Efficient Transformer Network for Visible and Ultraviolet Object Tracking

Qinghua Song[0009−0000−6260−3067] and Xiaolei Wang[0009−0006−6953−0626]

Intelligent Game and Decision Laboratory, Beijing, 100080, China
songziyao1122@outlook.com

Abstract. In recent years, more researchers in the field of multi-modal tracking have focused on various algorithms for RGB-T tracking, leveraging the complementary nature of RGB and TIR imaging to achieve good application results. However, their performance tends to degrade in specific scenarios where the target is color-camouflaged and TIR modality is ineffective. Our experiments reveal that ultraviolet (UV) sensors can effectively image certain camouflage materials. Therefore, in extreme scenarios where the target is color-camouflaged and TIR modality is ineffective, UV modality can serve as a supplementary means to RGB and TIR modalities, enhancing tracking performance. In this paper, we propose the first multi-modal object tracking network for visible light, thermal infrared, and ultraviolet, namely VTUTrack, which achieves better tracking performance in complex scenarios. Furthermore, to meet the needs of real-time tracking applications, we introduce an adaptive candidate elimination mechanism based on modality reliability within the ViT (Vision Transformer) backbone network, reducing the computational burden of multi-modal feature extraction and improving tracking inference speed. Extensive experiments further demonstrate the effectiveness of our proposed RGB-T-UV multi-modal object tracking method. The VUOT dataset is available at https://drive.google.com/file/d/1f2FflhUTvfDERJ_j18Plk4XoOur_BLXN/view?usp=drive_link.

Keywords: RGB-UV object tracking, Dataset, Transformer network, Adaptive feature selection mechanism.

1 Introduction

Given the initial position of a target, the visual object tracking is used to capture the target in subsequent frames [13], where the target can suffer out-of-view, occlusion, variation in illumination and blur. Current algorithms have addressed some of these challenges in the visible (RGB) mode. However, under extreme conditions where the target is actively camouflaged, the limited information provided by the RGB modality can result in color confusion between the target and the background, making it difficult to distinguish between them. Our experiments reveal that ultraviolet (UV) imaging is more sensitive to certain camouflage materials. As shown in Fig. 1, UV imaging outperforms RGB imaging in these scenarios. Therefore, UV can supplement RGB by leveraging its advantages to address such tracking problems.

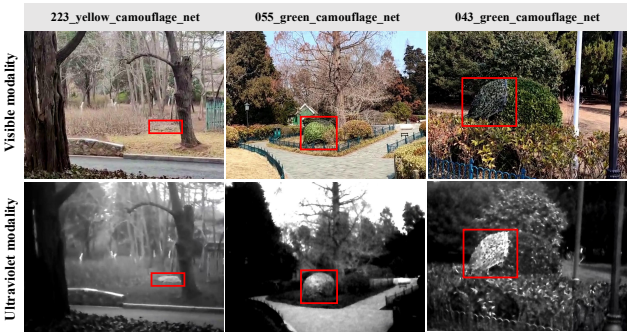

Fig. 1: Multiple RGB and UV modality image pairs sampled from the VUOT dataset, where the UV modality demonstrates superior imaging performance for the target.

With the advancement of UV imaging technology, UV lenses can now directly generate clear images of objects using only weak UV light sources. They typically capture UV spectrum wavelengths ranging from approximately 10 to 400 nanometers. In UV imaging, all objects reflect UV light to varying degrees. Since camouflage materials cannot match the reflectance of any natural environment, the difference in UV reflectance can be used to achieve ultraviolet-sensitive imaging of camouflage materials. Therefore, UV imaging is more effective than RGB imaging for some camouflage materials, and by combining both RGB and UV information, we can achieve more robust tracking performance in certain challenging scenarios.

However, the current lack of datasets is a major bottleneck in RGB and UV fusion tracking, particularly for data-driven deep learning models where datasets are crucial for training, evaluation, and tuning. Although there are several publicly available visible-thermal (RGB-T) benchmark datasets in the field of fusion tracking, including VTUAV [18], LasHeR [10], RGBT210 [11], and RGBT234 [9], there is a lack of publicly available datasets for RGB and UV object tracking. To achieve RGB-UV object tracking, this paper constructs a RGB-UV object tracking dataset that includes various extreme scenarios and proposes an efficient network for RGB-UV object tracking. The main contributions are as follows:

(1) We construct a dataset for visible and ultraviolet object tracking (VUOT) that includes various challenging scenarios with target-background confusion. To our knowledge, VUOT is the first dataset specifically designed for RGB and UV tracking. Furthermore, we consider data attributes such as target occlusion and out-of-view to achieve a comprehensive evaluation with wider applications. We also provide exquisite target annotations in frame levels, which can meet the requirement of training trackers.

