MATERIAL-GUIDED SIAMESE FUSION NETWORK FOR HYPERSPECTRAL OBJECT TRACKING

Zhuanfeng Li[1] *Fengchao Xiong*[1] *Jianfeng Lu*[1] *Jun Zhou*[2] *Yuntao Qian*[3]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2] School of Information and Communication Technology, Griffith University, Australia
[3] College of Computer Science, Zhejiang University, China

ABSTRACT

Hyperspectral videos (HSVs) have more potential in target tracking than color videos thanks to the material identification capability provided by abundant spectral bands. Due to limited HSVs for training, most current hyperspectral trackers are based on hand-crafted features rather than deeply learned ones, resulting in poor tracking performance. This paper introduces a material-guided Siamese fusion network (SiamF) for hyperspectral object tracking to make up this gap. Belonging to the Siamese tracker family and SiamF aims to model the appearance of hyperspectral objects using backbone networks trained on color images. Specifically, SiamF splits each hyperspectral frame into multiple groups of false-color images according to their band importance. Then SiamF employs a hyperspectral feature fusion (HFF) module with a dense connection architecture to integrate the extracted features from different layers and band groups, producing a multi-scale multilevel spatial-spectral representation of the targets. Instead of direct addition or concatenation, HFF employs global-local channel attention for feature fusion, so that yielded features capture the global and local structure of a specific object. Moreover, online spatial and material classifiers are developed to inject spatial and material appearance changes information into SiamF for adaptively online tracking. Experimental results demonstrate our tracker outperforms alternative methods.

***Index Terms*—** Hyperspectral object tracking, feature fusion, material unmixing, appearance modelling

1. INTRODUCTION

Visual object tracking is a fundamental task in computer vision and has achieved significant developments from manual features-based [1] to deep features-based [2, 3] methods. Compared with alternative videos, hyperspectral videos (HSVs) can capture the physical material property of the target thanks to acquired numerous spectral bands. The material information is very helpful to distinguish the object from the surrounding background [4]. Accordingly, tracking with HSVs is attracting more and more attention [5, 6].

Existing research mainly focuses on the visual appearance representation of the target. Xiong *et. al* [7] designed a spatial-spectral histogram of multi-dimensional gradients to describe the local spatial-spectral structure and fractional abundance to depict the global constitute material distribution of the scene. As hyperspectral object tracking suffers from a small training sample problem, most current deep learning (DL) based trackers employ the pre-trained backbone networks on color images for appearance modeling. For example, Uzkent *et. al* [8] converted each hyperspectral frame into a false-color image and then passed the converted image through the pre-trained VGGNet to extract deep features. Unfortunately, the converted image unavoidably loses useful spectral information, making the appearance modeling still not robust. Alternatively, Li *et. al* [9] proposed a deep hyperspectral tracker (BAE-Net) to split a hyperspectral frame into multiple band groups according to band-wise importance generated by the band attention module. These band groups were then fed into the pre-trained backbone network to produce several weak trackers for ensemble tracking. However, as we know, powerful feature extraction is always important to build a good tracker [10]. Each weak tracker of BAE-Net only extracts partial information of a hyperspectral object, making the extracted features still not powerful enough.

In this paper, we introduce a material-guided Siamese fusion network named SiamF to address the limitations of BAE-Net. As shown in Fig. 1, our SiamF includes a hyperspectral feature fusion (HFF) module, an online spatial classifier (OSC) module, and an online material classifier (OMC) module. Like BAE-Net, our SiamF splits a hyperspectral frame into multiple false-color images, i.e., band groups, so that ResNet-50 network [11] can be applied to extract features from each false-color image. HFF module is then set after each convolution block of ResNet-50 to consider the spatial-spectral information for improved feature extraction simulta-

This work was supported in part by the National Natural Science Foundation of China under Grant 62002169 and 62071421, Jiangsu Provincial Natural Science Foundation of China under Grant BK20200466 and 111 Program (No. B13022). (Corresponding author: Fengchao Xiong; Jianfeng Lu.)

neously. Armed with dense connection architecture, the HFF module aggregates the feature maps produced by each band group from two paths. One path integrates the extracted features from each band group to obtain spatial-spectral representation. The other path passes the fused features from shallow to deep layers to yield a multi-scale multilevel representation of the object. Moreover, instead of direct addition for feature fusion, we leverage a global-local channel attention mechanism that simultaneously focuses on global and local feature context for feature aggregation [12].

