remote sensing

MDPI

Article

A Cyclic Information–Interaction Model for Remote Sensing Image Segmentation

Xu Cheng [1,2,*], Lihua Liu [1,2] and Chen Song [1,2]

[1] School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; 20201220022@nuist.edu.cn (L.L.); 20191221015@nuist.edu.cn (C.S.)
[2] Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China
* Correspondence: xcheng@nuist.edu.cn

Abstract: Object detection and segmentation have recently shown encouraging results toward image analysis and interpretation due to their promising applications in remote sensing image fusion field. Although numerous methods have been proposed, implementing effective and efficient object detection is still very challenging for now, especially for the limitation of single modal data. The use of a single modal data is not always enough to reach proper spectral and spatial resolutions. The rapid expansion in the number and the availability of multi-source data causes new challenges for their effective and efficient processing. In this paper, we propose an effective feature information–interaction visual attention model for multimodal data segmentation and enhancement, which utilizes channel information to weight self-attentive feature maps of different sources, completing extraction, fusion, and enhancement of global semantic features with local contextual information of the object. Additionally, we further propose an adaptively cyclic feature information–interaction model, which adopts branch prediction to decide the number of visual perceptions, accomplishing adaptive fusion of global semantic features and local fine-grained information. Numerous experiments on several benchmarks show that the proposed approach can achieve significant improvements over baseline model.

Keywords: deep learning; image segmentation; transfer learning; remote sensing image

Citation: Cheng, X.; Liu, L.; Song, C. A Cyclic Information–Interaction Model for Remote Sensing Image Segmentation. *Remote Sens.* **2021**, *13*, 3871. https://doi.org/10.3390/rs13193871

Academic Editors: Thien Huynh-The and Sun Le

Received: 27 August 2021
Accepted: 23 September 2021
Published: 27 September 2021

1. Introduction

Recently, deep learning (DL) technology has attracted more and more attention in a variety of fields with satisfying results. The convolutional neural network (CNN), as one of the deep learning network models, has greatly promoted the progress of remote sensing technology. Remote sensing image analysis has been a hot topic, and it has been widely utilized in many fields such as urban planning, land-used management, and environmental surveillance. Many conventional approaches employed hand-crafted features to achieve object segmentation and tracking. However, these methods cannot achieve an important performance for the complicated appearance changes of the ground objects in very high resolution (VHR) aerial imagery. In the past decade, deep learning-based semantic segmentation has played an important part in the remote sensing applications. The block diagram of remote sensing systems framework is shown in Figure 1.

Deep learning-based remote sensing methods have many practical applications; the collected remote sensing data from sensors should be processed, and only the available informative could be recorded for future usages, such as event detection, human–computer interaction, video abstraction, object tracking [1], scene segmentation [2], urban management and planning [3], etc. Although much work has been done over recent years, effective object segmentation and tracking approaches for object segmentation and tracking in complex scenarios remain challenges.

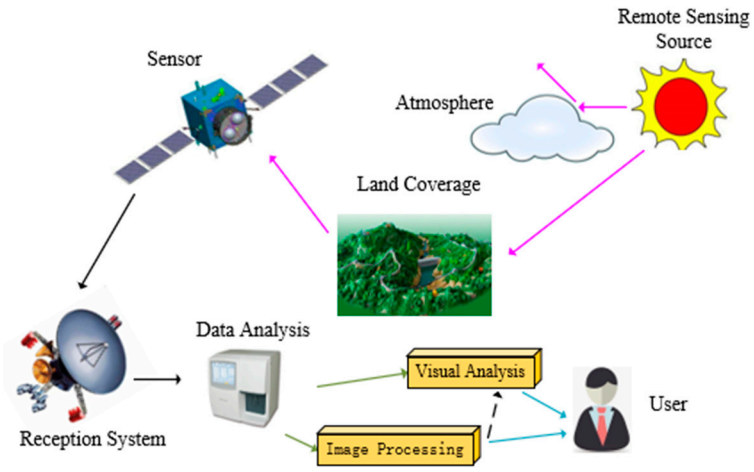


Figure 1. Block diagram of remote sensing systems framework.

Object segmentation and tracking technology are some of the important components in the field of computer vision. Deep learning (DL) technology and correlation filter (CF) technology have reported great potentials for real-time segmentation and tracking tasks. The original CF method performs excellent inference speed by utilizing element-wise multiplications in fast Fourier transform. However, the robustness of baseline approaches often drops considerably in several complex scenarios. In the last five years, many effective deep network architectures have been widely applied to remote sensing data processing and have become integral parts of our everyday lives. Currently, most existing DL methods rely heavily on the abundant training data. Additionally, the deep learning models are trained on large data offline via an end-to-end framework and aggressively learn the network model parameters online. These methods have shown promising applications in some challenging benchmarks.

However, there are two weaknesses for inaccurate object position prediction. One is the inadequate feature fusion of multilayer response maps. The other is the limitation of single modal. The use of single modal data is not always enough to reach proper spectral and spatial resolutions. Thus, multiple source data acquired by sensors onboard different platforms should be combined.

The growth of multimodal data poses a challenge for efficient data processing technology. The existing object detection deep network architectures are designed for various vision tasks. However, the targets contain few pixels in remote sensing images and exhibit arbitrary perspective transformations, thus many technical challenges are left open.

In this paper, we design a real-time convolutional framework for remote sensing data detection and segmentation. The proposed framework is capable of conducting mid-level fusion of multiple sources of data. Experiments were carried out on some common datasets. The contributions of the paper can be summarized as follows.

(1) We propose an effective feature information–interaction visual attention model for multimodal data fusion and enhancement, which utilizes channel information to weight self-attentive feature maps of multi-source data, completing extraction, fusion, and enhancement of global semantic feature with local contextual information of the object.

(2) To improve the effectiveness of multi-source feature fusion, we further develop an adaptively cyclic feature information–interaction model, which adopts branch prediction to decide the number of visual perceptions, accomplishing adaptive fusion of global semantic features and local fine-grained information.

(3) Our experiments reveal that the proposed appro6ach can provide competitive advantages with respect to baseline methods.

The rest of the paper is summarized as follows. In Section 2, we review related work. Section 3 introduces our approach for object segmentation tasks in detail. In Section 4,

we report results on some common datasets. Section 5 discusses the advantages of our approach. Section 6 summarizes this paper.

2. Related Works

Recently, extensive studies of object segmentation and tracking methods were surveyed. Deep learning based methods can produce satisfying results. In the following, we mainly review the key methods that related to our approach.

Deep learning approaches: Deep learning technology is utilized to enhance the robustness of visual tasks (e.g., segmentation, classification, and tracking). Some known methods combine DL models with CF to perform accuracy segmentation and tracking such as HCF (hierarchical convolutional features) [4], MOS (multiscale optimized segmentation) [5], DeepSRDCF (convolutional features for correlation filter) [6], ECO (efficient convolution operators) [7], benchmarking [8], SSCF (spatial semantic convolutional features) [9], RNT (residual network tracker) [10], GAN-RI (GAN re-identification) [11], and DRDN (deep residual dense network) [12]. Another approach utilizes classification and regression networks to formulate object segmentation and tracking tasks, such as SMIS (supervised methods image segmentation) [13], FCNT (fully convolutional networks) [14], DeepTrack [15], and CNN-SVM [16]. The benefit of the above methods is obviously that the high-level semantic features of deep network model are utilized to match objects. However, the computational complexity is not increased due to online update mechanisms of the object template.

In the last five years, several effective deep network models were trained on large classification datasets offline and employed to segment and track objects online, including MDNet (multi-domain network) [17], CFNet (correlation filter network) [18], ACFN (attentional correlation filter network) [19], etc. Recently, the Siamese network model [20–24] successfully solved the inaccurate performances. SINT (Siamese instance network tracker) [20] regards the tracking problem as a verification task and learns a similarity measure for object matching in each frame. The representative approaches contain SiamRPN++ [21], deeper and wider Siamese tracker [22], DaSiamRPN [23], and so on. In the VITAL method [25], hard samples are generated through utilizing adversarial learning, and an effective loss function is leveraged to address the class imbalance problem. These kinds of methods promote the development of deep learning models and obtain satisfying evaluations on several challenging datasets. However, most deep learning models suffer from under-fitting problems because of a lack of training samples.

