# Cross-Modal Object Tracking: Modality-Aware Representations and A Unified Benchmark

**Chenglong Li**[1,2,4], **Tianhao Zhu**[3*], **Lei Liu**[3*], **Xiaonan Si**[3], **Zilin Fan**[3], **Sulan Zhai**[5†]

[1]Information Materials and Intelligent Sensing Laboratory of Anhui Province, Hefei, China
[2]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Hefei, China
[3]School of Computer Science and Technology, Anhui University, Hefei, China
[4]School of Artificial Intelligence, Anhui University, Hefei, China
[5]School of Mathematical Sciences, Anhui University, Hefei, China
lcl1314@foxmail.com, zhutianhao79@gmail.com, liulei970507@163.com, 1065682519,1198791575@qq.com, 01044@ahu.edu.cn

## Abstract

In many visual systems, visual tracking often bases on RGB image sequences, in which some targets are invalid in low-light conditions, and tracking performance is thus affected significantly. Introducing other modalities such as depth and infrared data is an effective way to handle imaging limitations of individual sources, but multi-modal imaging platforms usually require elaborate designs and cannot be applied in many real-world applications at present. Near-infrared (NIR) imaging becomes an essential part of many surveillance cameras, whose imaging is switchable between RGB and NIR based on the light intensity. These two modalities are heterogeneous with very different visual properties and thus bring big challenges for visual tracking. However, existing works have not studied this challenging problem. In this work, we address the cross-modal object tracking problem and contribute a new video dataset, including 644 cross-modal image sequences with over 478K frames in total, and the average video length is more than 742 frames. To promote the research and development of cross-modal object tracking, we propose a new algorithm, which learns the modality-aware target representation to mitigate the appearance gap between RGB and NIR modalities in the tracking process. It is plug-and-play and could thus be flexibly embedded into different tracking frameworks. Extensive experiments on the dataset are conducted, and we demonstrate the effectiveness of the proposed algorithm in two representative tracking frameworks against 19 state-of-the-art tracking methods.

## Introduction

Visual tracking is an important problem in the field of computer vision and plays a critical role in many visual systems, such as visual surveillance, intelligent transportation, and robotics. However, existing tracking methods often base on RGB image sequences which are sensitive to illumination variations, and some targets are thus invalid in low-light conditions. In such scenarios, the tracking performance of existing methods might degrade significantly.

Some works introduce other modalities such as depth and infrared data to overcome imaging limitations of RGB source (Song and Xiao 2013; Li et al. 2016, 2019b). However, multi-modal imaging platforms usually require elaborate design and cannot be applied in many real-world applications at present. For example, the depth sensors can provide valuable additional depth information to improve tracking results by robust occlusion and model drift handling, but suffer from the limited range (e.g., 4-5 meters at most) and indoor environment (Song and Xiao 2013; Li et al. 2016). Thermal sensors are usually independent of RGB ones and their visual properties are very different. Therefore, a lot of efforts are needed in platform design and frames alignment (Li et al. 2016, 2019b).

Near-infrared (NIR) imaging becomes an essential part of many surveillance cameras, whose imaging is switchable between RGB and NIR based on the light intensity, as shown in Fig. 1(a). This kind of imaging system well handles imaging limitations of RGB source in low-light conditions while avoiding the imaging and platform problems introduced by existing multi-modal visual systems. From Fig. 1(b) we can also observe that these two modalities are heterogeneous with very different visual properties and the appearance of the target object is thus totally different in different modalities. Such an appearance gap brings big challenges for visual tracking, and existing tracking works have not studied this challenging problem.

In this work, we address the problem of cross-modal object tracking and aim to answer the following two questions. How to design a suitable algorithm, which could mitigate the appearance gap between RGB and NIR modalities and flexibly embedded into different tracking frameworks, for robust cross-modal object tracking? How to create a video benchmark dataset for the promotion of research and development of cross-modal object tracking?

First, we propose a **M**odality-**A**ware cross-**M**odal **O**bject **T**racking algorithm (MArMOT), which learns the modality-aware target representations to mitigate the appearance gap between RGB and NIR modalities in the tracking process. MArMOT is plug-and-play and could thus be flexibly embedded into different tracking frameworks. MArMOT in-

*Tianhao Zhu and Lei Liu contribute equally to the article.
†Sulan Zhai is the corresponding author.

