A New Multi-Source Light Detection Benchmark and Semi-Supervised Focal Light Detection: Supplementary material

In this supplementary material, we have provided more information of our YouTube Driving Light Detection (YDLD) dataset and detailed experimental results and the discussion of our method.

- In Sec. A, we have provided more details of YDLD dataset.
- In Sec. B, we have provided the detailed explanation of the process for aligning between pseudo-labels and predictions.
- In Sec. C, we have provided more details of our SS-FLD.
- In Sec. D, we have provided more details of the overlapping classes and experimental results between YDLD, SODA10M, KITTI, ImageNet, and MS-COCO with BDD100k.
- In Sec. E, we have provided a class-wise comparison with the state-of-the-art (SOTA) object detectors.
- In Sec. F, we have provided expanded experimental results.
- In Sec. G, we have provided the specific effects of SS-FLD.
- In Sec. H, we have provided the qualitative comparisons of the detection results.
- In Sec. I, we have provided the analysis of failure cases through confusion matrix and qualitative results.
- In Sec. J, we have provided the sensitivity analysis of attention prior term.
- In Sec. K, we have provided the discussion for the spatial attention prior.
- In Sec. L, we have provided the datasheet for our YDLD dataset.
- In addition, we have provided the supplementary video for our experimental result. The qualitative detection results of our SS-FLD method in the real-driving environment are contained.

A More details of YDLD dataset

In this section, we provide more details of our YDLD dataset. As we mentioned in our manuscript, our dataset consists of three classes: Carlight, TrafficSignallight, and Streetlight. Especially, the Carlight class contains headlights, rear lamps, and brake lamps. Table 1 shows the class-wise statistics. Our YDLD dataset has light sources of various sizes to reflect real-world traffic environment. Also, we illustrate the annotated samples of each class in Fig. 1. As shown in this figure, car lights are many variety of light sources depending on the vehicle design. Otherwise, TrafficSignallights are usually small. These light sources play an important role in driver's decision-making process. However, the most bulbs are circular with high pixel values. In the first row, Carlight and TrafficSignallight are very similar. Additionally, Streetlights have a high pixel intensity to illuminate the road, so they tend to produce a large glare. Because of these difficulties, Figure .12 demonstrates the failure cases of ConsistentTeacher [1]. We can show that the objects in bounding boxes are very similar to distinguish their classes.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

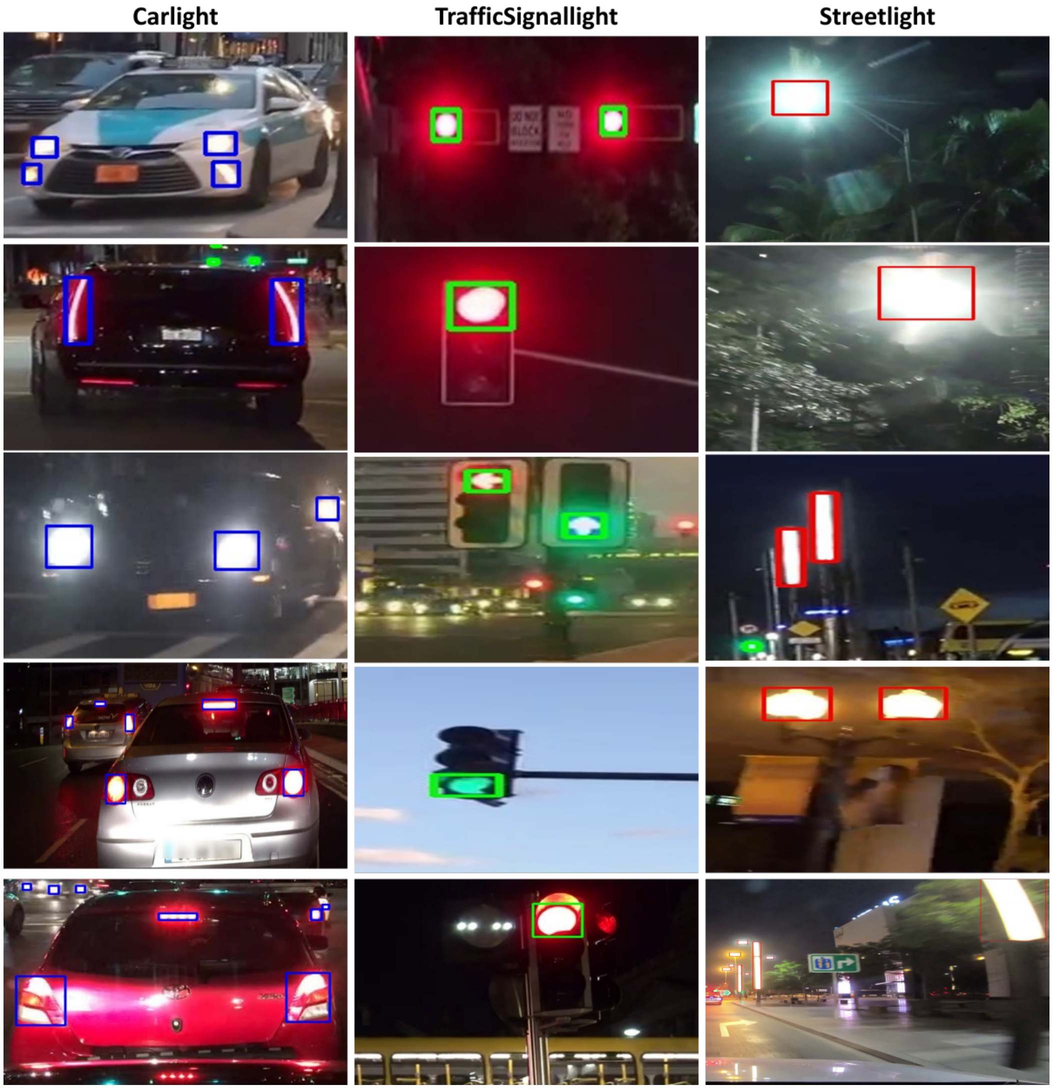


Figure 1: The samples of each class from our YDLD dataset. Boxes for each light source are represented with different colors.

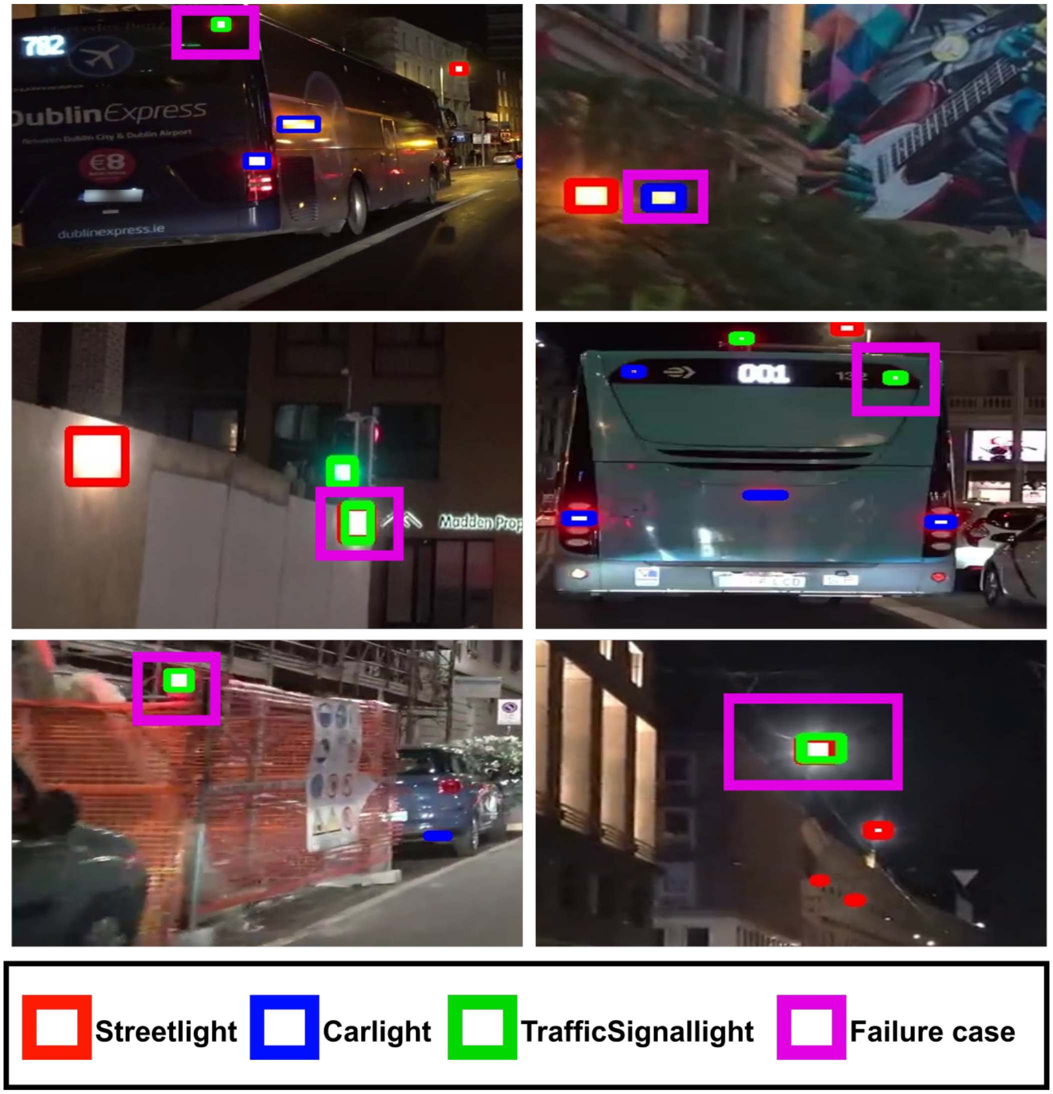


Figure 2: The failure cases of ConsistentTeacher. Red, blue, and green boxes represent Streetlight, Carlight, TrafficSignalLight class, respectively. We highlight the failure case using magenta boxes.

