Journal Pre-proof

Object Fusion Tracking Based on Visible and Infrared Images: A Comprehensive Review

Xingchen Zhang, Ping Ye, Henry Leung, Ke Gong, Gang Xiao

PII: S1566-2535(20)30265-7
DOI: https://doi.org/10.1016/j.inffus.2020.05.002
Reference: INFFUS 1233

To appear in: *Information Fusion*

Received date: 6 October 2019
Revised date: 6 March 2020
Accepted date: 2 May 2020

Please cite this article as: Xingchen Zhang, Ping Ye, Henry Leung, Ke Gong, Gang Xiao, Object Fusion Tracking Based on Visible and Infrared Images: A Comprehensive Review, *Information Fusion* (2020), doi: https://doi.org/10.1016/j.inffus.2020.05.002

Highlights

- A review of fusion tracking methods via visible and infrared images is presented
- Main RGB-infrared trackers are summarized and categorized into several groups
- Public RGB-infrared datasets are summarized and compared
- Main results on public datasets are summarized and analyzed in detail
- Future prospects of RGB-infrared fusion tracking are discussed and suggested

Object Fusion Tracking Based on Visible and Infrared Images: A Comprehensive Review

Xingchen Zhang[a], Ping Ye[a], Henry Leung[b], Ke Gong[a], Gang Xiao[a,*]

[a]*School of Aeronautics and Astronautics, Shanghai Jiao Tong University, No. 800 Dongchuan Road, Shanghai, 200240, China*

[b]*Department of Electrical and Computer Engineering, University of Calgary, 2500 University Drive NW Calgary, Alberta, Canada T2N 1N4*

Abstract

Visual object tracking has attracted widespread interests recently. Due to the complementary features provided by infrared images, the fusion tracking based on visible and infrared images can boost the tracking performance under adverse challenging conditions. RGB-infrared fusion tracking has become an active research topic and various algorithms have been proposed to perform RGB-infrared fusion tracking in recent years. In this paper, we present a review on RGB-infrared fusion tracking. We summarize all major RGB-infrared trackers in the literature and categorize them into several major groups for better understanding. We also discuss the development of RGB-infrared datasets, and analyze the main results on public large scale datasets. We observe that deep learning-based methods give the state-of-the-art performances. Also, the graph-based and correlation filter-based methods give a bit worse but still competitive performances. In conclusion, we give some suggestions on future research directions of fusion tracking based on our observations.

Keywords: fusion tracking, deep learning, object tracking, Siamese network, correlation filter

1. Introduction

Visual object tracking has received a great amount of attention in recent years due to its wide applications in many areas, such as robotics [1], autonomous vehicles [2], human-computer interface [3] and video surveillance [4]. According to the type of images included, it can be roughly classified into tracking based on visible images, tracking based on infrared images, and RGB-infrared fusion tracking. Among these types, the most popular is tracking based on visible images. Note that, in this paper we do not distinguish between visible and RGB (Red-Green-Blue) images, although the visible images also contain gray-scale images.

*Corresponding author
Email address: xiaogang@sjtu.edu.cn (Gang Xiao)

(a) Target in infrared images (bottom) is more clear (b) Target in visible images (top) is more clear

Figure 1: Examples of complementary information from visible and infrared images [17].

Currently, two main kinds of methods in visual object tracking are deep learning (DL)-based methods [5] and correlation filter (CF)-based approaches [6]. Tracking methods based on deep learning mainly utilize its strong feature representation ability to extract better features than handcrafted ones, thus these approaches can achieve good tracking results in many cases. Here handcrafted ones means the features that are designed manually, such as histogram of oriented gradients (HOG) [7] and scale invariant feature transform (SIFT) [8]. Previously, deep learning-based methods suffer from slow speed severely [9]. However, with the application of fully convolutional Siamese networks in tracking [5], recent deep learning-based trackers achieve high performance tracking results while maintaining real-time speeds [10–12]. In CF-based tracking algorithms, the model can be updated in real-time as the correlation operation can be efficiently implemented via the fast Fourier transform (FFT). Therefore, during tracking process, CF-based methods utilizing shallow features can run in real-time. However, recently some CF-based trackers use raw deep convolutional features which are of high dimensionality [13–15]. These trackers become slower and slower because the computational time for the correlation filters increases with the feature dimensionality [16].

However, due to the limitation of the imaging mechanism of visible images, tracking algorithms based on visible images may fail as they may be unreliable in certain circumstances. For example, when the illumination conditions are poor or change significantly. The infrared images detect thermal information of objects and are insensitive to these factors. They can provide complementary information to visible images, as shown in Fig. 1(a). In recent years, researchers also explore performing object tracking with infrared images [18–21]. However, the infrared images typically have low resolution and poor texture, and are also unreliable in certain conditions as shown in Fig. 1(b). Therefore, more researchers begin to investigate object tracking method based on the fusion of visible and infrared images to overcome the inherent shortcomings of the methods based on single-modal images. By fusing complementary information from visible and infrared images, the robustness of tracking algorithms can be greatly enhanced. As a result, in recent years, object

tracking based on RGB and infrared images have become a hot research topic. An increasing number of researches have been published in high quality journals or well-known conferences [22–30]. As a consequence, the well-known visual object tracking challenge (VOT) started a new RGB-infrared subchallenge in 2019 [1], aiming to attract researchers to evaluate the performances on provided video sequences. Note that since the appearance of tracking based on visible and infrared images, it did not have a consistent name. A large part of researchers used fusion tracking [31, 32] or tracking by fusion [33–35]. It was until 2017 that some researchers started using RGBT tracking [27]. In this review, we denote the object tracking based on the fusion of visible and infrared images as RGB-infrared fusion tracking, because we think it can cover this kind of methods better and is thus more suitable for a comprehensive review. Besides, by using this name we aim to emphasize the importance of fusion in this kind of methods.

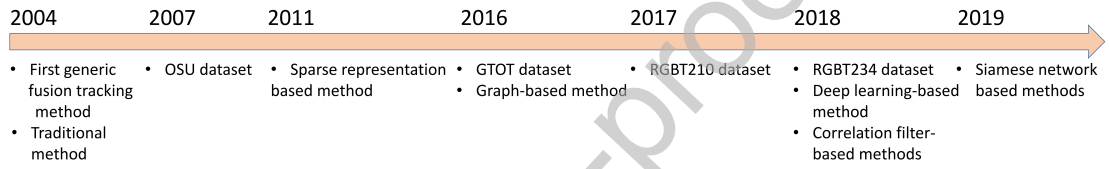

Figure 2: Development timeline of RGB-infrared fusion tracking.

The research on RGB-infrared fusion tracking has begun in 2000s and a development timeline of this field is given in Fig. 2. RGB-infrared fusion tracking can be categorized into different categories. According to the primary modality utilized during fusion tracking, there are infrared-assisted RGB tracking and RGB-assisted infrared tracking. In infrared-assisted RGB tracking, visible image is the primary modality. Infrared images are employed for assisting RGB tracking, especially when the visible images are not reliable [36, 37]. In these works, the evaluation metrics are evaluated based on the ground truth of visible images. In contrast, in RGB-assisted infrared tracking, infrared image is the primary modality and all evaluation metrics need to be computed based on the infrared ground truth [38]. In this paper, we broadly divide the RGB-infrared fusion tracking methods into five categories according to their adopted theories, namely traditional methods, sparse representation (SR)-based, graph-based, correlation filter-based and deep learning-based approaches. It is well known that effective and robust feature representation is crucial for a tracking algorithm. Before sparse representation-, graph-, correlation filter- and deep learning-based methods, researchers performed fusion tracking using traditional techniques such as such as mean shift, Camshift, Kalman filter, and particle filter. Traditional methods utilize handcrafted features to represent the target. Sparse representation-based

[1] http://www.votchallenge.net/vot2019/index.html

methods work on the basis of possible representation of the target with linear combinations of bases in overcomplete dictionaries. Graph-based approaches firstly divide the bounding box around the target to non-overlapping patches, and then build the relationship among these patches to work out a feature representation of the target. CF-based trackers learn correlation filters online efficiently to adapt to variation of the target. Deep learning-based methods leverage the strong feature representation ability of deep neural networks to learning robust feature representation of the target from a large amount of images. In all these methods, a key to good fusion tracking performance is the effective combination of visible and infrared features.

As can be seen from Fig. 2, RGB-infrared fusion tracking is developing very fast. However, to the best of our knowledge, there is still a lack of review on RGB-infrared fusion tracking in the literature that gives a comparison and evaluates the performance of these different techniques. This paper tries to fill this gap. The main contribution of this review is in several aspects. First, to the best of our knowledge, this is the first review on RGB-infrared fusion tracking. This manuscript systematically investigate the RGB-infrared fusion tracking methods, benchmark datasets, and evaluation metrics. Main RGB-infrared tracking algorithms are grouped into several types according to their corresponding theories and each kind is introduced in detail, including the main principles, representative methods as well as pros and cons. Second, main results on public datasets are presented and analyzed in this review to provide an objective comparison of the existing approaches. Third, based on the systematically review of almost all RGB-infrared fusion tracking methods and the performance comparison of different trackers, we give detailed discussions on future prospects and provide suggestions on promising research directions of this field.

The structure of this review is schematically illustrated in Fig. 3. Section 2 gives some background information. In Section 3, RGB-infrared fusion tracking methods will be discussed in detail, including key points in implementation and categorizing the main methods. In Section 4, we summarize the development of RGB-infrared datasets. Section 5 introduces the evaluation metrics. Section 6 presents experimental results and gives an analysis on the performances. Sections 7 discusses the future prospects. Finally, Section 8 concludes the paper.

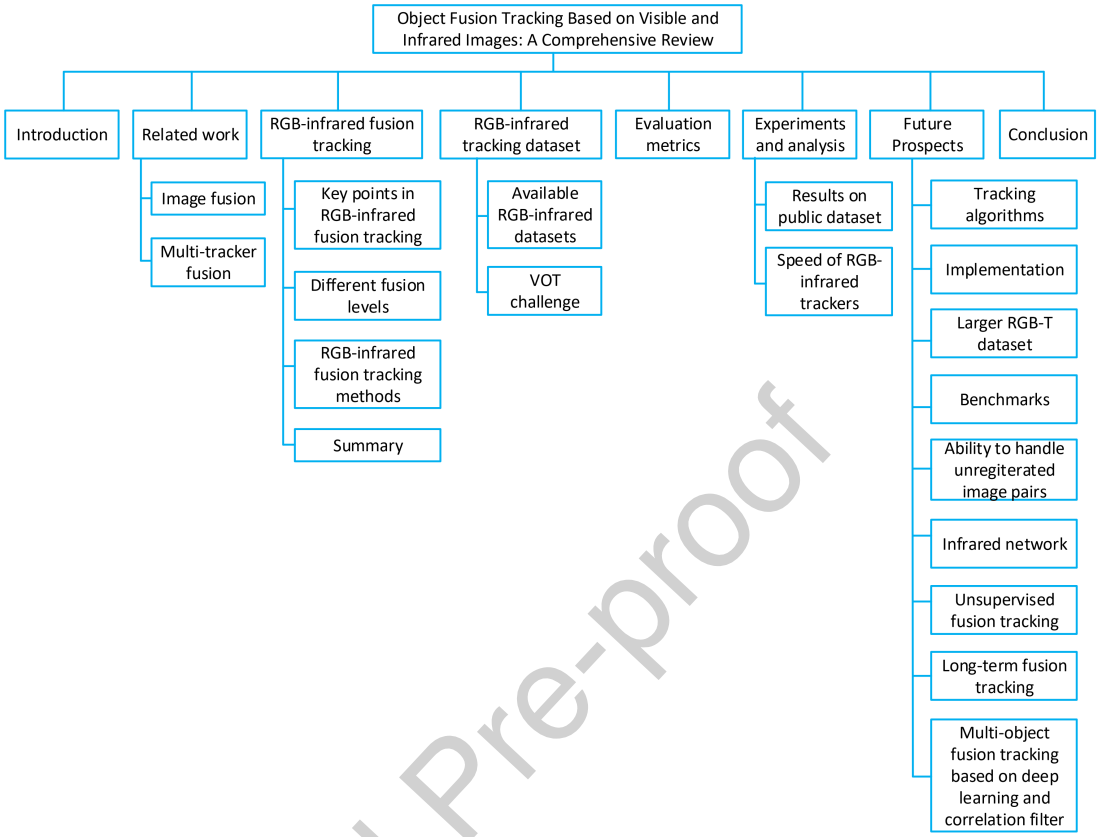

Figure 3: Structure of this review.

2. Related work

This section discusses some related work which is helpful for understanding and performing RGB-infrared fusion tracking.

2.1. Image fusion

Image fusion aims to combine information from multiple images into a single image 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 approaches. Also, image fusion can either be performed in the spatial domain or transform domain. Image fusion technology can be applied in various areas, such as medical image fusion [39], multi-focus image fusion [40], remote sensing image fusion [41], multi-exposure image fusion [42], visible and infrared image fusion [43].

Before deep learning is introduced to image fusion community, the main image fusion methods include weighted average method [44], wavelet transform based method [45], PCA-based method [46], sparse representation method [47] and compressed sensing method [48]. In the past few years, a number of image fusion methods based on deep learning have emerged [49–52]. Deep learning can help to solve several important problems in image fusion. First, deep learning can provide better features compared to handcrafted ones. Second, deep learning can learn adaptive weights in image fusion, which is crucial in many fusion rules. Regarding methods, convolutional neural network (CNN) [53–57], generative adversarial networks (GAN) [58], Siamese networks [59], autoencoder [60] have been explored to conduct image fusion. Apart from image fusion methods, the image quality assessment, which is critical in image fusion performance evaluation, has also benefited from deep learning [61]. It is foreseeable that in the future, image fusion technology will develop in the direction of machine learning, and an increasing number of research results will appear.

It should be noted that the aim of fusion tracking is different from image fusion. In image fusion algorithms, the goal is to produce a fused image which has better visual quality. However, in fusion tracking, only information about the target and its surroundings are important and thus should be extracted by tracking algorithms.

