FCN and Siamese Network for Small Target Tracking in Forward-looking Sonar Images

Xiufen Ye,Yue Sun,Chuanlong Li
College of Automation
Harbin Engineering University
Harbin, Heilongjiang Province, China
yexiufen@hrbeu.edu.cn,1286857718@qq.com,lcl@hrbeu.edu.cn

***Abstract*—In underwater forward-looking sonar images, a small moving target is susceptible to noise pollution, and the performance of object tracking is greatly affected by background disturbances, illumination changes and occlusion. Hence, we propose to combine FCN and Siamese network for small moving target tracking. In order to solve the problem of too few data sets, we use geometric transformation methods to extend the data sets. In the other side, we adopt the FCN network structure, it can accept any size of input forward-looking sonar images and make tracking more efficient. Moreover, by using the Siamese network structure and removing the last full connected layer, it enables tracking more accurately. The reduction in the number of network layers also greatly improves real-time performance. The experimental results show that our method is very suitable for small moving target tracking in forward-looking sonar images and there is no target tracking loss occurred. It overcomes the noise interference in forward-looking sonar images, and significantly improves the accuracy and real-time performance.**

Keywords—object tracking; deep learning; forward-looking sonar; siamese network; FCN

I. INTRODUCTION

With the advancement of science and technology, sonar technology has developed by leaps and bounds. At present, the relatively mature imaging sonar is side scan sonar and forward-looking sonar. In order to rationally develop and use marine resources, it is necessary to carry out follow-up research on underwater moving targets. Moving target tracking is to select one or more features that can uniquely identify the target according to the target's state and the environment in which the feature is located. That is, the selected features can distinguish the target and its background well. Target tracking technology has a profound impact on improving the performance of domestic underwater target detection and tracking systems, as well as the territorial sea defense capabilities. In the modern marine field, it is very realistic to track underwater targets, especially maneuvering targets such as submarines and crashed seaplanes.

The traditional small target tracking method based on forward-looking sonar images is mainly based on the principle of particle filtering. According to the imaging characteristics of forward-looking sonar, several image preprocessing methods are studied, including gray local enhancement, Otsu threshold segmentation, and binary morphology processing, edge extraction and etc[1]. Since 2015, deep learning technology has begun to enter the field of target tracking. Convolutional networks are powerful visual models that yield hierarchies of features. However, it does not combine deep information with shallow information. It reduces the tracking accuracy[1]. A correlation Filter-based trackers is proposed for tracking. However, since the model that they learn depends strongly on the spatial layout of the tracked object, they are notoriously sensitive to deformation[3]. Ning J and others present a simple yet efficient dual linear SSVM (DLSSVM) algorithm to enable fast learning and execution during tracking, but it is easy to lose the target[4]. Z. Hong and others adopt cognitive psychology principles to design a flexible representation that can adapt to changes in object appearance during tracking. However, the accuracy rate is low[5]. Since 2015, deep learning has begun to enter the field of target tracking, and has achieved remarkable results compared to the traditional methods. Therefore, this paper mainly proposes a method based on deep learning to solve these problems. The following section is composed of 3 parts. The second part is the preliminary preparation of the data and the data augementation, the third part is to study the deep learning method, the siamese network and FCN is proposed for target tracking. The last part gives the specific improved Siamese algorithm and the comparative experimental analysis.

II. DATA PREPARATION AND PROCESSING

A. Database Preparation

The concept of deep learning stems from the study of artificial neural networks. A multilayer sensor with multiple hidden layers is a deep learning structure. Deep learning creates more abstract high-level representation attribute categories or features by combining low-level features to discover distributed representations of data. A large part of the final results obtained by adopting the deep learning method depends on the preparation of the data. Adequate data preparation is a prerequisite for the network to be trained.

All the data used in our experiments was produced according to the standard VOT (visual object tracking) dataset format. Due to the limited number of data types in the VOT

978-1-5386-4814-8/18/$31.00 ©2018 IEEE

dataset itself, our experiments produced some of its own data. The images used were all taken underwater by the forward-looking sonar of type Blueview.

B. Data Augmentation

More training data in deep learning methods means that you can train more precise models with deeper networks. In order to avoid overfitting, we usually need to enter sufficient amount of data. In addition to collecting more useful data, commonly used data augmentation transformation methods are also required to make some changes to the original data and get more data. Here, we use geometric transformation for the data augmentation.