Correlation filter approaches: The approaches based on the correlation filter framework achieved promising results between speed and accuracy [26–34]. They are classified into two categories: baseline methods and improved regularization methods.

Several baseline methods are presented to improve speed and accuracy by utilizing scale prediction [27], spatial regularization [28,35], and long-term tracking [29]. Initially, the MOSSE method was first introduced into object tracking by using a single feature channel. Then, Henriques et al. [26] proposed an effective kernelized tracking method (KCF) using the circular correlation solution scheme for ridge regression. Danelljan et al. [27] trained DCF using a scale pyramid representation to handle the scale variations of the object. However, baseline methods are constrained in the detection region due to the equality of patch size and filter size.

To solve above problems, some improved regularization methods were developed, which include SRDCF [28], DeepSRDCF [6], ECO [7], STRCF [30], C-COT [31], CSR-DCF [32], ATOM [33], MCPF [36], etc. In ACFN [19], a subset is chosen from the associated CFs as an attention scheme to improve the performance. To alleviate the unwanted boundary effects, they further propose a spatial constraint [28] to penalize the correlation filter coefficients during the training process. Several approaches combine correlation filters with high-level semantic features, which produces a remarkable advance in performance. CSR-DCF [32] exploits color histograms as features to obtain a saliency response map in the Fourier domain, which trains the attention network model in an end-to-end way.

Transfer learning approaches: There have been many efforts to utilize the transfer learning for processing remote sensing imagery [37]. The first successful application of these models to object tracking was presented by Wang et.al. [38]. They pre-trained a stacked denoising auto-encoder (SDAE) from an ILSVRC dataset and then transferred it to an object tracking task. Since then, some supervised transfer learning-based approaches have been presented to segment and track object. They offline-train deep models using other datasets as source domains and then use the learned model online to obtain satisfying accuracy in the target domain. However, the high computational cost is an obvious deficiency. Some transfer learning-based methods utilize the features from different deep network layers to improve tracking performance. Gao et al. [39] exploited the extracted prior knowledge from the Gaussian processes learning to improve the robustness. In [40], an effective offline-trained meta-updater was presented to achieve robust tracking performance, which consisted of an online local tracker, a meta-updater, a re-detector, and an online verifier in the long-term tracking framework.

Other approaches: Benedek et al. [41] proposed a novel object-change modeling approach based on multitemporal marked point processes, which simultaneously exploits low-level change information between the time layers and the object-level building description to recognize and separate changed and unaltered buildings. Xu et al. [42] developed an image segmentation neural network based on the deep residual networks and used a guided filter to extract buildings in remote sensing imagery. Grinias et al. [43] proposed a novel segmentation algorithm based on a Markov random field model and obtained good classification performance. Shi et al. [44] constructed a convolutional network based on a generative adversarial network to discriminate between ground truth maps and generated maps by the segmentation model.

3. Methodology

In this section, we first introduce the problem and the motivation and then give a detailed description of deep network architecture and channel attention. Finally, we apply our deep network architecture to object segmentation and tracking.

3.1. Problems and Motivations

In the remote sensing field, the Single Shot MultiBox Detector (SSD) [45] is one of the most representative detection methods with respect to speed and accuracy trade-off. Nevertheless, some drawbacks limit the accuracy of the algorithm. First, the semantic information of shallow layers is weak, and it fails to capture global dependent information to predict small and dense clusters of objects in RS images. Second, the feature maps from medium layers present the problem of feature confusion, which makes it difficult to accurately regress bounding boxes. Finally, the deep layers have less object contextual information, making it fail to predict large objects confidently.

Inspired by information guidance between self-attentive models [37], we propose a feature information–interaction model, which introduces feature map channel weights on the self-attentive module and takes a weighted mechanism to focus on the regional block. On this basis, an adaptively cyclic information–interaction visual model is developed to solve the problem of insufficient feature fusion, which concentrates on the feature map more than once to distinguish the background clutter.

3.2. Feature Information–Interaction Model

As mentioned above, the existing self-attentive models associate the internal information of feature maps and concentrate on the local information of the object, ignoring the inter-channel feature information association, i.e., the global semantic feature information of the object. To tackle the problem, we propose a feature information–interaction model (FIM), where weighted channels of feature maps are proposed to perceive the global semantic and the local fine-grained features of the object. The overall structure of FIM is shown in Figure 2.

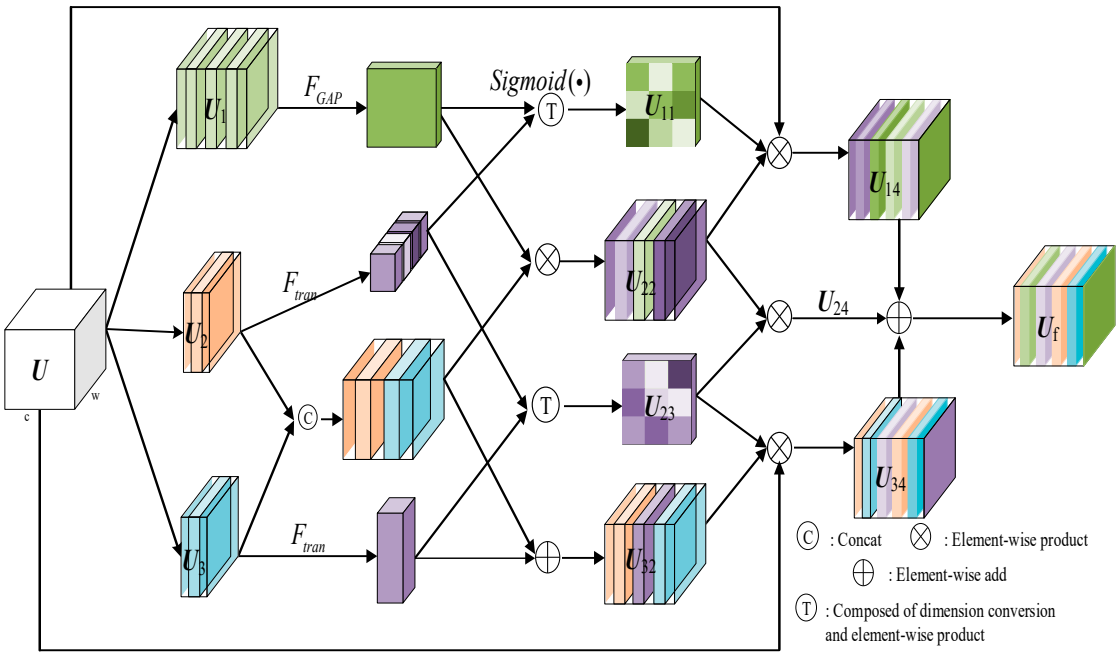


Figure 2. Feature information–interaction model.

Given the feature map, $\boldsymbol{U}\in\mathbf{R}^{H\times W\times C}$, $\boldsymbol{U}_1\in\mathbf{R}^{H\times W\times C}$, $\boldsymbol{U}_2\in\mathbf{R}^{H\times W\times C/8}$, and $\boldsymbol{U}_3\in\mathbf{R}^{H\times W\times C/8}$ are obtained through the convolution operator. The channel attention and the self-attentive modules are achieved by these feature maps. For the channel attention, global average pooling and sigmoid are utilized to get the feature map of the weighted channel, and the resulting feature map is $\boldsymbol{U}_{12}\in\mathbf{R}^{1\times 1\times C}$. For the self-attention module, we utilize dimension transformation to obtain the intermediate response map, and the weighted channel supervision information is utilized to assist the intermediate self-attention feature map.