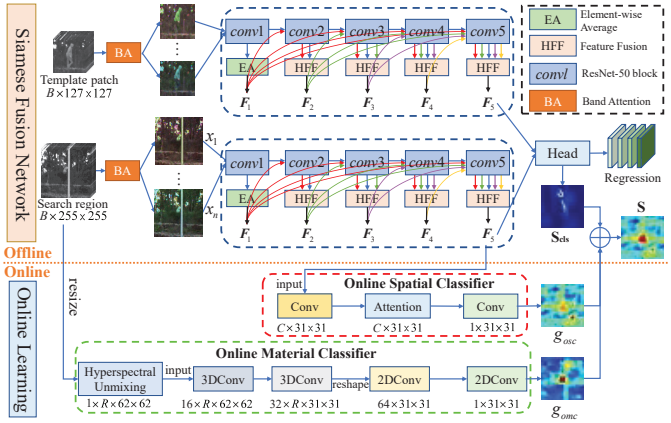


Fig. 1. The overall architecture of SiamF.

Object tracking is an online dynamic process where appearance changes of objects are not uncommon. To adapt SiamF for more adaptive tracking, two online classifiers are introduced. The online spatial classifier is the same as that in [13] and aims to describe the spatial appearance changes. In addition, there are also constituted material distribution changes with the movement of the target. The material information can improve the discrimination capability of the inter-object [14] and increase the robustness of a tracker to many challenging scenes such as background clutter. Motivated by this, we extract the material distribution of each frame via hyperspectral unmixing [15–17] and design an online material classifier to provide the material appearance changes of the target. Finally, these two online classifiers are integrated into SiamF to guide the object localization.

Our contributions can be summarized as follows: 1) we introduce a hyperspectral feature fusion module to adaptively integrate the feature extracted from each band group and features in different layers, yielding multi-scale multilevel spatial-spectral features for improved tracking. 2) We design an online material classifier module to complement SiamF so that the material appearance changes of the object are considered, increasing the robustness of our tracker to distractors and background noises.

2. PROPOSED METHOD

This section introduces the details of SiamF and the online spatial and material classifier modules.

2.1. Siamese Fusion Network

As shown in Fig. 1, SiamF consists of a band attention module, a hyperspectral feature fusion module, and box adaptive heads. As in [9], band attention module learns the band importance of a hyperspectral frame $\mathbf{X}$ and split $\mathbf{X}$ into multiple false-color images $\{x_1, \cdots, x_n\}$, i.e., band groups, for subsequent deep feature extraction. Hyperspectral feature fusion network extracts multi-scale multilevel spatial-spectral features. Given the extracted features $\varphi(z)$ and $\varphi(x)$ from template branch and search branch, box adaptive heads locates the object generates the classification map $\mathbf{S}_{cls}$ and regression map $\mathbf{S}_{reg}$ to locate the object, i.e.,

$$\begin{aligned} \mathbf{S}_{cls} &= [\varphi(x)]_{cls} \star [\varphi(z)]_{cls} \\ \mathbf{S}_{reg} &= [\varphi(x)]_{reg} \star [\varphi(z)]_{reg} \end{aligned} \tag{1}$$

$[\varphi(\cdot)]_{cls}$ and $[\varphi(\cdot)]_{reg}$ are copied from $\varphi(\cdot)$. $\star$ denotes the cross-correlation operation. The loss function of SiamF is a multi-task loss including cross-entropy loss for classification and intersection over union (IoU) loss for regression, i.e.,

$$\mathcal{L} = \beta_1 \mathcal{L}_{cls} + \beta_2 \mathcal{L}_{reg} \tag{2}$$

where $\mathcal{L}_{cls}$ is the cross-entropy loss and $\mathcal{L}_{reg}$ is the IoU loss. β_1 and β_2 balance two losses and are empirically set as 1.