(2) We propose a visible and ultraviolet object tracking transformer network, namely VUTrack, which fully leverages the complementary information from the RGB and UV spectra to significantly improve the tracking performance in

challenging scenarios with target-background confusion. Moreover, we introduce an adaptive feature selection mechanism based on modality reliability into the ViT [2] backbone network to reduce the impact of less efficient modalities on tracking inference speed, thereby further improving the overall performance of the tracker.

(3) Using the VUOT dataset, we conduct extensive experiments on leading multimodal fusion trackers. The experimental results demonstrate the effectiveness of both our VUOT dataset and the VUTrack network. Based on these findings, we further analyze the potential applications of UV in object tracking and provide fundamental insights.

2 Related Work

In this section, we briefly review different fusion tracking datasets and fusion tracking methods.

Fusion Tracking Dataset. With the rapid development of various optical imaging sensors, an increasing number of sensors are being applied to the field of object tracking. Currently, more researchers are focusing on the development of algorithms for RGB-T object tracking, and several public datasets have been constructed for this purpose. In 2016, Li *et al.* [8] release a gray-scale RGB-T dataset with 50 videos. Later, RGBT210 [11] and RGBT234 [9] are proposed, containing 210 and 234 test videos. In recent years, Li *et al* proposed the large-scale LasHeR [10] dataset for short-term RGBT tracking, and Zhang *et al* constructed a high-resolution, large-scale VTUAV [18] dataset. These datasets provide abundant data resources for the training and testing of RGB-T algorithms, significantly promoting the rapid development of the field. However, to the best of our knowledge, there is still a lack of publicly available datasets for visible and ultraviolet object tracking. This gap limits in-depth research in the field of visible and ultraviolet fusion, highlighting the urgent need for the development of new datasets to support algorithm development and application.

Fusion Tracking. Currently, there are no publicly available visible and ultraviolet fused tracking algorithms. Therefore, this paper primarily discusses RGB-T object tracking algorithms. Since both thermal infrared and ultraviolet images are grayscale and share similar data characteristics, this paper will extend existing RGB-T tracking algorithms to explore visible and ultraviolet fused tracking. Since the introduction of the Transformer architecture into the object tracking field by ViT in 2020, Transformer-based models have significantly improved tracking performance by replacing traditional CNN backbone networks. However, the high computational complexity and large number of network parameters of Transformers inevitably reduce inference speed, affecting the real-time performance of the algorithms. The latest RGB-T fused tracking algorithms, TBSI [5] and ViPT [20], adopt a powerful ViT as the base network architecture. They design the fusion module as an independent component that is inserted between two Transformer encoders to achieve better feature fusion between the template and search region. While this method enhances the in-

Table 1: Statistics comparison among existing multi-modality tracking datasets.

Modality	Datasets	Num.Seq	Avg.Frame	Min.Frame	Max.Frame	Total.Frame	Resolution	Year
RGB-T	GTOT [8]	50	157	40	376	7.8K	384*384	2016
	RGBT210 [11]	210	498	40	4140	104.7K	630*460	2017
	VOT2019 [7]	60	334	40	1335	40.2K	630*460	2019
	RGBT234 [9]	234	498	40	4140	116.7K	630*460	2019
RGB-UV	VUOT(Ours)	226	332	182	596	75.3K	960*720	2024

teraction capability between modalities during the feature extraction phase, it requires computation on all image tokens during feature extraction and correlation modeling, increasing computational complexity and making real-time performance nearly unattainable. Inspired by OSTrack [16], to achieve a more efficient ViT network for visible and ultraviolet fused tracking, we introduce a modality reliability-based adaptive feature selection mechanism within the backbone network. This mechanism retains more effective modality information in the early stages of the ViT network, reducing the computational burden of less effective modalities. This approach improves tracking inference speed, better balancing the trade-off between inference speed and tracking performance.

