Cross-Modal Pattern-Propagation for RGB-T Tracking

Chaoqun Wang, Chunyan Xu*, Zhen Cui*, Ling Zhou, Tong Zhang, Xiaoya Zhang, Jian Yang
PCALab, Key Lab of Intelligent Perception and Systems for High-Dimensional
Information of Ministry of Education, School of Computer Science and Engineering,
Nanjing University of Science and Technology

Abstract

Motivated by our observations on RGB-T data that pattern correlations are high-frequently recurred across modalities also along sequence frames, in this paper, we propose a cross-modal pattern-propagation (CMPP) tracking framework to diffuse instance patterns across RGB-T data on spatial domain as well as temporal domain. To bridge RGB-T modalities, the cross-modal correlations on intra-modal paired pattern-affinities are derived to reveal those latent cues between heterogenous modalities. Through the correlations, the useful patterns may be mutually propagated between RGB-T modalities so as to fulfill inter-modal pattern-propagation. Further, considering the temporal continuity of sequence frames, we adopt the spirit of pattern propagation to dynamic temporal domain, in which long-term historical contexts are adaptively correlated and propagated into the current frame for more effective information inheritance. Extensive experiments demonstrate that the effectiveness of our proposed CMPP, and the new state-of-the-art results are achieved with the significant improvements on two RGB-T object tracking benchmarks.

1. Introduction

Visual object tracking is a fundamental and challenging task in the field of computer vision, has achieved significant progresses over the past few years with the breakthrough of deep neural network [8, 9, 19, 36, 39]. However, there are still various existing difficulties in the scenes of low illumination, heavy occlusion and dark night, *etc*. As the essential loss of object information, the current RGB-based trackers are often overwhelmed in these scenes. On the contrary, thermal infrared images can greatly reduce the influences of lightings, and effectively compensate to RGB images for identifying objects. This refers to the dual-modal RGB-T tracking, which has arisen increasing attention more recently due to the ease of use of thermal camera.

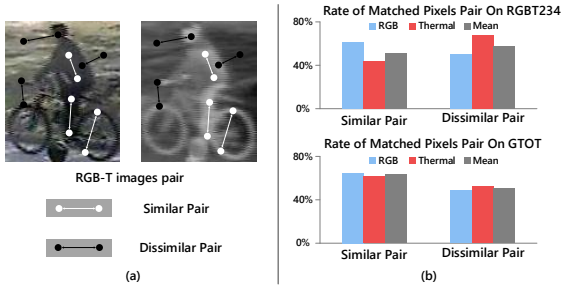


Figure 1: The statistical observation across visible RGB and thermal infrared images. (a) Visual exhibition. The points colored white are similar pairs across visible RGB and thermal infrared images, while the pairs of black points correspond to dissimilar pairs. (b) Statistical finding. Similar/dissimilar pairs across two modalities as well as their average statistics are with fairly high matching rates, which provide strong cues for cross-modal pattern-propagation.

The existing RGB-T tracking methods [23, 24, 25, 29, 30, 31, 32] usually follow the conventional weighted fusion on dual-modal (patch) features, or extend those classic RGB-based tracking techniques. For instance, the weights of modalities are integrated into sparse representation to be learnt in [24, 27]; after splitting candidate regions into patches, patch weights are adaptively learnt to construct target representation in [25, 31, 32]; some critical channels of all modalities are selected in [29, 46] according to their confidences. In [30], the part-based RGB tracking method is revised for dual-modal case, where the patches from two modalities are ranked in their importances for driving a confident tracking. However, all these methods do not deliberate also utilize the inner pattern-correlations cross modalities and even sequence frames, whist the pattern-correlations are high-frequently recurred therein.

To this end, we conduct statistical analysis about pattern correlations on two large RGB-T tracking datasets, RG-

*Corresponding author: {cyx; zhen.cui}@njust.edu.cn

BT234 [25] and GTOT [24]. The quantitative statistical results cross RGB and thermal images are shown in Fig. 1. The affinity of pair pixels is defined by the Euclidean metric on their gray values. The larger the affinity value is, the more similar this pair of pixels are. According to the affinity values, we define those similar pairs and dissimilar pairs within each RGB or thermal image. Further a counterpart of inter-modal pairs are determined on the condition of the same affinity values at the same space position. In other word, the counterpart defines a matched relation pattern, viewed as a second-order correlation. All matched inter-modal pairs (*i.e.*, pair counterparts) are summarized to produce the statistical ratios cross RGB and thermal modalities. From this figure, we can observe that the pattern counterparts of inter-modal pairs are high-frequently occurred for the similar and dissimilar pairs. This case is named as inter-modal pattern-correlation, which is a second-order relationship (*i.e.*, relation on relation). Besides, another explicit observation is that image patches tend to redundantly recur many times across adjacent sequence frames because of the continuity of sequence. Thus we can define the pattern affinities across sequence frames, which is one-order intra-modal pattern-correlation.

Just motivated by the above observations, in this paper, we propose a cross-modal pattern-propagation (CMPP) tracking framework to diffuse instance patterns across RGB and thermal infrared modalities, as well as within single modality. In virtue of the cross-modal second-order statistical correlations, we propose an inter-modal pattern-propagation method to transmit patterns across heterogenous modalities. To bridge them, the cross-modal correlations on within-image affinities are derived to reveal those associative patterns between different modalities. Through the correlations, the useful patterns may be mutually propagated between modalities so that feature information can be compensated for each other. In view of the temporal continuity of sequence frames, further, we extend the spirit of pattern propagation from cross-modal spatial domain to temporal domain to construct dynamical pattern propagations. In a sequence, long-term historical contexts are adaptively correlated and propagated into the current frame for more effective information inheritance during the online tracking. Extensive experiments demonstrate that our CMPP is greatly superior to those baselines as well as the state-of-the-art methods.

In summary, our contributions are three folds:

- Based on our findings on RGB-T images, we propose a cross-modal pattern-propagation framework to excavate and utilize those latent pattern cues for online RGB-T object tracking.