Hyperspectral Feature Fusion Module: Let $\mathbf{M}(x_i^l) \in \mathbb{R}^{C\times W\times H}$ be the feature map extracted from x_i with the l-th convolution block of ResNet-50. Each $\mathbf{M}(x_i^l)$ captures partial information of a hyperspectral frame. Hyperspectral feature fusion module aims to fuse $\{\mathbf{M}(x_i^l)\}_{i=1,\cdots,n}$ to produce the spatial-spectral representation of $\mathbf{X}$. Inspired by DenseNet [18], we also include the fused features in previous layers, i.e., $\{\mathbf{F}_k\}_{k=1\cdots l-1}$ so that multi-scale multilevel representation is achieved in every layer, facilitating subsequent object tracking. Formally, HFF module is formulated as

$$\mathbf{F}_l = fuse(\mathbf{M}(x_1^l), \cdots, \mathbf{M}(x_n^l), A(\mathbf{F}_1), \cdots, A(\mathbf{F}_{l-1})) \tag{3}$$

$A(\cdot)$ is an adaptor to make $\mathbf{F}_k$ maintain consistent size with $\mathbf{M}(x_i^l)$. For simplicity, we implement $A(\cdot)$ by passing $\mathbf{F}_k$ into the convolution blocks of ResNet-50. As there is no previous fused map in the first fusion, we directly perform element-wise averaging on $\{\mathbf{M}(x_i^l)\}_{i=1,\cdots,n}$ to produce $\mathbf{F}_1$.

Concatenation and summation are common practices for feature fusion but can only offer a fixed linear aggregation of feature maps, failing to consider the specific properties of objects. Instead, we employ a global-local channel attention module (GLCAM) in [12] to simultaneously consider global and local information of objects for attentional feature fusion. Specifically, GLCAM is produced by aggregating local and global contexts, i.e.,

$$\mathbf{W} = sigmoid(L(\overline{\mathbf{F}_l}) \oplus G(\overline{\mathbf{F}_l})) \tag{4}$$

where $\oplus$ represents the broadcasting addition and $\overline{\mathbf{F}_l}$ is obtained by averaging $\{\mathbf{M}(x_i^l)\}_{i=1\cdots n}$ and $\{A(\mathbf{F}_k)\}_{k=1\cdots l-1}$.

$L(\overline{\mathbf{F}_l}) \in \mathbb{R}^{C\times W\times H}$ represents the local contexts and is produced by conducting pixel-wise convolution on $\overline{\mathbf{F}_l}$, i.e,

$$L(\overline{\mathbf{F}_l}) = BN(conv_2(\sigma(BN(conv_1(\overline{\mathbf{F}_l}))))) \tag{5}$$

where $conv_1$ and $conv_2$ are 1×1 pixel-wise convolution with $\frac{C}{s}$ and C channels, s is the channel reduction ratio and BN represents batch normalization and σ is ReLU activation function. The global contexts $G(\overline{\mathbf{F}_l}) \in \mathbb{R}^C$ is obtained by summarizing the features of all the pixels, i.e.,

$$G(\overline{\mathbf{F}_l}) = BN(conv_4(\sigma(BN(conv_3(GAP(\overline{\mathbf{F}_l})))))) \tag{6}$$

where $GAP(\cdot)$ represents the global average pooling. $conv_3$ and $conv_4$ are 1×1 convolution with $\frac{C}{s}$ and C output channels, respectively. With $\mathbf{W}$, the attentional feature fusion is performed by

$$\begin{aligned}\mathbf{F}_l = \alpha_1\mathbf{W}\odot A(\mathbf{F}_{l-1}) + \sum_{i=1}^{n}\alpha_2(1-\mathbf{W})\odot\mathbf{M}(x_i^l) \\ + \sum_{k=1}^{l-2}\alpha_2(1-\mathbf{W})\odot A(\mathbf{F}_k)\end{aligned} \tag{7}$$

where $\odot$ represents the element-wise multiplication. α_1 is set as 1 to focus more on the fused feature in the latest layer. α_2 is given by $\frac{1}{n+l-2}$ to equally treat other feature maps.

