WED-PM-122

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

# Bridging Search Region Interaction with Template for RGB-T Tracking

Tianrui Hui[1,2,4] Zizheng Xun[3,5] Fengguang Peng[3,5] Junshi Huang[4]
Xiaoming Wei[4] Xiaolin Wei[4] Jiao Dai[1,2] Jizhong Han[1,2] Si Liu[3,5]

[1]IIE, CAS [2]SCS, UCAS [3]IAI, BUAA [4]Meituan [5]HII, BUAA

中国科学院信息工程研究所 INSTITUTE OF INFORMATION ENGINEERING,CAS
中国科学院大学 University of Chinese Academy of Sciences
北京航空航天大学
美团

# Preview

# Bridging Search Region Interaction with Template for RGB-T Tracking

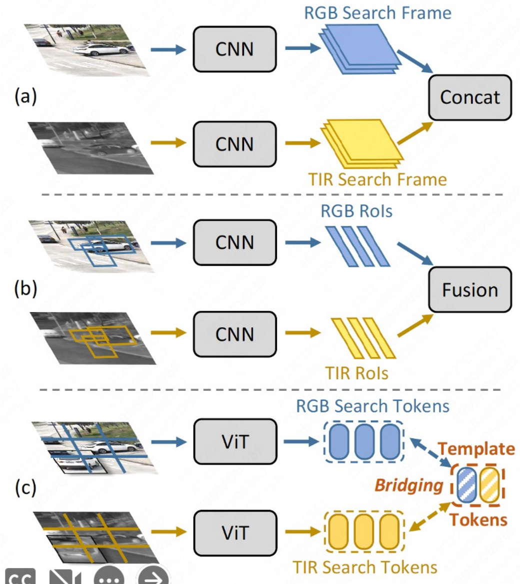
Tianrui Hui[1,2,4] Zizheng Xun[3,5] Fengguang Peng[3,5] Junshi Huang[4] Xiaoming Wei[4] Xiaolin Wei[4] Jiao Dai[1,2] Jizhong Han[1,2] Si Liu[3,5]

[1]IIE, CAS [2]SCS, UCAS [3]IAI, BUAA [4]Meituan [5]HII, BUAA

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

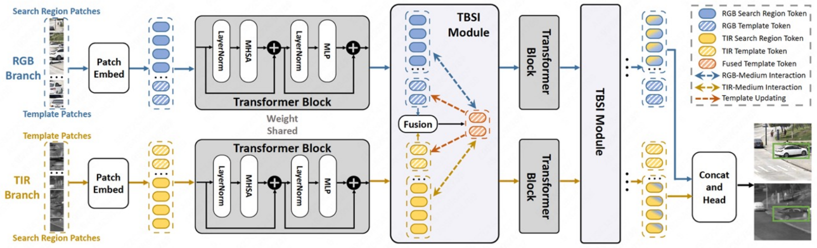
## Motivation

- As a multimodal vision task, the key to RGB-T tracking is how to perform effective cross-modal interaction
- Some previous methods **concatenate the RGB and TIR search region features directly** to perform a coarse interaction process with redundant background noises introduced
- Many other methods conduct fusion on **isolated pairs of RGB and TIR boxes**, which limits the cross-modal interaction within local regions and brings about inadequate context modeling
- We exploit **templates as the medium to bridge the cross-modal interaction between RGB and TIR search regions** by gathering and distributing target-relevant object and environment contexts




## Framework

- We extend the ViT architecture for **joint feature extraction, search-template matching, and cross-modal interaction**
- In TBSI module, bidirectional RGB and TIR search region interaction are bridged by the fused template, which serves as a medium to gather and distribute target-relevant contexts
- Original templates are also updated with the template medium




## Template-Bridged Search Region Interaction

- RGB and TIR template fusion

$$Z_m = [Z_r; Z_t]W_m,$$

- Bidirectional Template-Bridged Interaction
- Fused template gathers contexts from the TIR search region

$$D_t = \mathrm{Softmax}(\frac{(Z_m W_q^1)(X_t W_k^1)^{\mathrm{T}}}{\sqrt{C}})(X_t W_v^1),$$

$$Z'_m = \mathrm{LN}(Z_m + D_t),$$

$$\tilde{Z}_m = \mathrm{LN}(Z'_m + \mathrm{MLP}(Z'_m)),$$

$$D_{mt} = \mathrm{Softmax}(\frac{(X_r W_q^2)(\tilde{Z}_m W_k^2)^{\mathrm{T}}}{\sqrt{C}})(\tilde{Z}_m W_v^2).$$