At present, the most urgent shortage for the image field is the massive data source resources, this is one of the key factors that restrict the output of model effects. Therefore, using the data augmentation method can increase the data capacity, so as to achieve the purpose of improving the tracking model effect. However, how much the tracking model effect can improve depends mainly on the original data set, the scene complexity of the data set and some other factors.

In the context of different tasks, we can use the geometric transformation of the image to achieve data augmentation. An original forword-looking sonar image is shown in Fig. 1. After the rotation, translation, scaling, and shear transformations, the amount of input data are increased, the typical forward-looking sonar image database is expanded, as shown in Fig. 2. Using geometric transformations to perform data augmentation does not change the pixel values, but changes the location of the pixels and expands the range of the data set. As input of the network, it expects to learn more image invariant features.

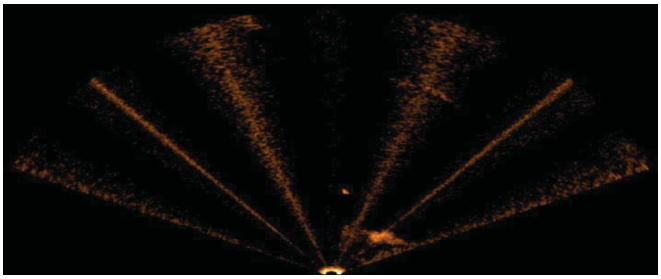

Fig. 1. An original forward-looking sonar images

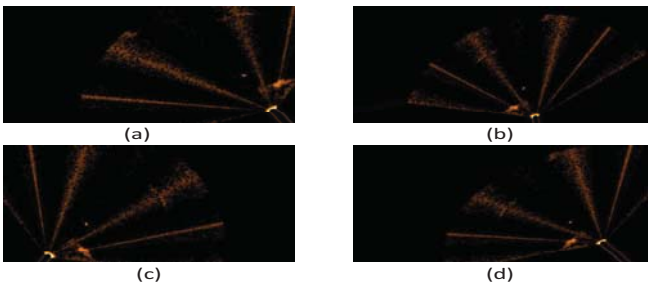

Fig. 2. Typical image instances augmented by geometric transformations: (a) cut (b) shift (c) zoom (d) rotati

III. Target Tracking Method in Forward-looking Sonar Images Based on Deep Learning

High-resolution imaging sonars have been successfully developed, but forward-looking sonar images still have their shortcomings compared to visible light and infrared images. For example, the resolution is low, the amount of target information is small, there is no clear and robust outline, and the noise is serious. All of these will have a bad influence on the tracking effect of the target, thus it causes failing to follow the target.

In this paper, we will introduce a forward-looking sonar target tracking method based on the Siamese model and FCN network. This method can overcome the shortcomings of the traditional target tracking method and significantly improve the effectiveness of the underwater tracking of the small target.

A. Analysis of Full Convolutional Network Model

FCN and CNN use full-connected layers to obtain fixed-length feature vectors for classification[1]. The size of the input image in the CNN network must be fixed. This limitation is derived from the last fully connected layer of the network. Since the network of the fully connected structure must be a fixed size input, the input size of the entire network cannot be changed.

In this paper, we propose a new tracking method using a full convolutional network structure, as shown in Fig. 3. FCN can accept input image of arbitrary size and use a deconvolution layer to upsample the feature map of the last convolutional layer, restoring it in the same size of the input image, so that a prediction can be made for each pixel, and at the same time, it also retains the spatial information in the original input image. Finally, the classification loss is calculated pixel by pixel on the upsampled feature map, and the feature classification is performed.

In this paper, the proposed FCN network structure can not only accept arbitrary sized input images, but also accept arbitrary sized image for training images and testing images. It also avoids the problem of duplicated storage and computational convolution due to the use of pixel blocks.

