The Visual Computer
https://doi.org/10.1007/s00371-023-02857-3

ORIGINAL ARTICLE

Multitask learning for image translation and salient object detection from multimodal remote sensing images

Yuanfeng Lian[1] · Xu Shi[1] · ShaoChen Shen[1] · Jing Hua[2]

Accepted: 9 March 2023

Abstract

This paper presents a novel and efficient multitask learning framework for image translation and saliency detection from remote sensing images, which mainly contains the image translation network-weight sharing attention GAN (WSA-GAN) and the salient object detection network-boundary guidance network (BGNet). WSA-GAN can be used to generate a large number of synthetic infrared remote sensing images (IRIs) or optical remote sensing images (ORIs) from the corresponding complementary modality images. Then, a new multimodal context-aware learning is proposed for feature extraction and to coordinate the entanglement of latent features in the multimodal context of ORIs and IRIs. Since convolutional neural networks do not perform well when the object has directional variance, our framework introduces the attention-aware CapsNet (AACNet) to alleviate the problem and enhance the feature expressiveness. In addition, knowledge distillation strategy is introduced in AACNet to reduce the model complexity. Finally, the multiscale feature learning network and the boundary-aware block are designed to generate more accurate saliency detection results with clear boundaries. Experimental results demonstrate that the presented image translation and salient object detection networks outperform other approaches.

Keywords Multitask learning · Image translation · Salient object detection · Remote sensing image · Context-aware learning

1 Introduction

Salient object detection is of great importance in resources exploration, environmental monitoring, and sea navigation. The fundamental challenges in salient object detection from optical remote sensing images (ORIs) lie in that ORIs span a large area with complex background and various noise interference. In addition, ORIs are vulnerable to influence from weather conditions such as wave disturbance, daylight and cloudy conditions. Compared to ORIs for salient object detection, IRIs provide better clues to search for salient objects but suffer from fuzzy edges. As a result, the salient object detection technology integrating both ORIs and IRIs can extract complement feature information and improve the accuracy of salient object detection. Unfortunately, the availability of IRIs is extremely limited. Currently, there are few public IRIs for research exploration. In contrast, ORIs are readily available and have the potential to augment IRIs if translation from ORIs to IRIs is viable. However, current image-to-image translation techniques are inefficient and lead to poor quality of generated images due to the imbalanced information from the two modality domains. The task of salient object detection in ORIs/IRIs remains very challenging.

In response to the aforementioned issues, we propose a multitask learning framework, based on novel image-to-image translation and salient object detection algorithms, to enhance the feature representation capabilities for better saliency analysis. More specifically, to address the issue of the limited availability of IRIs, we design the weight-sharing attention GAN (WSA-GAN) that can convert ORIs to IRIs (and vice versa). Thus, multimodal context information and latent feature can be generated even from a single input modality (i.e., either ORIs or IRIs). A novel feature extraction network, namely multimodal context-aware learning (MCL), is then used to extract and fuse the features of those two modalities. To further explore the features' representa-

✉ Yuanfeng Lian
lianyuanfeng@cup.edu.cn

Xu Shi
2020215940@student.cup.edu.cn

[1] Beijing Key Lab of Petroleum Data Mining, Department of Computer Science and Technology, China University of Petroleum, Beijing, China

[2] Wayne State University, Detroit, MI, USA

Published online: 04 May 2023

tion capabilities, we design a new salient object detection network, in which the attention-aware CapsNet (AACNet) offers a dynamic sigmoid routing and a EM routing with the attention-aware strategy for developing higher-level capsules to capture complex characteristics in salient objects, and finally, the boundary guidance network (BGNet) is used to separate salient items from the background. Our proposed network structure is shown in Fig. 1

Our main *contributions* can be summarized as follows:

- We propose a novel WSA-GAN to achieve high-quality image translations between ORIs and IRIs, in which the multimodal feature share block (MFSB) can effectively exchange the information of different modalities and multiadditive attention block (MAB) can further improve the feature reusability. In addition, we introduce the regulation terms and texture loss to construct the WSA-GAN loss function.
- A novel representation learning framework, MCL, is proposed, which learns global and local contextual information through three different context-aware blocks from the input/translated multimodal remote sensing images. To better entangle the latent feature vectors of the two modalities, we further introduce the inter-modal feature mapping loss in MCL.
- Due to the capabilities of the capsule network in describing the shape size, direction and deformation information, we design an improved CapsNet structure for further feature enhancement. We present AACNet with the feature fusion attention-aware module (FFAM). AACNet is able to narrow the search space when the route from low-level capsule to high-level capsule occurs, which indicates AACNet pays more attention to the feature information with high correlation. This significantly retains the intrinsic feature information while reducing the possibility of noise allocation for boundary guided saliency detection.
- Our method outperforms the state-of-the-art (SOTA) methods in the experiments on image translate and saliency detection. The ablation study shows there is great improvement when integrating MAB in WSA-GAN, attention-aware strategy and knowledge distillation strategy in AACNet.

2 Related work

Image translation, salient object detection and context-aware learning are very important research topics in computer vision. We review the classic methods below. Interested readers can refer to the survey articles [1–3] to further study other methods.

2.1 Image-to-image translation

Image-to-image translation techniques aim at learning a function to map an input image to the desired output image, which can be divided into three categories: image methods [4, 5], GAN approaches [6–12] and separating spaces methods [13–15]. Recent work mainly focused on mapping the source image into a common latent feature space through convolutions and decoding those latent features to target domain image by using transposed convolutions. Compared with traditional methods [4, 5], GAN-based approaches have achieved promising results which can be classified into two types of transfer learning: supervised learning [6, 7, 12] and unsupervised learning [8–11]. Isola et al. [6] proposed a pix2pix algorithm model for image-to-image conversion based on cGAN [16]. Wang et al. [7] proposed a method to generate high-resolution images using pix2pix. Pix2pix algorithm combines the antagonism loss and L1 loss between the source image and the target image. Therefore, the models input paired data sets and belongs to supervised learning. Yoo et al. [12] employed an additional discriminator in the original GAN to judge whether the image pairs from different domains are related, which can constrain the consistency between the generated image and the ground truth image.

However, the collection of paired data samples is extremely difficult, so the unsupervised learning image-to-image translate algorithm has received more attention. Zhu et al. [9] proposed CycleGAN, the first unsupervised image-to-image conversion model, which ensures that the generate image can retain the structure and content of the real source domain image to the greatest extent. Some recent work [8, 11] also used the same principle and modify the loss function to enhance the robustness of the system. More recently, Li and Tuzel introduced CoupleGAN (Cogan) [17], which learns the joint distribution of the two domains in the potential space to achieve unpaired image translation. Furthermore, Liu et al. [10] proposed an unsupervised image-to-image translation network (UNIT) based on the assumption of shared potential space. In addition, several image-to-image translation algorithms presume that the latent space of pictures may be split into a content space and a style space. Huang et al. [14] proposed MUNIT, a multimodal unsupervised image-to-image translation framework with two latent representations for style and content. Similarly, Lee et al. [13] proposed diverse image-to-image translation (DRIT), which is based on disentangled representation on unpaired data. DRIT divides the latent space into a domain-invariant content space and a domain-specific attribute space. PSC-GAN [15] converts the abstract representation of images into code representation through visual content disentanglement module (VCDM). However, the above image translation methods perform the image translation without the consideration of their interrelationship. Therefore, these methods can not semantically align and unify the features in latent space for downstream image analyses. Inspired by [6], we design the generator by using a share weight strategy which can take full advantage of multimodal information from the pairs of the remote sensing images. The network architecture and design are totally different from the existing WSA-GANs [18–20] in the literature. In our WSA-GAN, we introduced a MFSB module as a dynamic feature interaction mechanism, which can balance the gradients of the two network branches with the inter-modal feature dependencies to ensure the joint learning feature representation for the task of image transformation.

2.2 Context-aware learning

In image processing tasks, various approaches have been presented to incorporate the context-aware information. In some recent works [21–24], context cues were potentially incorporated into feature maps, such as deeper global context information and shallower local context features. Li et al. [21] proposed the multilayer feature context encoding network for the remote sensing image scene classification by using the multiscale spatial context information contained in the multilayer features. Wang et al. [22] extracted global and local structures from hyperspectral images for scene classification by incorporating contextual information into the classifier, because adjacent pixels are highly likely to belong to the same class. The extraction of useful context information is also critical to the salient object detection task. In [23], the relationship between regional context and dominant objects was adopted to detect salient objects from their surrounding context to guide advanced tasks. The work [24] presented a residual refinement network that first recovers high-level semantic context information, then strengthens features at all scales.

However, the existing saliency models rarely study the multimodal context-aware information of remote sensing images (e.g., in ORIs and IRIs) and their relational importance. To bridge this gap, we propose the multimodal context-aware learning for ORIs-IRIs image pairs to extract their context-aware relational information.

2.3 Salient object detection

Salient object detection (SOD) is a method that simulates human visual perception to locate the most important target in the scene. Traditional SOD approaches based on hand-crafted features are divided into two groups, including bottom-up methods [25, 26] and top-down methods [27, 28]. Readers can gain a comprehensive understanding of these methods from [2]. However, when capturing high-level semantics in complex scenes, these methods easily become ineffective.

On the other hand, CNN-based models [29–34] can be trained end-to-end using pixel-wise annotated saliency maps, which have broken the performance bottlenecks. In [29], an edge guidance network (EGNet) was proposed to simulate the two types of complementary information between salient edges and objects through a single network. By introducing short connections to the skip-layer structures within the holistically nested edge detector (HED) architecture, Hou et al. [30] introduced a succession of short connections between shallower and deeper side-output layers.

Compared with common scenes, Li et al. [31] proposed the salient object detection method based on deep learning for remote sensing images. Similarity, Zhang et al. [32] proposed a saliency adaptive multifeature fusion model based on low-rank matrix recovery to detect remote sensing images by integrating color, intensity, texture and other clues. Moreover, Hu et al. [33] proved that the use of deep contextual information effectively improves the accuracy of salient object detection. In [34], closure guided attention network (CGAN) and the coarse significance network (CSN) jointly supervise the feature channels to eliminate simplicity bias.

The above CNNs methods intend to extract the perceptual contexts for salient object detection. However, they ignore the intrinsic features of objects, such as scale and orientation, which often leads to incomplete segmentation of the saliency detection. To address this problem, the capsule networks use vectorized capsule neurons to encode feature information, and use weight matrices and dynamic routing algorithms to convey spatial relationships between feature objects with a higher level of abstraction modeling capability. Yu et al. [35] proposed a convolutional capsule network for detecting vehicles from high-resolution remote sensing images. Later, they [36] presented a sparse anchoring guided high-resolution capsule network (SAHR-CapsNet) for geospatial object detection. In [37], a multilevel CapsNet framework was proposed to achieve efficient military target recognition with the small training dataset.

In order to explore the routing mechanism of the capsule network and reduce the computational complexity of the capsule network. Sabour et al. [38] suggested a dynamic routing method to learn the intrinsic spatial distribution between the portion and the whole. Later in [39], Hinton et al. consolidated their findings by presenting the matrix CapsNet, which featured a pose matrix and an activation probability for each capsule. Based on [38, 39], some recent works [40–42] continued to innovate the routing mechanism. Feng et al. [40] designed a novel dual-routing mechanism to filter low-discriminative capsules which consists of inter-video and intra-video. Zhang et al. [41] implemented a dual flow strategy, called two-stream part-object relational network (TSPORTNet), to reduce network complexity and possible redundancy during capsule routing. Moreover, Mazzia et al. [42] introduced a non-iterative parallel routing algorithm to replace dynamic routing, which effectively reduces the number of capsules between subsequent layers by using the self-attention mechanism.

