This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multi-Object Tracking in the Dark

Xinzhe Wang Kang Ma Qiankun Liu Yunhao Zou Ying Fu [*]
Beijing Institute of Technology
{wangxinzhe, makang, liuqk3, zouyunhao, fuying}@bit.edu.cn

Abstract

*Low-light scenes are prevalent in real-world applications (*e.g. *autonomous driving and surveillance at night). Recently, multi-object tracking in various practical use cases have received much attention, but multi-object tracking in dark scenes is rarely considered. In this paper, we focus on multi-object tracking in dark scenes. To address the lack of datasets, we first build a* ***L****ow-light* ***M****ulti-****O****bject* ***T****racking (****LMOT****) dataset. LMOT provides well-aligned low-light video pairs captured by our dual-camera system, and high-quality multi-object tracking annotations for all videos. Then, we propose a low-light multi-object tracking method, termed as* ***LTrack****. We introduce the adaptive low-pass downsample module to enhance low-frequency components of images outside the sensor noises. The degradation suppression learning strategy enables the model to learn invariant information under noise disturbance and image quality degradation. These components improve the robustness of multi-object tracking in dark scenes. We conducted a comprehensive analysis of our LMOT dataset and proposed LTrack. Experimental results demonstrate the superiority of the proposed method and its competitiveness in real night low-light scenes. Dataset and Code: https://github.com/ying-fu/LMOT*

1. Introduction

Multi-object tracking (MOT) aims to locate and associate multiple objects in video sequences. It is widely used in many downstream applications, such as video recognition [7, 32], autonomous driving [18], and surveillance [11]. Recently, multi-object tracking in various practical use cases has garnered much attention [8, 9, 18, 34, 42], greatly advancing the development of MOT. However, these works are primarily tailored for high-quality inputs and overlook the prevalent low-light scenes in real-world scenes. Motivated by this, we study multi-object tracking in dark scenes.

Due to the physical limitations of existing cameras, acquiring high-quality videos under low-light conditions is difficult. One inherent difficulty in capturing consecutive video frames under such conditions is avoiding motion blur. Current camera technology typically requires short exposure times (usually just a few tens of milliseconds), but in low-light scenarios, the sensor struggles to capture an adequate number of photons within this limited duration. This limitation inevitably leads to degradation in image quality accompanied by higher noise levels. This presents two main challenges for MOT under low-light conditions. The first challenge is for collecting a low-light multi-object tracking dataset. Collecting and annotating a low-light MOT dataset is difficult and expensive. MOT requires dynamic object videos, but videos captured in low-light scenes have extremely low brightness, making it hard to recognize and annotate objects in the videos. The second challenge revolves around low-light multi-object tracking. The popular *tracking-by-detection* paradigm [3, 10, 29, 30, 62] generally consists of detector, motion-based association module, and appearance-based association module. These modules typically require high-quality input images. The poor quality of low-light images leads to severe performance degradation for both detectors and appearance-based correlation modules. A simple approach is to cascade low-light enhance modules [2, 23, 24, 47], but this introduces additional computational costs. Furthermore, images optimized for visual quality may be suboptimal for downstream tasks [6, 21, 27].

In this paper, we build a low-light multi-object tracking dataset (LMOT), specifically designed to address the challenges of multi-object tracking in dark scenes. To this end, we develop a dual-camera system that simultaneously captures well-lit and low-light video frames. The video pairs are highly aligned in both spatial and temporal dimensions, offering two key benefits. First, it enables us to annotate on the well-lit videos, resulting in high-quality annotations. Second, the well-lit videos can provide additional supervision information during the training phase, and strongly enhance the performance in the dark scenes. After careful annotation, we collect 32 video sequences (2.3× MOT17), over 35K frames (3.1× MOT17), and over 815K bounding boxes (2.8× MOT17). The RAW data is the output of

[*] Corresponding Author

the image sensor and is the input data of the image signal processor (ISP). It saves all information from the image sensor, which is crucial for capturing object information in dark scenes [6, 52, 67]. Therefore, we collect RAW videos for LMOT.

Additionally, we propose a low-light multi-object tracking method, termed as LTrack. The low-light video is characterized by substantial sensor noise and poor image quality, which significantly degrades both shallow and deep feature representations, leading to reduced tracking performance. We observe that the sensor noise in low-light images exhibits a similarity to adversarial attacks [35, 43]. To address this issue, *our main idea is to learn the invariant semantic information under noise disturbance quality degradation*. We present the adaptive low-pass downsample module (ALD). It employs spatial low-pass convolution to extract low-frequency components from images, excluding noises, and adaptively enhance the feature maps. We also present the degradation suppression learning strategy (DSL), which utilizes paired low-light videos to help the model suppress the noise disturbance and encourage image content response in the feature domain. We conduct a comprehensive analysis of our LMOT dataset and validate the superiority of our LTrack in real-world night scenes.

In summary, our main contributions are as follows:

- We build the first low-light multi-object tracking dataset using a carefully constructed dual-camera system. It provides well-aligned low-light videos in RAW format, and high-quality MOT annotations for all videos.
- We propose a low-light multi-object tracking method. It utilizes the adaptive low-pass downsample module and the degradation suppression learning strategy to learn to extract invariant features from low-light videos.
- We conduct a comprehensive analysis of our dataset and the proposed method. Experimental results demonstrate the superiority of the proposed method and its competitiveness on real night scenes.

2. Related Work

In this section, we first review the current research status of low-light enhancement and low-light datasets. Then, we summarized the present research for multi-object tracking and tracking in the dark scenes.

Low-light enhancement. Traditional methods for low-light enhancement are primarily based on histogram equalization and Retinex theory [20, 22, 25]. Recently, deep learning has been explored for many low-level tasks [14, 44, 45, 51, 60, 61, 65, 66], and achieved superior results on low-light enhancement [2, 4, 13, 15, 19, 24, 39, 53]. While these methods are capable of recovering images with high visual quality, they often require heavy computation and may not consider downstream tasks, resulting in suboptimal performance. In contrast, our approach focuses on directly learning multi-object tracking from low-light images, thus bypassing the low-light enhancement.