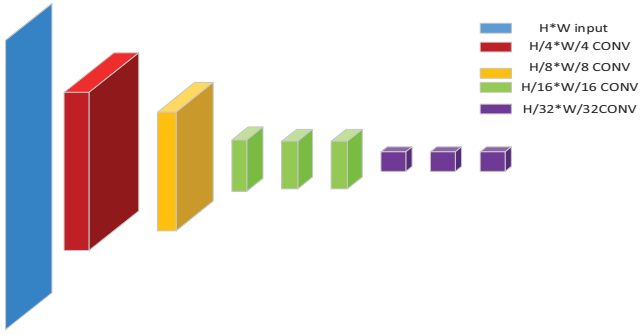

Fig. 3. FCN network structure

B. Siamese Network Model

The weights of the two channels in basic structure of the traditional siamese network are not shared[6], as shown in Fig.

4. The weights of the two channels of our improved siamese network are shared, as shown in Fig. 5. This means that the parameters in the network are greatly reduced and it improves the tracking speed. What's more, we remove the full connected layer, as shown in Fig. 6. So, it means removing a lot of redundant parameters and improving the tracking accuracy. In general, the improved Siamese network model uses CNN for feature extraction and constructs a loss function with eigenvectors, as shown in Fig. 7. In other words, the improved Siamese model uses neural network to extract the description operator to obtain the feature vector, and then uses the feature vector to determine the similarity.

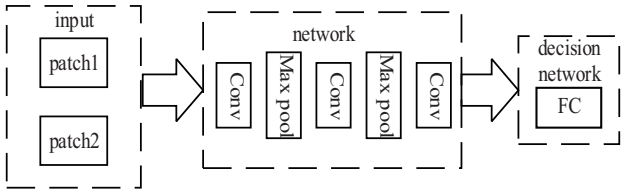

Fig. 4. Traditional Siamese network structure

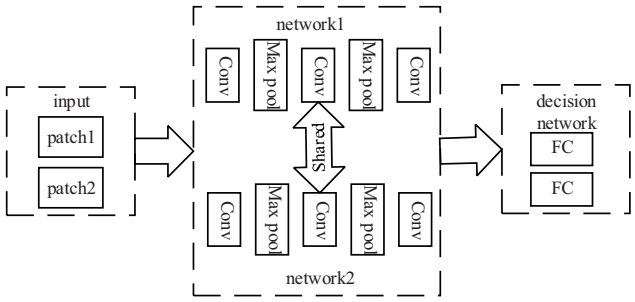

Fig. 5. Improved siamese network structure sharing weights

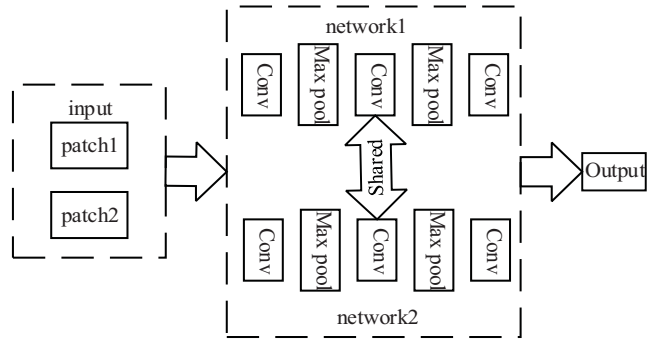

Fig. 6. Improved siamese network structure sharing weights and removing FC

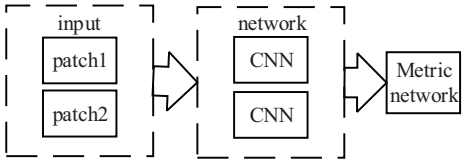

Fig. 7. Improved siamese network model

In this paper, the Siamese model is actually a process of extracting the feature operator of a image. The last layer of the network defines the loss function of the similarity between the feature vectors. The essence is a network structure with multiple branch parameters shared.

C. *Siamese Model for Forward-looking Sonar Target Tracking*

Different from the traditional target tracking method, deep learning is used as target tracking, and the learning process is completed online. After the model is determined, the model is no longer updated during tracking. The Siamese model algorithm is intended to learn the matching mechanism. From a large number of external images learning matching functions, the training images and the testing images have no intersection, focusing on the generalized target shape in the learning process. In the tracking process, the target remains unchanged, and the tracker combination and occlusion processing are not performed. In the newly occurring frame, the image block that most closely matches the original image block is found, as shown in Fig. 8.