- Gathered contexts are distributed to RGB search region

## Experiments

- Extensive ablation studies demonstrate the effectiveness of the components of our proposed method

| Layers 4 | Layers 7 | Layers 10 | Precision | NormPrec | Success |
|---|---|---|---|---|---|
| | | | 53.5 | 49.1 | 42.5 |
| ✓ | | | 60.5 | 56.9 | 47.8 |
| ✓ | ✓ | | 62.7 | 59.2 | 49.8 |
| ✓ | ✓ | ✓ | **63.8** | **60.2** | **50.6** |

| Method | Precision | NormPrec | Success |
|---|---|---|---|
| RGB Baseline | 50.1 | 45.4 | 40.1 |
| RGB-T Baseline | 53.5 | 49.1 | 42.5 |
| w/o Template Bridging | 59.6 | 55.9 | 47.4 |
| w/o RGB→TM→TIR | 58.7 | 55.1 | 46.6 |
| w/o Template Updating | 62.7 | 58.9 | 49.7 |
| Full Model (TBSI) | **63.8** | **60.2** | **50.6** |

| | APFNet† [39] | CMPP [37] | mfDiMP* [44] | TBSI |
|---|---|---|---|---|
| NO | 93.4/66.4 | 95.6/67.8 | **96.2**/69.4 | 96.1/**72.8** |
| PO | 85.0/58.7 | 85.5/60.1 | 86.6/60.9 | **88.7/64.7** |
| HO | 72.9/49.0 | 73.2/50.3 | 76.1/53.2 | **81.5/58.6** |
| LI | 82.3/54.4 | 86.2/58.4 | 84.2/58.0 | **89.2/63.6** |
| LR | 82.9/54.8 | **86.5**/57.1 | 82.1/53.0 | 85.1/**60.0** |
| TC | 82.1/57.3 | 83.5/58.3 | 84.8/58.9 | **85.8/63.2** |
| DEF | 77.1/54.6 | 75.0/54.1 | 81.5/60.2 | **84.1/63.7** |
| FM | 78.2/49.2 | 78.6/50.8 | 77.3/54.8 | **81.4/58.7** |
| SV | 82.1/56.5 | 81.5/57.2 | 87.1/63.7 | **89.9/66.8** |
| MB | 72.8/53.0 | 75.4/54.1 | 80.1/58.0 | **88.1/64.9** |
| CM | 76.3/54.5 | 75.6/54.1 | 84.0/60.3 | **88.0/65.0** |
| BC | 80.6/52.4 | 83.2/53.8 | 82.8/53.7 | **83.4/57.8** |

- Quantitative and qualitative comparison with state-of-the-art methods on three RGB-T tracking benchmarks

LasHeR

| | Method | Backbone | Pretraining | Precision | NormPrec | Success | FPS |
|---|---|---|---|---|---|---|---|
| Online | DAPNet [49] | VGG-M | ImageNet | 43.1 | 38.3 | 31.4 | - |
| | FANet [50] | VGG-M | ImageNet | 44.1 | 38.4 | 30.9 | - |
| | DAFNet [14] | VGG-M | ImageNet | 44.8 | 39.0 | 31.1 | 20.5 |
| | CAT [20] | VGG-M | ImageNet | 45.0 | 39.5 | 31.4 | - |
| | MANet [21] | VGG-M | ImageNet | 45.5 | - | 32.6 | 2.1 |
| | MANet++ [27] | VGG-M | ImageNet | 46.7 | 40.4 | 31.4 | - |
| | MaCNet [43] | VGG-M | ImageNet | 48.2 | 42.0 | 35.0 | 1.6 |
| | DMCNet [28] | VGG-M | ImageNet | 49.0 | 43.1 | 35.5 | - |
| | APFNet [39] | VGG-M | ImageNet | 50.0 | 43.9 | 36.2 | 1.9 |
| | mfDiMP [44] | ResNet-50 | SOT | 59.9 | - | 46.7 | 34.6 |
| Offline | TBSI | ViT-Tiny | ImageNet | 61.7 | 57.8 | 48.9 | **40.3** |
| | TBSI | ViT-Small | ImageNet | 62.4 | 58.6 | 49.4 | 39.1 |
| | TBSI | ViT-Base | ImageNet | 63.8 | 60.2 | 50.6 | 36.2 |
| | TBSI | ViT-Base | SOT | **69.2** | **65.7** | **55.6** | 36.2 |