2.2. Multi-tracker fusion tracking

Different from RGB-infrared fusion tracking, some researchers have also investigated multi-tracker fusion tracking. Here multiple trackers are all based on visible images. The main idea of multi-tracker fusion is to leverage complementary features of different trackers, based on the fact that different trackers can handle different kinds of challenges encountered during tracking. Therefore, by selecting a proper tracker for each video sequence can lead to a stronger tracker with better overall performance. For example, Baile et al. studied fusion tracking method that combines the tracking results of different tracking algorithms [62]. This method only requires tracking results of different tracking algorithms (in the form of target marker boxes) as the input. Biresaw et al. proposed a tracker-level fusion method, which aims to fuse the tracking results of different trackers to achieve better tracking results [63]. Vojir et al. proposed a tracking result fusion method based on Markov model [64]. Gundogdu et al. proposed a tracking algorithms that is tuned to IR data, by using an ensemble of base trackers [65]. At any time, only one base tracker actively learns the target appearance. In 2019, Xie et al. designed a multi-tracker fusion method via adaptive outliers detection [66]. Yang et al. proposed a parallel correlation filter-based trackers, which utilizes two correlation filter trackers and fuses the results of them according to the designed fusion rule [67]. Experiments show

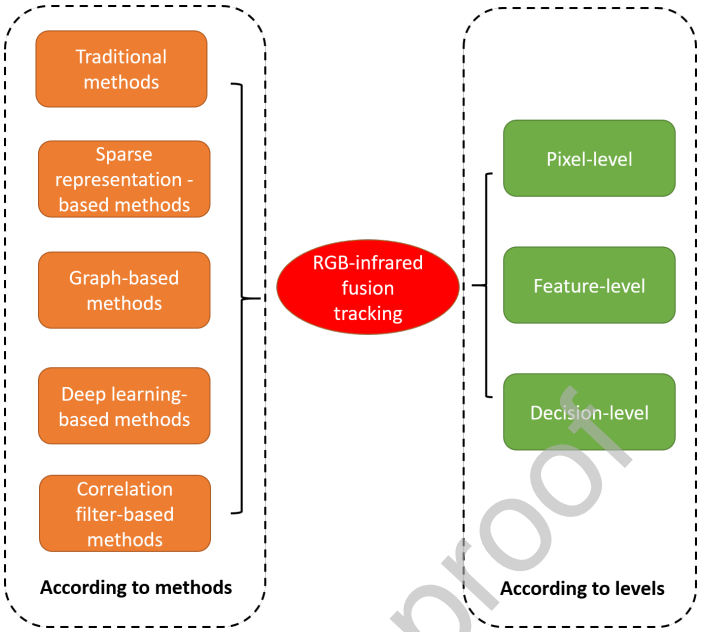

Figure 4: Categories of RGB-infrared fusion tracking methods.

that the method can achieve better performance than that of individual tracker. However, a shortcoming is the lack of reasonable basis for judging the reliability of different trackers.

Indeed, RGB-infrared fusion tracking is a special case of multi-tracker fusion tracking by fusing an RGB tracker and an infrared tracker. Therefore, some ideas of multi-tracker fusion can be applied in RGB-infrared fusion tracking, especially the decision-level RGB-infrared fusion tracking which will be introduced in Section 3.2.3.

3. RGB-infrared fusion tracking

In recent years, a lot of RGB-infrared fusion tracking algorithms have been proposed and some examples are listed in Table 1. In this section, we firstly discuss the key points to good fusion tracking performance. Then, we introduce the fusion levels which can be employed in fusion tracking. According to when the images are fused, they can be divided into pixel-level, feature-level and decision-level fusion tracking. We then give a comprehensive survey on RGB-infrared fusion tracking methods. These methods are divided into five categories according to their corresponding theories, and we review each kind in detail. Figure 4 shows categories of RGB-infrared fusion tracking.

Table 1: Examples of recent published researches on RGB-infrared fusion tracking

References	Year	Journal/Conference	Category
[25]	2016	TIP	SR-based
[68]	2016	MMM	SR-based
[69]	2017	IEEE TRANS. SYST., MAN, CYBERN., SYST.	SR-based
[27]	2017	ACM Multimedia	Graph-based
[70]	2018	Chinese Conference on Image and Graphics Technologies	SR-based
[24]	2018	PRL	SR-based
[28]	2018	ECCV	Graph-based
[71]	2018	arXiv	Graph-based
[29]	2018	TCSVT	Graph-based
[72]	2018	SPIC	Graph-based
[73]	2018	SPIC	Traditional
[35]	2018	Applications of Digital Image Processing XLI	CF-based
[74]	2018	PRCV	CF-based
[75]	2018	ICVR	DL-based
[76]	2018	CISP-BMEI	DL-based
[77]	2018	arXiv	DL-based
[26]	2018	Neurocomputing	DL-based
[23]	2018	AAAI	DL-based
[22]	2019	IEEE T IND ELECTRON	DL-based
[38]	2019	IEEE Access	DL-based
[78]	2019	Infrared Physics & Technology	Traditional
[30]	2019	Neurocomputing	CF-based
[79]	2019	Infrared Physics & Technology	CF-based
[80]	2019	Mathematical Problems in Engineering	CF-based
[81]	2019	Laser & Optoelectronics Progress	DL-based
[36]	2019	Fusion	DL-based
[37]	2019	Chinese Conference on Information Fusion	DL-based
[82]	2019	ACM International Conference on Multimedia	DL-based
[83]	2019	IEEE International Conference on Image Processing	DL-based
[84]	2019	ICCV Workshop	DL-based

3.1. Key points in RGB-infrared fusion tracking

There are several key points in RGB-infrared fusion tracking algorithms, which are vital for achieving good performance. These key points are discussed as follows:

- Registration of visible and infrared images. Currently, almost all RGB-infrared trackers in the literature require that the visible and infrared images being registered.
- Reliability of different modal images. To leverage the complementary information provided by visible and infrared images, one should know the reliability of different modal images. In other words, the algorithm needs to know when the visible images are reliable and when the infrared images are reliable.
- How to leverage complementary information effectively. This means when and how to fuse features of visible and infrared images. Fusion tracking can also be performed at different levels, which will determine the performance of fusion tracking. Besides, various fusion rules can be designed to fuse multi-modal information together.

- Speed of fusion tracking algorithm. The fusion of complementary features will take some CPU time, and the speed of fusion tracking becomes slower. How to fuse features efficiently and perform the fusion tracking quickly, is another point that should be noted when investigating RGB-infrared fusion tracking algorithms.

3.2. Different fusion levels

3.2.1. Pixel-level fusion tracking

Pixel-level fusion tracking, which is also called fusion-before-tracking approach, means that the images of different modalities are firstly fused into more informative images, then object tracking is conducted based on these fused images, as shown in Fig. 5. Pixel-level fusion tracking is easy to implement since there are many image fusion codes and trackers available online. However, pixel-level image fusion keeps most information from source images, thus it is relatively time consuming. As a result, if a pixel-level image fusion algorithm is time consuming, it will slow down the whole fusion tracking algorithm significantly. In addition, the image fusion approach has huge impact on tracking results [33, 85–87].

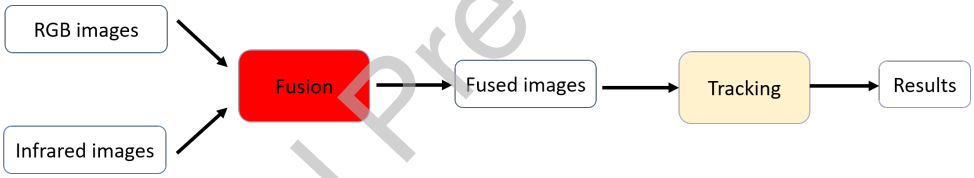

Figure 5: Pixel-level RGB-infrared fusion tracking.

3.2.2. Feature-level fusion tracking

The middle level fusion tracking algorithm is based on feature-level fusion. In feature-level fusion tracking, the features of RGB and infrared images are extracted and then fused according to designed fusion rule to obtain the fused feature. Tracking is then performed with this fused feature. Normally, the fused feature is more informative than the individual features, thus the performance can be improved. The basic process is illustrated in Fig. 6. The feature-level fusion tracking directly construct multi-modal features thus it is more straightforward than pixel-level method. The key of feature-level fusion tracking lies in the extraction of visible and infrared images, and the effective fusion of them.

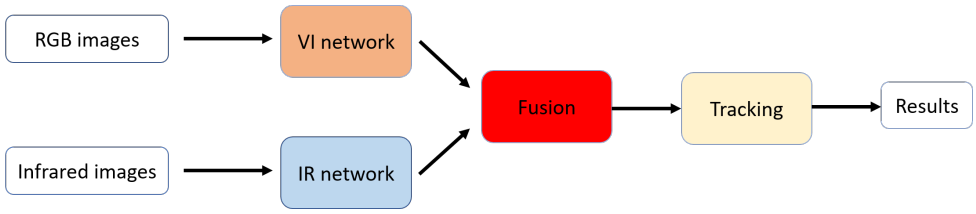

Figure 6: Feature-level RGB-infrared fusion tracking.

3.2.3. Decision-level fusion tracking

The highest level fusion tracking is the decision-level fusion tracking, which is also known as tracking-before-fusion approach. As shown in Fig. 7, tracking is performed in individual modalities and then the results are fused to obtain the final results. Decision-level fusion tracking algorithms have some advantages. First, different trackers can be chosen to perform tracking based on visible and infrared images, respectively. The only thing that most decision-level fusion tracking methods require are the bounding box around the target. Second, the computational cost is usually less compared to pixel-level and feature-level fusion tracking methods. As a consequence, the tracking speed could be faster than pixel and feature-level fusion tracking. In addition, decision-level fusion tracking has less requirements on the registration of visible and infrared images.

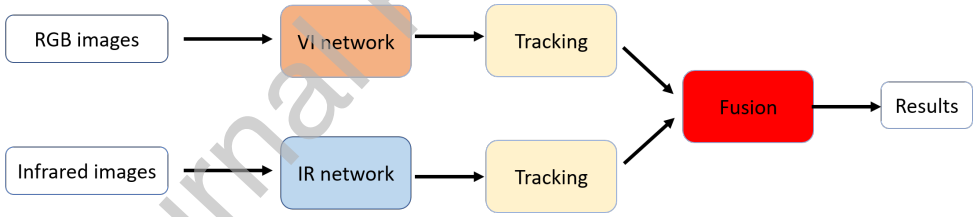

Figure 7: Decision-level RGB-infrared fusion tracking.

3.3. RGB-infrared fusion tracking methods

3.3.1. Traditional fusion tracking methods

Features play an important role in the performance of a tracking algorithms. When researchers started to investigate RGB-infrared fusion tracking in 2000s, they utilized handcrafted features such as histogram of oriented gradients (HOG), scale-invariant feature transform (SIFT), local binary pattern (LBP). Besides, in these methods, traditional tracking techniques, such as Kalman filter, particle filter and mean shift, are employed to perform tracking.

Kalman filtering can be utilized to perform fusion tracking. Bunyak et al. presented a moving object

detection and tracking system that robustly fused infrared and visible video within a level set framework [88]. The long-term trajectories for object clusters were estimated using Kalman filtering and watershed segmentation. Yun et al. proposed a compressive time-space Kalman fusion tracking algorithm to extend the compressive tracking (CT) method to the case of fusion tracking using visible and infrared images [89]. The proposed fusion model was completed in both space and time domains. The general framework of that method is given in Fig. 8. The shortcoming of Kalman filter is that it is limited to the case of linear dynamics and Gaussian noise. However, very few practical visual tracking problems belong to this case.

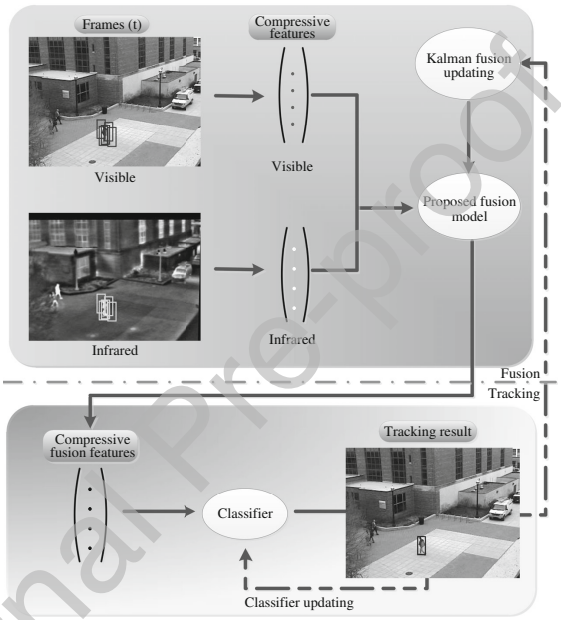

Figure 8: The framework of the compressive time-space Kalman fusion tracking algorithm [89]. Solid and dotted lines denote input and feedback directions, respectively.

Compared to Kalman filter, the particle filter relaxes the restriction of linearity and Gaussianity. This property is very useful in visual tracking, where most cases are non-linear and non-Gaussian. It was firstly used in object tracking by Isard et al. [90]. Several works have been published which utilized particle filter to perform RGB-infrared fusion tracking. For example, in 2007, Cvejic et al. studied the effect of pixel-level image fusion methods on object tracking results [91], as compared to tracking in single modality videos. In that work, tracking was performed via particle filter method. In 2008, Liu and Sun proposed a fusion tracking method based on sequential propagation algorithm [31]. In that work, the covariance feature was utilized to construct the likelihood function under the framework of particle filter. Also, the fusion was automatically realized by sequential belief propagation. Peteri et al. also presented a particle filter-based fusion tracking

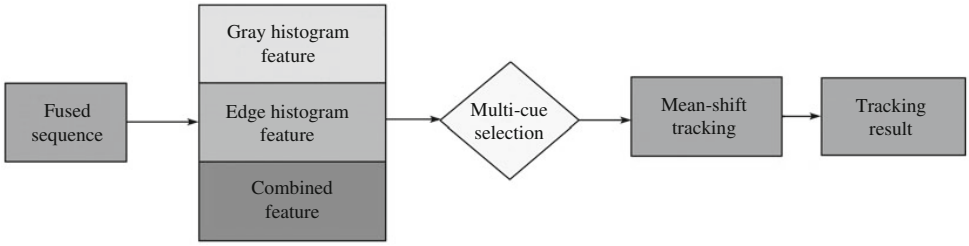

Figure 9: The multi-cue mean-shift tracking approach [95].

method with visible and infrared images [92]. The particle filter searched color or temperature features similar to the target in visible and infrared images, respectively. This method assumed that the visible and infrared images were strictly registered. Besides, it assumed that the target was trackable in at least one modality. Wang et al. proposed a local discrimination analysis based infrared and visible cooperative tracking approach, where a particle filter tracking framework was employed [93]. The Fisher linear discrimination theory was introduced to design the discriminative function between the target and background in local regions. Based on which, the fusion of visible and infrared images was conducted in feature-level. Xiao et al. proposed a tracking-before-fusion approach to perform fusion tracking based on visible and infrared images [94]. Specifically, tracking was performed in visible and infrared images separately, and then the tracking results were fused to obtain the final tracking results. Compared to methods which firstly fuse images and then perform tracking, the tracking-before-fusion method has less restricted requirement on the registration of image pairs. In that work, target tracking in visible images was performed by an improved particle filter method, and tracking in infrared images was conducted using an improved template matching scheme.

Mean shift method is also utilized in RGB-infrared fusion tracking. Conaire et al. proposed a framework that can efficiently combine features for robust tracking based on fusing the outputs of multiple spatiogram trackers, using mean-shift approach [96]. Zhang et al. also proposed a fusion tracking algorithm for visible and infrared object based on mean shift [97]. The method firstly weighted the similarity of the visible light object and the similarity of the infrared target to construct a new objective function. Object fusion tracking was then achieved through a core tracking method. However, mean-shift is a local deterministic search strategy, which is easy to be trapped into local minimal, and difficult to recover from tracking failure. Wang et al. extracted color information of the visible image and gray information of infrared image as the parameters of the target model, and used the mean shift algorithm to achieve tracking [98]. Xiao et al. developed a multi-cue mean-shift target tracking approach based on the fuzzified region dynamic image fusion [95]. In

that work, visible and infrared images were firstly fused into fused images, and then a multi-cue mean-shift tracking (MMT) algorithm was utilized to perform tracking. Here multi-cue included gray histogram feature, the edge histogram feature and the combined feature, as illustrated in Fig. 9.