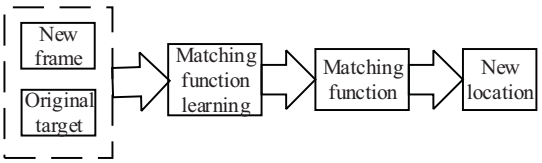

Fig. 8. Improved siamese network model work process

Using the learned matching function throughout the process, the tracker only needs to find the image block that most closely matches the original image block of the target in the first frame. The matching function is learned on the data set. In the training phase, the training goal is to make the distance between the positive and negative samples as close as possible. This is reflected in the loss function, as follows:

$$l(y,v)=\log\left(1+\exp(-yv)\right) \tag{1}$$

Where, v is real-valued score, y is groundtruth label.

The unimproved Siamese network model has five convolutional layers and one full-connection layer. Due to the existence of the last full-connection layer, the small target is lost during the tracking process, as shown in Fig. 9.

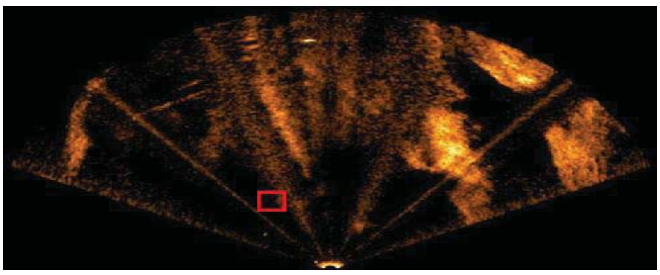

Fig. 9. Loss effect of target tracking with unimproved Siamese network model

D. An Improved Siamese Model Based on Alexnet

In forward-looking sonar images, the resolution is low and the noise is serious, there is little target feature information, and the outline of the object is obscure, hence it is easy to lose the target during the tracking process. This paper proposes an improved Siamese network structure based on Alexnet. The traditional Alexnet has 8 layers, the first 5 layers are convolutional layers, and the other 3 layers are full-connected layers. In this paper, the improved siamese network only retains the first five convolutional layers and reduces the number of pooling layers in the original network. This can significantly improve the spatial resolution of the image. Considering the time-consuming and labor-intensive problem of dealing with multiple candidate areas in a single process, so we use region pooling layer to quickly handle overlapping areas. The input of each branch is a full graph plus a series of bounding boxes. The first few layers of the network first process the entire image and extract the feature map, then the region pooling layer transforms the feature map of a specific area into a fixed-length representation and sends it to the higher layers of the network.

The deeper the network layer is, the more abstract the expression is, the lower-level features are more sensitive to intra-class differences, and the higher-level features are more sensitive to inter-class differences. Whether the high-level features are good or the low-level features are good in the tracking task is difficult to make a conclusion. Therefore, the features of high-level and low-level are adopted, and the multi-layer output characteristics are directly fed to the loss layer, so that the multi-layer features are comprehensively considered. This is the overall framework of the improved Siamese algorithm, as shown in Fig. 10.

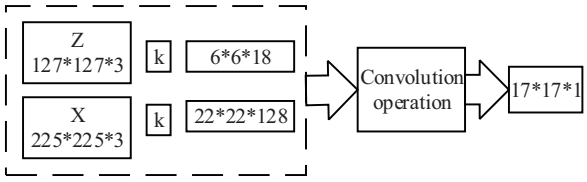

Fig. 10. Framework of the improved siamese algorithm

Where, Z is template image, X is candidate box search area in the tracked frame, k is a feature mapping operation. In this paper, it uses the convolution and pooling layers in CNN; 6*6*128 represents the dimension of characteristics of Z after k. It is a 128 channel 6*6 size feature. Similarly, 22*22*128 is the dimension of features of X after k. Let the convolution kernel perform the convolution operation, the features are changed from 22*22*128 to 6*6*128, and finally get a output with a size of 17*17*1.

In principle, the Siamese model has two identical parameters, weighted sharing CNN branches structure X and Z, respectively inputting two images, each obtaining an output feature vector x and z, and then constructing a distance metric of the two feature vectors as two images similarity function:

$$E(X,Z) = \| G(x) - G(z) \| \tag{2}$$

where $G(x)$ and $G(z)$ are independent network mapping functions.