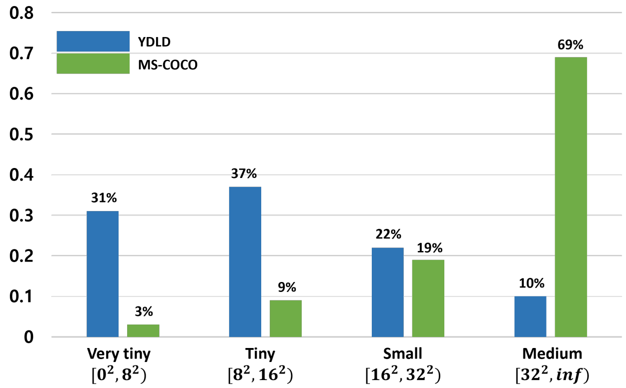


Figure 3: The comparison for the distribution of bounding box sizes between YDLD and MS-COCO.

Table 1: The class-wise statistics of YDLD.

	Carlight	**TrafficSignallight**	**Streetlight**
# GTs	49,059	14,373	52,596
Avg. of # GTs	13.95	4.09	14.96
Highest # GTs on an image	127	23	472
Avg. of bbox size	22.02×14.50	14.35×14.95	18.10×14.65
Largest bbox size	395×204	141×152	362×650
Very tiny	16,638	15,182	4,685
Tiny	19,952	17,027	5,420
Small	10,886	10,957	3,122
Medium	5,120	5,893	1,146

Table 2: The comparison of the statistics between YDLD and MS-COCO17.

	YDLD	**MS-COCO**
# images	3,156	123,287
# GTs	116,028	896,782
# Classes	3	80
Avg. of # GTs	33.00	7.27
Highest # GTs on an image	501	93
Avg. of bbox size	19.29×14.62	103.86×107.41
Largest bbox size	362×650	640×640
Smallest bbox size	1×1	0.23×3.64

Due to these reasons, the recent object detectors do not achieve satisfactory results on YDLD. In addition, we show the object statistics difference from MS-COCO17 [2]. Figure 3, Table 2, and Figure 4 demonstrate the different distribution between YDLD and MS-COCO. While MS-COCO medium-sized objects account for most, YDLD has many small and tiny objects. 68% of the light source objects in our YDLD fall within the very tiny ($[0^2, 8^2)$) and the tiny ($[8^2, 16^2)$) size categories, whereas only 12% of the objects in the MS-COCO dataset belong to these size ranges. As a result, the distribution of the object sizes is relatively even. As mentioned in our manuscript, these small objects easily degrade detection accuracy. Figure 5 shows a bunch of tiny object samples with specific size of bounding boxes.

Furthermore, the object density of YDLD is an additional challenge. As shown in Table 2, our dataset contains an average of 33 objects per image, while in the case of MS-COCO, there are 7 objects on average. This clearly shows that our YDLD is denser than a typical object detection dataset and has difficulty predicting many objects. Especially, Figure 6 illustrates that the objects are densely concentrated within a region in the image. Moreover, light sources tend to overlap each other due to glare, especially when concentrated. It poses challenges in accurately predicting their boundaries.

Additionally, YDLD have fewer distinctive visual features compared to typical detection objects. This is due to the similar appearance between light sources as shown in 1. Because of these problems, the overall mAP score is much lower about two times (*e.g.* DINO mAP: 25.6 % (YDLD), 50.1% (COCO)).

B Aligning pseudo-labels and predictions for accurate loss calculation

As we described in Sec. 4.3 and Fig. 3 in our manuscript, the teacher and the student networks exploit weak and strong augmentations for the unlabeled data, respectively. Since the augmentation techniques used for these networks differ, it is necessary to align the pseudo-labels from the teacher with prediction results of the student to calculate the loss accurately. To this end, we use the transformation matrices applied during each augmentation to correct the differences in bounding boxes. Specifically, we calculate the overall geometrical transformation by taking the inverse of the teacher's matrix and multiplying it by the student's matrix as follows:

$$T_{align} = T_s \cdot T_t^{-1} \tag{1}$$

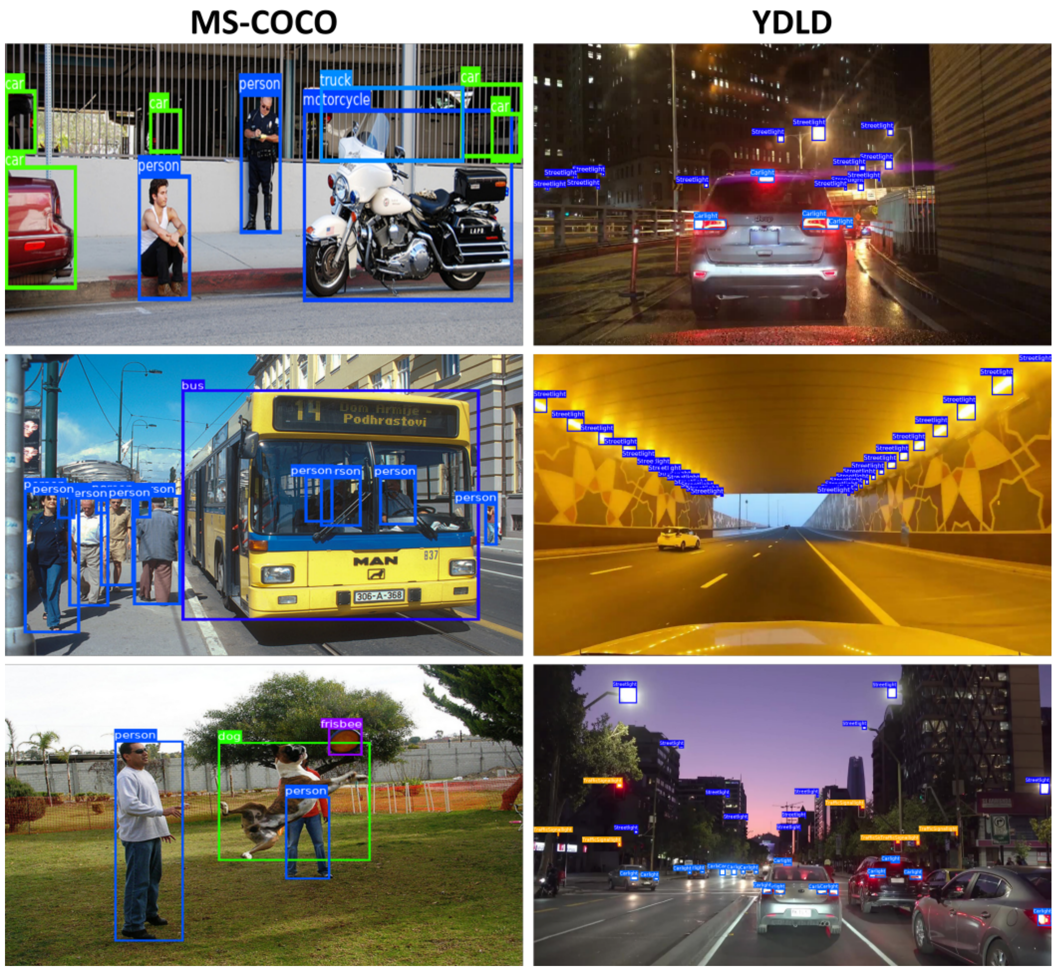


Figure 4: Bounding box comparison between MS-COCO and our YDLD.

where T_s and T_t are transformation matrices of the teacher and the student networks, respectively. As a result, this transformation matrix T_{align} allows us to map the bounding boxes of the teacher network to the transformed image coordinate of the student network. Once they are aligned, we compute the loss by comparing the predictions from the student with the transformed pseudo-labels.

C More details of SS-FLD

In this section, we provide the total loss functions and the overall architectures of our SS-FLD based on DINO and RFLA. To further explain our SS-FLD total detection loss defined in Eq. (3) of our manuscript, we first describe each total loss used in DINO and RFLA. Here, we omit the index l (c.f. u) denoting labeled data for simplicity unless other mentioned.

