—Supplementary Materials—
Lite DETR : An Interleaved Multi-Scale Encoder for Efficient DETR

Feng Li[1,2*], **Ailing Zeng**[2], **Shilong Liu**[2,3], **Hao Zhang**[1,2], **Hongyang Li**[2,4]
Lei Zhang[2†], **Lionel M. Ni**[1,5]
[1]The Hong Kong University of Science and Technology.
[2]International Digital Economy Academy (IDEA).
[3]Dept. of CST., BNRist Center, Institute for AI, Tsinghua University.
[4]South China University of Science and Technology.
[5]The Hong Kong University of Science and Technology (Guangzhou).

{fliay,hzhangcx}@connect.ust.hk
{liusl20}@mails.tsinghua.edu.cn
{eeli.hongyang}@mail.scut.edu
{leizhang}@idea.edu.cn
{ni}@ust.hk

In this supplementary material, we share more qualitative analyses compared with other related works and verify our design choices that are not presented in our main paper, including:

- Comparisons between sparse DETR and Lite deformable DETR in Sec. A.
- Comparisons between DINO-3scale and Lite DINO in Sec. B.
- Comparisons between deformable attention and KDA attention in Sec. C.
- More visualization of sampled locations between deformable attention and KDA attention in Sec. D.
- Our failure cases in Sec. E.

All the models are based on Deformable DETR and DINO, which are denoted as Lite-Deformable DETR and Lite DINO. Except for models in Sec. A that are based on Deformable DETR, other models are based on DINO. Please note that all boxes shown in these figures are selected from predicted boxes of the corresponding models with classification scores larger than 0.3.

A. Comparison between Sparse DETR and Lite Deformable DETR

To further analyze why our Lite-Deformable DETR outperforms Sparse DETR [1], we conduct a visualization of these two models in Fig. 1. In Fig. 1(a), we show that Sparse DETR may miss small objects in some cases. More importantly, in Fig. 1(b), (c), and (d), we demonstrate that Sparse DETR is inferior to our model in medium and large objects in that it tends to predict duplicate and wrong boxes. This phenomenon is also consistent with the results in Table 1, i.e., our model outperforms Sparse DETR by 1.6 AP in AP_L under comparable GFLOPs. As Sparse DETR only selects some tokens from multi-scale features, it breaks the structured feature organization, especially for high-level features with rich semantics. Therefore, it impacts large object detection and is difficult to plug in existing detection models as a general strategy.

B. Comparison between DINO-3scale and Lite DINO

As we claimed in the main paper, the high-resolution (low-level) map is redundant but important and should be preserved properly. Simply dropping these features will harm the performance of small object detection. In Fig. 2, we present the visualization comparisons of directly dropping

[*] This work was done when Feng Li was an intern at IDEA.
[†] Corresponding author.

