AN ANCHOR-FREE SIAMESE TARGET TRACKING NETWORK FOR HYPERSPECTRAL VIDEO

Zhenqi Liu[1]*, Xinyu Wang*[2]*, Meng Shu*[1]*, Guanzhong Li*[2]*, Chen Sun*[3]*, Ziying Liu*[2]*, Yanfei Zhong*[1]***

[1]The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University, P, R, China
[2]School of Remote Sensing and Information Engineering, Wuhan University, P, R, China
[3]School of Computer Science, Wuhan University, P, R, China

ABSTRACT

Hyperspectal target tracking is aimed at taking advantage of the spectral and spatial information in the target tracking. However, due to the limited training samples, the existing hyperspectral target trackers cannot exploit semantic information of hyperspectral image. In this paper, in order to solve this problem, we propose an anchor-free Siamese network for hyperspectral video target tracking (HA-Net). A spectral classification branch is introduced to the anchor-free Siamese network to increase the network's ability to identify objects. This branch exploits all the bands of the hyperspectral video for end-to-end training, to obtain more discriminative features. By fusing the classification response map of the spectral classification branch with the classification response map of the anchor-free Siamese network, the ability of the network to distinguish foreground and background can be enhanced. At the same time, the anchor-free tracking network can reduce the calculation time of the network. The experiments conducted on hyperspectral video showed that HA-Net can effectively exploit the spectral features and significantly improve the performance of the tracking network.

Index Terms— Siamese network, hyperspectral target tracking, anchor-free, spectral classification

1. INTRODUCTION

Hyperspectral target tracking is aimed at exploiting hyperspectral video for target tracking. The state information of the target is given in the first frame of the hyperspectral video, and then the state information of the target in the subsequent frames is predicted by the tracker. Compared with RGB target tracking, hyperspectral target tracking can exploit both the spectral and spatial information of the target to track the target, which means that hyperspectral target tracking has great potential in the video tracking field.

Deep learning based methods have been widely used in the field of RGB video target tracking, and have achieved very good performances [1], [2], [3]. However, due to the limited hyperspectral video training samples, handcrafted features are used in almost all of the current hyperspectral target trackers [4], [5], [6]. The methods proposed by Qian *et al*. [7] and Uzkent *et al*. [8] are not end-to-end frameworks for deep learning. Li *et al*. [4] proposed BAE-Net, which exploits a band selection network to input the spectral band groups into the VITAL [12] tracker for tracking

In this paper, we propose a novel anchor-free hyperspectral target tracking network, where the entire network can be trained end-to-end, and high-level semantic features can be exploited for the target tracking. The whole framework is divided into two parts. The first part is the RGB branch network that is trained by the RGB data set, which can be further divided into classification and regression parts. The second part is the hyperspectral classification network, which is trained by the hyperspectral training samples. The hyperspectral classification branch is only responsible for the classification task. Finally, the hyperspectral classification response map is merged with the classification response map of the RGB branch to enhance the discrimination of the foreground and background.

2. THE ANCHOR-FREE SIAMESE TARGET TRACKING NETWORK

The limited training samples also leads to another problem, in that the modern tracking methods, such as the anchor-based [9], [1] and anchor-free methods [2], [3], [10], cannot be applied to the field of hyperspectral video tracking. The methods based on anchors have shown good performances, but heuristic knowledge is required to design the anchors, which requires a lot of time to design the anchors. The anchor-free methods have shown similar performances to the methods based on anchors. However, because there is no need to design anchors, these methods have developed rapidly in the past two years.

2.1. HA-Net

Most of the current anchor-free tracking networks are based on a fully convolutional one-stage object detector (FCOS)

*Corresponding author: Y. Zhong. E-mail: zhongyanfei@whu.edu.cn.