- We jointly build inter-modal pattern-propagation via a second-order affinity correlations for the interaction across modalities, and long-term context-propagation through high-order dynamic correlations for historical information inheritance.

- We report the new state-of-the-art results on two RGB-T object tracking benchmarks with the significant improvement.

2. Related Work

Visual Object Tracking. Recently correlation filter based and CNN-based trackers achieve state-of-the-art performances on the public visual object tracking benchmarks [41, 42]. MOSSE [2] firstly used adaptive filter for visual tracking, then numerous correlation filter based trackers sprung up and performed excellent in the yield of visual object tracking. Henriques *et al.* [17] used kernel trick, Danelljan *et al.* [12] exploited color attributes as target features, and SAMF [33], KCF [18] used scale estimation to handle various target scale problems. CNN-based methods regard tracking task as detection process. MDNet [36] trained a general offline binary classification model using multi-domain learning to distinguish target from the background. RT-MDNet [19] introduced RoIAlign method to extract more accurate representation for target. Park *et al.* [37]exploited meta learning algorithm into MDNet, which adjusted initial model via temporal information in tracking sequence for quick optimization. These trackers performed well but overwhelmed in the conditions of low illumination, smog and dark night, *etc*., limited by the inherent defects of RGB images.

RGB-T Object Tracking. RGB-T object tracking is getting more and more attention due to the promotion of thermal infrared data in tracking performance [13]. Sparse representation based trackers performed well because of its capability of suppressing noises and errors [24, 40, 23]. Wu *et al.* [40] concatenated the candidate regions of multi-modal data and sparsely represented target in template space; Li *et al.* [24] proposed a collaborative sparse representation model to jointly optimize the sparse coefficients and modality weights. Furthermore, Li *et al.* [30] considered heterogeneous property between multi-modal source in cross-modal ranking model, to learn patch-based weights and construct target representation. Using neural network to adaptively fuse multiple data also achieves excellent performance. Li *et al.* [29] proposed a fusion net to select critical feature channels for confident tracking; Zhu *et al.* [45] proposed a feature aggregation network to weight multi-modal and multi-scale features for robust target representation. These methods utilize multi-modal data to construct target representation by extracting factors or balancing their confidences. They neglected the inner pattern-correlations across multi-modal data, which are high-frequently recurred and can be utilized for mutually complementing each other.

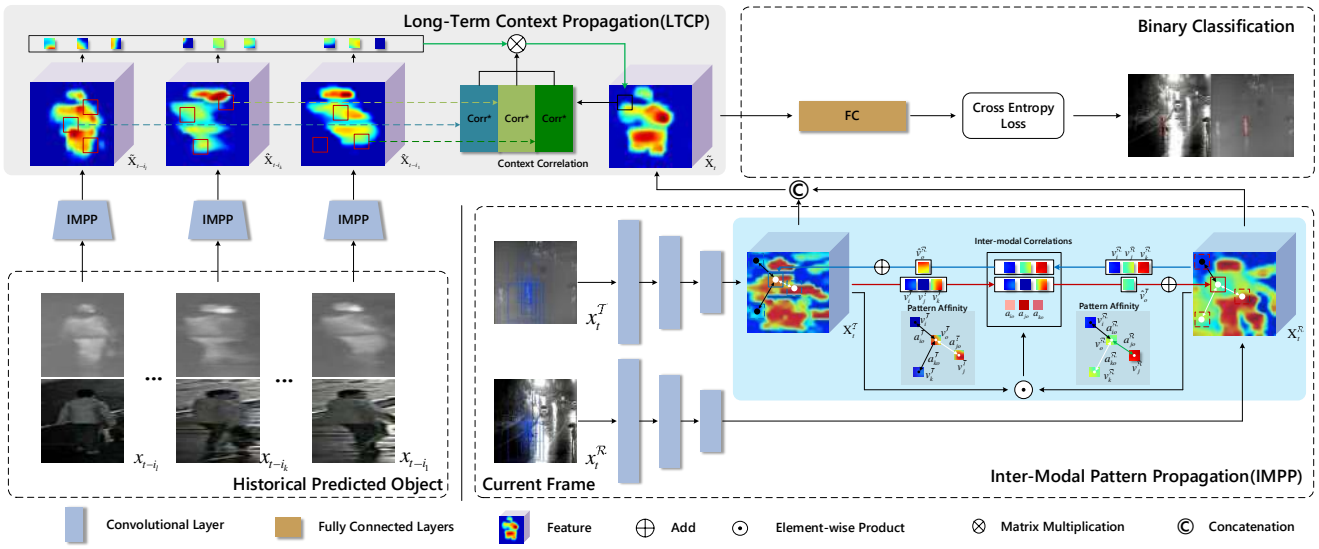


Figure 2: Our proposed CMPP framework for RGB-T object tracking. There are two main modules to fulfill spatially cross-modal and temporally sequent pattern propagation, corresponding to inter-modal pattern-propagation and long-term context propagation respectively. The former builds and utilizes the correlations between pairwise bundling patterns, while the latter leverages the continuity of pattern sequences. More details could be referred in Section 3. Zoom the figure for clear details.

3. Our Method

In this section, we introduce the proposed cross-modal pattern-propagation method. We first overview the overall architecture, and then illustrate how to perform inter-modal pattern-propagation process for cross-modal spatial fusion, finally expound long-term context propagation module for temporal sequence encoding.

3.1. The Architecture

The overall network architecture of our proposed CMPP is shown as Fig. 2. In the online RGB-T object tracking, given a pair of frames $\{x_t^{\mathcal{R}}, x_t^{\mathcal{T}}\}$ at time t, we need to estimate the localization of tracked object. The candidate bounding boxes are first sampled around an initial location of object, which is determined by the location at time $t-1$. To extract more discriminative representation, we adopt VGG-M network [3] to filter each candidate region and produce multi-channel convolution feature maps. By omitting the index of candidates, we abuse the notation $\{\mathbf{X}_t^{\mathcal{R}}, \mathbf{X}_t^{\mathcal{T}}\}$ as the convolution features of one pair of dual-modal candidates for more clear illustration below. Our aim is to predict whether the candidates $\{\mathbf{X}_t^{\mathcal{R}}, \mathbf{X}_t^{\mathcal{T}}\}$ are the most possible object.