The key difference between the methods mentioned above and our proposed approach lies in the fact that our method is using multimodal context-aware learning to make full use of optical, infrared, and more importantly, their relational information. The detection method BGNet combines the AACNet and utilizes context-aware information by MCL. The network architecture and design are totally different from the existing BGNets [43, 44] in the literature. In BGNet, we employed multiscale dilated convolution with capsule salient map to fuse feature information at different scales and enhance the directional feature representation capabilities for the edge of objects in remote sensing images. The latent space GAN transforms the information from the descriptor space to the latent feature space and, then, use the feature fusion attention-aware module (FFAM) to fuse the two-way routing to explore and unify the multimodal relational representations.

3 Methodology

In this work, we propose WSA-GAN to learn the conversion from ORIs to IRIs (and vice versa) and BGNet for salient object detection from ORIs and IRIs.

3.1 Overview

The entire end-to-end deep neural network framework consists of five parts, including WSA-GAN, MCL, latent space GAN, AACNet, and BGNet, as shown in Fig. 1. During image translation through WSA-GAN, we use multiadditive attention blocks (MABs) to filter features and multimodal feature share blocks (MFSBs) to share the inter-modal features. The schematic diagram of MABs and MFSBs is shown in Fig. 2. The network consists of compressed pathways for extracting localized features and extended paths for resampling the image mapping along with contextual information. To encourage more semantically relevant output, skip connections are utilized to mix high-resolution local characteristics with low-resolution global features. In feature learning with MCL, the input and translated ORIs-IRIs pairs are used as the input, and the structure of two MCLs is shown in Fig. 3. MCL contains three different context-aware blocks to capture contextual information from varying receptive fields. The output of two MCLs, which are unified in the latent space GAN, is used to construct the primary capsule in AACNet. The primary capsule is input to feature fusion attention module (FFAM) with an attention-aware strategy to build higher-level capsules. The schematic diagram of reciprocal attention module (RAM) and dual attention module (DAM) in FFAM is shown in Fig. 5. At the final stage, we design multiscale feature learning network (MFLN) to generate the multiscale aggregated feature and boundary-aware block to produce the edge feature, as illustrated in Fig. 1. To generate the saliency detection result which are close to the ground truth, we combine the multiscale aggregated feature, the edge feature and capsule salient map to derive the final saliency. In the following, we introduce the novel technical components of the proposed method: WSA-GAN, MCL, latent space GAN, AACNet, and BGNet.

3.2 Weight-sharing attention GAN

We build an image translation network, weight sharing attention GAN (WSA-GAN), to handle a relatively small number of ORIs-IRIs image pairs. Due to the fact that there exists correlations among scene contents between the tasks of generating paired ORI and IRI, the MFSB is introduced into WSA-GAN to extract features from the two shared weight generators to improve the accuracy of the output images. Sharing weights in our networks between two tasks, when learning for two related sets of data, increases network robustness and learning efficiency since generators can fully utilize context-aware feature vectors by sharing weights across the two generators. Both encoders are constrained to encode the same semantic items in different styles, and the method learns to use the similar encoding to represent two visually distinct domains by sharing the multimodal feature information (i.e., ORIs and IRIs domains).

Multimodal feature share block As a result, we construct the multimodal feature share block (MFSB) as depicted in Fig. 2, which consists of two feature branches: infrared branch f_{IRI} and optical branch f_{ORI}. Most of the existing cross modal feature fusion methods are designed on the basis of addition or concatenation operation, which often causes redundant information and complex structure. Inspired by the attention mechanism, this work uses element-wise multiplication to build MFSB which fuse the infrared feature f_{IRI} and optical feature f_{ORI}. Through the multiple feature fusion, infrared feature f_{IRI} and optical feature f_{ORI} will gradually absorb each other's useful information and reduce redundant information. Considering the noise of shallow features, this paper uses MFSB to cross fuse optical and infrared features in the deeper level of the generator.

Multiadditive attention block In the deep stage of coding, the network will extract features with rich semantic information. However, due to cascading convolution and nonlinearity, the loss of spatial details becomes severe, which will lead to poor final image conversion results. To solve this problem, we designed MAB which is incorporated into the basic U-Net architecture in order to highlight important aspects via skip connections. A gating signal g is used for each pixel to determine the focal region. First, we deployed convolution layer-batch normalization structure to treat the

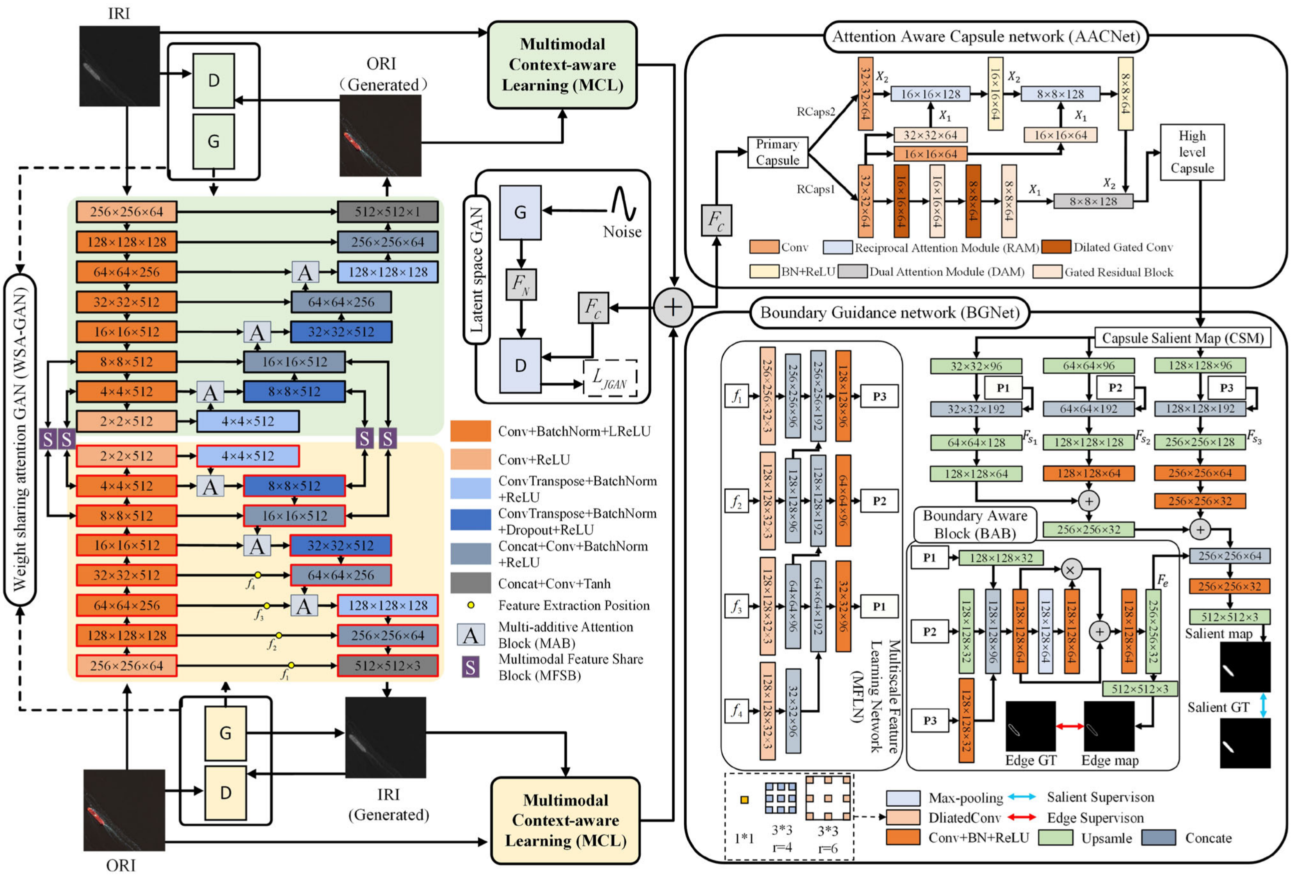

Fig. 1 The overall architecture of our multitask learning framework. The first stage is weight-sharing attention GAN (WSA-GAN), where ORIs-IRIs image pairs are input to train the network to have the ability to generate corresponding IRIs from ORIs (and vice versa). Multiadditive attention blocks (MABs) filter the features propagated by skipping connections. The features (f_1, f_2, f_3, f_4) extracted by the generator in the first four layers of the encoding part are used as input to the multiscale feature learning block (MFLN). MFLN outputs multiscale aggregated features ($P1$, $P2$, $P3$) to add accuracy for subsequent saliency detection. The second stage is multimodal context-aware learning (MCL), where we perform multimodal feature extraction and fusion on the ORIs-IRIs pairs generated in the first stage, and input these features into the attention-aware CapsNet (AACNet) in the third stage. Before the third stage, our Latent Space GAN can generate a joint potential space vector F_N through training, which includes optical and infrared modalities. F_c represents the features of MCL output. Next, AACNet processes the features F_c, and the primary capsules are transited to higher-level capsules by feature fusion attention-aware module (FFAM). Higher-level capsules and ($P1$, $P2$, $P3$) are used as inputs to the last phase of the boundary guidance network (BGNet). In order to better generate saliency detection results, we developed boundary-aware block (BAB), which can process ($P1$, $P2$, $P3$) to the edge features (F_e). Finally, BGNet will combine F_e, ($P1$, $P2$, $P3$) and capsule salient map to generate the final salient map

gating signal (g) and input features (x). Referring to the self-attention mechanism, we use max-pooling layer and element-wise multiplication operation to enhance the gating signal. Inspired by the spatial attention mechanism, we fuse the gating signal branch and the input feature branch using the addition operation and the result is further processed by the LeakyReLU activation-convolution layer-batch normalization-Sigmoid activation structure. The decoding layer output ($\widetilde{x}$) is the construct from gating signal (g) plus the element-wise product of attention coefficients (η) and input features (x).

In the GAN expanded stage, these trimmed features are concatenated with resampled output maps at a certain scale. Finally, the generator is trained to generate the ORIs or IRIs with a size of 512×512. The discriminator is used to determine whether or not the input picture is real or generated. The discriminator, given an input image, uses a series of convolution layer-batch normalization-LeakyReLU activation combination to extract the image's multiscale features. Finally, the discriminator outputs the result using the fully connected layer and the softmax activation function. The output result of the discriminator is the estimated probability of judging that the input image is a real image. As illustrated in Fig. 1, there are two sets of discriminators and generators introduced here, corresponding to IRIs-ORIs and ORIs-IRIs.