Other traditional methods were also utilized for performing RGB-infrared fusion tracking. For instance, Kumar et al. [99] proposed a pixel-level fusion tracking for UAV-based target tracking. An optical flow technique based on Horn-Schunck method was utilized for tracking. Schnelle et al. [33] conducted a feasibility study on fusing visible and infrared images using 13 spatial domains and pyramid-based pixel-level fusion algorithms. In that work, a tracker based on background subtraction was employed. The same tracker was also employed by Chan et al. [86, 87]. Background subtraction-based tracker was also used by Mangale et al. [100] to perform camouflaged target detection and tracking and by Laurent et al. [101] in an automatic context-independent video monitoring system with visible and infrared sensors. Besides, Ding et al. [73] proposed a fusion tracking method based on the locality-constrained linear coding (LLC) and the saliency map.

However, these traditional fusion tracking methods have some shortcomings, which limit their performances. First and most important, features utilized in tracking are manually extracted or designed, which may not be effective in many cases. The limitation of hand-crafted features have well been demonstrated by far. As a consequence, these trackers cannot deal with real challenges during tracking well, such as scale change and fast motion. Second, the above methods are computationally intensive, especially the particle filter-based ones. Almost none of the above-mentioned trackers can meet the real-time requirement. Last but not the least, in above-mentioned papers, they usually just test the proposed algorithms using several or even one single video, which is obviously not enough to fully evaluate the performance on different challenges.

3.3.2. Sparse representation-based methods

Sparse representation is an effective tool for characterizing the human visual system [52]. It was firstly introduced into object tracking by Mei and Ling [102], since it is capable of suppressing noise and errors. It has also been utilized in RGB-infrared fusion tracking since it can help to create effective joint features in fusion tracking. The key of sparse representation-based fusion tracking method lies in the sparse representation of joint features. Besides, the modality reliability is also important for achieving good tracking performance.

Wu et al. concatenated image patches from grayscale and thermal modalities into a one-dimensional vector, and then represented it sparsely in the target template space [103]. To he best of our knowledge, this is the first sparse representation-based fusion tracking method. Liu and Sun also proposed an approach based on joint sparse representation [32]. In this work, the joint sparse representation was used to design similarity

between samples and the likelihood function of particle filter fusion tracker. Li et al. [68] proposed an online grayscale-thermal tracking method via Laplacian sparse representation in Bayesian filtering framework. A generative multimodal feature model was induced by the Laplacian sparse representation, which utilized the local patch similarities to improve the robustness. Besides, the reverse representation and parallel computation were adopted to improve tracking speed. In 2018, Li et al. [70] proposed a fusion tracking algorithm based on the cross-modal sparse representation in the Bayesian filtering framework. In this work, both the intra and inter-modality constraints were taken into account in the sparse representation. Unlike previous sparse representation-based methods, the reconstruction residues and coefficients were employed together to define the likelihood probability for each candidate sample generated by the motion model. The object location was indicated by the candidate sample with the maximum likelihood probability. Lan et al. [24] designed a modality-correlation-aware sparse representation model for RGB-infrared object tracking. This method exploited the correlation of different modalities via the low rank regularization and adaptively selected representative templates to deal with appearance changes via the sparsity regularization, which made it more able to perform effective modality fusion and handle large appearance changes. However, in this method, modality reliability was not considered.

However, in these sparse representation-based approaches, modality reliability was not considered thus may limit the tracking performance in dealing with occasional perturbation or malfunction of individual modalities.

There are some sparse representation-based approaches that consider modality reliability to utilize complementary information more effectively. Li et al. [25] proposed a collaborative sparse representation (CSR) method for grayscale-thermal tracking in Bayesian filtering framework. In this work, modality reliability was considered by introducing the weight variables in the collaborative sparse representation. The sparse codes and weights variables were jointly optimized in an online way. Fig. 10 illustrates the modality weights computed in this method. As can be seen, the modality weights can indicate the reliability of visible and infrared images. An improved version of [68] via multitask Laplacian sparse representation was also presented by Li et al. [69], by taking modality weights into account. In this method, the modal reliability for each frame is computed based on the observation that more reliable modality is more discriminative between target object and its surrounding regions, as illustrated in Fig. 11.

One of the main shortcomings of sparse representation-based fusion trackers is the efficiency. Almost all reported sparse representation fusion tracking algorithms can not meet the real time requirement. This may because that the online optimization process is time-consuming in sparse representation-based method.

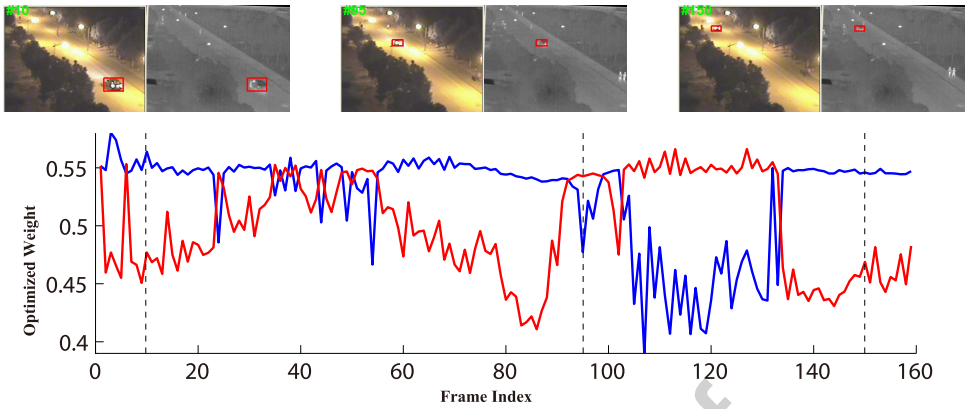

Figure 10: Illustrations of modality weights. The blue and red curve indicate the weights of visible and infrared images, respectively. It can be seen that the weights of visible and infrared images are almost consistent with their reliabilities. Therefore, the method can still work even though one modality has occasional perturbation [25].

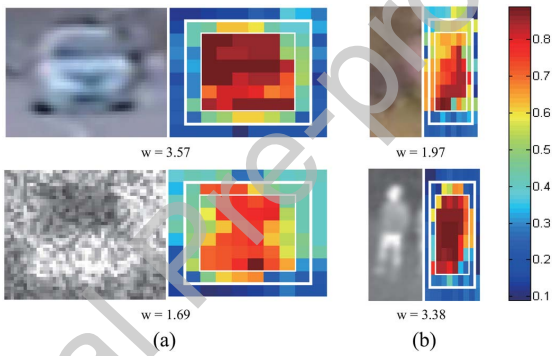

Figure 11: Examples of modality weights computed in [69]. The weight of each image is shown under the corresponding image. (a) In the top image, the target is more discriminative form its surroundings, thus having a larger weight. (b) In the bottom image, the target is more clear than the top image, thus having a larger weight. In each subfigure, the left column is image patches while the right column is the corresponding graphs. The graphs consist of nonoverlapping patches where the color of each small patch indicates the weight that the patch belongs to the foreground.

3.3.3. Graph-based methods

In the past few years, graph-based approaches have been applied to RGB-infrared fusion tracking. In graph-based methods, the bounding box of the target is firstly partitioned into non-overlapping patches. By doing this, the bounding box was represented with a graph with image patches as nodes. Then, each patch is assigned a foreground weight to suppress background information. In other words, the weight indicates whether the patch belongs to the foreground or background. Besides, weights are also assigned between each two patches to denote the relationship between them, termed edge weight. The graph and weights are dynamically learned during tracking to adapt to the variations of target. A general framework of graph-based RGB-infrared tracking method is illustrated in Fig. 12.

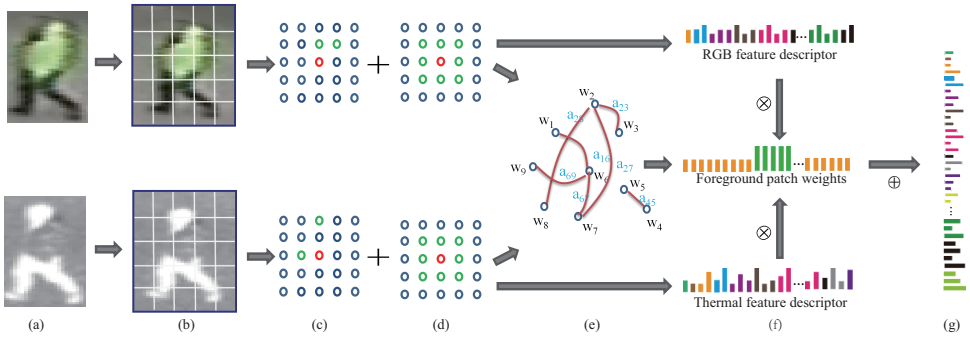

Figure 12: A general framework of graph-based RGB-infrared tracking method. (a) Input images. (b) Partitioned patches. (c) and (d) represents the illustration of local and global relationship, respectively, where the red circle is taken as an example, and the green circles have relations with the red one. (e) Illustration of graph construction, where 9 graph nodes are shown for clarity. w_i denotes the weight of the i-th node, and a_{ij} represents the edge weight of the i-th node and the j-th node. (f) Illustration of RGB-infrared features and foreground patch weights. (g) Final feature representation [29].

In graph-based fusion trackers, both the visible and infrared images should be represented using graph, and specific relationship between the visible graph and infrared graph has to be built. This relationship and the weights are key to graph-based methods. The main purpose of graph-based methods is to suppress background effects in RGB-infrared fusion tracking thus obtaining better feature representation.

Several studies have demonstrated that graph-based methods are effective in RGB-infrared fusion tracking. Li et al. [27] proposed a weighted sparse representation regularized graph to learn a robust object representation using visible and infrared images. To be best of our knowledge, this is the first graph-based RGB-infrared tracking method. In that paper, the graph affinity was optimized based on weighted sparse representation. Then, each patch weight was propagated from others along with graph affinity. In that work, the RGB feature and thermal feature were cascaded by considering the foreground patch weights. Li et al. [28] proposed a cross-modal ranking method with soft consistency for robust RGB-infrared tracking based on graph, as shown in Fig. 13. In this work, the patch weight was computed using the proposed cross-modal ranking algorithm in order to suppress background information. Also, the RGB and thermal feature were cascaded by considering the foreground patch weights. However, that method has two shortcomings. First, it does not consider modality weights in computing soft cross-modality consistency. This means that the method cannot distinguish the more reliable modality during tracking and thus would be affected by the imaging limitation of individual sources. Second, its speed is 8 FPS which does not meet the real-time requirement. Besides, Li et al. [29] proposed an approach to learn a local-global multi-graph descriptor, which automatically explores the intrinsic relationship among multi-spectral patches with both global and local cues for fusion tracking, as shown in Fig. 12. Unlike above-mentioned methods, in this method, low-rank

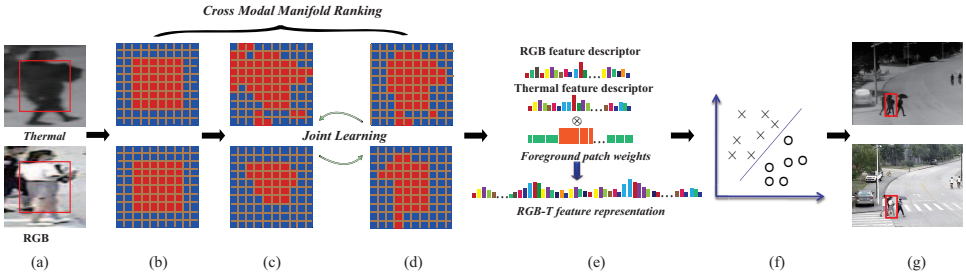

Figure 13: Pipeline of Cross-Modal Ranking for Robust RGB-infrared Tracking [28]. (a) Cropped regions, where the red bounding box indicates the region of initial patches. (b) Patch initilization indicated by red color. (c) Optimized results from initial patches. (d) Ranking results with the soft cross-modality consistency. (e) RGB-T feature representation. (f) Structured SVM. (g) Tracking results.

constraint was imposed on the joint representation coefficient matrix formed from multi-modal images for collaboratively incorporating multi-modal information. In addition, local neighboring information was imposed into the representation coefficients to encode the locally spatial cues. In addition, Li et al. [72] also proposed a two-stage modality-graphs regularized manifold ranking for RGB-infrared tracking. This work was similar to above-mentioned graph-based methods, but with two main differences. The first one was that in this method, a fixed-structure graph for each modality was constructed, therefore fewer variables need to be optimized which leads to faster convergence speed. The second was that a two stage ranking algorithm for computing the patch weights using multiple modalities was utilized.

Modality reliability has also been introduced to graph-based methods. Li et al. [71] proposed a graph-based fusion tracker which is similar to [28]. The difference is that, modality weights were incorporated for adaptive fusion of multiple modalities in [71].

All of these five works utilize the structured SVM [104] to carry out tracking. Similar graph-based idea has also been applied to the tracking based on visible images [105, 106].

3.3.4. Deep learning-based methods

Deep learning is well known for being able to learn effective feature representations from a large number of images. Compared to hand-crafted features, the learned deep features are more effective and robust, thus are beneficial for the tracking problem. Therefore, deep learning has been applied to RGB-infrared fusion tracking recently. It was in 2018 that deep learning was applied to RGB-infrared fusion tracking. Most deep learning-based fusion tracking algorithms are using feature-level fusion, while a small part is based on pixel-level and decision-level fusion.

In 2018, Xu et al. proposed a fusion tracking algorithm using visible and infrared images based on CNN [75]. To the best of our knowledge, it was the first work that performs RGB-infrared fusion tracking based on

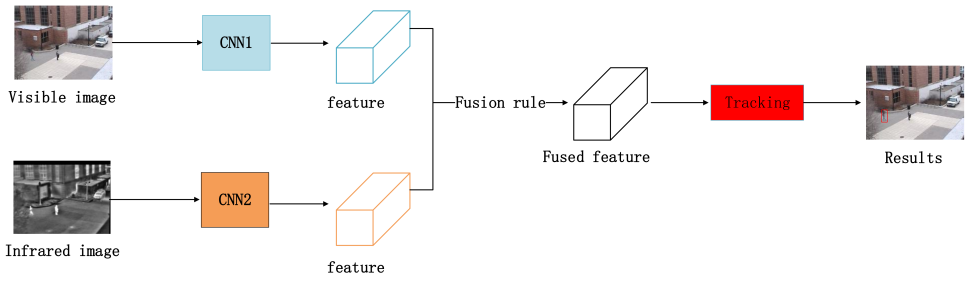

Figure 14: General framework of feature-level deep learning-based RGB-infrared fusion tracking algorithms.

