

Quality-aware Feature Aggregation Network for Robust RGBT Tracking

Yabin Zhu, Chenglong Li, Jin Tang, Bin Luo

Abstract—This paper investigates how to perform robust visual tracking in adverse and challenging conditions using complementary visual and thermal infrared data (RGBT tracking). We propose a novel deep network architecture called quality-aware Feature Aggregation Network (FANet) for robust RGBT tracking. Unlike existing RGBT trackers, our FANet aggregates hierarchical deep features within each modality to handle the challenge of significant appearance changes caused by deformation, low illumination, background clutter and occlusion. In particular, we employ the operations of max pooling to transform these hierarchical and multi-resolution features into uniform space with the same resolution, and use 1×1 convolution operation to compress feature dimensions to achieve more effective hierarchical feature aggregation. To model the interactions between RGB and thermal modalities, we elaborately design an adaptive aggregation subnetwork to integrate features from different modalities based on their reliabilities and thus are able to alleviate noise effects introduced by low-quality sources. The whole FANet is trained in an end-to-end manner. Extensive experiments on large-scale benchmark datasets demonstrate the high-accurate performance against other state-of-the-art RGBT tracking methods.

Index Terms—RGBT tracking, Feature aggregation, Quality-aware fusion.

I. INTRODUCTION

Thermal cameras recently become more economically affordable and have been applied to many computer vision tasks, such as object segmentation [1], [2], person Re-ID [3] and pedestrian detection [4], [5], to name a few. They are insensitive to lighting condition and have a strong ability to penetrate smog and haze [6], and thus able to provide strong complementary information for visible cameras. However, thermal infrared data are limited by the challenge of thermal crossover and lacking in color and texture details. Visible spectrum information can make up for it, and Fig. 1 shows some examples. Therefore, effective fusion of visible and thermal data has big potential in handling various challenges for robust visual tracking.

As a subbranch of visual tracking, RGBT tracking aims to take advantage of complementary information from visible and thermal infrared spectra to estimate the state of a specific instance in a video, given the ground truth bounding box in the first frame. Recent studies on RGBT tracking mainly focus on two aspects. The one is aiming to learn robust feature representation via the usage of RGB and thermal data [7], [8], [9] and achieves promising tracking performance. These works rely on either handcrafted features or highly

semantic deep features. Handcrafted features are too weak to handle the challenges of significant appearance changes caused by deformation, low illumination, background clutter and occlusion within each modality. Although highly semantic features are effective to distinguish target semantics, they are unable to meet the purpose of visual tracking, which is to locate objects accurately rather than just to infer their semantic categories. Therefore, relying on handcrafted features or highly semantic features only is difficult to handle various challenges in RGBT tracking.

The another one focuses on introducing modality weights to achieve adaptive aggregation of different source data [10], [11]. Early works employ the reconstruction residues [10], [7] to regularize modality weight learning. Lan *et al.* [11] uses the max-margin principle to optimize the modality weights according to classification scores. However, these works would fail when reconstruction residuals or classification scores do not reliably reflect modal reliability. For example, for the *BlurCar* sequence in GTOT dataset [10], the quality of RGB modality is much better than thermal modality. But the reconstruction residue on RGB modality computed by the model in [10] is 37.60, which is higher than the residue of thermal modality 32.67. Therefore, the uncertainty of reliability degree in different modalities is still an uncrossed obstacle for these type of works.

To solve these problems, we propose a novel architecture, namely quality-aware Feature Aggregation Network (FANet), for robust RGBT tracking. Within each modality, our FANet first aggregates hierarchical multi-resolution feature maps into a uniform space at the same resolution. Shallow features encode fine-grained spatial details such as target positions and help achieving precise target localizations, while deep features are more effective to capture target semantics. Existing methods [12], [13], [14], [15] usually select part of deep features extracted from pre-trained networks to learn correlation filter models for visual tracking. However, using only part of features and unchanged network parameters in training and tracking processes would limit tracking performance. Therefore, we propose an effective method to deploy all layer features and learn robust target representations in an end-to-end trained network. In a specific, we employ the operations of max pooling to convert their resolutions into uniform feature space, and use the convolution operations to suppress redundant features.

In different modalities, we elaborately design an adaptive aggregation subnetwork to integrate features from different modalities based on their reliability. Many efforts are devoted to calculate the modality weights that reflect the reliability

The authors are with School of Computer Science and Technology, Anhui University, Hefei 230601, China.

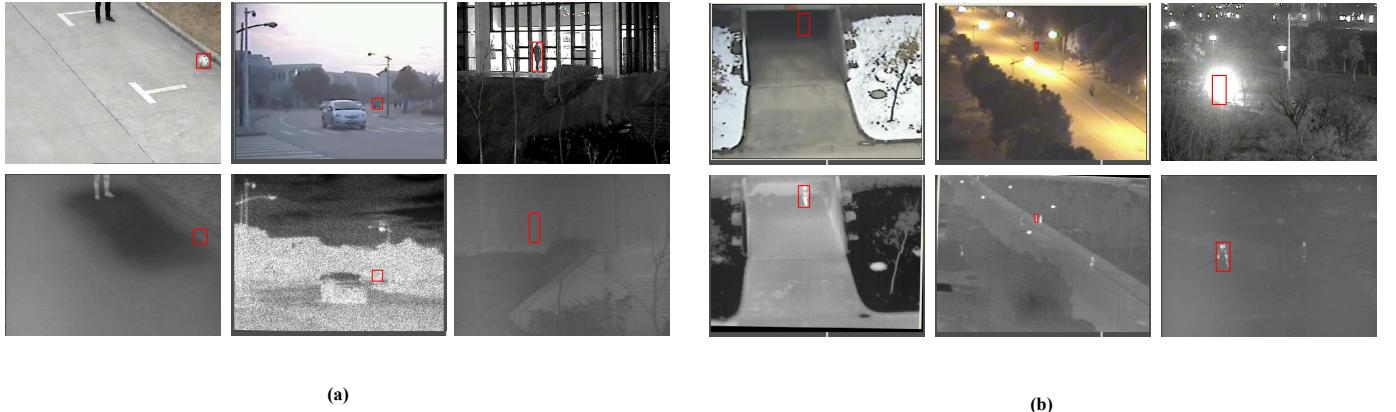


Fig. 1. Typical complementary cases of RGB and thermal data. (a) RGB data source is better than thermal ones, where thermal spectrum is disturbed by thermal crossover, noise and glass. (b) Thermal sources better than RGB ones, where visible spectrum is disturbed by low illumination and high illumination.

degrees to make adaptive aggregation of different modalities, and thus achieve significant improvement for tracking performance [10], [11]. In this paper, we come up with a more effective method to compute the modality weights in an end-to-end learning framework. Given the aggregated features, we first fuse all modalities feature using a concatenation operation and compress the global spatial information into a channel descriptor utilizing the global average pooling. Then, we use a fully connected layer to learn a nonlinear interaction between channels, and the fully connected layers and the SoftMax activation function are used to predict the modality weights that control the information flows of different modalities for adaptive aggregation. The object tracking is performed via the binary classification in the multi-domain learning network [16].

We summarize our major contributions of our work as follows.

- We propose a novel end-to-end trained deep network to learn a powerful feature representations for robust RGBT tracking. The network aggregates hierarchical deep features effectively and multi-modal information adaptively to handle the challenge of significant appearance changes caused by deformation, low illumination, background clutter and occlusion in RGBT tracking.
- We design a feature aggregation module and an adaptive aggregation module to fuse multiple features within each modality and in different modalities respectively. In particular, the operations of max pooling are employed to transform hierarchical deep features into a uniform space with the same resolution, and the operations of 1×1 convolution are used to compress feature dimensions for more effective hierarchical feature aggregation. In addition, the operations of global average pooling, fully connected layer and SoftMax activation function are used to predict the modality weights for adaptive aggregation.
- Extensive experiments on large-scale benchmark datasets prove that the FANet significantly better than other state-of-the-art tracking methods with a clear margin and achieves 19 FPS.