2.2. Online Spatial and Material Classifiers

Besides offline SimaF, we also introduce online spatial and material classifier modules as a supplement to learn appearance changes during tracking. Both modules formulate the following learning objective:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{m}\gamma_i\|g(\mathbf{F}_i;\mathbf{w}) - Y_i\|^2 + \sum_j \mu_j\|\mathbf{w}_j\|^2 \tag{8}$$

where Y_i is the classification confidence by sampling a Gaussian function at the center of the object location, $\mathbf{F}_i$ is the feature map of a training sample, and $g(\mathbf{F}_i;\mathbf{w})$ formulates an online classifier. The weight γ_i controls the impact of $\mathbf{F}_i$, and the μ_j controls the regularization on $\mathbf{w}_j$. We introduce the details of $g(\mathbf{F}_i;\mathbf{w})$ in the following.

Online Spatial Classifier: Online spatial classifier is a 2-layer fully convolutional neural network and aims to learn the spatial appearance changes of the objects by

$$\begin{aligned}g_{osc}(\mathbf{F}_l;\mathbf{W}_S,\text{Att}_c,\text{Att}_s) =& \sigma(\mathbf{W}_S^{(2)} * (\text{Att}_c \otimes \sigma(\mathbf{W}_S^{(1)} * \mathbf{F}_l) \\ &+ \text{Att}_s \otimes \sigma(\mathbf{W}_S^{(1)} * \mathbf{F}_l)))\end{aligned} \tag{9}$$

Following [19], we use spatial attention Att_s and channel attention Att_c to address the training sample imbalance problem between foreground positive pixels and background negative pixels. Att_s is implemented by two fully convolution layers followed by a GAP. Att_c consists of a softmax activation function followed by channel average. $\mathbf{W}_S^{(1)}$ and $\mathbf{W}_S^{(2)}$ are the convolution operators respectively size of $64\times 1\times 1$ and $1\times 4\times 4$. $*$ represents multi-channel convolution operation, $\otimes$ represents element-wise broadcasting multiplication.

Online Material Classifier: Besides the spatial information changes, changes also exist in the material constitution during tracking. The material constitution information can enhance the recognition ability of the system and improve the robustness of the system against challenging scenes such as background clutter and rapid changes of the target appearance [7]. To this end, we introduce an online material classifier to guide the localization of the target where the material abundance features $\mathbf{F}_M$ are obtained by hyperspectral unmixing as [7].

According to $\mathbf{F}_M$, our online material classifier module is a four-layer convolutional neural network, expressed as

$$g_{omc}(\mathbf{F}_M;\mathbf{W}_M) = \mathbf{W}_M^{(4)} * (\mathbf{W}_M^{(3)} * f(\sigma(\mathbf{W}_M^{(2)} * \sigma(\mathbf{W}_M^{(1)} * \mathbf{F}_M)))) \tag{10}$$

$\mathbf{W}_M^{(1)}$ and $\mathbf{W}_M^{(2)}$ are 3D convolutions with kernels of $3\times 3\times 3$. $\mathbf{W}_M^{(3)}$ and $\mathbf{W}_M^{(4)}$ represent the 2D convolution with similar architecture to OSC. $f(\cdot)$ reshapes the four-dimensional feature map into three-dimensional one and adjust the feature channel to 64 with 1×1 convolution. Finally, the above three classification maps are weighted sum to obtain the final classification score map $\mathbf{S}$, i.e.,

$$\mathbf{S} = \theta(\lambda g_{osc}(\varphi(\mathbf{X})) + (1-\lambda)\mathbf{S}_{\mathbf{cls}}) + (1-\theta)(g_{omc}(\mathbf{F}_M) \tag{11}$$

where λ and θ balance three classification maps.

