FANet: Quality-Aware Feature Aggregation Network for RGB-T Tracking

Yabin Zhu¹, Chenglong Li¹, Yijuan Lu², Liang Lin³, Bin Luo¹ and Jin Tang¹

¹School of Computer Science and Technology, Anhui University, Hefei, 230601, China
²Department of Computer Science, Texas State University, San Marcos, TX, 78666, USA
³School of Data and Computer Science, Sun Yat-sen University, Guangzhou, 510006, China,
{zhuyabin0726, lcl1314}@foxmail.com, lu@txstate.edu, linliang@ieee.org, {luobin, tj}@ahu.edu.cn

Abstract

This paper investigates how to perform robust visual tracking in adverse and challenging conditions using complementary visual and thermal infrared data (RGB-T tracking). We propose a novel deep network architecture "quality-aware Feature Aggregation Network (FANet)" to achieve quality-aware aggregations of both hierarchical features and multimodal information for robust online RGB-T tracking. Unlike existing works that directly concatenate hierarchical deep features, our FANet learns the layer weights to adaptively aggregate them to handle the challenge of significant appearance changes caused by deformation, abrupt motion, background clutter and occlusion within each modality. Moreover, we employ the operations of max pooling, interpolation upsampling and convolution to transform these hierarchical and multi-resolution features into a uniform space at the same resolution for more effective feature aggregation. In different modalities, we elaborately design a multimodal aggregation sub-network to integrate all modalities collaboratively based on the predicted reliability degrees. Extensive experiments on large-scale benchmark datasets demonstrate that our FANet significantly outperforms other state-of-the-art RGB-T tracking methods.

1. Introduction

Thermal cameras recently become more economically affordable and have been applied to many computer vision tasks, such as object segmentation [19], person ReID [33] and pedestrian detection [11, 35] etc. Thermal infrared modality is insensitive to lighting condition and has a strong ability to penetrate haze and smog [6]. Therefore, it has big potential in object tracking owing to its complementary benefits to traditional visual tracking. As a sub-branch of visual tracking, RGB-T object tracking aims to integrate complementary data from visual and thermal infrared spectrums to estimate the state of a specific instance in a video, given the ground truth bounding box in the first frame. RGB-T object tracking receives many attentions in recent years [16, 21, 15, 22], although it is still a new research line in visual tracking.

Recent studies on RGB-T tracking mainly focus on two aspects. One aims to learn robust feature representation via the usage of multimodal data [21, 20, 22] and achieves promising tracking performance. These works rely on either handcrafted features or highly semantic deep features only to localize objects. Handcrafted features are too weak to handle the challenges of significant appearance changes caused by deformation, abrupt motion, background clutter
and occlusion within each modality. Although highly semantic features are effective to distinguish target semantics (as shown in Fig. 1), they are unable to meet the objective of visual tracking, which is to locate targets precisely rather than to infer their semantic categories. Therefore, relying on handcrafted features or highly semantic features only is difficult to handle the challenges in RGB-T tracking and therefore limits potentials of RGB-T tracking.

The other one focuses on introducing modality weight to achieve adaptive fusion of different source data [16, 15]. Early works employed reconstruction residues [16, 21] to regularize modality weight learning. Lan et al. [15] used max-margin principle to optimize the modality weights according to classification scores. However, these works would fail when the reconstruction residues or classification scores are unreliable to reflect modal reliability. For example, for the BlurCar sequence in GTOT dataset [16], the quality of RGB modality is much better than thermal modality. But the reconstruction residue on RGB modality computed by the model in [16] 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 RGB-T tracking. Within each modality, our FANet first aggregates hierarchical multi-resolution feature maps into a uniform space at the same resolution. Shadow features encode fine-grained spatial details such as target positions and help achieve precise target localizations, while deep features are more effective to capture target semantics, as shown in Fig. 1. To utilize both advantages of these features, we employ the operations of max pooling and interpolation upsampling to convert their resolutions into the same one and then add convolution operations to transform them into a uniform feature space. In addition, features of different layers should contribute unequally to some certain video sequences [25]. Therefore, unlike manually setting the weights of different features in existing works [25], our FANet learns the layer weights intelligently to adaptively integrate them to highlight more discriminative features and suppress noisy ones, and thus achieve improved tracking performance.