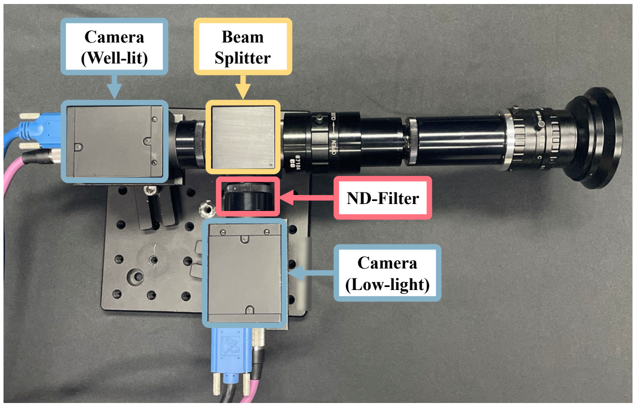


Figure 1. Our dual-camera system. It consists of two cameras, a beam splitter, and an ND-filter. Two cameras of identical models are meticulously engineered to achieve pixel-by-pixel alignment in the captured video data.

Low-light datasets. The long-short exposure is a widely used method to collect paired low-light images, but can only be used to collect low-light images for static scenes. [4, 5, 49]. To capture dynamic scenes, some works designed the mechatronic system. They obtain paired low-light data by repeating the motion twice [13, 47]. However, these mechatronic systems cannot be used to collect dynamic object videos in the wild. Jiang *et al.* [23] designed a dual-camera system that simultaneously captures paired well-lit and low-light videos, making it possible to capture dynamic scenes and dynamic object videos for multi-object tracking. Zou *et al.* [57] setup an optical system to collect paired videos and event streams. These works explore various ways to construct low-light datasets and inspire research on high-level vision tasks in dark scenes, such as LOD [21] for object detection, LIS [6] for instance segmentation, and ExPose [27] for human pose estimation. These datasets provide paired low-light data only for image tasks, and cannot be expanded for multi-object tracking in dark scenes.

Multi-object tracking datasets. MOT15 [26] is the first large-scale benchmark for multi-object tracking. MOT17 [34] stands as one of the most widely applied MOT benchmarks. MOT20 [9] focuses on very crowded scenes. These three datasets are for pedestrians. KITTI [18] and BDD100K [55] are for autonomous driving scenarios. DanceTrack [42] focuses on dancing scenes and is characterized by similar appearance and diverse motions. Recently, SportsMOT [8] aims to track athletes and encourage algorithms to promote both appearance and motion association. These datasets explore multi-object tracking in various practical use cases, but none of them consider multi-object tracking in dark scenes.

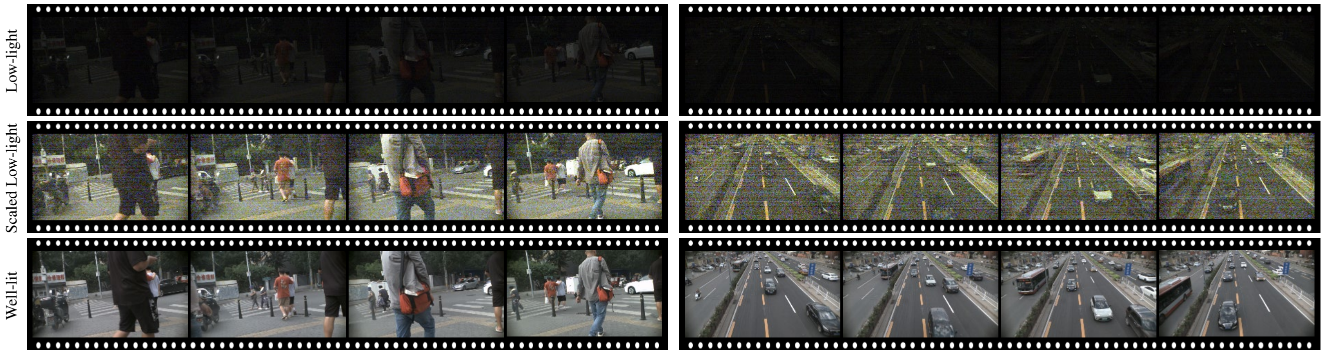


Figure 2. Two example videos from our LMOT dataset. It provides well-aligned low-light video pairs and MOT annotations for all videos. The time interval between adjacent frames is $1s$. The first row is the low-light video, the second row is the scaled low-light video and the last row is the well-lit video. Our LMOT dataset is collected from city outdoor scenes.

Dataset	Format	Videos	Frames	Length (s)	Bbox	Tracks
MOT17 [34]	sRGB	14	11,235	463	292,733	1,342
MOT20 [9]	sRGB	8	13,410	535	1,652,040	3,456
DanceTrack [42]	sRGB	100	105,855	5,292	-	990
SportMOT [8]	sRGB	240	150,379	6,015	1,629,490	3,401
KITTI [18]	sRGB	21	8,000	-	47,000	917
BDD100K [55]	sRGB	1,600	318,000	-	3,300,000	131,000
SWIR [36]	SWIR	-	7,309	-	57,221	-
LMOT	RAW	32	35,120	1,756	815,550	4,090

Table 1. Comparison of statistics between existing MOT datasets and our LMOT dataset

Object tracking in the dark. To track in low-light scenes, some methods explore to use multi-modal information for single-object tracking, such as event camera [58, 64], depth [40, 54] and thermal [46, 59] devices. Park *et al.* [36] proposed to use Short-Wave Infrared (SWIR) images for multi-object tracking, since its advantages in terms of robustness in low-light conditions. The common drawback of these methods is that they require additional hardware equipment and cannot be applied to the most widely used CMOS imaging systems.