$\boldsymbol{U}_{22}\in\mathbf{R}^{H\times W\times C}$, $\boldsymbol{U}_{23}\in\mathbf{R}^{1\times 1\times C}$, and $\boldsymbol{U}_{32}\in\mathbf{R}^{H\times W\times C}$ are obtained to represent the enhanced feature maps with global semantic features and local fine-grained information. Additionally, we merge the original feature map $\boldsymbol{U}$ into enhanced feature maps for enriching semantic feature information. Therefore, the adaptively weighted attention information can be obtained through threshold multiply and add operations in Equation (1).

$$\boldsymbol{U}_{14} = (1+\alpha)(\boldsymbol{U}_{12}\otimes\boldsymbol{U}_{22})\oplus(1-\alpha)\boldsymbol{U} \tag{1}$$

where $\boldsymbol{U}_{14}$ is the enhanced feature map; $\boldsymbol{U}_{12}$, $\boldsymbol{U}_{22}$, and $\boldsymbol{U}$ are the intermediate layer information, and α denotes the predicted threshold through convolution; $\oplus$ and $\otimes$ stand for element-add and element-multiply operations, respectively. Then, self-attentive feature maps can be obtained in spatial dimension by using Equations (2) and (3).

$$\boldsymbol{U}_{24} = \alpha\boldsymbol{U}_{22}\oplus\beta\boldsymbol{U}_{23} \tag{2}$$

$$\boldsymbol{U}_{34} = (1+\alpha)(\boldsymbol{U}_{23}\otimes\boldsymbol{U}_{32})\oplus(1-\alpha)\boldsymbol{U} \tag{3}$$

where $\boldsymbol{U}_{24}$ and $\boldsymbol{U}_{34}$ are the enhanced feature maps; $\boldsymbol{U}$, $\boldsymbol{U}_{22}$, $\boldsymbol{U}_{23,}$ and $\boldsymbol{U}_{32}$ are the intermediate layer information. α and β are the predicted thresholds, and we set $\alpha + \beta = 1$ empirically. Finally, the threshold weighting scheme is utilized to generate the resulting feature map through Equation (4).

$$\boldsymbol{U}_{\mathrm{f}} = \alpha(\boldsymbol{U}_{14})\oplus\beta(\boldsymbol{U}_{24})\oplus\gamma(\boldsymbol{U}_{34}) \tag{4}$$

where $\boldsymbol{U}_{\mathrm{f}}$ is the resulting feature map; $\boldsymbol{U}_{14}$, $\boldsymbol{U}_{24}$, and $\boldsymbol{U}_{34}$ are the intermediate enhanced features; α, β, and γ are adaptive thresholds.

3.3. Adaptively Cyclic Feature Information–Interaction Model

To enrich the global semantic feature and the local contextual information of the object, we propose an adaptively cyclic information–interaction model (ACFIM) to strengthen the ability of feature extraction. Concretely speaking, the convolutional prediction module is developed to control the location and the number of visual perceptions and adaptively concentrates on the feature map more than once, better distinguishing the similarities and the differences between objects.

Based on our knowledge and experience, the feature maps of the shallow layer have enriched local fine-grained information of the object, and the feature maps of the deep layer contain abundant global semantic features of the object.

Therefore, cycle times and locations of feature fusion between shallow, medium, and deep layers are different. For the feature maps of the shallow layer, we set three times default for the cyclic model to better enrich and represent the global feature flow of the object. For the feature maps of the medium and the deep layers, we set two times default for the cyclic model for fine-tuning the information flow between low-level fine-grained features and high-level semantic features. In addition, we set an intermediate threshold δ for deciding the cyclic location to better adapt the enhanced feature map. If δ is less than a predefined threshold (0.5 in the paper), we merge the enhanced feature map into the original feature map for loop initialization. Otherwise, we merge it into the intermediate feature map for information fusion and initialization.

3.4. Objective Loss Function

The overall loss function of SSD is defined as a weighted sum of the confidence loss and the localization loss; more detailed 6information can be referenced in [15]. In our model, we adopt focal loss for confidence to address the problem of class imbalance. In addition, we slightly adjust the parameters between the default anchors and the ground-truth boxes, as shown in Equations (5) and (6).

$$L_{loc}(x, p, g, d) = \sum_{i \in pos}^{N} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^{k} L_1(p_i^m - k_j^m) \tag{5}$$

$$\begin{aligned} k_j^{cx} &= (g_j^{cx} - d_i^{cx})/d_i^{w}/\mathrm{var}^1 \\ k_j^{cy} &= (g_j^{cy} - d_i^{cy})/d_i^{h}/\mathrm{var}^1 \\ k_j^{w} &= \log(\tfrac{g_j^{w}}{d_i^{w}} + \mathrm{var}^2) \\ k_j^{h} &= \log(\tfrac{g_j^{h}}{d_i^{h}} + \mathrm{var}^2) \end{aligned} \tag{6}$$

3.5. Online Segmentation and Tracking

Given the object location of the first frame with annotation, we construct an initial training set by using a data augmentation scheme, which includes 20 positive samples. Then, the backbone model is fine-tuned with training samples of the first frame.

The appearance of the object in the process of object data processing may change. To capture the appearance variations, we update the object template with the previous video observations. First, we define a fixed length unit L to store the object state at every frame and update the object template if the length unit reaches a fixed number of elements. The element with the maximum saliency score in the length unit is utilized to update the template of the object.

Thus, the updated template is expressed as:

$$\widetilde{\mathbf{c}}_{\mathrm{u}} = (1 - \eta)\mathbf{c}_{\mathrm{f}} + \eta \mathbf{c}_{\mathrm{p}} \tag{7}$$

where η denotes an empirical learning parameter; $\widetilde{\mathbf{c}}_{\mathrm{u}} \in \mathbf{R}^{kn \times 1}$ stands for a new object template, which consists of the initial object template $\mathbf{c}_{\mathrm{f}} \in \mathbf{R}^{kn \times 1}$ and the last updated

template $\mathbf{c}_p \in \mathbf{R}^{kn \times 1}$. The feature of the object is concatenated into a column vector as the initial template at the original frame. Then, the initial template of the object combines with the new updated template to alleviate the drift of the object.

In the test stage, in each frame, we crop several search regions centered at the object location of the last frame by using the multiple scales scheme. Then, these search regions are inputted to the ResNet-50 network model to extract the object features. The fine-grained features of the object are considered as the CF layer. The high-level object feature is obtained according to the channel attention module. Furthermore, we randomly draw the object candidates $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ based on the object location in the previous frames. Finally, the candidate states in search region are computed by Equation (8).

$$\mathbf{x}_t^* = \underset{j=1,\ldots,N}{\operatorname{argmax}} S(x_j) \tag{8}$$

The object states with the highest response value are regarded as the final tracking result. Then, we update the object template by using training samples obtained in previous sequences in every 20 frames.

4. Results

4.1. Implementation Details

In the study, we implemented our approach on the framework of Pytorch with an NVIDIA GTX 2080Ti GPU and an Intel i7 CPU (32G RAM) and utilized SSD as our baseline model. The proposed visual attention model was embedded into the extra four prediction modules. During the training stage, without bells and whistles, we followed the original SSD strategies, which included data augmentation, backbone network, scale, and aspect ratios for the predefined anchors. In the paper, the learning rate schedule was slightly changed to obtain a better performance.

4.2. PASCAL VOC2007

The PASCAL VOC dataset [46] contains 20 object categories. The pixel size of the image varies and is usually (horizontal view) 500 × 375 or (longitudinal view) 375 × 500. The mean average precision was used to measure the performance of the object detection network (http://host.robots.ox.ac.uk/pascal/VOC/) (accessed on 14 May 2021).

We trained our model on the PASCAL VOC2007 dataset and the VOC2012 trainval set and tested our model on the VOC2007 test set. We utilized a 10^{-3} learning rate for the first 80,000 iterations, then decreased it to 10^{-4} for the next 20,000 iterations and 10^{-5} for the remaining 20,000 iterations. In addition, we adopted a "warmup" strategy that gradually ramped up the learning rate, which contributed to stabilizing the training process. The momentum and the weight decay were set to 0.9 and 0.0005, respectively.

Table 1 shows experimental results on the PASCAL VOC test set, which were trained with VOC07 trainval and VOC12 trainval sets. The proposed approach obtained 79.7% mAP with 300 × 300 input images and 82.1% mAP with 512 × 512 input images, exceeding the latest SSD300* by 2.2 points and SSD512* by 2.6 points.

Table 1. Results on PASCAL VOC2007 test set (07 + 12 denotes data are trained with VOC07 and VOC12 trainval sets).

Method	Data	mAP
SSD300*	07 + 12	77.5
SSD512*	07 + 12	79.5
Ours300	07 + 12	79.7
Ours512	07 + 12	82.1

In Table 2, we compare the proposed method with other methods under the same baseline model. For fairness and simplicity, we simply replaced our module with other

visual attention models. Our method considers object feature information–interaction under visual perception and obtains the best accuracy among all models. Without bells and whistles, the baseline model SSD with FIM achieved 79.48% mAP on the VOC2007 test set, which proves the effectiveness of our method which concentrates on semantic feature and contextual information interaction.