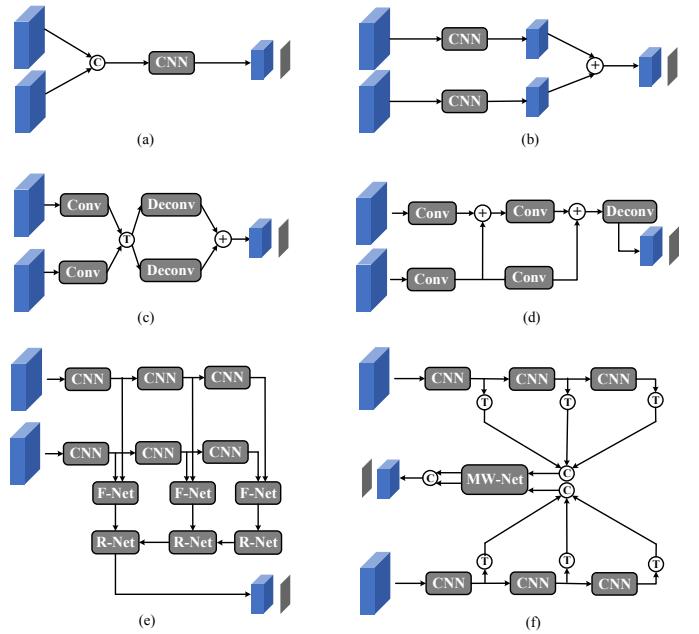


Fig. 2. Different existing architectures of multimodal fusion. (a) Early fusion. (b) Late fusion. (c) Architecture proposed by [17]. (d) Architecture proposed in [18]. (e) Architecture proposed in [19]. (f) Our architecture. ‘C’, ‘T’ and ‘+’ represent concatenation, transformation, and element-wise summation. And ‘F-Net’, ‘R-Net’ and ‘MW-Net’ indicate fusion, refinement and modality weight prediction networks, respectively.

II. RELATED WORK

Based on the relevance to our work, we review the related works from four research lines: RGB-Thermal fusion, hierarchical features for tracking, multi-domain network for tracking, and multimodal aggregation network architectures.

A. RGB-Thermal Fusion for Tracking

RGBT tracking become more and more attentions in computer vision community with the popularity of thermal infrared sensors [1], [3], [4], [5], [9]. One research stream is to introduce modality weights to achieve adaptive aggregation of different modalities [20], [10], [11]. Early works employ reconstruction residues [20], [10] to regularize the modality

weight learning and carry out object tracking in Bayesian filtering framework. Lan *et al.* [11] use the max-margin principle to optimize the modality weights according to classification scores. However, these methods would fail when reconstruction residuals or classification scores do not reliably reflect modal reliability.

The other research stream is to learn robust feature representations via the usage of multimodal data [7], [8], [9]. Li *et al.* [7] propose a weighted sparse representation regularized graph learning approach to construct a graph-based multimodal descriptor, and adopt structured SVM for tracking. And Li *et al.* [9] further consider the heterogeneous property between different modalities and noise effects of initial seeds in the cross-modal ranking model. Li *et al.* [8] design a FusionNet that select most discriminative feature maps from deep feature of CNN to avoid feature redundant and noises for adaptive aggregation of different modalities. These methods rely on hand-crafted features or highly semantically deep features to locate objects, therefore it is difficult to cope with the challenges of significant appearance changes caused by background clutter, low illumination, deformation and occlusion within each modality.

B. Hierarchical Features for Tracking

In recent literatures, several works started incorporating hierarchical deep features for the task of visual tracking [12], [13], [21], [22], [14], [15]. Ma *et al.* [12] interprets the hierarchical features of the convolutional layer as a non-linear correspondence of the representation of the image pyramid and explicitly uses these multiple levels of abstraction to represent the target object. In order to make full use of the multi-layer CNN features, Qi *et al.* [13] designs a Hedge method to hedge multiple trackers into more powerful ones. To take the interdependencies among different features into account, Zhang *et al.* [22] present a multi-task correlation filter to learn the correlation filters jointly. A combination of handcrafted low-level and hierarchical deep features is proposed by Danelljan *et al.* [14], [15] by employing an implicit interpolation model to construct the learning problem in the continuous spatial space, which makes multi-resolution feature maps integrated effectively. These methods, however, are based on the assumption that multiple features contribute equally or set the feature weights manually, therefore they cannot fully and effectively use features according to their reliabilities.

C. Multi-Domain Networks for Tracking

MDNet [16] has achieved the state-of-the-art performance on multiple datasets through multi-domain learning. Han *et al.* [23] proposes a random selection of a branches of CNN for online learning when target object appearance models need to be updated for better regularization, where each of the branches may have a different number of layers to maintain a variable level of abstraction of the target's appearance. The meta learning is introduced in MDNet to adjust the initialization of the deep networks [24], which could quickly adapt to robustly model a particular target in future frames.

Jung *et al.* [25] propose a real-time MDNet, where an improved RoIAlign technique extracts a more accurate feature representation of the target. All these methods are developed for single-modality tracking and we study them in the task of multi-modality tracking in this work.

D. Multimodal Fusion Network Architectures

Early fusion works simply concatenate input multimodal channels and then adopt the convolutional networks (CNN) to extract feature representations (*i.e.* Fig. 2 (a)). Some later fusion works additionally report the result of fusing multiple feature representations extracted from the CNN of each modality [8] (*i.e.* Fig. 2 (b)). Wang *et al.* [17] propose a structure for deconvolution of multiple modalities, in which an additional feature transformation network is introduced to associate two modalities by discovering common and modal-specific features (*i.e.* Fig. 2 (c)). It doesn't exploit any abundant intermediate feature of both modalities, and the training procedures are not end-to-end. Hazirbas *et al.* [18] proposes to use intermediate modal features because they simply add intermediate multimodal features, so the effective mid-level multimodal features cannot be fully utilized. Park *et al.* [19] proposes a network that can effectively utilize multi-level features simultaneously (*i.e.* Fig. 2 (e)). However, it neglect the reliability degrees of different modalities, and thus might be affected by noises of features and modalities. In this work, we take all above issues into account and propose a quality-aware feature aggregation network for effective multimodal fusion (*i.e.* Fig. 2 (f)).

III. FANET FRAMEWORK

In this section, we introduce the details of our FANet (quality-aware Feature Aggregation Network) framework, including network architecture, training procedure and tracker details.

A. Network Architecture

The overall network architecture of our FANet is shown in Fig. 3. FANet mainly consists of two parts, *i.e.*, a hierarchical feature aggregation subnetwork and an adaptive aggregation subnetwork. Followed by these two subnetworks, we add a binary classification layer with three fully connected layers and a SoftMax with Binary Cross Entropy (BCE) loss to carry out tracking task. The three fully connected layers (with 1024, 512, 2 output units respectively (see Fig. 3)) are combined with ReLUs and dropouts. The third fully connected layer with two output units include K branches, where each branch corresponds to a specific domain that is a video sequence for the tracking task. The last layer is a SoftMax with BCE loss to perform binary classification to distinguish the foreground(target object) and the background in each domain. And in the training phase, the instance embedding loss [25] function is also adopted to learn more discriminative target representation models. Unlike the feature extraction of MDNet [16], we extract the features of samples directly from the feature maps through a RoIAlign operation [25], which greatly reduces the computational burden and accelerates the running speed of FANet.

