Object Fusion Tracking Based on Visible and Infrared Images Using Fully Convolutional Siamese Networks

1st Xingchen Zhang
School of Aeronautics and Astronautics
Shanghai Jiao Tong University
Shanghai, China
xingchen@sjtu.edu.cn

2nd Ping Ye
School of Aeronautics and Astronautics
Shanghai Jiao Tong University
Shanghai, China
yeping2018@sjtu.edu.cn

3th Dan Qiao
College of Automation Engineering
Shanghai University of Electric Power
Shanghai, China
921362160@qq.com

4th Junhao Zhao
School of Aeronautics and Astronautics
Shanghai Jiao Tong University
Shanghai, China
jhzhao_16@sjtu.edu.cn

5th Shengyun Peng
College of Civil Engineering
Tongji University
Shanghai, China
pengshengyun@sina.com

6nd Gang Xiao*
School of Aeronautics and Astronautics
Shanghai Jiao Tong University
Shanghai, China
xiaogang@sjtu.edu.cn

***Abstract*—Visual tracking is of great importance and thus has attracted a lot of interests in recent years. However, tracking based on visible images may fail when visible images are not reliable, for example when the illumination conditions are poor or in foggy day. Infrared images reveal thermal information thus are insensitive to these factors. Due to their complementary features, object fusion tracking based on visible and infrared images has attracted great attention recently. In this paper, a pixel-level fusion tracking method based on fully convolutional Siamese Networks, which has shown great potential in RGB object tracking, is proposed. Visible and infrared images are firstly fused and then tracking is performed based on fused images. Extensive experiments on a large dataset which contains challenging scenarios have been conducted to evaluate tracking performances. The results clearly indicate that the proposed fusion tracking method can improve tracking performance compared to methods based on images of single modality.**

***Index Terms*—fusion tracking, object tracking, image fusion, Siamese networks, deep learning**

I. Introduction

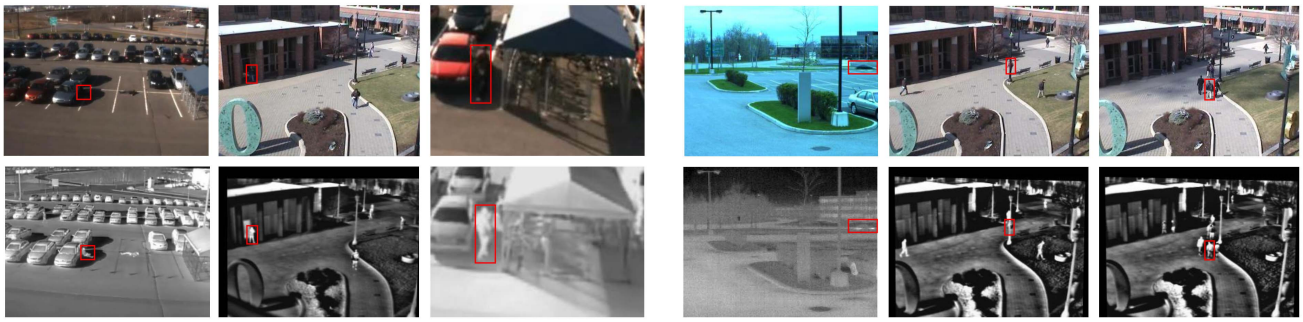
Object tracking has received great attention in recent years due to its wide applications in many areas, such as robotics, surveillance and human-machine interface. A lot of algorithms have been proposed to perform object tracking, among which the most popular ones are based on deep learning and correlation filters. Tracking methods based on deep learning, especially the convolutional neural networks (CNN), can produce good tracking performance due to the very strong feature representation ability of CNN. However, since the training and update of CNN model is time-consuming, thus normally one trains the CNN model offline and just use it during tracking. In contrast, the correlation filter based trackers can update model online due to the low computational cost of correlation filters. However, the correlation filter-based methods suffer from the boundary effect of correlation filter computation, which limits their applications.

Currently, most tracking algorithms are developed for tracking based on visible images [1]. Despite remarkable progress, tracking algorithms based on visible images may fail when visible images are not reliable, for instance when poor light condition, fog or haze present. In contrast to visible images, infrared images reveal thermal information of objects and are insensitive of these factors, thus can provide complementary benefits with visible images and show camouflaged objects under darkness etc., as shown in Fig. 1. By fusing complementary information from visible and infrared images, the application area and robustness of tracking algorithms can be greatly enhanced. Therefore, in recent years the object tracking based on visible and thermal infrared images have become a hot research topic, and is termed as RGB-Thermal (RGB-T) tracking [2] or fusion tracking [3].

