Available Online: 26 December 2025 DOI: 10.54254/2977-3903/2025.30878

Research on real-time target tracking method based on visible-thermal infrared multimodal fusion

Shiye Zhang

Gansu Police College, Lanzhou, China

374433287@qq.com

Abstract. This paper proposes a real-time suspicious person identification and tracking system based on visible-thermal infrared multimodal fusion. Aiming at the limitations of traditional tracking methods under extreme conditions, the system innovatively designs a multimodal feature fusion mechanism and a motion information-guided tracking compensation algorithm. Verified on the RGBT234 dataset, the system achieves a tracking success rate of 0.502, which is superior to existing mainstream methods, while maintaining a real-time processing speed of 15.3FPS, providing an effective technical solution for the field of public security.

Keywords: thermal infrared, real-time tracking, multimodal fusion, public security surveillance

1. Introduction

Online visual tracking is a core issue in computer vision and video processing, with wide applications in navigation, surveillance, robotics, traffic control, augmented reality and other fields. Over the past few decades, researchers have proposed various classic tracking frameworks, such as Incremental Visual Tracking (IVT) [1], Multiple Instance Learning (MIL) [2], Tracking-Learning-Detection (TLD) [3], Approximate Proximal Gradient L1 (APGL1) [4], Sparse Coding Model (SCM) [5], Adaptive Scale and Appearance Learning for Tracking (ASLAS) [6], Structured Output Tracking with Kernels (STRUCK) [7] and Kernelized Correlation Filters (KCF) [8]. These methods rely on handcrafted features or combine online learning algorithms, but handcrafted features have limited discriminative ability, resulting in poor tracking performance under extreme conditions, which is difficult to meet the needs of practical applications. In the past five years, deep learning has achieved remarkable results in computer vision, speech recognition, natural language processing and other fields by virtue of multi-layer nonlinear feature extraction. Trackers based on deep learning (such as Fully-Convolutional Networks for Tracking (FCNT) [9], Multi-Domain Network (MDNet) [10], Spatial-Temporal Context Tracking (STCT) [11]) have significantly improved tracking performance and repeatedly refreshed the records of the Object Tracking Benchmark (OTB)-100 dataset. In the Visual Object Tracking 2016 (VOT2016) competition, the top four trackers (Continuous Convolutional Operators for Tracking (C-COT) [12], Tracking Convolutional Neural Network (TCNN) [13], Saliency-Aware Siamese Tracking (SSAT) [14], Multi-Level Deep Feature (MLDF) [15]) are all based on deep networks. However, these methods are still prone to failure under conditions such as partial occlusion, illumination changes, complex backgrounds, motion blur or perspective changes, affecting robustness. This project designs a real-time suspicious person identification and tracking system based on visible-thermal infrared modalities. Based on the existing visible light tracking framework, the system proposes an efficient multimodal fusion mechanism to address the problems of large differences and difficult fusion of multimodal information; at the same time, aiming at the complexity of public security surveillance scenarios, it designs a motion information-guided tracking compensation mechanism to improve the robustness of tracking under extreme illumination and occlusion scenarios. Verified on existing surveillance data, the system performance has been effectively improved.

2. Invention purpose and basic ideas

For Red Green Blue-Thermal infrared (RGB-T) videos, this project realizes continuous tracking after calibrating the target in the first frame. Thermal infrared cameras perceive the infrared radiation of objects to measure temperature, enabling the system to effectively identify abnormal situations in complex environments. To improve accuracy and robustness, this study proposes a multimodal suspicious person identification and tracking system, which combines visible-thermal infrared spherical surveillance cameras to achieve automatic target tracking and identification. In terms of multimodal information fusion, an efficient Multimodal Feature Fusion Module (MMF) is designed to fully utilize the complementary advantages of different modalities; in terms of motion information utilization, a motion information-guided tracking compensation mechanism (AdaptAlign) is proposed to improve tracking accuracy in complex scenarios based on the historical motion of the target. The system also improves model optimization and computational efficiency to achieve real-time processing on edge devices. Experimental results show that the method has good robustness under harsh illumination, occlusion and complex weather conditions, providing effective technical support for public security surveillance and suspicious person detection.