Table 2. Comparison of mAP for different attention models (07 + 12 denotes data are trained with VOC07 and VOC12 trainval sets).

Attention Model	Data	Feature Fusion	mAP
Non-local Network [37]	07 + 12	Element-wise sum	78.12
GCNet [40]	07 + 12	Element-wise sum	78.11
BAM [47]	07 + 12	Element-wise sum	77.61
CBAM [48]	07 + 12	Element-wise sum	77.90
SKNet [49]	07 + 12	Element-wise sum	78.16
FIM (Ours)	07 + 12	Element-wise sum	79.48
ACFIM (Ours)	07 + 12	Element-wise sum	79.70

4.3. MS COCO

The COCO dataset [50] is a large, rich object detection, segmentation, and subtitle dataset. This dataset is mainly extracted from complex daily scenes. Targets in images are calibrated by accurate segmentation. Images include 91 categories of objects, 328,000 images, and 2500,000 labels. Thus far, the largest dataset with semantic segmentation provides 80 categories, more than 330,000 images, among which 200,000 are annotated (http://cocodataset.org/#home) (accessed on 15 May 2021).

To further verify the effectiveness of the proposed method, we trained our model on MS COCO. We utilized the trainval35 k (118,287 images) for training and evaluated the results on the minival. The batch size was set to 32 for 300 × 300 input and 16 for 512 × 512 input. We trained the model with 10^{-3} for the first 280,000 iterations, then 10^{-4} and 10^{-5} for the remaining 120,000 and 40,000 iterations. In Table 3, we observe that our method achieved 27.6% AP@[0.5:0.95], 46.8% AP@0.5, and 28.7% AP@0.75, which improved the baseline model SSD300* by 2.5, 3.7, and 2.9 points, respectively. Our model with 512 × 512 input images also outperformed the baseline SSD512*.

Table 3. Results on MS COCO benchmark.

Methods	Average Precision, IOU			Average Precision, Area		
	0.5:0.95	0.5	0.75	S	M	L
SSD300*	25.1	43.1	25.8	6.6	25.9	41.4
SSD512*	28.8	48.8	30.3	10.9	31.8	43.5
Ours (300)	27.6	46.8	28.7	8.9	29.6	42.8
Ours (512)	30.9	51.3	32.8	13.0	35.2	46.4

It is noticeable that our model with 300 × 300 and 512 × 512 input images achieved 8.9% AP and 13.0% AP for small objects, respectively. The proposed method is more powerful in detection of small objects. For medium and large objects, our method validates the effectiveness of the feature information–interaction scheme.

4.4. HRSC2016

To further verify the validity of our approach, we conducted the experiments on a remote sensing dataset collected from Google Earth and harbored it with the complex scenario, which is called high resolution ship collections 2016 (HRSC2016). On this dataset, the sizes of the image are between 300 × 300 and 1500 × 900, and the image resolutions range from 0.4 to 2 m. In addition, the inclined bounding box and te horizontal bounding box are provided as ground truth for each ship.

Figure 3 shows some representative detection approach evaluations when the threshold was set to 0.6. The experimental report shows the advantages of our approach in dense ship detection. Multiple-source data were fused, which made the fused features have a strong discriminative ability and overcame the limitation of a single modal. The proposed model achieved a higher accuracy and better generalization. In other words, our approach is more robust to serious background clutter and fragmentary ships by using multiple sources of fusion information.

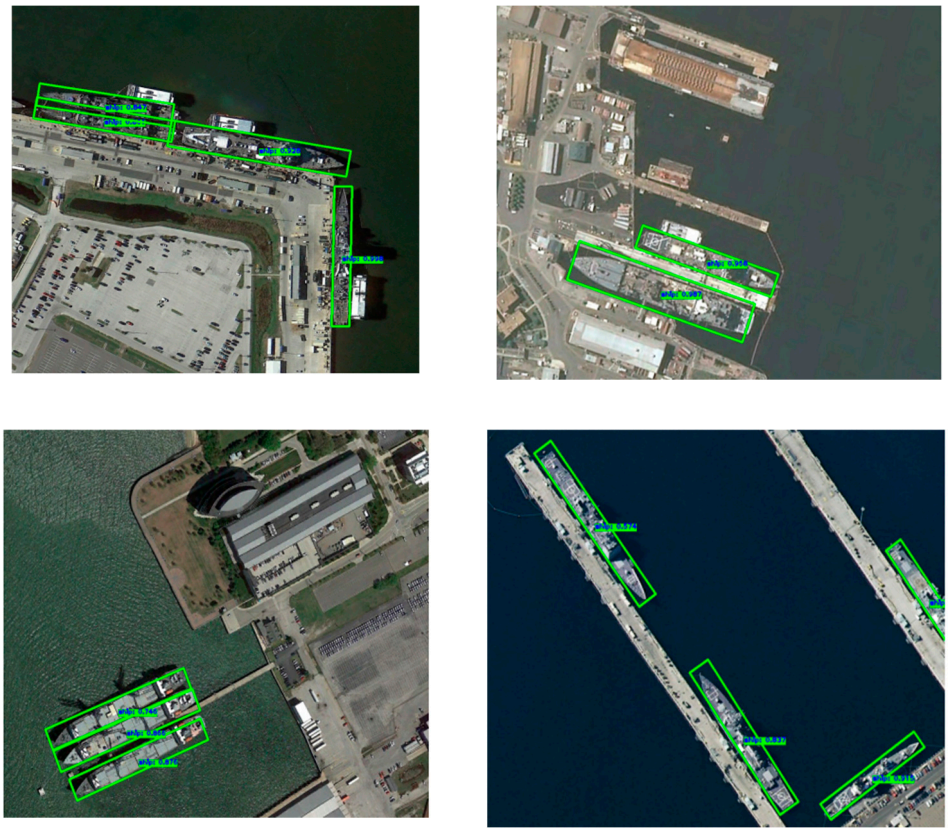


Figure 3. Multimodal data fusion results of the proposed method on the HRSC 2016 dataset.

4.5. LaSOT

The LaSOT dataset [51] is a single target tracking dataset with 1400 video sequences; each video has an average of 2512 frames, where the shortest video has 1000 frames, and the longest contains 11,397 frames. It is divided into 70 categories, each consisting of 20 video sequences (https://cis.temple.edu/lasot/) (accessed on 24 February 2021).

Our approach was also evaluated on the LaSOT benchmark including 280 videos, and there was an average of 2500 frames in the dataset, making the appearance change of an object an important challenge. Figure 4 reports the evaluation results.

The ATOM tracking method uses the ResNet-18 network model to separate the object from the image. The experiment evaluation results show that our method achieved an AUC score of 51.6% and had a lower failure rate of 15.1% while obtaining compatible robustness. Figure 5 gives the qualitative segmentation results of our method and other competing methods.

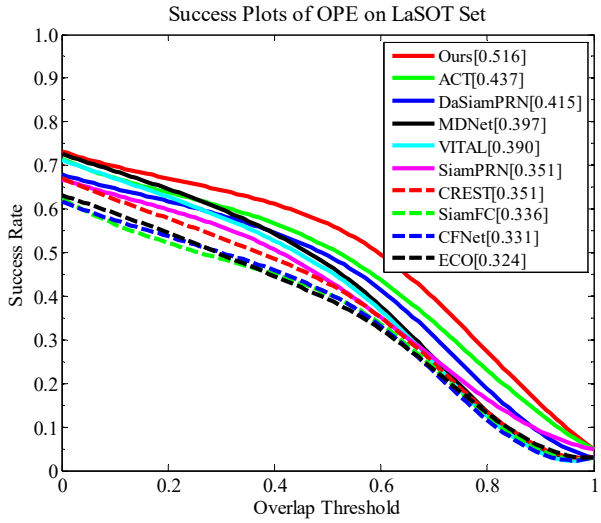


Figure 4. Success plot of OPE on LaSOT benchmark.