In view of candidates $\{\mathbf{X}_t^{\mathcal{R}}, \mathbf{X}_t^{\mathcal{T}}\}$, we propose an inter-modal pattern-propagation (IMPP) method to mutually diffuse patterns for each other. To bridge the difference between them, we attempt to leverage those latent cues of second-order correlations as observed in Fig 1. Concretely, the affinities of pairwise patterns within each candidate are computed to mine and utilize the cues of bundling patterns, which refers to intra-modal correlation computation. Next, the bundling patterns are compared across two modalities by using their affinities, and the inter-modal correlations are derived to favor cross-modal pattern propagation. The details could be found in Section 3.2.

For online object tracking, the historical sequent frames can benefit the localization of moving object. To this end, we further extend the spirit of pattern-propagation into the dynamic sequence, and propose long-term context propagation (LTCP, in Section 3.3) to adaptively inherit previous historical information. Concretely, given the predicted dual-modal objects $\{x_{t-i_l}, \cdots, x_{t-i_1}\}$ at the former l frames, we take IMPP to integrate them to produce the diffused features $\{\widetilde{\mathbf{X}}_{t-i_l}, \cdots, \widetilde{\mathbf{X}}_{t-i_1}\}$. The cross-modal diffused features are adaptively propagated into the current candidate and further integrated with its diffused features $\widetilde{\mathbf{X}}_t$ as the final response, which are fed into a binary classification network. The binary classification network has three fully connected layers followed by cross-entropy loss to estimate the possibility of the candidates as background or foreground. The first k confident candidates are used for regressing the final location of object as used in MDNet [36].

Besides, we consider that different depth convolutional layers carry with different level feature information, where the shallow convolutional layer expresses more local texture information, while deep convolutional layer captures more global semantic information. For this, integrating multi-scale pyramidal feature maps is customary to boost the performance of many tasks such as object detection [1, 22, 35], classification [38, 7, 34], and semantic seg-

mentation [5, 4, 44]. To this end, we downscale multi-layer convolution feature maps into the smallest-scale map according to the method proposed in [6], and then concatenate them to fed into our proposed propagation method.

3.2. Inter-Modal Pattern-Propagation

Suppose a candidate region contains n pixel points, each pixel position is associated with a multi-dimensional feature vector. We build graph topology on the candidate region, where each pixel is viewed as one vertex. Abstractly, we formulate inter-modal pattern propagation as

$$v_i^{\mathcal{T}} \leftarrow f(v_i^{\mathcal{T}}) + \lambda^{\mathcal{T}} \sum_{j \in \mathcal{N}(v_i)} w_{ij} f(v_j^{\mathcal{R}}), \tag{1}$$

$$v_i^{\mathcal{R}} \leftarrow f(v_i^{\mathcal{R}}) + \lambda^{\mathcal{R}} \sum_{j \in \mathcal{N}(v_i)} w_{ij} f(v_j^{\mathcal{T}}), \tag{2}$$

$$w_{ij} = \mathcal{C}(\mathcal{A}(f(v_i^{\mathcal{T}}), f(v_j^{\mathcal{T}})), \mathcal{A}(f(v_i^{\mathcal{R}}), f(v_j^{\mathcal{R}}))), \tag{3}$$

where $f(\cdot)$ are feature extractors of vertices $\{v_i^{\mathcal{T}}, v_i^{\mathcal{R}}\}$, $\mathcal{A}(\cdot,\cdot)$ is an affinity computation function between two vertices within a single-modal image, $\mathcal{C}(\cdot,\cdot)$ is a cross-modal correlation function between those pairwise vertices. The score/weight w_{ij} indicates a certain latent cue between one pattern-pair of RGB image and the other pattern-pair of the corresponding thermal infrared image at the same spatial position. The higher the weight is, the stronger the cue is among them. If $w_{ij} = 0$, the corresponding pairs have no any cues. Usually, we constrain the weight $w_{ij} > 0$ with the global sparsity, which means non-dense edge connections. Thus, we can construct the neighbor relationship $\mathcal{N}(v_i)$ based on the produced weights $\{w_{ij}\}$. The pattern propagation processes in Eqn. (1) and Eqn. (2) take the summation aggregation through weighting on those adjacent patterns of the counterpart modality, which followed by the pattern of modality therein. The hyperparameters $\lambda^{\mathcal{T}}$ and $\lambda^{\mathcal{R}}$ are the balance factors during the pattern propagation.

As multi-channel RGB-T features have the grid-shape structure, we convert the calculation into the matrix formulation. Given a pair of dual-modal candidates $\{\mathbf{X}^{\mathcal{R}}, \mathbf{X}^{\mathcal{T}} \in \mathbb{R}^{H\times W\times C}\}$, where we omit the time information t for simplification. Thus the number of vertices is $n = H \times W$. To conveniently define the matrix computation, we take a format operation $[\mathbf{X}^{\mathcal{T}}]$: $\mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{n\times C}$, which stacks spatial positions in a row-by-row way. The inter-modal pattern correlations relationship between visible RGB and thermal infrared features can be formulated as

$$\mathcal{C} = \sigma(\mathcal{A}^{\mathcal{T}} \odot \mathcal{A}^{\mathcal{R}})/\mathbf{F}, \tag{4}$$