3. Innovation points

Through multimodal feature fusion, the system deeply combines the high-resolution textural information of visible light with the illumination-insensitive temperature features of thermal infrared, realizing all-weather and accurate target positioning and tracking. When dealing with camouflaged, occluded and concealed targets in low-light environments, the system uses the body temperature characteristics of thermal infrared imaging to penetrate shallow shelters and identify human contours. Combined with the motion information-guided tracking compensation algorithm, it continuously locks the target trajectory in crowded scenes, effectively solving the detection blind spots of traditional surveillance under conditions of reflective clothing or shelters. At the same time, the system constructs a multi-dimensional feature database through spatiotemporal alignment of dual-modal data, synchronously recording target appearance, movement trajectory and body temperature information, providing reliable evidence support for crime scene reconstruction and judicial forensics. Furthermore, the system can be combined with a behavior recognition module to continuously analyze the behavior of suspicious targets, and use thermal infrared physical sign data to realize potential risk early warning, transforming from post-event tracing to pre-event intervention, and providing intelligent and real-time prevention and control means for border control, VIP protection and large-scale event security management.

4. Technical algorithms and main technical indicators

4.1. Technical benchmark algorithm

RT-MDNet is a real-time target tracking algorithm based on Multi-Domain Network (MDNet), aiming to improve tracking speed while maintaining accuracy. MDNet converts the tracking task into a foreground-background binary classification and adopts multi-domain learning: it shares the feature extraction network and fine-tunes it on different domains to enhance generalization ability, but its high computational complexity makes it difficult to meet real-time requirements. RT-MDNet is optimized on this basis, including using lightweight convolutional networks (such as VGG-M) for feature extraction, and designing independent fully connected layers for each domain for classification and regression. Its network structure is shown in Figure 1, including 3 convolutional layers, 1 adaptive RoIAlign layer and 3 fully connected layers: the input image first undergoes convolution to extract shared features, then RoIAlign is used to locate potential target regions, which are then sent to the fully connected layers to calculate confidence, and finally the region with the highest foreground confidence is selected as the tracking result. In terms of training strategy, the shared convolutional layers are trained on multiple domains to learn general features, and the fully connected layers are updated online during tracking to adapt to changes in target appearance. Data augmentation such as rotation and scaling is used to improve robustness. To optimize real-time performance, the algorithm prunes redundant convolutional kernels, quantizes floating-point parameters, and uses GPU parallel computing to achieve real-time tracking with both high accuracy and high efficiency.

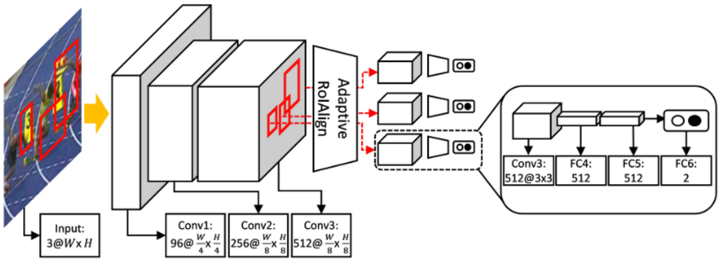

Figure 1. Network structure diagram of RT-MDNet

4.2. Main technical indicators

This project uses commonly used target tracking indicators to quantify model performance, including Tracking Success Rate and Tracking Precision. The Tracking Success Rate refers to the proportion of frames where the Intersection over Union (IoU) between the predicted target bounding box and the ground truth box exceeds a threshold t. The threshold is sampled between 0 and 1 to generate a success rate curve, and the Area Under the Curve (AUC) is used as the final indicator. The Tracking Precision measures the proportion of frames where the distance between the center point of the predicted box and the center point of the ground truth box is less than a set threshold. In the RGBT234 dataset, the center point distance threshold is set to 20 pixels to evaluate the precise positioning performance of the tracker.

5. Experimental design and results

This study uses the current mainstream dataset RGBT234 for evaluation. Comparisons are made with existing methods including JSR, KCF+RGBT, CFNet+RGBT, MEEM+RGBT, SOWP+RGBT, CSRDCF+RGBT [16], CFNet [17], MDNet [18]. The algorithm of this project achieves a tracking success rate of 0.502 on

RGBT234, which is higher than the comparison methods, proving its effectiveness. In addition, the proposed multimodal fusion module and the motion information-guided adaptive region of interest alignment module each achieve a 0.2% improvement in tracking success rate, further proving the effectiveness of the algorithm in this study.