CNN. This is a pixel-level fusion tracking method and uses a very simple image fusion method, i.e. simply utilizes the infrared image as the fourth channel of visible image. The method was based on the model proposed by Zhang et al. [107] and utilized a two-layer simple CNN which does not require a large amount of data for training. Experiments showed that the tracking performance of this method was better than traditional particle filter-based fusion trackers. However, the performance of this method was still limited, as visual artifacts were easily introduced which affect the extraction of effective features. In addition, the method was computationally intensive (about 5 FPS) and cannot meet real-time requirement. Xu et al. also proposed another pixel-level deep learning-based fusion tracking algorithms. In particular, they designed a fusion tracking algorithm based on the deep multi-view compression model [108]. Combining with the basic model of regional proposal network, this algorithm proposed an extended region proposal network, which can automatically change the position and scale of the object tracking window, and thus can effectively solve the problem of tracking fast moving target. Another pixel-level deep learning-based fusion tracking is proposed in [36], where Siamese networks are employed. Specifically, they firstly fused visible and infrared images into fused images, and then utilized these fused images as the input of the Siamese networks to perform tracking.

A general framework of feature-level deep learning-based fusion tracking algorithm is shown in Fig. 14. The features of both visible and infrared images are extracted by deep neural networks. These two features are then fused to form the joint feature using some rules. Then the tracking is performed using the joint feature. The key in deep learning-based fusion tracking lies in the joint feature representation and modality reliability.

Several methods have been reported which perform fusion tracking based on joint feature obtained from CNN. Zhang et al. [76] proposed a fusion tracking method based on the idea of MDNet [9]. The principle of this method is shown in Fig. 15. Basically, this method utilized a parallel structure, namely two shallow

CNNs were employed to handle visible and thermal infrared images, respectively. Then the visible and infrared features were concatenated directly and sent to domain-specific layers for binary classification and identification of the target. However, modality weights were not considered when combining visible and infrared features. Li et al. [26] proposed a two-stream fusion network to fuse the most efficient features generated by two sets of convolutional networks [26], as shown in Fig. 16. Basically, a CNN architecture consists of a two-stream CNN and a FusionNet were proposed. One CNN in the two-stream network was utilized to extract features from thermal images, and the other one was for dealing with visible images. The FusionNet was designed for adaptive fusion of two modalities and for removing redundant noise. The FusionNet would be updated during online tracking. However, the speed of this method was only around 15 FPS, which did not meet the real-time requirement. Besides, although the FusionNet tried to perform adaptive fusion of visible and infrared features by considering the significance of each feature, it did not take the reliability of visible and infrared images into account. Lan et al. presented a new tracking system which aimed to combine the information from RGB and infrared modalities for object tracking [22]. In that paper, they proposed a machine learning model, which can alleviate the modality discrepancy issue under the proposed modality consistency constraint from both representation patterns and discriminability, and generated discriminative feature templates for collaborative representations and discrimination in heterogeneous modalities.

There are some other works that consider the modality reliability during fusion tracking. Li et al. [77] proposed FANet, a quality-aware feature aggregation network for fusion tracking, as illustrated in Fig. 17. The key novelty of FANet was that it not only takes into the differences between multi-modal images into account, but also considered the differences of hierarchical features. During tracking, both layer weights and modality weights were learned. However, the speed of FANet was only only 1.3 FPS, which was far from the real-time requirement. Zhang et al. [109] presented SiamFT, which was based on SiamFC [5] and leveraged two Siamese networks to extract features of visible and infrared respectively. Visible and infrared features were fused to form the fused template feature and fused search feature. The similarity between these two features were then computed through cross-relation, which results in a final response map. Based on the final response map, the position and scale of the target can be derived. A modality weight computation method was proposed there based on the response map of each modality. The method can run at more than 28 FPS. Lan et al. [23] proposed a discriminative learning framework to perform RGB-infrared fusion tracking. The method can adaptively learn collaborative classifiers of each modality for target/background separation and reliability weights for each modality. Lan et al. also proposed [38] an online non-negative feature template learning model to perform RGB-assisted infrared tracking. In that work, an adaptive modality

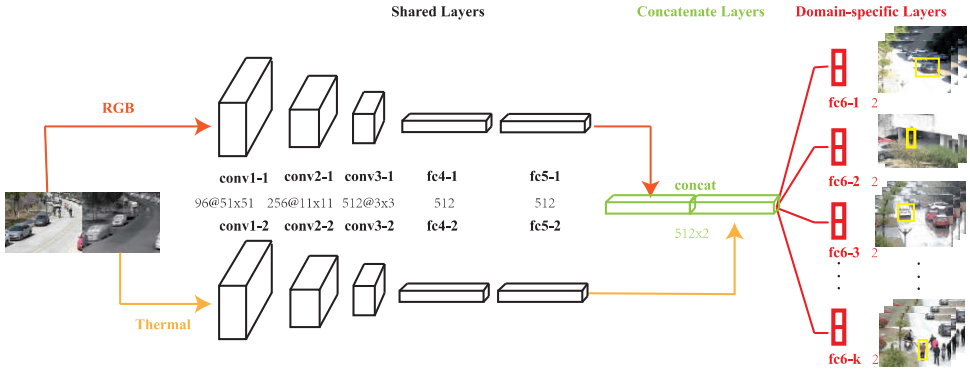

Figure 15: The architecture of the method proposed by Zhang et al. [76].

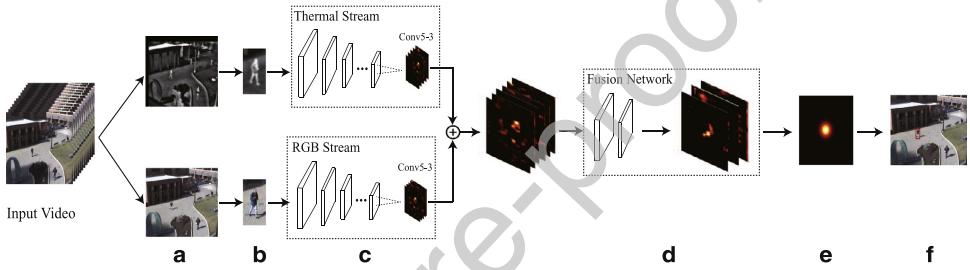

Figure 16: Pipeline of the method proposed by Li et al. [26]. (a) Input frames. (b) Target patch. (c) Generic network. (d) Feature map selection (e) Correlation filter. (f) Tracking results.

importance weight learning scheme was proposed and an iterative optimization algorithm was derived.

Regarding decision-level deep learning-based fusion tracking algorithms, Tang et al. [81] tracked the target in both visible and infrared images. Confidence values were generated for both visible and infrared images during tracking. The result with a larger confidence value was chosen as the fused result. Zhang et al. [37] utilized two Siamese networks to track the target in visible and infrared sequences, respectively. They used the modality weight computation method proposed in [109] to compute the reliability of results on two modalities. The result with a higher reliability degree was considered as the final result.

Apart from CNN, other deep learning techniques can also be applied in RGB-fusion tracking. Yun et al. proposed a fusion tracking algorithm for visible and infrared object based on multi-view multi-core fusion model [110]. The authors claimed that the fusion tracking based on visible and infrared tracking obey two major principles of multi-view learning, namely consensus and complementary, thus they utilized multi-view learning to perform fusion tracking. Besides, the multi-kernel framework was used to learn the importance of viewing features. The tracking was completed with naive Bayes classifier in sophisticated compressive feature domain.

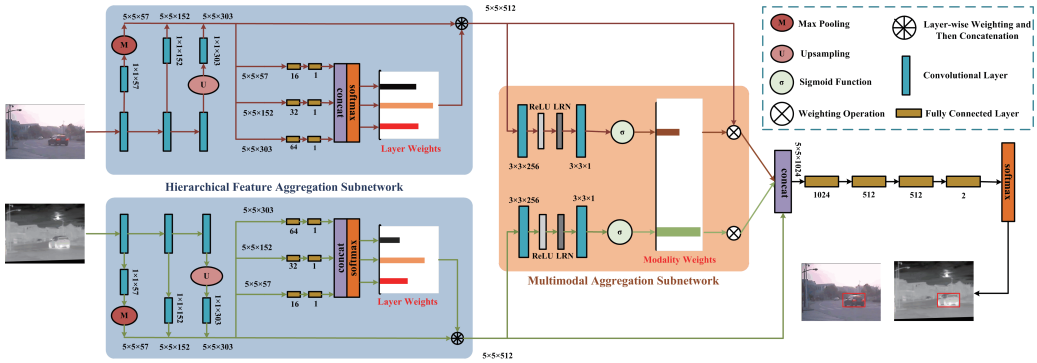

Figure 17: The flowchart of FANet [77].

To sum up, deep learning-based methods have shown very competitive performance in RGB-infrared fusion tracking. It can be expected that in the future more deep learning-based approaches will be reported. However, a major problem in deep learning-based method is the computational cost. Measures have to be taken to reduce the computation cost thus making the deep learning-based fusion trackers faster.

3.3.5. Correlation filters-based methods

Researchers have also begun to explore fusion tracking based on correlation filter techniques recently due to their good performance and high efficiency. By far all the CF-based RGB-infrared fusion tracking algorithms are based on decision-level fusion. In particular, the response maps obtained from visible images and infrared images are fused to obtain the final response map:

$$\mathbf{R} = \sum_{k=1}^{K} \alpha_k \mathbf{R}_k, \tag{1}$$

where $\mathbf{R}$ is the final response map, $\mathbf{R}_k$ is the response map of the k-th CF-based tracker, α_k is the weighted coefficient. The tracking results can be obtained based on the final response map. It should be noted that in reported trackers, a part of them utilize modality reliability as the weighted coefficient, while others do not.

To the best of our knowledge, Wang et al. [74] presented the first CF-based fusion tracking work. They proposed a fusion tracking method based on soft consistency correlation filter (SCCF), through which they took both collaboration and heterogeneity into account for visible and infrared modalities. Soft consistency means that on one hand, the learning filters for both modalities should have similar circular shifts, while on the other hand the filters can have sparse different elements to each other. They proposed a weighted fusion mechanism to fuse response map of both modalities to produce the final response map, based on which the

tracking result was obtained, as shown in Fig. 18. The weights were obtained according to the response map in the detection phase. Specifically, the average peak-to-correlation-energy (APCE) [111] was employed to compute weights. The speed of SCCF was 50 FPS, which meets the real-time requirement. Zhai et al. [30] proposed an RGB-infrared tracking algorithm via cross-modal correlation filters, as illustrated in Fig. 19. In that work, correlation filter was employed for each modality and then a low-rank constraint was utilized to learn filters jointly for cross-modal fusion. They observed that different modality features should have similar correlation filters to have consistent localization of the target object. The final response map was also obtained by fusing the response maps of each modality. However, modality weights were not considered in that work. The speed of this tracker was 224 FPS, which was the fastest RGB-infrared tracker by far to the best of our knowledge. Recently, Yun et al. [80] proposed a discriminative fusion correlation learning model (DFCL) to improve DCF-based tracking performance, which also applied a weighted fusion rule for the response map of visible and infrared images. However, the authors only gave the results on several sequences selected from different datasets, but did not show performance on large-scale datasets. It is hardly to evaluate its actual performance on large scale benchmarks as the source codes are also not available.

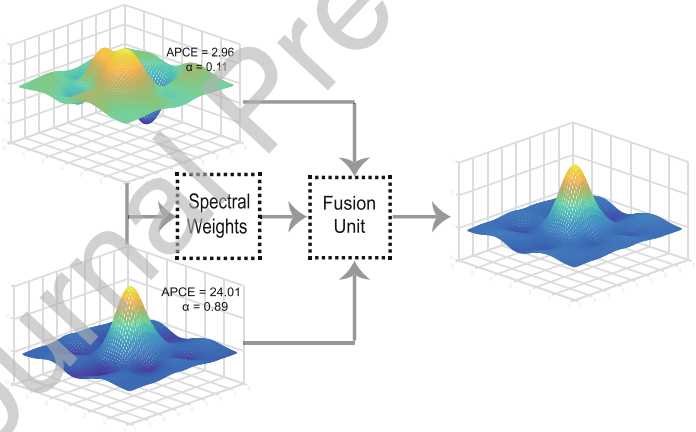

Figure 18: The proposed spectral fusion mechanism [74]. The spectral weights are obtained from the response map.

Apart from above works which purely utilize correlation filter to perform fusion tracking, there are also some studies which combine correlation filter with other methods to perform multi-modal tracking. Luo et al. [79] proposed a tracking-before-fusion framework which comprises two modules, namely a correlation filter-based tracking (CFT) module and histogram based tracking (HIST) module. The main idea of this work is illustrated in Fig. 20. In particular, the CFT module and HIST module implemented tracking individually, and then the final results were obtained through a decision-level fusion via an adaptive weighting

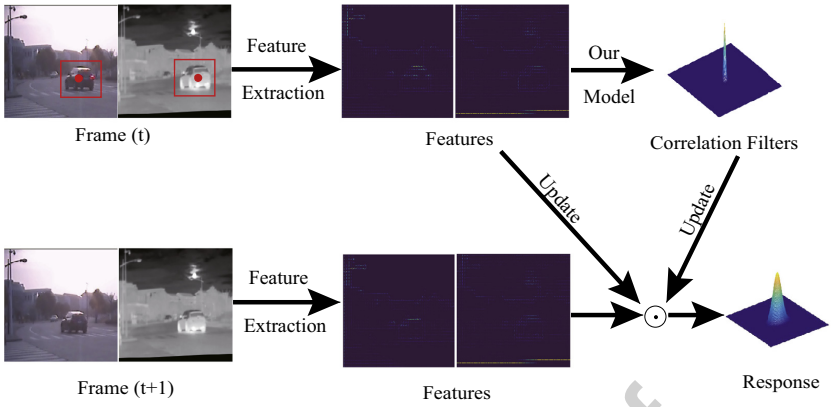

Figure 19: The fast RGB-infrared tracking via cross-modal correlation filter [30].

scheme. Specifically, temporal information was considered when determining the weights of two modules by employing Kullback-Leibler Divergence to determine the similarity between the current response map and response map of the last frame. Ren et al. [35] proposed a tracking framework which combines the CF-based visible tracker and Markov chain Monte Carlo (MCMC)-based thermal tracker (VIRF). The aim was to perform robust night target tracking. Different from above-mentioned CF-based trackers, VIRF handled asynchronous visible and infrared images. They employed different trackers for them because that thermal images and visible images have different imaging characteristics. A candidate region location-scale fusion rule was designed to obtain the final tracking result.