In different modalities, we elaborately design a multimodal aggregation sub-network to integrate all modalities collaboratively according to the predicted reliability degrees. Many efforts are devoted to calculate the modality weights that reflect the reliability degrees to make adaptive fusion of different modalities, and thus achieve significant improvement for tracking performance [16, 15]. Fig. 1 also demonstrates this point. In this work, we propose a more effective method to compute the modality weights. Given the aggregated features, we use the network to predict the modality weight for each modality and then combine them with the corresponding aggregated features to yield a robust target representation.

The object tracking is performed via binary classification in the multi-domain network [26]. Extensive experiments on large-scale benchmark datasets demonstrate that our FANet outperforms other state-of-the-art tracking methods with a clear margin. To our best knowledge, this is the first work to learn both layer and modality weights for the fusion of hierarchical multi-resolution deep features and multimodal information for RGB-T object tracking in an end-to-end trained deep learning framework. We summarize our major contributions of our work as follows.

- We propose a novel deep network architecture that is end-to-end trained for RGB-T tracking. The proposed network consists of two sub-networks: hierarchical feature aggregation sub-network and multimodal information aggregation sub-network. It is able to handle challenges of object tracking in significant appearance changes and adverse environmental conditions. The advantages of our network architecture over existing ones can be seen in Fig. 2. Extensive experiments show that the proposed method outperforms other state-of-the-art trackers on large-scale RGB-T tracking datasets. Source codes and experimental results would be available online for reproducible research.

- We present a new feature aggregation method to integrate hierarchical and multi-resolution deep features in an adaptive way. In particular, the operations of max pooling, interpolation upsampling and convolution are employed to transform deep features into a uniform space at the same resolution. And the layer weights are introduced for the adaptive fusion of deep features in different layers.

- To compute the reliability degree of each modality robustly, we design a multimodal aggregation sub-network that uses the aggregated features to predict modality weights. The modality weights are combined with the aggregated features to improve the performance of RGB-T tracking significantly.

2. Related Work

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

RGB-T tracking receives more and more attentions in computer vision community with the popularity of thermal infrared sensors [19, 33, 11, 35, 22]. One research stream is to introduce modality weights to achieve adaptive fusion of different modalities [18, 16, 15]. Early works employed reconstruction residues [18, 16] to regularize the modality weight learning and carried out object tracking in Bayesian filtering framework. Lan et al. [15] used the max-margin principle to optimize the modality weights according to classification scores. However, these methods would fail when the reconstruction residues or classification scores are unreliable to reflect the modal reliability.

The other research stream is to learn robust feature representations via the usage of multimodal data [21, 20, 22]. Li et al. [21] proposed weighted sparse representation regularized graph learning approach to construct a graph-based multimodal descriptor, and adopted structured SVM for tracking. And Li et al. [22] further considered heterogeneous property between different modalities and noise effects of initial seeds in the cross-modal ranking model. To adaptively fuse different modalities while avoiding redundant noises, Li et al. [20] designed a FusionNet to select most discriminative feature maps from the outputs of ConvNet. These methods rely on either handcrafted features or highly semantic deep features only to localize objects, and thus are difficult to handle the challenges of significant appearance changes caused by deformation, abrupt motion, background clutter and occlusion within each modality.