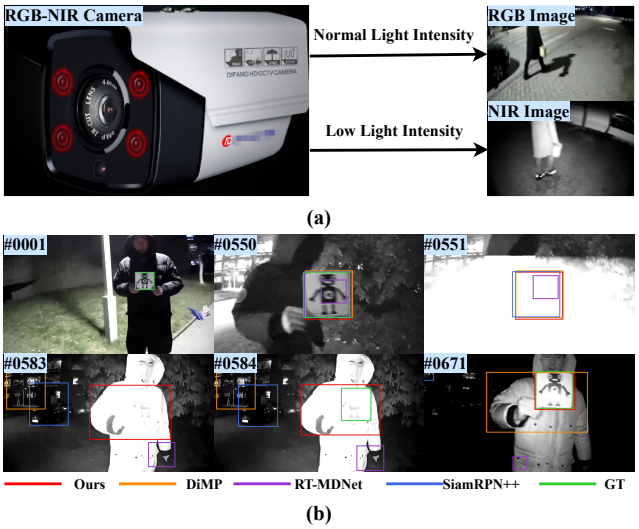



Figure 1: (a) Illustration of heterogeneous properties between RGB and NIR modalities. The visual camera changes RGB imaging to NIR when the light intensity becomes low from normal, and vice versa. (b) A comparison of our approach with state-of-the-art trackers in facing the challenge of modality switch, including DiMP (Bhat et al. 2019), SiamRPN++ (Li et al. 2019a) and RT-MDNet (Jung et al. 2018). The results show that our method handles this challenge well but the others trackers fail when the appearance of the target varies significantly caused by modality switch.

cludes two parallel CNN branches to learn modality-specific target representations using different sets of training samples. Besides, we do not know which modality appears in the tracking process. Thus, we design a ensemble module to adaptively incorporate effective features from both branches with any modality as input. In this way, the appearance gap between RGB and NIR modalities can be well addressed.

Second, to build a unified benchmark dataset, we collect 644 cross-modal object tracking sequences. The total number of video frames reaches over 478K, and the average video length and the maximum length of one sequence are more than 742 and 2037 frames. This dataset contains most of the real-world challenges in cross-modal object tracking task. Most importantly, it contains more challenges in adverse environmental conditions, as shown in Fig. 1(b), which easily triggers modality switch and significantly declines the capability of visual trackers.

The major contributions of this work can be summarized as follows. First, we introduce **a new task** called cross-modal object tracking that is very challenging but practical in many visual systems. Second, we propose **a novel algorithm** to mitigate the appearance gap of target object between different modalities for robust cross-modal object tracking, and integrate it into **two typical tracking frameworks** for effectiveness and generality validations. Third, we develop **a three-stage learning algorithm** to train the proposed tracking networks efficiently and effectively. Fourth, we create **a unified benchmark dataset** which contains most of the real-world challenges in cross-modal object tracking. Finally, we carry out **an extensive experiment** to demonstrate the effectiveness of the proposed approaches against the state-of-the-art trackers and clarify the research room on the cross-modal object tracking. Dataset, code, model and results are available at https://github.com/mmic-lcl/source-code.

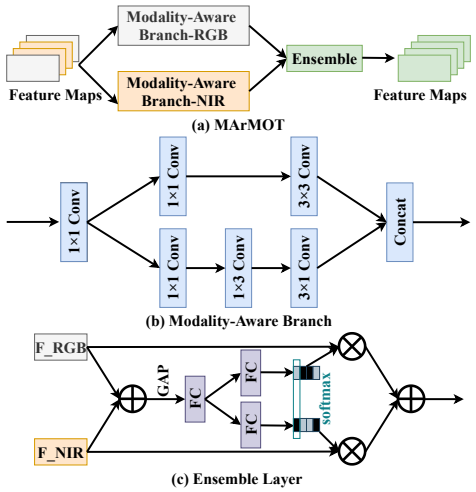



Figure 2: Details of MArMOT. The BN+ReLU layers after each Conv layer and the ReLU layers after each FC layer are omitted for clarity. Herein, $\oplus$ and $\otimes$ denote the operations of element-wise addition and multiplication respectively. F_RGB and F_NIR denote the output features of RGB and NIR after two parallel modality-aware branches respectively. GAP indicates the global average pooling.

## MArMOT Trackers

In this section, we first introduce the proposed **M**odality-**A**ware cross-**M**odal **O**bject **T**racking model (MArMOT), and then the tracking architectures with MArMOT including how embed the proposed plug-and-play MarMOT into two typical tracking frameworks. At last, the three-stage learning algorithm and the tracking details are provided.

### MArMOT Model

In the task of cross-modal object tracking, two modalities are heterogeneous with very different visual properties and thus bring big challenges for visual tracking. To solve this problem, we propose a new MArMOT which learns the modality-aware target representations to mitigate the appearance gap between RGB and NIR modalities in tracking process. Note that MArMOT is plug-and-play and can thus be flexibly embedded into different tracking frameworks.

MArMOT includes two parallel modality-aware branches to learn modality-specific target representations using different sets of training samples. Besides, we do not know which modality appears in the tracking process. Thus, we design an ensemble module to adaptively incorporate effective features from both branches with any modality as input. In this

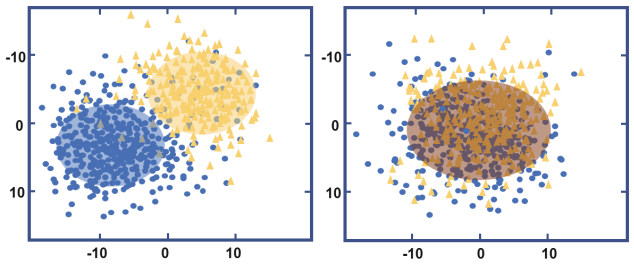



Figure 3: Visualization of target features in a sequence. (a) Projected features of the baseline tracker. (b) Projected features of our MArMOT with RT-MDNet as baseline. Herein, the blue circles and yellow triangles represent target features of RGB and NIR modalities respectively. From the results we can see that the heterogeneous gap of the target object between different modalities is mitigated to some extent.