RGBT234

| | Method | Precision | Success |
|---|---|---|---|
| Online | MDNet+RGBT [32] | 72.2 | 49.5 |
| | MaCNet [43] | 76.4 | 53.2 |
| | DAPNet [49] | 76.6 | 53.7 |
| | MANet [21] | 77.7 | 53.9 |
| | HDINet [30] | 78.3 | 55.9 |
| | FANet [50] | 78.7 | 55.3 |
| | JMMAC [46] | 79.0 | 57.3 |
| | M5L [35] | 79.5 | 54.2 |
| | MANet++ [27] | 79.5 | 55.9 |
| | DAFNet [14] | 79.6 | 54.4 |
| | CAT [20] | 80.4 | 56.1 |
| | ADRNet [45] | 80.7 | 57.0 |
| | CMPP [37] | 82.3 | 57.5 |
| | APFNet [39] | 82.7 | 57.9 |
| | DMCNet [28] | 83.9 | 59.3 |
| | mfDiMP [44] | 84.2 | 59.1 |
| Offline | SiamCDA [47] | 76.0 | 56.9 |
| | SiamIVFN [16] | 81.1 | 63.2 |
| | TBSI | **87.1** | **63.7** |

RGBT210

| | Method | Precision | Success |
|---|---|---|---|
| Online | TFNet [51] | 77.7 | 52.9 |
| | CAT [20] | 79.2 | 53.3 |
| | DMCNet [28] | 79.7 | 55.5 |
| | mfDiMP* [44] | 84.9 | 59.3 |
| Offline | DSiamMFT [48] | 64.2 | 43.2 |
| | TBSI | **85.3** | **62.5** |

(a) redcarcominginlight (b) rightdarksingleman (c) umbrella (d) ab_blkskirtgirl

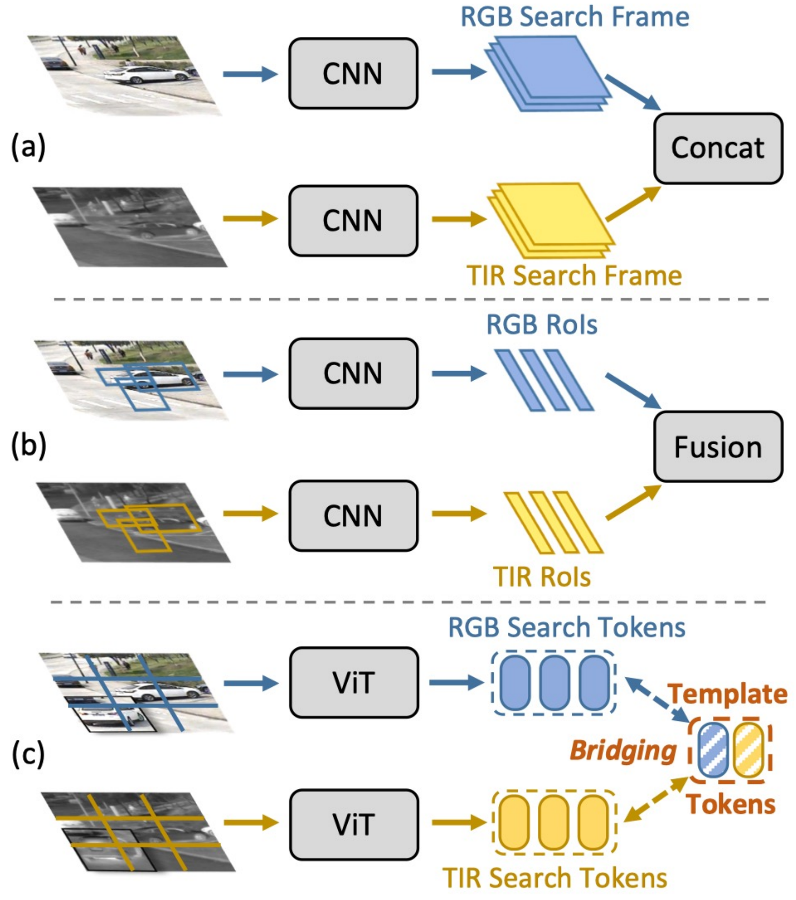
GT TBSI APFNet MANet++ CAT FANet MaCNet DAFNet DMCNet mfDiMP