2.2. Hierarchical Features for Tracking

In recent literatures, several works started incorporating hierarchical deep features for the task of visual tracking [25, 29, 14, 37, 5, 1]. Ma et al. [25] interpreted hierarchical features of convolutional layers as a nonlinear counterpart of an image pyramid representation and explicitly exploited these multiple levels of abstraction to represent target objects. Qi et al. [29] took full advantages of features from different CNN layers and used an adaptive Hedge method to hedge several CNN trackers into a stronger one. To take the interdependencies among different features into account, Zhang et al. [37] presented a multi-task correlation filter to learn the correlation filters jointly. A combination of handcrafted low-level and hierarchical deep features was proposed by Danelljan et al. [5, 1] by employing an implicit interpolation model to pose the learning problem in the continuous spatial domain, which enabled efficient integration of multi-resolution feature maps. These methods, however, are based on the assumption that multiple features contribute equally or set the feature weights manually, and therefore are unable to make best use of these features according to their reliabilities.

2.3. Multi-Domain Networks for Tracking

MDNet [26] achieved the state-of-the-art performance on multiple datasets by dealing with label conflict issue across videos through multi-domain learning. Han et al. [7] proposed to select a subset of branches in the CNN randomly for online learning whenever target appearance models need to be updated for better regularization, where each branch may have a different number of layers to maintain variable abstraction levels of target appearances. The meta learning was introduced in MDNet to adjust the initial deep networks [27], which could quickly adapt to robustly model a particular target in future frames. Jung et al. [12] proposed a real-time MDNet, where an improved RoIAlign technique was employed to extract more accurate representations for targets. All these methods were developed for single-modality tracking and we study them in the task of multi-modality tracking in this work.

2.4. Multimodal Fusion Network Architectures

Early fusion work simply concatenated input multimodal channels and then adopted the convolutional networks (CNN) to extract feature representations (i.e. Fig. 2 (a)). Some later fusion work additionally reported the result of fusing multiple feature representations extracted from the CNN of each modality [20] (i.e. Fig. 2 (b)). Wang et
al. [32] proposed a structure for deconvolution of multiple modalities, in which an additional feature transformation network was introduced to correlate two modalities by discovering common and modality specific features (i.e. Fig. 2 (c)). It didn’t exploit any informative intermediate feature of both modalities, and the training procedures were not end-to-end. Hazirbas et al. [9] proposed to exploit intermediate modal features (i.e. Fig. 2 (d)), which didn’t fully exploit effective mid-level multimodal features as they simply summed intermediate multimodal features. Park et al. [28] proposed a network that effectively exploited multi-level features simultaneously (i.e. Fig. 2 (e)). However, it neglected the reliability degrees of different layers and 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)).

3. FAkNet Framework

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

3.1. Network Architecture

The overall network architecture of our FAkNet is shown in Fig. 3. FAkNet mainly consists of two parts: a feature aggregation sub-network and a multimodal aggregation sub-network. Followed by these two sub-networks, we add a binary classification layer with four fully connected layers and softmax cross-entropy loss to carry out tracking task. The first three fully connected layers (with 1024, 512, 512 output units respectively (see Fig. 3)) are combined with ReLUs and dropouts. The fourth fully connected layer with two output units include \( K \) branches. Each branch corresponds to a specific domain that is a video sequence for the tracking task. The last layer is a softmax with the cross-entropy loss to perform binary classification to distinguish the target and the background in each domain. More details are referred to [26] and we make an emphasis on the details of the proposed sub-networks.

**Feature aggregation sub-network.** As shown in Fig. 3, we extract hierarchical features (Conv1-Conv3) using VGG-M Network [30] pre-trained on the ImageNet dataset for informative target representations. Features in different layers have different resolutions due to max pooling operation. Therefore, to convert these features to the same resolution, we employ max pooling operation to subsample high-resolution features and use bilinear interpolation operation to upsample low resolution features. Furthermore, to transform hierarchical features into a uniform space, we add a convolution operation for each layer and adopt a local response normalization (LRN) layer to normalize these feature maps to balance the influences of these features. Features of different layers should make different contributions to some certain video sequences as shadow features preserve the fine-grained details while deep features capture target semantics. Ma et al. [25] manually set different weights to different layers to integrate hierarchical features while Qi et al. [29] took full advantages of features from different CNN layers and used an adaptive Hedge method to hedge several CNN trackers into a stronger one. Therefore, we learn the layer weights to adaptively integrate hierarchical features to highlight more discriminative features and suppress noisy ones, and thus clearly improve tracking performance. In specific, features of each layer pass through two fully connected layers and then we concatenate the outputs to the softmax classifier to compute the weights that reflect the reliabilities of different features. Fig. 3 shows the details of the sub-network configuration and parameters.
And we summarize the operations of the feature aggregation sub-network for one modality in the following equations:

\[ F^1 = fc^{12}(f^{11}(F^{11})) \]
\[ F^2 = fc^{22}(fc^{21}(F^{12})) \]
\[ F^3 = fc^{32}(fc^{31}(F^{13})) \]
\[ F^{\text{concat}} = F^1 \oplus F^2 \oplus F^3 \]
\[ w^t = \text{softmax}(F^{\text{concat}}) \]
\[ F = w^t \otimes F^{\text{concat}} \]

where \( \oplus \) and \( \otimes \) denote the operations of concatenation and weighting, respectively. \( \text{softmax}(\cdot) \) is the softmax function, and \( fc(\cdot) \) is the fully connected module which includes dropout, fully connected and ReLU operations. \( w^t \) indicates the layer weight, and \( F^t \) denotes the transformed features of different layers. \( F \) denotes the features after aggregation.

**Multimodal aggregation sub-network.** Due to different imaging properties of RGB and thermal sensors, their reliability degrees, which contribute to the tracking performance, should be different. Several works attempted to introduce modality weights to achieve adaptive fusion of different source data [16, 21, 15]. They employed reconstruction residues [16, 21] or classification scores [15] to regularize modality weight learning, and thus would fail when the reconstruction residues or classification scores are unreliable to reflect modal reliabilities.

In this work, we elaborately design a multimodal aggregation sub-network to predict reliability degrees and then integrate all modalities collaboratively. Given the aggregated features, the sub-network predicts the quality weight for each modality and then combines them with the corresponding aggregated features to yield a robust target representation. In specific, the sub-network consists of two convolutional layers, a ReLU activation function, a LRN layer, and a sigmoid function. Fig. 3 shows the details of the sub-network configuration and parameters. And we summarize the operations of the multimodal aggregation sub-network in the following equations:

\[ F^1_R = LRN(ReLU(C^1_R \ast F_R + b^1_R)) \]
\[ w^t_R = \sigma(C^2_R \ast F^1_R + b^2_R) \]
\[ F^1_T = LRN(ReLU(C^1_T \ast F_T + b^1_T)) \]
\[ w^t_T = \sigma(C^2_T \ast F^1_T + b^2_T) \]
\[ F_{\text{final}} = (F_R \otimes w^T_R) \oplus (F_T \otimes w^T_T) \]

where \( \ast \) denotes the operation of convolution, and \( \sigma(\cdot) \) is the sigmoid function. \( R \) and \( T \) denote the RGB and thermal modalities, respectively. \( C \) and \( b \) denote the convolutional kernels and their biases, respectively. \( w^m \) indicates the modality weight. \( F_{\text{final}} \) denotes the input feature of multimodal aggregation sub-network.

### 3.2. 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 [30]. Then, we train the whole network by the Stochastic Gradient Descent (SGD) algorithm where each domain is handled separately. The detailed settings of training are presented as follows. In iterations, 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. 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 100 \( \times \) \( K \) iterations, where the learning rate is 0.0001 for convolutional layers and 0.001 for fully connected layers. The weight decay and momentum are fixed to 0.0005 and 0.9, respectively. We train our network using 77 video sequences randomly selected from RGBT234 dataset [17] and test it on GTOT dataset [16]. For another experiment, we train our network on all video sequences from GTOT dataset and test it on RGBT234 dataset.