The loss of DINO L_{DINO} is defined as $L_{DINO}(\mathbf{x}, \mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}[\lambda_{cls} \cdot L_{cls}(f(\mathbf{x}_i), \mathbf{y}_i) + \lambda_{L1} \cdot L_{L1}(f(\mathbf{x}_i), \mathbf{y}_i) + \lambda_{GIoU} \cdot L_{GIoU}(f(\mathbf{x}_i), \mathbf{y}_i)]$. $\mathbf{x}_i$ and $\mathbf{y}_i$ are the input image and class label of i-th sample. $f(\cdot)$ is the object detector network. λ_{cls}, λ_{L1}, and λ_{GIoU} are the coefficient of each loss. L_{DINO} exploits the focal loss [3] as the classification loss L_{cls}. L1 loss L_{L1} and GIoU (generalized Intersection over Union loss) loss [4] L_{GIoU} are used for the regression losses.

The detection loss of RFLA is defined as $L_{RFLA} = \frac{1}{N}\sum_{i=1}^{N}[L_{cls}(f(\mathbf{x}_i), \mathbf{y}_i) + L_{DIoU}(f(\mathbf{x}_i), \mathbf{y}_i)]$. It uses the focal loss and DIoU (Distance Intersection over Union) loss [5] L_{DIoU} for classification and regression, respectively.

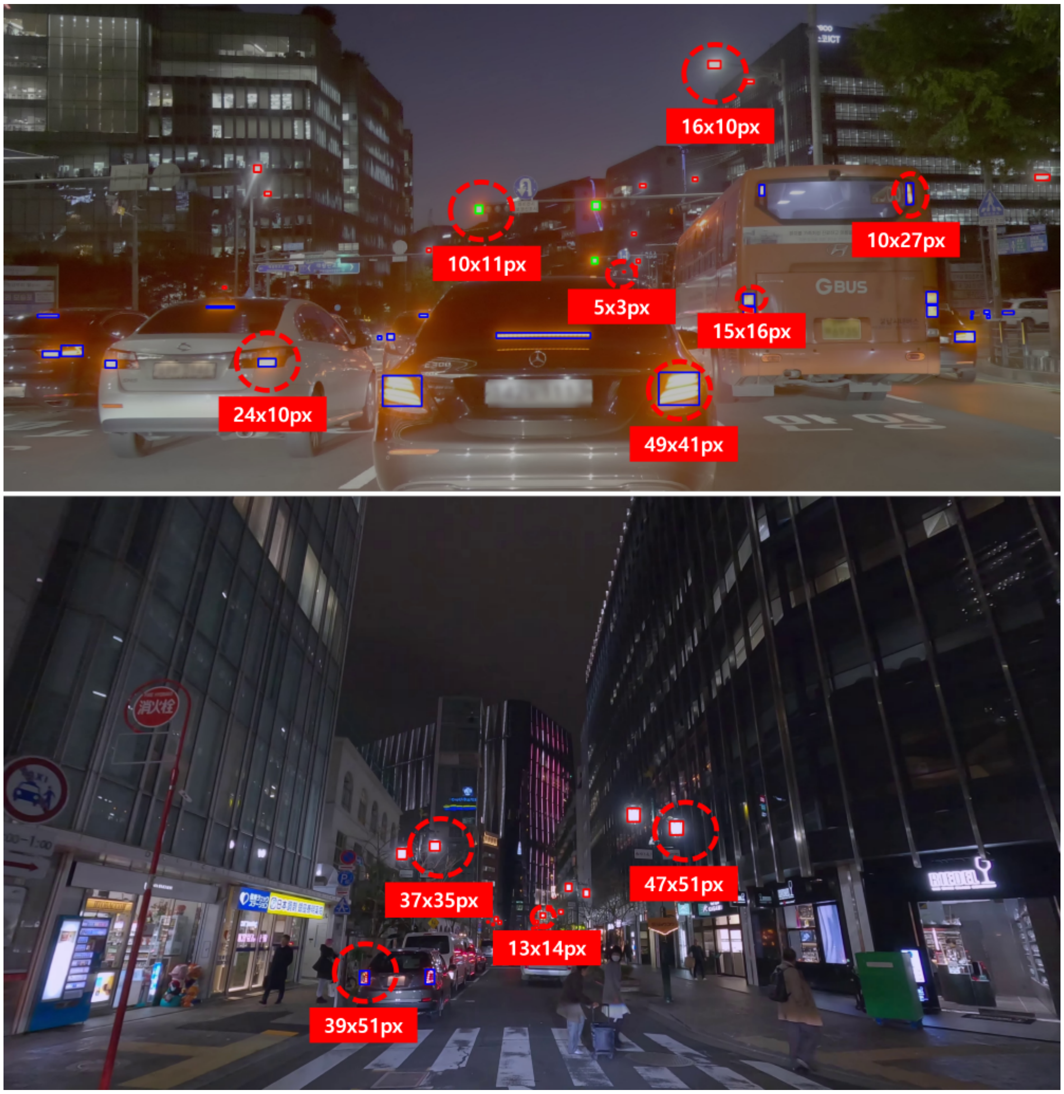


Figure 5: The object sizes of our YDLD dataset. Note that our light source objects have small sizes.

To apply SS-FLD for each detector, we replace L_{cls} with the lightness focal loss defined in Eq. (2) of our manuscript. In addition, we utilize these losses on labeled and unlabeled data to perform semi-supervised learning suited for light detection. The complete loss function of semi-supervision is in Eq. (3) of our manuscript. L_{reg} loss for the DINO and RFLA can be replaced with $L_{L1} + L_{GIoU}$ and L_{DIoU}. As a result, the overall frameworks of our SS-FLDs based on DINO and RFLA are illustrated in Fig. 7

D More details of overlapping object classes and experimental results

In this section, we provide more detailed overlapping classes between YDLD, KITTI [6], SODA10M [7], ImageNet [8], and MS-COCO [2] with the 10 classes of BDD100k [9]. MS-COCO has 9 overlapping classes as shown in Table 3. Therefore, 9 classes of BDD100k are matched except the rider class. On the other hand, our YDLD has one common class over the traffic light class. Moreover, we annotate the bulb area of the traffic light precisely, but MS-COCO and BDD100k annotate the entire area of the traffic light as described in Fig. 8. Despite of this annotation and overlapping class number differences, our YDLD have shown the remarkable effectiveness for transfer learning in

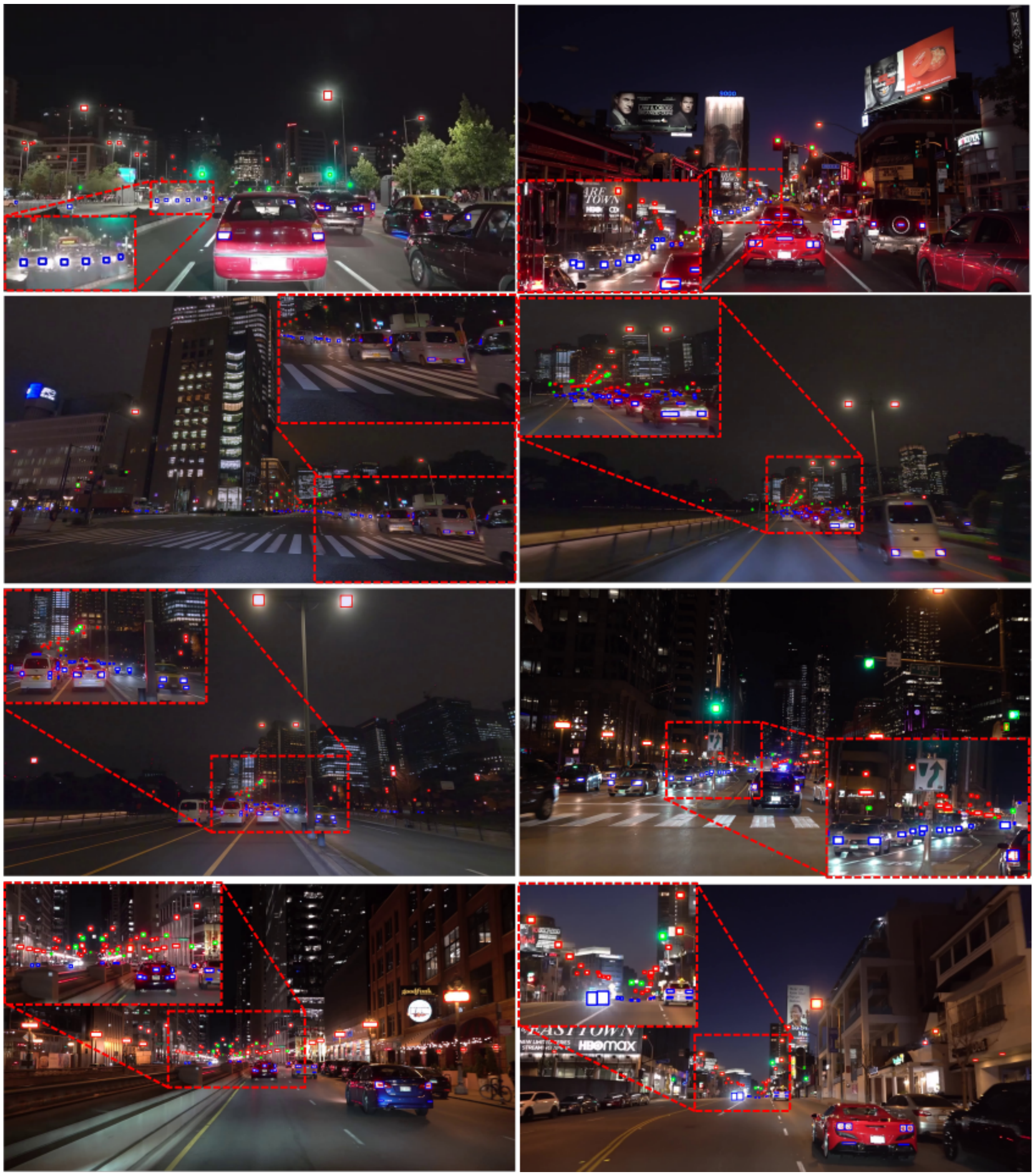


Figure 6: Examples of our YDLD dataset. Red, blue, and green boxes represent Streetlight, Carlight, and TrafficSignallight classes, respectively. They show light source objects are very small and located densely.