SORT [1] uses the Kalman Filter motion model and employs IoU for association. ByteTrack [62] enhances tracking performance by considering low-confidence bounding boxes. OS-SORT [3] enhances SORT by restoring lost targets. Recently, Transformer has been explored for MOT [16, 33, 41, 56, 63]. These methods have achieved high performance in many practical scenarios, but they do not consider dealing with low-light conditions. We focus on multi-object tracking under low-light conditions. Based on RAW videos, our method is highly practical with excellent performance and does not require additional hardware.

3. Low-light Multi-object Tracking Dataset

In this section, we first introduce our dual camera system and the details of collecting and annotating our low-light multi-object tracking (LMOT) dataset. Then, we analyzed the statistical characteristics of our LMOT dataset.

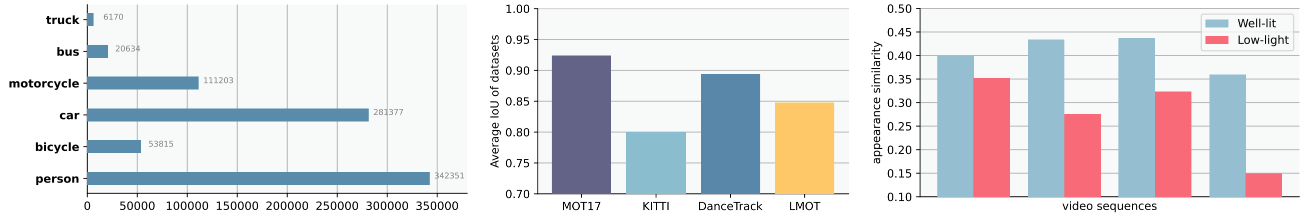
Dataset	Split	Videos	Bbox	Tracks	Paired Well-lit
LMOT-dual	train	11	309,466	1533	✓
	val	4	131,781	626	✓
	test	11	312,742	1644	✓
LMOT-real	real	6	61,561	287	

Table 2. Detailed statistics and data splits for LMOT.

3.1. Dataset Construction

Multi-object tracking requires dynamic scenes and object video. To collect low-light videos for multi-object tracking, we build a dual-camera system [23], which is illustrated in Fig. 1. It can simultaneously capture paired low-light and well-lit video pairs. Its main components include a beam splitter, a neutral density (ND) filter, and two *FLIR Grasshopper3 GS3-U3-23S6C* cameras. The beam splitter divides the incoming light into two separate paths. This arrangement allows one camera to capture well-lit images directly, while the other camera records low-light images, with the ND-filter attenuating the light intensity. To ensure temporal synchronization of the video frames, we employ the hardware interface to trigger the camera exposure events. Moreover, to avoid frame loss, our dual-camera system uses two independent hardware interfaces for data transmission and is equipped with a high-speed solid-state drive. Thanks to precise calibration, our dual-camera system can capture paired low-light and well-lit videos in real time for dynamic scenes and objects. More details about our dual-camera system are given in *supplementary materials*.

We save the video frames in RAW format before the images are processed by the camera image signal processor (ISP). In terms of camera settings, we set the exposure time for both cameras to $10ms$, and the frame rate is fixed at 20. This setting is feasible to avoid motion blur. We adjust the gain level for well-lit cameras to achieve optimal image quality. The gain for low-light cameras is consistently set to the maximum value, to simulate low-light capturing setup in real scenarios. We also collect a real low-light MOT dataset (LMOT-real) to evaluate performance in real-world



(a) Number of instances per category

(b) IoU on adjacent frames

(c) Cosine distance of appearance features

Figure 3. (a) Number of instances per category. LMOT consists of 6 categories, most of the instances are the person and car. (b) IoU on adjacent frames. Compared to MOT17, KITTI, and DanceTrack, LMOT has a roughly average score. This indicates that LMOT has a relatively normal movement speed. (c) Cosine distance of appearance features. The cosine distance is smaller under low-light conditions, indicating that the appearance distinguishability is decreased under low-light conditions.

dark scenes. These videos are captured using a single camera with the same camera settings.

Our LMOT dataset contains a variety of city outdoor scenes, including roads, overpasses, pedestrians, and intersection. The overpass scenes take an overhead shot of objects, while all other scenes are captured from the perspective of pedestrians. To account for the impact of camera motion, we introduce arbitrary horizontal rotations and vertical random movements to the camera. Fig. 2 shows two sampled video sequences from our LMOT dataset.

We annotate six types of moving objects, including car, person, bicycle, motorcycle, bus, and truck. The annotated labels include bounding boxes, identifications, and visibility status. For partly occluded objects, a full box is annotated. For a fully occluded object, an estimated box is annotated. Each object has a unique ID throughout the entire video. Thanks to well-aligned low-light and well-lit videos, we can annotate well-lit videos to simultaneously obtain labels for low-light videos. This greatly reduces the annotation difficulty and enhances quality. Lastly, we carefully review all the annotation results.

3.2. Dataset Statistic

LMOT is a large-scale dataset that focuses on multi-object tracking in dark scenes. We compare the statistics of LMOT with existing MOT datasets in Tab. 1. It can be seen that LMOT is approximately three times larger than MOT17 [34]. Compared to large-scale datasets such as DanceTrack [42], SportsMOT [25], KITTI [18] and BDD100K [55], the scale of LMOT is still considerable. It should be emphasized that these datasets are not for multi-object tracking in dark scenes and only provide sRGB images. Compared to SWIR, our LMOT dataset has approximately 5× frames and 14× bounding boxes. The detailed statistics and data splits for LMOT are shown in Tab. 2.

We present the number of instances for each category in Fig. 3 (a). The majority of instances are persons and cars. As shown in Fig. 3 (b), the average IoU on adjacent frames of LMOT is lower than MOT17 and DanceTrack, but higher than KITTI. This indicates that the motions in LMOT are fast but within normal range. Following [42], we use cosine distance of appearance features [1] to evaluate the appearance similarity. From Fig. 3 (c), we can see that the cosine distance of appearance features under low-light conditions is smaller than that under well-lit conditions. In other words, the appearance of objects will deteriorate under low-light conditions, making them harder to distinguish.