2.3. Training and Tracking

Offline Training Our SimaF was trained on the training datasets provided by hyperspectral object tracking competition (HOTC)[1] [7]. The spatial sizes of the template patch and search area are set to 127×127 and 255×255, respectively. To alleviate the imbalance between positive and negative samples, we collect at most 48 negative samples and 16 positive samples from each image pair for offline training [3].

Online Tracking: 30 initial training samples are generated in the first frame by performing blur, rotation and dropout on the template patch. After that, we perform unmixing on the training samples, producing material abundance features. These features are used to pre-train the material classification network with a learning rate of 0.01 and number of epochs setting to 100. We crop the search region in subsequent frames and extract the abundance feature as training samples to finetune the material classification network every 10 frames with the learning rate of 0.001 and 5 epochs. For the online spatial classifier, the pre-training and finetuning methods are similar to those of the OMC. Conjugate gradient descent algorithm and stochastic gradient descent are used to optimize OSC and OMC, respectively.

[1] https://www.hsitracking.com

3. EXPERIMENTS

3.1. Experiment settings

Our proposed SiamF was implemented in Python under the framework of Pytorch. For offline training of SiamF, the learning rate was set as 0.0005. λ was set as 0.8, the same as [19]. θ was given by 0.72. Precision plot, success plot, and area under the curve (AUC) of success plot were used to evaluate the performance of all the trackers. All the trackers were tested on the testing set of the HOTC dataset.

3.2. Results and Analysis

Table 1. Ablation study on each module.

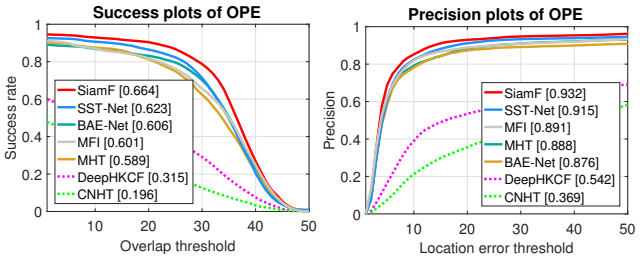
Baseline	HFF	OSC	OMC	AUC	DP
✓	-	-	-	0.626	0.873
✓	✓	-	-	0.640	0.899
✓	✓	✓	-	0.649	0.909
✓	✓	✓	✓	**0.664**	**0.932**

Ablation Study: We performed ablation studies on HFF, OSC, and OMC modules in Table 1 where distance precision (DP) was reported at 20 pixels. The baseline tracker splits the hyperspectral frame into multiple band groups for ensemble tracking, similar to BAE-Net. HFF module extracts multi-scale multilevel spatial-spectral features and significantly outperforms the baseline method with a gain of **0.014** in AUC score and **0.026** in DP score, which shows that powerful extraction is helpful for object tracking. OSC considers the spatial structure changes of the object and also improves the tracking performance to a large margin. It is worth noting that our OMC learns the changes of constitute material distribution, providing a gain of **0.015** in AUC score and **0.023** in DP score than only using OSC. Overall, the ablation study show the effectiveness of the proposed HFF and OMC modules.

Table 2. AUC comparison with state-of-the-art color trackers where "H" means hyperspectral and "F" means false-color.