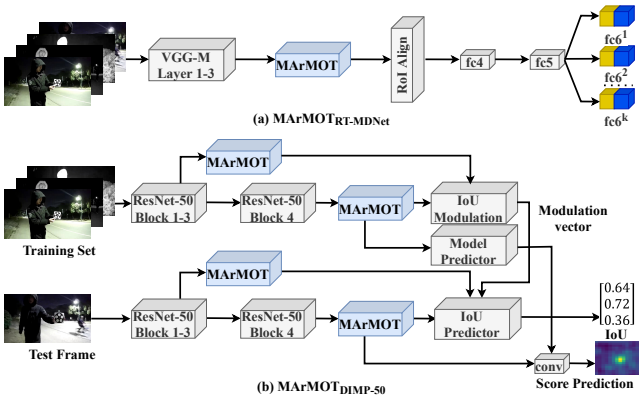



Figure 4: Visualization of the tracking architectures with MArMOT. (a) and (b) show the detailed structures of MArMOT combining with RT-MDNet and DiMP-50.

way, the appearance gap between RGB and NIR modalities can be well addressed, as shown in Fig. 2.

**Modality-Aware Branch**. Two parallel modality-aware branches are followed by backbone network, and used for learning modality-specific representations of the target in different modalities. As for the architecture of each branch, we use the inception-like network (Szegedy et al. 2016) for the effective and efficient computation. The details can be seen in Fig. 2(b). In each branch, the first $1 \times 1$ convolutional layer is used to capture modality-specific representations. Then it is divided into two flows by using another two $1 \times 1$ convolutional layers with half channels to decrease the dimensionality of the input feature and fed into two types of $3 \times 3$ convolution to increase the adaptability of the network to targets of different scales. Their outputs are concatenated together as the modality-specific representation.

**Ensemble Layer**. Due to the particularity of cross-modal object tracking, we design two parallel modality-aware branches to capture the modality-specific representations. However, in tracking process, we only have one modality as the input in each frame and do not know which modality is presented. To handle this problem, we design an ensemble layer to adaptively integrate features outputted from two branches given one modal input. By this way, we can obtain the effective features no matter which modality is input. Specifically, we utilize the SKNet (Li et al. 2019d) to fuse the features of the two parallel branches by weighting them using the normalized weights, and thus achieve adaptive fusion of these two branch features. The detailed design can be found in Fig. 2(c).

To visually demonstrate the effectiveness of our method, we present the features of an example obtained by the baseline tracker RT-MDNet (Jung et al. 2018) and by our tracker $\text{MArMOT}_{\mathbf{RT-MDNet}}$ after being projected to the 2D space via the t-SNE algorithm (Van der Maaten and Hinton 2008), respectively, as shown in Fig. 3. It can be found that the gap between target features of RGB and NIR can be eliminated well after the introduction of the proposed algorithm.

## Tracking Architectures with MArMOT

We embed the proposed plug-and-play MArMOT model into two tracking frameworks, i.e., RT-MDNet (Jung et al. 2018) and DiMP-50 (Bhat et al. 2019), named $\text{MArMOT}_{\mathbf{RT-MDNet}}$ and $\text{MArMOT}_{\mathbf{DiMP-50}}$ respectively, to verify the effectiveness and generalization of MArMOT. The overall tracking frameworks are shown in Fig. 4.

For each tracking framework, we first use the backbone network to extract deep feature representations of the target, then embed the proposed MArMOT model to mitigate the appearance gap of the target representations between different modalities, and finally send it to the classification branch and regression branch of target localization. Specifically, for RT-MDNet, it starts with several convolutional layers borrowed from VGG-M, which are used to capture common low-level information across modalities. Therefore, we insert MArMOT model after the last convolutional to learn the representations under corresponding modalities. In addition, such design can also reduce computational complexity since the size of feature map in last layer is smallest, which is shown in Fig. 4(a). For DiMP-50 tracker, the backbone network connects IoU predictor and model Predictor at the same time, which uses different layer features. Thus, we need to insert MArMOT modules after different layers for these two predictors, which is shown in Fig. 4(b). The consideration in reducing computational complexity is same with the design in RT-MDNet.