4. Low-light Multi-object Tracking

In this section, we peresnt our low-light multi-object tracking method (LTrack). Our main idea is to *learn the invariant semantic information under noise disturbance quality degradation*. The overall framework can be seen in Fig. 4.

4.1. Formulation and Motivation

In low-light scenes, the camera can capture only a small number of photons in a single exposure. Thus, the potential sensor noise is highlighted, resulting in significant degradation to the image quality [50]. We observed that directly feeding low-light images to the network without any special design leads to the feature map degradation and significantly reduces the performance of the model (see in Sec. 5.2). A direct solution is to apply low-light enhancement techniques, which focus on learning a mapping function from low-light images to clean well-lit images. Since it is a highly ill-posed problem, learning such a mapping function requires considerable computing and storage overhead. Although DNN-based methods have achieved excellent performance in low-light enhancement, images enhanced for visual quality may be suboptimal for downstream tasks.

In this work, we perform multi-object tracking from low-light images, bypassing the low-light enhancement. Leveraging RAW videos, the network obtains more original scene information compared to sRGB. To enhance the performance and robustness of the multi-object tracking model, our main idea is to earn the invariant semantic information under noise disturbance quality degradation. Thus, we present the adaptive low-pass downsampling (ALD) module

[1] We use UniTrack [12] to extract appearance features.

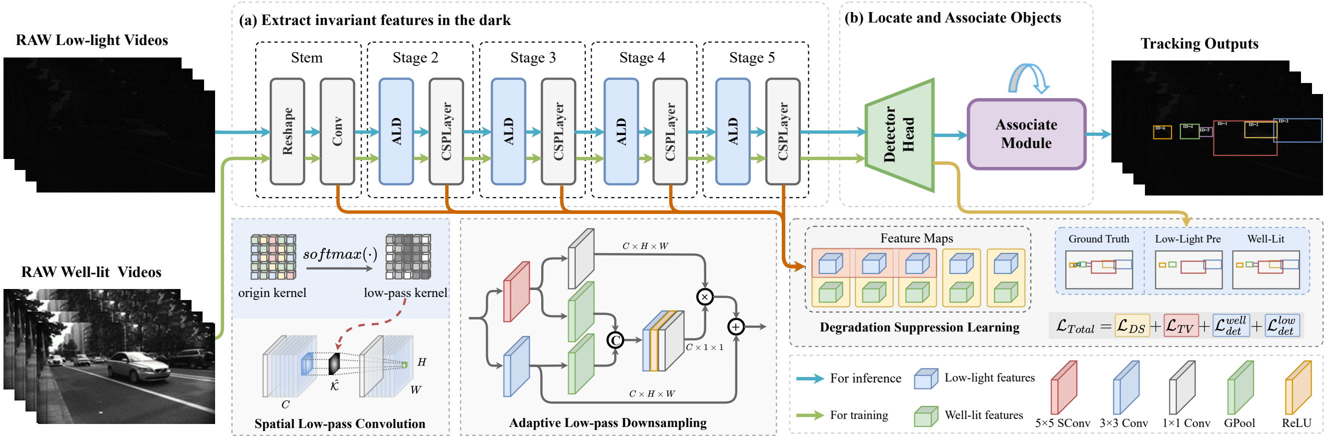


Figure 4. The overall framework of the proposed low-light multi-object tracking method, termed as **LTrack**. It employs adaptive low-pass downsample module and degradation suppression learning strategy, enabling the model to learn invariant features from low-light videos.

to enhance the low-frequency components of images and filter out high-frequency noise. We also propose the degradation suppression learning strategy (DSL), which utilizes paired low-light videos to help the model suppress image noise disturbance and encourage image content response in the feature domain.

4.2. Adaptive Low-pass Downsampling

The downsampling operation reduces the feature size while preserving the most important information. The noise in low-light images introduces high-frequency disturbance to the feature maps, which can mislead the preservation of object information. To weaken the impact of high-frequency noises and enhance the low-frequency part of the feature map, we introduce spatial low-pass convolution (SConv) to extract low-frequency features from the noised feature maps. The softmax function is used to constrain the original convolution kernel to be low-pass as

$$\hat{\mathcal{K}}_{i,j} = \frac{exp(\mathcal{K}_{i,j})}{\sum_{p,q} exp(\mathcal{K}_{p,q})} \tag{1}$$

where $\mathcal{K}$ and $\hat{\mathcal{K}}$ are the origin convolution kernel and low-pass convolution kernel. We initialize the spatial low-pass convolution kernel using a standard Gaussian kernel. The kernel size is set to 5 to obtain more spatial information. Then, the obtained low-frequency features are adaptively weighted and fused into the original features. We use global average pooling to obtain channel descriptors for both origin features and low-frequency features. These descriptors are then fed into a fully connected layer to compute the weight values.

4.3. Degradation Suppression Learning

Given that a low-light image and well-lit image pair share the same content, the model should exhibit identical feature responses to them. However, the low-light image results in shallow features full of noise, and the deep feature exhibits lower responses to objects (as shown in Fig. 5). To address this, our idea is to suppress image noise in shallow features and use well-lit images to help model learning disturbance invariant information from low-light images. The degradation suppression loss can be expressed as

$$\mathcal{L}_{DS} = \sum_{l} ||\mathbf{F}_l^{well} - \mathbf{F}_l^{low}||_2^2 \tag{2}$$

where $\mathbf{F}_l^{well}$ and $\mathbf{F}_l^{low}$ denotes l-th feature map corresponding to well-lit and low-light image, respectively. To further suppress the noise in shallow features, we also introduce the Total Variation (TV) loss [38] to features as

$$\mathcal{L}_{TV} = \sum_{l} ||\mathbf{G}^{row}\mathbf{F}_l^{low}||_2^2 + ||\mathbf{G}^{col}\mathbf{F}_l^{low}||_2^2 \tag{3}$$

where $\mathbf{G}^{row}$ and $\mathbf{G}^{col}$ are the first derivative matrix to role and column. The TV loss adds spatial smoothing constraints to features, which helps model learning to extract noise invariant features. Both the well-lit and low-light images are used to train the model by a common detection loss $L_{det}(\cdot)$. The total loss is $\mathcal{L}_{Total}$, *i.e.*,

$$\mathcal{L}_{Total} = \mathcal{L}_{det}^{well} + \alpha\mathcal{L}_{det}^{low} + \beta\mathcal{L}_{DS} + \gamma\mathcal{L}_{TV} \tag{4}$$