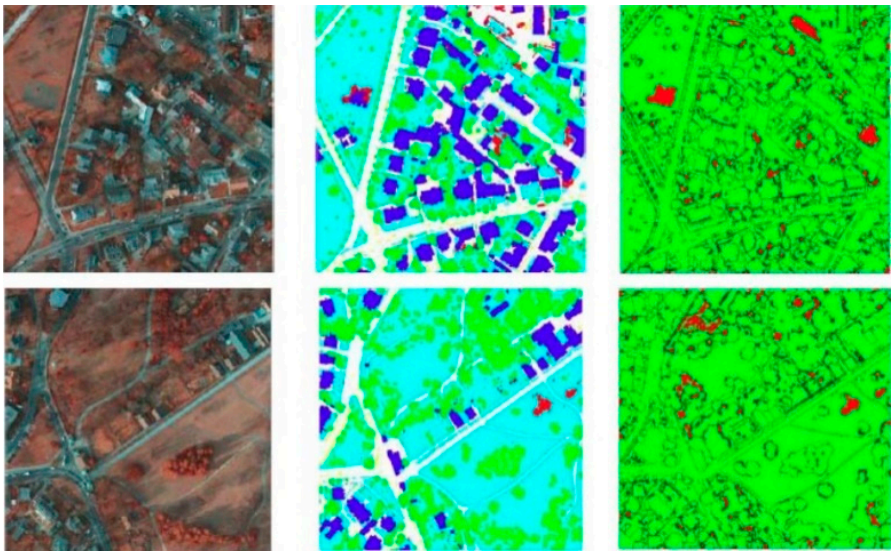
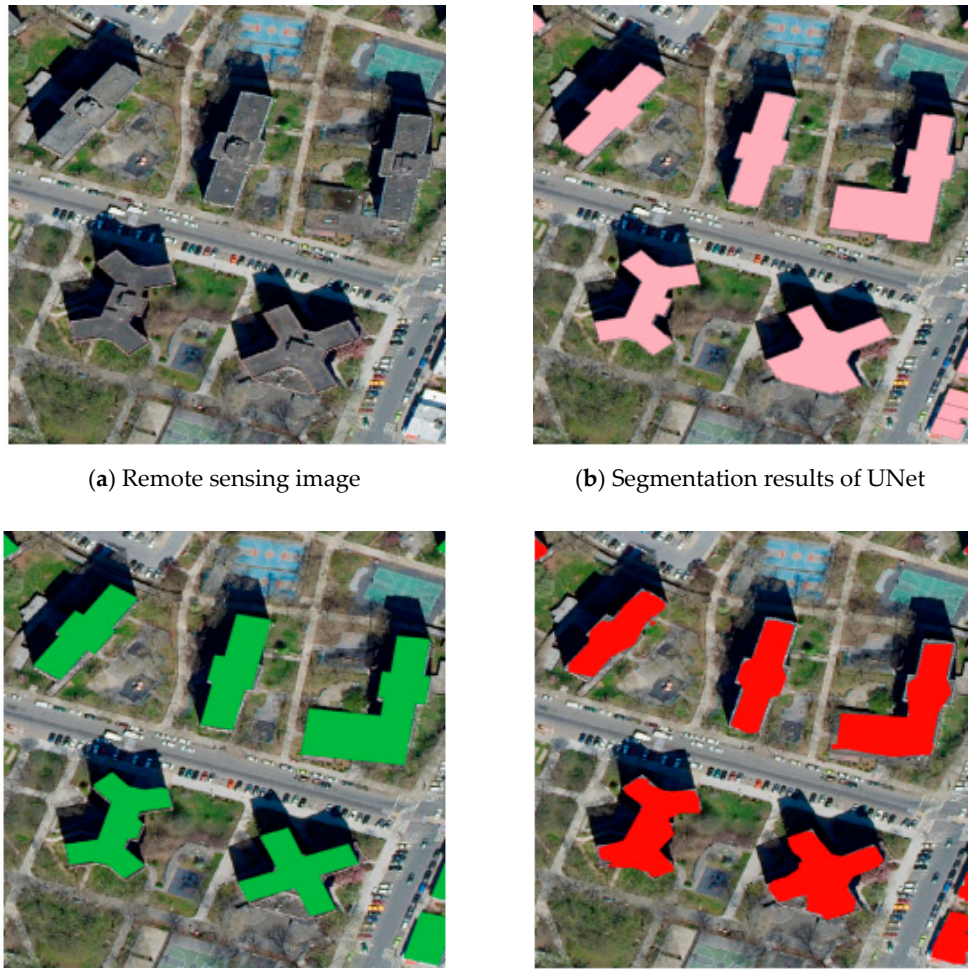


Figure 5. Qualitative segmentation results of our method and other competing methods.

4.6. Visualization Analysis

The comprehensive experiments were implemented on the ISPRS dataset. The dataset contains 38 images and consists of a true orthophoto (TOP) obtained from a larger TOP mosaic and is divided into six land cover classes. On this dataset, 24 classification label images are provided. The ground truth of the remaining scenes remains unreleased, and the benchmark is used for test verification. We utilized 15 images for training and nine images for testing. The network was trained by a data augmentation strategy that is a major method with rotation and scale variations of images.

Figure 6 reports the visualization results. The first column is the original and the resized true orthoimages for fair evaluation. The second column is the segmented output of our proposed approach. The last column indicates wrongly classified pixels via red/green image.



(**a**) Remote sensing image (**b**) Segmentation results of UNet

(**c**) Segmentation results of ResNet50 (**d**) Segmentation results of ours

Figure 6. Representative evaluation of our approach and other competing approaches.

Figure 7 illuminates the results of the proposed method using some challenging videos. In the first row of sequences, the object experienced illumination variation and scale changes. We can see that the proposed approach was able to cope with these challenging factors and kept in touch with the object successfully, which was attributed to our approach using both the transfer learning model updating strategy and the attention mechanism. However, other methods failed to match the object during the tracking process due to illumination changes and scale changes. MDNet suffered from the illumination changes and gradually missed the object. ECO did not perform well at the 78th frame.