# Task

- RGB-T tracking aims to leverage the mutual enhancement and complement ability of RGB and TIR modalities for improving the tracking process in various scenarios

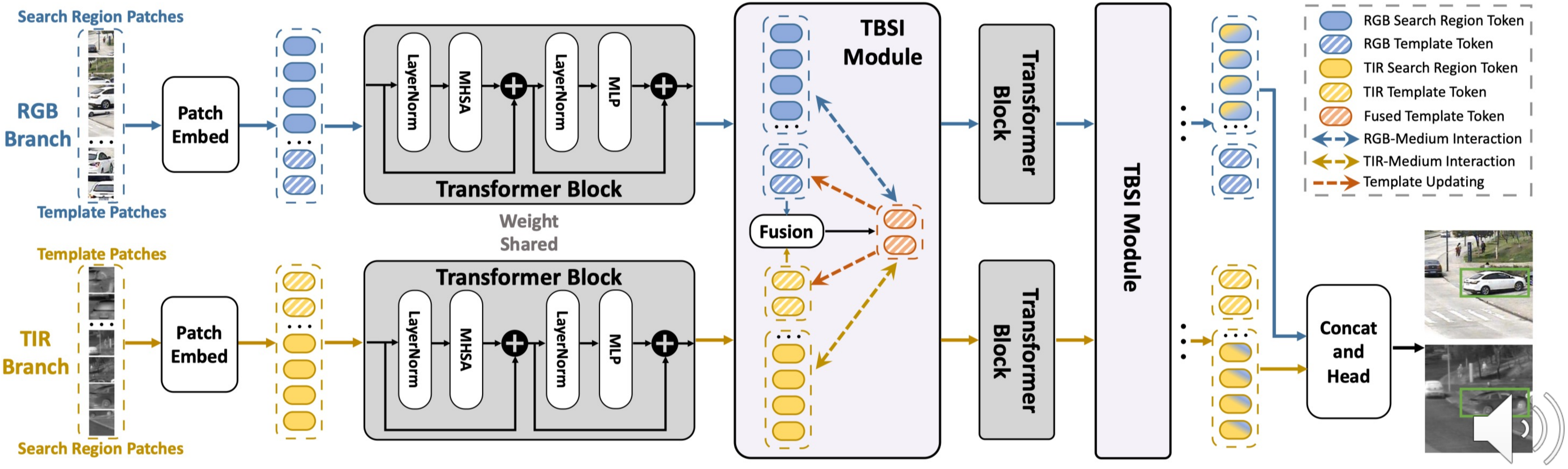
# Motivation

- As a multimodal vision task, the key to RGB-T tracking is how to perform effective **cross-modal interaction**
- Some previous methods **concatenate the RGB and TIR search region features directly** to perform a coarse interaction process with redundant background noises introduced
- Many other methods conduct fusion on **isolated pairs of RGB and TIR boxes**, which limits the cross-modal interaction within local regions and brings about inadequate context modeling
- We exploit **templates as the medium to bridge the cross-modal interaction between RGB and TIR search regions** by gathering and distributing target-relevant object and environment contexts



# Framework

- We extend the ViT architecture for joint feature extraction, search-template matching, and cross-modal interaction
- In TBSI module, bidirectional RGB and TIR search region interaction are bridged by the fused template, which serves as a medium to gather and distribute target-relevant contexts
- Original templates are also updated with the template medium