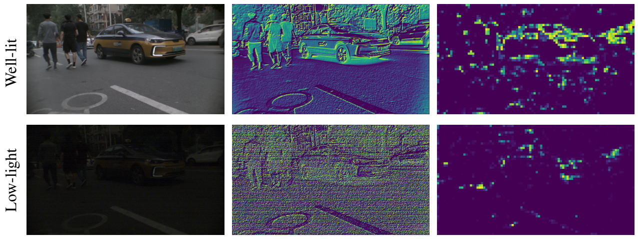
where is the $\mathcal{L}_{det}^{well}$ and $\mathcal{L}_{det}^{low}$ are the detection loss for well-lit and low-light images, α and β are loss weights. We set α, β and γ to 1, 1 and 0.01, respectively.

5. Experiments

5.1. Experiment Setup

Dataset Split. In structuring our dataset, we randomly split the videos into training, validation, and testing sets, consisting of 11, 4, and 11 videos respectively. We also provide LMOT-real dataset that is captured in real-world night scenes with 6 videos. Detailed dataset split and statistical information are shown in Tab. 2.

Evaluation Metrics. Following [8, 42], we recommend using HOTA [31] as the main evaluation metric to simulta-



(a) Input (b) Shallow feature (c) Deep feature

Figure 5. Visualization of shallow and deep features for well-lit and low-light images. It can be seen that, under low-light conditions, the shallow feature is full of noise, and the deep feature exhibits lower responses to objects.

neously evaluate the performance of detection and association. We also employ AssA and IDF1 [37] to evaluate association performance, MOTA, and Deta to evaluate detection performance. There are two ways to combine metrics for all classes into a single score. One is by averaging metrics over the class values, and the other is by over the detection values. To avoid possible result bias caused by some categories with fewer samples (such as Bus and Truck), we combine scores by averaging over the detection values.

Implementation Details. Following ByteTrack [62] and OC-SORT [3], we use YOLOX [17] as our detector. The trackers are pre-trained on the COCO [28] dataset and then trained on our LMOT dataset for 24 epochs. We apply data augmentation strategies including random flip, scale jitter of resizing, and Mosaic. We also use the physical-based noise model [50] for RAW image augment. We use SGD optimizer with weight decay 10^{-4} and cosine learning rate schedule, the initial learning rate is 10^{-4} and gradually reduces to 10^{-5}. We apply linear interpolation as post-processing to all trackers, with the maximum gap set to 20.

5.2. Analysis under low-light conditions

We analyze the impact of low-light conditions on multi-object tracking using the LMOT validation set. It should be emphasized that the low-light and well-lit video pairs are perfectly aligned in LMOT.

Impact to detectors. We first analyze the impact of lighting conditions on the detector. We select YOLOX as the detector since it is widely used in MOT areas [3, 8, 62]. We train the detector using well-lit images (WL), low-light images (LL), and all the images (AL). Then, test them on both well-lit and low-light images. The results are shown in Tab. 3. From the table, we can see that the model trained by well-light images achieves the best result on well-light images, but its performance significantly decreased on low-light images. Further, we visualized the feature maps under these two lighting conditions in Fig. 5. It can be seen that both the shallow feature and deep features of the low-light image are significantly degraded due to the sensor noises. We also observed that using low-light images can largely improve the performance of the decoder under low-light conditions, and use all images to achieve the best performance. But this result is much lower than that under well-lit conditions.

Training data		Well-lit			Low-light		
WL	LL	mAP	mAR	AP50	mAP	mAR	AP50
✓		**37.0**	**45.0**	**65.5**	3.1	4.8	6.6
	✓	23.0	30.8	40.8	16.9	23.7	30.3
✓	✓	28.7	35.2	49.1	**17.9**	**24.1**	**32.2**

Table 3. Analysis on LMOT validation set for detector (YOLOX [17]). WL and LL indicate the well-lit and low-light, respectively. It shows that it is hard to detect objects under low-light conditions.

Cond.	Mot.	App.	HOTA	AssA	IDF1	MOTA	DetA
-	✓		86.1	**79.5**	72.9	82.6	93.4
WL		✓	**87.3**	77.1	82.5	**98.8**	**98.4**
WL	✓	✓	83.7	78.0	**85.6**	96.8	90.0
LL		✓	53.8	32.1	40.0	67.3	90.2
LL	✓	✓	83.4	77.5	85.0	96.0	89.8

Table 4. Analysis on LMOT validation set for different association models. Cond, Mot, and App indicate the light condition, motion, and appearance, respectively. The detection boxes are ground-truth boxes. It indicates that appearance information is effective for association under well-lit conditions, but dose not as effective under low-light conditions.

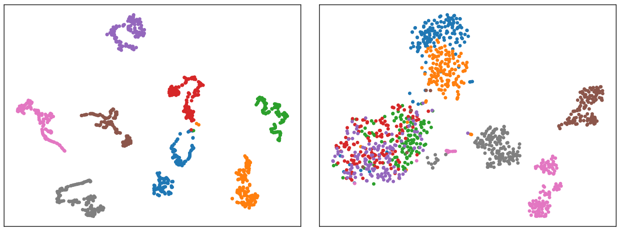
Impact to association modules. We analyze the impact of low-light conditions on the object association modules. Motion and appearance are important for object association. Both of them rely on detection boxes to locate the objects. To separate the impact of the detector, we use the ground-truth boxes as the detection boxes. The results are shown in Tab. 4. It can be seen that using only appearance matching under well-lit conditions achieves the best result while using only appearance matching under low-lit conditions results in very poor performance. This indicates that the objects in LMOT have obvious visual distinguishability, but the distinguishability is significantly reduced under low-light conditions. In Fig. 6, we visualize the appearance feature of objects in the LMOT dataset under both well-lit and low-light conditions. We can observe that under well-lit conditions, LMOT is very distinguishable in the feature space. However, under low-light conditions, this discrimination of LMOT decreased significantly.