A lot of fusion tracking algorithms have been proposed to utilize complementary information in visible and infrared during tracking. Before deep learning and correlation filters become popular, researchers perform fusion tracking with traditional tracking methods [3], [4]. In recent years, researchers have turned to fusion tracking algorithms based on deep learning [5] and correlation filters [6]. In these studies, researchers either firstly fuse features then perform tracking [5], [7], or firstly perform tracking with images of different modalities individually and then fuse tracking results [8], [9]. However, the *fusion then tracking* method has not been fully explored, which may be a promising research direction of fusion tracking.

In this study, we aim to explore the *fusion then tracking* or pixel-level method, namely firstly fuse visible and infrared

* Corresponding author



(a) Target in infrared images is more clear

(b) Target in visible images is more clear

Fig. 1. Examples of complementary information from visible and infrared images.

images, and then perform tracking based on fused images. Specifically, a Siamese network [10] is proposed to perform tracking after obtaining fused images. Also, in this study several image fusion methods are chosen in order to investigate the influence of fusion methods on tracking performance.

In summary, the main contributions of this paper are as follows:

- Siamese network is utilized to perform pixel-level fusion tracking. To the best of our knowledge, this is the first work to perform fusion tracking based on Siamese networks.
- Several image fusion methods are implemented in the pixel-level fusion tracking to investigate the influence of image fusion methods on tracking performance.
- Extensive experiments have been conducted on a large scale visible and infrared image dataset to verify the significance of the proposed fusion tracking method.

The rest of the paper is organized as follows. Section II introduces some related work and Section III discusses the proposed pixel-level fusion tracking algorithm. Then, experiment details and results will be presented in Section IV and Section V, respectively. In Section VI some discussion on results will be given and finally Section VII concludes the paper.

II. Related work

A. Visual object tracking

At the moment, two main kinds of methods in visual object tracking are based on deep learning [10] and correlation filter [11]. Methods based on deep learning mainly utilize its strong feature representation ability compared to hand-crafted features. However, since the update of deep learning model is time-consuming, normally one trains the model offline and keep it fixed when performing online tracking. Siamese works have been widely applied to object tracking since 2016 due to its high performance and computational efficiency [10], [12], [13]. In contrast, correlation filters-based methods are computationally efficient thus it can be updated online. However, correlation filters-based methods suffer from boundary effect etc.

A benchmark and dataset which can be used to compare results is of vital importance in visual tracking. Wu et al. [14], [15] proposed the benchmarks of visual tracking (OTB) which greatly promote the development of object tracking. In addition, the VOT challenge [16] also provides a nice platform for the community to compare tracking performance under the same standard. Thanks to these datasets and benchmarks, the researches on object tracking have been boosted.

B. Image fusion

Image fusion aims to combine information, especially complementary information, from multiple images into a single image. This single image is able to provide better data source for applications. Many image fusion algorithms have been proposed, which can be generally divided into pixel-level, feature-level and decision level fusion. Also, image fusion can either be performed in spatial domain or transform domain. Before deep learning is introduced to image fusion community, the main image fusion methods contain weighted average method, wavelet-based method, PCA-based method, sparse representation-based method etc.

After 2016, researchers began to perform image fusion based on deep learning methods, including multi-focus image fusion [17], medical image fusion [18], visible and infrared image fusion [19], multi-exposure [20] etc. Regarding methods, CNN [21], Generative Adversarial Networks (GAN) [22], Siamese networks [23], autoencoder [24] have been explored to perform image fusion.

C. Fusion tracking

In past three years, fusion tracking has attracted a lot of interests and an increasing number of researches have been published in high quality journals or well-known conferences. Some of them are listed in Table I. As can be seen, although most of them are based on deep learning methods, correlation filters have also been introduced into fusion tracking in 2018. Generally speaking, nowadays deep learning-based and correlation filters-based methods are two main kinds of fusion tracking methods.

Deep learning have shown great potential in RGB tracking, thus researchers have started to applied deep learning into fusion tracking. For example, Xu et al. [32] presented a fusion tracking method based on CNN. In that work, a two-layer simple CNN was utilized and the infrared channel was simply