# Template-Bridged Search Region Interaction

- RGB and TIR template fusion

$$\boldsymbol{Z}_m = [\boldsymbol{Z}_r; \boldsymbol{Z}_t]\boldsymbol{W}_m,$$

- Bidirectional Template-Bridged Interaction
- Fused template gathers contexts from the TIR search region

$$\boldsymbol{D}_t = \mathrm{Softmax}(\frac{(\boldsymbol{Z}_m\boldsymbol{W}_q^1)(\boldsymbol{X}_t\boldsymbol{W}_k^1)^{\mathrm{T}}}{\sqrt{C}})(\boldsymbol{X}_t\boldsymbol{W}_v^1),$$

$$\boldsymbol{Z}'_m = \mathrm{LN}(\boldsymbol{Z}_m + \boldsymbol{D}_t),$$

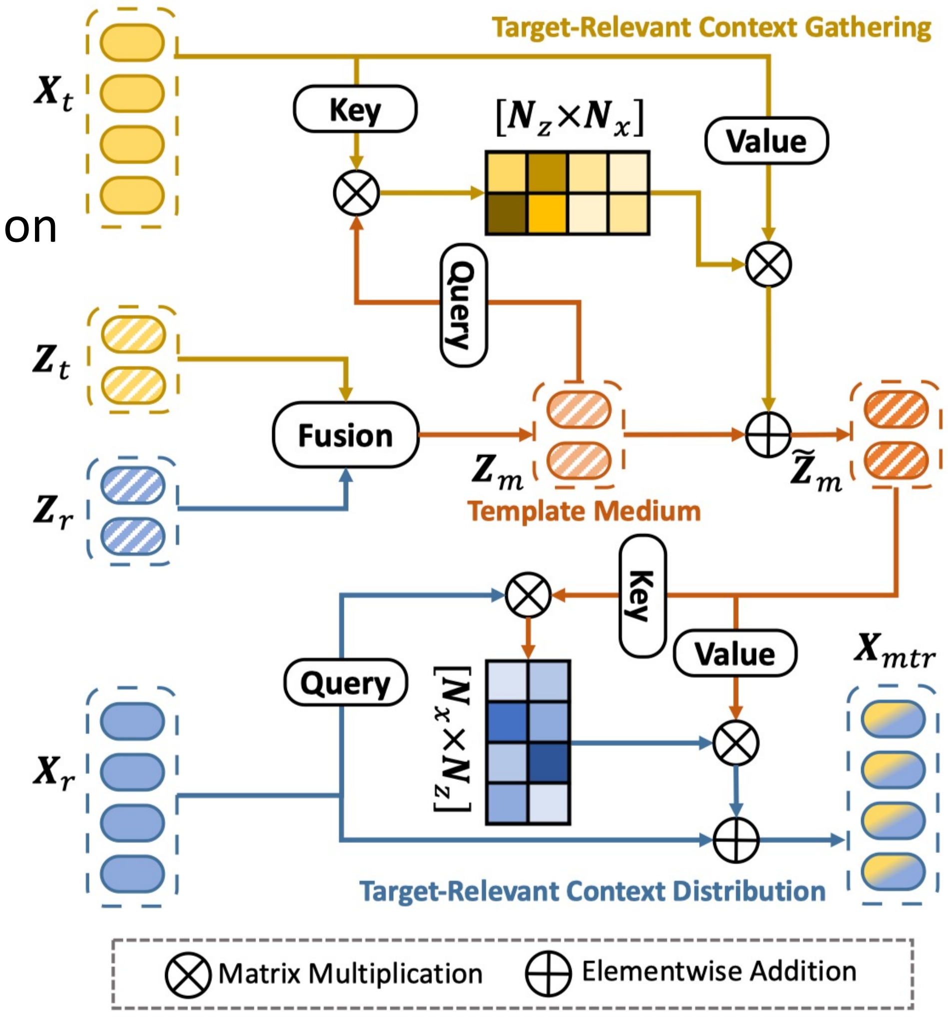
$$\tilde{\boldsymbol{Z}}_m = \mathrm{LN}(\boldsymbol{Z}'_m + \mathrm{MLP}(\boldsymbol{Z}'_m)),$$

$$\boldsymbol{D}_{mt} = \mathrm{Softmax}(\frac{(\boldsymbol{X}_r\boldsymbol{W}_q^2)(\tilde{\boldsymbol{Z}}_m\boldsymbol{W}_k^2)^{\mathrm{T}}}{\sqrt{C}})(\tilde{\boldsymbol{Z}}_m\boldsymbol{W}_v^2).$$

$$\boldsymbol{X}'_r = \mathrm{LN}(\boldsymbol{X}_r + \boldsymbol{D}_{mt}),$$

$$\boldsymbol{X}_{mtr} = \mathrm{LN}(\boldsymbol{X}'_r + \mathrm{MLP}(\boldsymbol{X}'_r)).$$