Table 3: The overlapped classes of YDLD, ImageNet, MS-COCO, KITTI, and SODA10M with BDD100k.

BDD100k	**Pedestrian**	**Rider**	**Car**	**Truck**	**Bus**	**Train**	**Motorcycle**	**Bicycle**	**Traffic light**	**Traffic sign**
ImageNet			✓	✓	✓	✓	✓		✓	
COCO	✓		✓	✓	✓	✓	✓	✓	✓	✓
KITTI	✓	✓	✓	✓		✓				
SODA10M	✓	✓	✓	✓		✓				
YDLD									✓	

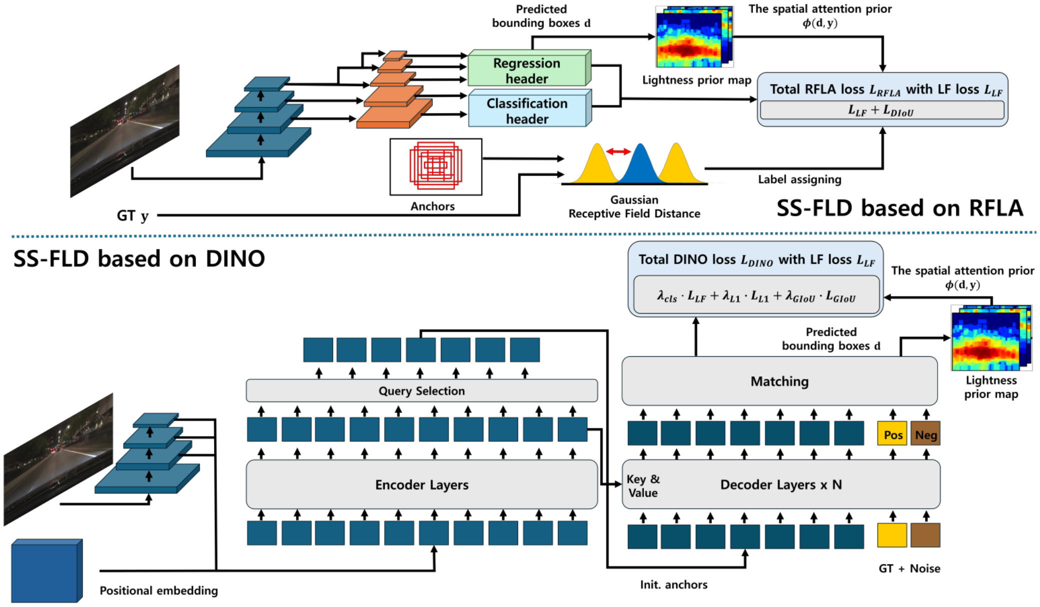

Predicted bounding boxes d
The spatial attention prior $\phi(\mathbf{d},\mathbf{y})$
Regression header
Classification header
Lightness prior map
Total RFLA loss L_{RFLA} with LF loss L_{LF}
$L_{LF} + L_{DIoU}$
Anchors
Gaussian Receptive Field Distance
Label assigning
GT y
SS-FLD based on RFLA
SS-FLD based on DINO
Total DINO loss L_{DINO} with LF loss L_{LF}
$\lambda_{cls} \cdot L_{LF} + \lambda_{L1} \cdot L_{L1} + \lambda_{GIoU} \cdot L_{GIoU}$
The spatial attention prior $\phi(\mathbf{d},\mathbf{y})$
Predicted bounding boxes d
Query Selection
Matching
Lightness prior map
Pos
Neg
Encoder Layers
Key & Value
Decoder Layers x N
Positional embedding
Init. anchors
GT + Noise

Figure 7: The overall architectures of our SS-FLDs based on RFLA (the top row) and DINO (the bottom row). For simply describing the role of our LF loss, we omit the teacher models, denoting labeled, and unlabeled data.

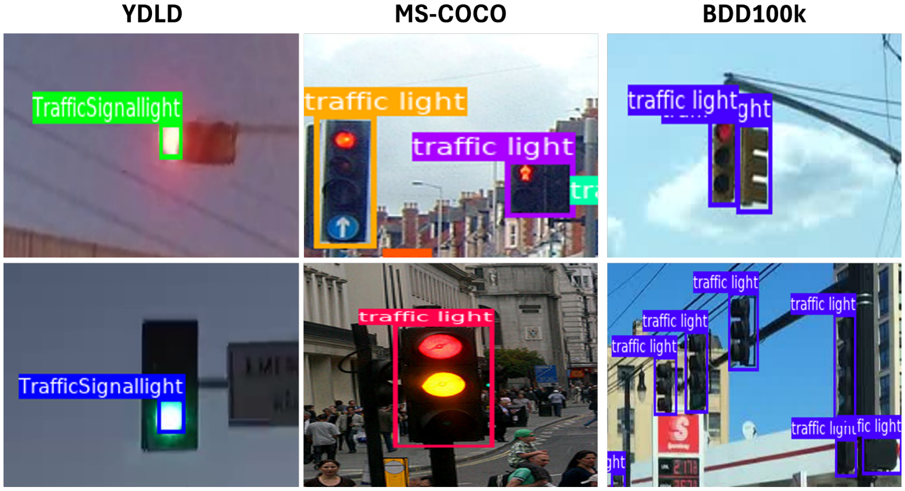


Figure 8: Comparison of the traffic light annotation areas between YDLD, MS-COCO, and BDD100k datasets.

Table 4 and Sec. 5.1 of our manuscript. Additionally, to show the importance of light source, we conduct the additional comparison that performs light detection on BDD100k. To this end, we have generated the subset of BDD100k that retains only the traffic light class from annotations since it is only the light source in BDD100k. Table 4 shows the experimental results. The YDLD and MS-COCO achieve the highest mAP scores since only these datasets include the traffic light class. In contrast, KITTI and SODA10M show marginal differences compared to ImageNet as the canonical dataset.

Table 4: Transfer learning effects of light detection on BDD100k.

Dataset	mAP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
ImageNet	27.3	73.5	13.3	24.8	42.3	37.9
COCO	28.0	74.4	14.5	25.8	43.2	49.3
KITTI	27.1	73.2	13.7	24.8	42.7	46.2
SODA10M	27.1	73.4	13.5	24.8	42.2	47.0
YDLD	28.2	74.6	14.5	25.6	44.1	45.3

E Class-wise comparison with SOTA methods

Table 5: Class-wise benchmark results with recent detectors on YDLD dataset. We highlight the best record with **red bold**. † represents our implementation.

Detectors	**Publications**	**Year**	mAP	**Streetlight**	**Carlight**	**TrafficSignallight**
Two-stage object detectors						
Libra R-CNN [10]	CVPR	2019	8.6	7.8	8.6	9.3
Faster R-CNN [11]	NeurIPS	2015	8.9	8.0	9.3	9.4
Cascade R-CNN [12]	CVPR	2018	10.2	9.0	10.7	10.7
Single-stage object detectors						
YOLOF [13]	CVPR	2021	5.7	6.1	5.6	5.5
RetinaNet [3]	ICCV	2017	8.1	7.3	5.2	11.8
ATSS [14]	CVPR	2020	15.6	15.5	13.3	18.1
YOLOX [15]	CVPR	2021	17.9	17.8	15.0	20.9
PAA [16]	ECCV	2020	21.6	22.8	17.0	25.0
Transformer-based object detectors						
Deformable DETR [17]	ICLR	2021	16.7	16.0	14.1	19.8
DINO [18]	ICLR	2023	22.6	23.8	18.3	25.7
Semi-supervised object detectors						
SoftTeacher [19]	ICCV	2021	8.4	6.9	9.3	8.8
MeanTeacher [20, 1]	NeurIPS	2017	14.4	12.8	9.5	20.9
ConsistentTeacher [1]	CVPR	2023	19.1	18.7	16.5	22.7
Tiny object detectors						
CEASC [21]	ICCV	2023	7.5	6.7	4.1	11.7
FASNet [22]	IEEE TGRS	2022	10.4	10.8	7.9	12.4
RFLA w/t RetinaNet [23]	ECCV	2022	10.8	11.4	4.7	16.4
NWD-RKA [24]	ISPRS P&RS	2022	16.6	16.8	12.0	21.0
RFLA w/t PAA †	ECCV	2022	21.6	22.9	16.2	25.8
SS-FLD w/t DINO	Proposed		25.6	26.0	**22.1**	28.7
SS-FLD w/t RFLA	Proposed		**26.0**	**26.3**	21.5	**30.3**

We have provided a more detailed benchmark results with SOTA object detectors. To show more details, we compare the class-wise mAP results of each detector. The result is shown in Table 5. The experiment settings are the same as Sec. 5.2 in the main manuscript. Our SS-FLDs achieve the best mAP scores for all three classes. Especially, compared with DINO, our method achieves 2.2%, 3.8%, and 3.0% higher mAP scores in each class. Compared with, PAA w/t RFLA, our method shows a higher 3.4%, 5.3%, and 4.5% improvement for Streetlight, Carlight, and TrafficSignallight classes, respectively.