IV. Experimental Results

In this paper, in order to verify the effectiveness of the improved siamese network, we conduct several comparative experiments.

The first set of comparative experiments is shown in Fig. 11. In this paper, the tracking goal is small. In the case of target occlusion, the traditional siamese network algorithm does not have the function of weight sharing between the two networks, it will cause the loss of target tracking. In the improved siamese network algorithm, the weights between the two networks are shared. It means that during the tracking phase, the network will automatically adjust and it can accurately track if the small target is occluded by other objects.

The second set of comparative experiments is shown in Fig. 12. The noise is serious in the forward-looking sonar images. In the traditional siamese network algorithm, the complete bounding box can not even appear when tracking small targets. It can not find the target to track. However, in the improved siamese network algorithm, it can still accurately track the small targets when background noise is serious.

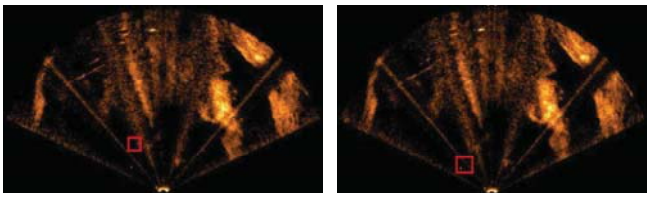

Fig. 11. Traditional siamese and improved siamese tracking comparison: (a) unimproved siamese network (b) improved siamese network

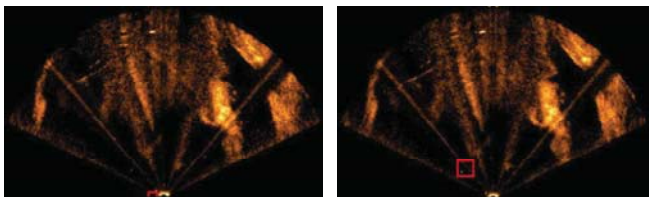

Fig. 12. Traditional siamese and improved siamese tracking comparison: (a) unimproved siamese network (b) improved siamese network

In the third set of comparative experiments, select another forward-looking sonar images to illustrate the effect. In the traditional siamese network algorithm, there are two fully connected layers in the network. It has strict requirements on the size of the input sonar image. After changing a set of sonar images, it will cause the loss of target tracking, but in the improved siamese network algorithm, we remove the fully connected layers, it has no requirements on the size of the input

sonar image. And it can accurately track the small target without loss, as shown in Fig. 13.

In the fourth set of comparative experiments, in the case of target occlusion, it causes the loss of target tracking with the traditional siamese network algorithm, while it can accurately track the small target and it does not appear to lose with the improved siamese network algorithm., as shown in Fig. 14.

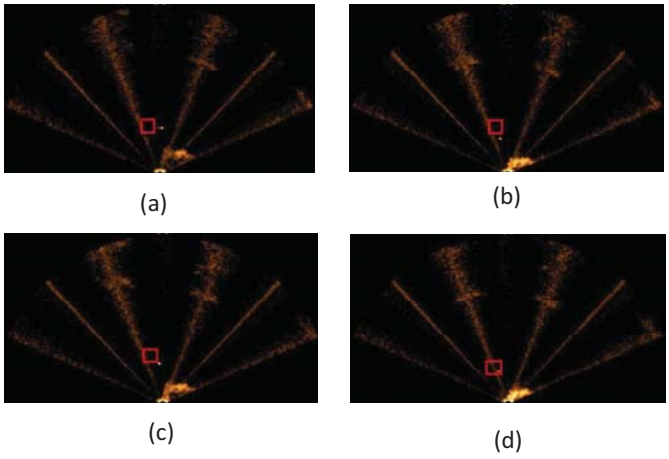

Fig. 13. Traditional siamese and improved siamese tracking comparison: (a) at time t, unimproved siamese network (b) at time t+1, unimproved siamese network (c) at time t+2, unimproved siamese network (d) improved siamese network