3 VUOT Dataset

3.1 Overview of the Dataset

The VUOT data collection used a multi-sensor lens platform with two degrees of rotational freedom. The RGB camera used is a Sony Starvis camera with a resolution of 1920x1080, while the UV camera is a UV230M imager that captures UV spectrum in the wavelength range of 20-400nm, with a resolution of 960x720. We maintained a distance of 20-50 meters between the collection platform and the target to obtain appropriate target sizes. We collected 226 sequences, comprising 75,370 image pairs. All images were downscaled to a resolution of 960*720 and stored in JPG format, with a sampling rate of 30 fps, resulting in a total data size of 31.5GB. Of these, 160 sequences are designated as the training set and 66 sequences as the test set. All sequences are provided with frame-by-frame precise target annotations, including bounding boxes that indicate the true position of the target, totaling 75,370 frame-level annotations. As shown in Table 1, the scale of this dataset is comparable to earlier RGB-T datasets, providing robust data support for algorithmic research.

3.2 Image Alignment and Target Annotation

Image Alignment. Due to differences in the magnification and imaging range of different sensors, even though preliminary optical axis alignment was performed during the data collection process, view discrepancies still arise during optical fusion, leading to imperfect alignment of multimodal images. In multimodal object detection tasks [1,17,6], multiple targets need to be located and globally aligned, but radial distortion can affect alignment, especially near the image

Table 2: Object and Scene Attributes.

Object Categories	Number	Modality Attributes		Sequence Attributes				
		UV-S	RGB-S	PO	OV	DEF	LR	CM
Yellow Camouflage Net	37	19	12	3	3	0	11	37
Green Camouflage Net	157	82	39	38	18	0	24	157
Person in Camouflage	42	10	23	3	3	4	4	20

boundaries. Unlike these tasks, visual tracking focuses more on the local region of the target object. Therefore, although radial distortion issues still exist in our dataset construction, our focus is on ensuring precise alignment of the target coverage area in each frame. In our VUOT dataset, we first perform image alignment on the initial frame of each video, and then inspect each frame of the video for misalignment, manually correcting any discrepancies. Through this process, the image pairs in all video sequences achieve a good level of alignment.

Target Annotation. To construct a high-quality dataset, we adopt a combination of automatic and manual annotation methods to provide dense and precise bounding boxes for the targets. Initially, we use an automatic annotation method to convert videos into image sequences at a rate of 30 frames per second. The target position information of the first frame is input into the high-performance tracker OSTrack [16], which generates tracking predictions for all subsequent frames, thereby achieving preliminary batch target annotation. Subsequently, we employ manual annotation methods using custom annotation software to scrutinize and manually correct each frame and its initial annotations. This combination improves both the efficiency and accuracy of the annotations.

3.3 Data Attributes

Object and Scene Attributes. The VUOT dataset includes various challenging scenarios where the target foreground and background colors are confused. As shown in Fig. 2, the scenes we collected are primarily natural environments, including grasslands and jungles of different colors. The tracked targets are divided into three categories: green camouflage nets, yellow camouflage nets, and people in camouflage clothing. For stationary targets, we collected videos by moving the camera. To account for real-world tracking challenges and enrich the diversity of the dataset, we defined five sequence-level challenge attributes: Partial Occlusion (PO), Out of View (OV), Low Resolution (LR), Camera Motion (CM), and Deformation (DEF). Some sequences may contain multiple challenge attributes. Additionally, to facilitate the evaluation of the effectiveness of the two modalities, we defined two modality attributes based on visual contrast effects: "UV-Salient (UV-S)" and "Visible-Salient (RGB-S)". Detailed data statistics are shown in Table 2.

Bounding Box Attributes. We conducted a detailed analysis of the distribution position and size of the bounding boxes. First, from the distribution of the first frames of the video sequences, as shown in Fig. 3 (a), the targets

Fig. 2: (a) Target with yellow camouflage net, (b) Target with green camouflage net, (c) Target with person in camouflage.

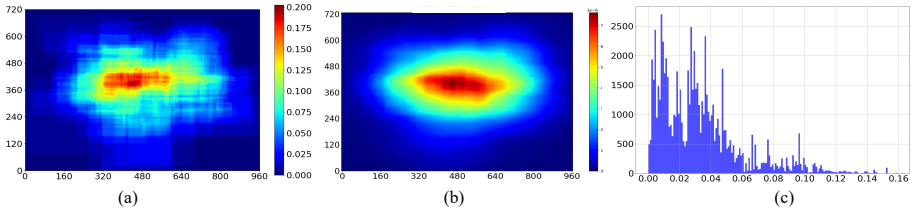

Fig. 3: (a) First Frame Bounding Box Distribution Map, (b) Distribution Map of All Bounding Boxes, (c) The Ratio of Bounding Boxes to Images.

in the initial frames are primarily concentrated in the center of the image, but there is also a certain distribution in the boundary areas of the image. This broad distribution helps to enrich the diversity of the data. We also analyzed the target distribution across all frames of the entire dataset. As shown in Fig. 3 (b), the targets generally move around the center of the image but also exhibit wide-ranging movements. This movement pattern effectively reflects the real-world relationship between the target and the distance to the image. Moreover, this data collection mainly covers close and medium-distance scenes, with the distance between the collection platform and the target ranging from 20 to 50 meters. The ratio distribution of target sizes is quite extensive, increasing the diversity of target sizes and further enriching the characteristics of the targets, as shown in Fig. 3 (c).