The project has significant advantages in scientificity and advancement: first, it designs a suspicious person identification and tracking system based on visible-thermal infrared modalities. By improving existing multimodal fusion methods, it proposes an efficient information fusion mechanism to realize the complementary advantages of different modal information. Compared with existing methods, it improves tracking efficiency while maintaining low computational overhead, providing technical support for the deployment of intelligent surveillance systems; second, it proposes a motion information-guided tracking compensation mechanism, which uses the historical motion information of the target to achieve accurate tracking in complex scenarios such as partial occlusion, illumination changes, cluttered backgrounds, motion blur and perspective changes, significantly improving the all-time and all-weather surveillance robustness; finally, the system is continuously optimized during model construction and selection, enabling the tracking system to achieve real-time processing in real surveillance scenarios, expanding the application scope of the method, and verifying the effectiveness and practicality of the model.

5.1. Technical analysis explanation

Based on RT-MDNet, this project proposes a robust multi-domain network based on multimodal fusion for visible-thermal infrared specific target tracking/suspicious person identification. This study proposes a multimodal feature fusion module and a motion information-guided adaptive region of interest alignment module respectively. Compared with the benchmark algorithm, this paper uses two different VGG-M backbone networks to extract features from visible light and thermal infrared images, obtaining visible light image features and thermal infrared image features respectively. The feature extraction process can be expressed as:

$$F_{rgb} = \varphi_{rgb}(F_{rgb}) \tag{1}$$

$$F_t = \varphi_t(I_t) \tag{2}$$

The model includes two backbone networks for extracting features of visible light and thermal infrared modalities respectively. The extracted features are then input into the multimodal feature fusion module, which adopts an adaptive feature weighted fusion strategy: the research team assumes that the richer the information carried by the image, the greater the weight of its features should be. Therefore, the importance of each modal feature is first estimated through information entropy, and then weighted fusion is performed to realize the interaction between different modal features, thereby enhancing the feature representation ability and ensuring tracking performance. After processing by this module, the fused features are obtained, which can be expressed as:

$$F_{fusion} = MMF(F_{rgto}F_{\tau}) \tag{3}$$

On the basis of the Region of Interest (RoI) alignment method of the benchmark algorithm, the research team improved it and proposed a motion information-guided Adaptive Region of Interest Alignment (AdaptAlign) module. This module introduces the historical motion information of the target in the candidate region sampling process. First, the possible position of the target is predicted according to the target's speed and position, and sampling is performed near this position, thereby improving sampling accuracy, reducing the frequency of background region sampling, helping the model's online learning and reducing computational overhead. Subsequently, the extracted region of interest features are sent to the tracking head network to calculate the foreground confidence score to determine the precise position of the target.

5.2. Multimodal feature fusion module

Multimodal information can provide complementary information to significantly improve tracking robustness. Therefore, the research team first designed a Multimodal Feature Fusion Module (MMF), which adaptively generates fusion weights by evaluating the information volume of the two modalities, improving the flexibility of the fusion process. First, the research team obtained the features of the two modalities respectively, and then calculated the information entropy of the features to represent the accuracy of the data distribution. Its calculation method is expressed as:

$$H(F) = E[I(F)] = E[-ln(P(F))] \tag{4}$$

Among them, $P(F)$ represents the probability mass function, E represents the expected function of the variable, $I(F)$ represents the information volume of the feature, and $H(F)$ represents the accuracy of the feature. We believe that the greater the accuracy of the information, the greater the weight it should occupy in the fused features. Therefore, we perform a weighted average of the information entropies H_{rgb} and H_t of the two modalities as the fusion weights of the two modalities, which is expressed as follows:

$$W_{rgb} = \frac{H_{rgb}}{H_{rgb}+H_c} \tag{5}$$

$$W_t = \frac{H_t}{H_{rgb}+H_t} \tag{6}$$

Using the fusion weights of the two modalities, we can obtain the fused features F_{fusion} :

$$F_{fusion} = W_{rgb} \times F_{rgb} + W_t \times F_t \tag{7}$$

The adaptive feature fusion module proposed by the research team can dynamically evaluate the reliability of the current frame image, dynamically fuse visible-thermal infrared features, and significantly improve the fusion effect of the algorithm. Compared with traditional feature concatenation and feature averaging, this module can realize information fusion more flexibly, providing support for building an accurate and robust suspicious person identification system.