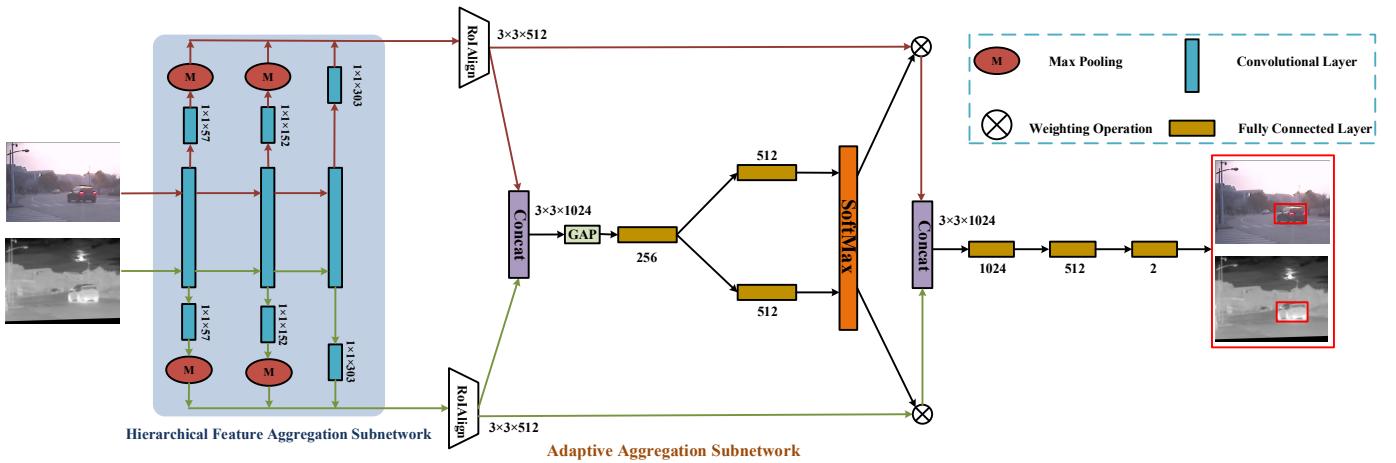


Fig. 3. Diagram of our FANet architecture that consists of a hierarchical feature aggregation subnetwork and an adaptive aggregation subnetwork.

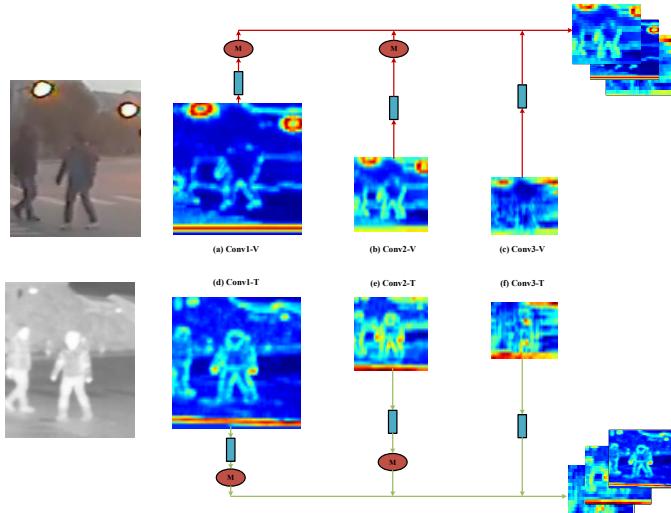


Fig. 4. Illustration of hierarchical feature aggregation.

B. Hierarchical Feature Aggregation Subnetwork

Feature aggregation plays an important role in improving tracking performance since diverse features are able to provide complementary advantages which could handle various challenges. However, existing methods [12], [13], [14], [15] usually select part of deep features extracted from pre-trained networks to learn correlation filter models for visual tracking. There are two major limitations. 1) They are unable to make the best use of hierarchical features since only part of features are employed. 2) These features are hard to handle all challenges well because the parameters of the used networks remain unchanged in training and tracking processes. To handle these problems, we propose an effective approach to aggregate the shallow-to-deep features of all layers in an end-to-end trained network, as shown in Fig. 3 and Fig. 4.

In a specific, we extract hierarchical features (*Conv1*-*Conv3*) using the improved VGG-M Network [26] pre-trained on the ImageNet dataset for informative target representations. In order to increase the receptive field and improve quality of representations of ROI, we refine the VGG network. Specif-

ically, we remove the max pooling layer followed by *conv2* layer and modify *conv3* to be a dilated convolutions [27] with rate $r = 3$. Features in different layers have different resolutions due to pooling operations. Therefore, to convert these features to the same resolution, we employ the max pooling operation to subsample high-resolution features. Furthermore, to transform hierarchical features into a uniform space and compress features, we add a convolution operation for each layer and adopt a local response normalization (LRN) layer to normalize these feature maps. It is worth noting that different modalities share the parameters in the first three convolutional layers (i.e., *Conv1*-*Conv3*).

That's because different modalities should share a large number of parameters to represent their collaboration information, and the shared structure can avoid many redundant parameters. Features of different layers should make different contributions to some certain video sequences as shallow features preserve the fine-grained details while deep features capture target semantics. Therefore, our feature aggregation subnetwork aggregates the spatial and semantic information of the two modalities from shallow to deep, and also compresses the feature channels so that more rich and effective feature representation can be obtained. Fig. 3 shows the details of the subnetwork configuration and partial parameters.

C. Adaptive Aggregation Subnetwork

Duo to different imaging properties of RGB and thermal sensors, their reliability degrees, representing the contributions to the tracking performance, should be different. Several works attempt to introduce modality weights to achieve adaptive aggregation of different source data [10], [7], [11]. They employ the reconstruction residues [10], [7] or classification scores [11] to regularize modality weight learning, and thus would fail when reconstruction residuals or classification scores do not reliably reflect modal reliability.

In this work, inspired by [28], we elaborately design an adaptive aggregation subnetwork to predict the modality weights in an end-to-end trained network. Note that our motivation is significantly different from [28] from the following

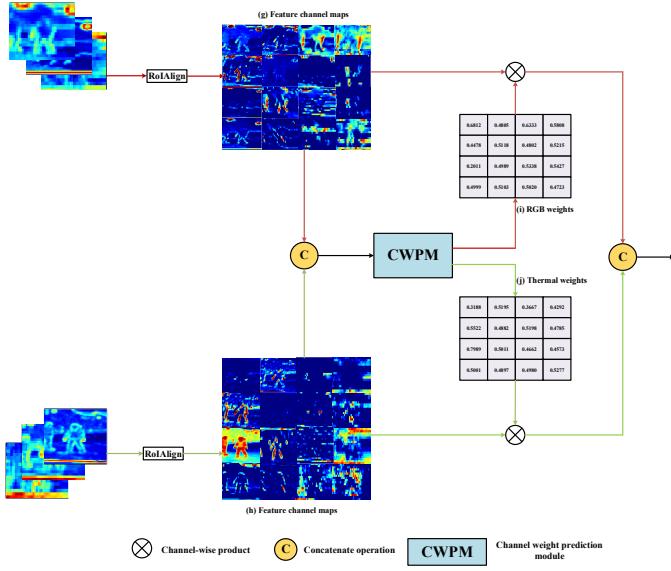


Fig. 5. Illustration of adaptive aggregation. (g) and (h) depict the feature channel maps, (i) and (j) denote the channel weights calculated from the feature maps. The details of the channel weight prediction module (CWPM) can be found in Fig. 3

two aspects. First, [28] presents a nonlinear approach to aggregate information from multiple kernels to realize the adaptive receptive field sizes of neurons. However, we are trying to adaptively aggregate the information of different modalities, and do not consider the different receptive fields of the same feature layer. Second, [28] aggregates the feature maps of different receptive fields with additive operation, while we integrate different modalities with concatenate operation, because the addition operation is likely to cause negative effects when the quality of some modal features is low. Our adaptive aggregation subnetwork has two steps including RGBT feature embedding and adaptive weighting.