5.3. Results on LMOT dataset

We compare the proposed method with potential low-light multi-object tracking methods. We first train three baseline models using low-light videos, well-lit videos, and both of them, without any modifications to baseline trackers. They are denoted by Base-low, Base-well, and Base-all. The second type of method is tracking after low-light enhancement. These methods use low-light enhancement methods as pre-processing module. Both the enhanced videos and well-lit

Method	ByteTrack[62]			OC-SORT[3]			Detection			Params (M)	FLOPS (G)
	HOTA	AssA	IDF1	HOTA	AssA	IDF1	mAP	AP50	AP75		
Base-low	28.0	42.9	10.9	28.8	40.3	34.3	18.8	32.4	18.9	99.00	793.29
Base-well	13.5	36.8	32.9	13.8	31.7	12.0	5.5	9.7	5.5	99.00	793.29
Base-all	28.1	42.6	32.9	28.9	39.8	34.6	19.0	32.5	19.1	99.00	793.29
LLFlow [48]	28.8	42.9	34.0	29.4	41.0	35.0	22.4	38.3	22.6	116.42	8755.91
SDSD [47]	29.1	41.1	**35.5**	29.3	36.9	**36.8**	19.8	36.2	19.3	103.30	1136.39
DNF [24]	27.6	40.9	33.0	28.1	36.4	35.3	19.1	35.6	18.2	101.83	907.33
SMOID [23]	28.0	39.7	33.9	28.7	36.7	35.6	19.1	35.5	18.2	120.88	7286.17
RAOD [52]	26.1	42.9	29.5	26.1	**42.9**	29.5	18.3	31.3	18.4	99.07	897.04
LTrack (Ours)	**29.4**	**43.2**	35.2	**29.8**	39.0	36.7	**23.4**	**40.6**	**23.5**	100.43	800.92

Table 5. Experimental results on LMOT dataset. Base-low, Base-well and Base-all indicate the baseline model trained on low-light, well-lit and all of them, respectively. The number of parameters and prediction latency of each method are reported along with the accuracy


(a) Well-lit

(b) Low-light

Figure 6. Visualization of appearance features of LMOT dataset using t-SNE. The same object is coded by the same color. It indicates that the appearance of objects under well-lit is distinguishable, but significantly reduced under low-light conditions

videos are used for training. We select four different low-light enhancement techniques for comparison. LLFlow [48] and SDSD [47] are low-light enhancement methods based on RGB images, while DNF [24] and SMOID [23] are based on RAW low-light enhancement methods. LLFlow and DNF take images as input, SDSD and SMOID take videos as input. The RAW object detection method is also trained together on well-lit and low-light videos. Its output bounding boxes are directly fed to the tracker. We test all the potential methods on two state-of-the-art trackers ByteTrack [62] and OC-SORT [3]. We also show mAP, AP50, and AP75 to highlight the detector performance.

From Tab. 5 we can see that, the proposed LTrack achieves the best HOTA and is highly competitive on all metrics. For example, the proposed method improves HOTA 1.3 with almost the same parameters and computation as base methods. This strongly proves the effectiveness of the proposed method. Compared with tracking after low-light enhancement, these methods have a large amount of additional parameters and computation but still perform worse than the proposed method. For example, LLFlow delivers 8 times FLOPS but still performs worse than our LTrack. In addition, We observed that there is no significant difference in results between RAW-based image enhancement methods and RGB-based image enhancement methods. This is not consistent with what our method has observed. The reason may be that low light enhancement focuses on image quality restoration and may mislead downstream tasks. As for RAOD [52], it has almost the same number of parameters and computational load as our method, but its performance is much lower than the proposed LTrack. Despite addressing HDR scenes through a preprocessing module for RAW input, it does not perform well under low-light conditions.

Method	HOTA	AssA	IDF1	MOTA	Det
Base-low	32.5	33.4	39.3	42.4	27.1
Base-well	28.0	30.4	34.1	36.9	23.3
Base-all	34.6	38.1	40.0	44.8	30.2
LLFlow[48]	33.6	38.0	38.4	43.1	29.7
SDSD[47]	31.6	34.2	36.7	38.8	27.5
DNF[24]	33.2	**39.0**	36.9	44.4	30.1
SMOID[23]	31.2	34.1	36.6	41.2	26.9
RAOD[52]	32.2	33.5	39.4	40.9	26.5
LTrack (Ours)	**35.1**	38.9	**40.4**	**45.2**	**30.7**

Table 6. Experimental results on LMOT-real datase with OC-SORT[3]. The best result are shown in boldface

5.4. Results on Real World

To validate the performance of our method in real-world low-light scenes at night, we evaluate all methods on LMOT-real dataset, using OC-SORT [3]. As shown in Tab. 6, the proposed LTrack performs much better than all comparison methods. Both the methods of tracking after low-light enhancement and the RAW detection method encounter generalizability problems and are not even better than Base-all. This strongly demonstrates the robustness of

DSL	ALD	HOTA	AssA	IDF1	MOTA	DetA
		28.1	42.6	32.9	23.8	18.7
✓		29.0	42.4	35.0	25.8	20.1
	✓	28.3	42.4	33.1	24.6	19.3
✓	✓	**29.4**	**43.2**	**35.2**	**26.1**	**20.3**

Table 7. Ablation on adaptive low-pass downsampling (ALD) and degradation suppression learning (DSL).