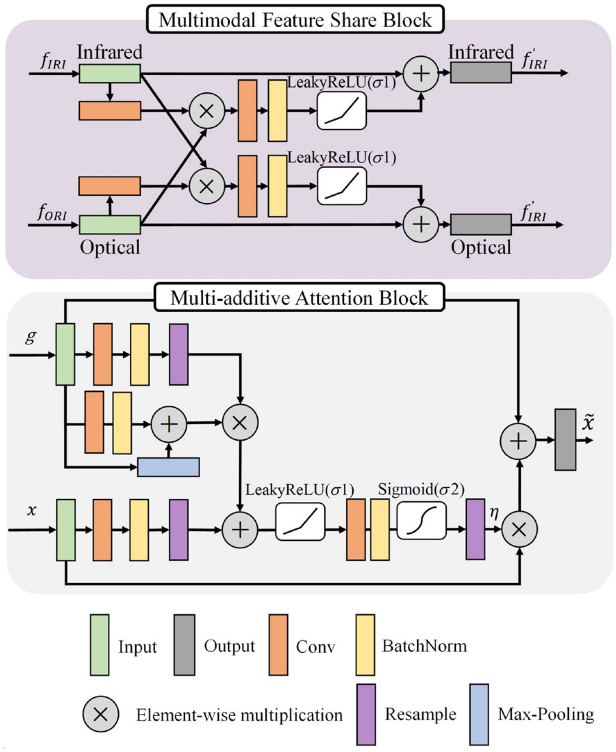

Fig. 2 Diagram of the multimodal feature share block (MFSB) and multiadditive attention block (MAB). MFSB: First, the channels numbers of f_{IRI} and f_{ORI} are compressed by a convolution layer, which is convenient for the subsequent element-wise multiplication to fuse and transform the features, and then, recover the channels numbers of fused features by the convolution layer, batch normalization and LeakyReLU activation. Finally, we add the fused feature to the original feature to realize the sharing of multimodal feature information. MAB: To transfer important information to the decoding layer output ($\tilde{x}$), the input features (x) are scaled with attention coefficients (η). Contextual information is provided by the coarser gating signal (g). Trilinear interpolation is used to grid resample attention coefficients

WSA-GAN Loss Although the learning process is automatic, we still need to carefully design the objective function to determine the goal of optimization. In WSA-GAN, the objective function is a combination of the conditional GAN loss and the regularization term and texture loss. The conditional GAN loss can be formatted as follows:

$$\begin{aligned} L(G, D) = & E_{x_{ORI}, y_{ORI}}[\log D(x_{ORI}, y_{ORI})] + \\ & E_{x_{ORI}}[\log(1 - D(x_{ORI}, G(x_{ORI})))] + \\ & E_{x_{IRI}, y_{IRI}}[\log D(x_{IRI}, y_{IRI})] + \\ & E_{x_{IRI}}[\log(1 - D(x_{IRI}, G(x_{IRI})))], \end{aligned} \tag{1}$$

where x_{ORI} and x_{IRI} represent the source images of ORIs and IRIs, y_{ORI} and y_{IRI} are the corresponding target images. G is the generator, D is the discriminator, and E represents the expectation of the discrimination outcomes. The source images x_{ORI} and x_{IRI} are used as the condition term entered in the discriminator. The discriminator's goal is to maximize the expected value, while the generator tries to minimize the expected value, $G^* = arg\min_{G}\max_{D} L(G, D)$.

Regularization Terms In order for the image created by the generator to be close to the ground truth image in the target domain, the conditional GAN loss must be minimized. To encourage the generated and target images to have similar styles, the regularization terms are mixed as GAN losses, which can be defined as follows:

$$\begin{aligned} L_{\mathrm{r}}(G) = & E_{x_{\mathrm{ORI}}, y_{\mathrm{ORI}}}[\| y_{\mathrm{ORI}} - G(x_{\mathrm{ORI}}) \|_1] + \\ & E_{x_{\mathrm{IRI}}, y_{\mathrm{IRI}}}[\| y_{\mathrm{IRI}} - G(x_{\mathrm{IRI}}) \|_1], \end{aligned} \tag{2}$$

where $\|\cdot\|_1$ denotes the L1 distance between the generated images and the target images. Compared with the L2 distance, the L1 distance fosters sparsity and less blurring than the L2 distance.

Texture Loss To constrain the internal drawing task and ensure fine-grained textures, we employ a local binary pattern (LBP)-based [45] loss function to minimize the difference between the generated textures and the ground truth texture, the equation as follows:

$$\begin{aligned} L_t = & \| LBP(\mathrm{Gray}(G(x_{\mathrm{ORI}}))) - LBP(\mathrm{Gray}(y_{\mathrm{ORI}})) \|_1 + \\ & \| LBP(\mathrm{Gray}(G(x_{\mathrm{IRI}}))) - LBP(\mathrm{Gray}(y_{\mathrm{IRI}})) \|_1 \, , \end{aligned} \tag{3}$$

where $Gray(\cdot)$ is a function that converts a color image into a grayscale image and $LBP(\cdot)$ is a differentiable LBP layer that acquires a grayscale image and outputs a LBP image.

Multitask loss weight optimization A common way to simplify multitask optimization is to balance different loss functions. The final objective function is a weighted combination of conditional GAN loss, regularization terms and texture loss.

$$G^* = arg\min_{G}\max_{D} L(G, D) + \lambda_1 L_r(G) + \lambda_2 L_t, \tag{4}$$

where λ_1 is the regularization coefficient and λ_2 is the texture coefficient, which are calculated by Eq. (5). With the regularization constraint and texture constraint on the conditional GAN, the generated image can not only deceive the discriminator, but also have the same intensity, texture and structure as the ground truth image. The weight by the relationship for the loss weighting and task schedule is assigned as:

$$\lambda_i = 1 + \left(\mathrm{softsign}(\bar{P} - P_{\mathrm{i}})) \min(\theta, (\max_{j} P_{\mathrm{j}})^{\gamma} \left| \bar{P} - P_{\mathrm{i}} \right|^{\delta} \right), \tag{5}$$

where P_{i} is the ratio of the current validation performance to the target validation performance, $\bar{P}$ is the average value of P_{i}. θ limits the excessive difference between the task weights, γ adjusts the speed and intensity of weight deviation from uniformity and δ adjusts the emphasis on deviations of the current ratio from the mean ratio. The softsign function is introduced to solve the gradient problem; therefore, the network convergence is faster and less prone to saturation. We use the grid search method to train the hyper-parameters (θ, γ, δ) by the validation set at $\{1, 5, 10, 15\}$, $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. The hyper-parameters (θ, γ, δ) are set as (10,0.1,0.1) in our experiments.

3.3 Multimodal context-aware learning

In order to further explore the spatial context information embedded in CNN multilayer features, this paper proposes multimodal context-aware learning (MCL) that effectively integrates both multimodal information. Figure 3 shows two sets of MCLs, and we first take one of them to explain the processing flow. The MCL consists of three parts: the feature processing of IRIs, the feature processing of ORIs and the multimodal feature fusion processing.

Feature extraction phase In the feature extraction phase, we employ a convolution layer, a batch normalization, a max-pooling layer and Res-blocks to extract feature. The max-pooling layer is used for down-sampling, with a 3×3 kernel and astride of 2.

Context-aware block The main purpose of the context-aware blocks is to learn the spatial context information in the feature. We design three different context-aware blocks which have different structure to capture context information from varying receptive fields. The different scale convolution layers in three context-aware blocks are designed to enhance the scalability of the complementary representation. The first block contains a 3×3 convolution layer, a batch normalization and a LeakyReLU activation. The second block contains two residual blocks. The residual block consists of three convolution layers each followed by a batch normalization and a ReLU activation. The output feature of the last 1×1 convolution layer is added to the input feature as the final output feature. The third block has two branches. One branch has 3×3 average pooling layer and 1×1 convolution layer, and the other branch has 3×3 convolution layer and 1×1 convolution layer. Each 1×1 convolution layer followed by a batch normalization and a ReLU activation, and the two branches finally merged by addition operation.

Feature superposition module The inter-modal feature mapping loss($L_{\text{opt}-\text{inf}}$, $L_{\text{inf}-\text{opt}}$) is deployed prior to the feature superposition module, which enforce the features from the optical branch and the infrared branch to be as similar as possible. The entanglement of multimodal latent features is to synergistically and complementarily map the features from different modalities (i.e., optical or infrared) into a unified joint latent space, which reconstructs the encoded information (targets or background) from the pair of the remote sensing images to improve the accuracy of saliency detection. The feature superposition module entangles the features from the ORIs branch and the IRIs branch in MCL to improve the efficacy of the feature information and decrease divergence. We build the basic modules based on convolution layer–batch normalization–pooling layers which can realize the compactness of image feature representation. In the end, the three processed features are spliced by the concatenation operation and mapped from the pool layer to their respective feature space. We deploy the addition operation to fuse F_{IO} and F_{OI}. Context awareness is used to enhance the spatial and semantic relationships of feature maps from different modalities. Multimodal feature fusion processing parts for IRIs and ORIs enhance the semantic relationship between features of different modalities. And the pooling layers at different scales aim to sample the features at different dimensions to obtain spatial information at different locations and scales in the image. The pooling layer with smaller size can reflect more information about the spatial details of an image, while the larger size pooling layer can reflect information about larger subareas of an image.

Inter-modal feature mapping loss We designed the inter-modal feature mapping loss ($L_{\text{opt}-\text{inf}}$, $L_{\text{inf}-\text{opt}}$) prior to the addition operation to better entangle the latent vectors of the two modalities. The goal of modal feature mapping loss is to entangle the feature vectors of IRIs and ORIs into a unified joint latent space, which forces the outputs from the two modalities to be as similar as possible. The detail of two loss functions is expressed by:

$$\begin{aligned} L_{\text{opt}-\text{inf}} = {} & E_{A'_{\text{IRI}} \sim INF'}\left[\left\| M_{\text{I}}(F_{\text{e}}(A'_{\text{IRI}})) - B_{\text{ORI}} \right\|_1\right] + \\ & E_{A_{\text{ORI}} \sim OPT}\left[\left\| M_{\text{O}}(F_{\text{e}}(A_{\text{ORI}})) - B'_{\text{IRI}} \right\|_1\right], \end{aligned} \tag{6}$$

$$\begin{aligned} L_{\text{inf}-\text{opt}} = {} & E_{A'_{\text{ORI}} \sim OPT'}\left[\left\| M_{\text{O}}(F_{\text{e}}(A'_{\text{ORI}})) - B_{\text{IRI}} \right\|_1\right] + \\ & E_{A_{IRI} \sim INF}\left[\left\| M_{\text{I}}(F_{\text{e}}(A_{\text{IRI}})) - B'_{\text{ORI}} \right\|_1\right], \end{aligned} \tag{7}$$

where $A_{\text{IRI}} \sim INF$ and $A'_{\text{IRI}} \sim INF'$ represent the feature vectors of IRIs GT and IRIs generated, respectively. $A_{\text{ORI}} \sim OPT$ and $A'_{\text{ORI}} \sim OPT'$ denote the feature vectors of ORIs GT and ORIs generated, respectively. In Fig. 3, F_e denotes the feature extraction, and M_O and M_I denote the three context-aware block for ORIs and IRIs, respectively. E represents the expectation of the outcomes. $B_{\text{IRI}} = M_I(F_{\text{e}}(A_{\text{IRI}}))$, $B'_{\text{ORI}} = M_{\text{O}}(F_{\text{e}}(A'_{\text{ORI}}))$, $B'_{\text{IRI}} = M_{\text{I}}(F_{\text{e}}(A'_{\text{IRI}}))$ and $B_{\text{ORI}} = M_{\text{O}}(F_{\text{e}}(A_{\text{ORI}}))$.