RGBT feature embedding. We aim to adaptively integrate features from RGB and thermal sources according to their reliabilities to tracking performance. The basic thought is to utilize the gate mechanism to control the information flows from different modalities. To this end, we first fuse all modality features (shown in Fig. 3 and Fig. 5) using a concatenate operation. Then we compress global spatial information into a channel descriptor utilizing the global average pooling. Specifically, the c -th element of the compressed features f is calculated by:

$$f_c = \frac{1}{W \times H} \sum_{j=1}^W \sum_{k=1}^H X_c(j, k) \quad (1)$$

where $X = [X_1, \dots, X_C] \in \mathbb{R}^{C \times H \times W}$ is the input feature maps, and $X = [X_{rgb}, X_t]$. X_{rgb} and X_t are the aggregated feature maps of RGB and thermal sources, respectively. C denotes the number of channels, and H and W denote the height and width of feature maps respectively.

Adaptive weighting. A compact features $\mathbf{z} \in \mathbb{R}^{d \times 1}$ are created to learn nonlinear interactions between channels. This

is done through a fully connected layer with the reduction of dimensionality for better efficiency:

$$\mathbf{z} = \mathcal{F}(f) = \delta(Wf) \quad (2)$$

where δ denotes the *ReLU* function, and $\mathcal{F}(\cdot)$ indicates a fully connected operation. $W \in \mathbb{R}^{d \times C}$ denotes the filter of the fully connected layer, and in this paper d is set to 256.

Further, to adaptively aggregate information of different modalities, the fully connected layer and the SoftMax activation function are used. Note that we use a fully connected layer for each modality to weigh different modality features as follows:

$$\begin{aligned} \mathbf{V}_{rgb} &= \delta(\mathbf{W}_{21}\mathbf{z}), \mathbf{V}_t = \delta(\mathbf{W}_{22}\mathbf{z}), \\ a_c &= \frac{e^{\mathbf{V}_{rgb}^c}}{e^{\mathbf{V}_{rgb}^c} + e^{\mathbf{V}_t^c}}, b_c = \frac{e^{\mathbf{V}_t^c}}{e^{\mathbf{V}_{rgb}^c} + e^{\mathbf{V}_t^c}} \end{aligned} \quad (3)$$

where $\mathbf{V}_{rgb}, \mathbf{V}_t \in \mathbb{R}^{C \times 1}$ and a, b denote the attention vector for X_{rgb} and X_t , respectively. $\mathbf{V}_{rgb}^c, \mathbf{V}_t^c \in \mathbb{R}^{1 \times 1}$ and a^c, b^c are the c -th row of $\mathbf{V}_{rgb}, \mathbf{V}_t$ and a, b . $\mathbf{W}_{21}, \mathbf{W}_{22} \in \mathbb{R}^{C \times d}$ is filter for the fully connected layer operation.

Finally, the aggregated feature maps can be obtained:

$$\mathbf{F}_c = a_c \cdot X_{rgb}^c + b_c \cdot X_t^c \quad (4)$$

where $\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_C], \mathbf{F}_c \in \mathbb{R}^{H \times W}$.

D. Training Procedure

The whole network is trained in an end-to-end manner. We first initialize the parameters of the first three convolutional layers using the pre-trained model of the VGG-M network [26]. Then, we train the whole network by the Adam optimization algorithm where each domain is handled separately. The detailed settings of training are presented as follows. In each iteration, a minibatch is constructed from 8 frames which are randomly chosen in each video sequence. And we draw 32 positive and 96 negative samples from each frame and generate 256 positive and 768 negative samples together in a minibatch. Herein, the samples whose the IoU overlap ratios with the ground-truth bounding box are larger than 0.7 are treated as positive, and the negative samples have less than 0.5 IoUs. For multi-domain learning with K training sequences, we train the network with $200 \times K$ iterations, where the learning rate is 0.0001 for convolutional layers and 0.001 for fully connected layers. The weight decay is fixed to 0.0005, respectively. We train our network on all video sequences from RGBT234 dataset [29] and test it on GTOT dataset [10]. For another experiment, we train our network on all video sequences from GTOT dataset and test it on RGBT234 dataset.

In this training phase, the loss function of binary cross entropy (BCE) is used for binary classification (foreground and background) for each domain (*i.e.* each video sequence), and the BCE loss can be formulated as:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^N y_i \log p_i + (1 - y_i) \log (1 - p_i) \quad (5)$$

where N is the number of samples, and p_i is the predicted probabilistic value of the i -th sample from FANet. y_i is the

ground truth label of corresponding sample, where y_i is 1 for positive samples, otherwise 0.

In addition, the loss function of instance embedding [25] is also adopted to learn a more discriminative embedding of target objects with similar semantics. Unlike BCE loss, this instance embedding loss is for all domains and treats each domain as a category, and only positive samples of each domain are used in calculating this loss. The instance embedding loss can learn discriminative representations of the unseen target objects in new test sequences by forcing target objects in different sequences to be embedded far away from each other. It can be formulated as :

$$\mathcal{L}_{inst} = -\frac{1}{N} \sum_{i=1}^N \sum_{d=1}^D y_{i,d} \log p_{i,d} \quad (6)$$

where D is the number of domains, and $y_{i,d}$ is the ground truth label of the i -th sample of the d -th domain. $p_{i,d}$ is the predicted probabilistic value of the i -th sample of the d -th domain from FANet.

The final loss function for the optimization of our FANet is formulated as:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \cdot \mathcal{L}_{inst} \quad (7)$$

where α is a hyper-parameter that controls the balance between the two loss terms and in this paper we set it to $\alpha = 0.1$.

E. Tracker Details

In tracking, we replace the K branches of domain specific (the last fc layer) with a single branch for each test sequence. The newly added branch is trained in the first frame pair and updated in subsequent frame pairs. Given the first frame pair with the ground truth of target object, we draw 500 positive (IoU with ground truth is larger than 0.7) and 5000 negative samples (IoU with ground truth is smaller than 0.5), and train the new branch with 30 iterations, where the learning rate of the last fc layer is set to 0.001 and others are 0.0001. Given the t -th frame, we draw a set of candidates $\{\mathbf{x}_t^i\}$ from a Gaussian distribution of the previous tracking result \mathbf{x}_{t-1}^* , where the mean of Gaussian function is set to $\mathbf{x}_{t-1}^* = (x_{t-1}, y_{t-1}, s_{t-1})$ and the covariance is set as a diagonal matrix $diag\{0.09r^2, 0.09r^2, 0.25\}$. (x, y) and s indicate the location and scale respectively and r is the mean of (x_{t-1}, y_{t-1}) . For the i -th candidate \mathbf{x}_t^i , we compute its positive and negative scores using the trained network as $f^+(\mathbf{x}_t^i)$ and $f^-(\mathbf{x}_t^i)$, respectively. The target location of current frame is the candidate with the maximum positive score as:

$$\mathbf{x}_t^* = \arg \max_{\mathbf{x}_t^i} f^+(\mathbf{x}_t^i), \quad i = 1, 2, \dots, N, \quad (8)$$

where N is the number of candidates. We also apply bounding box regression technique [16] to improve target localization accuracy. We only train the regenerator in the first frame to avoid potential unreliability of other frames. If the estimated target state is reliable enough, *i.e.*, $f^+(\mathbf{x}_t^*) > 0.5$, we adjust the target locations using the regression model. During the tracking process, our network has a long-term update strategy and a short-term update strategy. When $f^+(\mathbf{x}_t^i) < 0$, the short-term update is used, and the last 20 positive samples and 20

negative samples are adapted to fine-tune network parameters. Correspondingly, a long-term update is performed at 10 frames interval, and the last 100 positive samples and 20 negative samples are adapted to fine-tune the parameters of our tracker. To reduce the computational burden and avoid over-fitting, we only update the network parameters of fully connected layers.

IV. PERFORMANCE EVALUATION

To validate the effectiveness of our quality-aware Feature Aggregation Network (FANet), we test it on two large-scale RGB-T benchmarks, including GTOT dataset [10] and RGBT234 dataset [29] and analyze the tracking performance.

A. Evaluation Setting

Datasets. We evaluate our FANet on two large-scale benchmark datasets, *i.e.*, GTOT dataset [10] and RGBT234 dataset [29].