3.4 Evaluation Metrics

In our dataset, we use three widely recognized tracking algorithm evaluation metrics [10]: Precision Rate (PR), Normalized Precision Rate (NPR), and Success Rate (SR).

• Precision rate (PR). The precision rate is to calculate the percentage of frames where the distance between the predicted position and the ground truth is within a certain threshold range. In this work, due to the close-range scenes and relatively large targets in our dataset, we set the threshold to 10 pixels to compute a representative PR score.

• Normalized precision rate (NPR). Since the precision metric is easily affected by the image resolution and the size of the bounding box, we further normalized the precision as the second metric. For detailed calculation of NPR, please refer to [14].

• Success rate (SR). The success rate is to calculate the ratio of successful frames where the overlap between the predicted bounding box and the ground truth is greater than a certain threshold. In this work, we employ the area under curve to compute the representative SR score.

These metrics are utilized to evaluate the performance of the proposed VUTrack method and other trackers on the VUOT dataset, as shown in Table 3. Additionally, the impact of the adaptive feature selection mechanism on these metrics is analyzed in Table 4.

4 Visible and Ultraviolet Object Tracking Network

In this part, we introduce an object tracking network based on visible and ultraviolet, namely VUTrack. The overall framework of our method is shown in Fig. 4. First, the input visible (RGB) and ultraviolet (UV) search region and template images are segmented and flattened as a sequence of patches (tokens), which are then input into a series of shared Transformer blocks for feature extraction and search-template matching within each modality. We propose an adaptive feature selection mechanism based on modal reliability, which is inserted between the Transformer blocks. This mechanism selects the modality information most suitable for the current tracking scene and discards the less efficient modality information. Finally, the tracking head obtains the combined RGB and UV search region features from the backbone network and completes the target prediction.

4.1 Multimodal ViT for RGB-UV Tracking

Based on the powerful feature extraction capabilities of ViT, this algorithm extends the ViT network as the backbone for multimodal object tracking and performs joint feature extraction and search template matching for RGB and UV images through multiple layers of Transformer blocks. Here, $X_{rgb}^{image}, X_{uv}^{image} \in \mathbb{R}^{H_x \times W_x \times 3}$ represent the RGB and UV search region images, and $Z_{rgb}^{image}, Z_{uv}^{image} \in$

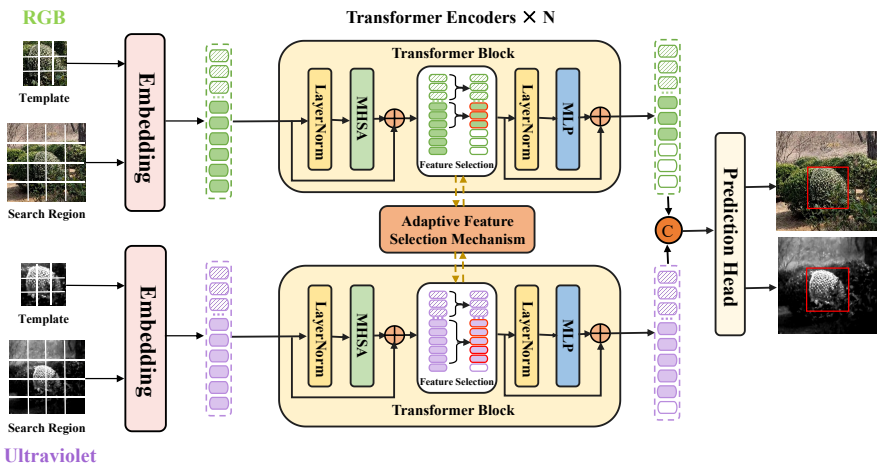

Fig. 4: The overall architecture of VUTrack. The template and search region are split, flattened, and linear projected through dual embedding layers. Image emeddings are inputted into the Transformer blocks for feature extraction and search-template matching within each modality. The adaptive feature selection mechanism retains more reliable modality features, reducing the interference of less efficient modality features. Finally, the combined RGB and UV search region features are fed into the tracking head to complete the target prediction.