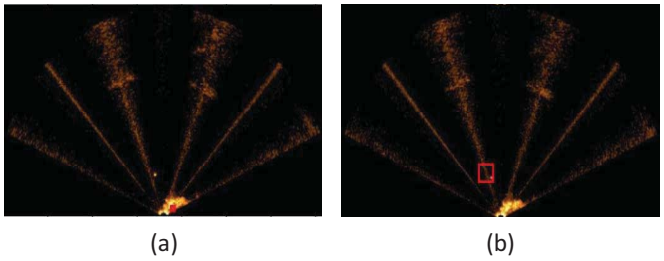

Fig. 14. Traditional siamese and improved siamese tracking comparison: (a) unimproved siamese network (b) improved siamese network

Table I introduces the tracking performance. IOU is the overlapping rate of the window produced by the model and the real target, and the function is:

$$IOU = \frac{area\ \ of\ \ overlap}{area\ \ of\ \ union} \tag{3}$$

AUC is the good and bad evaluation index of model classification:

$$AUC = \frac{\sum_{i \in positiveClass} rank - \frac{M(1+M)}{2}}{M \times N} \tag{4}$$

where M is the number of positive samples and N is the number of negative samples.

TABLE I. THE TRACKING PERFORMANCE INDEX

Target Tracking Network	Performance			
	Accuracy	*AUC*	*IOU*	*FPS*
Siamese network	13.5	0.32	4.85	30
Siamese+Alexnet network	57.0	0.55	16.98	34
Improved Siamese network	89.0	0.88	49.48	40

From Table I, some performance indexes were got through the experiments. Compared with the traditional siamese network, the accuracy of our proposed method significantly increases from 13.5% to 89%; the AUC of our method increases from 0.32 to 0.88; and the IOU increases from 4.85% to 49.48%. Thus, the proposed method allows us to accurately track the targets of interest, and it is more robust. The frame rate is increased from 30 frames per second to 40 frames per second. It greatly improves the tracking efficiency and the real-time performance is greatly guaranteed. Due to the complexity of the underwater environment, there are many unavoidable disturbance factors, and the noise interference in the forward-looking sonar images is serious. These factors have caused the difficulty in tracking the small target of the underwater forward-looking sonar images. And our proposed method not only improves the accuracy but also improves the real-time performance in tracking small targets.

V. CONCLUSIONS

In this paper, we propose a moving target tracking method in forward-looking sonar images based on FCN network and Siamese network model. We designed a 5-layer siamese network to achieve a good tracking effect. The comparison experiment shows that the method proposed in this paper does not lose in the small target tracking, even if the background noise is serious, the target is occluded and the environment is different. And it also improves the accuracy and real-time performance of tracking.

ACKNOWLEDGMENT

This work was supported by the National key research and development program of China (Grant No. 2017YFC0306000), the State Key Program of National Natural Science Foundation of China (Grant No.61633004), Development Project of Applied Technology in Harbin (Grant No.2016RAXXJ071) and the Fundamental Research Funds for the Central Universities(Grant No. HEUCFP201728 and HEUCFP201746).

REFERENCES

[1] N. Wang, J. Shi, D.-Y. Yeung, and J. Jia. Understanding and diagnosing visual tracking systems. In ICCV, 2015.

[2] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

[3] Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.S.: Staple:Complementary learners for real-time tracking. CVPR 2016.

[4] Ning, J., Yang, J., Jiang, S., Zhang, L., Yang, M.H.: Object tracking via dual linear structured svm and explicit feature map. In: CVPR 2016.

[5] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In CVPR, 2015.

[6] Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML 2015 Deep Learning Workshop. (2015)

[7] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In ECCV, 2014.

[8] J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. In CVPR, 2015.

[9] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In CVPR, 2013.

[10] Vedaldi, A., Lenc, K.: MatConvNet – Convolutional Neural Networks for MATLAB.(2015).

[11] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking.2015

[12] D. Huang, L. Luo, M. Wen, Z. Chen, and C. Zhang. Enable scale and aspect ratio adaptability in visual tracking with detection proposals. In BMVC, 2015.

[13] Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., Moreno-Noguer, F.:Discriminative learning of deep convolutional feature point descriptors. In: ICCV 2015. (2015) 118–126.

[14] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: FlowNet: Learning optical flow with convolutional networks. In: ICCV 2015.

[15] Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernandez, G.,Vojir, T., Hager, G., Nebehay, G., Pflugfelder, R.: The Visual Object Tracking VOT2015 Challenge results. In: ICCV 2015 Workshop. (2015) 1–23.