5.3. Motion information-guided adaptive region of interest alignment module

As shown in Figure 2, the research team observed that the Region of Interest (RoI) alignment method in the benchmark algorithm uses Gaussian distribution for sampling when extracting target regions, that is, the sampling probability near the target is higher, while the sampling probability at positions far from the target is lower. However, this random sampling strategy will introduce a large number of redundant candidate boxes (blue boxes, corresponding to background regions irrelevant to the target), increasing the difficulty of model learning, making the model prone to learning background noise, and reducing online learning efficiency. To improve the learning ability and efficiency of the model, the research team proposed a motion information-guided Adaptive Region of Interest Alignment Module (AdaptAlign) to improve the RoI alignment method in RT-MDNet. Specifically, the module first uses the target's historical trajectory information to predict the target position in the future frame through enhanced Kalman filtering, assuming that the target's motion acceleration is constant, thereby estimating the possible position of the target. Subsequently, Gaussian distribution sampling is performed centered on this expected position to improve the relevance of candidate regions, reduce background redundancy, and thus improve the accuracy of feature extraction and tracking performance.

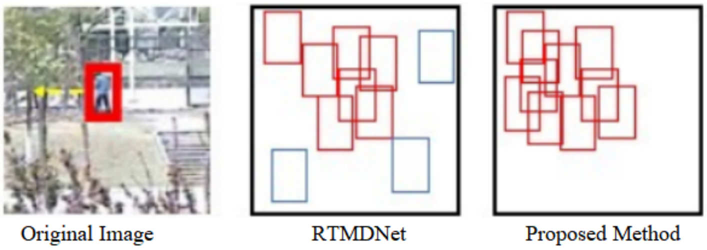

Figure 2. Comparison diagram of sampling strategies between RT-MDNet and proposed method

5.4. Experimental analysis

This study verified the proposed method on the public visible-thermal infrared tracking dataset RGBT234. The RGBT234 dataset was constructed by Li's research group from Anhui University in 2019, including 234 visible-thermal infrared videos with a total of about 233,000 frames. The longest video length can reach 8,000 frames, which is the largest RGB-T tracking dataset at that time. All videos are collected from surveillance cameras, covering public scenes such as urban traffic and schools, providing data support for the construction of public security systems. Data collection uses a rotatable pan-tilt equipped with a thermal infrared camera and a CCD camera (as shown in Figure 3(a)). The two cameras use the same imaging parameters, and the optical axis synchronization is achieved through a collimator (as shown in Figure 3(b)); the data registration method is shown in Figure 3(c)(d), ensuring the spatial alignment of visible light and thermal infrared modalities, providing a reliable foundation for multimodal tracking research.