Data Type	HOTA	AssA	IDF1	MOTA	DetA
sRGB	29.2	42.7	34.9	26.4	**20.5**
RAW 8-bit	29.0	42.3	34.5	26.1	20.2
RAW 10-bit	29.3	42.9	34.8	**26.5**	20.4
RAW 12-bit	**29.4**	**43.2**	**35.2**	26.1	20.3

Table 8. Results of different image format and dynamic range on LMOT test set.

Methods	HOTA	AssA	IDF1	MOTA	DetA
Gaussian-Possion	29.9	43.1	35.7	26.6	21.1
Physical-based [50]	30.4	**44.1**	36.0	26.6	21.2
LMOT (Ours)	**35.1**	38.9	**40.4**	**45.2**	**30.7**

Table 9. Results with low-light synthesis methods and LMOT dataset on LMOT-real set.

our LTrack in real-world dark scenes.

5.5. Exploration and Discussion

In this section, we conduct extensive analysis and discussion on LMOT datasets and the proposed method.

Ablation Study. We conduct ablation experiments to validate the effectiveness of our improvements, the results are shown in Tab. 7. It can be seen that all improvements contribute effectively to enhanced performance. Among them, the degradation suppression learning strategy demonstrates the most significant effect, resulting in an approximate 1-point enhancement in HOTA. The best results are achieved when employing all strategies simultaneously. This demonstrates the effectiveness of all our contributions.

RAW *vs*. sRGB. We analyze the impact of input data formats, the results are shown in Tab. 8. The 12-bit RAW format achieves significantly better results than sRGB. Because the RAW format saves much more potential information and is helpful for MOT in low-light scenes. We also observed that higher bitwidth is beneficial for performance, which has also been observed in other vision tasks [21, 52].

LMOT *vs*. Synthetic data. We compare our LMOT dataset with the synthetic low-light data to further demonstrate the value of LMOT. We apply the Gaussian-Possian based and Physical-based [50] low-light data synthesis method to synthesize low-light videos from well-lit videos. We train our LTrack on these types of data and evaluate their performance in real low-light scenes using LMOT-real. As shown in Tab. 9, the tracker trained on our LMOT dataset has much better performance in real night scenes, which strongly demonstrates the value of our LMOT dataset.

Categories	HOTA	AssA	IDF1	MOTA	DetA
Person	24.2	30.8	29.9	25.1	19.2
Bicycle	16.9	40.3	15.6	8.9	7.1
Car	37.5	53.3	47.2	33.7	26.5
Motorcycle	19.8	35.0	22.6	15.2	11.4
Bus	44.3	59.9	51.8	36.7	32.9
Truck	6.9	15.6	6.4	3.2	3.1

Table 10. Results of the proposed method for different categories on LMOT test set.

Analysis on different category. We also analyze the performance for different categories. From Tab. 10, we can see that cars and buses achieved the best performance because they have regular shapes and larger areas. Trucks exhibited the poorest performance, because they have the fewest instances, making it challenging for the model to learn accurate identification. The person achieves relatively average scores. Bicycles and motorcycles have close scores since they have similar appearance and motion patterns.

6. Conclusion

In this work, we investigate the multi-object tracking in the dark scenes. We build a new low-light multi-object tracking (LMOT) dataset, which provides well-aligned low-light video pairs and high-quality multi-object tracking annotations. We observed that low-light images are significantly degraded by the sensor noises, which also degrades the feature maps and significantly deteriorates the model performance. To learn the invariant semantic formation under noise disturbance and quality degradation, we present the adaptive low-pass downsample module and degradation suppression learning. These improvements greatly enhance the robustness of our method in real-world low-light scenes.

Limitations. We focus on multi-object tracking in the dark scenes. But we do not consider other degradation environments in the real world, such as rainy and foggy days. In our future work, we will consider exploring multi-object tracking in more real-world scenarios, promoting the development of MOT in real-world applications.

Acknowledgments

This work was supported by the National Key R&D Program of China (2022YFC3300704), the National Natural Science Foundation of China (62331006,62171038, and 62088101), the R&D Program of Beijing Municipal Education Commission (KZ202211417048), and the Fundamental Research Funds for the Central Universities.

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, pages 3464–3468, 2016. 3

[2] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. *arXiv preprint arXiv:2303.06705*, 2023. 1, 2

[3] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *CVPR*, pages 9686–9696, 2023. 1, 3, 6, 7

[4] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, pages 3291–3300, 2018. 2

[5] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *ICCV*, pages 3185–3194, 2019. 2

[6] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *IJCV*, pages 1–21, 2023. 1, 2

[7] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, pages 215–230, 2012. 1

[8] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. *arXiv preprint arXiv:2304.05170*, 2023. 1, 2, 3, 5, 6

[9] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 1, 2, 3

[10] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 1

[11] Anna Ellis and James Ferryman. Pets2010 and pets2009 evaluation of results using individual ground truthed single views. In *2010 7th IEEE international conference on advanced video and signal based surveillance*, pages 135–142, 2010. 1

[12] Wang Zhongdao et al. Do different tracking tasks require different appearance models? 2021. 4

[13] Huiyuan Fu, Wenkai Zheng, Xicong Wang, Jiaxuan Wang, Heng Zhang, and Huadong Ma. Dancing in the dark: A benchmark towards general low-light video enhancement. In *ICCV*, pages 12877–12886, 2023. 2

[14] Ying Fu, Jian Chen, Tao Zhang, and Yonggang Lin. Residual scale attention network for arbitrary scale image super-resolution. *Neurocomputing*, 427:201–211, 2021. 2

[15] Ying Fu, Yang Hong, Linwei Chen, and Shaodi You. Le-gan: Unsupervised low-light image enhancement network using attention module and identity invariant loss. *Knowledge-Based Systems*, 240:108010, 2022. 2

[16] Ruopeng Gao and Limin Wang. Memotr: Long-term memory-augmented transformer for multi-object tracking. In *ICCV*, pages 9901–9910, 2023. 3