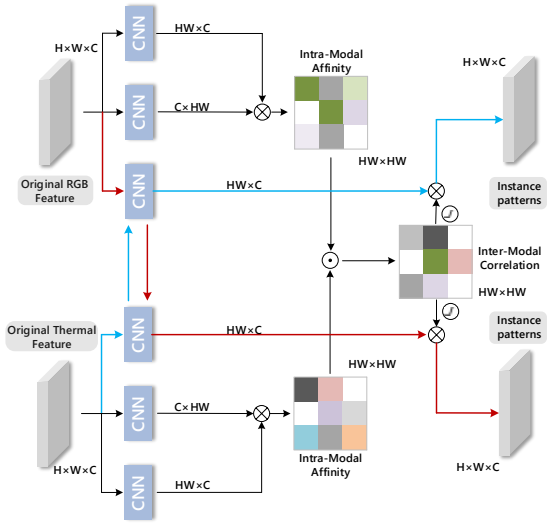
$$\mathcal{A}^{\mathcal{T}} = \mathcal{S}([f_1^{\mathcal{T}}(\mathbf{X}^{\mathcal{T}})] \times [f_2^{\mathcal{T}}(\mathbf{X}^{\mathcal{T}})]^{\intercal}), \tag{5}$$

$$\mathcal{A}^{\mathcal{R}} = \mathcal{S}([f_1^{\mathcal{R}}(\mathbf{X}^{\mathcal{R}})] \times [f_2^{\mathcal{R}}(\mathbf{X}^{\mathcal{R}})]^{\intercal}), \tag{6}$$

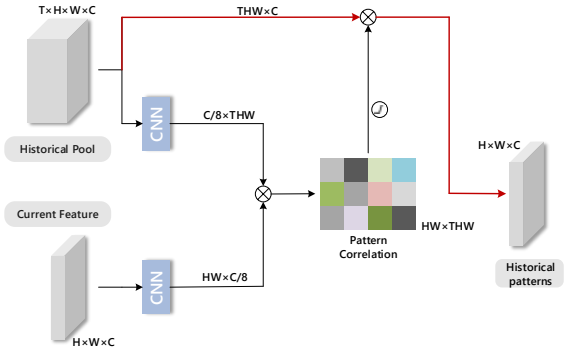
where $\odot$ denotes the element-wise multiplication, σ is the exponential function, $\mathbf{F} = \sigma(\mathcal{A}^{\mathcal{T}} \odot \mathcal{A}^{\mathcal{R}}) \times \mathbf{1} \times \mathbf{1}^{\intercal}$ is the normalization factor, $f_1^{\mathcal{T}}, f_2^{\mathcal{T}}, f_1^{\mathcal{R}}, f_2^{\mathcal{R}}$ are the convolutional layers with 1×1 kernel to be learnt in the network, the affinity matrices $\mathcal{A}^{\mathcal{T}}, \mathcal{A}^{\mathcal{R}}$ are computed by the inner production on the feature matrices therein, and $\mathcal{S}(\cdot)$ is the sparsity operation to choose those important pairs for the construction of sparse graph (*e.g.*, set a threshold τ to mask those values less than τ as zeros). Further, we define the pattern propagation across the two modalities as

$$[\widetilde{\mathbf{X}}^{\mathcal{T}}] = [\mathbf{X}^{\mathcal{T}}] + \lambda^{\mathcal{T}} \times \mathcal{C} \times [g^{\mathcal{R}}(\mathbf{X}^{\mathcal{R}})], \tag{7}$$

$$[\widetilde{\mathbf{X}}^{\mathcal{R}}] = [\mathbf{X}^{\mathcal{R}}] + \lambda^{\mathcal{R}} \times \mathcal{C} \times [g^{\mathcal{T}}(\mathbf{X}^{\mathcal{T}})], \tag{8}$$



(a) Illustration of inter-modal pattern-propagation process.



(b) Illustration of long-term context-propagation process.

Figure 3: Two critical pattern-propagation modules. (a) Inter-modal pattern correlations are defined on intra-modal correlations, which compute the affinities of paired bundling patterns in single modality. Based the correlations, the patterns of one modality could be propagated into the other modality, or oppositely. (b) The patterns of historical frames are correlated and propagated into the candidate at the current frame feature for the use of long-term contexts. $\times$ and $\odot$ denote matrix multiplication and element-wise product respectively.

where the nonlinear transformation $g^{\mathcal{T}}, g^{\mathcal{R}}$ are the convolutional layers with 1×1 kernel, $\lambda^{\mathcal{R}}, \lambda^{\mathcal{T}}$ are adaptively learnt balance factors. At last, we concatenate the two new-fused features $\widetilde{\mathbf{X}}^{\mathcal{T}}, \widetilde{\mathbf{X}}^{\mathcal{R}}$ to form the RGB-T response output $\widetilde{\mathbf{X}}$. The concrete network module is shown in Fig. 3(a).

3.3. Long-Term Context Propagation

Context information is very important in tracking task for avoiding model drift [47, 14]. To this end, we construct a historical information pool, which consists of those objects with high confidence and is usually dynamically updated with sequence variations. We expect that the historical context patterns can be propagated into the candidates of the current frame to boost online object tracking performance. Given anyone pair of dual-modal candidates $\{x^{\mathcal{R}}, x^{\mathcal{T}}\}$, we can derive their fused feature $\widetilde{\mathbf{X}}$ based on the above IMP-P module. Suppose the historical pool is the feature set of l objects, $\{\widetilde{\mathbf{X}}_{t-i_l}, \widetilde{\mathbf{X}}_{t-i_{l-1}} \cdots, \widetilde{\mathbf{X}}_{t-i_1}\}$, we can define the pattern correlations between them and the current candidate $\widetilde{\mathbf{X}}_t$. As they are all in the same feature space, we directly compare inter-frame patterns. Formally,

$$[\widetilde{\mathbf{X}}_t] \leftarrow [\widetilde{\mathbf{X}}_t] + \frac{\gamma}{\mathbf{G}} \sum_{k=1}^{l} \mathcal{C}_{t,t-i_k} [\widetilde{\mathbf{X}}_{t-i_k}], \tag{9}$$

$$\mathcal{C}_{t,t-i_k} = \mathcal{S}([h_1(\widetilde{\mathbf{X}}_t)] \times [h_2(\widetilde{\mathbf{X}}_{t-i_k})]^{\intercal}), \tag{10}$$

where the nonlinear transformation h_1, h_2 are designed as the convolutional layers with 1×1 kernel to be learnt, $\mathbf{G}$ is the normalization factor, which scales the sum of correlation weights to be equal to 1 for each spatial position, γ is a balance factor, and the correlation weight $\mathcal{C}_{t,t-i_k}$ represents the relationship between the current candidate and the previous historical frame $t - i_k$. By adaptively learning the weights, the historical patterns can be properly propagated into the current candidate to enhance the feature information for more robust object tracking. The concrete network module is provided in Fig. 3(b).