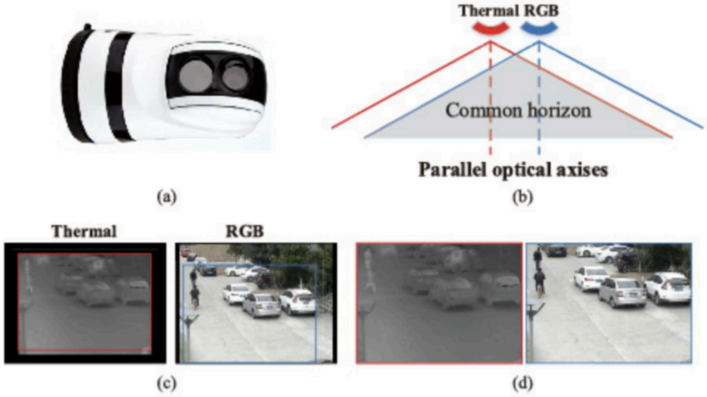

Figure 3. Schematic diagram of dataset acquisition system

The RGBT234 dataset has significant advantages: first, the ultra-large annotation scale (total frames of about 234,000, the longest single sequence can reach 8,000 frames) supports large-scale evaluation of tracking algorithms; second, the advanced imaging mechanism realizes precise alignment of visible light and thermal infrared cross-modality without preprocessing such as stereo matching [6, 7] or color correction [8]; third, the dataset provides three-level annotation states: no occlusion, partial occlusion and severe occlusion, facilitating the analysis of tracker performance under different occlusion conditions; finally, the matched camera parameters and parallel optical axis design can support both static and dynamic shooting while maintaining alignment accuracy, making it the current mainstream evaluation dataset for RGB-T tracking models.

5.5. Comparison results with existing methods

The method in this paper is compared with existing visible light target tracking methods and visible-thermal infrared target tracking methods. The visible-thermal infrared tracking methods include JSR [18], KCF+RGBT [8], CFNet+RGBT [17], MEEM+RGBT [19], SOWP+RGBT [20], CSRDCF+RGBT [21]. The visible light comparison methods include CFNet [17] and MDNet [22]. The experimental results are shown in the following table. The method in this paper achieves the highest performance among the comparison methods, with a tracking success rate of 0.502, which is a 0.4% improvement compared with RT-MDNet, as shown in Table 1.

Table 1. Comparison of our method and existing methods on the RGBT234 dataset

Method	Success Rate
JSR [18]	0.234
KCF+RGBT [8]	0.305
CFNet [17]	0.380
CFNet+RGBT [17]	0.390
MEEM+RGBT [19]	0.405
SOWP+RGBT [20]	0.451
CSRDCF+RGBT [21]	0.490
RT-MDNet [22]	0.498
Ours	0.502

5.6. Method analysis

In this section, we conduct a quantitative analysis of each module of the proposed method. The MDNet method using a single modality achieves tracking success rates of 0.498 and 0.496 respectively. The multimodal fusion module proposed in this paper can achieve a 0.2% improvement in tracking success rate. In addition, after adding the motion information-guided adaptive region of interest alignment module, the method in this paper further achieves a 0.2% improvement, as shown in Table 2. Compared with the tracking speed, the computational load brought by adding the two modules is negligible. The method in this paper achieves a speed of 15.3FPS. Compared with the benchmark algorithm, the method in this paper can improve tracking performance with a slight reduction in tracking speed.

Table 2. Ablation study analysis of our method

Method	Success Rate	Tracking Speed
MDNet-rgb	0.498	20.7
MDNet-t	0. 496	18.2
Our-MMF	0.500	16.5
Ours-MMF+ AdaptAlign	0.502	15.3

6. Typical application scenario analysis

Thermal imaging systems have significant advantages at night or in low visibility conditions, which can penetrate completely dark environments and identify human bodies (about 36.5℃) and vehicle engines (80–

120 ℃) heat sources within 200 meters. For example, in a coastal anti-smuggling operation in 2021, investigators used vehicle-mounted thermal imaging to find abnormal engine heat distribution of camouflaged fishing boats, revealing the modification facts. In rainy and foggy weather (visibility < 50 meters), the 8–14μm thermal infrared band is less affected by aerosol scattering. Combined with fog penetration algorithms, the monitoring distance can reach more than three times that of visible light systems. For stolen goods hiding scenarios, thermal infrared can detect abnormal heat conduction in confined spaces. For example, experiments show that the surface temperature difference of luggage containing electronic equipment can reach 2–3℃; in a cross-border money laundering case in 2022, investigators quickly located a metal money box buried 1.5 meters underground through thermal imaging, which had a continuous temperature difference of about 0.8℃ from the surrounding soil. In vegetation-covered environments, thermal radiation feature matching increases the human detection rate from 43% of visible light to 89%. In crowded places (such as exchanges and auction houses), the multispectral fusion system combines thermodynamic behavior models to identify sudden changes in body temperature (decrease by 0.5–1℃ when nervous) and local high-temperature areas (such as frequent operation of electronic equipment). Combined with visible light facial micro-expression analysis, it forms a multi-dimensional behavior feature vector. According to the data from the intelligent monitoring platform of a provincial economic investigation corps, this technology increases the accuracy of suspicious transaction identification by 37% and reduces the false alarm rate to 2.1%, fully reflecting the application value of thermal imaging in public security and suspicious behavior monitoring.