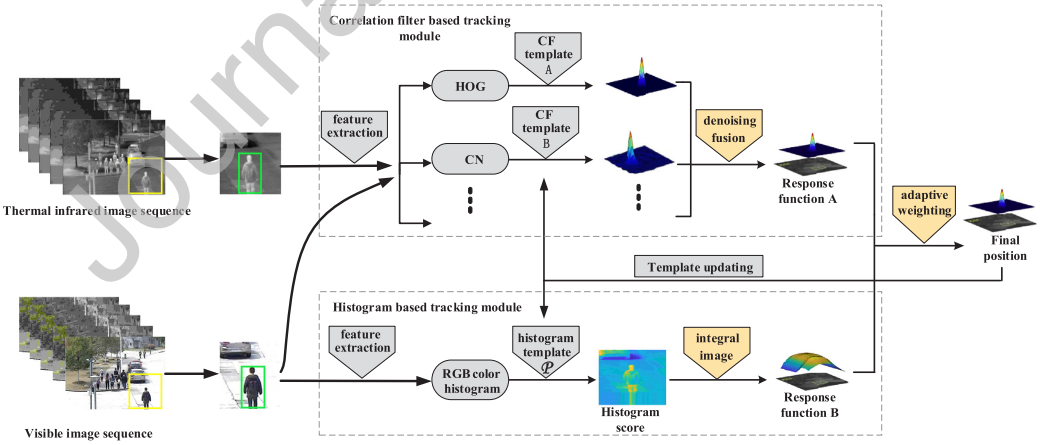

Figure 20: The hybrid framework consists of two modules [79].

To sum up, although the research of CF-based RGB-infrared fusion tracking began in 2018 and is still at

its very early stage, their highly competitive performances and efficiency make them a promising research direction in future.

3.3.6. Multi-object fusion tracking

In practical applications, it is frequent to track multiple objects simultaneously. For instance, in video surveillance, security staffs need to track all people in the scene to find the suspect ones. In addition, multi-object tracking (MOT) can be applied in other applications such as human-computer interface and virtual reality [112]. However, to the best of our knowledge, almost all published RGB-infrared fusion tracking algorithms, including those mentioned previously, were designed for tracking a single object. Only very few works discussed multi-object fusion tracking (MOFT). For example, Bunyak et al. [88] proposed a moving object detection and tracking system using infrared and visible images, which can track multiple objects based on the Kalman filter. Mangale et al. [100] proposed a method based on background subtraction to detect all moving objects in the scene. Similarly, Qiu et al. [78] proposed a method to extract moving targets based on the fusion of visible and infrared images. Different from [88, 100], this algorithm utilized a spatio-temporal local binary pattern (LBP) algorithm.

As can be seen, these MOFT methods utilize traditional techniques such as background subtraction to detect and tracker moving objects. Therefore, there is still much room for improvement in tracking performance by utilizing advanced techniques such as deep learning.

3.4. Summary

Although RGB-infrared fusion tracking has evolved for more than ten years, the development was relatively slow before new techniques are applied in this field in recent years, such as sparse representation, deep learning and correlation filter. These new techniques have significantly boosted the performance of RGB-infrared fusion tracking. Regarding fusion level, by far almost all sparse presentation-based and graph-based methods belong to feature-level fusion, while all CF-based methods utilize decision-level fusion. Furthermore, deep learning has been applied to all three levels of fusion tracking.

4. RGB-infrared tracking dataset

4.1. Available RGB-infrared dataset

Large-scale datasets are of vital importance in RGB-infrared fusion tracking, since they are not only beneficial for training algorithms, but are also crucial for testing algorithms and comparing performance

among trackers. Before large-scale datasets are available, in most RGB-infrared fusion tracking publications, the experimental part only employs several visible and infrared video pairs to verify the algorithm, or even one single video pair. For example, the OTCBVS dataset [113] containing 6 pairs of visible and infrared videos is a frequently utilized test set. However, this very few sets of videos are unable to cover enough challenging situations, such as illumination changes, similar object interference, object deformation, and occlusion. Therefore, the performance of those algorithms under various conditions cannot be fully evaluated, making it extremely difficult, if not impossible, to evaluate performance of various trackers reasonably. In addition, those videos do not have uniform annotations. Researchers need to make their own annotations, resulting in different accuracy of the data and also different initialization state of the algorithm. This makes it difficult to objectively evaluate algorithms under a unified standard. Therefore, in some of the aforementioned RGB-infrared fusion tracking studies, especially those published before 2016, the algorithms cannot be compared uniformly. In this section, several publicly available fusion tracking datasets are summarized.

A good dataset should have following attributes:

- Contain a large number of aligned visible and infrared video frames, which can be used to test tracking performance comprehensively.
- Videos should have annotations on tracking targets, i.e. the bounding box which shows the position and size of target.
- Videos should cover a wide range of working conditions, such as low illumination, fast motion and occlusion. In this case, the dataset can be used to evaluate performance of trackers under different conditions. This is also helpful for developing tracking algorithms which work well in specific conditions.

To the best of our knowledge, currently there are in total five fusion tracking datasets available, namely OTCBVS dataset [17, 113], LITIV [114], GTOT [25], RGBT210 [27], RGBT234 [71].

In 2005, Davis et al. [113] captured six thermal/color video sequence pairs from two different locations at the Ohio State University campus. These six videos contain around 17,060 frames and three of them have illumination changes across the scene. However, the bounding boxes are not available and no attributes are annotated for these videos. In 2012, Torabi et al. [114] captured nine visible and thermal video pairs with different zoom settings and at different locations. However, there are also no bounding boxes available in that dataset.

The above-mentioned datasets are not created for tracking originally, and the videos are not enough to cover various challenging scenarios and different kinds of targets. Therefore, by using these videos the

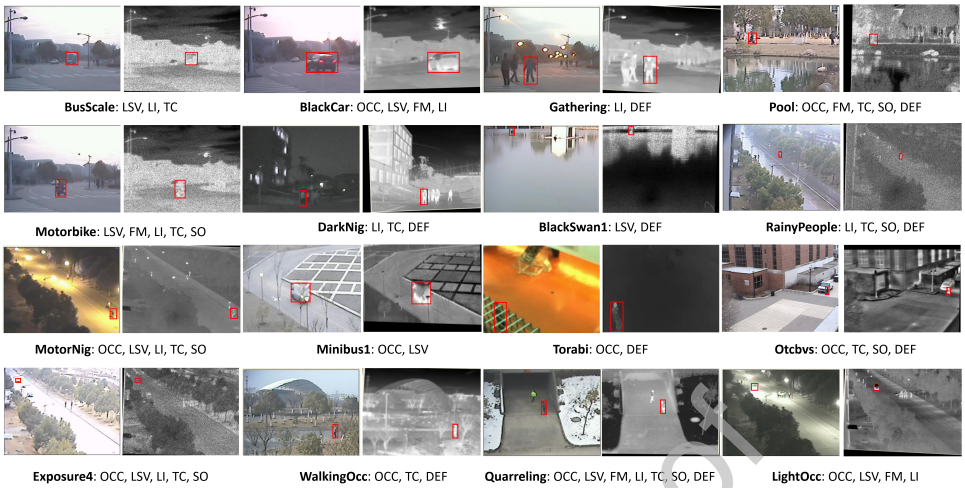

Figure 21: Video examples and corresponding attribute annotations in the GTOT dataset [25]. In each image pair, the left one is visible image while the right one is infrared image. Red box is the bounding box of the target in the first frame. The words under each image pair give the name and annotated attributes of that image sequence.

performance of fusion tracking algorithms may not be evaluated comprehensively. As a consequence, the development of RGB-infrared fusion tracking is hindered to some extent.

It was in 2016 that large scale RGB-infrared datasets became available. Li et al. produced a grayscale-thermal object tracking dataset (GTOT) for gray image and infrared image [25]. GTOT contains fifty pairs of registered grayscale and infrared videos captured under different scenarios and conditions. It also contains manual annotated data, including bounding boxes around target and attributes indicating challenging conditions as listed in Table 2. To the best of our knowledge, this dataset is the first large-scale dataset in the field of RGB-infrared fusion tracking. Examples of videos in GTOT are given in Fig. 21. In 2017, Li et al. produced a larger dataset RGBT210 [27] for RGB-infrared fusion tracking, containing 210 pairs of visible and infrared videos. This dataset contains in total around 210K frames which is sufficiently large for performance evaluation. Besides, more attributes (12 in total) are annotated for evaluating challenge-based performances, as listed in Table 3. In 2018, the dataset RGBT210 was expanded to RGBT234 by adding another 24 video pairs [71].

The comparison of above-mentioned give datasets is listed in Table 4. The emergence of GTOT, RGBT210 and RGBT234 has played an important role in promoting research of RGB-infrared fusion tracking. However, these datasets also have some drawbacks. First of all, their videos are collected mainly by the same device (RGBT210 and RGBT234), so the resolution and imaging characteristics of the image are the same. This makes it inconvenient to correctly evaluate the various resolutions that fusion tracking algorithms may

Table 2: Attributes annotated in GTOT [25].

Attribute	Description	Attribute	Description
OCC	Partial or full occlusion	TC	Thermal crossover
LSV	Large scale variation	SO	Small object
FM	Fast motion	DEF	Deformation
LI	Low illumination		

Table 3: Attribute information of RGBT210 and RGBT234 dataset [27, 71]

Attribute	Description	Attribute	Description
NO	No Occlusion	DEF	Deformation
PO	Partial Occlusion	FM	Fast Motion
HO	Heavy Occlusion	SV	Scale Variation
LI	Low Illumination	MB	Motion Blur
LR	Low Resolution	CM	Camera Moving
TC	Thermal Crossover	BC	Background Clutter

encounter in practical applications. Second, the video attributes annotation are not comprehensive. For example, the attributes do not include 'illumination change', which is one of the original objectives of investigating RGB-infrared fusion tracking algorithm.

Table 4: Comparison of RGB-infrared fusion tracking datasets.

Name	Videos	Frames (In total)	Attributes	Groundtruth	Video type	Resolution	Year
OTCBVS	6	17,060	No	No	RGB, T	320× 240	2005
LITIV	9	4,300	No	No	RGB, T	320× 240	2012
GTOT	50	15,800	7	Yes	Gray, T	Various	2016
RGBT210	210	210,000	12	Yes	RGB, T	630×460	2017
RGBT234	234	234,000	12	Yes	RGB, T	630×460	2018

Note that in addition to the aforementioned datasets, there are some other datasets also containing aligned visible and infrared videos, such as KAIST [115]. However, although it contains around 95,000 aligned visible and infrared video frames, it is designed for pedestrian detection thus is not suitable for generic single object tracking. A similar case is the dataset captured by Bilodeau et al. [116].

Another thing should be noted is that, these datasets are mainly used for testing trackers and comparing performances. Because the number of video frames are not large enough, it is difficult to train algorithms using them from the beginning.

4.2. VOT challenge

Due to the increasing popularity of RGB-infrared fusion tracking, in 2019, the well-known visual object tracking (VOT) challenge introduced a new subchallenge, namely the VOT-RGBT challenge. The VOT-RGBT challenge provides a dataset consisting of 60 visible and infrared video pairs. However, after careful checking, we find that these 60 video pairs are selected from RGBT234 [71]. Therefore, this dataset does not add new data. One of the differences is that the VOT-RGBT challenge utilizes thermal images as the primary modality and thus the evaluation metrics are computed based on the groundtruth of thermal images. Whereas in RGBT234, the groundtruth of both modalities are utilized (see Section 5).

We believe that the VOT-RGBT challenge will greatly promote researchers to further investigate RGB-infrared fusion tracking. As a consequence, we expect that the number of RGB-infrared fusion tracking algorithms will increase quickly in the coming years.

5. Evaluation metrics

In recent years, several well-recognized evaluation metrics have been proposed to evaluate tracking performance based on visible images. These include precision rate (PR), success rate (SR), accuracy, robustness and Expected Average Overlap (EAO). These evaluation metrics can also be applied in RGB-infrared fusion tracking.

5.1. Precision rate

Precision means that the center location error (CLE) between the predicted bounding box and the ground truth is smaller than a chosen threshold. The precision rate means the percentage of frames whose CLE is within the threshold. Success plot shows the trends when threshold changes from small to large. The threshold is set to 20 pixels usually.

However, precision has a few shortcomings. First, it does not consider the scale change of target. For examples, an error of 5 pixels is not much compared to a large object, but it is severe for a small object which just occupies several pixels in size. Second, the error will continue to increase even when the object has lost.

Li et al. [71] proposed using maximum precision rate (MPR) to evaluate fusion tracking performance. Specifically, for every frame, the Euclidean distance between predicted bounding box and ground truth on both visible and infrared modalities is computed, and smaller distance is chosen then to compute precision score. It is certain that MPR value will be larger than PR for the same case.

5.2. Success rate

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

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

where a and b indicates the predicted bounding box and ground truth, 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. The success plot is a better evaluation metric compared to precision, therefore in many papers only the success plot is shown. In addition, researchers usually rank different trackers according to their success scores.

Li et al. [71] proposed using maximum success rate (MSR) to evaluate fusion tracking performance. Specifically, for every frame, the overlapping between predicted bounding box and ground truth on both visible and infrared modalities is computed, and larger overlapping value is chosen to calculate success score then. It is certain that MSR value will be larger than PR for the same case.

5.3. TRE and SRE

The temporal robustness evaluation (TRE) is designed for evaluating temporal robustness of tracking algorithms. The idea is that, each tracker is evaluated many times starting from different frames across a video, with the initialization of the corresponding ground truth object state. The tracking results of all the tests are averaged to generate the TRE score. In OTB [117], the TRE is evaluated for 20 times.

The spatial robustness evaluation (SRE) is designed to evaluate whether a tracking method is sensitive to initialization state by running several tests. The idea is to initialize the tracker by slightly shifting or scaling the ground truth bounding box of a target object. The SRE score is the average of 12 evaluations in OTB [117].

5.4. Accuracy

Similar to success rate, accuracy measures how well the predicted bounding box given by tracker overlaps with the ground truth bounding box. It calculates the average overlaps between the predicted and ground truth bounding boxes during successful tracking periods.

5.5. Robustness

When the overlap between predicted and ground truth bounding boxes is zero, the tracking is considered as a failure. Robustness measures the times that a tracker loses the target during tracking.

5.6. Expected Average Overlap

Expected Average Overlap (EAO) is firstly utilized in the well-known VOT challenge to evaluate tracking performance, and now it is the most important metric in VOT challenge [118]. EAO combines the raw values of per-frame accuracies and failures in a principled manner. It measures the expected no-reset overlap of a tracker on a short-term sequence. In other words, it estimates how accurate the estimated bounding box is after a certain number of frames are processed since initialization.

There are still other evaluation metrics, but PR and SR are the most frequently used in literature by far.

6. Benchmark results and analysis

In this section, we present results on available public fusion tracking datasets. The results are either collected from the published literature or produced by the authors. The aim is to facilitate the research of this direction, and make it easier for researchers to compare tracking results with the state-of-the-arts. It should be mentioned that many RGB-infrared trackers are not open-source and their results on public dataset have not been reported [22–24, 38, 79, 80]. As a consequence, it is not feasible to summarize their results here.

6.1. Results on public datasets

6.1.1. Results on GTOT dataset

As the first large-scale fusion tracking dataset, GTOT has been chosen in several studies to evaluate performance of fusion tracking algorithms. Table 5 presents the reported results of 13 trackers on the GTOT dataset. Note that some data are not available so far, thus we denote it with a hyphens.