### 3.3. Tracker Details

In tracking, the \( K \) branches of domain-specific layers (the last fc layer) are replaced 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 \( \{ x_t^i \} \) from a Gaussian distribution of the previous tracking result \( x_{t-1}^i \), where the mean of Gaussian function is set to \( x_{t-1}^i = (x_{t-1}, y_{t-1}, s_{t-1}) \) and the covariance is set as a diagonal matrix \( \text{diag}\{0.09r^2, 0.09r^2, 0.25\} \). \( x, y \) 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 \( x_t^i \), we compute its positive and negative scores using the trained network as \( f^+(x_t^i) \) and \( f^-(x_t^i) \), respectively. The target location of current frame is the candidate with the maximum positive score as:

\[ x_t^* = \arg \max_{x_t^i} f^+(x_t^i), \quad i = 1, 2, ..., N, \quad (3) \]

where \( N \) is the number of candidates. We also apply bounding box regression technique [26] to improve target local-
4. Performance Evaluation

To validate the effectiveness of our quality-aware Feature Aggregation Network (FANet), we test it on two large-scale RGB-T benchmarks: GTOT dataset [16] and RGBT234 dataset [17] and analyze its performance.

4.1. Evaluation Setting

Datasets. We evaluate our FANet on two large-scale benchmarks: GTOT dataset [16] and RGBT234 dataset [17]. GTOT dataset is a recently proposed standard benchmark dataset for RGB-T tracking. It has 50 RGB-T video clips with ground truth object locations under different scenarios and conditions. It is annotated with seven attributes, and thus partitioned into seven subsets for analyzing the attributed-sensitive performance of RGB-T tracking approaches. RGBT234 dataset is a large-scale RGB-T tracking dataset extended from RGBT210 dataset [21]. It contains 234 RGB-T videos and each video has a RGB video and a thermal video. Its total number of frames reach about 234,000 and the number frames of the longest video pair reaches 8,000. To analyze the attribute-based performance of different tracking algorithms, it is annotated with 12 attributes.

Evaluation metrics. On these two datasets, we use precision rate (PR) and success rate (SR) 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 the frames whose overlap ratio between the output bounding box and the ground truth bounding box is larger than a threshold. We set the threshold to be 5 and 20 pixels for GTOT and RGBT234 datasets respectively (since the target object in the GTOT dataset is generally small) to obtain the representative PR. And we employ the area under the curves of success rate as the representative SR for quantitative performance evaluation.

4.2. Evaluation on GTOT Dataset

On the GTOT dataset [16], we compare our approach with 16 state-of-the-art trackers. The top seven are MDNet [26]+RGBT, SGT [21], CSR [16], Struck [8]+RGBT, CN [4]+RGBT, SCM [38]+RGBT and KCF [10]+RGBT, where we concatenate features used in trackers from RGB and thermal modalities as RGB-T input of corresponding tracking algorithms [16]. Fig. 4 shows our method performs better than other state-of-the-art trackers on the GTOT dataset. Specifically, our method improves the MDNet+RGBT and SGT by a large margin, i.e., 8.5%/6.1% and 3.4%/7.0% in PR/SR. The overall promising performance of our method can be explained by the fact that the proposed FANet makes fully use of hierarchical deep features and multimodal information to well handle the challenges of significant appearance changes and adverse environmental conditions.

4.3. Evaluation on RGBT234 Dataset

Overall performance. For more comprehensive evaluation, we report the evaluation results on the RGBT234 dataset [17], as shown in Fig. 5. The comparison trackers include seven RGB ones (MDNet [26], SOWP [13], SRDCF [3], CSR-DCF [24], DSST [2], CFnet [31] and SAMF [23]) and eight RGB-T ones (MDNet+RGBT, SGT [21], SOWP+RGBT, CSR-DCF+RGBT, MEEM [36]+RGBT, CFnet+RGBT, KCF [10]+RGBT and L1-PF [34]). From the results, we can see that the performance of our FANet clearly outperforms the state-of-the-art RGB and RGB-T methods in all metrics. It demonstrates the importance of thermal information and quality-aware feature aggregations proposed in our method. In particular, our FANet achieves 5.4%/4.2% performance gains in PR/SR over the second best RGB tracker MDNet, and achieves 4.2%/3.7% gains over the second best RGB-T tracker MDNet+RGBT.