$\mathbb{R}^{H_z \times W_z \times 3}$ represent the RGB and UV target template images, where the image resolutions for the different modalities are assumed to be aligned. First, these images are divided into patches of size $P \times P$ and each is flattened into a sequence of 4 patches, $X_{rgb}, X_{uv} \in \mathbb{R}^{N_x \times (3P^2)}$ and $Z_{rgb}, Z_{uv} \in \mathbb{R}^{N_z \times (3P^2)}$, where $N_x = \frac{H_x W_x}{P^2}, N_z = \frac{H_z W_z}{P^2}$ denote the number of patches for the search region and template, $H_x W_x$ and $H_z W_z$ denote the height and width of the search region and template, respectively. Then, we use linear projection layers $E_{rgb} \in \mathbb{R}^{(3P^2) \times D}$ and $E_{uv} \in \mathbb{R}^{(3P^2) \times D}$ to project $X_{rgb}, X_{uv}, Z_{rgb}, Z_{uv}$ to a D-dimensional latent space, while adding the positional information P^x_{rgb}, P^x_{uv} and P^z_{rgb}, P^z_{uv} to the patches. The result of the projection is illustrated as follows:

$$Z'_{rgb} = [Z^1_{rgb}E_{rgb}, Z^2_{rgb}E_{rgb}, \ldots, Z^{N_z}_{rgb}E_{rgb}] + P^z_{rgb}, \tag{1}$$

$$Z'_{uv} = [Z^1_{uv}E_{uv}, Z^2_{uv}E_{uv}, \ldots, Z^{N_z}_{uv}E_{uv}] + P^z_{uv}, \tag{2}$$

$$X'_{rgb} = [X^1_{rgb}E_{rgb}, X^2_{rgb}E_{rgb}, \ldots, X^{N_x}_{rgb}E_{rgb}] + P^x_{rgb}, \tag{3}$$

$$X'_{uv} = [X^1_{uv}E_{uv}, X^2_{uv}E_{uv}, \ldots, X^{N_x}_{uv}E_{uv}] + P^x_{uv}. \tag{4}$$

Then, the RGB and UV tokens are concatenated to form $H_{rgb} = [X'_{rgb}; Z'_{rgb}] \in \mathbb{R}^{(N_x+N_z) \times C}$ and $H_{uv} = [X'_{uv}; Z'_{uv}] \in \mathbb{R}^{(N_x+N_z) \times C}$, and they are respectively input into a series of Transformer blocks for joint feature extraction and search template matching for the multimodal features. Since the operations on the RGB and UV modalities are similar, we describe ViT in detail using the RGB modality as an example. For simplicity, the subscript _rgb is omitted hereafter.

In each Transformer block, matrix H first undergoes three projections to obtain the query Q, key K, and value V. Then attention weights are calculated using matrix multiplication to aggregate features, as shown below:

$$A = Softmax\left(\frac{[Q_z; Q_x][K_z; K_x]^T}{\sqrt{d_k}}\right) \tag{5}$$

$$= Softmax\left(\frac{\left([Q_z K_z^T, Q_z K_x^T; Q_x K_z^T, Q_x K_x^T]\right)}{\sqrt{d_k}}\right) \tag{6}$$

Here, W_{zx} represents the similarity measure between the target template and the search region, while the rest are classified similarly. The output A can further be expressed as:

$$A = [W_{zz}V_z + W_{zx}V_z; W_{xz}V_z + W_{xx}V_x] \tag{7}$$

On the right side of Equation (7), $W_{xz}V_z$ is responsible for the aggregation of features between the images, which corresponds to the feature extraction process of the search region. Therefore, as the stack of Transformer blocks continues, the features and matching relationships between the search region and the target template tokens are gradually extracted, facilitating the localization of the target object within each modality. To simplify the model and reduce computational overhead, the parameters of the Transformer blocks are shared among the RGB and UV modalities, avoiding redundant calculations and model redundancy.

4.2 Adaptive Feature Selection Mechanism

We propose an adaptive feature selection mechanism based on modality reliability, which helps the ViT network in the early stages to select more suitable modality search region features for the current tracking scene and remove more background features of less efficient modalities. This approach alleviates the computational burden and reduce the negative impact of background noise from less efficient modalities on feature learning. We repeatedly insert the adaptive feature selection mechanism between the Transformer blocks of the dual-stream ViT backbone, controlling the feature selection modules of the two modalities by computing modality reliability scores. Fig. 5 illustrates the adaptive feature selection mechanism based on modality reliability.