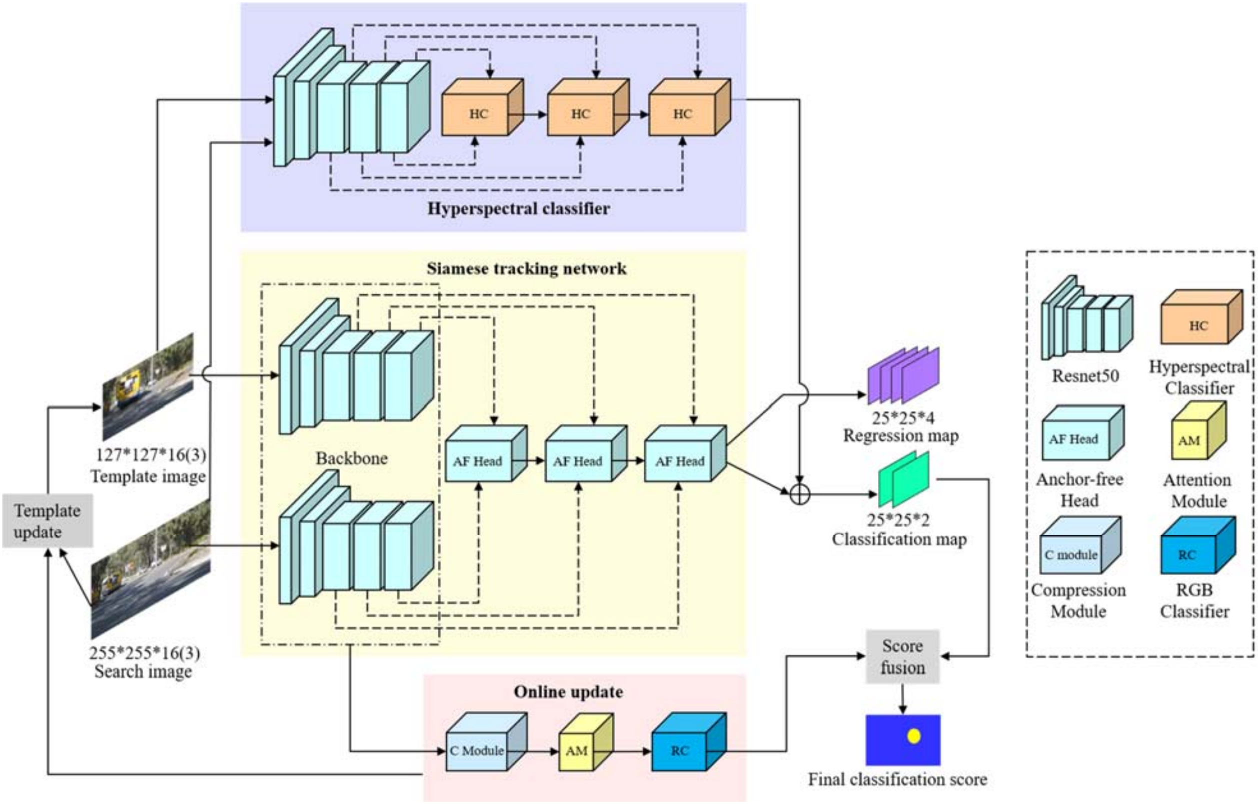


Fig. 1. The structure of HA-Net.

[11]. Typical examples are the methods of Ocean [3], SiamCAR [2], and SiamFC++ [3], all of which are based on a Siamese tracking network. The proposed HA-Net method is also an anchor-free method based on a Siamese tracking network. HA-Net includes a Siamese tracking network, a hyperspectral classifier, and online update, as shown in Fig. 1. The Siamese tracking network contains two branches: a template branch and a search branch. The template branch receives a template patch as input, and the search branch receives a search patch as input. The two branches share parameters to measure the similarity of the two inputs. The anchor-free Siamese tracking network divides the target detection into two tasks, namely, classification and regression. The classification head predicts the target category, and the regression head predicts the bounding box (BBOX) of the target. In Figure 2, the classification head and the regression head are integrated into the AF head. Unlike the anchor-based approach, the anchor-free approach does not require us to set anchors in advance, but instead directly predicts a BBOX. In the Siamese network, the feature obtained from the template branch is denoted as $\varphi(z)$, and the feature obtained from the search branch is denoted as $\varphi(x)$. The results of these two features from the classification head and regression head can be denoted as：