Especially, the Carlight class shows a lower mAP rather than other classes. We visualize the miss-classified results of the Carlight class (magenta circle) in Fig. 9. We can observe that there are many

Table 6: YDLD benchmark evaluation with additional AP_{50}, AP_{75}, and AP_m scores. We highlight the best record with **red bold**. † represents our implementation.

Detectors	**Publications**	**Year**	mAP	AP_{50}	AP_{75}	AP_{vt}	AP_t	AP_s	AP_m
Two-stage object detectors									
Libra R-CNN [10]	CVPR	2019	8.6	16.6	8.2	0.2	6.1	10.2	40.3
Faster R-CNN [11]	NeurIPS	2015	8.9	17.1	8.3	0.1	6.2	10.5	41.5
Cascade R-CNN [12]	CVPR	2018	10.2	18.7	10.1	0.3	7.3	12.5	45.8
Single-stage object detectors									
YOLOF [13]	CVPR	2021	5.7	13.3	4.3	0.0	3.7	7.2	27.4
RetinaNet [3]	ICCV	2017	8.1	20.2	5.2	2.2	4.8	20.6	37.2
ATSS [14]	CVPR	2020	15.6	33.7	12.5	4.7	11.7	25.3	42.7
YOLOX [15]	CVPR	2021	17.9	42.3	12.1	8.1	14.4	27.5	39.1
PAA [16]	ECCV	2020	21.6	48.0	16.1	9.7	18.0	34.6	43.6
Transformer-based object detectors									
Deformable DETR [17]	ICLR	2021	16.7	41.9	9.7	7.3	14.0	26.4	34.4
DINO [18]	ICLR	2023	22.6	51.6	16.3	10.5	19.0	35.2	47.2
Semi-supervised object detectors									
SoftTeacher [19]	ICCV	2021	8.4	15.9	7.9	0.3	6.4	9.3	39.4
MeanTeacher [20, 1]	NeurIPS	2017	14.4	32.5	10.7	9.0	16.0	25.2	16.2
ConsistentTeacher [1]	CVPR	2023	19.1	42.6	14.7	7.4	15.0	31.7	42.7
Tiny object detectors									
CEASC [21]	CVPR	2023	7.5	19.3	4.4	2.3	4.7	18.5	33.1
FSANet [22]	IEEE TGRS	2022	10.4	24.9	7.1	2.3	7.7	16.1	34.1
RFLA w/t RetinaNet [23]	ECCV	2022	10.8	30.4	5.1	5.9	9.8	16.9	21.4
NWD-RKA [24]	ISPRS P&RS	2022	16.6	40.4	11.0	6.6	12.7	29.7	37.7
RFLA w/t PAA †	ECCV	2022	21.6	50.8	14.7	11.0	19.4	33.4	40.2
SS-FLD w/t DINO	Proposed		25.6	57.6	**19.1**	12.7	23.1	38.6	**48.0**
SS-FLD w/t RFLD	Proposed		**26.0**	**58.3**	19.0	**12.8**	**24.8**	**39.2**	43.4

Table 7: Effects of our semi-supervised focal loss (SS-FLD) detection on the YDLD.

Baseline	**SS-FLD**	mAP	AP_{50}	AP_{75}	AP_{vt}	AP_t	AP_s	AP_m
RFLA w/t RetinaNet		10.8	30.4	5.1	5.9	9.8	16.9	21.4
RFLA w/t RetinaNet	✓	18.6 [↑ 7.8]	50.0 [↑ 19.6]	8.7 [↑ 3.6]	8.5 [↑ 2.6]	17.0 [↑ 7.2]	29.4 [↑ 12.5]	32.7 [↑ 11.3]
RFLA w/t PAA		21.6	50.8	14.7	11.0	19.4	33.4	40.2
RFLA w/t PAA	✓	26.0 [↑ 4.4]	58.3 [↑ 7.5]	19.0 [↑ 4.3]	12.8 [↑ 1.8]	24.8 [↑ 5.4]	39.2 [↑ 5.8]	43.4 [↑ 3.2]
ATSS		15.6	33.7	12.5	4.7	11.7	25.3	42.7
ATSS	✓	17.7 [↑ 2.1]	38.2 [↑ 4.5]	14.3 [↑ 1.8]	6.0 [↑ 1.3]	13.3 [↑ 1.6]	29.2 [↑ 3.9]	43.7 [↑ 1.0]
PAA		21.6	48.0	16.1	9.7	18.0	34.6	43.6
PAA	✓	25.6 [↑ 4.0]	55.1 [↑ 7.1]	20.4 [↑ 4.3]	11.9 [↑ 2.2]	22.4 [↑ 4.4]	40.1 [↑ 5.5]	46.9 [↑ 3.3]
Deformable DETR		16.7	41.9	9.7	7.3	14.0	26.4	34.4
Deformable DETR	✓	18.0 [↑ 1.3]	43.3 [↑ 1.4]	11.6 [↑ 1.9]	8.7 [↑ 1.4]	15.8 [↑ 1.8]	28.4 [↑ 2.0]	35.2 [↑ 0.8]
DINO		22.6	51.6	16.3	10.5	19.0	35.2	47.2
DINO	✓	25.6 [↑ 3.0]	57.6 [↑ 6.0]	19.1 [↑ 2.8]	12.7 [↑ 2.2]	23.1 [↑ 4.1]	38.6 [↑ 3.4]	48.0 [↑ 0.8]

overlapped predicted results on a single object. However, our SS-FLD achieves the highest mAP score 22.1% in Carlight class.

F Detailed experimental results

In this section, we have provided more experimental results of Table 5 and 6 in our manuscript with AP_{50}, AP_{75}, and AP_m metrics used for the conventional object detection. Table 6 demonstrates the expanded benchmark evaluation. Conventional detectors show high AP_m scores similar to canonical object detection datasets (*e.g.* MS-COCO [2] and PASCAL VOC [25]). However, the very tiny and tiny objects occupy 68% of our YDLD dataset as shown in Fig. 3. Therefore, they show low mAP scores. In contrast, our SS-FLDs achieve the highest scores in all metrics. Compared with DINO, our SS-FLD improves 6.0, 2.8, and 0.8 points in AP_{50}, AP_{75}, and AP_m, respectively. Compared with RFLA w/t PAA, the SS-FLD achieves 7.5, 4.3, and 3.2 points improvements in each metric.

Table 7 shows the expanded effects of our SS-FLD with additional AP_{50}, AP_{75}, and AP_m scores. In Table 6 and 7, tiny object detectors show low performance on AP_s and AP_m scores. However, our SS-FLD significantly increases AP_s and AP_m scores about 12.5 and 11.3 points for RFLA w/t