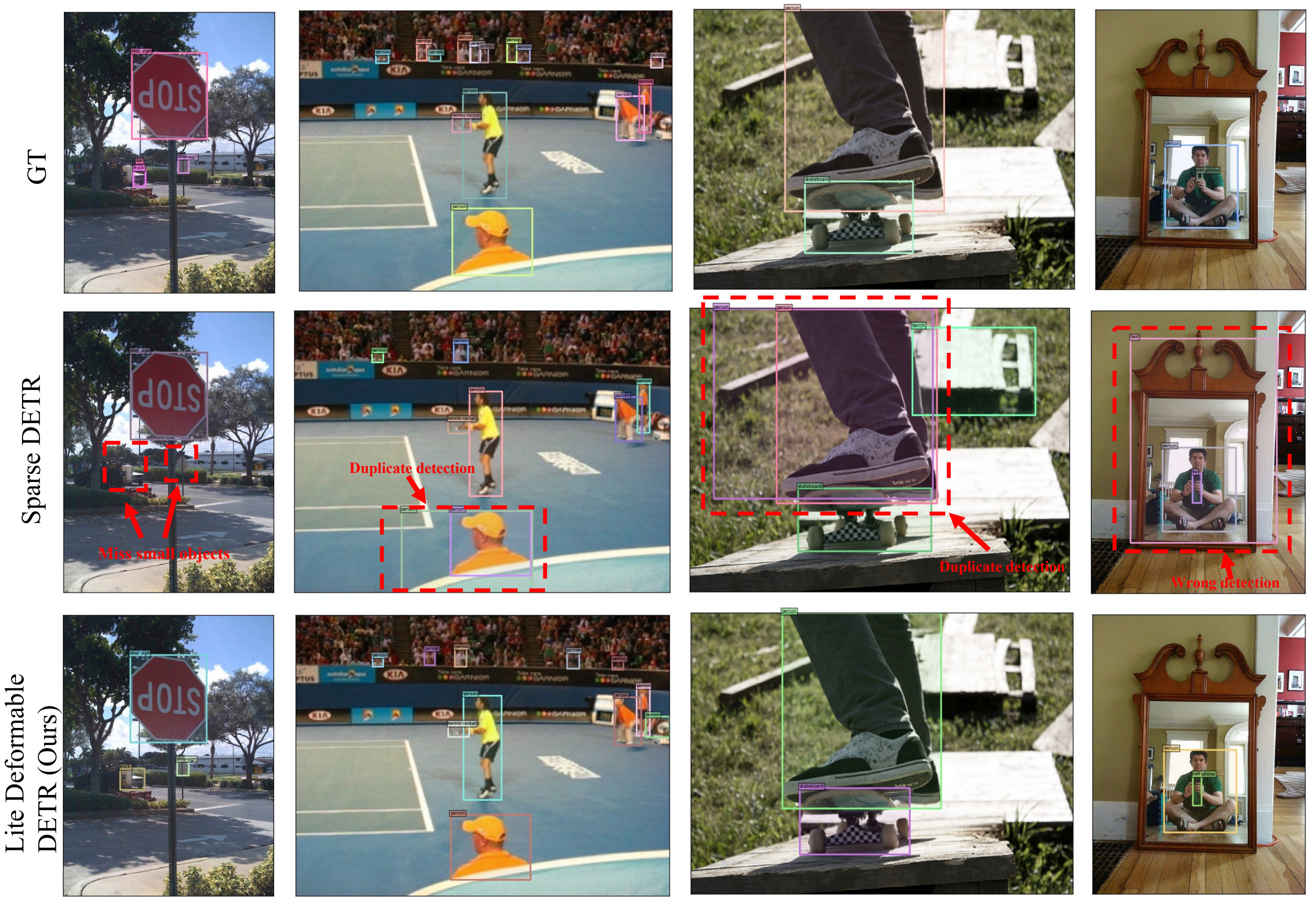
Model	#epochs	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$	GFLOPs	Encoder GFLOPs	Params
Deformable DETR† [3]	50	46.8	66.0	50.6	29.8	49.7	62.0	177	90	40M
Lite-Deformable DETR H3L1-(2+1)x3(25%, ours)	50	**46.7**	66.1	50.6	29.1	49.7	**62.2**	123	39	41M
Sparse DETR*-rho-0.3 [3]	50	46.0	65.9	49.7	29.1	49.1	60.6	127	40	41M

Table 1. Results for Sparse DETR and Lite-Deformable DETR under the ResNet-50 backbone. * Sparse DETR is based on an improved Deformable DETR baseline that combines the components from Efficient DETR [2]. 'rho' is the keeping ratio of encoder tokens in Sparse DETR. Value in the parenthesis indicates the percentage of our high-level tokens compared to the original features. † we adopt the result from the official Deformable DETR codebase.

the high-resolution map (DINO-3scale) and our proposed Lite DINO. Detecting small objects would be more difficult for DINO-3scale. Though Lite DINO slightly increases the GFLOPs compared to DINO-3scale, it maintains comparable performance as the original DINO.

C. Comparison between Deformable and KDA Attention

To enhance the lagged low-level feature update in the Lite DETR framework, we visualize how our KDA attention outperforms the original deformable attention in Fig. 3. KDA attention shows its superiority in improving small object



(a) (b) (c) (d)

Figure 1. Visualization of detection results in Sparse DETR and our Lite-Deformable DETR. (a) shows that Sparse DETR may miss small objects in some cases. (b), (c) and (d) demonstrate that Sparse DETR is inferior to our model in medium and large objects, where it tends to predict duplicate and wrong boxes.