- Gathered contexts are distributed to RGB search region
- Original RGB and TIR templates are also updated by the enriched contexts from fused template




*TIR→Medium→RGB*

# Experiments

| | Method | Backbone | Pretraining | Precision | NormPrec | Success | FPS |
|---|---|---|---|---|---|---|---|
| Online | DAPNet [49] | VGG-M | ImageNet | 43.1 | 38.3 | 31.4 | - |
| | FANet [50] | VGG-M | ImageNet | 44.1 | 38.4 | 30.9 | - |
| | DAFNet [14] | VGG-M | ImageNet | 44.8 | 39.0 | 31.1 | 20.5 |
| | CAT [20] | VGG-M | ImageNet | 45.0 | 39.5 | 31.4 | - |
| | MANet [21] | VGG-M | ImageNet | 45.5 | - | 32.6 | 2.1 |
| | MANet++ [27] | VGG-M | ImageNet | 46.7 | 40.4 | 31.4 | - |
| | MaCNet [43] | VGG-M | ImageNet | 48.2 | 42.0 | 35.0 | 1.6 |
| | DMCNet [28] | VGG-M | ImageNet | 49.0 | 43.1 | 35.5 | - |
| | APFNet [39] | VGG-M | ImageNet | 50.0 | 43.9 | 36.2 | 1.9 |
| | mfDiMP [44] | ResNet-50 | SOT | 59.9 | - | 46.7 | 34.6 |
| Offline | TBSI | ViT-Tiny | ImageNet | 61.7 | 57.8 | 48.9 | **40.3** |
| | TBSI | ViT-Small | ImageNet | 62.4 | 58.6 | 49.4 | 39.1 |
| | TBSI | ViT-Base | ImageNet | 63.8 | 60.2 | 50.6 | 36.2 |
| | TBSI | ViT-Base | SOT | **69.2** | **65.7** | **55.6** | 36.2 |

Table 1. Comparison with state-of-the-art methods on LasHeR testing set. "SOT" denotes pretraining on the joint splits of COCO, LaSOT, GOT-10k, and TrackingNet, which is a common practice for training SOT methods. We also adopt this setting for a fair comparison. We only report the inference speeds of previous methods whose codes are available.

| | Method | Precision | Success |
|---|---|---|---|
| Online | MDNet+RGBT [32] | 72.2 | 49.5 |
| | MaCNet [43] | 76.4 | 53.2 |
| | DAPNet [49] | 76.6 | 53.7 |
| | MANet [21] | 77.7 | 53.9 |
| | HDINet [30] | 78.3 | 55.9 |
| | FANet [50] | 78.7 | 55.3 |
| | JMMAC [46] | 79.0 | 57.3 |
| | M5L [35] | 79.5 | 54.2 |
| | MANet++ [27] | 79.5 | 55.9 |
| | DAFNet [14] | 79.6 | 54.4 |
| | CAT [20] | 80.4 | 56.1 |
| | ADRNet [45] | 80.7 | 57.0 |
| | CMPP [37] | 82.3 | 57.5 |
| | APFNet [39] | 82.7 | 57.9 |
| | DMCNet [28] | 83.9 | 59.3 |
| | mfDiMP [44] | 84.2 | 59.1 |
| Offline | SiamCDA [47] | 76.0 | 56.9 |
| | SiamIVFN [16] | 81.1 | 63.2 |
| | TBSI | **87.1** | **63.7** |

Table 2. Comparison with state-of-the-art methods on RGBT234 dataset. Our method outperforms both online and offline ones.

| | Method | Precision | Success |
|---|---|---|---|
| Online | TFNet [51] | 77.7 | 52.9 |
| | CAT [20] | 79.2 | 53.3 |
| | DMCNet [28] | 79.7 | 55.5 |
| | mfDiMP* [44] | 84.9 | 59.3 |
| Offline | DSiamMFT [48] | 64.2 | 43.2 |
| | TBSI | **85.3** | **62.5** |

Table 3. Comparison with state-of-the-art methods on RGBT210 dataset. * means results are reproduced by us.