4. Implementation Details

We implement our proposed CMPP on the PyTorch platform with E5-2650@2.20GHz CPU and NVIDIA GeForce GTX 2080Ti GPU. We use MDNet [36] as the backbone to build our network, and elaborate the training as well as online tracking strategy as follows.

4.1. Training Procedure

Pre-training. In this stage, we firstly remove LTCP module and initialize the parameters of first three sequential convolutional layers using the pre-trained VGG-M [3] network. Then we crop positive and negative samples in training sequences and optimalize the cross-entropy loss by the Stochastic Gradient Descent (SGD) [20] algorithm where each domain is handled separately. In the iteration, we randomly choose 8 frames, from which we crop 32 positive (IoU in $0.7 \sim 1.0$) and 96 negative (IoU in $0 \sim 0.5$) samples and construct minibatch in each video sequence. For multi-domain learning, we set K fully connected layer branches for K video sequences and train the network for 100K iterations. We set the learning rate to 0.0001 for parameters of the IMPP module and 0.001 for the rest. The weight decay and momentum are fixed to 0.0005 and 0.9, respectively.

Training. We initialize our model using the pre-trained network and train the whole model in this stage. For a video sequence, we crop 32 positive and 96 negative samples from a single frame as minibatch, and then crop 16 positive samples from 4 previous frames as the historical objects. We fine-tune the parameters of IMPP module and three sequential convolutional layers with learning rate of 1e-6 while set it to 0.0001 and 0.001 for parameters of LTCP module and the rest respectively.

4.2. Online Tracking

In tracking, the K branches for multi-domain learning are replaced by a single branch. Then the CMPP fine-tunes the pre-trained network on the first frame. In the fine-tuning stage, we crop 500 positive and 5000 negative samples with the given ground-truth bounding box and train the model for 30 iterations. The learning rate is 0.0005 for parameters of the first two fully connected layers and 0.005 for parameters of the last one. After that we utilize the output features to train a ridge regression module for bounding box regression. For the t-th frame, we crop 256 sample regions as candidates $\{x_t^i\}$ under the guidance of predicted result in the $(t-1)$-th frame. Then we can obtain their positive and negative scores. We find k candidates ($k = 5$) with the maximum positive scores and employ the bounding box regressor to improve target localization accuracy, their mean value can be seen as the optimal target state x^*, more details can be referred in [36]. Moreover, historical pool is constructed when online tracking begins. In each frame, we insert the reliable prediction region whose positive score over zero into the tail of historical pool, and pop the top one if there are more than l frames in historical pool.

5. Experiment

5.1. Datasets and Evaluation Metrics

We implement our proposed CMPP framework on two large RGB-T tracking benchmarks, GTOT [24] and RGBT234 [25], to demonstrate its effectiveness. The GTOT dataset contains 50 spatially and temporally aligned visible RGB and thermal infrared sequences, while the RGBT234 dataset includes 234 RGB-T videos and 12 annotated attributes, which totally reach 234K frames. We select the whole GTOT dataset as the training set when evaluating on

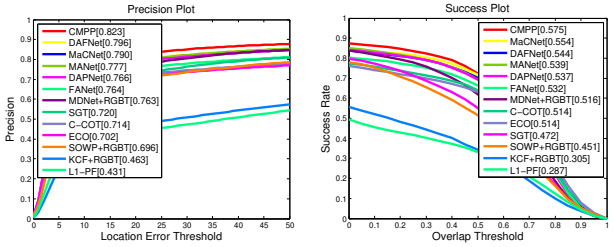
RGBT234 and randomly select 70 videos of the RGBT234 dataset as the training set when evaluating on GTOT.

We utilize two widely used metrics, precision rate (PR) and success rate (SR), to evaluate the tracking performance. In a specific, PR is the percentage of frames whose Euclidean distance between the predicted bounding box and ground-truth within a manually set threshold (5 pixels in GTOT and 20 pixels in RGBT234). SR is the percentage of frames whose overlap ratio between the predicted bounding box and ground-truth larger than a specified threshold, and we compute the SR score by the area under curves (AUC).

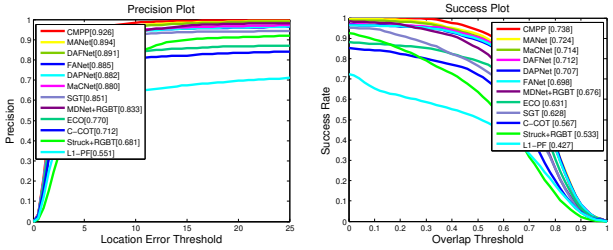
5.2. Comparison with State-of-the-art Trackers

Overall performance. We implement our CMPP on GTOT and RGBT234 benchmarks, and compare the tracking performances with state-of-the-art trackers, *i.e.*, RGB trackers (including C-COT [11] and ECO [10]) and RGB-T trackers (including MANet [26], DAPNet [46], DAFNet [15], MaCNet [43], FANet [45], SGT [28], Struck [16]+RGBT, SOWP [21]+RGBT, KCF [18]+RGBT, L1-PF [40], and MDNet [36]+RGBT). MDNet+RGBT is our baseline. We directly concatenate the multi-scale RGB and thermal infrared features and train the model on the training set. The overall tracking performances are shown in Fig. 4. Our CMPP achieves significant outstanding performance over other state-of-the-art trackers for all metrics on both two benchmarks. Specifically, on the RGBT234 benchmark, our CMPP achieves 82.3%/57.5% in PR/SR, and has 6.0%/5.9% promotion over baseline, while on the GTOT benchmark, our CMPP achieves 92.6%/73.8% in PR/SR, and has 9.3%/6.2% improvement against the baseline. The excited performance and significant promotion demonstrate the effectiveness of our proposed framework.