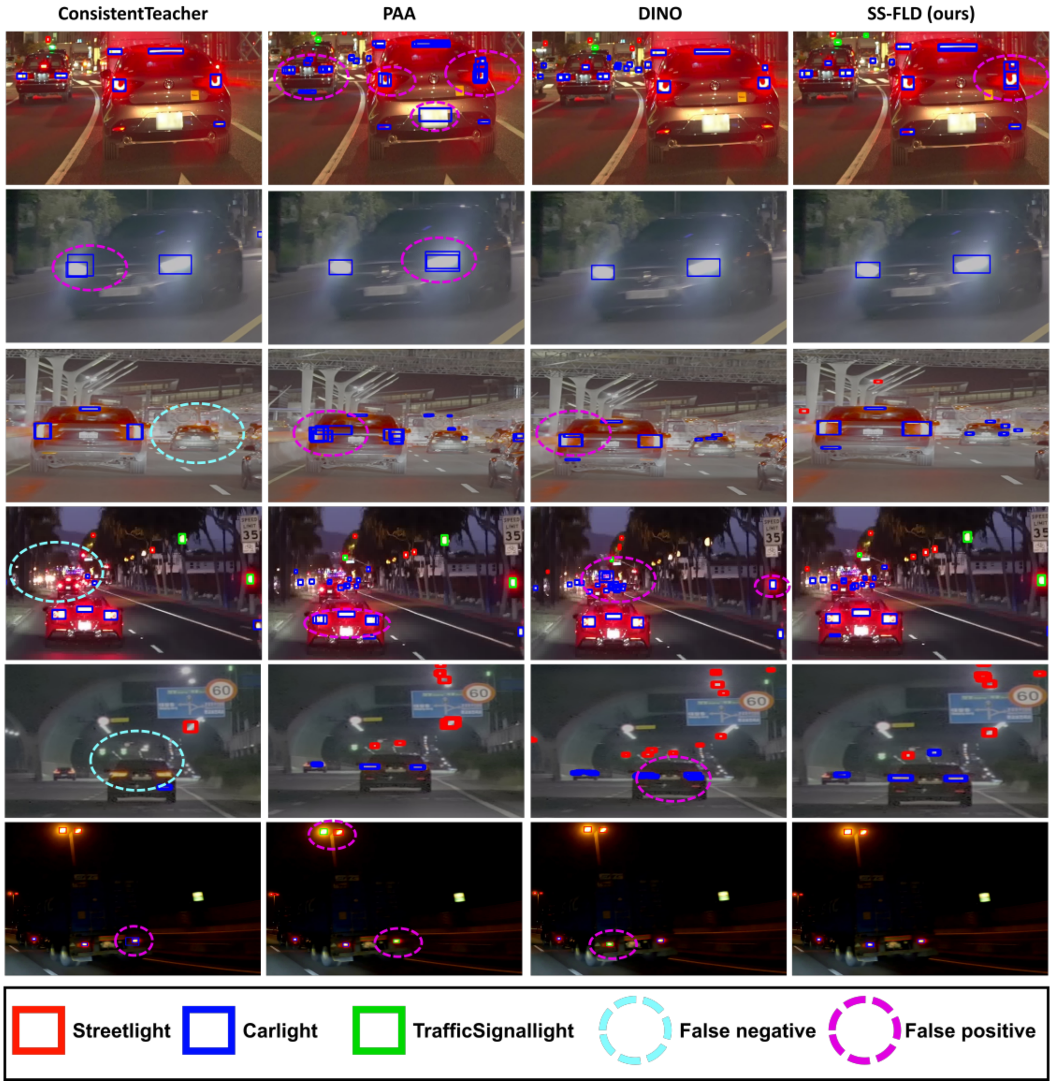


Figure 9: Detection results of Carlight class using our SS-FLD, ConsistentTeacher, PAA, and DINO are compared.

RetinaNet and 5.8 and 3.2 points for RFLA w/t PAA. Compared with other baselines, our SS-FLD consistently improves performance in all metrics.

G Detailed effects of SS-FLD

To describe effects of SS-FLD more specifically, we compare attention maps of RFLA w/t and w/o our SS-FLD. In Fig. 10, RFLA w/o SS-FLD mainly focuses on the small regions centered at light sources similar to traditional light detections based on hand-craft features. As a result, most detectors tend to produce excessive false positives for similar light sources. On the other hand, our SS-FLD attends not only light sources, but also related structures such as the pole, frames of the traffic light, and other parts of the car to encode rich contexts. These results show that our SS-FLD can handle the intrinsic and practical challenges of light detection.

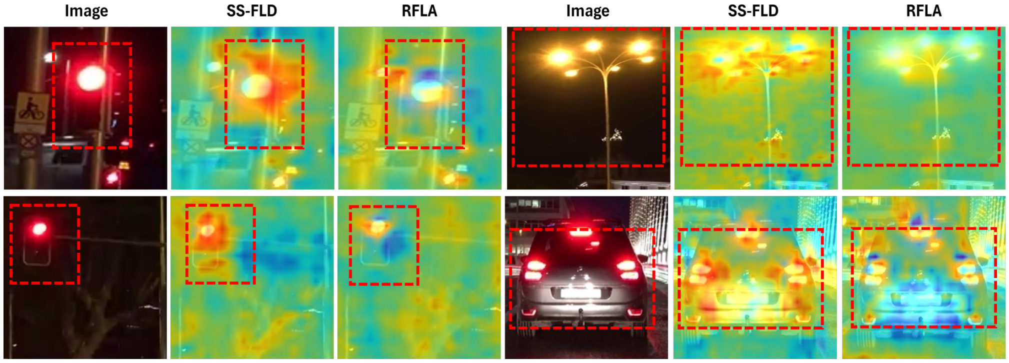


Figure 10: We visualize attention maps of RFLA w/t and w/o SS-FLD using Grad-CAM [26] (red indicates higher, but blue lower scores). We highlight a region with distinct attention differences with red dotted boxes.

H More qualitative comparison

We provide a more qualitative comparison between our proposed SS-FLD, ConsistentTeacher, PAA, and DINO in Fig. 11. In this figure, red, blue, and green boxes represent Streetlight, Carlight and TrafficSignallight classes, respectively. We highlight false negatives and false positives with cyan and magenta circles, respectively. Also, we emphasize the reflected lights and glare regions using yellow and pink boxes, respectively. In this comparison, we show better detection results compared with other detectors. Especially, our detector successfully avoids detecting reflected lights and glare regions.

I Analysis of failure cases through confusion matrix and qualitative results

In this section, we provide more failure cases and the confusion matrices to analysis the performance degradation on our YDLD in depth. In Fig. 12, false negatives usually happen for very tiny and twisted lights. To remedy this, increase the generality of our YDLD to contain these hard samples. On the other hand, false positives often occur in areas with strong glare or when the light source is tiny.

To understand false positive and negative cases in depth, we have presented a confusion matrix in Fig. 13. To show the effects of our method, we make the comparison of confusion matrices for DINO and DINO w/t SS-FLD. In this evaluation, we consider a true positive when the predicted and a GT box has more than a certain IoU score. Following the standard mAP evaluation on MSCOCO, we change the IoU threshold at [0.5 (start):0.95(end):0.05(step)] and evaluate a confusion matrix per IoU threshold. We then present the confusion matrices in terms of precision and recall rates in Fig. 13-(a) and (b).

For DINO, the average false discovery rate (FDR; $\frac{FP}{TP+TP}$) and false negative rate (FNR; $\frac{FN}{TP+FN}$) are 69.39% and 72.50%, respectively. On the other hand, for DINO with SS-FLD, the FDR and FNR are 62.72% and 69.69%. Thus, our SS-FLD reduces the rates by 3.67% and 2.81%. Remarkably, our method improves these rates for all the classes. We assert that these gains originated from our LF loss and semi-supervision. Interestingly, both detectors showed few false detections between light source classes and more confusion for the discrimination between light source and background. It might be due to light reflection, scattering, complex scene clutters, tiny light sizes, etc.

J Sensitivity analysis of attention prior term

We perform a sensitivity analysis for η within our LF loss Eq. (2) in our manuscript. To this end, we train different SS-FLDs w/t PAA by changing η within a range of $[1, 5]$ and compare their mAP scores as shown in Table 8. Albeit we obtain the best score with $\eta = 4$, the mAP difference of the interval $\eta = [2, 5]$ is so marginal. This means that our LF loss is not sensitive to η. When $\eta = 1$, some score reduction occurs due to the loss $LF = 0$ when $\phi(\mathbf{d}_c) = 1$.