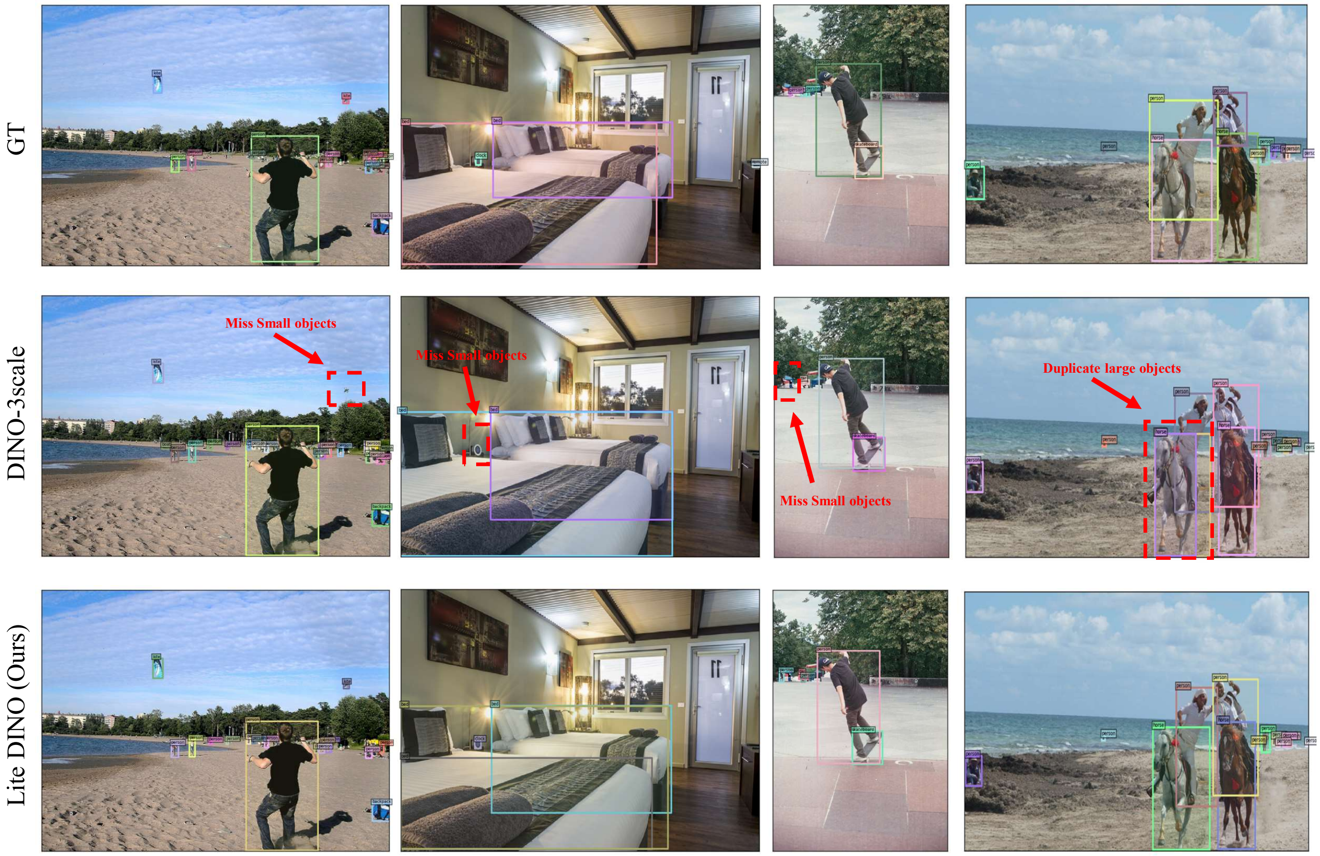


Figure 2. Visualization of detection results in DINO-3scale and our proposed Lite DINO. Directly dropping the high-resolution map (DINO-3scale) will make small object detection difficult, while Lite DINO can maintain comparable performance as the original DINO.

detection and reducing duplicate detection, indicating it can mitigate the effects of asynchronous features exposed in the Lite DETR framework.

D. More Visualization of Sampled Locations between Deformable and KDA Attention

In Fig. 5 in our paper, we provide a few visualization maps of deformable and KDA attention under Lite DINO. To better show their differences, we further provide more attention to visualization results in our interleaved encoder. Following Fig. 5 in our paper, we show the top 200 sampling locations on all scales (S1, S2, S3, and S4) for all query tokens in Fig. 4 and 5. In high-level feature maps (S1, S2), KDA and deformable attention focus on similar regions. However, kDA can sample more meaningful locations on low-level maps with high resolution (S3, S4), which indicates KDA can better extract local features from low-level maps and improve average precision in small object detection.

E. Failure Cases in Lite DETR

In Fig. 6, we analyze the cases when our method fails, including occlusion, blur, ambiguity, and reflection. These cases are quite difficult to detect, even for a human, and we will leave it for future work to address these cases.