Attribute-based performance. To further demonstrate the effectiveness of our proposed CMPP method, we plot the attribute-based performance on RGBT234. RGBT234 dataset contains 12 annotated attributes, including 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 tracking results are shown in Table 1. We can observe that the proposed method significantly outperforms other trackers in the most annotated attributions. Specifically, in the challenge of low illumination, low resolution, thermal crossover and fast motion, our CMPP achieves around 10% promotion in PR/SR against baseline. In these scenes, single modal data is unreliable, such as RGB images with low illumination and thermal infrared images with thermal crossover. Considering all intra-modal affinity relationships, our CMPP can obtain confident inter-modal correlations, which are utilized to extract reliable instance patterns from the high quality images



(a) comparison on RGBT234



(b) comparison on GTOT

Figure 4: Overall performances comparison with state-of-the-art trackers on RGBT234 (a) and GTOT (b).

and propagate them into the ropey ones to enhance their discrimination. Besides, in the challenge of thermal crossover (TC), the single modal tracker C-COT achieves the best performance because the thermal infrared data is extremely unreliable and is viewed as noise. Even so, our CMPP can still achieve a comparable performance after fusing the branch of unconfident thermal infrared data, which again demonstrates the effectiveness of our CMPP.

5.3. Ablation Study

Single/dual-modal data. To demonstrate the effectiveness of multi-modal data, we plot the tracking performances of CMPP+RGB, CMPP+T, and CMPP. CMPP+RGB and CMPP+T denote the experiments of our method with RGB and thermal infrared modality respectively. The tracking performances are shown in Fig. 5(a). The CMPP is significantly better than the experiments with single modality input, and demonstrates the great promotion of multiple modalities data in the visual object tracking task.

Prune experiments. To validate the effectiveness of our major contributions, we implement two variants, including 1) w/o-LTCP, that prunes long-term context propagation module, 2) w/o-IMPP, that prunes inter-modal pattern-propagation module. The comparison results on RGBT234 are shown in Fig. 5(b). We draw two conclusions from the results. 1) The w/o-LTCP and w/o-IMPP frameworks outperform baseline(MDNet+RGBT) as well as other state-of-the-art RGB-T trackers, demonstrating the effectiveness of our proposed IMPP and LTCP modules respectively, 2) The CMPP method performs better than the pruned frameworks, demonstrating that the IMPP and LTCP can improve track-

Table 1: Attribute-based PR/SR scores(%) on RGBT234 dataset against state-of-the-art trackers. The best, second and third performances are represented in red, green, and blue respectively.

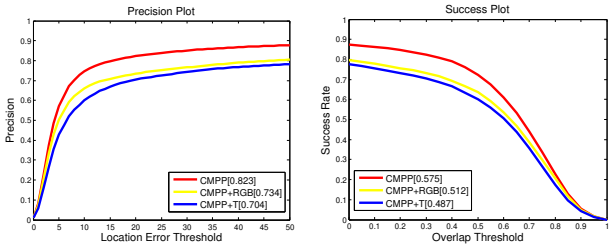
	ECO [10]	C-COT [11]	FANet [45]	DAPNet [46]	MaCNet [43]	DAFNet [15]	MDNet+RGBT	CMPP
NO	88.0/65.5	88.8/65.6	84.7/61.1	90.0/64.4	92.7/66.5	90.0/63.6	89.5/62.6	95.6/67.8
PO	72.2/53.4	74.1/54.1	78.3/54.7	82.1/57.4	81.1/57.2	85.9/58.8	79.6/53.5	85.5/60.1
HO	60.4/43.2	60.9/42.7	70.8/48.1	66.0/45.7	70.9/48.8	68.6/45.9	67.0/44.9	73.2/50.3
LI	63.5/45.0	64.8/45.4	72.7/48.8	77.5/53.0	77.7/52.7	81.2/54.2	74.5/48.1	86.2/58.4
LR	68.7/46.4	73.1/49.4	74.5/50.8	75.0/51.0	78.3/52.3	81.8/53.8	75.8/48.9	86.5/57.1
TC	82.1/60.9	84.0/61.0	79.6/56.2	76.8/54.3	77.0/56.3	81.1/58.3	73.9/51.1	83.5/58.3
DEF	62.2/45.8	63.4/46.3	70.4/50.3	71.7/51.8	73.1/51.4	74.1/51.5	70.8/50.0	75.0/54.1
FM	57.0/39.5	62.8/41.8	63.3/41.7	67.0/44.3	72.8/47.1	74.0/46.5	66.4/43.3	78.6/50.8
SV	74.0/55.8	76.2/56.2	77.0/53.5	78.0/54.2	78.7/56.1	79.1/54.4	76.8/52.0	81.5/57.2
MB	68.9/52.3	67.3/49.5	67.4/48.0	65.3/46.7	71.6/52.5	70.8/50.0	67.8/48.0	75.4/54.1
CM	63.9/47.7	65.9/47.3	66.8/47.4	66.8/47.4	71.7/51.7	72.3/50.6	71.0/50.1	75.6/54.1
BC	57.9/39.9	59.1/39.9	71.0/47.8	71.7/48.4	77.8/50.1	79.1/49.3	74.4/48.9	83.2/53.8
ALL	70.2/51.4	71.4/51.4	76.4/53.2	76.6/53.7	79.0/55.4	79.6/54.4	76.3/51.6	82.3/57.5

ing performance from different aspects and jointly promote the tracking results to a certain extent.