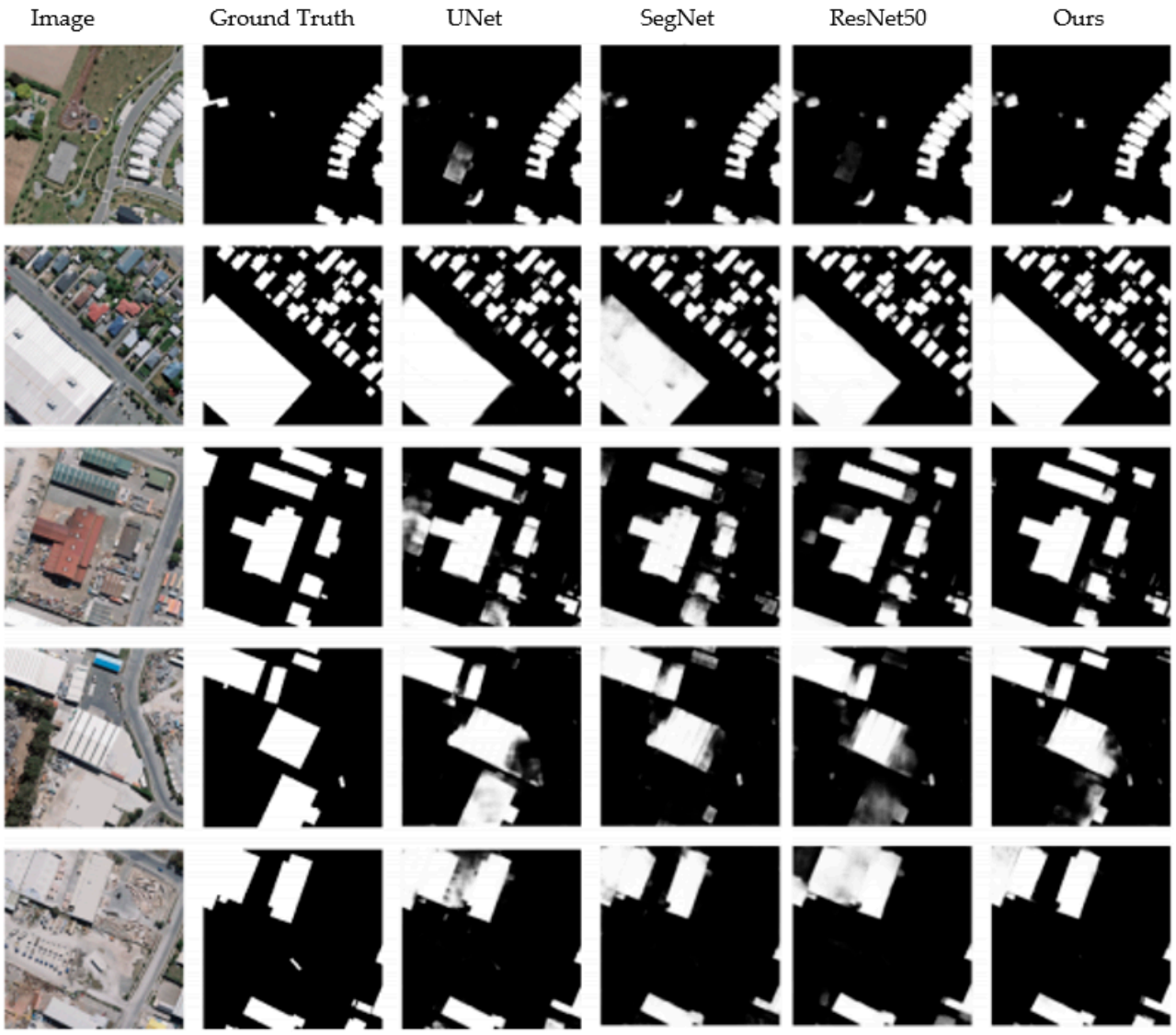


Figure 7. Some representative results of comparisons on the benchmark.

In the second and the third rows, these sequences encountered the challenges of scale variation, occlusion, and low contrast. These challenging factors greatly increased the difficulty for robust tracking. We can see that the proposed approach performed more robustly compared with other trackers, which drifted away from the object due to scale variations and occlusion. The proposed method performed well when the object was occluded in a complex scenario.

In the rest of the video sequences, most of the trackers missed the object and drifted. Our algorithm performed with the better accuracy in this scenario. This was mainly due to the proposed transfer learning model updating strategy and the attention module, which made the learned network concentrate on the robust object features and reduced the influence of background clutter within the image area. Overall, the proposed approach could track the object well in these challenging sequences.

4.7. Quantitative Evaluation

In order to further verify the effectiveness and the feasibility of the proposed method, we carried out the quantitative evaluation by calculating the accuracy evaluation indicators on the test set. The accuracy evaluation indicators included the precision ratio (PR), the recall ratio (RR), and the F_1-score.

The PR is the ratio of true positives to the sum of true positives and false positives, which is defined as:

$$PR = \frac{TP}{TP + FP} \tag{9}$$

The RR represents the ratio of true positives to the sum of true positives and false negatives, which is written as:

$$RR = \frac{TP}{TP + FN} \tag{10}$$

The F_1 score integrates PR and RR. The higher the F_1 score is, the better is the result of the model prediction.

$$F_1 = \frac{2 \times PR \times RR}{PR + RR} \tag{11}$$

where TP (true positive) and TN (true negative) denote the total number of object pixels and non-object pixels correctly predicted, respectively. FP (false positive) and FN (false negative) denote the total number of pixels with an incorrect outcome from the object and the non-object regions, respectively. Total denotes the total number of pixels. The precision and the recall measures both range from 0 to 1. An F_1 score reaches its best value at one and its worst at zero.

We compared our method with state-of-the-art methods, including GAN [44], FCN [52], and SegNet [53] on the ISPRS dataset. The evaluation results are reported in Table 4. Our method outperformed all compared methods on the dataset. We can see that all evaluation indicators of the proposed method improved compared to state-of-the-art methods. The main reason is that the proposed method benefits from the feature information–interaction model (FIM). By introducing FIM, our method weighted channels of feature maps to perceive the global semantic and the local fine-grained features of the object. Moreover, these deep learning methods can make decisions at multiple layers to improve the accuracy.

Table 4. A comparison between the proposed method with the existing models.

Methods	*PR*	*RR*	F_1
GAN [44]	0.9310	0.8544	0.8616
FCN [52]	0.9326	0.8645	0.8862
SegNet [53]	0.9499	0.9011	0.9249
Ours	0.9618	0.9307	0.9470

The proposed method achieved state-of-the-art results on the Vaihingen and the Potsdam datasets in Table 5. It can be clearly observed that the results support the idea that it is beneficial to use the cyclic feature information–interaction model.

Table 5. The average accuracy for precision, recall, as well as F1 score for buildings in Potsdam and Vaihingen (OA denotes overall accuracy).

Dataset	OA	*PR*	*RR*	F_1
Postdam	0.9699	0.9623	0.9167	0.9406
Vaihingen	0.9786	0.9613	0.9446	0.9534

5. Discussion

We further carried out the evaluation experiments to explain the contributions of different modules and different layer features. The AUC scores are reported using different backbone networks in Table 6.

Table 6. Ablation experiments of our approach on two datasets. Fine-tune denotes that the network model is offline-trained.

BackBone	Fine-Tune	Conv5	Conv4	VOT2018	OTB-100
AlexNet				0.355	0.666
ResNet50	√		√	0.347	0.679
	√	√		0.337	0.675
ResNet50		√	√	0.392	0.676
	√	√	√	0.408	0.700

Feature selection: Different layer features play a significant part in tracking tasks. We found that the type of network layers and the number of parameters directly affected the tracking performance. First, ResNet-50 and AlexNet networks were considered as backbone networks to evaluate the accuracy of the proposed approach on two popular benchmarks. The proposed approach and SiamRPN++ exhibited stronger performance, benefiting from the deeper learning model. In other words, our approach achieved an obvious improvement by fine-tuning network parameters. In addition, the evaluation outputs report that conv4 alone achieved the satisfying performance with 0.347 in EAO. Low-level and high-level features from deep network architecture performed with 5% drops. Unsurprisingly, significant improvement could be obtained through combining Conv4 and Conv5.

Effectiveness of different components: Our method includes CF module, S module, and A module, and they denote CF layer, SiamRPN, and channel attention component, respectively. To verify the effectiveness of these modules, the variants of the proposed approach were implemented: (1) ours (S) denotes the tracking method only utilizing SiamRPN to predict the location of the object at each frame; (2) ours (CF) is the proposed approach by combining the shallow layer CF and the deep features representation to predict the object state at each frame; (3) ours (A) stands for our method with an attention scheme; and (4) ours (S + CF + A) denotes our proposed approach in the paper. The contribution of different modules is reported in Table 7.

Table 7. Contribution of different modules for the proposed approach.

Approach	Ours (S)	Ours (CF)	Ours (A)	Ours (S + CF + A)
F-score	0.553	0.583	0.597	0.603
Pr	0.551	0.584	0.607	0.613
Re	0.541	0.557	0.565	0.596
FPS	34.7	30.6	21.4	24.8

Table 8 reports the evaluation performance of the variants and verifies that all components improved the tracking accuracy. Removal of the channel attention module from our approach led to a 3.1% precision drop; the precision of the variant dropped by 7.8% without the correlation filter layer. We can see that the performances of both variants were comparable to the S module, but the failure examples increased during tracking and segmentation processing. This is because the channel attention module could focus on the important part of the image. Thus, the attention scheme was very important to achieve reliable tracking and segmentation. Our method resulted in a 9.5% EAO and an 8.7% accuracy improvement due to the rotated bounding box estimation.

Table 8. The impact of different loss terms on five benchmarks.

	$L_{SiamRPN}$	L_{low}	L_{high}	All
AUC	0.607	0.579	0.586	0.619
Precision	0.878	0.842	0.851	0.887

Impact of different loss function terms: We compared different loss function terms, and their impacts are shown in Figure 8. It is quite clear that every loss term made its contribution to the performance of our approach. Meanwhile, Table 8 illustrates the performance of every loss term by showing that our approach outperformed every variant. The loss function indicates the importance of end-to-end training.