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, KCF+RGBT, MEEM+RGBT, SOWP+RGBT, CSR-DCF+RGBT, CFNet+RGBT, MDNet+RGBT, SGT and MDNet+RGBT, as shown in Table 1. 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, LR, FM, and SV. 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 1.
First, although SGT and MDNet+RGBT perform well when no occlusion occurs, their performance drop a lot when partial or heavy occlusions happen. Our FANet keeps high tracking performance with partial or heavy occlusions, which proves the aggregated hierarchical features are able to improve the tracking robustness in presence of occlusions. Second, in adverse lighting conditions and thermal crossover, 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 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 low-resolution video sequences by our FANet. It may attribute to the hierarchical feature aggregation is unable to help improve feature representation for the target object with less appearance information.

**Qualitative Performance.** The qualitative comparison of our FANet versus four state-of-the-art RGB-T trackers on four video sequences is presented in Fig. 6, including MDNet [26]+RGBT, SGT [21], MEEM [36]+RGBT and CSR-DCF [24]+RGBT. Overall, our method is effective in handling the challenges of occlusion, background clutter, appearance variation, thermal crossover, low resolution and illumination variation. For example, in Fig. 6 (b), our method performs well in presence of partial and heavy occlusions and large appearance changes while other trackers lose the target when occlusion happens. In Fig. 6 (c), the football is totally invisible in thermal source but the visible images can provide reliable information to distinguish the target from the background. In this case, all compared methods lose the target. But our FANet method tracks it robustly by adaptively incorporating useful information from these two modalities.

**4.4. Analysis of Our Network**

**Ablation study.** To justify the significance of the main components, we implement two special versions of our method for comparative analysis and evaluate them on both GTOT and RGBT234 datasets. These two variants are: 1) FANet-FA, that only uses the layer weights and removes the multimodal aggregation network; and 2) FANet-MA, that only uses the modality weights and removes the layer.
weights. Table 2 presents the comparison results. The superior performance of our FANet over FANet-FA and FANet-MA justifies the effectiveness of the proposed networks to predict the layer and modality weights.

**Feature analysis.** To analyze the effectiveness of hierarchical feature aggregation in the proposed network, we compare the tracking results using features of different convolutional layers of the VGG-M network [30] on the RGBT234 dataset. We first use each single convolutional layer (C1, C2 and C3) to represent objects in the proposed network. Fig. 7 shows better tracking performance can be obtained when a deeper layer is used, which can be attributed to the semantic abstraction from a deeper network. Moreover, we also use combinations of C1, C2 and C3 (C12, C13 and C23) to represent objects in the proposed network. Fig. 7 shows the evaluation results on the RGBT234 dataset.

The superior performance of our FANet over others demonstrates the effectiveness of the used hierarchical features. Fig. 7 shows better tracking performance can be obtained when a deeper layer is used, which can be attributed to the semantic abstraction from a deeper network. FANet-C13 and FANet-C23 methods perform better than FANet-C1, FANet-C2 and FANet-C3. The quantitative evaluation results show that hierarchical cues from multiple CNN layers do help improve tracking performance.

**Runtime analysis.** Finally, we present the runtime of our FANet against the baseline methods, MDNet [26] and MDNet+RGBT with their tracking performance on the RGBT234 dataset in Table 3. Our implementation runs on the platform of PyTorch with 4.2 GHz Intel Core i7-7700K and NVIDIA GeForce GTX 1080Ti GPU, and the average tracking speed is 1.3 FPS. Overall, the results demonstrate that our framework clearly outperforms baseline methods, MDNet (7.3%/6.5% performance gains in PR/SR on GTOT dataset and 5.4%/4.2% on RGBT234 dataset) with an acceptable impact on the frame rate (1.3 FPS versus 4.0 FPS), and MDNet+RGBT (8.5%/6.1% on GTOT dataset and 4.2%/3.7% on RGBT234 dataset) with a modest impact on the frame rate (1.3 FPS versus 2.1 FPS).