3.4 Latent space GAN

Our latent space GAN, which contains a generator G and a discriminator D, operates in the joint latent space. As seen in Fig. 1, the input is a random noise vector and generator produce F_N to deceive discriminator. Taking a pair of IRI-

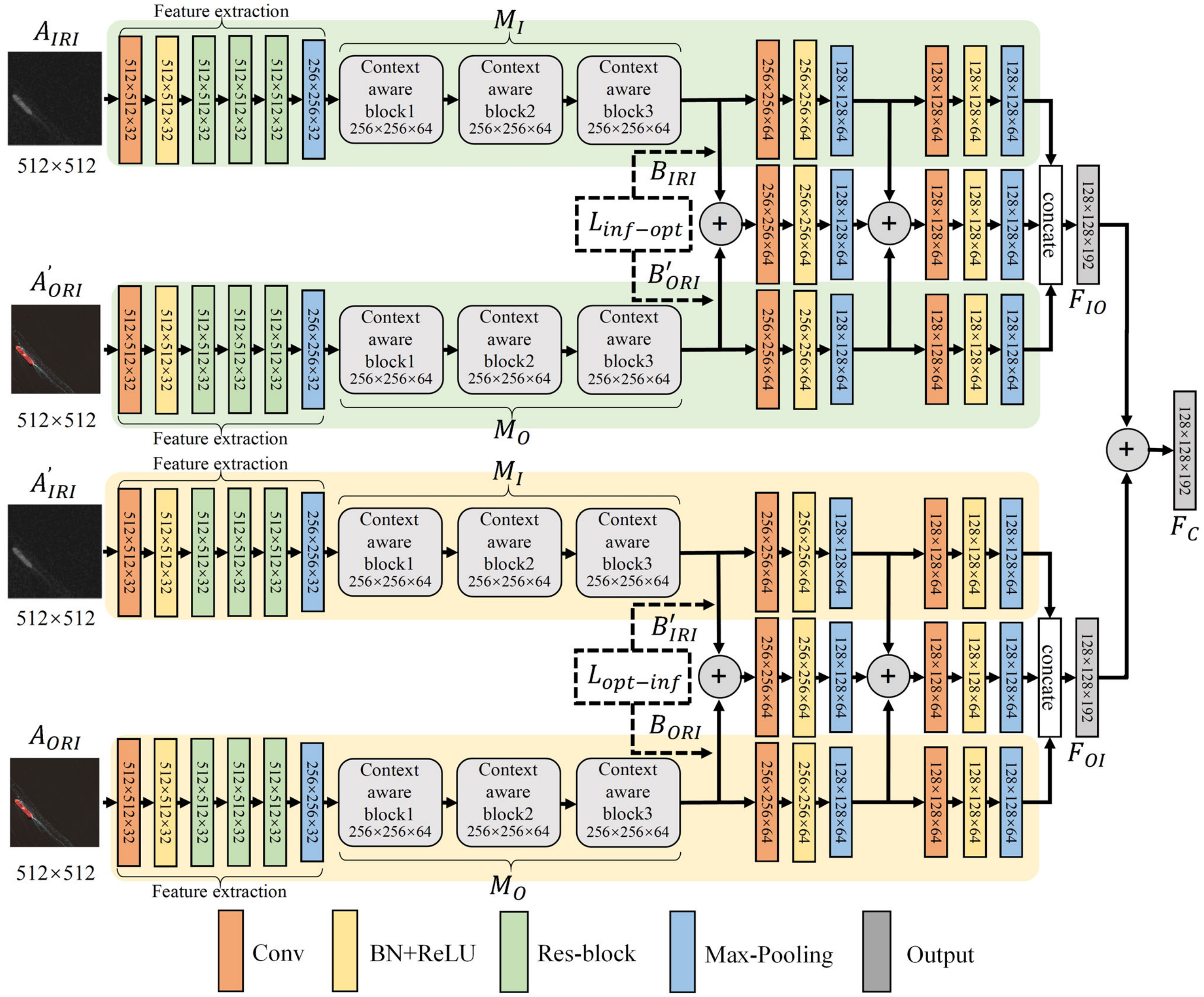

Fig. 3 The architecture of two multimodal context-aware learnings (green and yellow). A_{IRI} and A'_{IRI} represent the IRIs GT and generated IRIs, respectively. A_{ORI} and A'_{ORI} represent the ORIs GT and generated ORIs, respectively. M_O and M_I denote the processing flow of three context-aware blocks for ORIs and IRIs, respectively. $L_{opt-inf}$ and $L_{inf-opt}$ are detailed in Eqs. (6) and (7). F_{IO} and F_{OI} mean the concatenation features of green MCL and yellow MCL, respectively. F_C is the final output feature of two MCLs

ORI, we encode them into one joint latent vector F_C in the joint latent space INF-OPT. During training, the multimodal context-aware learning is fixed. We used Wasserstein GAN [46] in the joint adversarial loss:

$$\begin{aligned} L_{\text{JGAN}} = & E_{F_N \sim \text{Noise}}[D(F_\text{N})] - \\ & E_{F_C \sim \text{INF-OPT}}[D(F_\text{C})] + \mu L_{\text{rgp}}, \end{aligned} \tag{8}$$

where L_{rgp} represents the regularization gradient penalty loss and μ is a scalar weight (default is 10). F_N is the generator vector in the latent space $Noise$. F_C is the latent vector in the latent space INF−OPT.

3.5 Attention-aware CapsNet

To effectively extract the information in the remote sensing image, it is necessary to redesign the feature extraction module in the original CapsNet. The proposed AACNet is dedicated to segmenting salient items from the background and is used to investigate the relationship between objects in the input images.

Knowledge Distillation Strategy We import knowledge distillation strategy in AACNet, which can decrease the number of network parameters while preserve the model performance. As shown in Fig. 4, the original AACNet is defined as teacher network and the pruned AACNet is defined as

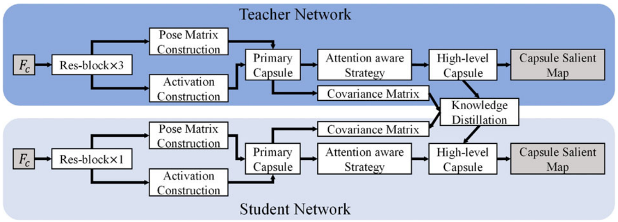

Fig. 4 Our knowledge distillation strategy in attention-aware capsule net (AACNet). F_c means the context-aware features from multimodal context-aware learning (MCL)

student network. The difference between teacher network and student network is the number of Res-block. We deploy three Res-blocks to extract the deeper features with stronger semantic information from multimodal context-aware feature prior to the operations of pose matrix construction and activation construction in teacher network. In contrast, the student network only remains one Res-block, which means that the parameters of student network are less. The specific process of our knowledge distillation strategy is as follows. First, we pre-train the teacher model with cross entropy loss, then we train the student model with the help of privileged information learned by teacher model. The lower-level capsule contains local features, and its different dimensions capture different aspects of the information space [38]. The length of the output vectors of the high-level capsule represents the probability of the existence of the entity, and the direction represents the instantiated parameters. In order to make full use of the capsule information, we build knowledge distillation loss L_{kd} to constrain the inter-dimension correlation similarity of lower-level capsules (primary capsules) and the difference of information distribution of higher-level capsules between teacher network and student network [47]. The covariance matrices are used to calculate the similarity between them, which is defined as:

$$M_t = C_{\mathrm{t}}^T \cdot C_{\mathrm{t}};\ M_{\mathrm{s}} = C_{\mathrm{s}}^T \cdot C_{\mathrm{s}}, \tag{9}$$

where M_{t}, M_{s} denotes the covariance matrices of lower level capsules of the teacher network (C_{t}) and student network (C_{s}) respectively. The knowledge distillation loss which contains the lower-level capsules loss and higher-level capsules loss is as follows:

$$\begin{aligned} L_{\mathrm{kd}} = & \left\| \frac{M_{\mathrm{t}}}{\|M_{\mathrm{t}}\|_2} - \frac{M_{\mathrm{s}}}{\|M_{\mathrm{s}}\|_2} \right\|_F^2 + \\ & \mathrm{KL}[\log(\sigma(\|h_{\mathrm{s}}\| / \tau)), \sigma(\|h_{\mathrm{t}}\| / \tau)]\tau^2, \end{aligned} \tag{10}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\sigma(\cdot)$ represents the softmax activation, $KL[\cdot]$ means the KL divergence and τ denotes the temperature parameter [48]. h_{s} and h_{t} are the higher-level capsules of student network and teacher network, respectively.

The training loss function of AACNet L_{AACNet} contains two-part, knowledge distillation loss and capsule salient map (CSM) cross-entropy loss L_{ce}:

$$\begin{aligned} L_{\mathrm{ce}}(CSM) = - \sum_{i=1}^{W \times H} & (G_{\mathrm{s}1}(i) \log(CSM(i)) + \\ & G_{\mathrm{s}0}(i) \log(1 - CSM(i))), \end{aligned} \tag{11}$$

$$L_{\mathrm{AACNet}} = \alpha L_{\mathrm{ce}}(CSM) + \beta L_{\mathrm{kd}}, \tag{12}$$

where $G_{\mathrm{s}1}$ and $G_{\mathrm{s}0}$ represent the salient object pixels and background pixels, respectively, in the salient ground truth. CSM denotes the capsule salient map generated by high-level capsules through up-sampling layer. α and β are the hyperparameters for cross-entropy loss and knowledge distillation loss, respectively.

Construction of primary capsules The features generated by the multimodal context-aware learning are first transferred into 8 capsules, each of which is consisted of an activation value and a 4×4 pose matrix. Two Conv-BN-LeakyReLU activations are used to downsample the multimodal feature maps $F_m(128 \times 128 \times 192)$ into feature map $F_d(64 \times 64 \times 128)$ for the subsequent operation. First, we construct the pose matrix $M_p(64 \times\ 64 \times 8 \times 16)$ by the convolution layer and reshape layers. Then, we compute the activation information $A(64 \times 64 \times\ 8 \times 1)$ of 8 capsules via one convolution layer. In the end, we combine the pose matrixes M_p and the activation information A to set up the primary capsules $C_p(64 \times 64 \times 8 \times 17)$.

Attention-aware strategy We split the primary capsules into two capsule groups $C_{r1}(64 \times 64 \times 4 \times 17)$ and C_{r2} $(64 \times 64 \times 4 \times 17)$. To get a capsule with more advanced features, We reshape the two sets of capsules to $C_1^{'}(2048 \times 36 \times 17)$ and $C_2^{'}(2048 \times 36 \times 17)$ via a convolution layer with the step size of 2 and the channel number of 9. Then sent C_{r1} and C_{r2} into two routes (EM routing and Sigmoid dynamic routing), respectively. After that, we input the pose matrix and the transformation matrix to calculate the number of votes $V(2048 \times 36 \times 8 \times 16)$ from the low-level capsules to the adjacently high-level capsules. At last, we assign low-level and high-level capsules to each other. The difficulty of allocating each portion to the total can be solved by discovering tight voting clusters from each part. The establishment of RCaps2 is identical to RCaps1 but differs in one way. Instead of EM routing, we employ Sigmoid dynamic routing [38] in RCaps2. Jia and Huang [49] proved that the probability of sending features to potential capsules is almost equal, which may lead to wrong classification. Therefore, we use the sigmoid function to calculate the coupling coefficient θ_{ij}, which no longer represents the distribution probability of the final capsule, but represents the correla-

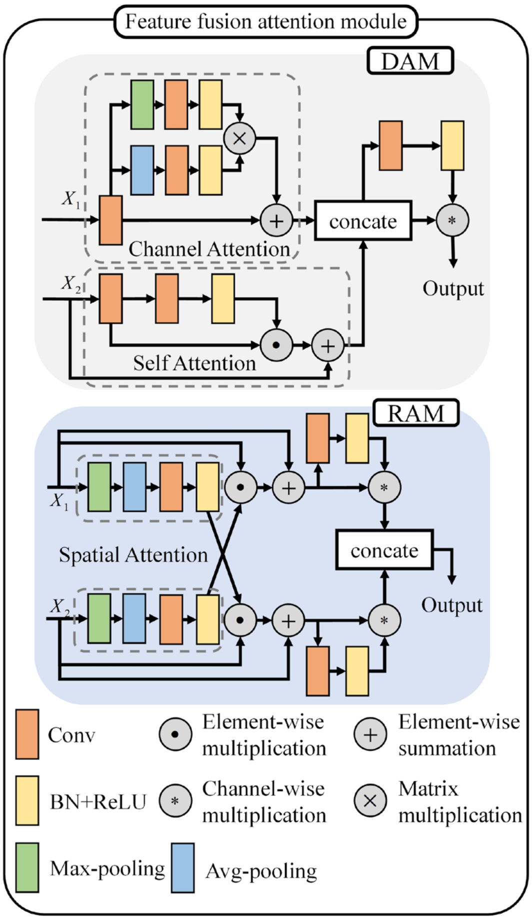

Fig. 5 The reciprocal attention module (RAM) and dual attention module (DAM) in feature fusion attention-aware module (FFAM). The RCaps1 is in charge of extracting background characteristics with a broad receptive field. The RCaps2 is designed to extract salient objects, and the RAM is used to improve the features on several levels, which is inspired by spatial attention. Then, the output of the two routing is fused via the DAM which is inspired by self-attention and channel attention

tion strength between the primary capsule and the potential capsule. We apply feature clustering for the final feature $F_{R1}(32\times32\times128)$ and $F_{R2}(32\times32\times128)$ after the routing process. For dual-routing attention-aware fusion, we employ the Feature fusion attention-aware module (FFAM), which consists of two branches, depicted in Fig. 5. Based on the spatial attention structure, we create the reciprocal attention module (RAM) and applied it to several layers of the RCaps2 branch. RAM can emphasize the spatial information related to salient object areas due to the contribution of spatial attention structure.