GTOT is a standard benchmark dataset for RGBT tracking. It contains 50 RGBT video sequences under different scenes and conditions, and the target box for each frame is annotated. In addition, it is annotated with 7 attributes for analyzing the attributed-sensitive performance of RGBT tracking approaches. The RGBT234 dataset is currently the largest RGBT tracking dataset, which is extended from the rgbt210 dataset [7]. It contains 234 rgbt videos with a frame count of approximately 234,000 and a maximum video pair frame size of 8000. In order to analyze the attribute-based performance of different tracking algorithms, 12 attributes are used for annotation.

Evaluation metrics. On these two datasets, we use success rate (SR) and precision rate (PR) for quantitative performance evaluation. PR is the percentage of frames whose output location is within a threshold distance of the ground truth. SR is the percentage of frames whose overlap rate between the output boundary box and the ground truth boundary box is larger than a threshold. The threshold of PR is set to 5 and 20 pixel for GTOT and RGBT234 datasets respectively (because the target object in the GTOT dataset is generally small). And the area under the curves of success rate is adopted to represent SR for quantitative performance evaluation.

Baseline. In order to verify the validity of our RGBT tracking algorithm, we compare our method with 29 popular trackers, some of which are from GTOT and RGBT234 benchmark.

B. Evaluation on GTOT Dataset

Comparison with RGB trackers. To verify the superiority of the proposed RGBT tracking method compared to RGB trackers, we first evaluate our method with 12 state-of-the-art RGB trackers, including MDNet [16], DAT [30], ECO [15], RT-MDNet [25], SRDCF [31], BACF [32], ACFN [33], ACT [34], SiamDW [35], CFnet [36] and KCF [37], SiameseFC [38]. Fig. 6(a) demonstrates the effectiveness of introducing thermal information in visual tracking. In particular, our tracker outperforms DAT, RT-MDNet and ECO with 12.0%/11%, 14.6%/11.5% and 12.1%/9.7% in PR/SR, respectively.

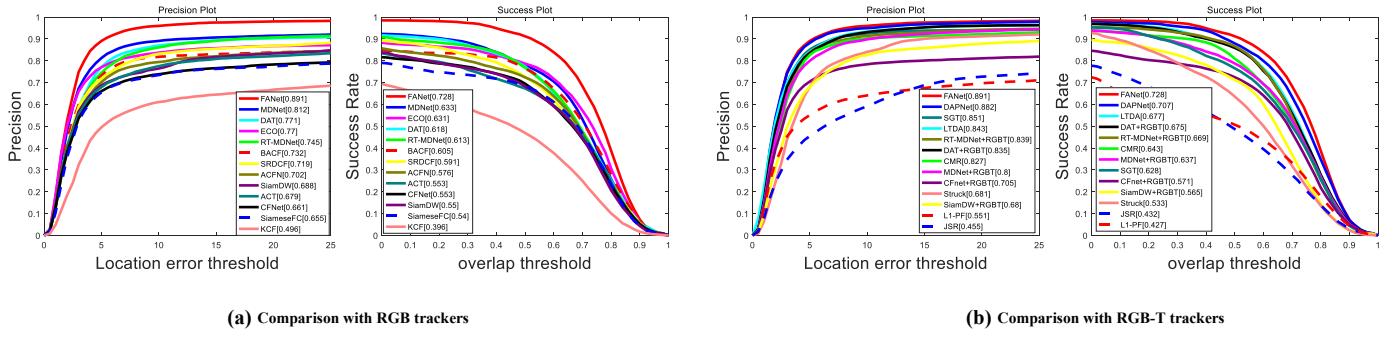


Fig. 6. The evaluation results on GTOT dataset. The representative scores of PR/SR is presented in the legend.

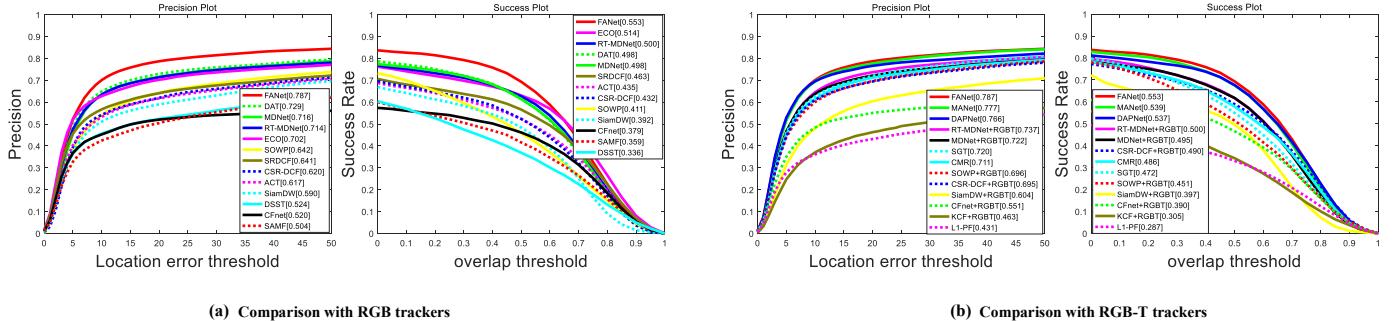


Fig. 7. The evaluation results on RGBT234 dataset. The representative scores of PR/SR is presented in the legend. We separate RGB and RGB-T trackers in (a) and (b).

TABLE I

ATTRIBUTE-BASED PR/SR SCORES (%) ON RGBT234 DATASET AGAINST WITH EIGHT RGB-T TRACKERS. THE BEST AND SECOND RESULTS ARE IN RED AND GREEN COLORS, RESPECTIVELY.

	SOWP+RGBT	CFNet+RGBT	CMR	L1-PF	CSR-DCF+RGBT	DAPNet	SGT	RT-MDNet+RGBT	FANet
NO	86.8/53.7	76.4/56.3	89.5 /61.6	56.5/37.9	82.6/60.0	90.0 /64.4	87.7/55.5	86.7/61.1	88.2 / 65.7
PO	74.7/48.4	59.7/41.7	77.7/53.5	47.5/31.4	73.7/52.2	82.1 / 57.5	77.9/51.3	81.3/55.4	86.6 / 60.3
HO	57.0/37.9	41.7/29.0	56.3/37.7	33.2/22.2	59.3/40.9	66.0 / 45.7	59.2/39.4	60.3/39.7	66.5 / 45.8
LI	72.3/46.8	52.3/36.9	74.2/49.8	40.1/26.0	69.1/47.4	77.5 / 53.0	70.5/46.2	71.3/47.4	80.3 / 54.8
LR	72.5/46.2	55.1/36.5	68.7/42.0	46.9/27.4	72.0/47.6	75.0 / 51.0	75.1 /47.6	74.6/48.8	79.5 / 53.2
TC	70.1/44.2	45.7/32.7	67.5/44.1	37.5/23.8	66.8/46.2	76.8 / 54.5	76.0/47.0	72.1/50.8	76.6 / 55.1
DEF	65.0/46.0	52.3/36.7	66.7/47.2	36.4/24.4	63.0/46.2	71.7 / 51.9	68.5/47.4	64.8/46.4	72.2 / 52.7
FM	63.7/38.7	37.6/25.0	61.3/38.2	32.0/19.6	52.9/35.8	67.0 / 44.5	67.7 /40.2	64.6/40.2	68.1 / 43.8
SV	66.4/40.4	59.8/43.3	71.0/49.3	45.5/30.6	70.7/49.9	78.0 / 54.2	69.2/43.4	73.9/50.5	78.5 / 56.3
MB	63.9/42.1	35.7/27.1	60.0/42.7	28.6/20.6	58.0/42.5	65.3 / 46.7	64.7/43.6	63.5/45.1	70.0 / 50.3
CM	65.2/43.0	41.7/31.8	62.9/44.7	31.6/22.5	61.1/44.5	66.8 / 47.4	66.7/45.2	64.6/44.7	72.4 / 52.3
BC	64.7/41.9	46.3/30.8	63.1/39.7	34.2/22.0	61.8/41.0	71.7 / 48.5	65.8/41.8	64.2/40.8	75.7 / 50.3
ALL	69.6/45.1	55.1/39.0	71.1/48.6	43.1/28.7	69.5/49.0	76.6 / 53.7	72.0/47.2	73.7/50.0	78.7 / 55.3

Comparison with RGBT trackers. On the GTOT dataset [10], we compare our approach with 16 state-of-the-art trackers, some of which are from GTOT benchmark.