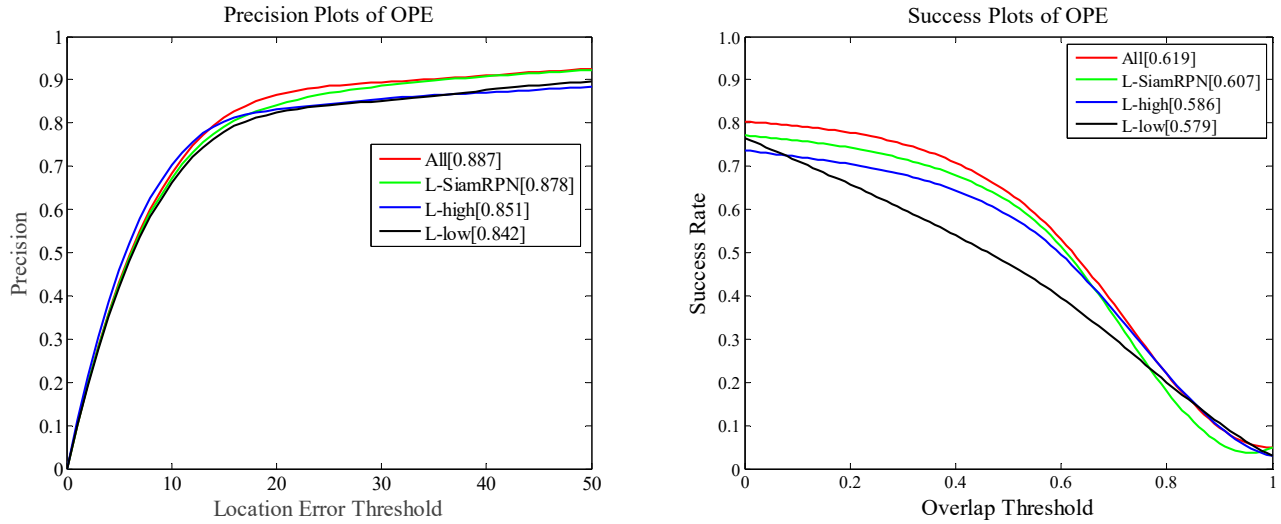
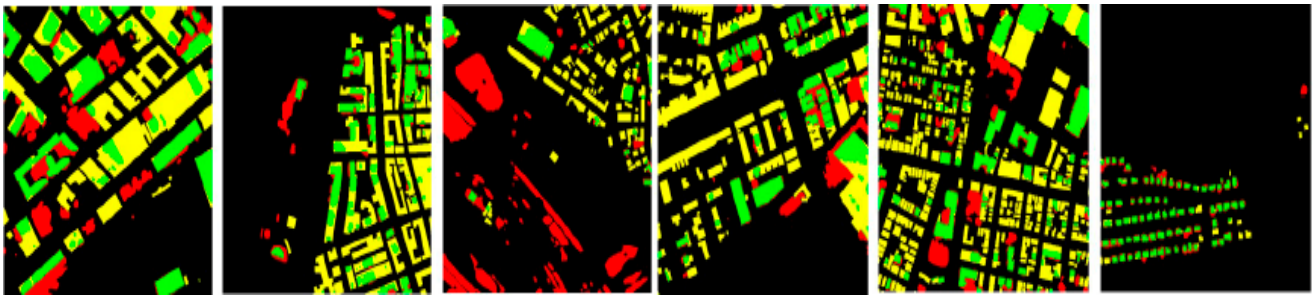


Figure 8. The impact of different losses on five benchmarks.

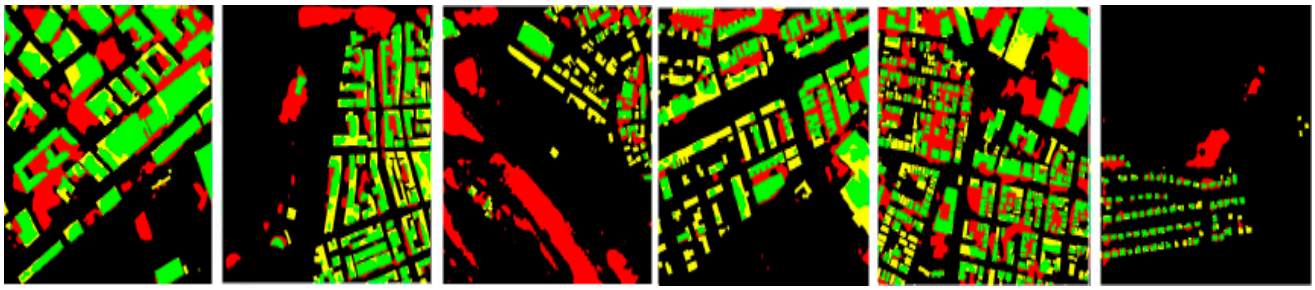
Figure 9 reports several building segmentation results of the UNet [54] and the ResNet-50. The yellow, the green, and the red pixels denote "false negative", "true positive", and "false positive", respectively. As shown in Figure 10, the flyover was wrongly labeled as building by using the ResNet50 and the UNet network models. We can see that the proposed method could remove most false alarms due to the introduction of the attention mechanism into the deep network model, which resulted in generating more precise segmentation results; this indicates the integration of the attention channel module into the remote sensing image process helped to improve the performance. Our network architecture yielded satisfying segmentation results. In addition, generalization ability was improved due to a data augmentation strategy.

(**a**) Input image

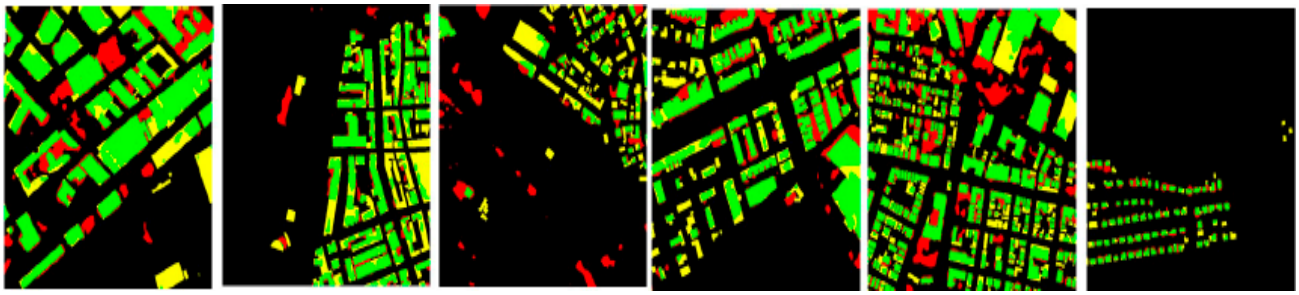
Figure 9. *Cont.*



(**b**) UNet



(**c**) ResNet50



(**d**) Our method

Figure 9. Some building segmentation results of different methods.

Figure 10. Cases of failure detection results. Red boxes, blue boxes, and green boxes denote accurately detected areas, ignored labeled detected areas, and wrongly detected areas, respectively.

Failure cases: Although the proposed method could obtain good performance on public datasets, our method did not achieve the desired results for some farmlands. As

shown in Figure 10, our method had difficulty in some cases, which were wrongly detected in the urban construction change areas bounded by green boxes. This might be due to the ground surface changing that usually happens in farmlands. Moreover, the noises in labeling might further result in performance drop.

6. Conclusions

In this paper, we proposed a feature information–interaction model for multi-source data fusion under visual perception, which adopts channel information of feature maps to weight self-attention feature maps of multiple-source data, completing extraction, fusion, and enhancement of global semantic feature with object contextual information. Then, we presented an adaptively cyclic feature information–interaction model, which adopts a branch prediction mechanism to decide the number of visual perceptions, accomplishing adaptive fusion of global semantic features and local detailed information repeatedly. Experimental results demonstrate that the proposed approach significantly improves the accuracy of the baseline model.

However, our method still needs to be improved in terms of speed and real time. How to balance the computational complexity and the accuracy remains a big challenge. In the future, we would like to discover a lower computational complexity. Additionally, better pre-trained models will be applied to the research with the development of deep networks.

Author Contributions: Conceptualization, X.C.; Methodology, X.C. and L.L.; Validation, X.C. and L.L.; Software, L.L. and C.S.; Formal Analysis, X.C. and L.L.; Writing and Original Draft Preparation, X.C and C.S.; Writing-Review and Editing, X.C., L.L. and C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded in part by the National Natural Science Foundation of China (Grant No. 61802058, 61911530397), in part by the China Postdoctoral Science Foundation (Grant No. 2019M651650) and in part by the China Scholarship Council (CSC) (Grant No. 201908320175).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Datasets are obtained from OTB Tracking Evaluation Web Site and are available at http://www.visual-tracking.net (accessed on 14 April 2021).