We create a dual attention module (DAM) based on self-attention and channel attention to efficiently merge the two branches. Self-attention is employed to represent the distant dependence of pixels in different regions and get global context information in the RCaps1 branch. Dilated gated convolution is used in the EM routing branch (RCaps1) to provide appropriate content by extract features with large receptive field. The sigmoid routing branch (RCaps2) makes advantage of RAM to improve edge features and the gated residual block to make the branch more detailed. Finally, we combine the two branches using DAM to create higher-level capsules. Based on the multimodal context-aware features, attention-aware capsule net constructs two attention branches to avoid the attention mechanism focusing on a single region, which makes the final generated capsule salient map more accurate and the model more robust.

3.6 Boundary guidance network

The proposed boundary guidance network (BGNet) is shown in Fig. 1, which can generate accurate saliency detection map based on the guidance of boundary-aware block (BAB) and multiscale feature learning network (MFLN). First, we use MFLN to output multiscale aggregated features $(P1, P2, P3)$ based on the shallow features (f_1, f_2, f_3, f_4) extracted by WSA-GAN, which will be used by the BAB to generate edge map for supervised training. Although the boundary of the object inside in capsule salient map (CSM) is blurred, irregular bright spots appear in the background area, and the smoothness of the salient parts is low. We can use the multiscale aggregation features $(P1, P2, P3)$ extracted by MFLN, and the edge features extracted by BAB to generate the more accurate saliency detection map through the guidance of concatenation and convolution operations and conduct supervised training with the saliency GT, as shown in Fig. 1.

Multiscale feature learning network We designed multiscale feature learning network (MFLN) to further process the extracted shallow features (f_1, f_2, f_3, f_4). As shown in Fig. 1, from f_1 to f_4 are divided into four groups, and each group stacked up the features treated by three different parameters of dilated convolution layers. The multiscale aggregated features $(P1, P2, P3)$ are extracted using the MFLN which allows rich contextual information to be captured without increasing the kernel size. In addition, we collect context data across multiple scales and combine deep features with shallow features since low-level feature maps help to capture details, while high-level feature maps help to capture semantic knowledge.

Boundary-aware block By reviewing the existing ORI and IRI saliency detection models, we find that in the current work, there are few studies that focus on boundary information. Therefore, we propose the edge-aware block that focuses on boundary information, which enables our model to overcome the fuzzy boundary provided by CSM and generate a salient prediction map with clear boundary. First, we select the multiscale aggregated features $(P1, P2, P3)$ from MFLN and up-sample them to the same dimension for concatenation operation. Then, we use convolution layer

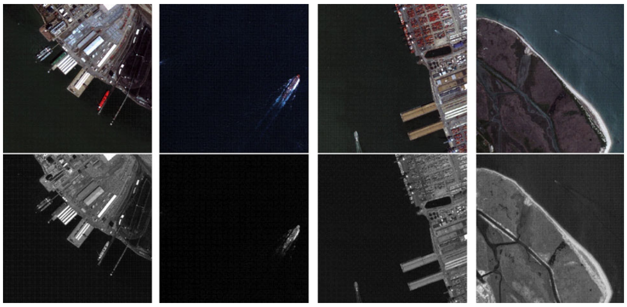

Fig. 6 Our Dataset. The optical remote sensing images (ORIs) are in the first row, and the infrared remote sensing images (IRIs) are in the second row

to further process the concatenated feature and the result is called $F_{e_{tmp}}$. At last, based on the channel attention mechanism, the temporary edge features $F_{e_{tmp}}$ are further processed to generate clearer edge information by:

$$F_{\mathrm{e}} = \mathrm{Up}(\mathrm{Conv}(\mathrm{Conv}(\mathrm{Max}P(F_{e_{tmp}})) \times F_{e_{tmp}} + F_{e_{tmp}})), \tag{13}$$

where Max$P$ represents the max-pooling layer, Up$(\cdot)$ represents upsampling operation, Conv$(\cdot)$ denotes convolution layer followed by a batch normalization layer and a ReLU activation, "$+$" is element-wise addition operation and "$\times$" is element-wise multiplication operation.

With the aid of edge features (F_{e}) and the multiscale aggregated features (P_1,P_2,P_3), the process of generating the multiscale saliency detection features from CSM is as follows:

$$F_{S_i} = \mathrm{Up}(\mathrm{Concat}(\mathrm{Up}(CSM), Pi)), \tag{14}$$

where $i \in 1,2,3$, Concat$(\cdot)$ is the concatenation layer.

Then, based on edge feature F_{e}, the salient features F_{S_i} are further processed to generate the final saliency detection map S_m by:

$$\begin{aligned} S_{\mathrm{m}} = {} & \mathrm{Up}(\mathrm{Conv}(\mathrm{Concat}(F_{\mathrm{e}}, (\mathrm{Conv}(\mathrm{Conv}(F_{S_3})) \\ & + Up(Conv(F_{S_2}) + Up(F_{S_1})))))), \end{aligned} \tag{15}$$

where $Conv(\cdot)$ denotes convolution layer followed by a batch normalization layer and a ReLU activation.

Cross-entropy loss We use cross entropy loss to realize the edge supervision and saliency supervision. The edge map E_m is generated from edge-aware block by up-sampling edge features F_e. The cross-entropy edge loss is defined as:

$$\begin{aligned} L_{\mathrm{edge}}(E_{\mathrm{m}}) = - & \sum_{i=1}^{W \times H} (G_{e1}(i) \log(E_m(i)) + \\ & G_{e0}(i) \log(1 - E_m(i))), \end{aligned} \tag{16}$$

$$\begin{aligned} L_{\mathrm{object}}(S_{\mathrm{m}}) = - & \sum_{j=1}^{W \times H} (G_{\mathrm{s}1}(j) \log(S_{\mathrm{m}}(j)) + \\ & G_{s0}(j) \log(1 - S_{\mathrm{m}}(j))), \end{aligned} \tag{17}$$

where G_{e1} and G_{e0} indicate the edge pixels and background pixels respectively in the edge ground truth. G_{s1} and G_{s0} denote the salient object pixels and background pixels, respectively, in the salient ground truth.

The overall loss function is $L_{\mathrm{BGNet}} = L_{\mathrm{edge}}(E_{\mathrm{m}}) + L_{\mathrm{object}}(S_{\mathrm{m}})$. Therefore, the final salient features contain low-level spatial information and high-level semantic information and are guided by edge information, and we can generate high-quality and more accurate saliency detection maps.

In order to realize multimodal image-to-image translation and saliency detection of multimodal remote sensing images, with an end-to-end fashion, we use the different weight parameters to fuse the above loss function (except for $L_J GAN$, the latent space GAN is to ensure the feature reproducibility and we train it separately.) for the complete model training. The complete model loss function is as follows:

$$\begin{aligned} L_{CM} = {} & \mu_1 G^* + \mu_2 (L_{opt-inf} + L_{inf-opt}) \\ & + \mu_3 L_{AACNet} + \mu_4 L_{BGNet}, \end{aligned} \tag{18}$$

where μ_1, μ_2, μ_3 and μ_4, respectively, represent the coefficients of each loss function. We set $\mu_1 = 0.3$, $\mu_2 = 0.2$, $\mu_3 = 0.2$ and $\mu_4 = 0.3$ in our experiments.

Table 1 Quantitative evaluation of image translation

Evaluation metrics	MAE (↓)	PSNR (↑)	SSIM (↑)	Model size (MB)	Test time (ms)
Pix2pix	100.137	17.802	0.639	210	**70**
Pix2pixHD	102.025	16.896	0.621	713	87
CycleGAN	100.923	17.564	0.627	**43**	243
MUNIT	104.755	16.060	0.501	127	211
DRIT	103.895	16.152	0.567	717	301
DualGAN	98.436	17.449	0.637	270	294
Ours	**96.497**	**18.402**	**0.682**	879	85

Bold values indicate the best data for the corresponding evaluation indicator

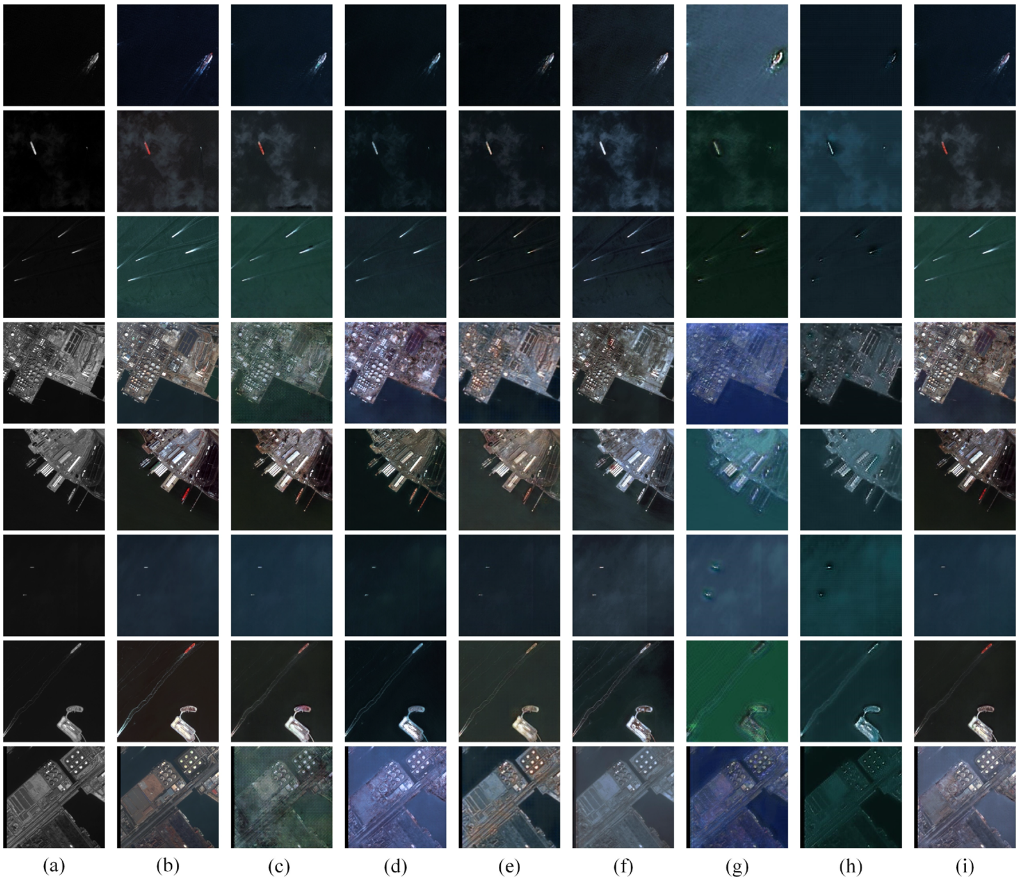

Fig. 7 Image-to-image translation results of IRIs-to-ORIs with different methods. **a** Input IRIs; **b** Ground truth of ORIs; **c** Pix2pix **d** CycleGAN **e** DualGAN **f** Pix2pixHD **g** MUNIT **h** DRIT and **i** Our method

4 Experiments and implementation

In this section, we conduct extensive experiments and analysis to verify the effectiveness and advantages of our multitask learning framework. We compared our methods on the two tasks, image translation and salient object detection, with the state-of-the-art methods on different datasets qualitatively and quantitatively. All models in this paper are trained on an NVIDIA GeForce GTX1080Ti GPU, and the model is trained using Pytorch 1.2.0 with momentum set to 0.9 and weight decay set to 5×10^{-4}. The initial learning rate was 10^{-4}, the number of iterations of the model is 800 epochs, and the batch size is 2. The Adam optimizer [50] is used to train our model with an initial learning rate of 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The source code of our framework and the dataset will be released shortly.