Video	SiamF	SiamBAN	SiamRPN++	DROL-RPN	BACF	MCCT
Color	n/a	**0.630**	0.609	**0.628**	0.519	0.569
H/F	**0.664**	0.598	0.569	**0.607**	0.544	0.528

Comparison with State-of-the-art Color Trackers: In this section, we compare the SiamF with several state-of-the-art color trackers. These trackers include deep feature-based trackers, such as SiamBAN [3], SiamRPN++ [2], and DROL-RPN [19], and hand-crafted feature-based trackers, such as BACF [20] and MCCT [21]. Our proposed SiamF was run on the hyperspectral videos, while the color trackers were run on color and false-color videos with the same scene. As shown in Table 2, the deep feature-based trackers surpass all hand-crafted feature-based trackers thanks to their stronger feature representation. Attributing to the multi-scale multi-level spatial-spectral features extracted by the HFF module and the consideration of spatial and material changes, our SiamF tracker significantly outperforms all color trackers and achieves the highest AUC score of **0.664**.



(a) Success plot (b) Precision plot

Fig. 2. Comparison with hyperspectral trackers.

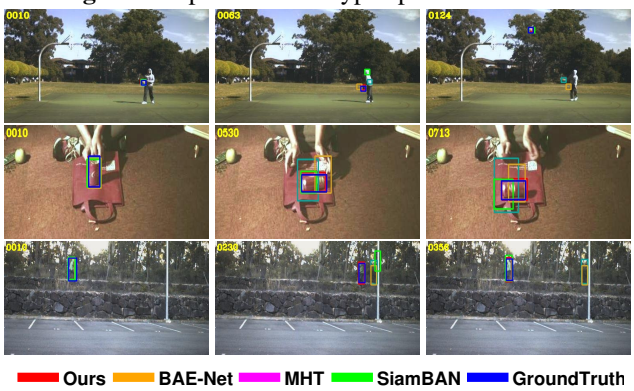


Fig. 3. Demonstrations of tracking results.

Comparison with Hyperspectral Trackers: We further compare the SiamF with several hyperspectral trackers such as SST-Net [22], BAE-Net [9], MFI [23], CNHT [24], MHT [7], and DeepHKCF [8]. Fig. 2 shows the performance of all trackers with respect to precision plot and success plot. The lack of negative samples limits the discriminative ability of CNHT, yielding the lowest AUC score. Simultaneous consideration of spatial and spectral information helps MHT, BAE-Net, and SST-Net achieve higher tracking performance. Our SiamF achieves the highest AUC score and DP score compared to other hyperspectral trackers. Fig. 3 visualizes the tracking results in SiamF, BAE-Net, SiamBAN, MHT on *basketball*, *coke*, and *forest2* sequences, where SiamBAN is run on false-color videos. As can be seen, our SiamF provides the most accurate bounding box, again demonstrating its effectiveness in hyperspectral tracking.

4. CONCLUSION

This paper introduces a SiamF method for hyperspectral tracking. SiamF divides each hyperspectral frame into multiple band groups and uses the strong representation ability of color video to model the appearance of hyperspectral objects. Hyperspectral feature fusion module is embedded in SiamF to generate the multi-scale and multi-level spatial-spectral representation of the target. An online material classifier is introduced to guide target location by considering the changes of material distribution. Experiments demonstrate that our proposed SiamF achieves higher tracking performance than other hyperspectral and color trackers.

5. REFERENCES

[1] Xi Li, Liming Zhao, Wei Ji, Yiming Wu, Fei Wu, Ming-Hsuan Yang, Dacheng Tao, and Ian Reid, "Multi-task structure-aware context modeling for robust keypoint-based object tracking," *IEEE TPAMI*, vol. 41, no. 4, pp. 915–927, 2019.

[2] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *IEEE CVPR*, 2019, pp. 4282–4291.

[3] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji, "Siamese box adaptive network for visual tracking," in *IEEE CVPR*, 2020, pp. 6668–6677.

[4] Jie Liang, Jun Zhou, Lei Tong, Xiao Bai, and Bin Wang, "Material based salient object detection from hyperspectral images," *PR*, vol. 76, pp. 476–490, 2018.

[5] Fengchao Xiong, Jun Zhou, Jocelyn Chanussot, and Yuntao Qian, "Dynamic material-aware object tracking in hyperspectral videos," in *IEEE WHISPERS*, 2019, pp. 1–6.