## Three-stage Learning Algorithm

There are two problems in training the entire tracking frameworks. First, the loss of a training sample with any modality will be backwardly propagated to two modality-aware branches. Thus, there is no guarantee that the two modality-aware branches will learn the corresponding modality-specific representation of the target. Second, the modality information is available in training stage but unavailable in testing stage. Therefore, we need to train an ensemble layer to simulate the modal agnostic situation in tracking process. To handle these two problems, we design an effective three-

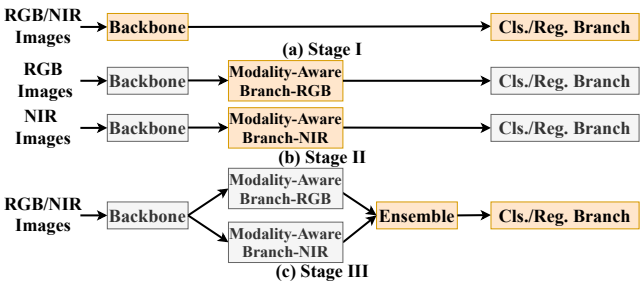



Figure 5: The visualization of three-stage training method. The parameters learned in each stage are shown in orange.

stage training algorithm.

- Stage I : Fine-tune the parameters of the baseline network on our dataset. Note that our dataset is the first cross-modal tracking dataset. To adapt the tracker to the cross-modal scenario, we first need to fine-tune the parameters of the baseline network pre-trained on other large-scale datasets on our training set. The learning rate of the network parameters is set to one-tenth of the default learning rate of the baseline network, and the number of iterations remains the same.
- Stage II : Train two parallel modality-aware branches. To enable two parallel modality-aware branches to learn the modality-specific representations of the target in different modalities, we first divide the training set into two subsets according to modality type, and use the corresponding sub-datasets to learn the parameters of the corresponding modality-aware branches. Since the baseline network has been adapted to the cross-modal tracking task in the first stage, thus, in this stage, we only learn the parameters of the two modality-aware branches, and the rest of the parameters are fixed (except for the parameters of the fc6 layer based on the RT-MDNet (Jung et al. 2018) framework). The initial learning rate is set to $1e-6$ and $1e-4$, and the number of iterations is set to $50$ and $1000$, respectively.
- Stage III : Train ensemble layer and fine-tune the parameters of the baseline network on our dataset again. After the first two stages of training, the baseline network has been able to adapt to the tracking of cross-modal scenarios, and the two parallel modality-aware branches have also learned the modality-specific representations of the target in different modalities. Since which modality in each frame is unknown in the tracking process, the deep features extracted by the backbone need to be sent to two parallel modality-aware branches to extract the corresponding modality-specific representations. To simulate the modality-unknowable situation in the tracking process, we train the ensemble layer at this stage to perform weighted fusion of the features of the two branches, and fine-tune the parameters of the network to adapt to the situation after embedding the proposed MArMOT. At this stage, we only learn the parameters of the ensemble layer and fine-tune the parameters of the baseline network except the backbone. The learning rate of the ensemble layer is the same as the modality-aware branch of the second stage, and the learning rate of the tracker is the same as the first stage, and the number of iterations is set to the same as the second stage.

Fig. 5 shows more details, and the learned part is indicated in the orange color.

| Benchmark | Video | Min frames | Mean frames | Max frames | Total frames |
|---|---|---|---|---|---|
| **CMOTB (train)** | **430** | **85** | **755** | **2037** | **325K** |
| **CMOTB (test)** | **214** | **101** | **721** | **1838** | **153K** |

Table 1: The details of our CMOTB Dataset

### Online Tracking

The tracking processes and parameter settings of our trackers during online tracking are the most same as the baseline trackers. The only difference is that the deep features extracted by the backbone networks (VGG-M in $\text{MArMOT}_{\mathbf{RT-MDNet}}$ and ResNet$50$ in $\text{MArMOT}_{\mathbf{DiMP-50}}$) are sent to the proposed MArMOT model to mitigate the appearance gap of the target representations under different modalities. The outputs of MArMOT are as the inputs of the classifiers ($\text{fc}4-\text{fc}6$ in $\text{MArMOT}_{\mathbf{RT-MDNet}}$ and model predictor in $\text{MArMOT}_{\mathbf{DiMP-50}}$) and the regressor (IoU predictor module in $\text{MArMOT}_{\mathbf{DiMP-50}}$). The details of tracking processes of our trackers are shown in Fig. 4.

## CMOTB Benchmark

Large-scale dataset are crucial in cross-modal object tracking because they are not only useful for training deep trackers, but also for evaluating different tracking algorithms. To this end, we provide a large-scale cross-modal object tracking benchmark, called CMOTB. In this section, we introduce CMOTB with detailed analysis.

### Data Collection and Annotation

**Large-scale collection.** Current object tracking field lacks cross-modal video data, and we introduce our CMOTB benchmark. Our goal is to provide large-scale and high-diverse cross-modal object tracking benchmark for real-world scenarios and challenges. To this end, we use hand-held cameras to capture video data in a large range of scenes and background complexities. Unlike traditional visual tracking data, we need to consider the variations of light intensity that trigger modality switch in data creation. Therefore, we carefully select some environmental conditions to simulate real-world applications such as visual surveillance, intelligent transportation and self-driving systems. Fig. 1 shows a typical example of CMOTB dataset, and we can see that the imaging is switching with several times between RGB and NIR modalities. By this way, we collect $644$ cross-modal image sequences with over $478$K frames in total and the average video length of is more than $742$ frames.