4.1 Datasets

Our dataset The experimental data of the self-made dataset includes 800 ORIs and 800 IRIs from the Landsat-8 satellite. We crop these images into image slices of 512*512 pixels, which have a resolution of 0.5−2.0 m/pixel. We manually annotate the salient objects of the image at the pixel level. Compared with other datasets, our image targets are smaller and contain more complex scenes. Figure 6 shows some of these examples.

ORSSD dataset ORSSD dataset [31] collected 800 ORIs from Google Earth or other datasets, and the images are manually labeled at the pixel level for a variety of significant targets (islands, ships, vehicles, etc.). We conduct separate training for this dataset, and the visual and quantitative comparisons are described in this section.

EORSSD dataset EORSSD dataset [51] is built based on ORSSD dataset, which has more types of significant targets and a larger number of images. In addition to the original 800 ORIs, the dataset includes another 1200 ORIs from Google Earth. We also conduct separate training for this dataset, and the visual and quantitative comparisons are presented in this section.

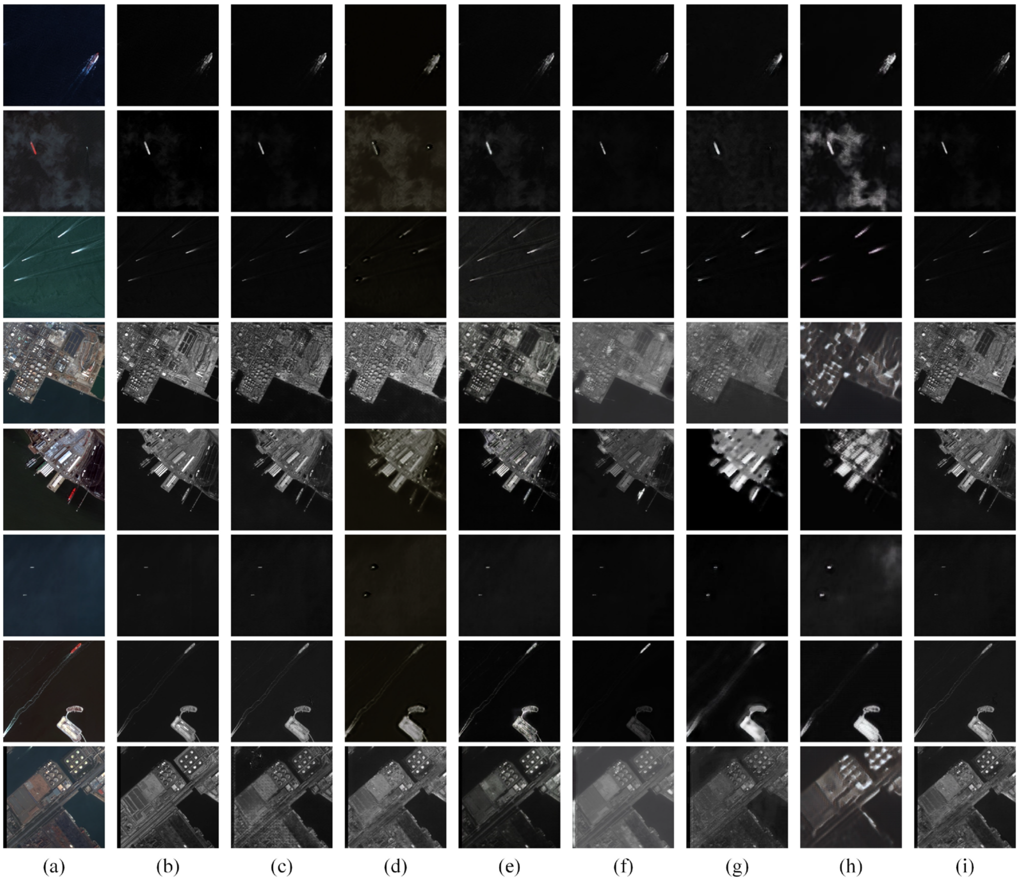

Fig. 8 Image-to-image translation results of ORIs-to-IRIs with different methods. **a** Input ORIs; **b** Ground truth of IRIs; **c** Pix2pix **d** CycleGAN **e** DualGAN **f** Pix2pixHD **g** MUNIT **h** DRIT and **i** Our method

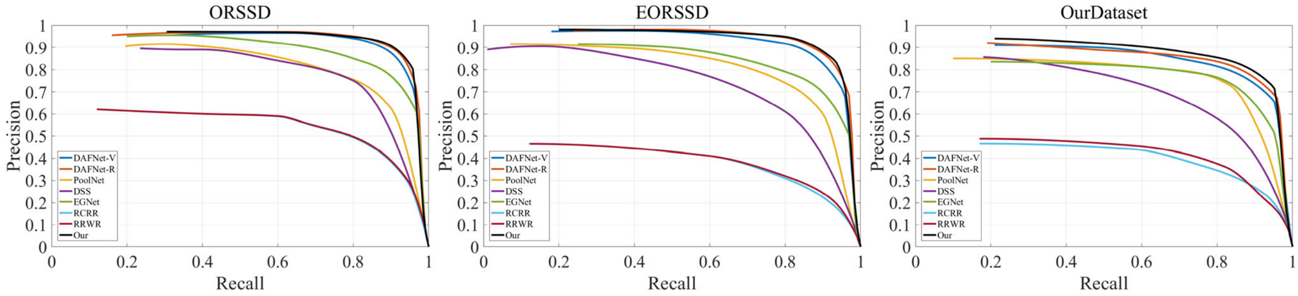

Fig. 9 P-R curves of different methods on the testing subset of the ORSSD, EORSSD and our dataset

4.2 Image-to-image translation

To demonstrate the efficacy of our proposed model, we compare it to Pix2pix [6], CycleGAN [9], Pix2pixHD [7], MUNIT [14], DualGAN [8], and DRIT [13]. The above methods are retrained with our dataset under the default parameter settings of the corresponding model. Both the visual and quantitative comparison will be taken into account.

Evaluation metrics In this subsection, we use PSNR (peak signal-to-noise ratio), MAE (mean absolute error), and SSIM (structural similarity index) to evaluate the image conversion quality. The index PSNR for image quality evaluation, which is measured in db (decibels), depends on the mean square error (MSE). Mathematically, the PSNR is defined according to the error between the corresponding pixels as follows:

$$PSNR = 10 \times \log_{10} \frac{(2^n - 1)^2}{MSE}, \tag{19}$$

where n is the number of bits per pixel, and MSE is calculated by:

Table 2 Quantitative comparisons with different methods on the testing subset of the ORSSD, EORSSD and our datasets

Dataset	ORSSD			EORSSD			Our Dataset				
Evaluation metrics	$F_\beta(\uparrow)$	MAE(↓)	$S_m(\uparrow)$	$F_\beta(\uparrow)$	MAE(↓)	$S_m(\uparrow)$	$F_\beta(\uparrow)$	MAE(↓)	$S_m(\uparrow)$	Model size(MB)	Test time(ms)
DAFNet-V [51]	0.9174	0.0125	0.9191	0.8922	0.0060	0.9167	0.8681	0.0153	0.9121	110	**204**
DAFNet-R [51]	0.9235	**0.0106**	0.9188	0.9060	0.0053	**0.9185**	0.8855	0.0113	0.9164	115	221
PoolNet [52]	0.7911	0.0358	0.8403	0.7812	0.0209	0.8218	0.7536	0.0379	0.8008	265	205
DSS [30]	0.7838	0.0363	0.8262	0.7158	0.0186	0.7874	0.6467	0.0397	0.7379	237	312
EGNet [29]	0.8438	0.0216	0.8721	0.8060	0.0109	0.8602	0.7578	0.0268	0.8493	427	255
RCRR [53]	0.5944	0.1277	0.6849	0.4495	0.1644	0.6013	0.3687	0.1879	0.5128	0.25	3110
RRWR [25]	0.5950	0.1324	0.6835	0.4495	0.1677	0.5997	0.3705	0.1923	0.5076	**0.23**	3092
OURS	**0.9368**	0.0113	**0.9254**	**0.9101**	**0.0047**	0.9172	**0.9081**	**0.0097**	**0.9276**	879	283

Bold values indicate the best data for the corresponding evaluation indicator

$$MSE = \frac{1}{H \times W}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}[I_y(i,j) - I_{G(x)}(i,j)]^2, \qquad (20)$$

where H and W represent the height and width of images, respectively; I_y stands for a GT image and $I_{G(x)}$ is a generated image.

MAE uses pixel measures and relies on the error between the predicted image and the real image, which is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|G(x)_i - y_i|, \qquad (21)$$

where n is the number of pixels in the image, and $G(x)_i$ and y_i are the values of the pixels for the generated image and the GT image, respectively.

The standard SSIM is commonly used to evaluate the structural similarity between grayscale images in terms of three factors: luminance, contrast and structure, defined as:

$$SSIM(y, G(x)) = \frac{(2\alpha_y\alpha_{G(x)} + c_1)(2\delta_{yG(x)} + c_2)}{(\alpha_y^2 + \alpha_{G(x)}^2 + c_1)(\delta_y^2 + \delta_{G(x)}^2 + c_2)}, \qquad (22)$$

where y and $G(x)$, respectively, stand for the pixels of the GT image and the generated one; α_y and $\alpha_{G(x)}$ are their average values; δ_y and $\delta_{G(x)}$ represent their standard deviation, and $\delta_{yG(x)}$ is the covariance of these two images; c_1 and c_2 are two parameters to avoid the denominator being 0. The larger the SSIM value, the more similar the structure between the two images.

Comparison with other methods The results of these three metrics on the dataset are shown in Table 1. From this table, it can be seen that our method shows the smallest MAE, the largest PSNR and the largest SSIM compared to the other six methods. In addition, the reduction value of MAE reaches 1.939 and the increase values of PSNR and SSIM reach 0.6 and 0.043, respectively, compared to the results of the second best method. The results of the image transformation are plotted in Figs. 7 and 8, and the selected results have a variety of targets, such as ships, islands and Oil tanks. As can be seen from the figures, our method is closer to the actual image, although Pix2pix does generate competitive results, in which our method has better results in terms of details.

