16] in a scenario involving two downstream tasks. In Table 5, we compare the memory overhead, FLOPS, and inference latency. Since our VIFA shares the fusion backbone and compresses the task-specific components to only one layer, the memory overhead and FLOPS have been significantly reduced, and the inference latency has improved by 1.85x-1.98x. Besides, we also break down the results and show how the task-specific components contribute to the inference overhead in Table 6. The task-specific components account for only 0.45% to 28.9%of the total memory overhead, 0.46% to 21.1% of the total FLOPS, and 0.87% to 8.2% of the total inference time. Please note that PIAFusion has two layers in the task-specific component, as the output layer of PIAFusion has only a few hundred parameters. The results demonstrate that the task-specific layer imposes a negligible overhead while significantly enhancing task performance.

4.6 Ablation Study

We verified the effectiveness of each strategy in Table 7. The fusion networks exhibit a significant decline in performance when evaluated on the M3FD dataset in the absence of any strategy. This phenomenon, known as catastrophic forgetting, has been observed in previous studies. The catastrophic forgetting of the fusion models is mitigated when only key channel distillation is employed, without the introduction of downstream task feedback to adjust the fusion model. This finding corroborates the hypothesis that task-specific knowledge is prone to conflict, as previously discussed. However, the lack of guidance from downstream tasks results in the fusion network failing to optimize its performance on relevant tasks. When only the reference model result distillation is employed, the learning for the new

Table 7. Results of Ablation Experiments.

Configurations	Tardal		PIAFusion	
	M3FD	MSRS	M3FD	MSRS
Seq.	0.711	0.652	0.723	0.658
key channel-constrained distillation	0.783	0.624	0.769	0.623
reference model-guided distillation	0.813	0.653	0.773	0.656
Both	0.822	0.65	0.791	0.656

dataset reaches its upper limit of performance. However, the lack of constraints on the backbone of the fusion network allows the backbone to change considerably, resulting in the forgetting of previously learned knowledge. This leads to a decrease in performance on the M3FD dataset compared to the reference model. The combination of both approaches allows for the efficient isolation of potential conflicts between tasks, while ensuring that the shared backbone network provides robust feature extraction for multiple downstream tasks. This results in model performance that is near the upper limit on both the old and new datasets. In conclusion, our approach is based on a stepwise optimization of intermediate results, with each step contributing to the improvement of the final fusion results.

5 conclusion

In this work, we proposed a scalable multi-task architecture for visible-infrared image fusion and downstream tasks. We analyzed the catastrophic forgetting issues and feature clustering characteristics in the fusion model, and based on that, we partitioned the fusion network into a knowledge-sharing fusion backbone and task-specific components. In the knowledge-sharing component, we proposed a key channel-constrained distillation method to retain the learned knowledge in the key channels, thus improving the generalizability of the fusion backbone. Besides, we proposed a reference model-guided distillation method to compress the task-specific network to one layer, thus improving the multi-task inference efficiency. We conducted experiments on multiple representative fusion networks that covered the main types of architectures, including CNN, GAN, AutoEncoder, and Transformer. Experimental results show that our approach is capable of supporting multiple downstream tasks with a shared fusion backbone and achieving 1.85x-1.98x speedups for inference on Jetson Nano.

References

1. Ariffin, S.M.Z.S.Z., Jamil, N., Rahman, P.N.M.A.: Can thermal and visible image fusion improves ear recognition? In: 2017 8th International Conference on Information Technology (ICIT). pp. 780–784. IEEE (2017)
2. Cha, H., Lee, J., Shin, J.: Co2l: Contrastive continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9516–9525 (October 2021)

3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision (2018), https://api.semanticscholar.org/CorpusID:3638670
4. Doshi, K., Yilmaz, Y.: Continual learning for anomaly detection in surveillance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2020)
5. Douillard, A., Chen, Y., Dapogny, A., Cord, M.: Plop: Learning without forgetting for continual semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4040–4050 (June 2021)
6. Huang, Z., Liu, J., Fan, X., Liu, R., Zhong, W., Luo, Z.: Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In: European conference on computer Vision. pp. 539–555. Springer (2022)
7. Kang, M., Park, J., Han, B.: Class-incremental learning by knowledge distillation with adaptive feature consolidation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16071–16080 (2022)
8. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017). https://doi.org/10.1073/pnas.1611835114, https://www.pnas.org/doi/abs/10.1073/pnas.1611835114
9. Lin, H., Zhang, B., Feng, S., Li, X., Ye, Y.: Pcr: Proxy-based contrastive replay for online class-incremental continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24246–24255 (June 2023)
10. 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. pp. 5802–5811 (2022)
11. Liu, J., Liu, Z., Wu, G., Ma, L., Liu, R., Zhong, W., Luo, Z., Fan, X.: Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8115–8124 (2023)
12. Liu, Y., Schiele, B., Vedaldi, A., Rupprecht, C.: Continual detection transformer for incremental object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23799–23808 (June 2023)
13. Liu, Y., Schiele, B., Vedaldi, A., Rupprecht, C.: Continual detection transformer for incremental object detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23799–23808 (2023). https://doi.org/10.1109/CVPR52729.2023.02279
14. Liu, Z., Liu, J., Wu, G., Ma, L., Fan, X., Liu, R.: Bi-level dynamic learning for jointly multi-modality image fusion and beyond. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. IJCAI '23 (2023). https://doi.org/10.24963/ijcai.2023/138, https://doi.org/10.24963/ijcai.2023/138
15. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y.: Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica **9**(7), 1200–1217 (2022)
16. NVIDIA Corporation: NVIDIA Jetson Nano Developer Kit. NVIDIA, Santa Clara, CA, USA (2021), https://developer.nvidia.com/embedded/jetson-nano-developer-kit