# Experiments

| Method | Precision | NormPrec | Success |
|---|---|---|---|
| RGB Baseline | 50.1 | 45.4 | 40.1 |
| RGB-T Baseline | 53.5 | 49.1 | 42.5 |
| w/o Template Bridging | 59.6 | 55.9 | 47.4 |
| w/o RGB→TM→TIR | 58.7 | 55.1 | 46.6 |
| w/o Template Updating | 62.7 | 58.9 | 49.7 |
| Full Model (TBSI) | **63.8** | **60.2** | **50.6** |

Table 4. Ablation studies of our proposed TBSI module. "TM" denotes the template medium for bridging interaction.

| Layers 4 | Layers 7 | Layers 10 | Precision | NormPrec | Success |
|---|---|---|---|---|---|
| | | | 53.5 | 49.1 | 42.5 |
| ✓ | | | 60.5 | 56.9 | 47.8 |
| ✓ | ✓ | | 62.7 | 59.2 | 49.8 |
| ✓ | ✓ | ✓ | **63.8** | **60.2** | **50.6** |

Table 5. Inserting layers of the proposed TBSI module.

| | APFNet† [39] | CMPP [37] | mfDiMP* [44] | TBSI |
|---|---|---|---|---|
| NO | 93.4/66.4 | 95.6/67.8 | **96.2**/69.4 | 96.1/**72.8** |
| PO | 85.0/58.7 | 85.5/60.1 | 86.6/60.9 | **88.7/64.7** |
| HO | 72.9/49.0 | 73.2/50.3 | 76.1/53.2 | **81.5/58.6** |
| LI | 82.3/54.4 | 86.2/58.4 | 84.2/58.0 | **89.2/63.6** |
| LR | 82.9/54.8 | **86.5**/57.1 | 82.1/53.0 | 85.1/**60.0** |
| TC | 82.1/57.3 | 83.5/58.3 | 84.8/58.9 | **85.8/63.2** |
| DEF | 77.1/54.6 | 75.0/54.1 | 81.5/60.2 | **84.1/63.7** |
| FM | 78.2/49.2 | 78.6/50.8 | 77.3/54.8 | **81.4/58.7** |
| SV | 82.1/56.5 | 81.5/57.2 | 87.1/63.7 | **89.9/66.8** |
| MB | 72.8/53.0 | 75.4/54.1 | 80.1/58.0 | **88.1/64.9** |
| CM | 76.3/54.5 | 75.6/54.1 | 84.0/60.3 | **88.0/65.0** |
| BC | 80.6/52.4 | 83.2/53.8 | 82.8/53.7 | **83.4/57.8** |

Table 6. Attribute-based Precision/Success scores on RGBT234 dataset. † denotes that the values are obtained by evaluating the authors' released raw tracking results. * means results are reproduced by us since raw results are unavailable.