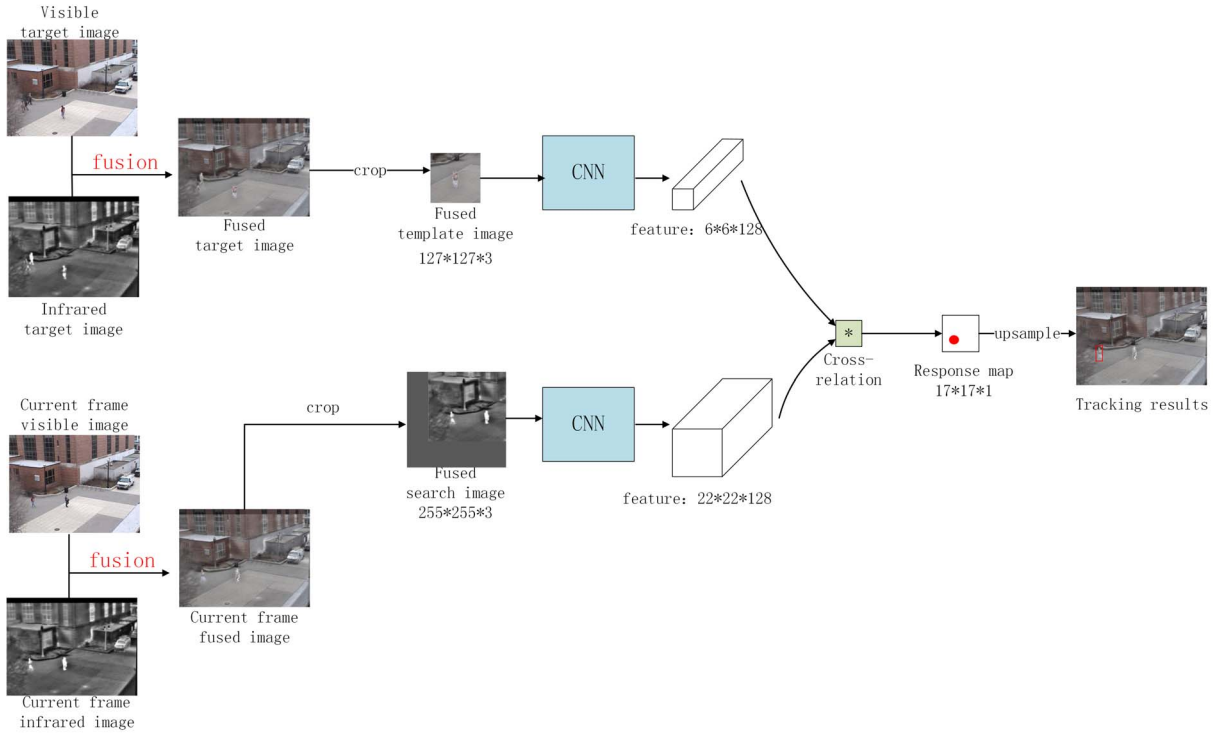


Fig. 2. Flowchart of the proposed pixel-level fusion tracking algorithms.

TABLE I
EXAMPLES OF RECENT PUBLISHED RESEARCHES ON RGB-T FUSION TRACKING

Author	Year	Where	Type
Li et al. [25]	2016	TIP	DL-based
Li et al. [26]	2017	ACM Multimedia	DL-based
Zhu et al. [7]	2018	arXiv	DL-based
Lan et al. [27]	2018	AAAI	DL-based
Lan et al. [28]	2018	PRL	DL-based
Li et al. [2]	2018	arXiv	DL-based
Li et al. [5]	2018	Neurocomputing	DL-based
Li et al. [29]	2018	ECCV	Structured SVM
Wang et al. [6]	2018	PRCV	CF-based
Zhai et al. [30]	2019	Neurocomputing	CF-based
Lan et al. [31]	2019	IEEE T IND ELECTRON	DL-based

regarded as the 4th channel of the RGB image. Li et al. [5] proposed a two-stream CNN for fusion tracking, which employed two CNNs to process visible and infrared images, respectively.

Apart from deep learning methods, correlation filters have also produced promising performance in fusion tracking due to their good performance and high efficiency. To the best of our knowledge, Wang et al. [6] presented the first fusion tracking work based on correlation filters. After that, Zhai et al. [30] proposed a fast RGB-T tracking via cross-modal correlation filters. Although the research of fusion tracking based on correlation filters just began in 2018, their highly competitive performances make them a promising research direction in future.

III. METHODS

A. Pixel-level fusion tracking

Similar to image fusion, fusion tracking can also be performed in pixel-level, feature-level and decision-level, depending on when and how images are fused in the whole process. In this work, we focus on the pixel-level fusion tracking method.