As can be seen from Table 5, deep learning-based methods achieve the best results in terms of both PR and SR. In particular, the top-3 methods in both PR and SR are all based on deep learning and they outperform other methods with a clear margin. This clearly demonstrates the effectiveness of deep learning in boosting RGB-infrared fusion tracking performance. This is due to the strong feature representation

capability of deep neural networks. However, none of these deep learning-based approaches can run at real-time speed. The framerate of most deep learning-based fusion trackers is around 1 FPS or ever less, which is too slow for practical applications.

Apart from deep learning-based methods, CF-based approaches obtain competitive performance. For example, SCCF ranks fifth in PR and fourth in SR among all trackers. Furthermore, SCCF can run at real-time speed, which is much faster than its deep learning-based counterparts. It is obvious that CF-based methods strike a better balance between tracking precision and speed compared to the deep learning-based methods.

The results of graph-based methods are slightly worse than SCCF but are better than that of Zhai et al. [30]. However, although graph-based methods achieve competitive tracking precision and success rate against CF-based ones, their speed is much slower, namely with the framerate less than 8 FPS. Sparse representation-based approaches achieve comparable success rate and worse precision rate compared to graph-based methods and CF-based methods. However, their speed is even slower and is around 1.5 FPS.

Based on these results on the GTOT dataset, we think that deep learning-based methods and CF-based methods will be two most important research directions in RGB-infrared fusion tracking. Specifically, the speed of deep learning-based methods should be further enhanced while the precision of CF-based trackers should be improved.

Table 5: Precision rate (PR %), Success rate (SR %) and running speed (FPS) on the GTOT dataset. The best three results are shown in red, green and blue, respectively. Best viewed in color.

	MANet [84]	DAPNet [82]	Yang[83]	Li [26]	FANet [77]	LGMG [29]	Li [72]	Li [28]	SGT [27]	Zhai [30]	SCCF [74]	CSR [25]	Li [70]
PR	89.4	88.2	84.3	85.2	88.5	82.9	84.2	82.7	85.1	77	85	75	-
SR	72.4	70.7	67.7	62.6	69.8	65.5	62.2	64.3	62.8	63.2	68.1	62	66.5
Type	DL	DL	DL	DL	DL	Graph	Graph	Graph	Graph	CF	CF	SR	SR
Speed	1.11	-	0.349	15	1.3	7	7	8	5	227	50	1.6	1.5

6.1.2. Results on RGBT234 dataset

Since the corresponding paper on RGBT234 is formally published in 2019, the available results on it so far are relatively rare compared to that of GTOT. Table 6 lists the reported results of 8 trackers on RGBT234 dataset, showing both the overall performance and attribute-based performance. Some of these results are collected from literatures which present new fusion tracking methods, such as MANet [84], DAPNet [82], Yang[83], FANet [77] and Li [71]. Others are provided by other researchers, like SGT [27], CSR [25] and JSR [32]. Note that some data are not available so far, thus we denote it with a hyphens. Besides, it should

Table 6: Maximum precision rate, maximum success rate (MPR/MSR %) and speed (FPS) on the RGBT234 dataset. The best three results are shown in red, green and blue, respectively. Best viewed in color.

	MANet [84]	DAPNet [82]	Yang[83]	FANet [77]	SGT [27]	Li [71]	CSR [25]	JSR [32]
NO	-/64.6	90.0/64.4	-/-	84.7/61.1	87.7/55.5	87.4/55.0	56.7/41.5	56.7/41.5
PO	-/56.6	82.1/57.4	-/-	78.3/54.7	77.9/51.3	77.9/51.0	49.4/34.9	37.8/25.2
HO	-/46.5	66.0/45.7	-/-	70.8/48.1	59.2/39.4	58.8/39.2	38.4/26.8	25.9/18.1
LI	-/51.3	77.5/53.0	-/-	72.7/48.8	70.5/46.2	71.9/46.8	39.3/27.3	38.1/26.4
LR	-/51.5	75.0/51.0	-/-	74.5/50.8	75.1/47.6	75.4/47.7	41.3/25.9	39.9/23.9
TC	-/54.3	76.8/54.3	-/-	79.6/56.2	76.0/47.0	70.4/44.3	44.4/32.5	33.8/20.9
DEF	-/52.4	71.7/51.8	-/-	70.4/50.3	68.5/47.4	67.5/47.0	44.8/33.1	27.8/20.2
FM	-/44.9	67.0/44.3	-/-	63.3/41.7	67.7/40.2	62.5/36.9	34.9/22.0	25.6/15.7
SV	-/54.2	78.0/54.2	-/-	77.0/53.5	69.2/43.4	69.5/43.5	50.9/37.3	35.5/23.8
MB	-/51.6	65.3/46.7	-/-	67.4/48.0	64.7/43.6	64.2/43.5	37.9/27.0	24.2/17.2
CM	-/50.8	66.8/47.4	-/-	66.8/47.4	66.7/45.2	66.2/44.6	41.5/30.1	29.1/21.0
BC	-/48.6	71.7/48.4	-/-	71.0/47.8	65.8/41.8	66.4/42.2	38.8/25.3	33.2/21.2
Overall	77.7/53.9	76.6/53.7	78.7/54.5	76.4/53.2	72.0/47.2	71.8/46.9	46.3/32.8	34.3/23.4
Type	DL	DL	DL	DL	Graph	Graph	SR	SR
Speed	1.11	-	0.349	1.3	5	5	1.6	-

be mentioned that the evaluation metrics on RGBT234 are different from that of GTOT. Specifically, the maximum precision rate (MPR) and minimum success rate (MSR) are utilized by considering the groundtruth of both modalities.

Table 6 clearly shows that the deep learning-based methods achieve the leading performance on RGBT234 by outperforming graph-based and sparse representation-based approaches with a very clear margin. In particular, the top 4 results are all produced by deep learning-based methods. This indicates that deep learning techniques can improve RGB-infrared fusion tracking performance significantly. This is due to the superior feature representation ability of deep learning and the huge number of images used to train these models. Besides, although the performance of graph-based methods are worse than deep leaning-based ones, they beat sparse representation-based approaches with a huge gap. This indicates that the recently emerged graph-based approaches is a promising research direction as well. This is further demonstrated by the attribute-based performance. Specifically, the graph-based methods shows leading performance in some challenges. For instance, SGT achieves the best precision rate in fast motion and the tracker of Li et al. [71] produces the best precision rate in low resolution.

Regarding running speed, although graph-based approaches are relatively faster, all these methods do not meet the real time requirement, namely with a framerate less than 5 FPS.

By far we have not found the results of CF-based RGB-infrared fusion trackers on RGBT234, thus it is still not clear how the CF-based fusion trackers perform on this data. Although the reported results on RGBT234 dataset are relatively rare currently, we believe that more trackers will be run on this dataset in the future. At that time, a more thorough comparison can be performed.

6.2. Speed Comparison of RGB-infrared trackers

As mentioned in Section 3.1, running speed is a key factor in RGB-infrared trackers. Table 7 lists reported speed of some fusion tracking approaches. As can be seen, the running time of different fusion tracking algorithms varies significantly. Generally speaking, the CF-based methods are most efficient and can easily achieve real time speed even using CPU. This is due to the efficient implementation of correlation filter via FFT. The deep learning-based methods are much slower even running with GPU. The speed varies from 0.349 FPS to around 30 FPS, depending on implementations. By far the only deep learning-based approach which can meet the real-time requirement is the SiamFT proposed by Zhang et al. [109]. Besides, the graph and sparse representation-based methods are slow and are not able to run at real time speed currently. This is because that these two kinds of methods need to update or optimize the model online, which is time consuming.

7. Future prospects

Despite the remarkable progress that has been achieved in RGB-infrared fusion tracking, several issues remain for future work. In this section, we give detailed discussions on specific trends of RGB-infrared fusion tracking based on the review of existing approaches.

7.1. Tracking algorithms

7.1.1. Deep learning-based algorithms

As discussed in earlier sections, some deep learning-based fusion tracking algorithms have been proposed, which have achieved the leading performance. However, the performance of deep learning-based fusion tracking algorithms can still be improved, in terms of both precision and speed. Furthermore, compared to video tracking based on visible images, the application of deep learning in fusion tracking is still at its very early stage. Therefore, a lot of techniques in tracker based on visible images can be applied to fusion tracking by considering cross-modal characteristics.

We believe that one of the most promising development directions of RGB-infrared fusion tracking is based on Siamese networks, which are renowned for their high performance and efficiency in visual

Table 7: Speed of some RGB-infrared fusion tracking algorithms.

Name/Reference	Framerate (FPS)	CPU or GPU	Category
Zhai [30]	227	CPU	CF-based
SCCF [74]	50	CPU	CF-based
Yun [89]	32	CPU	Traditional
Zhang [36]	28-32	GPU (GTX 1080Ti)	DL-based
Li [68]	17	CPU	SR-based
Luo [79]	15	CPU	CF-based
Li [26]	15	GPU (TITAN X)	DL-based
Li [69]	9.3	CPU	SR-based
Li [28]	8	CPU	Graph-based
LGMG [29]	7	CPU	Graph-based
Li [72]	7	CPU	Graph-based
Li [27]	5	CPU	Graph-based
Xu [75]	5	CPU	DL-based
Lan [24]	3	PRL	SR-based
Li [25]	1.6	CPU	SR-based
Li [70]	1.5	CPU	SR-based
FANet [77]	1.3	GPU (GTX 1080Ti)	DL-based
MANet [84]	1.11	GPU	DL-based
Lan [22]	0.7	CPU	ML-based
Yang[83]	0.349	GPU	DL-based

tracking. For example, the SiamFC [5] has been well-recognized for its performance and efficiency, and has been utilized as the baseline of many trackers. Zhang et al. [36]performed RGB-infrared fusion tracking at pixel-level [36], feature-level [109] and decision-level [37]. These works demonstrate the feasibility of Siamese networks in providing good tracking performance at real time speed.

Recently, some Siamese-based trackers achieved the state-of-the-art performance with real-time speed, for example DaSiamRPN [119], SiamRPN [10], SiamRPN++[11], SiamMask [120]. However, to the best of our knowledge, the principles of these trackers have not been applied to RGB-infrared fusion tracking. This is a research direction that worth exploring.

7.1.2. Correlation filters-based algorithms

Although some correlation filter-based fusion tracking algorithms have been proposed as summarized in Section 3.3.5, they are still at their early stage. We believe that correlation filter-based methods will be an important trend in the future. The major advantage is the efficiency. As pointed out earlier, correlation filter-based fusion tracking algorithms can easily meet the real-time requirement. Besides, a lot of advanced

correlation filters for visible images have been proposed which have achieved excellent performance, such as BACF [121], ASRCF [122], PRCF [123] and SACF [124]. The principle of these trackers may also beneficial for fusion tracking.

7.2. Implementations

7.2.1. Modification on the basis of existing RGB trackers

The number of RGB trackers is increasing quickly, and the performance is also continuously improved. A straightforward way to implement an RGB-infrared fusion tracking method is to make modifications based on existing RGB trackers. For example, Zhang et al. [36] performed pixel-level RGB-infrared fusion tracking based on the SiamFC proposed by Bertinetto et al. [5]. They employed the SiamFC directly without modification and replaced the input with fused images. Actually, as claimed in that paper, other RGB trackers can also be utilized to perform fusion tracking by using fused image as input.

7.2.2. New fusion tracking model

Although the fusion tracking methods can be implemented based on existing RGB trackers, the performances may be limited. One of the main reasons is that the RGB trackers may not well suit for infrared images, because the characteristics of infrared images are quite different from that of visual images. To solve this problem, one can retrain the model using infrared images. However, the available infrared images are not sufficient to train an infrared network from the beginning. Therefore, one can choose specific methods which can handle the limited number of training data, namely transfer learning or fine-tuning.

Another reason is that by using existing RGB trackers, the relationship between visual and infrared images are not fully leveraged, which may be useful if it is well utilized. To handle this, new algorithms need to be developed.

7.3. Larger RGB-infrared datasets

Based on the discussion of available RGB-infrared datasets in Section 4, it is necessary to develop a larger RGB-infrared dataset. To be more specific, this new dataset should have the following properties:

- Large enough. This dataset should provide enough visual and infrared image pairs. Ideally, it should be divided into training subset, validation subset and test subset. The training subset should contain a large number of images to train a RGB-infrared tracker from the scratch if learning-based methods are employed.

- More diverse. Diverse videos that cover more challenging situations should be contained. The creator of this dataset should consider as more as possible the challenging situations that RGB-infrared trackers may encounter in practical applications, such as occlusion, illumination change, poor light condition.
- The ground truth should be available, including the bounding boxes and attribute annotations.

7.4. Benchmarks

In the community of tracking based on visible images, several benchmarks have been proposed, such as OTB and VOT. These benchmarks make it convenient for researchers to compare results among different trackers. However, by far the source codes of most RGB-infrared fusion tracking algorithms are not available. Compared to tracking based on visible images, RGB-infrared fusion tracking currently has some features, which are summarized as follows:

- Most trackers are not open-source, thus are difficult for other researchers to obtain, run and modify the source code. This is very different from that of visual tracking based on RGB images, where most trackers are open-source, such as SiamFC [5], CFNet [125], DCFNet [126], KCF [6], ECO [14], SiamRPN++ [11], MDNet [9]. This severely hinders the development of fusion tracking field.
- There is still a lack of benchmark, by using which one can compare the performance with other RGB-infrared trackers. Li et al. [71] made an early effort to handle this by building the RGBT234 datasets. They also tried to provide results of some trackers. However, they are still not enough, because the results of most published RGB-infrared trackers on this datasets are missing.

As a consequence, it is quite difficult for researchers in the RGB-infrared field to compare with the state-of-the-arts. For example, in some recently published RGB-infrared tracking studies [22–24], different sequences were employed to conduct experiments. Therefore, it is almost impossible for other researchers to compare with them.

Based on these discussions, it is quite necessary to build a RGB-infrared fusion tracking benchmark which has following features:

- have enough visual and infrared videos which cover as more as possible challenging scenarios. This is crucial for the extensive and comprehensive evaluation of RGB-infrared trackers.
- Have some open-source RGB-infrared trackers and corresponding results, thus everyone in the community can compare their trackers with results in the benchmark.

- Have evaluation metrics. Aforementioned precision score, success score and EAO can be chosen.

The benchmark will make it convenient for the community to compare results. What is more, it will greatly promote the development of RGB-infrared fusion tracking field.

7.5. Ability to handle unregistered image pairs

Currently, most, if not all, RGB-infrared trackers require that the visible and infrared images are well registered. However, this is not easy to achieve in practice. For example, Figure 22 presents some examples of videos captured by the authors where the visible and infrared images are not strictly aligned. In addition, although claimed well aligned, a large part of videos in aforementioned datasets are not strictly registered. Besides, the VOT-RGBT challenge committee states that a part of images in the dataset they provided go out of sync [2].

(a) Sequence *gateBike* (b) Sequence *gateMan*

Figure 22: Examples of unregistered visible and infrared images.