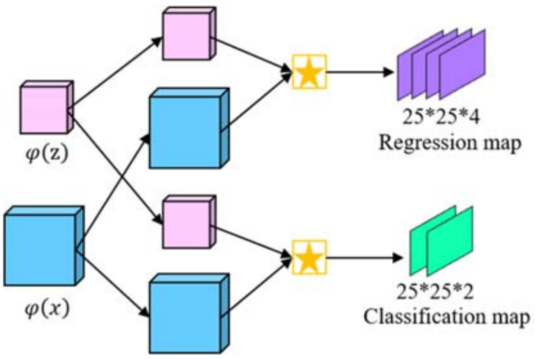
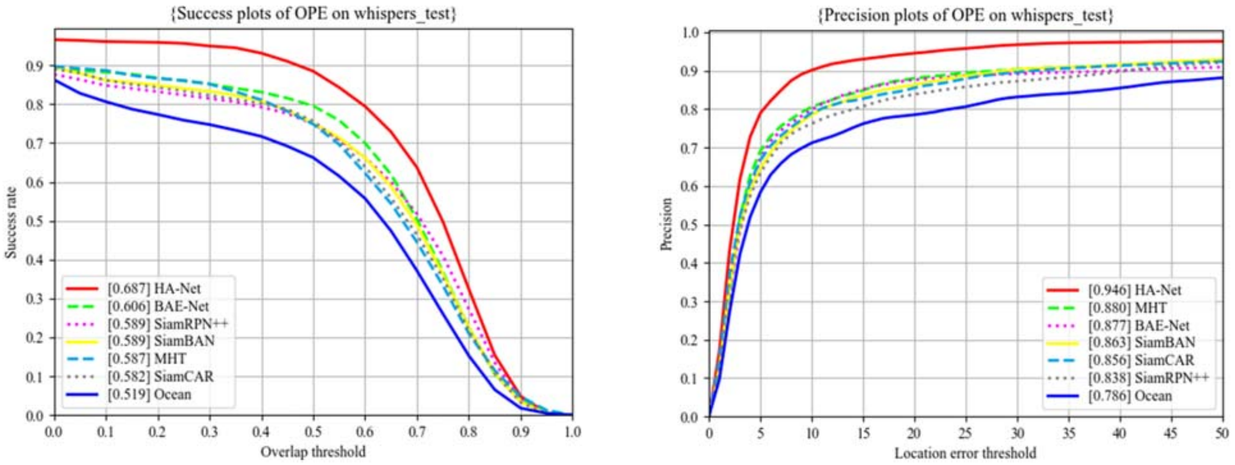


Fig. 2. The structure of anchor-free (AF) head

$$R^{cls} = [\varphi(x)]^{cls} \star [\varphi(z)]^{cls} \qquad (1)$$

$$R^{reg} = [\varphi(x)]^{reg} \star [\varphi(z)]^{reg} \qquad (2)$$

where R^{cls} and R^{reg} represent the classification response map and regression response map. $\star$ represents the cross-correlation operation, which uses $[\varphi(z)]^{cls}$ or $[\varphi(z)]^{reg}$ as



(a) Success plots (b) Precision plots

Fig. 3. The success plots and precision plots of all the trackers.

Table 1. AUC comparison of the strong benchmark deep learning based RGB trackers and hyperspectral trackers.

Video	HA-Net	BAE-Net [4]	MHT [6]	SiamRPN ++ [1]	SiamCAR [2]	Ocean [3]	SiamBAN [16]
Hyperspectral/false-color	0.687	0.606	0.587	0.589	0.582	0.519	0.589

the convolution kernel to perform convolution on $[\varphi(x)]^{cls}$ or $[\varphi(x)]^{reg}$. R^{cls} can predict the foreground (target) and background in the search patch through a soft-max function. R^{reg} predicts four offset values from the point of the corresponding search patch to the BBOX. With these four offset values, the BBOX can be calculated.

2.2. Hyperspectral Classifier

Due to the limited training samples, the performance of the current modern target tracking networks which directly exploit hyperspectral samples for the training can be very poor. Compared with RGB images, hyperspectral images have multi-band spectral information. Under the same spatial resolution, hyperspectral images can perform better in classification tasks. By combining different bands and inputting different band combinations into the VITAL tracker [12], BAE-Net has obtained higher accuracies than the original VITAL tracker, which proves that different band combinations can enhance the tracking performance of the network. Unfortunately, BAE-Net cannot exploit the semantic features of hyperspectral video to enhance the tracking performance of the network. This is due to the limited hyperspectral training samples. The tracking performance of a model trained directly using limited spectral data cannot reach the performance of a model trained with RGB data. As mentioned above, the current tracking networks divide the state prediction of the target into classification and regression operations. Therefore, we use a hyperspectral classifier to enhance the anchor-free Siamese tracking network, making the entire network more discriminative. ResNet-50 [13] is exploited as the backbone network to extract the hyperspectral features. The extracted hyperspectral features are then fed into the classification network to obtain a hyperspectral classification response map. The fused classification map can be denoted as:

$$R_{fus}^{cls} = \alpha R^{cls} + \beta R_{HSI}^{cls} \tag{3}$$

where R^{cls} and R_{HSI}^{cls} represent the RGB classification response map and the hyperspectral classification response map, respectively; and α and β respectively represent the weights of each classification map.