Pixel-level fusion tracking means that the images of different modalities are firstly fused into a more informative image, then object tracking is conducted based on the fused image. The basic principle is that the fused image should contain complementary information or features from different images, thus is beneficial for tracking algorithm. The main advantages of pixel-level fusion tracking are as follows:

- It is easy to implement. There are many image fusion codes available online thanks to their authors. Therefore, it is convenient for researchers to produce fused images.
- It is possible to employ advanced RGB trackers in pixel-level fusion tracking. Object tracking is a very hot topic in computer vision, a lot of advanced RGB trackers have been and are being proposed, which can achieve good tracking performance. These advanced trackers are beneficial for pixel-level fusion tracking, for instance one

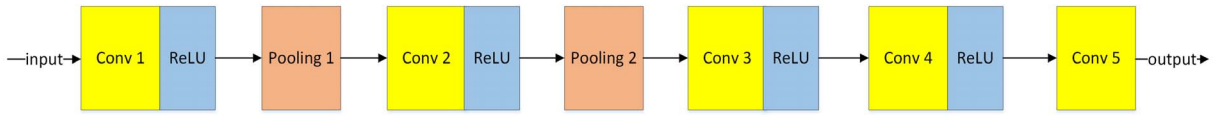


Fig. 3. The network architecture of CNN in SiamFC.

can utilize the fused images as input of RGB trackers without modifying network architecture.
- Tracking results can be visualized in fused images, which is more user-friendly and clear especially when the object is under darkness.

B. Network architecture

In this study, the SiamFC proposed by Bertinetto et al. [10] is utilized as the backbone of pixel-level fusion tracking framework because of its good performance in both tracking results and speed. However, one should note that the principle of this study is generic, which means that other Siamese networks-based tracking methods can also be employed as the backbone of fusion tracking process.

The flowchart of the proposed pixel-level fusion tracking algorithm is illustrated in Fig. 2. Basically, one firstly needs to fuse the first frame visible and infrared image to obtain the fused target image. Then, based on this fused target image, a fused template image is generated by cropping. Besides, one need to fuse current frame visible and infrared image to produce the current frame fused image, from which the fused search image is generated. Note that both the template and search images are centered at the tracking target object, and their size are $127 \times 127 \times 3$ and $255 \times 255 \times 3$, respectively. If the target is very close to the boundary, then one needs to fill in the image using mean pixel value after cropping. The next step is to feed the template and search image to two branches of the Siamese network, which will produce two convolutional features. Then by computing the cross-relation between these two features, one can obtain the response map which reflects the position of target. Finally, the position of the target in current frame can be obtained by upsampling the response map. The algorithm of pixel-level fusion tracking based on Siamese network is illustrated in Algorithm 1.

Denote the CNN with φ, fused search image with x_f, fused template image with z_f, then the response map of the pixel-level fusion tracking method proposed in this study is:

$$responseMap = \varphi(x_f) * \varphi(z_f), \tag{1}$$

where $*$ indicates the convolution operation. Note that in this work, the cross-relation is implemented using convolution efficiently.

The structure of the CNN is given in Fig. 3. As can be seen, a ReLU layer is followed after each convolution layer except the last one. Also, pooling layers are only utilized after the first two convolution layers. In addition, this CNN is fully convolutional such that there is no restrict requirements on the size of input images.

Algorithm 1: Pixel-level fusion tracking algorithm

Input: Registered visible and infrared images
Output: Predicted position and size of object in each frame
***Initialization*:**
Fuse the first frame visible and infrared image, $I_f = I_v \oplus I_i$
Crop the fused image to obtain fused template image z_f
Tracking:
For each frame i ***do***
    Fuse visible and infrared image, $I_{f_i} = I_{v_i} \oplus I_{i_i}$
    Crop the fused image to obtain fused search image x_{f_i}
    Feed template image z_f and search image x_{f_i} into two branches of the Siamese network, to obtain template features and search features
    Compute cross-relation of template and search features to produce the response map
    Upsample the response map to obtain the predicted position of target
end for

IV. Experiments

To test the performance of pixel-level fusion tracking, a lot of experiments are conducted. All experiments in this study are conducted using a desktop with a NVIDIA GTX 1080Ti GPU and i7-8700K CPU. The Siamese network is pretrained using the ImageNet dataset [33].

A. Datasets

Large datasets are critical in fusion tracking. A good dataset should have following attributes:

- Have a large number of aligned visible and infrared video frames.
- Groundtruth is available, showing the position and size of target object.
- Videos should cover a wide range of working conditions, such as low light condition, fast motion, occlusion.

In this paper, a recently released large-scale RGB-T dataset, namely RGBT234 [2] is chosen. This dataset contains 234 pairs of visible and infrared videos. Besides, it provides groundtruth and attribute annotations as shown in Table II.

B. Evaluation metrics

In recent years, several well-recognized evaluation metrics have been proposed to evaluate tracking performance based on visible images. In this study, we chose two of them, namely success plot and precision plot to evaluate fusion tracking performance.