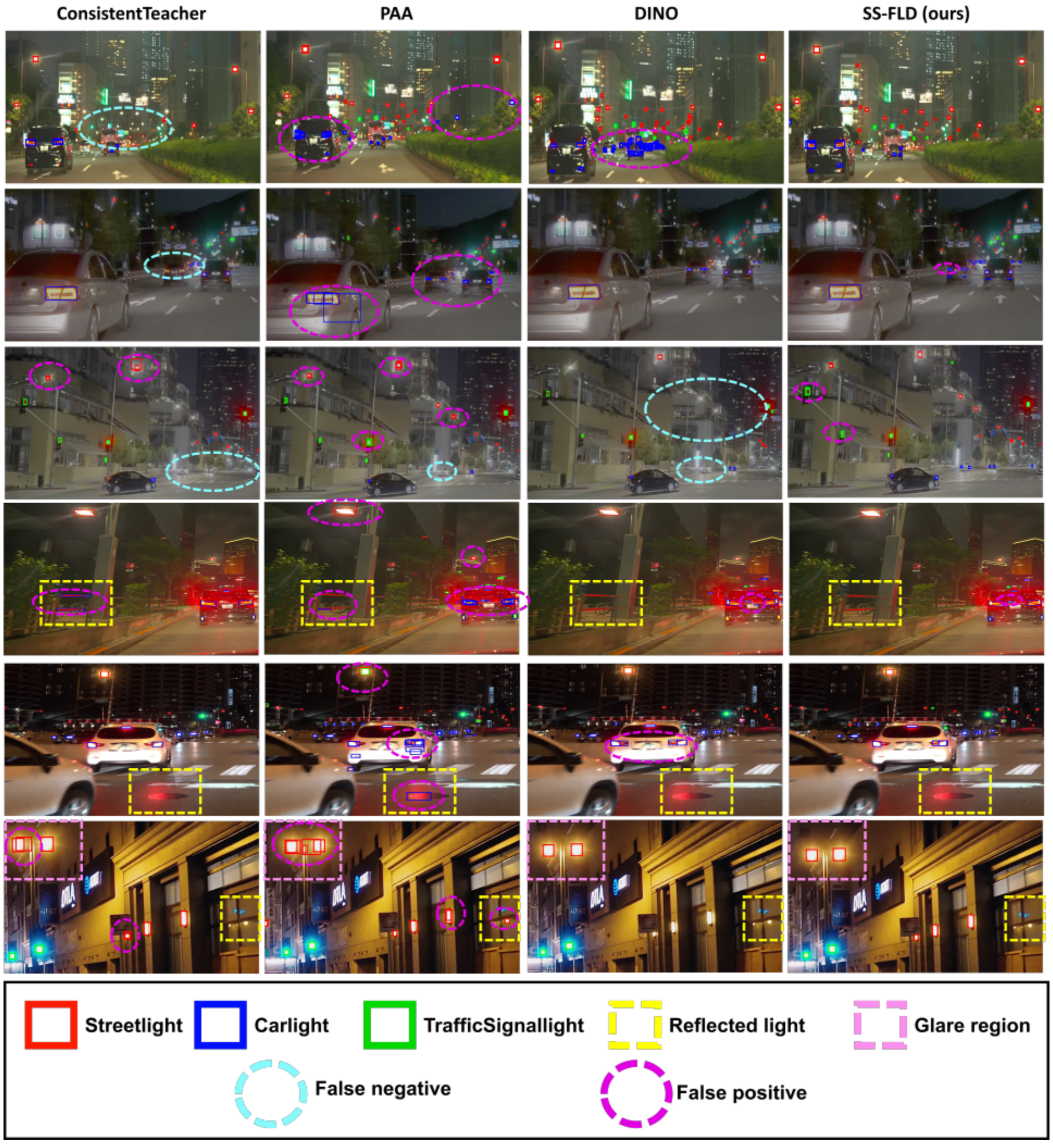


Figure 11: Qualitative comparison results between our SS-FLD, ConsistentTeacher, PAA, and DINO.

Table 8: Comparison with different η of the lightness prior.

η	mAP	AP_{50}	AP_{75}	AP_{vt}	AP_t	AP_s	AP_m
$\eta = 1$	24.8	52.4	20.2	11.3	21.4	39.2	46.9
$\eta = 2$	25.4	54.5	20.5	11.7	22.3	39.5	47.7
$\eta = 3$	25.5	54.8	20.4	11.9	22.5	39.8	47.2
$\eta = 4$	25.6	55.1	20.4	11.9	22.4	40.1	46.9
$\eta = 5$	25.4	54.7	20.3	12.1	22.6	39.8	46.3

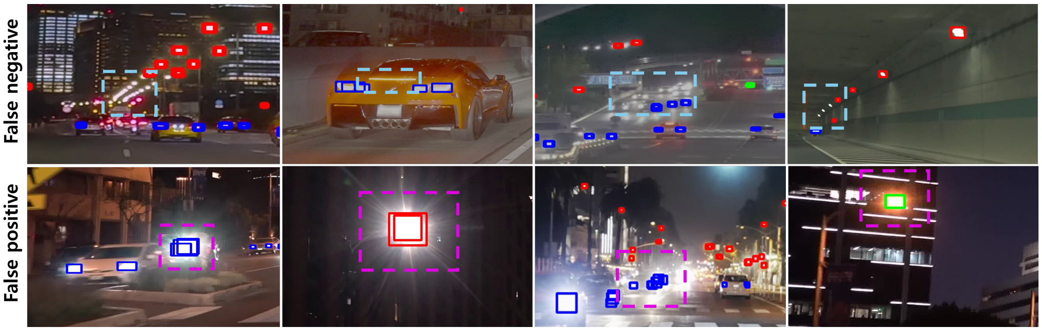
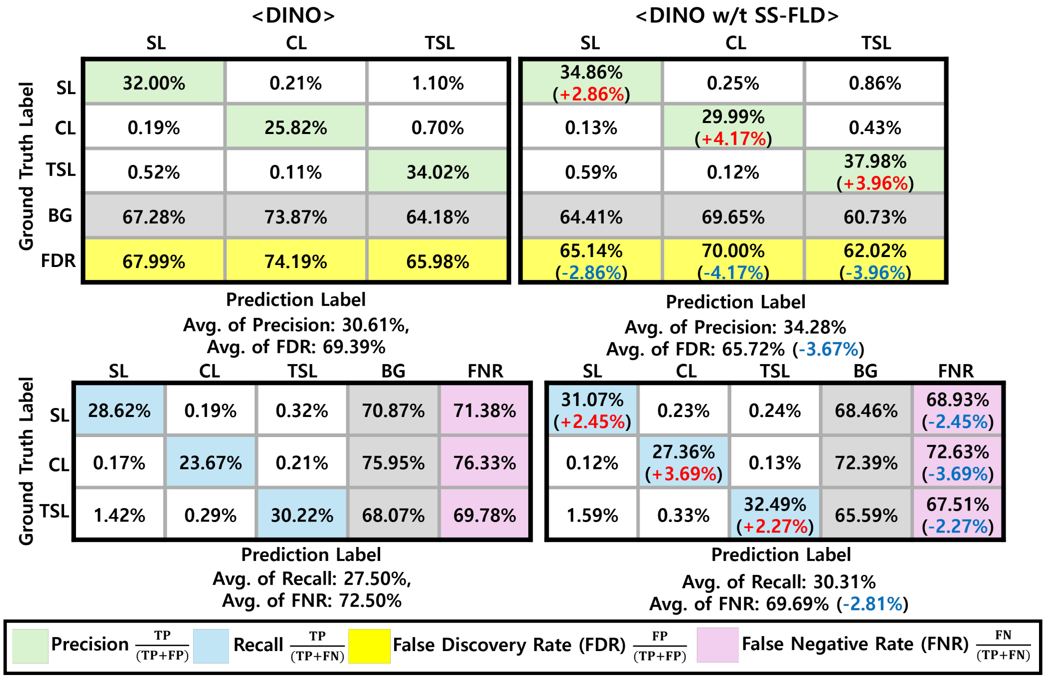


Figure 12: The failure cases of our SS-FLD. Red, blue, and green boxes represent Streetlight, Carlight, and TrafficSignallight classes, respectively. We highlight the false negative and false positive using the cyan and the magenta boxes, respectively.



<DINO>

Ground Truth Label	SL	CL	TSL
SL	32.00%	0.21%	1.10%
CL	0.19%	25.82%	0.70%
TSL	0.52%	0.11%	34.02%
BG	67.28%	73.87%	64.18%
FDR	67.99%	74.19%	65.98%

Prediction Label
Avg. of Precision: 30.61%,
Avg. of FDR: 69.39%

<DINO w/t SS-FLD>

Ground Truth Label	SL	CL	TSL
SL	34.86% (+2.86%)	0.25%	0.86%
CL	0.13%	29.99% (+4.17%)	0.43%
TSL	0.59%	0.12%	37.98% (+3.96%)
BG	64.41%	69.65%	60.73%
FDR	65.14% (-2.86%)	70.00% (-4.17%)	62.02% (-3.96%)

Prediction Label
Avg. of Precision: 34.28%
Avg. of FDR: 65.72% (-3.67%)

Ground Truth Label	SL	CL	TSL	BG	FNR
SL	28.62%	0.19%	0.32%	70.87%	71.38%
CL	0.17%	23.67%	0.21%	75.95%	76.33%
TSL	1.42%	0.29%	30.22%	68.07%	69.78%

Prediction Label
Avg. of Recall: 27.50%,
Avg. of FNR: 72.50%

Ground Truth Label	SL	CL	TSL	BG	FNR
SL	31.07% (+2.45%)	0.23%	0.24%	68.46%	68.93% (-2.45%)
CL	0.12%	27.36% (+3.69%)	0.13%	72.39%	72.63% (-3.69%)
TSL	1.59%	0.33%	32.49% (+2.27%)	65.59%	67.51% (-2.27%)

Prediction Label
Avg. of Recall: 30.31%
Avg. of FNR: 69.69% (-2.81%)

Precision $\frac{TP}{(TP+FP)}$ Recall $\frac{TP}{(TP+FN)}$ False Discovery Rate (FDR) $\frac{FP}{(TP+FP)}$ False Negative Rate (FNR) $\frac{FN}{(TP+FN)}$

Figure 13: The confusion matrices between original DINO and DINO w/t SS-FLD. SL, CL, TSL, and BG mean Streetlight, Carlight, TrafficSignallight, and background, respectively.

K Discussion

Based on the assumption in Sec. 4.2 of our manuscript, we design the spatial prior under the flat road. However, special situations such as highly inclined roads can affect the prior. To resolve this, a simple remedy does not use the prior in the LF loss, and we show that our SS-FLD works well without the prior by comparing the (M2) and (M4) of Table 8 in our manuscript. To learn the more generalized spatial prior of handling that, our future solution is to track VPs or saliency feature point motions on an image sequence. Then, we can update the spatial prior maps by reflecting the geometrical transformation on the prior map.