2.3 Online Update

During the tracking, as the target changes, the difference between the initial template and the changed target becomes greater, resulting in a decrease in the performance of the tracker. By updating the template online, this allows the tracker to find a reasonable target template which can improve the performance of the tracker. Online updating has been successfully used in trackers such as Ocean [3], DiMP [14], and ATOM [15]. Therefore, we also adopted an online update module, which selects the image with the highest confidence as the template Z_{online} every 10 frames. The

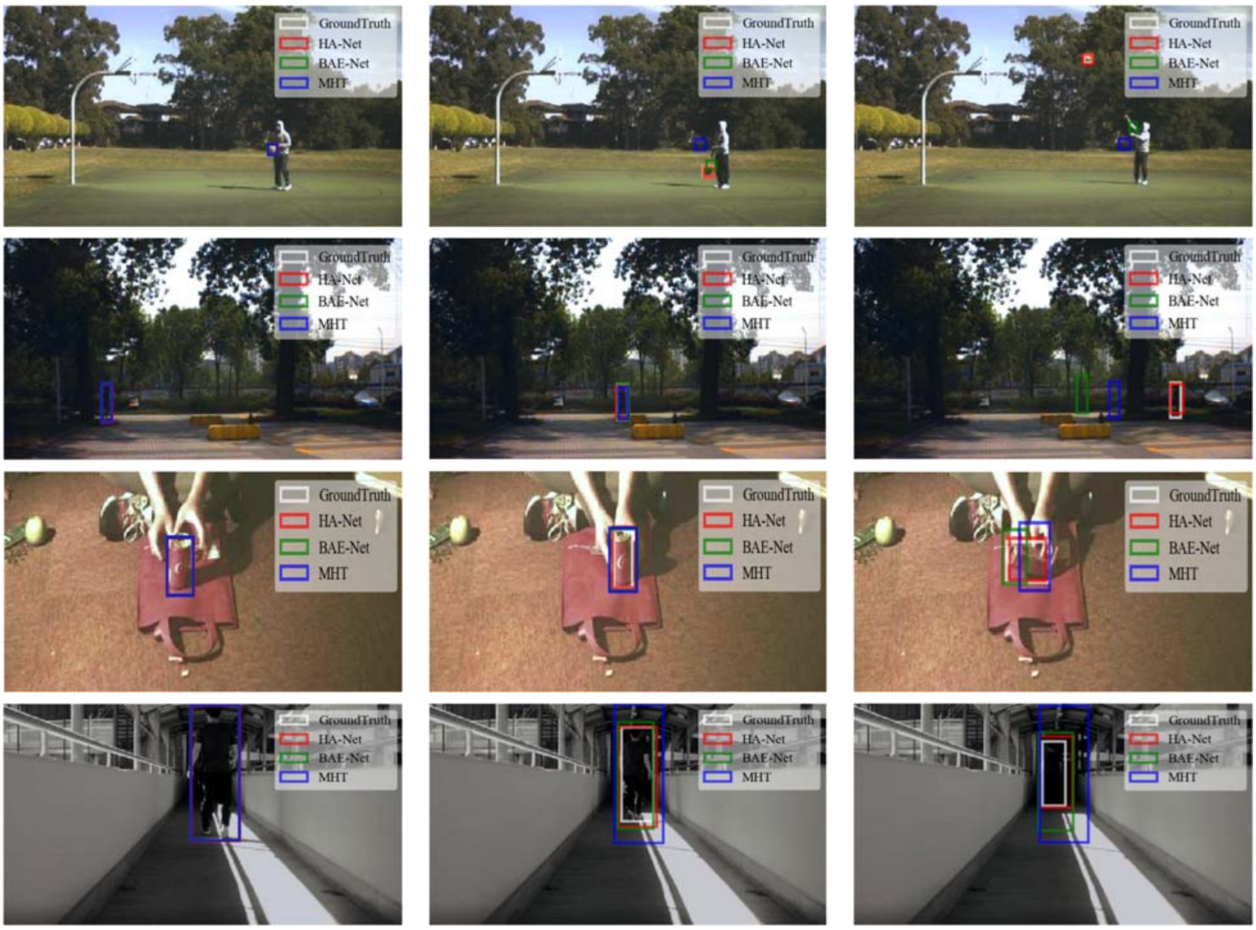


Fig. 4. Visualization results of the hyperspectral trackers (scenes: basketball, pedestrian2, coke, student).

classification score obtained by template Z_{online} and the classification score obtained by the original template Z are then fused as the final classification score.

3. EXPERIMENTS

3.1. Experimental Settings

The experiments were undertaken using the 35 videos provided by the competition website[1]. The training data set provided by the competition website was used to train the hyperspectral image classification network. The proposed network was trained with stochastic gradient descent (SGD) with a minibatch of 32 pairs. The learning rate was decayed from 0.001 to 0.0005 over a total of 20 epochs.