[6] Lulu Chen, Yongqiang Zhao, Jiaxin Yao, Jiaxin Chen, Ning Li, Jonathan Cheung-Wai Chan, and Seong G. Kong, "Object tracking in hyperspectral-oriented video with fast spatial-spectral features," *Remote Sensing*, vol. 13, no. 10, 2021.

[7] Fengchao Xiong, Jun Zhou, and Yuntao Qian, "Material based object tracking in hyperspectral videos," *IEEE TIP*, vol. 29, pp. 3719–3733, 2020.

[8] Burak Uzkent, Aneesh Rangnekar, and Matthew J Hoffman, "Tracking in aerial hyperspectral videos using deep kernelized correlation filters," *IEEE TGRS*, vol. 57, no. 1, pp. 449–461, 2018.

[9] Zhuanfeng Li, Fengchao Xiong, Jun Zhou, Jing Wang, Jianfeng Lu, and Yuntao Qian, "BAE-Net: A band attention aware ensemble network for hyperspectral object tracking," in *IEEE ICIP*, 2020, pp. 2106–2110.

[10] Naiyan Wang, Jianping Shi, Dit-Yan Yeung, and Jiaya Jia, "Understanding and diagnosing visual tracking systems," in *IEEE ICCV*, 2015, pp. 3101–3109.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[12] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard, "Attentional feature fusion," in *IEEE WACV*, 2021, pp. 3560–3569.

[13] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg, "Atom: Accurate tracking by overlap maximization," in *IEEE CVPR*, 2019, pp. 4660–4669.

[14] Muhammad Uzair, Arif Mahmood, and Ajmal Mian, "Hyperspectral face recognition with spatiospectral information fusion and pls regression," *IEEE TIP*, vol. 24, no. 3, pp. 1127–1137, 2015.

[15] Jos M. Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE JSTARS*, vol. 5, no. 2, pp. 354–379, 2012.

[16] Fengchao Xiong, Jun Zhou, Shuyin Tao, Jianfeng Lu, and Yuntao Qian, "SNMF-Net: Learning a deep alternating neural network for hyperspectral unmixing," *IEEE TGRS*, vol. 60, pp. 1–16, 2022.

[17] Fengchao Xiong, Jun Zhou, Minchao Ye, Jianfeng Lu, and Yuntao Qian, "NMF-SAE: An interpretable sparse autoencoder for hyperspectral unmixing," in *IEEE ICASSP*, 2021, pp. 1865–1869.

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, 2017, pp. 4700–4708.

[19] Jinghao Zhou, Peng Wang, and Haoyang Sun, "Discriminative and robust online learning for siamese visual tracking," in *AAAI*, 2020, vol. 34, pp. 13017–13024.

[20] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey, "Learning background-aware correlation filters for visual tracking," in *IEEE CVPR*, 2017, pp. 1135–1143.

[21] Ning Wang, Wengang Zhou, Qi Tian, Richang Hong, Meng Wang, and Houqiang Li, "Multi-cue correlation filters for robust visual tracking," in *IEEE CVPR*, 2018, pp. 4844–4853.

[22] Zhuanfeng Li, Xinhai Ye, Fengchao Xiong, Jianfeng Lu, Jun Zhou, and Yuntao Qian, "Spectral-spatial-temporal attention network for hyperspectral tracking," in *IEEE WHISPERS*, 2021, pp. 1–5.

[23] Zhe Zhang, Kun Qian, Juan Du, and Huixin Zhou, "Multi-features integration based hyperspectral videos tracker," in *IEEE WHISPERS*, 2021, pp. 1–5.

[24] Kun Qian, Jun Zhou, Fengchao Xiong, Huixin Zhou, and Juan Du, "Object tracking in hyperspectral videos with convolutional features and kernelized correlation filter," in *ICSM*. Springer, 2018, pp. 308–319.