In addition, we further improve the SS-FLD using motion. To show the possibility, we conduct the additional evaluation, which leverages motion for re-weighting lightness p_c. We detect $\mathbf{d}_{c,t}$ and track $\hat{\mathbf{d}}_{c,t}$ lights by using our SS-FLD and Kalman Filtering (KF) at frame t. Then, we can determine the optimal matching pairs between $\mathbf{d}_{c,t}$ and $\hat{\mathbf{d}}_{c,t}$. For each pair, we can update $p_{c,t}$ as $\hat{p}_{c,t} = \sigma(p_{c,t} \times w_{s,t} \times w_{m,t})$ with a normalization function σ , where $w_{s,t} = \phi(\mathbf{d}_{c,t}) + 0.5$ and $w_{m,t} = 1.0 + \text{IoU}(\mathbf{d}_{c,t}, \hat{\mathbf{d}}_{c,t|t-1})$. $\phi(\mathbf{d}_{c,t})$ and $\hat{\mathbf{d}}_{c,t|t-1}$ are the spatial prior as shown in Sec. 4.1

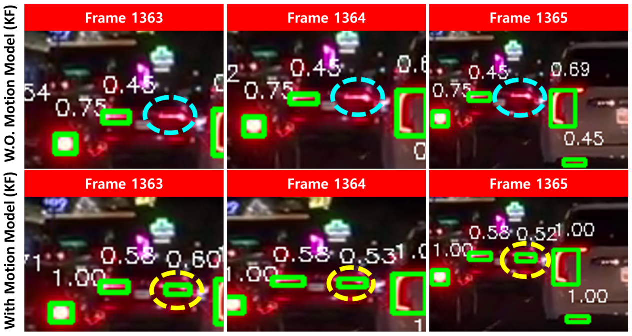


Figure 14: Qualitative comparison before/after re-weighting.

in our manuscript and the predicted motion. In Fig. 14, we compare detection results before/after re-weighting using KF. Some missed detections can be successfully detected by our re-weighted $\hat{p}_{c,t}$.

L Datasheet for Datasets

L.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description

- Our YDLD dataset is proposed to detect multi-source light sources that include a wide range of scenarios and objects, such as traffic lights, car lights, and streetlights. As we mentioned in Sec. 3.1 of our manuscript, accurate light source detection is essential for recognizing objects, predicting behaviors, and understanding the environment for autonomous driving and surveillance systems.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

- The YDLD dataset is created by the authors of this paper at Inha University.

What support was needed to make this dataset? (e.g., who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

- This work was supported in part by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. NRF-2022R1C1C1009208) and funded by the Ministry of Education (No.2022R1A6A1A03051705); supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No.2022-0-00448/RS-2022-II220448: Deep Total Recall, 30%, No.RS-2022-00155915: Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)); supported in part by INHA UNIVERSITY Research Grant

Any other comments?

- No.

L.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

- Instance of the YDLD dataset is comprised of images and their associated light source annotation data.

How many instances are there in total (of each type, if appropriate)?

- Our YDLD dataset contains 3,516 images and 116,028 light source instances. We provide the detailed statistics of our YDLD dataset in Table 1 of our manuscript.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

- The dataset contains all instances.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description

- Instances of the dataset consist of images.

Is there a label or target associated with each instance?

- Yes, we provide the light source label json file. As we mentioned in Sec. 3.2 of our manuscript, our label is followed MS-COCO format.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

- No, information is not missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

- No, individual instances of this dataset are treated as independent.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

- Yes, we split our YDLD dataset for the training and testing. The detailed information is provided in Table 1 of our manuscript.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

- No, there are no redundancies in this dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

- No.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

- No, the YDLD dataset does not contain confidential protected data.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

- No, our dataset does not contain offensive, insulting, and threatening data.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

- No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

- No, we do not identify people by subpopulation.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

- No. In the YDLD dataset, as mentioned in Sec. 3.2 of the main manuscript, it is not possible to identify elements that could reveal personal information (*e.g.* faces and license plates).

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

- No.

Any other comments?

- No.

L.3 Collection

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how

- The YDLD dataset is collected from the YouTube videos. We mentioned our collection process in Sec. 3.2 of our manuscript.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

- We manually collected our data. The collection procedure is mentioned in Sec. 3.2 of our manuscript.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

- The authors of this paper and students participated in the data collection for this study. Each author and student received appropriate payments and compensations in accordance with internal regulations.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.

- The data was collected from August 2022 to February 2023.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

- No.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

- No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

- We collected data from YouTube, not from individuals in directly.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

- No, we did not notify any individuals about the data collection.

Any other comments?

- No.

L.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

- Yes, we perform the labeling for our YDLD dataset. This process is described in Sec. 3.2 of our manuscript.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

- No.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

- We use the coco annotation tools for labeling as described in Sec. 3.2 of our manuscript.

Any other comments?

- No.

L.5 Use

Has the dataset been used for any tasks already? If so, please provide a description.

- No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

- No.

What (other) tasks could the dataset be used for?

- We encourage researchers to use the YDLD dataset for light source detection research, developing technologies related to autonomous driving, and other related studies.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

- No.

Are there tasks for which the dataset should not be used? If so, please provide a description.

- No.

Any other comments?

- No.

L.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

- Yes, the YDLD dataset will be published in public.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

- The dataset is available through https://github.com/YDLD-dataset/YDLD.

When will the dataset be distributed?

- We initially opened a portion of the YDLD dataset on June 13, 2024, and the full data will be released on October 30, 2024.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

- The YDLD dataset is distributed under the CC BY-NC-ND 4.0 license, meaning it can be used by anyone for non-commercial research purposes.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

- No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation

- No.

Any other comments?

- No.

L.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

- The authors of this paper will support hosting and maintaining of this dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

- The user of this dataset can contact us by GitHub https://github.com/YDLD-dataset/YDLD.

Is there an erratum? If so, please provide a link or other access point.

- There are no erratum in our initial dataset release. Any future corrections will be documented in the dataset's GitHub repository.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

- No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

- Yes, our dataset will be continued to support in our GitHub page.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

- Yes, they can directly contact us via GitHub issue. We will also notify the contactable E-mail soon.

Any other comments?

- No.

References

[1] X. Wang, X. Yang, S. Zhang, Y. Li, L. Feng, S. Fang, C. Lyu, K. Chen, and W. Zhang, "Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection," in *CVPR*, pp. 3240–3249, 2023.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, pp. 740–755, 2014.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, pp. 2980–2988, 2017.

[4] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *CVPR*, pp. 658–666, 2019.

[5] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-iou loss: Faster and better learning for bounding box regression,” in *AAAI*, vol. 34, pp. 12993–13000, 2020.

[6] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite,” in *CVPR*, pp. 3354–3361, 2012.

[7] J. Han, X. Liang, H. Xu, K. Chen, L. Hong, J. Mao, C. Ye, W. Zhang, Z. Li, X. Liang, and C. Xu, “SODA10M: A large-scale 2d self/semi-supervised object detection dataset for autonomous driving,” in *NeurIPS*, 2021.

[8] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*, pp. 248–255, 2009.

[9] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, “BDD100K: A diverse driving dataset for heterogeneous multitask learning,” in *CVPR*, pp. 2633–2642, 2020.

[10] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, “Libra r-cnn: Towards balanced learning for object detection,” in *CVPR*, pp. 821–830, 2019.

[11] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *NeurIPS*, vol. 28, 2015.

[12] Z. Cai and N. Vasconcelos, “Cascade r-cnn: Delving into high quality object detection,” in *CVPR*, pp. 6154–6162, 2018.

[13] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, “You only look one-level feature,” in *CVPR*, pp. 13039–13048, 2021.

[14] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, “Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection,” in *CVPR*, pp. 9759–9768, 2020.

[15] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “Yolox: Exceeding yolo series in 2021,” *arXiv*, 2021.

[16] K. Kim and H. S. Lee, “Probabilistic anchor assignment with iou prediction for object detection,” in *ECCV*, pp. 355–371, 2020.

[17] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” in *ICLR*, 2021.

[18] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H. Shum, “DINO: DETR with improved denoising anchor boxes for end-to-end object detection,” in *ICLR*, 2023.

[19] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, “End-to-end semi-supervised object detection with soft teacher,” *ICCV*, 2021.

[20] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *NeurIPS*, pp. 1195–1204, 2017.

[21] B. Du, Y. Huang, J. Chen, and D. Huang, “Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images,” in *CVPR*, pp. 13435–13444, 2023.

[22] J. Wu, Z. Pan, B. Lei, and Y. Hu, “Fsanet: Feature-and-spatial-aligned network for tiny object detection in remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[23] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, “Rfla: Gaussian receptive field based label assignment for tiny object detection,” in *ECCV*, pp. 526–543, 2022.

[24] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, “Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 79–93, 2022.

[25] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, pp. 98–136, 2015.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *CVPR*, pp. 618–626, 2017.