Success means that the overlapping between the predicted bounding box and groundtruth is larger than a threshold, where the overlapping is defined as:

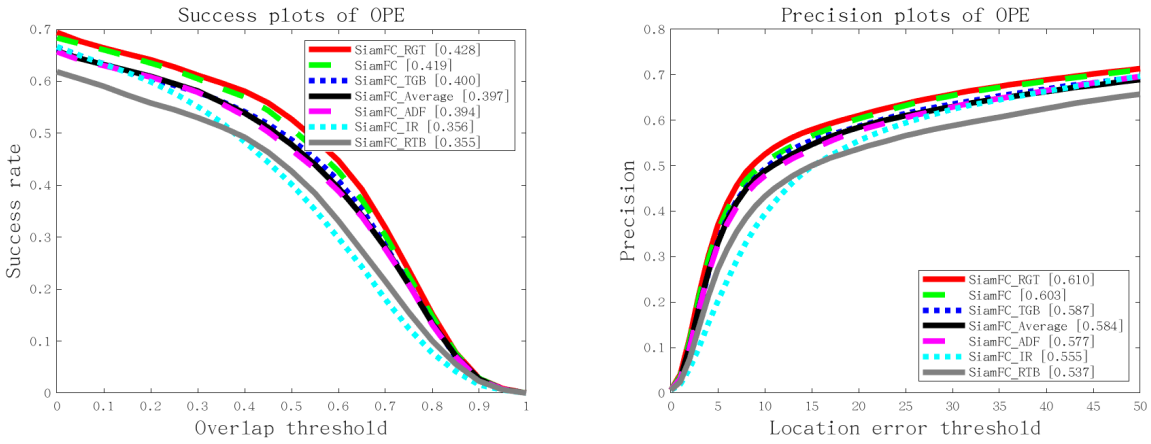


Fig. 4. Success (left) and precision (right) plot of results using different fusion methods

TABLE II
ATTRIBUTE INFORMATION OF RGBT234 DATASET [2]

Attribute	Description	Number of videos
NO	No Occlusion	41
PO	Partial Occlusion	96
HO	Heavy Occlusion	96
LI	Low Illumination	63
LR	Low Resolution	50
TC	Thermal Crossover	28
DEF	Deformation	76
FM	Fast Motion	32
SV	Scale Variation	120
MB	Motion Blur	55
CM	Camera Moving	89
BC	Background Clutter	54

$$O(a,b) = \frac{|a \bigcap b|}{|a \bigcup b|} \tag{2}$$

where a and b indicates the predicted bounding box and groundtruth, respectively. The success plot shows the trends of success rate when the threshold changes from 0 to 1. The area under curve (AUC) is employed to rank different methods effectively.

Precision means that the center location error (CLE) between the predicted bounding box and the groundtruth is smaller than a chosen threshold. The precision plot shows the trends when the threshold changes from small to large. Threshold is chosen as 20 pixels to rank algorithms.

C. Fusion methods

To investigate the influence of fusion methods on tracking performance, several fusion methods are chosen as listed in Table III. In the first method, we simply compute the average between visible and infrared images as