Since there are few RGBT trackers [39], [40], [41], [42], [10], [7], [9], some RGBT tracking methods are derived by extending the RGB method. Specifically, we implement rgbt tracking by directly concatenate RGB and thermal features or considering thermal as an additional channel to input RGB tracking. In particular, RT-MDNet+RGBT, DAT+RGBT, SiamDW+RGBT and MDNet+RGBT are extended RGBT

tracking algorithms that directly concatenate RGB and thermal channels as input. The top 12 are DAPNet [40], DAT [30]+RGBT, CMR [9], MDNet [16]+RGBT, SGT [7], LTDA [39], Struck [43]+RGBT, RT-MDNet [25]+RGBT, CFnet [36]+RGBT and SiamDW [35]+RGBT, JSR [42] and L1-PF [44]). Fig. 6(b) shows that our tracker is superior to other state-of-the-art trackers on the GTOT dataset. Specifically, our method improves the CMR, RT-MDNet+RGBT and SGT by a large margin, *i.e.*, 6.4%/8.5%, 5.2%/5.9% and 4.0%/10.0% in PR/SR. In addition, our method is slightly bet-

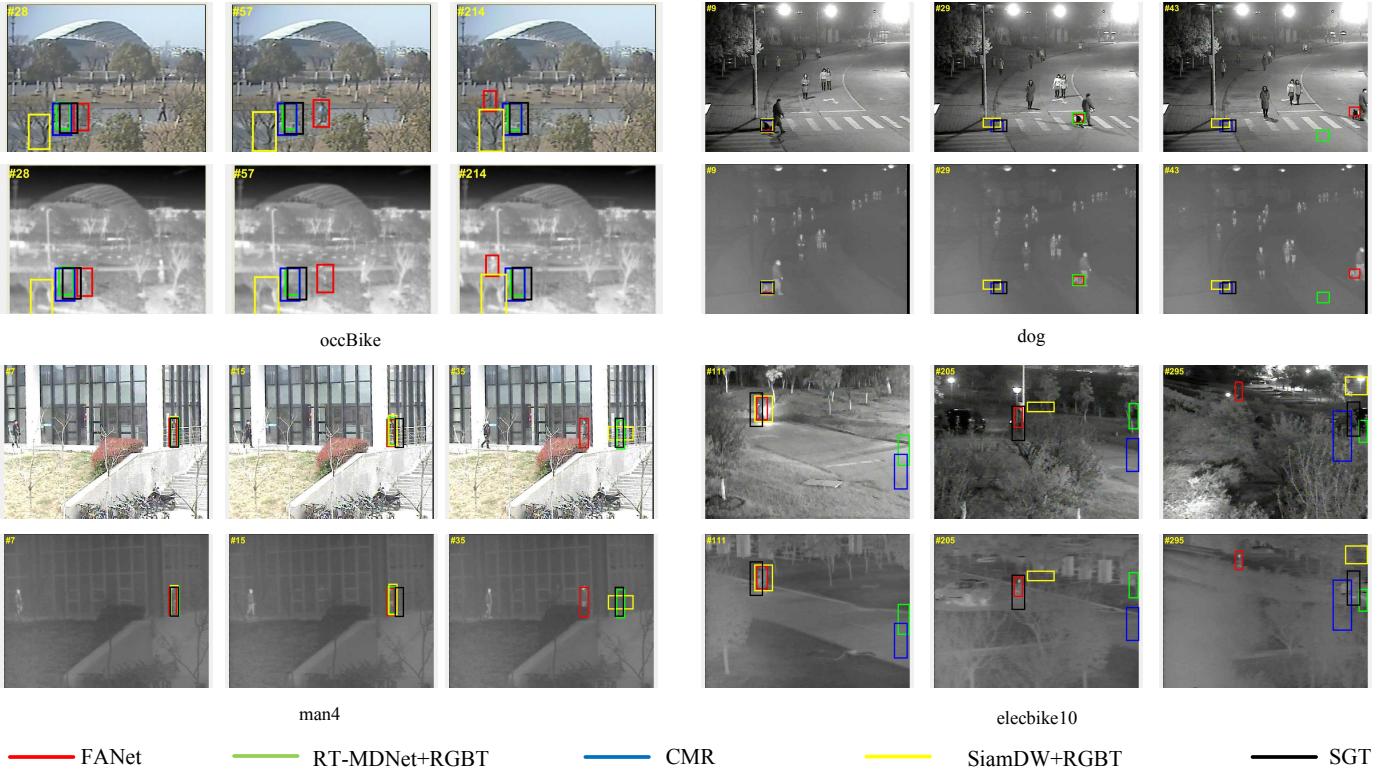


Fig. 8. Qualitative comparison of our FANet versus four state-of-the-art RGB-T trackers on four video sequences.

ter than DAPNet [40] 0.9%/2.1% in PR/SR, and our algorithm runs 6 times faster. The overall promising performance of our method can be explained by the fact that FANet makes fully use of hierarchical deep features and RGB-T information to well handle the challenges of significant appearance changes and adverse environmental conditions.

C. Evaluation on RGBT234 Dataset

Overall performance. For more comprehensive evaluation, we report the evaluation results on the RGBT234 dataset [29], as shown in Fig. 7. The comparison trackers include 12 RGB ones, i.e., (DAT [30], RT-MDNet [25], ACT [34] SiamDW [35], ECO [15], MDNet [16], SOWP [45], SRDCF [31], CSR-DCF [46], DSST [47], CFnet [36] and SAMF [48]), and 12 RGBT ones, i.e., (MANet [49], MDNet+RGBT, CMR [9], SGT [7], SOWP+RGBT, CSR-DCF+RGBT, MEEM [50]+RGBT, CFnet+RGBT, KCF [37]+RGBT and L1-PF [44]). From the results, we can see that the performance of our FANet clearly better than the advanced RGB and RGBT methods in all metrics. It demonstrate the importance of thermal information and feature aggregations proposed in our method. In particular, our FANet achieves 5.8%/5.5% performance gains in PR/SR over the second best RGB tracker DAT, achieves 1.0%/1.4% gains over the second best RGBT tracker MANet, and our algorithm runs 15 times faster than MANet.

Attribute-based performance. We also report the attribute-based results of our FANet versus other state-of-the-art RGB-T trackers, including L1-PF, SOWP+RGBT, CSR-DCF+RGBT,

CFNet+RGBT, RT-MDNet+RGBT, SGT, CMR and DAPNet, as shown in Table I. The attributes include no occlusion (NO), partial occlusion (PO), heavy occlusion (HO), low illumination (LI), low resolution (LR), thermal crossover (TC), deformation (DEF), fast motion (FM), scale variation (SV), motion blur (MB), camera moving (CM) and background clutter (BC). The results show that our method performs the best in terms of most challenges except for NO, TC and FM. It demonstrates the effectiveness of our FANet in handling the sequences with the appearance changes and adverse conditions. The following major observations and conclusions can be drawn from Table I.