4.3 Salient object detection

To demonstrate the advantages of our proposed model, we compare it with the SOTA methods, PoolNet [52], DAFNet-V [51], DAFNet-R [51], DSS [30], EGNet [29], RCRR [53] and RRWR [25]. Since our dataset is the only one containing ORIs-IRIs pairs, the IRIs saliency detection experiment use

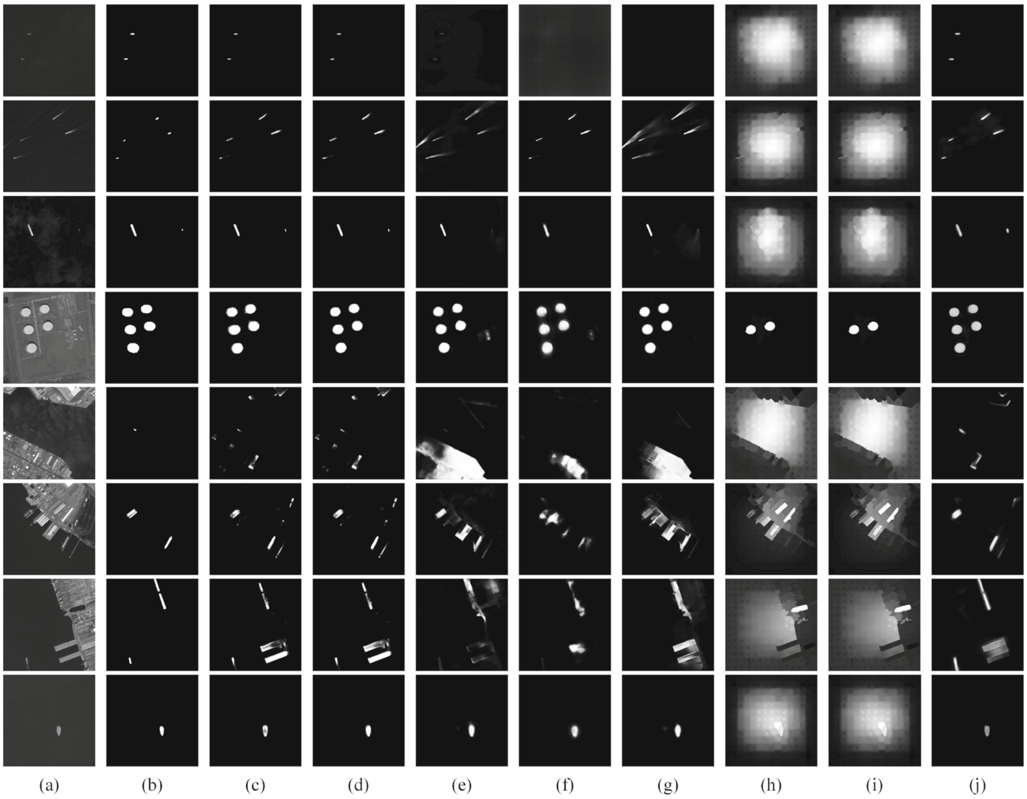

Fig. 10 Salient object detection result of IRIs with different method. **a** Input IRIs; **b** Ground truth images; **c** DAFNet-V **d** DAFNet-R; **e** PoolNet **f** DSS **g** EGNet **h** RCRR **i** RRWR **j** Our method

our dataset. For the input with only optical modal (Fig. 11), we use the optical branch method to compare with other method. The above methods based on deep learning use our dataset, ORSSD dataset and EORSSD dataset for retraining under the default parameter settings of the corresponding model. Both visual comparison and quantitative comparison will be taken into account.

Evaluation metrics We use the precision–recall (PR) curve, MAE score, F-measure and S-measure to evaluate the performance of different methods. The saliency map can be thresholded by integers ranging from 0 to 255 into some binary saliency masks and then compared with the real values to obtain precision and recall. Different combinations of precision and recall scores are used to draw the PR curve.

F-measure is an overall performance indicator, which is computed by

$$F_\beta = \frac{(1+\beta^2)\text{Precision} \times \text{Recall}}{\beta^2\text{Precision} + \text{Recall}}, \tag{23}$$

It is recommended to set β^2 to 0.3 in [2].

MAE is defined as Eq. (21). However, in this phase of the experiment, the generated image was replaced with saliency map. S-measure [27] S_m is computed by:

$$S_m = \alpha S_o + (1-\alpha) S_r, \tag{24}$$

where S_o and S_r represent the object-aware and region-aware structure similarities between the prediction and the ground truth, respectively. α is set to 0.5 as in [54].

Comparison with other methods On the two most prominent remote sensing salient object detection datasets and our own dataset, Fig. 9 depicts the PR curves of our method in comparison to other methods. If we observe the upper right corner of the PR curve, our method will produce higher accuracy when the recall score is close to 1, which indicates that the false positive rate is low. The advantages as shown by the curve also suggest that our resulting image is closer to ground truth. Table 2 lists the average values of the F-measure, MAE, and S-measure for different methods. Our model embodies competitive performance. On average, our method outperforms other approaches in the three quantitative metrics on our datasets and maintains competitive performance on ORSSD and EORSSD datasets. In addition, the performance of training methods based on deep learning is significantly better than traditional methods. On the ORSSD dataset, DAFNet-R (second-best method) reaches 0.9235 in F-measure, on the EORSSD dataset, it is 0.9060,

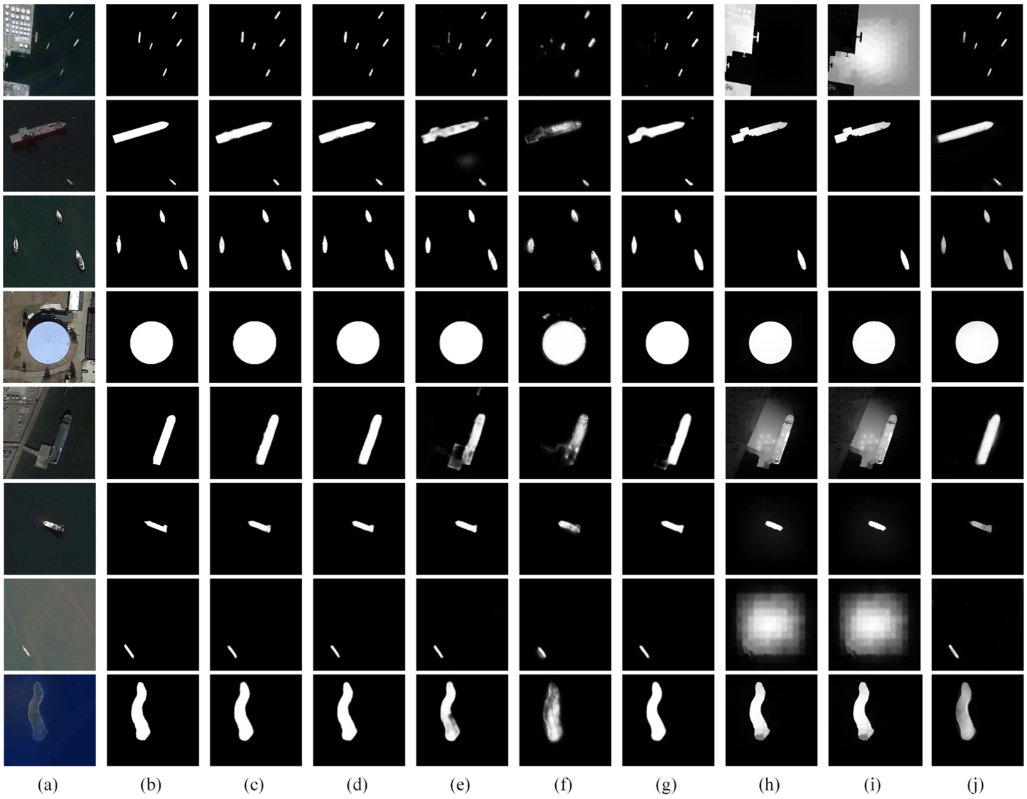

Fig. 11 Salient object detection result of ORIs with different method. **a** Input ORIs; **b** Ground truth images; **c** DAFNet-V **d** DAFNet-R **e** PoolNet **f** DSS **g** EGNet **h** RCRR **i** RRWR **j** Our method

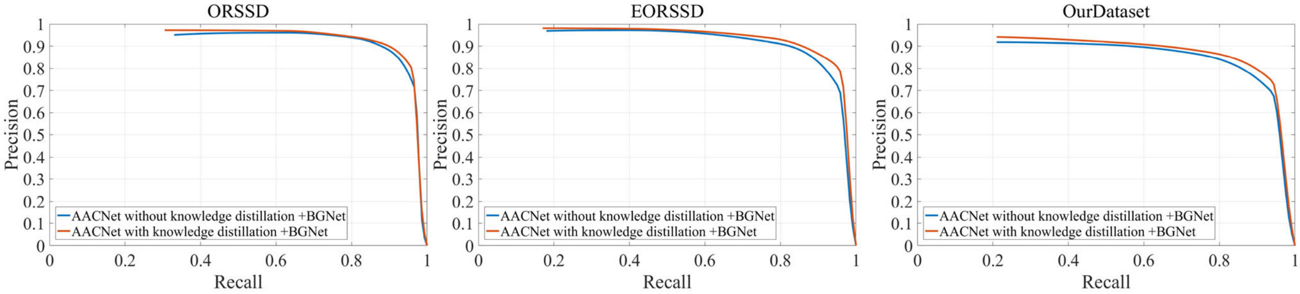

Fig. 12 P-R curves of ablation study of knowledge distillation strategy on the testing subset of the ORSSD, EORSSD and our dataset

and on our dataset, it is 0.8855. We have a percentage gain of 2.26 percent in F-measure, 0.16 percent in MAE scores, and 1.12 percent in S-measure when compared to the second-best method in our dataset. Our method has achieved greater percentage improvements on the ORIs-IRIs dataset in terms of the three quantitative metrics, i.e., F-measure, MAE and S-measure, which demonstrates that the multimodal fusion plays a significant role. Because the ORSSD and EORSSD datasets are optical remote sensing datasets, compared with the complete model structure used on our dataset, the model on the ORSSD and EORSSD only retains one WSA-GAN and MCL branch.

Figures 10 and 11 depict the running results of a variety of graphics with different properties, such as small targets, large targets, shore targets, offshore targets, and center deviations. Our model takes into account the majority of scenarios and performs well with photos of various properties. In general, our detection results are closer to the ground truth in many situations than the SOTA methods.

4.4 Model analysis and ablation study

To demonstrate the importance of including the knowledge distillation strategy, attention-aware strategy, and multiad-

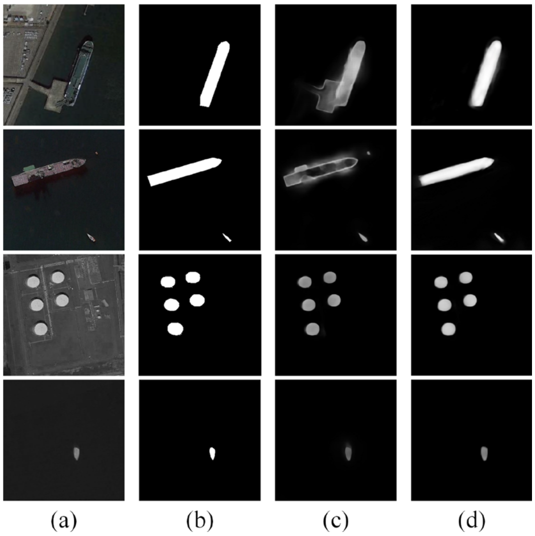

Fig. 13 Visual comparisons for the different routing strategy in salient object detection ablation study. **a** Image; **b** GT; **c** Single-routing CapsNet + BGNet; **d** Attention-aware CapsNet + BGNet

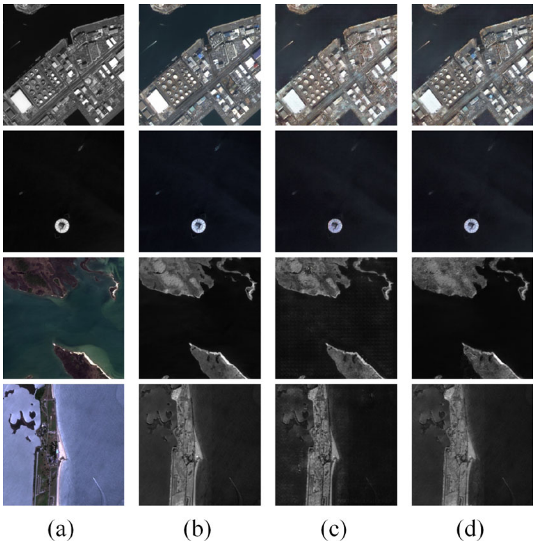

Fig. 14 Visual comparisons for the different GAN structure(MAB) in image translation ablation study. **a** Image; **b** GT; **c** Weight-sharing GAN (without MAB) **d** Weight-sharing attention GAN (contain MAB)

ditive attention blocks (MABs), we conduct model analysis and ablation experiments.

Knowledge distillation strategy In order to verify the importance of knowledge distillation strategy, we compared the PR curves of the following two structures on ORSSD, EORSSD and our dataset: AACNet with knowledge distillation strategy+BGNet and AACNet without knowledge distillation strategy (using student network)+ BGNet. It can be seen from Fig. 12 that on the three different datasets, the knowledge distillation strategy makes our model produce higher accuracy, which means the saliency detection map is closer to the ground truth.