### 5. 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, abrupt motion, 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 our FANet achieves the state-of-the-art tracking performance. In our future work, we will integrate more diverse features, such as color names and HOG, into our framework and explore more prior knowledge to learn robust modality weights for RGB-T tracking.

References

[1] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. Eco: Efficient convolution operators for tracking. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3
[2] M. Danelljan, G. Häger, F. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In Proceedings of British Machine Vision Conference, 2014. 6
[3] 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. 6
[4] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer. Adaptive color attributes for real-time visual tracking. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014. 6
[5] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In ECCV, 2016. 3
[6] R. Gade and T. B. Moeslund. Thermal cameras and applications: A survey. Machine Vision and Applications, 25(1):245–262, 2014. 1
[7] B. Han, J. Sim, and H. Adam. Branchout: Regularization for online ensemble tracking with convolutional neural networks. In CVPR, 2017. 3
[8] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In ICCV, 2011. 6
[9] C. Hazirbas, L. Ma, C. Domokos, and D. C. Fusen. Incorporating depth into semantic segmentation via fusion-based cnn architecture. In ACCV, 2016. 3, 4
[10] 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, 37(3):583–596, 2015. 6
[11] 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. 1, 3
[12] I. Jung, J. Son, M. Baek, and B. Han. Real-time mdnet. In ECCV, 2018. 3
[13] 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. 6
[14] T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: Towards accurate region proposal generation and joint object detection. In CVPR, 2016. 3
[15] 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. 1, 2, 3, 5
[16] 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, 25(12):5743–5756, 2016. 1, 2, 3, 5, 6
[17] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang. Rgb-t object tracking: Benchmark and baseline. arXiv: 1805.08982, 2018. 5, 6
[18] 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, 47(4):673–681, 2017. 3
[19] 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, 27(4):725–738, 2017. 1, 3
[20] C. Li, X. Wu, N. Zhao, X. Cao, and J. Tang. Fusing two-stream convolutional neural networks for rgb-t object tracking. IEEE Transactions on Information Theory, 281:78–85, 2018. 1, 3
[21] 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. 1, 2, 3, 5, 6, 7
[22] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang. Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking. In Proceedings of European Conference on Computer Vision, 2018. 1, 3
[23] Y. Li and J. Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In European Conference on Computer Vision, pages 254–265. Springer, 2014. 6
[24] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan. Discriminative correlation filter with channel and spatial reliability. In CVPR, 2016. 6, 7
[25] 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. 2, 3, 4
[26] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In CVPR, 2016. 2, 3, 4, 5, 6, 7, 8
[27] E. Park and A. C. Berg. Meta-tracker: Fast and robust online adaptation for visual object trackers. In ECCV, 2018. 3
[28] S.-J. Park, K.-S. Hong, and S. Lee. Rdfnet:Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In ICCV, 2017. 3, 4
[29] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang. Hedged deep tracking. In CVPR, 2016. 3, 4
[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015. 4, 5, 8
[31] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr. End-to-end representation learning for correlation filter based tracking. In CVPR, 2017. 6

[32] 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 ECCV, 2016. 3

[33] A. Wu, W.-S. Zheng, H. Yu, S. Gong, and J. Lai. Rgb-infrared cross-modality person re-identification. In Proceedings of IEEE International Conference on Computer Vision, 2017. 1, 3

[34] 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. 6

[35] 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. 1, 3

[36] 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. 6, 7

[37] T. Zhang, C. Xu, and M.-H. Yang. Multi-task correlation particle filter for robust object tracking. In CVPR, 2017. 3

[38] W. Zhong, H. Lu, and M.-H. Yang. Robust object tracking via sparsity-based collaborative model. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012. 6