# Experiments

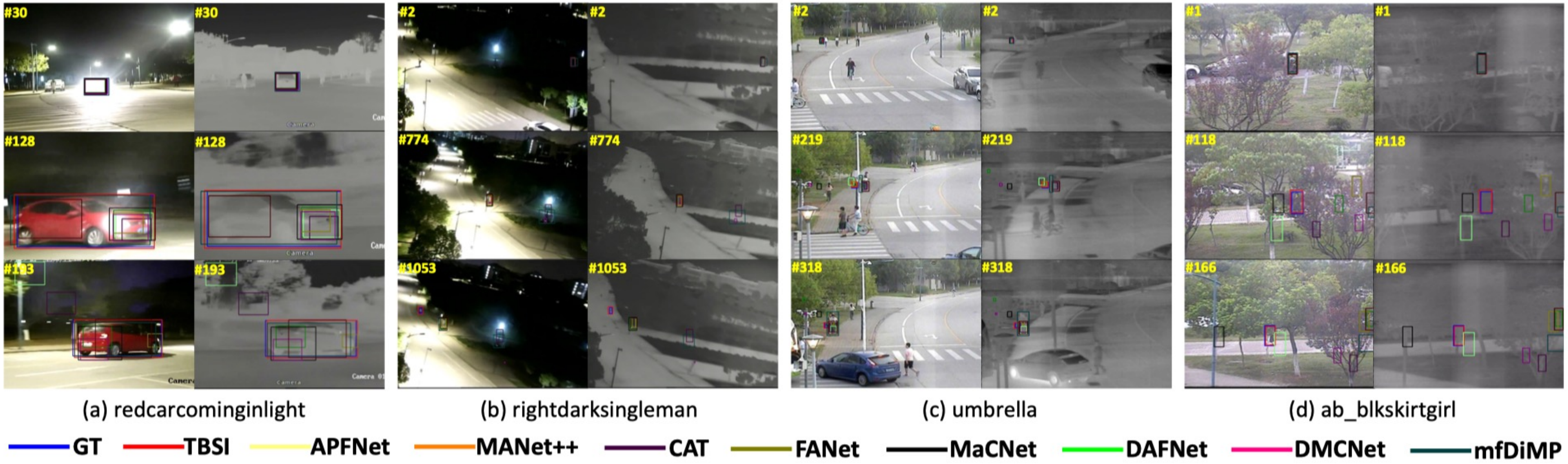



Figure 4. Qualitative comparison between our method and other RGB-T trackers on four representative sequences from LasHeR dataset.

# Experiments

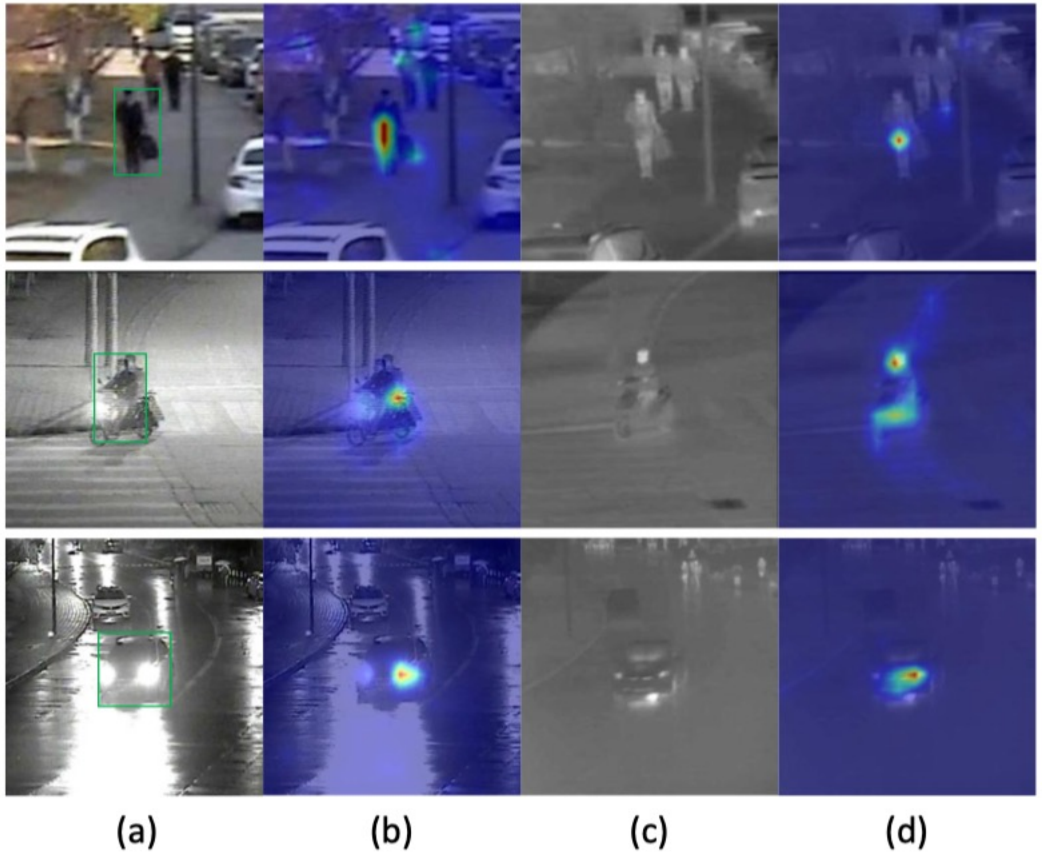



Figure 5. Visualization of attention maps between template medium tokens and search region tokens in our TBSI module. (a) RGB search region. (b) RGB attention map. (c) TIR search region. (d) TIR attention map.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

# Thank You!

Code will be released at https://github.com/RyanHTR/TBSI

中国科学院信息工程研究所 INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学 University of Chinese Academy of Sciences

北京航空航天大学

美团