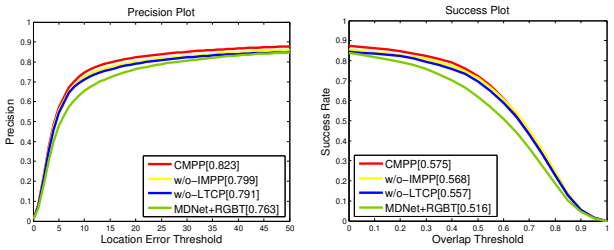
Model parameters comparison. The length of historical frames l and balance factor γ in the long-term context propagation module are critical hyperparameters in our proposed framework. We manually set l={8,16,32,64} and γ={0.01,0.02,0.03,0.04,0.05} to analyze their influences. The tracking performances on RGBT234 are shown in Table 2. The tracker achieves the best performance with γ=0.02 and l=32. By observing the tracking performances, we can find that the tracker performs better when there are more historical frames involved, because the LTCP module can properly propagate more historical confident information from longer previous frames to the current frame. Conversely, the performances are worse when the balance factor γ is too large (γ=0.05) or too small (γ=0.01), because too much propagation historical information might overwhelm the current feature, while too less propagation historical information tends to result in the model drift.

Modality-missed experiments. Our CMPP module can obtain reliable instance patterns and propagate across multiple modalities to enhance target representation, even when a branch of modality data is lost. The main reason is that the unharmed modality can timely compensate for the lost one through the use of inter-modal correlations as shown in Fig. 3(a). To test this case, we conduct the modality-missed experiments by disturbing an image pair with a certain probability. Concretely, we randomly mask RGB or thermal infrared image at a time step to simulate the modality losing, which is named modality-missed experiment and performed on RGBT234 dataset.

We manually set mask ratios r={0.1,0.2,0.3,0.4,0.5} and plot the success ratio (SR) variation curve of our CMPP and baseline. The experiment results are shown in Fig. 7.



(a) single/dual-modalities experiments



(b) prune experiments

Figure 5: Performances comparison of single/dual-modalities (a) and prune (b) experiments on RGBT234.

Our CMPP obviously decreases slower than baseline with the increase of mask ratio. Compared with the case of no mask, the decline of our CMPP framework and baseline are 2.7% and 4.8% respectively, even when masking 50% frames. The experiments demonstrate the robustness of our CMPP to the case of modality losing.

5.4. Qualitative performances

The visual comparisons between our proposed CMPP method and the other state-of-the-art trackers are shown in Fig. 6, including MDNet [36]+RGBT, SOWP [21]+RGBT, SGT [28], and C-COT [11]. Our approach performs obvi-

Figure 6: Qualitative comparison between CMPP and other state-of-the-art trackers on four video sequences.

Table 2: Performances (PR/SR) comparison of different parameters on RGBT234. The best, second and third performance are represented in red, green, and blue respectively.

PR/SR(%)	l=8	l=16	l=32	l=64
γ=0.01	81.1/56.6	79.1/55.7	80.7/56.3	80.1/56.0
γ=0.02	80.2/56.1	80.8/56.3	82.3/57.5	82.2/57.2
γ=0.03	79.5/55.7	81.4/56.9	80.8/56.6	80.2/56.2
γ=0.04	80.3/56.1	80.7/56.7	81.6/57.0	82.1/57.3
γ=0.05	81.3/56.7	80.1/56.1	81.1/56.4	79.7/56.0

ously better than other trackers in various challenges, such as heavy occlusion, low illumination, low resolution, and background clutter. For instance, Fig. 6(a) and Fig. 6(b) show the tracking results of video sequences with low illumination and background clutter. Our tracker can localize the target well while others lose the tracked target when the light condition sharply changes. In Fig. 6(c) and Fig. 6(d), which have background clutter, low resolution and heavy occlusion attributes, our approach can benefit from the joint intra-modal and inter-modal pattern propagation strategies to achieve excellent object localization.

5.5. Efficiency Analysis

Our method mainly contains two modules: IMPP and LTCP. The complexities are $\mathcal{O}(H^2W^2C + EHWC)$ and $\mathcal{O}(H^2W^2CL)$ respectively, where H, W, C, L are height, width, channel number and memory pool size, and $E \ll HW$ is the nonzero-value number in the sparse matrix $\mathcal{C}$. In practice, our CMPP is 1.3FPS on RGBT234 while the baseline is 1.5FPS, on par with MDNet-based RGBT trackers such as MANet (1.1FPS) and FANet (1.3FPS). Actually, the backbone network (MDNet) dominates the running time.

6. Conclusion

In this paper, we proposed a cross-modal pattern-propagation (CMPP) RGB-T tracking method, consisting of an inter-modal pattern-propagation (IMPP) and a long-term context propagation (LTCP) module. In virtue of intra-modal affinity relationships of dual modalities, the IMPP module can derive confident inter-modal correlations to obtain reliable instance patterns and propagate them across dual modalities to enhance original features. Such an IMPP process effectively integrates dual-modal information, and results in the robustness of tracking even for modality losing. In LTCP module, long-term context information can be well inherited under the guidance of correlations between previous historical frames and the current frame, and thus can relieve model drift to some extent in online tracking phase. We conducted extensive experiments on two large RGB-T datasets, and validated the effectiveness of our method as well as the main modules therein. In the future, we will extend our idea to more modalities to further boost tracking performance.

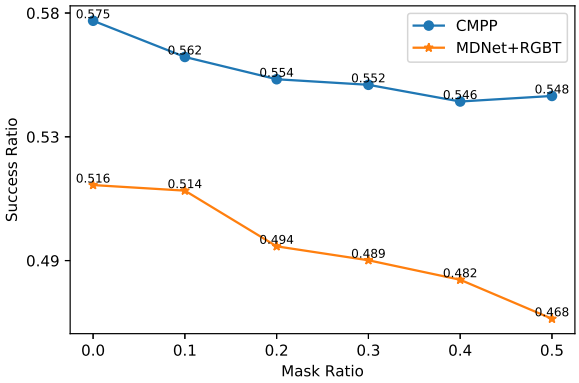


Figure 7: Performances (SR) comparison between CMPP and MDNet+RGBT with a specified mask ratio.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grants Nos. 61972204, 61772276, 61906094, U1713208) and the Natural Science Foundation of Jiangsu Province (Grants Nos. K20191283, BK20190019).