Feature Selection Module. Previous trackers retain all candidate objects during feature extraction and relation modeling, recognizing background regions only in the final output of the network. However, we adopt the candidate elimination idea from OSTrack [16], retaining the target candidate features of each modality in the early stages of ViT and gradually eliminating features that belong to the background. We use multi-head self-attention in ViT to generate multiple similarity scores. We average the similarity scores of all attention heads to obtain the final similarity scores for the target and candidate regions of the two modalities. A candidate is more likely to be a background region if its similarity score with the target is relatively small. Therefore, we only retain the candidate feature objects corresponding to the top k similarity scores (where

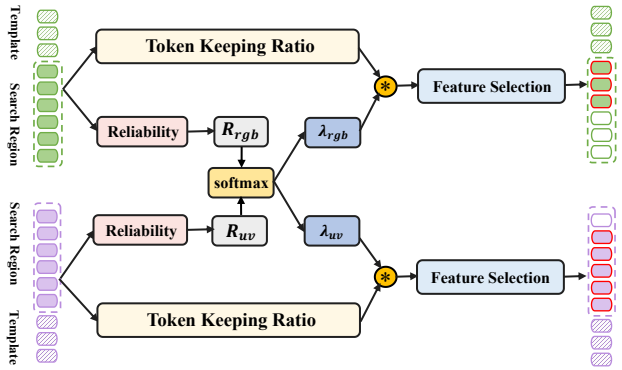

Fig. 5: Adaptive Feature Selection Mechanism

k is a hyperparameter defined as the retention rate $\rho = \frac{k}{n}$), and discard the remaining candidate features. Additionally, we record the original order of all candidates so that, after completing the feature selection, we can restore the original order of the selected feature objects and fill the missing positions with zeros.

Adaptive Feature Selection Mechanism. To fully utilize the modality-specific feature information in the current scene, this paper proposes an adaptive feature selection mechanism. As shown in Equation (7), during the feature selection stage, the similarity S between the target template and the search region is calculated through multi-head attention, and the average of the lowest N_z candidate regions with the highest similarity is selected to calculate the mean difference score, where N_z represents the number of tokens in the target template. The smaller the mean difference, the higher the reliability of the target token within the background region. Thus, the reliability score R_{rgb} and R_{uv} for different modalities is defined as:

$$R = \frac{\sum_{i=1}^{N_z}(s_i^{max} - s_i^{min})}{N_z} \tag{8}$$

By calculating the reliability scores R_{rgb} and R_{uv} for different modalities and then applying $softmax$ processing, we obtain the adaptive weights λ_{rgb} and λ_{uv}, as follows:

$$\lambda_{rgb}, \lambda_{uv} = softmax(R_{rgb}, R_{uv}) \tag{9}$$

The adaptive weights λ_{rgb} and λ_{uv} are combined with the initial modality retention rates ρ_{rgb} and ρ_{uv} to calculate the adaptive feature retention rates ρ'_{rgb} and ρ'_{uv}:

$$\rho'_{rgb} = \lambda_{rgb} * (\rho_{rgb} + \rho_{uv}) \tag{10}$$

$$\rho'_{uv} = \lambda_{uv} * (\rho_{rgb} + \rho_{uv}) \tag{11}$$

Table 3: Comparison with state-of-the-art methods on the VUOT Dataset. The values within parentheses represent the changes in scores after introducing the UV modality. The best two results are shown in red and blue fonts.

Method	Source	RGB Modality			RGB-UV Modality			FPS
		SR	PR	NPR	SR	PR	NPR	
SiamCDA[19]	TCSVT'21	62.8	51.9	63.9	66.6(+3.9%)	59.2(+7.3%)	68.3(+4.3%)	26.2
HMFT[18]	CVPR'22	63.2	51.4	64.9	66.9(+3.7%)	60.4(+9.0%)	68.6(+3.7%)	31.7
VIPT[20]	CVPR'23	64.4	54.1	65.5	72.5(+8.1%)	63.0(+8.9%)	71.9(+6.4%)	35.3
TBSI[5]	CVPR'23	70.2	58.0	68.7	74.2(+4.0%)	65.3(+7.3%)	75.8(+7.2%)	40.0
UVTrack	Ours	72.1	63.8	71.1	76.7(+4.6%)	68.3(+5.5%)	77.3(+6.2%)	76.9

To prevent the excessive elimination of features, the number of retained features is set to a maximum of N_z. The adaptive feature selection mechanism can retain more effective modality features, reducing the interference of ineffective modalities, and thereby improving inference speed while maintaining high inference performance.