We present the detailed information of CMOTB in Table 1. Note that there is no other cross-modal object tracking dataset, therefore, we divide CMOTB into training set

| Modality switch | 1 | 2 | 3 | >=4 |
|---|---|---|---|---|
| Number of sequences | 410 | 181 | 48 | 15 |

Table 2: Distribution of the number of modality switch.

and testing set for facilitating the training of deep trackers for cross-modal object tracking.

**High-quality dense annotation.** We use a minimum bounding box to represent object states including position and scale, and annotate each frame for training and evaluation set. Since the labeling process is time-consuming and labor-intensive, we design an auxiliary labeling tool based on the ViTBAT (Biresaw et al. 2016). The tool allows the labeling of their states manually or semi-automatically through a simple and friendly user interface in a time-efficient manner. The generated bounding boxes are accurate in most situations. However, when the object undergoes drastic appearance variations, the generated bounding boxes might be not quite accurate. For these bounding boxes, we manually adjust them carefully.

To ensure high-quality annotations, we train $4$ professional annotators to learn consistent annotation standards. Moreover, we let professional checkers to carry out a frame-by-frame inspection to prevent wrong and inaccurate marking. Due to the special challenges brought by modality switch, some objects are sometimes temporarily invisible, which may result in losing a few frames or more than a dozen frames. For such scenarios, we will keep ground truths unchanged of target object until it is visible.

## Attributes

Existing multi-modal tracking datasets, e.g., RGBD (Song and Xiao 2013) and RGBT (Li et al. 2016, 2019b), include two-modal data in each frame, while our dataset has only one modality in each frame but might occur modality switch. This is the major difference from existing multi-modal tracking datasets. The modality switch means that the imaging is changed from one modality to another one caused by light intensity variation. In such scenarios, the appearance of target object usually varies significantly so that trackers are easily failed. Note that the number of modality switch in a sequence is a key factor in affecting trackers. Therefore, we take the switch times in data creation and report the data distribution on switch times in Table 2.

According the modality switch, a new attribute, i.e., modality adaptation, is thus introduced in CMOTB. The modality adaptation means that some frames have high intensity due to imaging adaptation to the environment in modality switch. It does not always occur when imaging is switch, and thus we take it as an attribute. To enable attribute-based performance analysis of trackers, we annotate each sequence with several attributes from the total $11$ attributes, including Scale Variation (SV), Aspect Ratio Change (ARC), Fast Motion (FM), Out-of-View (OV), Modality Adaptation (MA), Motion Blur (MB), Background Clutter (BC), Similar Object(SO), In-Plane Rotation (IPR), Partial Occlusion (PO) and Full Occlusion (FO). The Table 3 shows the video distribution of attributes on the testing set.

## Statistics

CMOTB consists of $644$ video sequences, which cover most of the challenges in real-world scenarios. According to the holdout method (Kohavi et al. 1995), we randomly split the testing and training sets of our dataset with the ratio of $1:2$. And We have counted the distribution of attributes on testing set in Table 3. The total number of frames of CMOTB reaches $478$K, and the average length of our video sequence and the maximum number of frames reach $742$ and $2037$ frames respectively. More details are shown in Table 1.

## Discussion

**Differences from relevant tasks**. We discuss the differences of our new task from the task of multi-modal visual object tracking. Existing work usually introduce thermal infrared or depth information to achieve multi-modal visual object tracking, called RGBT tracking (Li et al. 2019b, 2020) and RGBD tracking (Song and Xiao 2013). Comparing with multi-modal visual object tracking, our task has the following differences and advantages. First, our task is more practical. Many visual cameras have equipped with NIR imaging, but RGBT or RGBD tracking requires two cameras. Second, our task is more cost-effective. Thermal cameras are usually very expensive and depth sensors have limited imaging range and environment, but our task only relies on surveillance cameras and thus does not have these limitations. Finally, the multi-modal data in our task do not have any alignment error. Both RGBT and RGBD tracking tasks involve two cameras and the alignment cross different modalities is needed, while our imaging system only includes one camera whose imaging is switchable between RGB and NIR modalities.

**Acquisition of modality switch signals**. To the best of our knowledge, the commercial sensors do not provide signals of modality switch, but they might be obtained by customizing the sensor. Therefore, the study to handling the scenario of unknown modality switch is essential.

# Experiment

The experiments are run on two platforms with Intel(R) Xeon(R) Silver $4210$ CPU (32G RAM), GeForce RTX $3090$ GPU and Intel(R) Xeon(R) Silver $4210$ CPU (32G RAM), GeForce RTX $1080$Ti GPU for DiMP-$50$ and RT-MDNet respectively. Our $\text{MArMOT}_{\mathbf{DiMP-50}}$ and $\text{MArMOT}_{\mathbf{RT-MDNet}}$ perform 25 FPS and 24 FPS respectively.