$$I_f = 0.5 \times I_v + 0.5 \times I_i, \tag{3}$$

TABLE III
IMAGE FUSION METHODS EMPLOYED IN THIS STUDY

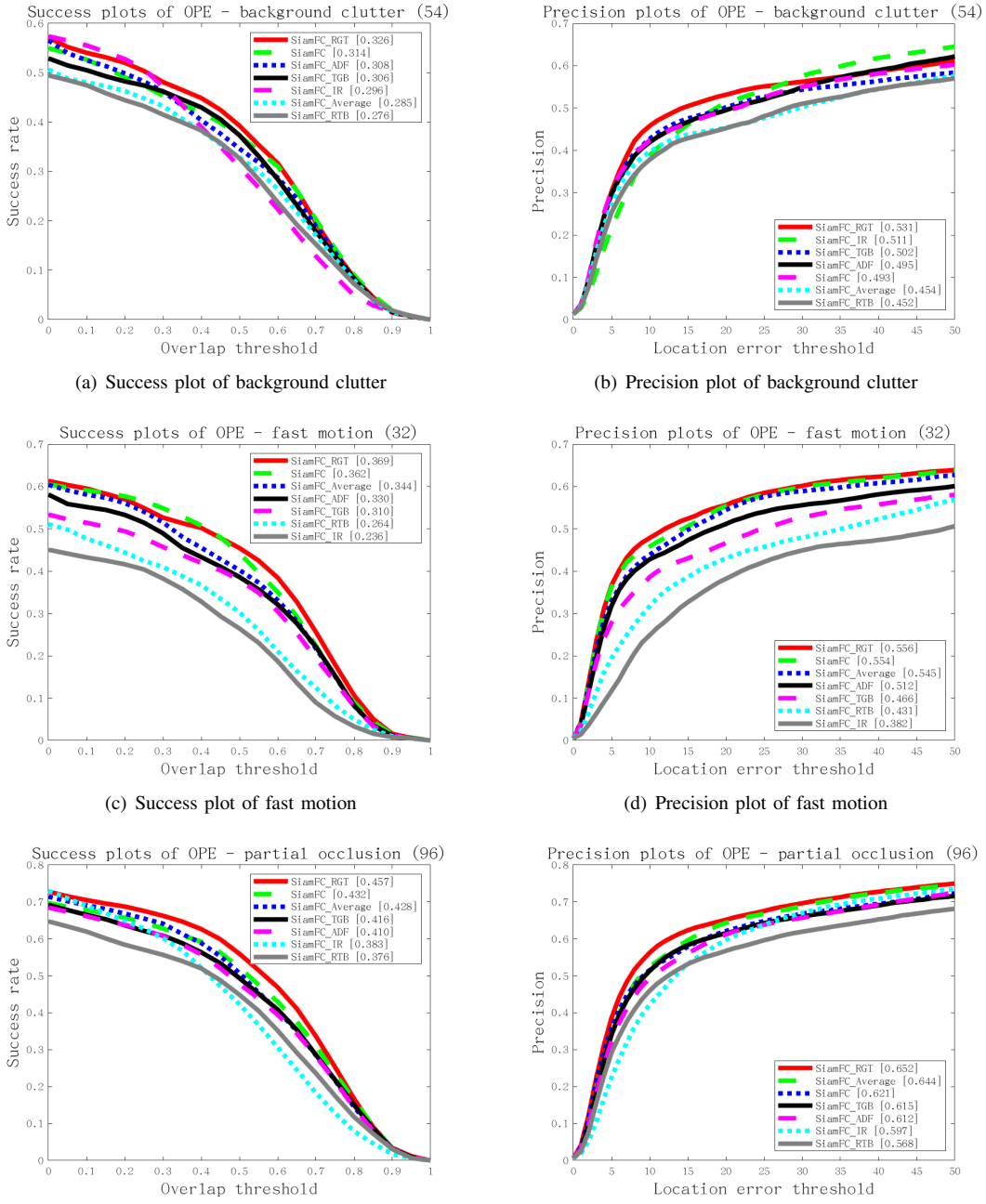
Name	Description	Method Denotation
Fused 1	Average	SiamFC_Average
Fused 2	TGB	SiamFC_TGB
Fused 3	RTB	SiamFC_RTB
Fused 4	RGT	SiamFC _RGT
Fused 5	ADF	SiamFC_ADF

where I_f denotes fused images, I_v and I_i are the visible and infrared images, respectively. From Fused 2 to Fused 4 methods, we replace one channel in visible images using the corresponding infrared image, resulting in TGB, RTB and RGT images. The fifth method is called ADF which is proposed in [34]. Default settings proposed by the authors are chosen. Note that this method can only produce grayscale fused images. It worth mentioning that some fusion algorithms are very time-consuming thus is not feasible for online pixel-level fusion tracking. For instance, it takes about 80 seconds to fuse one image pair using the latent low-rank representation (LatLRR) [19], which is absolutely not feasible.

V. RESULTS

A. Comparison of different fusion methods

As the first step, we present the comparison among methods using different fusion approaches. Figure 4 shows the overall precision and success plot. As can be seen, in both success plot and precision plot, SiamFC_RGT shows the best overall performance, outperforming both SiamFC and SiamFC_IR. This clearly indicates that by effectively fusing complementary information from visible and infrared images, the tracking performance can be improved. However, one can also observe that SiamFC outperforms the other four fusion tracking methods. This may because that these fusion tracking algorithms are not able to extract and exploit complementary features effectively. Furthermore, this indicates that it is critical



(a) Success plot of background clutter

(b) Precision plot of background clutter

(c) Success plot of fast motion

(d) Precision plot of fast motion

(e) Success plot of partial occlusion

(f) Precision plot of partial occlusion

Fig. 5. Success and precision plots of different attributes.

to develop an effective image fusion algorithm to produce better tracking performance. An improper fusion algorithm may impair the tracking performance.

B. Results of different attributes

An algorithm may have different performances in various working conditions. Figure 5 shows success and precision plot in three typical situations (background clutter, fast motion, partial occlusion). As can be seen, in these three situations SiamFC_RGT outperforms all other methods, indicating that the effective fusion of visible and infrared images can solve several challenges encountered by images of a single modality.

The success rate of these methods on all attributes are given in Table IV. Clearly, this tables indicates that SiamFC_RGT achieves the best attribute-based performance in this study, by having 9 best and 2 second best value. Specifically, in all cases SiamFC_RGT is better than SiamFC_IR, and it beats SiamFC in 10 attributes out of 12. In terms of precision rate, Table V presents the results for all attributes. Although slightly worse than that in the success rate, SiamFC_RGT still gives

TABLE IV
SUCCESS RATE. THE BEST THREE RESULTS ARE SHOWN IN RED, GREEN AND BLUE, RESPECTIVELY. BEST VIEWED IN COLOR.