5 Experiments

5.1 Evaluation on VUOT Dataset

To our knowledge, there are currently no publicly available trackers based on visible and ultraviolet modalities. Through experiments, we found that dual-modality RGB-T trackers can also achieve good modality complementarity effects on the VUOT dataset. Therefore, to demonstrate the effectiveness of our VUOT dataset, we selected four popular dual-modality RGB-T trackers (SiamCDA [19], HMFT [18], VIPT [20], TBSI [5]) for testing. As shown in Table 3, we didn't modify any parameters of the four RGB-T trackers and used the published network models of each tracker to retrain and test on the VUOT dataset, evaluating the trackers' precision rate, normalized precision rate, and success rate. Additionally, to compare the effectiveness of different modalities in the VUOT dataset, we analyzed and compared two modes during the testing phase (the first mode: RGB only, the second mode: RGB and UV dual modalities). The experimental results show that the introduction of the UV modality achieves an average improvement of 7.99%, 5.75%, and 4.85% in PR, NPR, and SR scores for the four trackers (SiamCDA, HMFT, VIPT, TBSI), respectively. These experimental results demonstrate the effectiveness of the UV modality in the VUOT dataset, showing that the UV modality has great potential in the field of object tracking.

5.2 Experimental Analysis of VUTrack

Our model is implemented using PyTorch [15].To ensure fairness, our method and other methods were experimented on using an NVIDIA RTX 4090 GPU. We adopt AdamW as the optimizer, with a weight decay of 1e-4 and a learning rate

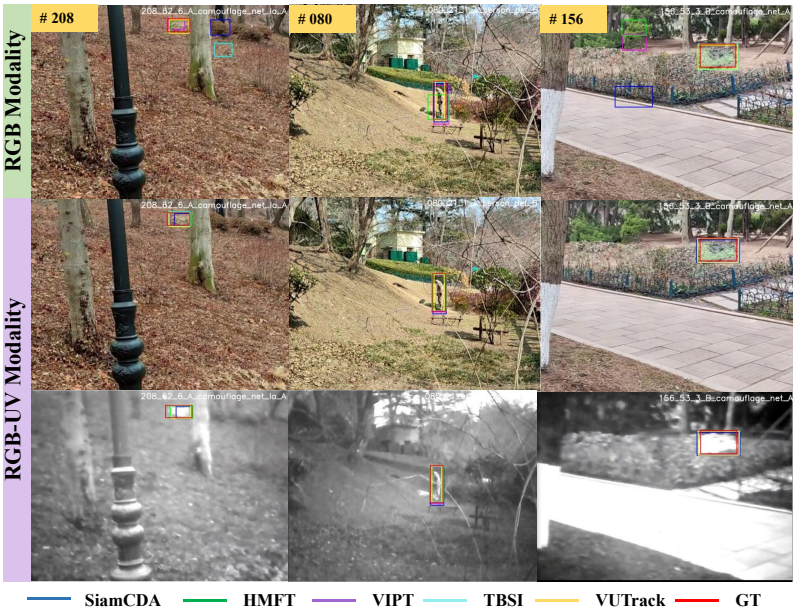

Fig. 6: Visual comparison between our method and other trackers using data from the VUOT dataset. The RGB modality indicates that the tracker is tested using only RGB data, while the RGB-UV modality indicates that the tracker is tested using RGB data supplemented with UV data. Best viewed in color.

of 4e-4 for the backbone. The size of the search region was adjusted to 256×256, and the template size was adjusted to 128×128. Each batch size is set to 16, and each epoch consisted of 12k image pairs. We pretrained the network on RGB tracking datasets such as COCO [12], LaSOT [3], GOT-10k [4], and TrackingNet [14], then trained on the VUOT training set and tested on the VUOT test set. AS shown in Table 3 and Fig. 6, our VUTrack network outperformed the other 4 trackers in both performance and speed. Compared to the state-of-the-art method TBSI, we achieved improvements of 2.5%/3.0%/1.5% in SR/PR/NPR, and our inference speed is 1.9 times that of TBSI. The experimental results fully demonstrate the effectiveness and efficiency of VUTrack. It is worth noting that, to the best of our knowledge, VUTrack is the first method for visible light and ultraviolet object tracking. Our method shows excellent adaptability to the dynamic relationship between the RGB and UV modalities.