Acknowledgments: The authors acknowledge the academic editors and the anonymous reviewers for their insightful comments and suggestions, helping to improve quality and acceptability of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Hu, Q.; Ling, H. Vision meets drones: Past, present and future. *arXiv* **2020**, arXiv:2001.06303.
2. Shadman Roodposhti, M.; Lucieer, A.; Anees, A.; Bryan, B.A. A robust rule-based ensemble framework using mean-shift segmentation for hyperspectral image classification. *Remote Sens.* **2019**, *11*, 2057. [CrossRef]
3. Wen, D.; Huang, X.; Liu, H.; Liao, W.; Zhang, L. Semantic classification of urban trees using very high resolution satellite imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 1413–1424. [CrossRef]
4. Ma, C.; Huang, J.B.; Yang, X.; Yang, M.H. Hierarchical convolutional features for visual tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015.
5. Xiao, P.; Zhang, X.; Zhang, H.; Hu, R.; Feng, X. Multiscale optimized segmentation of urban green cover in high resolution remote sensing image. *Remote Sens.* **2018**, *10*, 1813. [CrossRef]
6. Danelljan, M.; Hager, G.; Shahbaz Khan, F.; Felsberg, M. Convolutional features for correlation filter based visual tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision Workshops (ICCVW), Santiago, Chile, 11–12, 17–18 December 2015.
7. Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; Felsberg, M. Eco: Efficient convolution operators for tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–25 July 2017.
8. Mikeš, S.; Haindl, M.; Scarpa, G.; Gaetano, R. Benchmarking of remote sensing segmentation methods. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2240–2248. [CrossRef]

9. Zhang, J.; Jin, X.; Sun, J.; Wang, J.; Sangaiah, A.K. Spatial and semantic convolutional features for robust visual object tracking. *Multimed. Tools Appl.* **2020**, *79*, 15095–15115. [CrossRef]
10. Zhang, J.; Sun, J.; Wang, J.; Yue, X.G. Visual object tracking based on residual network and cascaded correlation filters. *J. Ambient Intell. Humaniz. Comput.* **2021**, *12*, 8427–8440. [CrossRef]
11. Zhou, S.; Ke, M.; Luo, P. Multi-camera transfer GAN for person re-identification. *J. Vis. Commun. Image Represent.* **2019**, *59*, 393–400. [CrossRef]
12. Wei, W.; Yongbin, J.; Yanhong, L.; Ji, L.; Xin, W.; Tong, Z. An advanced deep residual dense network (DRDN) approach for image super-resolution. *Int. J. Comput. Intell. Syst.* **2019**, *12*, 1592–1601. [CrossRef]
13. Costa, H.; Foody, G.M.; Boyd, D.S. Supervised methods of image segmentation accuracy assessment in land cover mapping. *Remote Sens. Environ.* **2018**, *205*, 338–351. [CrossRef]
14. Wang, L.; Ouyang, W.; Wang, X.; Lu, H. Visual tracking with fully convolutional networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–13 December 2015.
15. Li, H.; Li, Y.; Porikli, F. Deeptrack: Learning discriminative feature representations online for robust visual tracking. *IEEE Trans. Image Process.* **2015**, *25*, 1834–1848. [CrossRef]
16. Hong, S.; You, T.; Kwak, S.; Han, B. Online tracking by learning discriminative saliency map with convolutional neural network. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 June 2015.
17. Nam, H.; Han, B. Learning multi-domain convolutional neural networks for visual tracking. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
18. Valmadre, J.; Bertinetto, L.; Henriques, J.; Vedaldi, A.; Torr, P.H. End-to-end representation learning for correlation filter based tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–25 July 2017.
19. Choi, J.; Jin Chang, H.; Yun, S.; Fischer, T.; Demiris, Y.; Young Choi, J. Attentional correlation filter network for adaptive visual tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–25 July 2017.
20. Tao, R.; Gavves, E.; Smeulders, A.W. Siamese instance search for tracking. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
21. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
22. Zhang, Z.; Peng, H. Deeper and wider siamese networks for real-time visual tracking. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
23. Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; Hu, W. Distractor-aware siamese networks for visual object tracking. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
24. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
25. Song, Y.; Ma, C.; Wu, X.; Gong, L.; Bao, L.; Zuo, W.; Yang, M.H. Vital: Visual tracking via adversarial learning. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
26. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 583–596. [CrossRef]
27. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Discriminative scale space tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1561–1575. [CrossRef]
28. Danelljan, M.; Hager, G.; Shahbaz Khan, F.; Felsberg, M. Learning spatially regularized correlation filters for visual tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–13 December 2015.
29. Ma, C.; Yang, X.; Zhang, C.; Yang, M.H. Long-term correlation tracking. In Proceedings of the 2015 IEEE Conference on Computer Vision And Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
30. Li, F.; Tian, C.; Zuo, W.; Zhang, L.; Yang, M.H. Learning spatial-temporal regularized correlation filters for visual tracking. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
31. Danelljan, M.; Robinson, A.; Khan, F.S.; Felsberg, M. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
32. Lukezic, A.; Vojir, T.; Cehovin Zajc, L.; Matas, J.; Kristan, M. Discriminative correlation filter with channel and spatial reliability. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–25 July 2017.
33. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. Atom: Accurate tracking by overlap maximization. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
34. Xiang, L.; Shen, X.; Qin, J.; Hao, W. Discrete multi-graph hashing for large-scale visual search. *Neural Process. Lett.* **2019**, *49*, 1055–1069. [CrossRef]

35. Gui, Y.; Zeng, G. Joint learning of visual and spatial features for edit propagation from a single image. *Vis. Comput.* **2020**, *36*, 469–482. [CrossRef]
36. Zhang, T.; Xu, C.; Yang, M.H. Multi-task correlation particle filter for robust object tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–25 July 2017.
37. Lu, H.; Ma, L.; Fu, X.; Liu, C.; Wang, Z.; Tang, M.; Li, N. Landslides information extraction using object-oriented image analysis paradigm based on deep learning and transfer learning. *Remote Sens.* **2020**, *12*, 752. [CrossRef]
38. Wang, N.; Yeung, D.Y. Learning a deep compact image representation for visual tracking. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 657–664.
39. Gao, J.; Wang, Q.; Xing, J.; Ling, H.; Hu, W.; Maybank, S. Tracking-by-fusion via Gaussian process regression extended to transfer learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 939–955. [CrossRef]
40. Dai, K.; Zhang, Y.; Wang, D.; Li, J.; Lu, H.; Yang, X. High-performance long-term tracking with meta-updater. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14–19 June 2020.
41. Benedek, C.; Descombes, X.; Zerubia, J. Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 33–50. [CrossRef]
42. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
43. Grinias, I.; Panagiotakis, C.; Tziritas, G. MRF-based segmentation and unsupervised classification for building and road detection in peri-urban areas of high-resolution satellite images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 145–166. [CrossRef]
44. Shi, Q.; Liu, X.; Li, X. Road detection from remote sensing images by generative adversarial networks. *IEEE Access* **2017**, *6*, 25486–25494. [CrossRef]
45. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8 October 2016.
46. Everingham, M.; Gool, L.V.; Williams, C.K.I.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 2010, 88: 303-338. *Int. J. Comput.Vis.* **2010**, *88*, 303–338. [CrossRef]
47. Chen, B.; Wang, D.; Li, P.; Wang, S.; Lu, H. Real-time' Actor-Critic' Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
48. He, C.; Sun, L.; Huang, W.; Zhang, J.; Zheng, Y.; Jeon, B. TSLRLN: Tensor subspace low-rank learning with non-local prior for hyperspectral image mixed denoising. *Signal. Process.* **2021**, *184*, 108060. [CrossRef]
49. Sun, L.; He, C.; Zheng, Y.; Tang, S. SLRL4D: Joint restoration of subspace low-rank learning and non-local 4-d transform filtering for hyperspectral image. *Remote Sens.* **2020**, *12*, 2979. [CrossRef]
50. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
51. Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H. LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
52. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
53. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
54. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.