References

[1] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2874–2883, 2016.

[2] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2544–2550. IEEE, 2010.

[3] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[5] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[6] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019.

[7] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE international conference on computer vision*, pages 2590–2600, 2017.

[8] Zhen Cui, Youyi Cai, Wenming Zheng, Chunyan Xu, and Jian Yang. Spectral filter tracking. *IEEE Transactions on Image Processing*, 28(5):2479–2489, 2018.

[9] Zhen Cui, Shengtao Xiao, Jiashi Feng, and Shuicheng Yan. Recurrently target-attending tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1449–1458, 2016.

[10] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6638–6646, 2017.

[11] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *European Conference on Computer Vision*, pages 472–488. Springer, 2016.

[12] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost Van de Weijer. Adaptive color attributes for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1097, 2014.

[13] Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: a survey. *Machine vision and applications*, 25(1):245–262, 2014.

[14] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4649–4659, 2019.

[15] Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. Deep adaptive fusion network for high performance rgbt tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[16] Sam Hare, Stuart Golodetz, Amir Saffari, Vibhav Vineet, Ming-Ming Cheng, Stephen L Hicks, and Philip HS Torr. Struck: Structured output tracking with kernels. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2096–2109, 2015.

[17] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *European conference on computer vision*, pages 702–715. Springer, 2012.

[18] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014.

[19] Ilchae Jung, Jeany Son, Mooyeol Baek, and Bohyung Han. Real-time mdnet. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 83–98, 2018.

[20] Nikhil Ketkar. Stochastic gradient descent. In *Deep learning with Python*, pages 113–132. Springer, 2017.

[21] Han-Ul Kim, Dae-Youn Lee, Jae-Young Sim, and Chang-Su Kim. Sowp: Spatially ordered and weighted patch descriptor for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3011–3019, 2015.

[22] Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun. Hypernet: Towards accurate region proposal generation and joint object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 845–853, 2016.

[23] Xiangyuan Lan, Mang Ye, Shengping Zhang, Huiyu Zhou, and Pong C Yuen. Modality-correlation-aware sparse representation for rgb-infrared object tracking. *Pattern Recognition Letters*, 2018.

[24] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12):5743–5756, 2016.

[25] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: benchmark and baseline. *Pattern Recognition*, 96:106977, 2019.

[26] Chenglong Li, Andong Lu, Aihua Zheng, Zhengzheng Tu, and Jin Tang. Multi-adapter rgbt tracking. *arXiv preprint arXiv:1907.07485*, 2019.

[27] Chenglong Li, Xiang Sun, Xiao Wang, Lei Zhang, and Jin Tang. Grayscale-thermal object tracking via multitask laplacian sparse representation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(4):673–681, 2017.

[28] Chenglong Li, Xiao Wang, Lei Zhang, Jin Tang, Hejun Wu, and Liang Lin. Weighted low-rank decomposition for robust grayscale-thermal foreground detection. *IEEE Transactions*

on Circuits and Systems for Video Technology, 27(4):725–738, 2016.

[29] Chenglong Li, Xiaohao Wu, Nan Zhao, Xiaochun Cao, and Jin Tang. Fusing two-stream convolutional neural networks for rgb-t object tracking. *Neurocomputing*, 281:78–85, 2018.

[30] Chenglong Li, Chengli Zhu, Yan Huang, Jin Tang, and Liang Wang. Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 808–823, 2018.

[31] Chenglong Li, Chengli Zhu, Jian Zhang, Bin Luo, Xiaohao Wu, and Jin Tang. Learning local-global multi-graph descriptors for rgb-t object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[32] Chenglong Li, Chengli Zhu, Shaofei Zheng, Bin Luo, and Jing Tang. Two-stage modality-graphs regularized manifold ranking for rgb-t tracking. *Signal Processing: Image Communication*, 68:207–217, 2018.

[33] Yang Li and Jianke Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *European conference on computer vision*, pages 254–265. Springer, 2014.

[34] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-scale triplet cnn for person re-identification. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 192–196. ACM, 2016.

[35] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[36] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4293–4302, 2016.

[37] Eunbyung Park and Alexander C Berg. Meta-tracker: Fast and robust online adaptation for visual object trackers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 569–585, 2018.

[38] Pierre Sermanet, Soumith Chintala, and Yann LeCun. Convolutional neural networks applied to house numbers digit classification. *arXiv preprint arXiv:1204.3968*, 2012.

[39] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019.

[40] Yi Wu, Erik Blasch, Genshe Chen, Li Bai, and Haibin Ling. Multiple source data fusion via sparse representation for robust visual tracking. In *14th International Conference on Information Fusion*, pages 1–8. IEEE, 2011.

[41] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013.

[42] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015.

[43] Hui Zhang, Lei Zhang, Li Zhuo, and Jing Zhang. Object tracking in rgb-t videos using modal-aware attention network and competitive learning. *Sensors*, 20(2):393, 2020.

[44] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4106–4115, 2019.

[45] Yabin Zhu, Chenglong Li, Yijuan Lu, Liang Lin, Bin Luo, and Jin Tang. Fanet: Quality-aware feature aggregation network for rgb-t tracking. *arXiv preprint arXiv:1811.09855*, 2018.

[46] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. Dense feature aggregation and pruning for rgbt tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 465–472, 2019.

[47] Zheng Zhu, Wei Wu, Wei Zou, and Junjie Yan. End-to-end flow correlation tracking with spatial-temporal attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 548–557, 2018.