[17] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 6

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 1, 2, 3, 4

[19] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, pages 1780–1789, 2020. 2

[20] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 26(2):982–993, 2016. 2

[21] Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *BMVC*, page 3, 2021. 1, 2, 8

[22] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007. 2

[23] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *ICCV*, pages 7324–7333, 2019. 1, 2, 3, 7

[24] Xin Jin, Ling-Hao Han, Zhen Li, Chun-Le Guo, Zhi Chai, and Chongyi Li. Dnf: Decouple and feedback network for seeing in the dark. In *CVPR*, pages 18135–18144, 2023. 1, 2, 7

[25] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. Properties and performance of a center/surround retinex. *IEEE TIP*, 6(3):451–462, 1997. 2, 4

[26] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 2

[27] Sohyun Lee, Jaesung Rim, Boseung Jeong, Geonu Kim, Byungju Woo, Haechan Lee, Sunghyun Cho, and Suha Kwak. Human pose estimation in extremely low-light conditions. In *CVPR*, pages 704–714, 2023. 1, 2

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6

[29] Qiankun Liu, Qi Chu, Bin Liu, and Nenghai Yu. Gsm: Graph similarity model for multi-object tracking. In *IJCAI*, pages 530–536, 2020. 1

[30] Qiankun Liu, Dongdong Chen, Qi Chu, Lu Yuan, Bin Liu, Lei Zhang, and Nenghai Yu. Online multi-object tracking with unsupervised re-identification learning and occlusion estimation. *Neurocomputing*, 483:333–347, 2022. 1

[31] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *IJCV*, 129:548–578, 2021. 5

[32] Kang Ma, Ying Fu, Dezhi Zheng, Chunshui Cao, Xuecai Hu, and Yongzhen Huang. Dynamic aggregated network for gait recognition. In *CVPR*, pages 22076–22085, 2023. 1

[33] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, pages 8844–8854, 2022. 3

[34] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 2, 3, 4

[35] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, pages 1765–1773, 2017. 2

[36] Jihun Park, Jinseok Hong, Wooil Shim, and Dae-Jin Jung. Multi-object tracking on swir images for city surveillance in an edge-computing environment. *Sensors*, 23(14):6373, 2023. 3

[37] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35, 2016. 6

[38] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 5

[39] Aashish Sharma and Robby T Tan. Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects. In *CVPR*, pages 11977–11986, 2021. 2

[40] Shuran Song and Jianxiong Xiao. Tracking revisited using rgbd camera: Unified benchmark and baselines. In *ICCV*, pages 233–240, 2013. 3

[41] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 3

[42] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, pages 20993–21002, 2022. 1, 2, 3, 4, 5

[43] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[44] Ye Tian, Ying Fu, and Jun Zhang. Plug-and-play algorithm for under-sampling fourier single-pixel imaging. *Science China. Information Sciences*, 65(10):209303, 2022. 2

[45] Ye Tian, Ying Fu, and Jun Zhang. Local-enhanced transformer for single-pixel imaging. *Optics Letters*, 48(10): 2635–2638, 2023. 2

[46] Chaoqun Wang, Chunyan Xu, Zhen Cui, Ling Zhou, Tong Zhang, Xiaoya Zhang, and Jian Yang. Cross-modal pattern-propagation for rgb-t tracking. In *CVPR*, pages 7064–7073, 2020. 3

[47] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *ICCV*, pages 9700–9709, 2021. 1, 2, 7

[48] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *AAAI*, pages 2604–2612, 2022. 7

[49] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 2

[50] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *CVPR*, pages 2758–2767, 2020. 4, 6, 8

[51] Kaixuan Wei, Angelica Aviles-Rivero, Jingwei Liang, Ying Fu, Hua Huang, and Carola-Bibiane Schönlieb. Tfpnp: Tuning-free plug-and-play proximal algorithms with applications to inverse imaging problems. *Journal of Machine Learning Research*, 23(16):1–48, 2022. 2

[52] Ruikang Xu, Chang Chen, Jingyang Peng, Cheng Li, Yibin Huang, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Toward raw object detection: A new benchmark and a new model. In *CVPR*, pages 13384–13393, 2023. 2, 7, 8

[53] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *CVPR*, pages 17714–17724, 2022. 2

[54] Jinyu Yang, Shang Gao, Zhe Li, Feng Zheng, and Aleš Leonardis. Resource-efficient rgbd aerial tracking. In *CVPR*, pages 13374–13383, 2023. 3

[55] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2636–2645, 2020. 2, 3, 4

[56] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *ECCV*, pages 659–675, 2022. 3

[57] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *CVPR*, pages 4967–4976, 2021. 2

[58] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *ICCV*, pages 13043–13052, 2021. 3

[59] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *CVPR*, pages 8886–8895, 2022. 3

[60] Tao Zhang, Ying Fu, and Cheng Li. Hyperspectral image denoising with realistic data. In *ICCV*, pages 2248–2257, 2021. 2

[61] Tao Zhang, Ying Fu, and Cheng Li. Deep spatial adaptive network for real image demosaicing. In *AAAI*, pages 3326–3334, 2022. 2

[62] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21, 2022. 1, 3, 6, 7

[63] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *CVPR*, pages 22056–22065, 2023. 3

[64] Zhiyu Zhu, Junhui Hou, and Dapeng Oliver Wu. Cross-modal orthogonal high-rank augmentation for rgb-event

transformer-trackers. In *ICCV*, pages 22045–22055, 2023. 3

[65] Yunhao Zou and Ying Fu. Estimating fine-grained noise model via contrastive learning. In *CVPR*, pages 12682–12691, 2022. 2

[66] Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to reconstruct high speed and high dynamic range videos from events. In *CVPR*, pages 2024–2033, 2021. 2

[67] Yunhao Zou, Chenggang Yan, and Ying Fu. Rawhdr: High dynamic range image reconstruction from a single raw image. In *ICCV*, pages 12334–12344, 2023. 2