First, although CMR and DAPNet perform well when no occlusion occurs in PR, their performance drop a lot when partial or heavy occlusions happen. Our FANet keeps high tracking performance with partial or heavy occlusions, which suggests that the aggregated hierarchical features can improve the tracking robustness in presence of occlusions. Second, in adverse lighting conditions and low resolution, FANet outperforms all other trackers. It can be explained that adaptive incorporation from RGB and thermal information can boost tracking performance significantly (*e.g.* FANet versus RT-MDNet+RGBT). And our strategy to predict the modality weights is more robust than other existing methods (*e.g.*, FANet versus SGT). Third, our framework is robust to the significant appearance changes of target object and the effects of cluttered background by observing its performance under deformation, scale variation, camera moving and background clutter. Finally, unsatisfying results are usually generated in fast motion video sequences by our FANet. It might be that the rapid movement of the target object is beyond the scope

TABLE II

PR/SR SCORES (%) OF DIFFERENT VARIANTS INDUCED FROM OUR NETWORK ON GTOT AND RGBT234 DATASETS.

		RT-MDNet+RGBT	FANet-FA	FANet-MA	FANet
GTOT	PR	83.9	88.5	88.7	89.1
	SR	66.9	72.7	72.6	72.8
RGBT234	PR	73.7	77.0	76.7	78.7
	SR	50.0	53.0	54.0	55.3

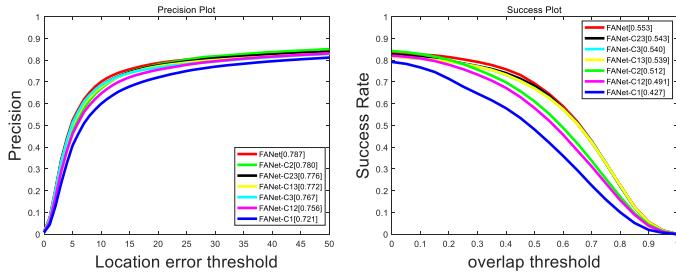


Fig. 9. Performance evaluation using different convolutional layers on RGBT234 dataset.

of the searching windows.

Qualitative performance. The qualitative comparison of our FANet versus 4 state-of-the-art RGB-T trackers on 4 video sequences is presented in Fig. 8, including RT-MDNet [25]+RGBT, SGT [7], CMR [9] and SiamDW [35]+RGBT. Overall, our method is effective in handling the challenges of occlusion, background clutter, appearance variation, low resolution and low illumination variation. For example, in Fig. 8 (occBike), our method performs well in presence of partial and heavy occlusions and background clutter while other trackers lose the target when occlusion happens. In Fig. 8 (dog), the target object moves so fast that most algorithms can't track it, and only our method can do it. In Fig. 8 (man4), we can see that our method has a relatively good treatment for heavy occlusion. In Fig. 8 (elecbike10), the *elecbike10* is partly invisible in visible source but the thermal images can provide relatively reliable information to distinguish the target object from the background. Moreover, in the case of low resolution and occlusions, all compared methods lose the targets, while our FANet method is able to track it robustly by adaptively incorporating useful information from these two modalities. From the above qualitative comparison, we can find that our algorithm is more capable of dealing with complex challenges in real scenes.

D. Analysis of Our Network

Ablation study. In order to prove the importance of the main components, we apply two special versions of our method for comparative analysis and evaluation them on both GTOT and RGBT234 datasets. These two variants are: 1) FANet-FA, that only uses the layer weights and removes the adaptive aggregation subnetwork; and 2) FANet-MA, that only uses the modality weights to integrate the feature of Conv3 of two modalities and removes the hierarchical feature aggregation

module. Table II presents the comparison results. The performance of our FANet-FA, FANet-MA are obviously better than baseline(RT-MDNet+RGBT), which proves the validity of our sub-networks. In addition, the performance of our FANet over FANet-FA and FANet-MA demonstrates the effectiveness of the proposed entire network.

Aggregation Analysis. To further verify the validity of our aggregation method, we extend RT-MDNet algorithm for RGBT tracking of dual-modal inputs, and have three aggregate forms. The first one is that we directly concatenate two modalities of data channels to form 6 channels of input data, then input the original RT-MDNet algorithm for tracking, which is named RT+RGBT-EARLY (equivalent to RT-MDNet+RGBT). The second one is that we concatenate the feature maps of the modalities at Conv1, and named RT+RGBT-MID. The last one is that we extract the convolution features of the two modalities separately, and concatenate the feature maps of Conv3, named RT+RGBT-LATE. Table III shows the results compared with the baseline on GTOT and RGBT234 datasets. From the Table III, we can see that our method that aggregates two modalities is obviously superior than RT-MDNet, which proves that the fusion of two modal information is beneficial to visual tracking. In addition, on GTOT and RGBT234 datasets, the FANet outperforms the best baseline method with 4.5%/4.5%, 5.0%/5.3% in PR/SR, respectively. It shows that our FANet is very effective for representing and fusing different modalities in visual tracking.

Feature analysis. To analyze the effectiveness of hierarchical feature aggregation in the proposed network, we compare the tracking results using deep features from different convolutional layers on the RGBT234 dataset. We first use each individual convolutional layer (C1, C2 and C3) to represent objects in the network. The tracking performance is better when a deeper layer is used, which can be attributed to the discriminative semantic abstraction from deeper layers. Moreover, we also use combinations of C1, C2 and C3 (C12, C13 and C23) to represent objects in the proposed network. Fig. 9 shows the evaluation results on the RGBT234 dataset. The superior performance of our FANet over others demonstrates the effectiveness of the used hierarchical features. From PR/SR scores, we can find that FANet-C13 and FANet-C23 methods perform better than FANet-C1, FANet-C2 and FANet-C3. The results show that using multi-layer network features simultaneously can improve tracking performance.

Runtime analysis. Finally, we present the runtime of our FANet against the state-of-the-art trackers, MDNet [16]+RGBT, MANet [49], DAPNet [40], CMR [9], SGT [7] with their tracking performance on the RGBT234 dataset in Table IV. Our implementation is on the platform of PyTorch0.41 with 2.1 GHz Intel(R) Xeon(R) CPU E5-2620 and NVIDIA GeForce GTX 2080Ti GPU, and the average tracking speed is 19 FPS. Overall, the results demonstrate that our framework clearly outperforms other tracking methods.

Failure cases We show some cases of tracking failure by FANet in Fig. 10. The *car3* sequence in Fig. 10(a), when the object scale variation too much, the proposed method fails to

TABLE III
PR(%) AND SR(%) SCORES OF OUR ALGORITHM ON GTOT AND RGBT234 COMPARING WITH DIFFERENT FUSION STAGES.

		RT-MDNet	RT+RGBT-EARLY	RT+RGBT-MID	RT+RGBT-LATE	FANet
GTOT	PR	74.5	84.6	79.4	82.9	89.1
	SR	61.3	68.3	67.2	66.7	72.8
RGBT234	PR	71.4	73.7	72.2	73.4	78.7
	SR	50.0	50.0	49.3	48.3	55.3

TABLE IV
PERFORMANCE AND RUNTIME OF OUR FANET AGAINST THE BASELINE METHODS ON RGBT234 DATASETS.