## Evaluated Algorithms

We evaluate $19$ most advanced and representative trackers on our benchmark. These trackers cover mainstream tracking algorithms from $2016$ to $2020$, and they are MDNet (Nam and Han 2016), RT-MDNet (Jung et al. 2018), SiamFC (Bertinetto et al. 2016), SPLT (Yan et al. 2019), GradNet (Li et al. 2019c), TACT (Choi, Kwon, and Lee 2020), SiamMask (Wang et al. 2019), VITAL (Song et al. 2018), GlobalTrack (Huang, Zhao, and Huang 2020), SiamRPN++ (Li et al. 2019a), ATOM (Danelljan et al. 2019), DiMP-$50$(Bhat et al. 2019), SiamBAN (Chen et al.

| Attribute | SV | BC | ARC | SO | FM | IPR | OV | PO | MA | FO | MB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Number | **35** | **59** | **17** | **64** | **19** | **126** | **25** | **104** | **97** | **40** | **47** |

Table 3: Distribution of attribute-based sequences on CMOTB testing set.

2020), SiamDW (Zhang and Peng 2019), LTMU (Dai et al. 2020), TransT (Chen et al. 2021), TrDiMP (Wang et al. 2021), Ocean (Zhang et al. 2020) and DaSiamRPN (Zhu et al. 2018). Note that all algorithms are evaluated on our testing set using the model provided by authors.

## Evaluation Metrics

To evaluate the performance of different trackers, we employ the widely used tracking evaluation metrics including precision rate (PR), normalized precision rate (NPR) and success rate (SR) (Muller et al. 2018) for quantitative performance evaluation. PR is the percentage of frames whose distance of the estimated bounding boxes with the ground truth is below a predefined threshold, to rank the trackers, we set the distance threshold to 20 pixels to compute the representative PR. However, since the PR is very sensitive to the target size, thus, we normalize the PR on the size of the ground truth to calculate the normalized precision rate, distance threshold is also set to 20 pixels to compute the representative NPR. SR is the percentage of frames where the overlap rate between the estimated bounding boxes and the ground truth are greater than a threshold, we set the overlap ratio to $0.5$ and use the area under curve of SR plots to compute two kinds of representative SR scores respectively, denoting SR-I score and SR-II score for the clarity.

## Overall Performance

We present the tracking performance in terms of precision plots, normalized precision plots and success plots in Fig. 6, and the representative scores are shown in the legends.

**Regression based deep trackers**. Regression-based trackers such as DiMP-50, LTMU, SiamRPN++, SiamBAN, ATOM, SiamMask, have achieved high performance while running at real-time speed. They are usually offline trained to learn a powerful regressor from large-scale datasets to locate the target. However, their performance is limited in cross-modal object tracking due to the existence of large heterogeneous gap between RGB and NIR modalities, as shown in Fig. 6. To verify the effectiveness of the proposed method, we insert the proposed MArMOT into the DiMP-50 tracking framework, namely MArMOT$_{\mathbf{DiMP-50}}$, and train the entire framework through the proposed multi-stage training method. Our MArMOT$_{\mathbf{DiMP-50}}$ outperforms the baseline tracker DiMP-50 with $15.9\%$/$14.1\%$/$17.8\%$/$13.7\%$ gains in PR/NPR/SR-I/SR-II, and has excellent performance gains compared with all comparison methods.

**Classification based deep trackers**. Classification based deep trackers such as MDNet, RT-MDNet and VITAL, usually employ online learning to train binary classifiers using positive and negative samples, and thus have a good generalization ability. To verify the effectiveness and generalization of the proposed method, we also insert MArMOT into the RT-MDNet framework, namely MArMOT$_{\mathbf{RT-MDNet}}$, and train the entire framework through the proposed multi-stage training method. And our MArMOT$_{\mathbf{RT-MDNet}}$ outperforms the baseline tracker RT-MDNet with $16.1\%$/$16.8\%$/$21.5\%$/$15.5\%$ gains in PR/NPR/SR-I/SR-II. We can find that although the performance of the baseline tracker is very low, but it surpasses the performance of all classification-based tracking frameworks after introducing our proposed model, which proves the effectiveness of our proposed method.

| Trackers | PR | NPR | SR-I | SR-II |
|---|---|---|---|---|
| **DiMP-50*** | 0.715 | 0.692 | 0.777 | 0.651 |
| **RT-MDNet*** | 0.517 | 0.523 | 0.534 | 0.444 |
| **LTMU*** | 0.561 | 0.563 | 0.628 | 0.526 |
| **SiamRPN++*** | 0.574 | 0.552 | 0.612 | 0.503 |
| **SiamMask*** | 0.556 | 0.529 | 0.571 | 0.486 |
| **MDNet*** | 0.534 | 0.526 | 0.547 | 0.461 |
| **GlobalTrack*** | 0.495 | 0.482 | 0.555 | 0.477 |
| **MArMOT$_{\mathbf{DiMP-50}}$** | 0.731 | 0.705 | 0.803 | 0.671 |
| **MArMOT$_{\mathbf{RT-MDNet}}$** | 0.566 | 0.574 | 0.603 | 0.490 |

Table 4: Comparison of deep trackers re-trained on CMOTB dataset, where * indicates the tracker is re-trained using CMOTB training dataset

| Trackers | PR | NPR | SR-I | SR-II |
|---|---|---|---|---|
| **DiMP-50-RGBT** | 0.681 | 0.468 | 0.523 | 0.446 |
| **DiMP-50*-RGBT** | 0.757 | 0.507 | 0.602 | 0.502 |
| **MArMOT$_{\mathbf{DiMP-50}}$-RGBT** | 0.767 | 0.534 | 0.639 | 0.529 |

Table 5: Tracking results of trackers on the synthetic RGBT dataset, where * indicates the tracker is re-trained using synthetic RGBT training dataset.