7. Conclusion

The research and verification of this system provide a practical reference paradigm for the application of multimodal fusion technology in the field of public security. Its design idea that balances accuracy and efficiency can provide stable and reliable technical support for scenarios such as security surveillance and key area protection. In the future, the dynamic fusion strategy of multimodal features can be further optimized, more challenging test scenarios such as complex weather and large-scale crowds can be expanded, and in-depth integration with edge computing and intelligent early warning systems can be explored to continuously improve the environmental adaptability and practical deployment value of the system, injecting new technical momentum into building a more intelligent and efficient public security protection system.

References

[1] Ross, D., Lim, J., Lin, R.-S., & Yang, M.-H. (2008). Incremental learning for robust visual tracking. *International Journal of Computer Vision, 77*(1), 125–141.

[2] Babenko, B., Yang, M.-H., & Belongie, S. (2009). Visual tracking with online multiple instance learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 983–990).

[3] Kalal, Z., Mikolajczyk, K., & Matas, J. (2012). Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(7), 1409–1422.

[4] Bao, C., Wu, Y., Ling, H., & Ji, H. (2012). Real-time robust L1 tracker using accelerated proximal gradient approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1830–1837).

[5] Zhong, W., Lu, H., & Yang, M.-H. (2012). Robust object tracking via sparsity-based collaborative model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1858–1865).

[6] Jia, X., Lu, H., & Yang, M.-H. (2012). Visual tracking via adaptive structural local sparse appearance model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1822–1829).

[7] Hare, S., Saffari, A., & Torr, P. H. S. (2011). Struck: Structured output tracking with kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(10), 2096–2109.

[8] Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2015). High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(3), 583–596.

[9] Wang, L., Ouyang, W., Wang, X., & Lu, H. (2015). Visual tracking with fully convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3119–3127).

[10] Nam, H., & Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4293–4302).

[11] Wang, L., Ouyang, W., Wang, X., & Lu, H. (2016). STCT: Sequentially training convolutional networks for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1373–1381).

[12] Danelljan, M., Robinson, A., Khan, F. S., & Felsberg, M. (2016). Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *Proceedings of the European Conference on Computer Vision* (pp. 472–488).

[13] Nam, H., Baek, M., & Han, B. (2016). Modeling and propagating CNNs in a tree structure for visual tracking. In *Proceedings of the European Conference on Computer Vision* (pp. 171–187).

[14] Teng, Z., Xing, J., Wang, Q., Lang, C., Feng, S., & Jin, Y. (2017). Robust object tracking based on temporal and spatial deep networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1144–1153).

[15] Wu, L., Xu, T., Zhang, Y., Wu, F., Xu, C., Li, X., & Wang, J. (2021). Multi-channel feature dimension adaption for correlation tracking. *IEEE Access, 9*, 63814–63824.

[16] Danelljan, M., Hager, G., Shahbaz Khan, F., & Felsberg, M. (2015). Learning spatially regularized correlation filters for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4310–4318).

[17] Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., & Torr, P. H. (2017). End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5000–5008).

[18] Liu, H., & Sun, F. (2012). Fusion tracking in color and infrared images using joint sparse representation. *Information Sciences, 55*(3), 590–599.

[19] Zhang, J., Ma, S., & Sclaroff, S. (2014). MEEM: Robust tracking via multiple experts using entropy minimization. In *Proceedings of the European Conference on Computer Vision* (pp. 188–203).

[20] Kim, H.-U., Lee, D.-Y., Sim, J.-Y., & Kim, C.-S. (2015). SOWP: Spatially ordered and weighted patch descriptor for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3011–3019).

[21] Lukezic, A., Vojir, T., Cehovin, L., Matas, J., & Kristan, M. (2017). Discriminative correlation filter with channel and spatial reliability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4847–4856).

[22] Jung, I., Son, J., Baek, M., & Han, B. (2018). Real-time MDNet. In *Proceedings of the European Conference on Computer Vision* (pp. 83–98).