Therefore, it is quite important for the RGB-infrared trackers to be able to handle unregistered image pairs. To the best of our knowledge, so far only the work of Xiao et al. [94] has considered the misalignment between visible and infrared images. They proposed methods to deal with scale, translation and rotation, respectively. Apart from this work, in all deep learning or correlation filter-based RGB-infrared trackers, the misalignment has not been considered by far.

Based on previous discussion on different levels of fusion tracking, we believe that the decision-level fusion tracking methods are more likely to handle this well, since they have less restricted requirement on the alignment of image pairs. One method to handle unregistered image pairs is to employ the image preprocessing to make the images well-aligned. Another way is to work out the transformation between visible and infrared images firstly, and then transform the bounding box of one modality to another according to the transformation. For instance, if visible image is the primary modality, then one can transform the bounding box obtained from infrared images to the visible modality.

[2] http://www.votchallenge.net/vot2019/participation.html

7.6. Infrared network

As discussed in Section 7.2.1, an infrared network is crucial for better fusion tracking performance. However, currently there is still a lack of datasets which are large enough to train RGB-infrared fusion tracking algorithms from the scratch. Besides, although thermal infrared object tracking is also an active research topic, most thermal object trackers are trained using visible images. For instance, Li et al. [19] proposed a hierarchical spatial-aware Siamese network for infrared object tracking. However, that network was trained using visible images from the large video detection dataset (ILSVRC2015) [127]. Similarly, Liu et al. [20] also trained their infrared tracker using visible images.

Although the network trained using visible images can track objects in infrared videos in some cases [19], their performance is not as good as a network trained using infrared images. This is because that the visible images differ from the infrared images in imaging mechanism, thus the network trained using visible images may not be able to extract effective infrared features.

To obtain an infrared network, one can apply transfer learning [128] to tackle the lack of infrared images. Specifically, one can firstly train the model using visible images and then fine-tune the model with infrared images. For instance, the infrared images from GTOT [25], RGBT234 [71] or the thermal object tracking benchmark [18] can be utilized. Alternatively, one can capture more infrared images and annotate them as training data. However, this process is time-consuming, labor-intensive and tedious.

Apart from above-mentioned methods, one can also try to generate infrared images from RGB images, and then use the generated infrared images to train the infrared model. This was not feasible until the emergence of GAN. In 2019, Zhang et al. [21] proposed generating infrared images from RGB images using image-to-image translation models and training an end-to-end infrared network using these synthetic images, as shown in Fig. 23. Their work is based on the fact that there are abundant RGB images available, such as the ImageNet dataset and many other large-scale datasets [129, 130]. Specifically, they employed two GAN model, namely pix2pix [131] and CycleGAN [132] to generate infrared images. Although the generated infrared images were not visually good, that study demonstrated that they can be utilized to train an infrared tracking model. Better performance was achieved than that of the model trained on available real infrared data.

7.7. Unsupervised fusion tracking

As discussed before, currently there is a lack of datasets which are large enough to train RGB-infrared fusion tracking algorithms from the scratch. Therefore, we have suggested creating larger RGB-infrared

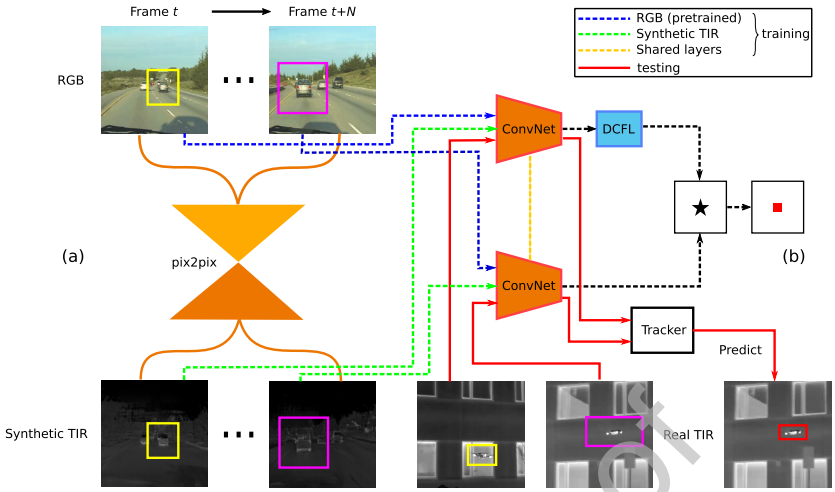

Figure 23: Training thermal infrared trackers with generated infrared images using GAN [21]. (a)Image-to-image translation component for generating a large labeled synthetic infrared tracking dataset. Blue dashed line represents the baseline RGB training model and the green dashed line represents the synthetic data training model. (b)Two-branch architecture for training the network to obtain adaptive features for infrared tracking.

datasets in Section 7.3. However, even though a large number of aligned visible and infrared images are available, the annotations of these images are tedious, time-consuming and labor-intensive. The reason is that by far almost all deep learning-based trackers are trained using a supervised manner, which requires ground truth. Furthermore, to the best of our knowledge, there are no published unsupervised RGB-infrared fusion tracking algorithms till now. Therefore, we think that the unsupervised fusion tracking will be a key research direction in the future.

Indeed, in other computer vision fields, such as object detection and segmentation, a current active research topic and also a crucial trend in the future is the weakly-supervised methods [133–136] and even unsupervised methods [137, 138]. In object tracking based on visible images, researchers have also started to develop unsupervised trackers. For examples, in 2019, Wang et al. [139] proposed an unsupervised deep tracking method (UDT). That method was based on the forward and backward computation of a tracking process. That work is among the earliest unsupervised RGB trackers. Although the tracking performance was not so good as supervised trackers, the idea is innovative and helpful in the development of unsupervised video object tracking.

7.8. Long-term fusion tracking

To the best of our knowledge, all published RGB-infrared tracking algorithms by far are designed for short-term fusion tracking. In addition, the well-known VOT challenge only has short-term RGBT subchal-

lenge currently. However, in practice, it is frequent that objects go inside and outside a scene. Therefore, existing RGB-infrared fusion trackers cannot handle this since they are not able to re-detect and recognize the same object. Besides, these trackers cannot handle full occlusion and failure of tracking.

A tracker must be able to handle long-term tracking task to be applied in practical applications [140]. Therefore, we believe that another trend in the development of RGB-infrared fusion tracking is the long-term RGB-infrared tracking algorithms.

7.9. Multi-object fusion tracking based on deep learning and correlation filter

Although there are several multi-object fusion tracking algorithms as introduced in Section 3.3.6, these methods are all traditional fusion tracking approaches as discussed in Section 3.3.1, and none of them have utilized deep features or correlation filter. Therefore, their performances are limited especially in challenging environments. We believe that the multi-object fusion tracking will be an important research direction of fusion tracking in the future, especially those algorithms based on deep learning and correlation filter.

8. Conclusion

Fusion tracking based on infrared and visible images (RGB-infrared fusion tracking) has attracted considerable attention and made significant progress in the past few years. Thus, we comprehensively survey existing RGB-infrared fusion tracking methods in the literature. These approaches can be divided into five categories: traditional methods, sparse representation-based, graph-based, correlation filter-based, and deep learning-based methods. Each category is introduced and summarized according to core idea and representative methods. Dataset is of vital importance in the training and test of fusion trackers, thus, we briefly review existing datasets containing visible and infrared sequence pairs. We then summarize several frequently used tracking performance evaluation metrics of fusion trackers. Furthermore, we summarize and analyze main results on public large-scale datasets to potentially provide an objective performance reference for researchers in the field of RGB-infrared fusion tracking. We observe that the deep learning-based methods give the leading performance and thus providing the most promising research direction in RGB-infrared fusion tracking. Finally, we propose some prospects based on our observation. This paper provides interested readers with an organized overview of the RGB-infrared fusion tracking and can serve a starting point for researchers who are interested in this field.

Acknowledgment

This paper is sponsored by National Program on Key Basic Research Project (2014CB744903), National Natural Science Foundation of China (61973212, 61673270), Shanghai Industrial Strengthening Project (GYQJ-2017-5-08), Shanghai Science and Technology Committee Research Project (17DZ1204304), Civil Aviation Pre-Research Projects and Shanghai Engineering Research Center of Civil Aircraft Flight Testing.

References

[1] L. Liu, J. Xing, H. Ai, X. Ruan, Hand posture recognition using finger geometric feature, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), IEEE, 2012, pp. 565–568.

[2] V. A. Laurense, J. Y. Goh, J. C. Gerdes, Path-tracking for autonomous vehicles at the limit of friction, in: 2017 American Control Conference (ACC), IEEE, 2017, pp. 5586–5591.

[3] J. Severson, Human-digital media interaction tracking, uS Patent 9,713,444 (Jul. 25 2017).

[4] A. Ali, A. Jalil, J. Niu, X. Zhao, S. Rathore, J. Ahmed, M. A. Iftikhar, Visual object trackingclassical and contemporary approaches, Frontiers of Computer Science 10 (1) (2016) 167–188.

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

[6] J. F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE transactions on pattern analysis and machine intelligence 37 (3) (2015) 583–596.

[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, 2005.

[8] T. Lindeberg, Scale invariant feature transform.

[9] H. Nam, B. Han, Learning multi-domain convolutional neural networks for visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4293–4302.

[10] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with siamese region proposal network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8971–8980.

[11] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, Siamrpn++: Evolution of siamese visual tracking with very deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[12] W. Zhou, L. Wen, L. Zhang, D. Du, T. Luo, Y. Wu, Siamman: Siamese motion-aware network for visual tracking, arXiv preprint arXiv:1912.05515.

[13] M. Danelljan, G. Hager, F. Shahbaz Khan, M. Felsberg, Convolutional features for correlation filter based visual tracking, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 58–66.

[14] M. Danelljan, G. Bhat, F. Shahbaz Khan, M. Felsberg, Eco: efficient convolution operators for tracking, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6638–6646.

[15] M. Danelljan, A. Robinson, F. S. Khan, M. Felsberg, Beyond correlation filters: Learning continuous convolution operators for visual tracking, in: European Conference on Computer Vision, Springer, 2016, pp. 472–488.

[16] J. Choi, H. Jin Chang, T. Fischer, S. Yun, K. Lee, J. Jeong, Y. Demiris, J. Young Choi, Context-aware deep feature compression for high-speed visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 479–488.

[17] J. W. Davis, V. Sharma, Fusion-based background-subtraction using contour saliency, in: Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2005, pp. 11–11.

[18] A. Berg, J. Ahlberg, M. Felsberg, A thermal object tracking benchmark, in: 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2015, pp. 1–6.

[19] X. Li, Q. Liu, N. Fan, Z. He, H. Wang, Hierarchical spatial-aware siamese network for thermal infrared object tracking, Knowledge-Based Systems 166 (2019) 71–81.

[20] Q. Liu, X. Lu, Z. He, C. Zhang, W.-S. Chen, Deep convolutional neural networks for thermal infrared object tracking, Knowledge-Based Systems 134 (2017) 189–198.

[21] L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, M. Danelljan, F. S. Khan, Synthetic data generation for end-to-end thermal infrared tracking, IEEE Transactions on Image Processing 28 (4) (2018) 1837–1850.

[22] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, H. Zhou, Learning Modality-Consistency Feature Templates: A Robust RGB-Infrared Tracking System, IEEE Transactions on Industrial Electronics 66 (12) (2019) 9887–9897.

[23] X. Lan, M. Ye, S. Zhang, P. C. Yuen, Robust collaborative discriminative learning for rgb-infrared tracking, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 7008–7015.

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

[25] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, L. Lin, Learning collaborative sparse representation for grayscale-thermal tracking, IEEE Transactions on Image Processing 25 (12) (2016) 5743–5756.

[26] C. Li, X. Wu, N. Zhao, X. Cao, J. Tang, Fusing two-stream convolutional neural networks for rgb-t object tracking, Neurocomputing 281 (2018) 78–85.

[27] C. Li, N. Zhao, Y. Lu, C. Zhu, 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.

[28] C. Li, C. Zhu, Y. Huang, J. Tang, L. Wang, Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking, in: Proceedings of ECCV, 2018, pp. 808–823.

[29] C. Li, C. Zhu, J. Zhang, B. Luo, X. Wu, J. Tang, Learning local-global multi-graph descriptors for rgb-t object tracking, IEEE Transactions on Circuits and Systems for Video Technology.

[30] S. Zhai, P. Shao, X. Liang, X. Wang, Fast RGB-T tracking via cross-modal correlation filters, Neurocomputing 334 (2019) 172–181.

[31] H. Liu, F. Sun, Fusion tracking in color and infrared images using sequential belief propagation, in: 2008 IEEE International Conference on Robotics and Automation, IEEE, 2008, pp. 2259–2264.

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

[33] S. R. Schnelle, A. L. Chan, Enhanced target tracking through infrared-visible image fusion, in: 14th International Conference on Information Fusion, IEEE, 2011, pp. 1–8.

[34] A. Leykin, R. Hammoud, Pedestrian tracking by fusion of thermal-visible surveillance videos, Machine Vision and Applications 21 (4) (2010) 587–595.

[35] K. Ren, X. Zhang, Y. Han, Y. Hou, Robust night target tracking via infrared and visible video fusion, in: Applications of Digital Image Processing XLI, Vol. 10752, International Society for Optics and Photonics, 2018, p. 1075206.

[36] X. Zhang, G. Xiao, P. Ye, D. Qiao, J. Zhao, S. Peng, Object fusion tracking based on visible and infrared images using

fully convolutional siamese networks, in: Proceedings of the 22nd International Conference on Information Fusion, IEEE, 2019.

[37] X. Zhang, P. Ye, J. Liu, K. Gonge, G. Xiao, Decision-level visible and infrared fusion tracking via siamese networks, in: Proceedings of the 9th Chinese Conference on Information Fusion, 2019.

[38] X. Lan, M. Ye, R. Shao, B. Zhong, D. K. Jain, H. Zhou, Online non-negative multi-modality feature template learning for rgb-assisted infrared tracking, IEEE Access.

[39] A. P. James, B. V. Dasarathy, Medical image fusion: A survey of the state of the art, Information Fusion 19 (2014) 4–19.

[40] Z. Wang, Y. Ma, J. Gu, Multi-focus image fusion using pcnn, Pattern Recognition 43 (6) (2010) 2003–2016.

[41] H. Ghassemian, A review of remote sensing image fusion methods, Information Fusion 32 (2016) 75–89.

[42] K. Ma, K. Zeng, Z. Wang, Perceptual quality assessment for multi-exposure image fusion, IEEE Transactions on Image Processing 24 (11) (2015) 3345–3356.

[43] J. Ma, C. Chen, C. Li, J. Huang, Infrared and visible image fusion via gradient transfer and total variation minimization, Information Fusion 31 (2016) 100–109.

[44] H. Yin, Tensor sparse representation for 3-d medical image fusion using weighted average rule, IEEE Transactions on Biomedical Engineering 65 (11) (2018) 2622–2633.