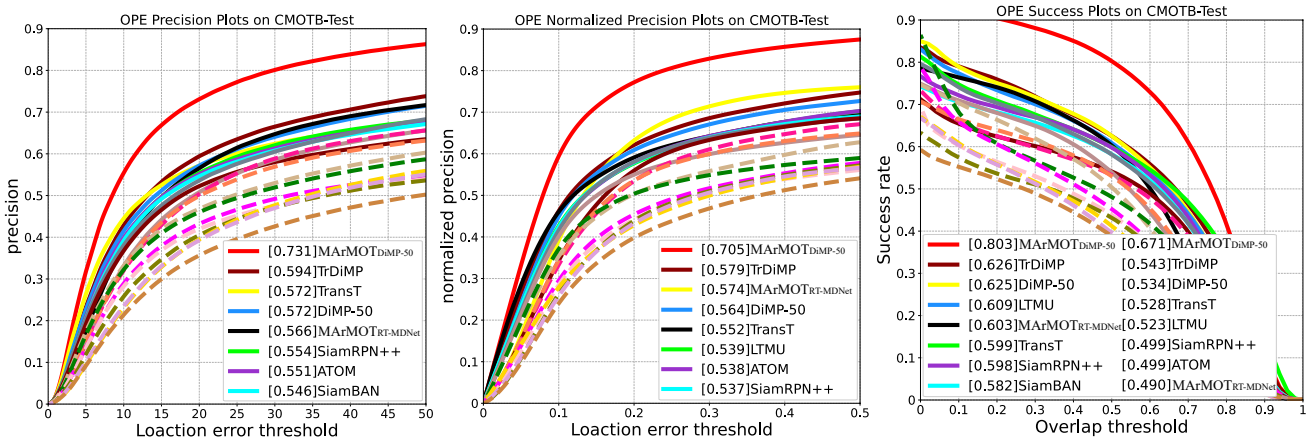
## Training Dataset Validation

We select seven representative trackers including DiMP-50, RT-MDNet, LTMU, SiamRPN++, SiamMask, MDNet and GlobalTrack to demonstrate the effectiveness of our training dataset in the training of deep models. The results are shown in Table 4, which shows that all the re-trained deep trackers achieve obvious improvements and verify the necessity of proposing this dataset for the study of cross-modal object tracking. In addition, after adding our proposed model to the DiMP-50 and RT-MDNet frameworks, we can see that the performance has been further improved, i.e., $1.6\%$/$1.3\%$/$2.6\%$/$2.0\%$ and $4.9\%$/$5.1\%$/$6.9\%$/$4.6\%$ gains on the PR/NPR/SR-I/SR-II respectively, which proves the effectiveness of MArMOT.

# Analysis of MArMOT

## Evaluation on Synthetic Data

To further validate the effectiveness of our MArMOT model, we construct a synthetic RGB-thermal cross-modal




(a) (b) (c)

Figure 6: Tracking curves of 19 trackers on the CMOTB testing dataset. For clarity, we only show the performance of the top 8 trackers in the legends. In (c), we show two kinds of representative SR scores in the legends, and the left and right ones are SR-I and SR-II respectively.

dataset from existing RGBT datasets, including GTOT (Li et al. 2016) and RGBT234 (Li et al. 2019b). To simulate cross-modal tracking tasks more accurately, each frame in synthetic videos is generated by selecting one modality from corresponding RGB and thermal images, according to the challenge labels of illumination various and thermal crossover. In particular, each sequence starts from RGB modality, and switches to another modality only when modality-specific challenges occur, including illumination various and thermal crossover.

For this synthetic dataset, we choose RGBT234 dataset as the training set and GTOT dataset as the testing set, and retrain the entire network for the cross-modal RGBT tracking task with proposed three-stage training method. The experiment results are shown in Table 5. It can be seen from the results that our MArMOT model can well deal with the appearance gap between RGB and thermal modalities in tracking process, which further prove the generalization and effectiveness of our method in handling the different cross-modal tracking tasks.