5.3 Ablation Study and Analysis

To verify the effectiveness of the adaptive feature selection mechanism based on modality reliability, we conducted comparative experiments with and without the adaptive feature selection mechanism. The effectiveness of the adaptive feature selection module in VUTrack was assessed in terms of inference speed

(FPS), Multiply-Accumulate Operations (MAC), and tracking performance. As shown in Table 4, the introduction of the adaptive feature selection mechanism reduced MACs by 25.6%, increased tracking speed by 70.2%, increased SR by 0.12%, PR by 0.2%, and NPR by 0.17%. The results indicate that incorporating an adaptive feature selection mechanism can significantly improve inference speed while ensuring better tracking performance.

Table 4: Comparison with and without the Adaptive Feature Selection Mechanism.

Method	FPS	MACs(G)	SR	PR	NPR
Without Adaptive Feature Selection Mechanism	45.21	59.16	76.60	68.09	77.10
Adaptive Feature Selection Mechanism	76.94	44.02	76.72	68.29	77.27

6 Conclusion

In this paper, we constructed the first dataset for visible and ultraviolet object tracking (VUOT). By utilizing the characteristics of UV imaging, we used UV as a complementary means to RGB to address tracking issues caused by target and background color confusion under target camouflage. Additionally, we conducted exploratory research in the field of RGB-UV fusion object tracking and proposed an efficient transformer network for visible and ultraviolet object tracking (VUTrack). By introducing an adaptive feature selection mechanism based on modality reliability, we can achieve efficient inference speed while ensuring better tracking performance. Analysis of the research results indicates that the UV modality can supplement RGB with additional information in complex outdoor environments. This has potential for broad applications in the military in the future, playing a critical role in areas such as drone reconnaissance, object tracking, and military operations.

Limitations. The UV sensor is highly sensitive to lighting conditions; under low-light conditions, the UV imaging performance is significantly affected. In the future, we plan to explore incorporating infrared into our VUTrack algorithm. By leveraging the complementary information from three modalities, we aim to develop a multimodal fusion tracking method that can adapt to complex scenes and deliver improved tracking performance.

References

1. Choi, H., Kim, S., Park, K., Sohn, K.: Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In: 2016 23rd International conference on pattern recognition (ICPR). pp. 621–626. IEEE (2016)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5374–5383 (2019)
4. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE transactions on pattern analysis and machine intelligence **43**(5), 1562–1577 (2019)
5. Hui, T., Xun, Z., Peng, F., Huang, J., Wei, X., Wei, X., Dai, J., Han, J., Liu, S.: Bridging search region interaction with template for rgb-t tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13630–13639 (2023)
6. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1037–1045 (2015)
7. Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kamarainen, J.K., ˇCehovin Zajc, L., Drbohlav, O., Lukezic, A., Berg, A., et al.: The seventh visual object tracking vot2019 challenge results. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp. 0–0 (2019)
8. Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., Lin, L.: Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Transactions on Image Processing **25**(12), 5743–5756 (2016)
9. Li, C., Liang, X., Lu, Y., Zhao, N., Tang, J.: Rgb-t object tracking: Benchmark and baseline. Pattern Recognition **96**, 106977 (2019)
10. Li, C., Xue, W., Jia, Y., Qu, Z., Luo, B., Tang, J., Sun, D.: Lasher: A large-scale high-diversity benchmark for rgbt tracking. IEEE Transactions on Image Processing **31**, 392–404 (2021)
11. Li, C., Zhao, N., Lu, Y., Zhu, C., Tang, J.: Weighted sparse representation regularized graph learning for rgb-t object tracking. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1856–1864 (2017)
12. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
13. Lu, H., Wang, D.: Online visual tracking. Springer (2019)
14. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the European conference on computer vision (ECCV). pp. 300–317 (2018)
15. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

16. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: European conference on computer vision. pp. 341–357. Springer (2022)
17. Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z.: Weakly aligned cross-modal learning for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5127–5137 (2019)
18. Zhang, P., Zhao, J., Wang, D., Lu, H., Ruan, X.: Visible-thermal uav tracking: A large-scale benchmark and new baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8886–8895 (2022)
19. Zhang, T., Liu, X., Zhang, Q., Han, J.: Siamcda: Complementarity-and distractor-aware rgb-t tracking based on siamese network. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1403–1417 (2021)
20. Zhu, J., Lai, S., Chen, X., Wang, D., Lu, H.: Visual prompt multi-modal tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9516–9526 (2023)