Attention-aware strategy To better understand the superiority of the attention-aware strategy, we investigate two architectures, including "BGNet+Attention-aware CapsNet" and "BGNet+Single-routing CapsNet", where the latter is achieved by directly adopting the original CapsNet. As shown in Fig. 13, the single-routing CapsNet incorrectly marks some background areas as part of the salient objects, which indicates that the single-routing strategy introduces some noisy capsule assignments. In contrast, due to the involvement of the attention-aware strategy, our CapsNet successfully mitigates these noisy capsule assignments, helping to cluster the correct salient parts together, thus constituting the entire salient object. From Table 3, we can see that, after changing to the attention-aware CapsNet+BGNet structure, all evaluation indicators have improved. In our dataset, F_β increased by 3.41 percent, MAE decreased by 0.32 percent and SSIM increased by 0.24 percent.

Multiadditive attention blocks In order to better understand the superiority of the MABs, we studied two architectures, including weight-sharing attention GAN and weight-sharing GAN. The latter structure does not add the MABs structure, but uses the original U-Net structure in the corresponding positions. As shown in Fig. 14, without MABs, the image transformation is not effective and some details do not match the real image. In contrast, the use of MABs makes the image conversion better and the details are closer to the real image. From Table 4, we can see that, after adding the MAB module, all evaluation indicators have improved. Among them, MAE decreased by 3.322, PSNR increased by 0.476 and SSIM increased by 0.035.

5 Conclusions

In this paper, we have presented an attention GAN network with a weight sharing strategy to synthesize infrared

Table 3 Table caption

Dataset	ORSSD			EORSSD			Our Dataset		
Evaluation metrics	$F_\beta(\uparrow)$	MAE(↓)	$S_m(\uparrow)$	$F_\beta(\uparrow)$	MAE(↓)	$S_m(\uparrow)$	$F_\beta(\uparrow)$	MAE(↓)	$S_m(\uparrow)$
Single-routing CapsNet+BGNet	0.9154	0.0138	0.9192	0.8844	0.0068	0.9040	0.8740	0.0129	0.9252
Attention-aware CapsNet+BGNet	**0.9368**	**0.0113**	**0.9254**	**0.9101**	**0.0047**	**0.9172**	**0.9081**	**0.0097**	**0.9276**

Bold values indicate the best data for the corresponding evaluation indicator

Table 4 Quantitative evaluation of different GAN structures in image translation ablation study

Evaluation metrics	MAE(↓)	PSNR(↑)	SSIM(↑)
Weight-sharing GAN	99.819	17.926	0.647
Weight-sharing Attention GAN	**96.497**	**18.402**	**0.682**

Bold values indicate the best data for the corresponding evaluation indicator

remote sensing images from optical remote sensing images (and vice versa) in order to address the problem of a lack of IRIs and build complementary visual features for more accurate saliency detection. To fully explore context-aware information of IRIs and ORIs, we have designed a context-aware learning that co-encodes the entanglement and de-entanglement of features from the multimodal of IRIs and ORIs. Then, we construct the attention-aware CapsNet which can further enhance feature representations and correlation them through latent space, to address the problem that some targets (e.g., ships, vehicles, etc.) have directionality, yet CNN is not sensitive to the directionality. Finally, the boundary-aware block is proposed to generate final saliency detection result through the multiscale feature learning network, boundary-aware block and capsule salient map. In the image translation and salient object detection multitask, we noticed that our large model requires high memory and high computation cost. In the future, we will focus on how to reduce the cost of the computational tasks and model size while maintaining accuracy, such as lightweight capsule networks.

Funding This research was funded by NSFC 61972353, NSF IIS-1816511, OAC-1910469 and Strategic Cooperation Technology Projects of CNPC and CUPB: ZLZX2020-05.

Data availability The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study. If you have any questions, please consult the corresponding author.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Kaji, S., Kida, S.: Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging. Radiol. Phys. Technol. **12**(3), 235–248 (2019)
2. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: a benchmark. IEEE Trans. Image Process. **24**(12), 5706–5722 (2015)
3. Hasanov, A., Laine, T.H., Chung, T.S.: A survey of adaptive context-aware learning environments. J. Ambient Intell. Smart Environ. **11**(5), 403–428 (2019)
4. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
5. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 327–340 (2001)
6. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
7. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8798–8807 (2018)
8. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2849–2857 (2017)
9. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
10. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. Adv. Neural Inform. Process. Syst. **30** (2017)
11. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning, pp. 1857–1865. PMLR (2017)
12. Yoo, D., Kim, N., Park, S., Paek, A.S., Kweon, I.S.: Pixel-level domain transfer. In: European Conference on Computer Vision, pp. 517–532. Springer (2016)
13. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 35–51 (2018)
14. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172–189 (2018)
15. Lan, J., Ye, F., Ye, Z., Xu, P., Ling, W.K., Huang, G.: Unsupervised style-guided cross-domain adaptation for few-shot stylized face translation. Visual Comput. pp. 1–15 (2022)
16. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
17. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. Adv. Neural Inf. Process. Syst. **29** (2016)
18. Tang, H., Xu, D., Sebe, N., Yan, Y.: Attention-guided generative adversarial networks for unsupervised image-to-image translation. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2019)
19. Li, J., Zeng, H., Peng, L., Zhu, J., Liu, Z.: Learning to rank method combining multi-head self-attention with conditional generative adversarial nets. Array **15**, 100205 (2022)
20. Heo, Y.J., Kim, B.G., Roy, P.P.: Frontal face generation algorithm from multi-view images based on generative adversarial network. J. Multimed. Inf. Sys. **8**(2), 85–92 (2021)
21. Ruoyao, L., Bo, Z., Bin, W.: Remote sensing image scene classification based on multi-layer feature context coding network. J. Infrared Millim. Waves **40**(4), 530 (2021)
22. Wang, Q., He, X., Li, X.: Locality and structure regularized low rank representation for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **57**(2), 911–923 (2018)
23. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(10), 1915–1926 (2011)

24. Li, T., Song, H., Zhang, K., Liu, Q.: Learning residual refinement network with semantic context representation for real-time saliency object detection. Pattern Recogn. **105**, 107372 (2020)
25. Li, C., Yuan, Y., Cai, W., Xia, Y., Dagan Feng, D.: Robust saliency detection via regularized random walks ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2710–2717 (2015)
26. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166–3173 (2013)
27. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 353–367 (2010)
28. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2083–2090 (2013)
29. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer vision, pp. 8779–8788 (2019)
30. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3203–3212 (2017)
31. Li, C., Cong, R., Hou, J., Zhang, S., Qian, Y., Kwong, S.: Nested network with two-stream pyramid for salient object detection in optical remote sensing images. IEEE Trans. Geosci. Remote Sens. **57**(11), 9156–9166 (2019)
32. Zhang, L., Liu, Y., Zhang, J.: Saliency detection based on self-adaptive multiple feature fusion for remote sensing images. Int. J. Remote Sens. **40**(22), 8270–8297 (2019)
33. Hu, X., Fu, C.W., Zhu, L., Wang, T., Heng, P.A.: Sac-net: Spatial attenuation context for salient object detection. IEEE Trans. Circuits Syst. Video Technol. **31**(3), 1079–1090 (2020)
34. Das, D.K., Shit, S., Ray, D.N., Majumder, S.: Cgan: closure-guided attention network for salient object detection. Vis. Comput. **38**(11), 3803–3817 (2022)
35. Yu, Y., Gu, T., Guan, H., Li, D., Jin, S.: Vehicle detection from high-resolution remote sensing imagery using convolutional capsule networks. IEEE Geosci. Remote Sens. Lett. **16**(12), 1894–1898 (2019)
36. Yu, Y., Wang, J., Qiang, H., Jiang, M., Tang, E., Yu, C., Zhang, Y., Li, J.: Sparse anchoring guided high-resolution capsule network for geospatial object detection from remote sensing imagery. Int. J. Appl. Earth Obs. Geoinf. **104**, 102548 (2021)
37. Janakiramaiah, B., Kalyani, G., Karuna, A., Prasad, L., Krishna, M.: Military object detection in defense using multi-level capsule networks. Soft Comput. pp. 1–15 (2021)
38. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Proceedings of Advances in Neural Information Processing Systems, pp. 3856–3866 (2017)
39. Hinton, G.E., Sabour, S., Frosst, N.: Matrix capsules with em routing. In: International Conference on Learning Representations (2018)
40. Feng, Y., Gao, J., Xu, C.: Learning dual-routing capsule graph neural network for few-shot video classification. IEEE Transactions on Multimedia (2022)
41. Liu, Y., Zhang, D., Zhang, Q., Han, J.: Part-object relational visual saliency. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
42. Mazzia, V., Salvetti, F., Chiaberge, M.: Efficient-capsnet: Capsule network with self-attention routing. Sci. Rep. **11**(1), 1–13 (2021)
43. Park, H.J., Choi, Y.J., Lee, Y.W., Kim, B.G.: ssfpn: Scale sequence (s^2) feature based feature pyramid network for object detection. arXiv preprint arXiv:2208.11533 (2022)
44. Chen, T., Xiao, J., Hu, X., Zhang, G., Wang, S.: Boundary-guided network for camouflaged object detection. Knowl.-Based Syst. **248**, 108901 (2022)
45. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006)
46. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Adv. Neural Inf. Process. Syst. **30** (2017)
47. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1365–1374 (2019)
48. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint **2**(7) arXiv:1503.02531 (2015)
49. Jia, B., Huang, Q.: De-capsnet: a diverse enhanced capsule network with disperse dynamic routing. Appl. Sci. **10**(3), 884 (2020)
50. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
51. Zhang, Q., Cong, R., Li, C., Cheng, M.M., Fang, Y., Cao, X., Zhao, Y., Kwong, S.: Dense attention fluid network for salient object detection in optical remote sensing images. IEEE Trans. Image Process. **30**, 1305–1317 (2020)
52. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3917–3926 (2019)
53. Yuan, Y., Li, C., Kim, J., Cai, W., Feng, D.D.: Reversion correction and regularized random walk ranking for saliency detection. IEEE Trans. Image Process. **27**(3), 1311–1322 (2017)
54. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4548–4557 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Yuanfeng Lian is an associate professor of computer science in Beijing Key Laboratory of Petroleum Data Mining and College of Information Science and Engineering at China University of Petroleum, Beijing, China. He received his Ph.D. degree from Beihang University, China, in 2012 and M.S. degree from Changchun University of Technology, China, in 2003. His current research interests include computer vision and remote sensing images processing.

Xu Shi is a M.Eng. candidate in the College of Information Science and Engineering in China University of Petroleum. He received his B.Eng. degree in the Software Engineering from Beijing Technology and Business University, China, in 2020. His current research interests include saliency detection and remote sensing images processing.

Shaochen Shen received his M.Eng. degree in the Computer Technology in China University of Petroleum, China, in 2019. He received his B.Eng. degree in the Computer Science and Technology from Hebei University, China, in 2015. His current research interests include computer vision.

Jing Hua is a Professor of Computer Science and the founding director of Computer Graphics and Imaging Lab (GIL) and Vision Lab (VIS) at Computer Science at Wayne State University (WSU). He received his Ph.D. degree (2004) in Computer Science from the State University of New York at Stony Brook. His research interests include computer graphics, visualization, image analysis and informatics, computer vision, etc. He has authored over 100 papers in the above research fields. He received the Gaheon Award for the Best Paper of International Journal of CAD/CAM in 2009, the Best Paper Award at ACM Solid Modeling 2004 and the Best Demo Awards at GENI Engineering Conference 21 (2014) and 23 (2015), respectively. His research is funded by the National Science Foundation, National Institutes of Health, Michigan Technology Tri-Corridor, Michigan Economic Development Corporation and Ford Motor Company.