		MDNet+RGBT	MANet	DAPNet	CMR	SGT	FANet
RGBT234	PR	72.2	77.7	76.6	71.1	72.0	78.7
	SR	49.5	53.9	53.7	48.6	47.2	55.3
	FPS	3	1.5	2	8	5	19

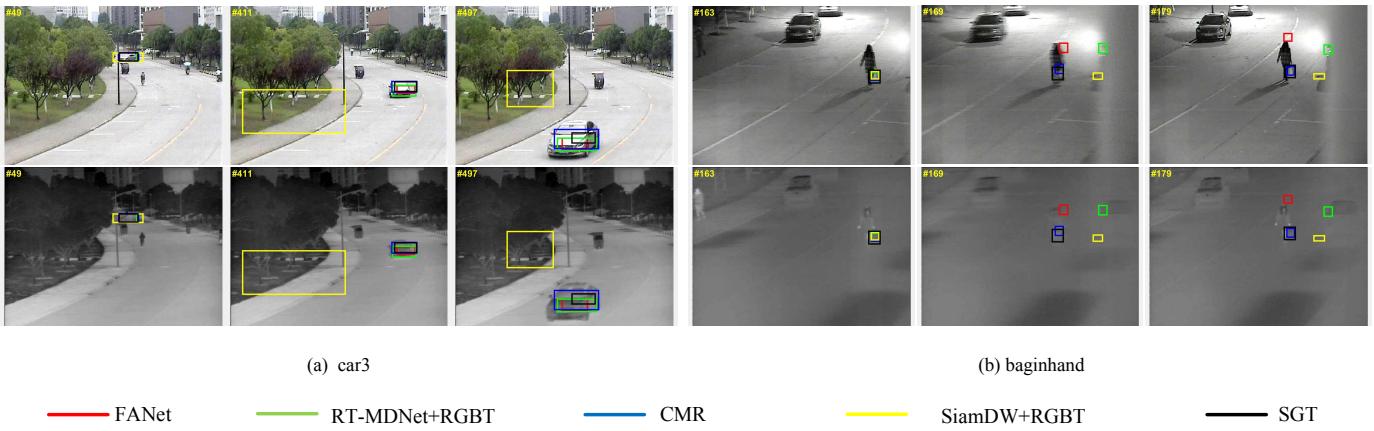


Fig. 10. Failure cases on the car3 and baginhand (RGBT234).

located the target. The reason for the failure might be that our tracker only learns a linear scale regression estimation parameter in the first frame, which is difficult to cope with large-scale variations. The *baginhand* sequence in Fig. 10(b), the sudden camera shake can cause our tracker to not successfully track the target object. The sudden large shake of cameras can cause our tracker to not successfully track the target object. There are two possible reasons. 1) Sudden camera shake will make the target exceed our limited search range. 2) Violent camera shake will blur the image, and blurred appearance and sudden changes would make our classifier difficult to distinguish the target object from the background.

V. CONCLUSION

In this paper, we propose an end-to-end trained FANet for robust RGB-T tracking. FANet aggregates hierarchical multi-resolution feature maps to handle significant appearance changes caused by deformation, low illumination, background clutter and occlusion within each modality, and predict the reliability degrees to address the problem of uncertainties in different modalities. Extensive experiments on large-scale benchmark datasets show that the FANet gains excellent tracking result. In future work, we will develop more efficient

and effective scale handling methods, such as IOUNet [51], into our framework for powerful RGBT tracking.

REFERENCES

- [1] C. Li, X. Wang, L. Zhang, J. Tang, H. Wu, and L. Lin, “Weld: Weighted low-rank decomposition for robust grayscale-thermal foreground detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 725–738, 2017.
- [2] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, “Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation,” *arXiv*, 2019.
- [3] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, “Rgb-infrared cross-modality person re-identification,” in *Proceedings of IEEE International Conference on Computer Vision*, 2017.
- [4] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, “Multispectral pedestrian detection: Benchmark dataset and baseline,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [5] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, “Learning cross-modal deep representations for robust pedestrian detection,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [6] R. Gade and T. B. Moeslund, “Thermal cameras and applications: a survey,” *Machine vision and applications*, vol. 25, no. 1, pp. 245–262, 2014.
- [7] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, “Weighted sparse representation regularized graph learning for rgb-t object tracking,” in *Proceedings of ACM International Conference on Multimedia*, 2017.
- [8] C. Li, X. Wu, N. Zhao, X. Cao, and J. Tang, “Fusing two-stream convolutional neural networks for rgb-t object tracking,” *Neurocomputing*, vol. 281, pp. 78–85, 2018.

- [9] L. Chenglong, Z. Chengli, H. Yan, T. Jin, and W. Liang, "Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking," in *Proceedings of European Conference on Computer Vision*, 2018.
- [10] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5743–5756, 2016.
- [11] X. Lan, M. Ye, S. Zhang, and P. C. Yuen, "Robust collaborative discriminative learning for rgb-infrared tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.
- [12] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2015.
- [13] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, "Hedged deep tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [14] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proceedings of European Conference on Computer Vision*, 2016.
- [15] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "Eco: Efficient convolution operators for tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [16] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [17] J. Wang, Z. Wang, D. Tao, S. See, and G. Wang, "Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks," in *Proceedings of European Conference on Computer Vision*, 2016.
- [18] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian Conference on Computer Vision*, 2016.
- [19] S.-J. Park, K.-S. Hong, and S. Lee, "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4980–4989.
- [20] C. Li, X. Sun, X. Wang, L. Zhang, and J. Tang, "Grayscale-thermal object tracking via multi-task laplacian sparse representation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 4, pp. 673–681, 2017.
- [21] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [22] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [23] B. Han, J. Sim, and H. Adam, "Branchout: Regularization for online ensemble tracking with convolutional neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [24] E. Park and A. C. Berg, "Meta-tracker: Fast and robust online adaptation for visual object trackers," in *Proceedings of European Conference on Computer Vision*, 2018.
- [25] I. Jung, J. Son, M. Baek, and B. Han, "Real-time mdnet," in *Proceedings of European Conference on Computer Vision*, 2018.
- [26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of International Conference on Learning Representations*, 2015.
- [27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [28] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [29] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "Rgb-t object tracking: benchmark and baseline," *Pattern Recognition*, vol. 96, p. 106977, 2019.
- [30] S. Pu, Y. Song, C. Ma, H. Zhang, and M. H. Yang, "Deep attentive tracking via reciprocative learning," in *Advances in Neural Information Processing Systems*, 2018.
- [31] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.
- [32] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2017.
- [33] J. Choi, H. J. Chang, S. Yun, T. Fischer, and Y. C. Jin, "Attentional correlation filter network for adaptive visual tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [34] B. Chen, D. Wang, P. Li, S. Wang, and H. Lu, "Real-time actor-critic tracking," in *Proceedings of European Conference on Computer Vision*, 2018.
- [35] Z. Zhipeng and P. Houwen, "Deeper and wider siamese networks for real-time visual tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [36] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [37] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.
- [38] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proceedings of European Conference on Computer Vision*, 2016.
- [39] Y. Rui, Z. Yabin, W. Xiao, L. Chenglong, and J. Tang, "Learning target-oriented dual attention for robust rgb-t tracking," in *Proceedings of IEEE International Conference on Image Processing*, 2019.
- [40] Y. Zhu, C. Li, B. Luo, J. Tang, and X. Wang, "Dense feature aggregation and pruning for rgbt tracking," in *Proceedings of ACM International Conference on Multimedia*, 2019.
- [41] C. Li, S. Hu, S. Gao, and J. Tang, "Real-time grayscale-thermal tracking via laplacian sparse representation," in *Proceedings of International Conference on Multimedia Modeling*, 2016.
- [42] H. Liu and F. Sun, "Fusion tracking in color and infrared images using joint sparse representation," *Information Sciences*, vol. 55, no. 3, pp. 590–599, 2012.
- [43] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, and P. H. Torr, "Struck: Structured output tracking with kernels," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2096–2109, 2015.
- [44] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, "Multiple source data fusion via sparse representation for robust visual tracking," in *Proceedings of International Conference on Information Fusion*, 2011.
- [45] H.-U. Kim, D.-Y. Lee, J.-Y. Sim, and C.-S. Kim, "Sowp: Spatially ordered and weighted patch descriptor for visual tracking," in *Proceedings of IEEE International Conference on Computer Vision*, 2015.
- [46] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [47] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proceedings of British Machine Vision Conference*, 2014.
- [48] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proceedings of European Conference on Computer Vision*, 2014, pp. 254–265.
- [49] C. Li, A. Lu, Z. Zheng, Z. Tu, and J. Tang, "Multi-adapter rgbt tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.
- [50] J. Zhang, S. Ma, and S. Sclaroff, "Meem: robust tracking via multiple experts using entropy minimization," in *Proceedings of European Conference on Computer Vision*, 2014.
- [51] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proceedings of the European Conference on Computer Vision*, 2018.