[45] P. Hill, M. E. Al-Mualla, D. Bull, Perceptual image fusion using wavelets, IEEE transactions on image processing 26 (3) (2016) 1076–1088.

[46] C. He, Q. Liu, H. Li, H. Wang, Multimodal medical image fusion based on ihs and pca, Procedia Engineering 7 (2010) 280–285.

[47] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Information Fusion 24 (2015) 147–164.

[48] T. Wan, Z. Qin, An application of compressive sensing for image fusion, International Journal of Computer Mathematics 88 (18) (2011) 3915–3930.

[49] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, K. He, A survey of infrared and visual image fusion methods, Infrared Physics & Technology 85 (2017) 478–501.

[50] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: A survey of the state of the art, Information Fusion 33 (2017) 100–112.

[51] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, X. Wang, Deep learning for pixel-level image fusion: Recent advances and future prospects, Information Fusion 42 (2018) 158–173.

[52] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: A survey, Information Fusion 45 (2019) 153–178.

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

[54] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Information Fusion 36 (2017) 191–207.

[55] X. Yan, S. Z. Gilani, H. Qin, A. Mian, S. Member, S. Z. Gilani, H. Qin, A. Mian, Unsupervised Deep Multi-focus Image Fusion (2018) 1–11arXiv:1806.07272.

[56] K. Xia, H. Yin, J. Wang, A novel improved deep convolutional neural network model for medical image fusion, Cluster Computing (2018) 1–13.

[57] K. R. Prabhakar, V. S. Srikar, 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.

[58] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Information Fusion 48 (June 2018) (2019) 11–26.

[59] Y. Liu, X. Chen, J. Cheng, H. Peng, Z. Wang, Infrared and visible image fusion with convolutional neural networks, International Journal of Wavelets, Multiresolution and Information Processing 16 (03) (2018) 1850018.

[60] H. Li, X. Wu, Densefuse: A fusion approach to infrared and visible images, IEEE Transactions on Image Processing 28 (5) (2018) 2614–2623.

[61] Q. Yan, D. Gong, Y. Zhang, Two-stream convolutional networks for blind image quality assessment, IEEE Transactions on Image Processing 28 (5) (2018) 2200–2211.

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

[63] T. A. Biresaw, A. Cavallaro, C. S. Regazzoni, Tracker-level fusion for robust bayesian visual tracking, IEEE Transactions on Circuits and Systems for Video Technology 25 (5) (2015) 776–789.

[64] T. Vojir, J. Matas, J. Noskova, Online adaptive hidden markov model for multi-tracker fusion, Computer Vision and Image Understanding 153 (2016) 109–119.

[65] E. Gundogdu, H. Ozkan, H. Seckin Demir, H. Ergezer, E. Akagunduz, S. Kubilay Pakin, Comparison of infrared and visible imagery for object tracking: Toward trackers with superior ir performance, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 1–9.

[66] C. Xie, N. Wang, W. Zhou, W. Li, H. Li, Multi-tracker fusion via adaptive outlier detection, Multimedia Tools and Applications 78 (2) (2019) 2227–2250.

[67] Y. Yang, Y. Zhang, D. Li, Z. Wang, Parallel correlation filters for real-time visual tracking, Sensors 19 (10) (2019) 2362.

[68] C. Li, S. Hu, S. Gao, J. Tang, Real-time grayscale-thermal tracking via laplacian sparse representation, in: International Conference on Multimedia Modeling, Springer, 2016, pp. 54–65.

[69] C. Li, X. Sun, X. Wang, L. Zhang, J. Tang, Grayscale-thermal object tracking via multitask laplacian sparse representation, IEEE Transactions on Systems, Man, and Cybernetics: Systems 47 (4) (2017) 673–681.

[70] L. Li, C. Li, Z. Tu, J. Tang, A fusion approach to grayscale-thermal tracking with cross-modal sparse representation, in: Chinese Conference on Image and Graphics Technologies, Springer, 2018, pp. 494–505.

[71] C. Li, X. Liang, Y. Lu, N. Zhao, J. Tang, Rgb-t object tracking: benchmark and baseline, Pattern Recognition (2019) 106977.

[72] C. Li, C. Zhu, S. Zheng, B. Luo, J. Tang, Two-stage modality-graphs regularized manifold ranking for rgb-t tracking, Signal Processing: Image Communication 68 (2018) 207–217.

[73] M. Ding, Y. Yao, L. Wei, Y. Cao, Visual tracking using locality-constrained linear coding and saliency map for visible light and infrared image sequences, Signal Processing: Image Communication 68 (2018) 13–25.

[74] Y. Wang, C. Li, 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.

[75] N. Xu, G. Xiao, X. Zhang, 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.

[76] X. Zhang, X. Zhang, X. Du, X. Zhou, J. Yin, Learning multi-domain convolutional network for rgb-t visual tracking, in: 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, 2018, pp. 1–6.

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

[78] S. Qiu, J. Luo, S. Yang, M. Zhang, W. Zhang, A moving target extraction algorithm based on the fusion of infrared and visible images, Infrared Physics & Technology.

[79] C. Luo, B. Sun, K. Yang, T. Lu, W.-C. Yeh, Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme, Infrared Physics & Technology 99 (2019) 265–276.

[80] X. Yun, Y. Sun, X. Yang, N. Lu, Discriminative fusion correlation learning for visible and infrared tracking, Mathematical Problems in Engineering 2019.

[81] C. Tang, Y. Ling, H. Yang, X. Yang, W. Tong, Decision-level fusion tracking for infrared and visible spectra based on deep learning, Laser & Optoelectronics Progress 56 (2019) 71502–1.

[82] Y. Zhu, C. Li, B. Luo, J. Tang, X. Wang, Dense feature aggregation and pruning for rgbt tracking, arXiv preprint arXiv:1907.10451.

[83] R. Yang, Y. Zhu, X. Wang, C. Li, J. Tang, Learning target-oriented dual attention for robust rgb-t tracking, arXiv preprint arXiv:1908.04441.

[84] C. Li, A. Lu, A. Zheng, Z. Tu, J. Tang, Multi-adapter rgbt tracking, arXiv preprint arXiv:1907.07485.

[85] C. O'Conaire, N. E. O'Connor, E. Cooke, A. F. Smeaton, Comparison of fusion methods for thermo-visual surveillance tracking, in: 2006 9th International Conference on Information Fusion, IEEE, 2006, pp. 1–7.

[86] A. L. Chan, S. R. Schnelle, Target tracking using concurrent visible and infrared imageries, in: Signal Processing, Sensor Fusion, and Target Recognition XXI, Vol. 8392, International Society for Optics and Photonics, 2012, p. 83920P.

[87] A. L. Chan, S. R. Schnelle, Fusing concurrent visible and infrared videos for improved tracking performance, Optical Engineering 52 (1) (2013) 017004.

[88] F. Bunyak, K. Palaniappan, S. K. Nath, G. Seetharaman, Geodesic active contour based fusion of visible and infrared video for persistent object tracking, in: 2007 IEEE Workshop on Applications of Computer Vision (WACV'07), IEEE, 2007, pp. 35–35.

[89] X. Yun, Z. Jing, G. Xiao, B. Jin, C. Zhang, A compressive tracking based on time-space kalman fusion model, Science China Information Sciences 59 (1) (2016) 1–15.

[90] M. Isard, A. Blake, Condensationconditional density propagation for visual tracking, International journal of computer vision 29 (1) (1998) 5–28.

[91] N. Cvejic, S. G. Nikolov, H. D. Knowles, A. Loza, A. Achim, D. R. Bull, C. N. Canagarajah, The effect of pixel-level fusion on object tracking in multi-sensor surveillance video, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–7.

[92] R. Péteri, O. Šiler, Object tracking using joint visible and thermal infrared video sequences.

[93] J. Wang, D. Chen, S. Li, Y. Yang, Infrared and visible fusion for robust object tracking via local discrimination analysis, Journal of Computer-Aided Design & Computer Graphics 26 (2014) 870–878.

[94] X. Gang, Y. Xiao, J. Wu, A new tracking approach for visible and infrared sequences based on tracking-before-fusion, International Journal of Dynamics & Control 4 (1) (2016) 40–51.

[95] G. Xiao, X. Yun, J. Wu, A multi-cue mean-shift target tracking approach based on fuzzified region dynamic image fusion, Science China Information Sciences 55 (3) (2012) 577–589.

[96] C. Ó. Conaire, N. E. OConnor, A. Smeaton, Thermo-visual feature fusion for object tracking using multiple spatiogram trackers, Machine Vision and Applications 19 (5-6) (2008) 483–494.

[97] B. Cai, C. l. Zhang, Z. Li, Tracking infrared-visible target with joint histogram, Journal of Guangxi Normal University (Natural Science Edition) 35 (3) (2018) 37–44.

[98] K. Wang, H. Wei, C. Chen, K. Cao, Target tracking based on infrared and visible light fusion, Computer Systems and Applications 27 (1) (2018) 149–153.

[99] K. S. Kumar, G. Kavitha, R. Subramanian, G. Ramesh, Visual and thermal image fusion for uav based target tracking, in: MATLAB-A Ubiquitous Tool for the Practical Engineer, IntechOpen, 2011.

[100] S. Mangale, M. Khambete, Camouflaged target detection and tracking using thermal infrared and visible spectrum imaging, in: The International Symposium on Intelligent Systems Technologies and Applications, Springer, 2016, pp. 193–207.

[101] L. S. Laurent, D. Prevost, X. P. Maldague, Context-independant video monitoring of mobile objects with color/thermal sensor, in: Optics and Photonics for Counterterrorism and Crime Fighting, Vol. 5616, International Society for Optics and Photonics, 2004, pp. 16–25.

[102] X. Mei, H. Ling, Robust visual tracking using 1 minimization, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 1436–1443.

[103] Y. Wu, E. Blasch, G. Chen, L. Bai, H. Ling, Multiple source data fusion via sparse representation for robust visual tracking, in: 14th International Conference on Information Fusion, IEEE, 2011, pp. 1–8.

[104] J. Ning, J. Yang, S. Jiang, L. Zhang, M.-H. Yang, Object tracking via dual linear structured svm and explicit feature map, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4266–4274.

[105] C. Li, L. Lin, W. Zuo, J. Tang, Learning patch-based dynamic graph for visual tracking, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[106] C. Li, L. Lin, W. Zuo, J. Tang, M.-H. Yang, Visual tracking via dynamic graph learning, IEEE transactions on pattern analysis and machine intelligence.

[107] K. Zhang, Q. Liu, Y. Wu, M.-H. Yang, Robust visual tracking via convolutional networks without training, IEEE Transactions on Image Processing 25 (4) (2016) 1779–1792.

[108] N. Xu, G. Xiao, F. He, X. Zhang, D. P. Bavirisetti, Object tracking via deep multi-view compressive model for visible and infrared sequences, in: Proceedings of the 21st International Conference on Information Fusion (FUSION), IEEE, 2018, pp. 941–948.

[109] X. Zhang, P. Ye, S. Peng, J. Liu, K. Gong, G. Xiao, Siamft: An rgb-infrared fusion tracking method via fully convolutional siamese networks, IEEE Access 7 (2019) 122122–122133.

[110] X. Yun, Z. Jing, B. Jin, Visible and infrared tracking based on multi-view multi-kernel fusion model, Optical Review 23 (2) (2016) 244–253.

[111] M. Wang, Y. Liu, Z. Huang, Large margin object tracking with circulant feature maps, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4021–4029.

[112] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, K. Schindler, Mot16: A benchmark for multi-object tracking, arXiv preprint arXiv:1603.00831.

[113] J. W. Davis, V. Sharma, Background-subtraction using contour-based fusion of thermal and visible imagery, Computer vision and image understanding 106 (2-3) (2007) 162–182.

[114] A. Torabi, G. Massé, G.-A. Bilodeau, An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications, Computer Vision and Image Understanding 116 (2) (2012) 210–221.

[115] S. Hwang, J. Park, N. Kim, Y. Choi, I. So Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1037–1045.

[116] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, D. Riahi, Thermal–visible registration of human silhouettes: A similarity measure performance evaluation, Infrared Physics & Technology 64 (2014) 79–86.

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

[118] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Hager, A. Lukezic, A. Eldesokey, et al., The visual object tracking vot2016 challenge results, in: Proceedings in European Conference on Computer Vision (ECCV) workshops, 2016.

[119] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 101–117.

[120] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P. H. Torr, Fast online object tracking and segmentation: A unifying approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1328–1338.

[121] H. Kiani Galoogahi, A. Fagg, S. Lucey, Learning background-aware correlation filters for visual tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1135–1143.

[122] K. Dai, D. Wang, H. Lu, C. Sun, J. Li, Visual tracking via adaptive spatially-regularized correlation filters, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4670–4679.

[123] Y. Sun, C. Sun, D. Wang, Y. He, H. Lu, Roi pooled correlation filters for visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5783–5791.

[124] M. Zhang, Q. Wang, J. Xing, J. Gao, P. Peng, W. Hu, S. Maybank, Visual tracking via spatially aligned correlation filters network, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 469–485.

[125] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, P. H. Torr, End-to-end representation learning for correlation filter based tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2805–2813.

[126] Q. Wang, J. Gao, J. Xing, M. Zhang, W. Hu, Dcfnet: Discriminant correlation filters network for visual tracking, arXiv preprint arXiv:1704.04057.

[127] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, 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.

[128] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (10) (2010) 1345–1359.

[129] L. Huang, X. Zhao, K. Huang, Got-10k: A large high-diversity benchmark for generic object tracking in the wild, arXiv preprint arXiv:1810.11981.

[130] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, H. Ling, Lasot: A high-quality benchmark for large-scale single object tracking, arXiv preprint arXiv:1809.07845.

[131] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

[132] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[133] H. Bilen, M. Pedersoli, T. Tuytelaars, Weakly supervised object detection with convex clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1081–1089.

[134] H. Bilen, A. Vedaldi, Weakly supervised deep detection networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2846–2854.

[135] D. Pathak, P. Krahenbuhl, T. Darrell, Constrained convolutional neural networks for weakly supervised segmentation, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1796–1804.

[136] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, S. Yan, Stc: A simple to complex framework for weakly-supervised semantic segmentation, IEEE transactions on pattern analysis and machine intelligence 39 (11) (2016) 2314–2320.

[137] E. Haller, M. Leordeanu, Unsupervised object segmentation in video by efficient selection of highly probable positive features, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5085–5093.

[138] Y. Yang, A. Loquercio, D. Scaramuzza, S. Soatto, Unsupervised moving object detection via contextual information separation, arXiv preprint arXiv:1901.03360.

[139] N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, H. Li, Unsupervised deep tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1308–1317.

[140] H. Lu, D. Wang, Online Visual Tracking, Springer, 2019.

Credit Author Statement

Xingchen Zhang: Conceptualization, Investigation, Writing- Original draft preparation, Writing - Review & Editing
Ping Ye: Visualization, Investigation, Data Curation Henry Leung: Reviewing and Editing
Ke Gong: Visualization, Investigation Gang Xiao: Supervision, Reviewing and Editing, Funding acquisition, Project administration

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.