Attribute	SiamFC	SiamFC_IR	SiamFC_Average	SiamFC_TGB	SiamFC_RTB	SiamFC_RGT	SiamFC_ADF
NO	0.521	0.450	0.532	0.520	0.451	0.524	0.551
PO	0.432	0.383	0.428	0.416	0.376	0.457	0.410
HO	0.359	0.287	0.304	0.330	0.291	0.355	0.307
LI	0.354	0.412	0.336	0.355	0.322	0.375	0.338
LR	0.377	0.310	0.375	0.379	0.317	0.395	380
TC	0.472	0.245	0.423	0.440	0.347	0.426	0.454
DEF	0.400	0.345	0.399	0.379	0.346	0.406	0.392
FM	0.362	0.236	0.344	0.310	0.264	0.369	0.330
SV	0.438	0.362	0.421	0.425	0.371	0.441	0.426
MB	0.399	0.315	0.347	0.368	0.325	0.402	0.356
CM	0.388	0.352	0.368	0.362	0.329	0.398	0.367
BC	0.314	0.296	0.285	0.306	0.276	0.326	0.308
Overall	0.419	0.356	0.397	0.400	0.355	0.428	0.394

TABLE V
PRECISION RATE. THE BEST THREE RESULTS ARE SHOWN IN RED, GREEN AND BLUE, RESPECTIVELY. BEST VIEWED IN COLOR.

Attribute	SiamFC	SiamFC_IR	SiamFC_Average	SiamFC_TGB	SiamFC_RTB	SiamFC_RGT	SiamFC_ADF
NO	0.729	0.695	0.750	0.735	0.666	0.719	0.767
PO	0.621	0.597	0.644	0.615	0.568	0.652	0.612
HO	0.528	0.451	0.448	0.490	0.446	0.518	0.457
LI	0.504	0.648	0.507	0.528	0.497	0.531	0.509
LR	0.590	0.575	0.609	0.613	0.538	0.607	0.628
TC	0.677	0.433	0.595	0.645	0.527	0.596	0.655
DEF	0.562	0.509	0.574	0.535	0.497	0.561	0.552
FM	0.554	0.382	0.545	0.466	0.431	0.556	0.512
SV	0.613	0.550	0.610	0.612	0.565	0.611	0.611
MB	0.574	0.469	0.495	0.514	0.463	0.560	0.497
CM	0.558	0.524	0.533	0.517	0.466	0.557	0.520
BC	0.493	0.511	0.454	0.502	0.452	0.531	0.495
Overall	0.603	0.555	0.584	0.587	0.537	0.610	0.577

the best overall performance by having 3 best and 4 second best value. Again, these experimental results demonstrate that by leveraging complementary information from visible and infrared images effectively, the tracking performance can be improved.

In this study, the tracking speed of SiamFC is around 35 frames per second (FPS), which can be used in real time. However, it takes time to fuse image, thus the speed of other fusion tracking methods degrade to different extent. Therefore, it is very helpful and vital to develop efficient fusion methods such that the pixel-fusion tracking can be performed in real time.

VI. DISCUSSION

The main aim of this study it to propose a generic pixel-level fusion tracking method, but not to propose a specific fusion algorithm that gives the best performance. Therefore, we choose several simple image fusion methods instead of more advanced ones. The other reason of just choose simple fusion methods is that, advanced fusion methods such as those based on CNN [23] or LatLRR [19] are computational expensive, thus are infeasible to achieve real time fusion tracking.

The most important shortcoming of pixel-level fusion tracking is that the computational costs of image fusion algorithms will greatly affect the computational efficiency of fusion tracking algorithms. Therefore, one has to develop efficient image fusion algorithms to make the pixel-level fusion tracking algorithm efficient.

Unlike image fusion which aims to obtain a high-quality image with as much inherited information as possible, visual object tracking do not need all information in source images. instead, it only takes certain information to achieve a good tracking result. This may be a future research direction to develop lightweight and efficient fusion algorithms that are particular suitable for tracking.

VII. Conclusion

In this paper, a pixel-level fusion tracking method based on fully convolutional Siamese Networks is proposed. Visible and infrared images are firstly fused and then fed into the Siamese network to perform object tracking. To investigate the influence of image fusion methods on tracking performance, several image fusion algorithms are chosen. Extensive experiments have been conducted, which clearly indicate that the proposed fusion tracking method can improve tracking performance especially when images of a single modality meets challenge, such as partial occlusion, low illumination. Experiments also show that an effective and efficient image fusion method is critical for the pixel-level fusion tracking, in terms of both tracking precision and speed. We believe that a lightweight yet effective image fusion rule should be one of future research directions in the field of object fusion tracking.