3.1. Results and Analysis

Comparison with SiamBAN[16]: In Figure 1, the Siamese tracking network we use is SiamBAB[16]. After adding hyperspectral classification module and online update module to SiamBAN, the final model is HA-Net. SiamBAN were tested on false-color videos. The proposed HA-Net method was tested on hyperspectral videos. Compared with SiamBAN, HA-Net surpasses SiamBAN in both the success rate plot and the precision plot.

Comparison with RGB trackers: We compared the proposed algorithm with recent RGB trackers which have are all been proposed in the last 2 years, i.e., SiamRPN++ [1], SiamCAR [2], and Ocean [3]. All the RGB trackers were tested on false-color videos. The proposed HA-Net method

[1] https://www.hsitracking.com/

was tested on hyperspectral videos. As shown in Fig. 3 and Table 1, HA-Net achieved the best performance.

Comparison with hyperspectral trackers: We also compared the proposed HA-Net method with BAE-Net [4] and MHT [6] using hyperspectral videos. BAE-Net and MHT are the most advanced hyperspectral trackers. From the experimental results shown in Table 1 and Fig. 3, it can be seen that the performance of HA-Net exceeded that of BAE-Net and MHT in terms of the success plot, precision plot, and area under the ROC curve (AUC) metric. The inference speed of HA-Net reached 14FPS, which we believe to be the fastest inference speed achieved by a hyperspectral tracker to date.

4. CONCLUSION

In this paper, we have proposed the HA-Net method, which is an advanced anchor-free Siamese target tracking network for hyperspectral video. HA-Net exploits the anchor-free Siamese network to avoid the need for a multi-scale search scheme and complicated hyperparameter search process. At the same time, HA-Net exploits the hyperspectral video classification network to extract the semantic information of the hyperspectral video, to enhance the classification results. As HA-Net effectively solves the problem of hyperspectral target tracking methods being unable to exploit the semantic information, HA-Net achieved a superior performance. In our future work, we will aim to improve the inference speed of the hyperspectral target tracker even further, which will allow the hyperspectral target tracker to process hyperspectral video in real time.

5. REFERENCES

[1] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing and J. Yan, “SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[2] D. Guo, J. Wang, Y. Cui, Z. Wang and S. Chen, “SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking,” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.

[3] Z. Zhang and H. Peng, “Ocean: Object-aware Anchor-free Tracking,” *European Conference on Computer Vision (ECCV)*, UK, 2020.

[4] Z. Li, F. Xiong, J. Zhou, J. Wang, J. Lu and Y. Qian, “BAE-Net: A Band Attention Aware Ensemble Network for Hyperspectral Object Tracking,” *2020 IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, 2020.

[5] Hien Van Nguyen, A. Banerjee and R. Chellappa, “Tracking via object reflectance using a hyperspectral video camera,” *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, San Francisco, CA, 2010.

[6] F. Xiong, J. Zhou and Y. Qian, “Material Based Object Tracking in Hyperspectral Videos,” *IEEE Transactions on Image Processing*, vol. 29, pp. 3719-3733, Jan. 2020.

[7] K. Qian, J. Zhou, F. Xiong, and H. Zhou., “Object tracking in hyperspectral videos with convolutional features and kernelized correlation filter,” *Proc. International Conference on Smart Multimedia*, Toulon, France, 2018.

[8] B. Uzkent, A. Rangnekar and M. J. Hoffman, “Tracking in Aerial Hyperspectral Videos Using Deep Kernelized Correlation Filters,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 449-461, Jan. 2019.

[9] B. Li, J. Yan, W. Wu, Z. Zhu and X. Hu, “High Performance Visual Tracking with Siamese Region Proposal Network,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, 2018.

[10] Y. Xu, Z. Wang, Z. Li, et al. “SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines,” *AAAI*, USA, 2020.

[11] Z. Tian, C. Shen, H. Chen and T. He, “FCOS: Fully Convolutional One-Stage Object Detection,” *2019 IEEE/CVF International Conference on Computer Vision (ICCV),* Seoul, Korea (South), 2019.

[12] Y. Song et al., “VITAL: VIsual Tracking via Adversarial Learning,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, 2018.

[13] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016.

[14] G. Bhat, M. Danelljan, L. Van Gool and R. Timofte, “Learning Discriminative Model Prediction for Tracking,” *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019.

[15] M. Danelljan, G. Bhat, F. S. Khan and M. Felsberg, "ATOM: Accurate Tracking by Overlap Maximization," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[16] Z. Chen, B. Zhong, G. Li, S. Zhang and R. Ji, "Siamese Box Adaptive Network for Visual Tracking," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.