17. Razakarivony, S., Jurie, F.: Vehicle detection in aerial imagery : A small target detection benchmark. Journal of Visual Communication and Image Representation **34**, 187–203 (2016). https://doi.org/https://doi.org/10.1016/j.jvcir.2015.11.002, https://www.sciencedirect.com/science/article/pii/S1047320315002187
18. Smith, J.S., Tian, J., Halbe, S., Hsu, Y.C., Kira, Z.: A closer look at rehearsal-free continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2409–2419 (2023)
19. Sun, H., Liu, Q., Wang, J., Ren, J., Wu, Y., Zhao, H., Li, H.: Fusion of infrared and visible images for remote detection of low-altitude slow-speed small targets. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 2971–2983 (2021)
20. Sun, Y., Cao, B., Zhu, P., Hu, Q.: Detfusion: A detection-driven infrared and visible image fusion network. In: Proceedings of the 30th ACM international conference on multimedia. pp. 4003–4011 (2022)
21. 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. Information Fusion **82**, 28–42 (2022)
22. Tang, L., Yuan, J., Zhang, H., Jiang, X., Ma, J.: Piafusion: A progressive infrared and visible image fusion network based on illumination aware. Information Fusion **83**, 79–92 (2022)
23. Ultralytics: ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation. https://github.com/ultralytics/yolov5.com (2022). https://doi.org/10.5281/zenodo.7347926, https://doi.org/10.5281/zenodo.7347926, accessed: 7th May, 2023
24. Xu, H., Ma, J., Le, Z., Jiang, J., Guo, X.: Fusiondn: A unified densely connected network for image fusion. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12484–12491 (2020)
25. Yan, Y., Ren, J., Zhao, H., Sun, G., Wang, Z., Zheng, J., Marshall, S., Soraghan, J.: Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos. Cognitive Computation **10**, 94–104 (2018)
26. Yang, Z., Li, R., Ling, E., Zhang, C., Wang, Y., Huang, D., Ma, K.T., Hur, M., Lin, G.: Label-guided knowledge distillation for continual semantic segmentation on 2d images and 3d point clouds. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18555–18566 (2023). https://doi.org/10.1109/ICCV51070.2023.01705
27. Ye, F., Bors, A.G.: Self-evolved dynamic expansion model for task-free continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22102–22112 (October 2023)
28. Zhai, M., Chen, L., Tung, F., He, J., Nawhal, M., Mori, G.: Lifelong gan: Continual learning for conditional image generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2759–2768 (2019)
29. Zhang, L., Zhang, L., Mou, X., Zhang, D.: Fsim: A feature similarity index for image quality assessment. IEEE Transactions on Image Processing **20**(8), 2378–2386 (2011). https://doi.org/10.1109/TIP.2011.2109730
30. Zhang, X., Ye, P., Qiao, D., Zhao, J., Peng, S., Xiao, G.: Object fusion tracking based on visible and infrared images using fully convolutional siamese networks. In: 2019 22th International Conference on information fusion (FUSION). pp. 1–8. IEEE (2019)
31. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Van Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality

image fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5906–5916 (2023)
32. Zhao, Z., Xu, S., Zhang, C., Liu, J., Li, P., Zhang, J.: Didfuse: Deep image decomposition for infrared and visible image fusion. arXiv preprint arXiv:2003.09210 (2020)
33. Zhou, M., Xiao, J., Chang, Y., Fu, X., Liu, A., Pan, J., Zha, Z.J.: Image de-raining via continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4907–4916 (2021)
34. Zhu, L., Chen, T., Yin, J., See, S., Liu, J.: Continual semantic segmentation with automatic memory sample selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3082–3092 (June 2023)