Acknowledgement

This paper is sponsored by National Program on Key Basic Research Project (2014CB744903), Shanghai Science and Technology Committee Research Project (17DZ1204304).

References

[1] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental comparison," *Pattern Recognition*, vol. 76, pp. 323–338, 2018.

[2] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: benchmark and baseline," *arXiv preprint arXiv:1805.08982*, 2018.

[3] H. Liu and F. Sun, "Fusion tracking in color and infrared images using joint sparse representation," *Science China Information Sciences*, vol. 55, no. 3, pp. 590–599, 2012.

[4] G. Xiao, X. Yun, and J. Wu, "A new tracking approach for visible and infrared sequences based on tracking-before-fusion," *International Journal of Dynamics and Control*, vol. 4, no. 1, pp. 40–51, 2016.

[5] C. Li, X. Wu, N. Zhao, X. Cao, and J. Tang, "Fusing two-stream convolutional neural networks for rgb-t object tracking," *Neurocomputing*, vol. 281, pp. 78–85, 2018.

[6] Y. Wang, C. Li, and J. Tang, "Learning Soft-Consistent Correlation Filters for RGB-T Object Tracking," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 295–306.

[7] Y. Zhu, C. Li, Y. Lu, L. Lin, B. Luo, and J. Tang, "FANet: Quality-Aware Feature Aggregation Network for RGB-T Tracking," *arXiv preprint arXiv:1811.09855*, 2018.

[8] C. Bailer, A. Pagani, and D. Stricker, "A superior tracking approach: Building a strong tracker through fusion," in *European Conference on Computer Vision*. Springer, 2014, pp. 170–185.

[9] T. A. Biresaw, A. Cavallaro, and C. S. Regazzoni, "Tracker-level fusion for robust bayesian visual tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 776–789, 2015.

[10] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European conference on computer vision*. Springer, 2016, pp. 850–865.

[11] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[12] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic siamese network for visual object tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1763–1771.

[13] A. He, C. Luo, X. Tian, and W. Zeng, "A twofold siamese network for real-time object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4834–4843.

[14] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2411–2418.

[15] ——, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[16] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 2137–2155, Nov 2016.

[17] X. Yan, S. Z. Gilani, H. Qin, A. Mian, S. Member, S. Z. Gilani, H. Qin, and A. Mian, "Unsupervised Deep Multi-focus Image Fusion," pp. 1–11, 2018.

[18] Y. Liu, X. Chen, J. Cheng, H. P. I. F. (Fusion), and U. 2017, "A medical image fusion method based on convolutional neural networks," *Information Fusion (Fusion), 2017 20th International Conference on*, pp. 1–7, 2017.

[19] H. Li and X.-J. Wu, "Infrared and visible image fusion using latent low-rank representation," *arXiv preprint arXiv:1804.08992*, 2018.

[20] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *2017 IEEE International Conference on Computer Vision (ICCV). IEEE*, 2017, pp. 4724–4732.

[21] H. Hermessi, O. Mourali, and E. Zagrouba, "Convolutional neural network-based multimodal image fusion via similarity learning in the shearlet domain," *Neural Computing and Applications*, pp. 1–17, 2018.

[22] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Information Fusion*, vol. 48, no. June 2018, pp. 11–26, 2019.

[23] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 03, p. 1850018, 2018.

[24] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *arXiv preprint arXiv:1804.08361*, 2018.

[25] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5743–5756, 2016.

[26] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 1856–1864.

[27] X. Lan, M. Ye, S. Zhang, and P. C. Yuen, "Robust collaborative discriminative learning for rgb-infrared tracking," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, "Modality-correlation-aware sparse representation for rgb-infrared object tracking," *Pattern Recognition Letters*, 2018.

[29] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 808–823.

[30] S. Zhai, P. Shao, X. Liang, and X. Wang, "Fast RGB-T tracking via cross-modal correlation filters," *Neurocomputing*, 2019.

[31] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, and H. Zhou, "Learning Modality-Consistency Feature Templates: A Robust RGB-Infrared Tracking System," *IEEE Transactions on Industrial Electronics*, 2019.

[32] N. Xu, G. Xiao, X. Zhang, and D. P. Bavirisetti, "Relative object tracking algorithm based on convolutional neural network for visible and infrared video sequences," in *Proceedings of the 4th International Conference on Virtual Reality*. ACM, 2018, pp. 44–49.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of Computer Vision and Pattern Recognition (CVPR). IEEE Conference on*. IEEE, 2009, pp. 248–255.

[34] D. P. Bavirisetti and R. Dhuli, "Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform," *IEEE Sensors Journal*, vol. 16, no. 1, pp. 203–209, 2016.