### Effectiveness of Modality-Aware Representations

To verify the effectiveness of the proposed modality-aware representations and proposed three-stage training method, we implement the variant trackers, named MArMOT$_{\mathbf{DiMP-50}}$-one-stage, by using one-stage training method to train the entire networks together on the CMOTB.

| Trackers | PR | NPR | SR-I | SR-II |
|---|---|---|---|---|
| **DiMP-50** | 0.572 | 0.564 | 0.625 | 0.534 |
| **DiMP-50*** | 0.715 | 0.692 | 0.777 | 0.651 |
| **MArMOT$_{\mathbf{DiMP-50}}$-one-stage** | 0.718 | 0.694 | 0.784 | 0.659 |
| **MArMOT$_{\mathbf{DiMP-50}}$** | 0.731 | 0.705 | 0.803 | 0.671 |

Table 6: Comparison of several variants of our MArMOT, where * indicates the tracker is re-trained using CMOTB training dataset.

The results are shown in Table 6. The experimental results show that proposed three-stage training method outperforms the one-stage training method with 1.3%/1.1%/1.9%/1.2% gains in PR/NPR/SR-I/SR-II, which can prove that the proposed three-stage learning method is benefit to modality-aware branches to learn corresponding modality-specific target representations. In addition, we can also find that the performance of the one-stage training method is still better than DiMP-50*, which can verify the effectiveness of the proposed MArMOT for mining cross-modal information.

## Conclusion

We provide a large-scale cross-modal object tracking benchmark with high-quality dense bounding box annotations. And we also propose a simple yet effective method based on a modality-aware feature learning algorithm for cross-modal object tracking purpose. Extensive experiments on the dataset demonstrate the effectiveness of the proposed method against state-of-the-art trackers. By releasing this dataset, we believe that it will help the research and developments of cross-modal object tracking. In the future, we will study more effective tracking algorithms to solve the cross-modal tracking problem and extend the dataset to cover more real-world scenarios.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61976003, 61876002, 62176085), the Natural Science Foundation for the Higher Education Institutions of Anhui Province (No. KJ2020A0061, KJ2019A0005), the China Postdoctoral Science Foundation (No. 2020M681989), and the Anhui Energy Internet Joint Fund Project (No. 2008085UD07).

## References


Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016. Fully-convolutional siamese networks for object tracking. In *Proceedings of the European Conference on Computer Vision*, 850–865. Springer.

Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 6182–6191.

Biresaw, T. A.; Nawaz, T.; Ferryman, J.; and Dell, A. I. 2016. Vitbat: Video tracking and behavior annotation tool. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, 295–301. IEEE.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8126–8135.

Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; and Ji, R. 2020. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6668–6677.

Choi, J.; Kwon, J.; and Lee, K. M. 2020. Visual tracking by tridentalign and context embedding. In *Proceedings of the Asian Conference on Computer Vision*.

Dai, K.; Zhang, Y.; Wang, D.; Li, J.; Lu, H.; and Yang, X. 2020. High-performance long-term tracking with meta-updater. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6298–6307.

Danelljan, M.; Bhat, G.; Khan, F. S.; and Felsberg, M. 2019. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4660–4669.

Huang, L.; Zhao, X.; and Huang, K. 2020. Globaltrack: A simple and strong baseline for long-term tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11037–11044.

Jung, I.; Son, J.; Baek, M.; and Han, B. 2018. Real-time mdnet. In *Proceedings of the European Conference on Computer Vision*, 83–98.

Kohavi, R.; et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 14, 1137–1145. Montreal, Canada.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019a. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4282–4291.

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

Li, C.; Liang, X.; Lu, Y.; Zhao, N.; and Tang, J. 2019b. RGB-T object tracking: benchmark and baseline. *Pattern Recognition*, 96: 106977.

Li, C.; Liu, L.; Lu, A.; Ji, Q.; and Tang, J. 2020. Challenge-aware rgbt tracking. In *Proceedings of the European Conference on Computer Vision*, 222–237. Springer.

Li, P.; Chen, B.; Ouyang, W.; Wang, D.; Yang, X.; and Lu, H. 2019c. GradNet: Gradient-guided network for visual object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 6162–6171.

Li, X.; Wang, W.; Hu, X.; and Yang, J. 2019d. Selective kernel networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 510–519.

Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; and Ghanem, B. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision*, 300–317.

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

Song, S.; and Xiao, J. 2013. Tracking revisited using RGBD camera: Unified benchmark and baselines. In *Proceedings of the IEEE International Conference on Computer Vision*, 233–240.

Song, Y.; Ma, C.; Wu, X.; Gong, L.; Bao, L.; Zuo, W.; Shen, C.; Lau, R. W.; and Yang, M.-H. 2018. Vital: Visual tracking via adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8990–8999.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Wang, N.; Zhou, W.; Wang, J.; and Li, H. 2021. Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1571–1580.

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

Yan, B.; Zhao, H.; Wang, D.; Lu, H.; and Yang, X. 2019. 'Skimming-Perusal'Tracking: A framework for real-time and robust long-term tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 2385–2393.

Zhang, Z.; and Peng, H. 2019. Deeper and wider siamese networks for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4591–4600.

Zhang, Z.; Peng, H.; Fu, J.; Li, B.; and Hu, W. 2020. Ocean: Object-aware anchor-free tracking. In *Proceedings of the European Conference on Computer Vision*, 771–787. Springer.

Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; and Hu, W. 2018. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European Conference on Computer Vision*, 101–117.