References

[1] Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse DETR: Efficient End-to-End Object Detection with Learnable Sparsity. *arXiv preprint arXiv:2111.14330*, 2021. 1

[2] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient DETR: Improving End-to-End Object Detector with Dense Prior. *arXiv preprint arXiv:2104.01318*, 2021. 2

[3] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021. 2

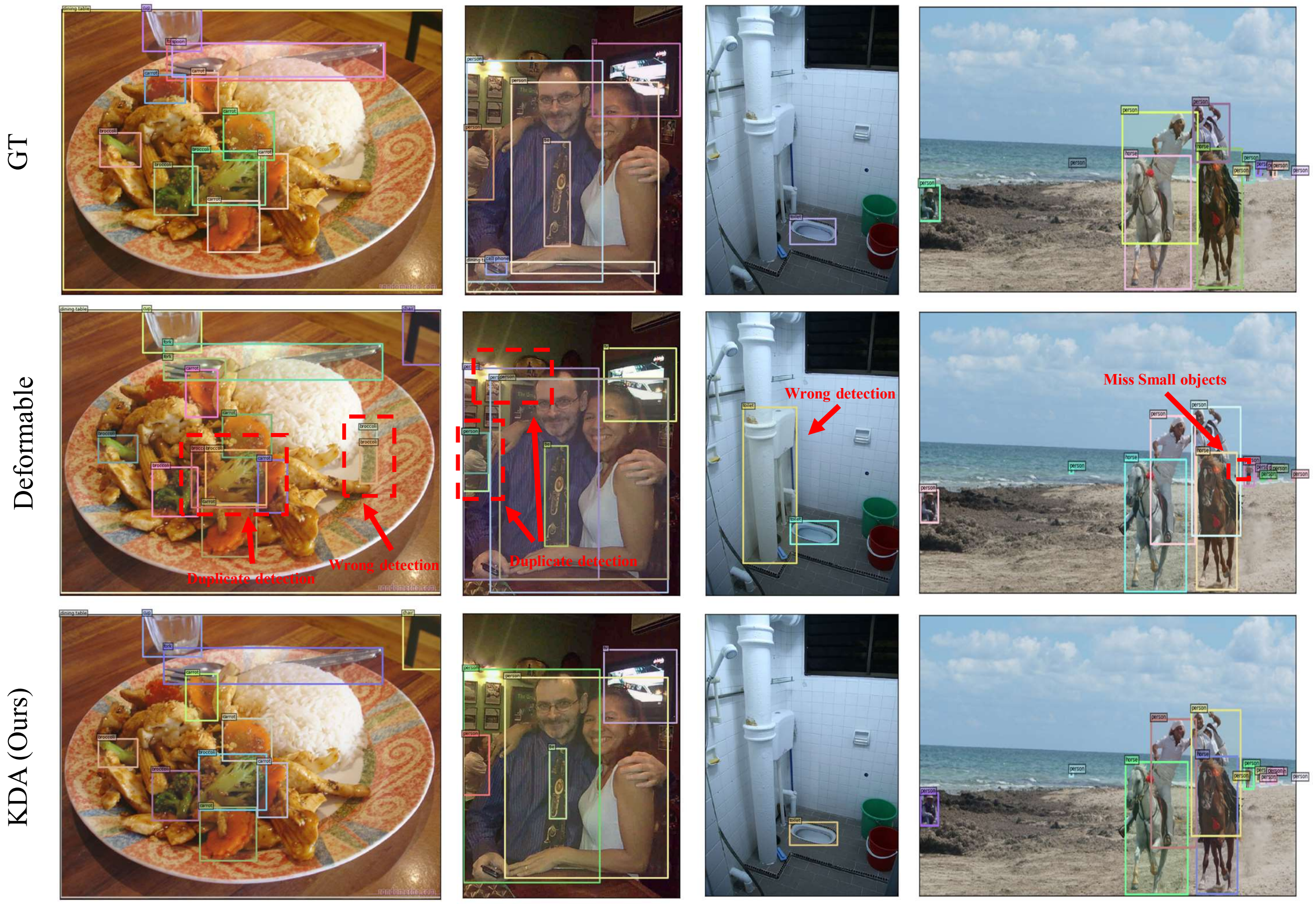


Figure 3. Visualization of detection results in our Lite DINO by using deformable attention and the proposed KDA attention. KDA attention shows its superiority in improving small object detection and reducing duplicate detection.

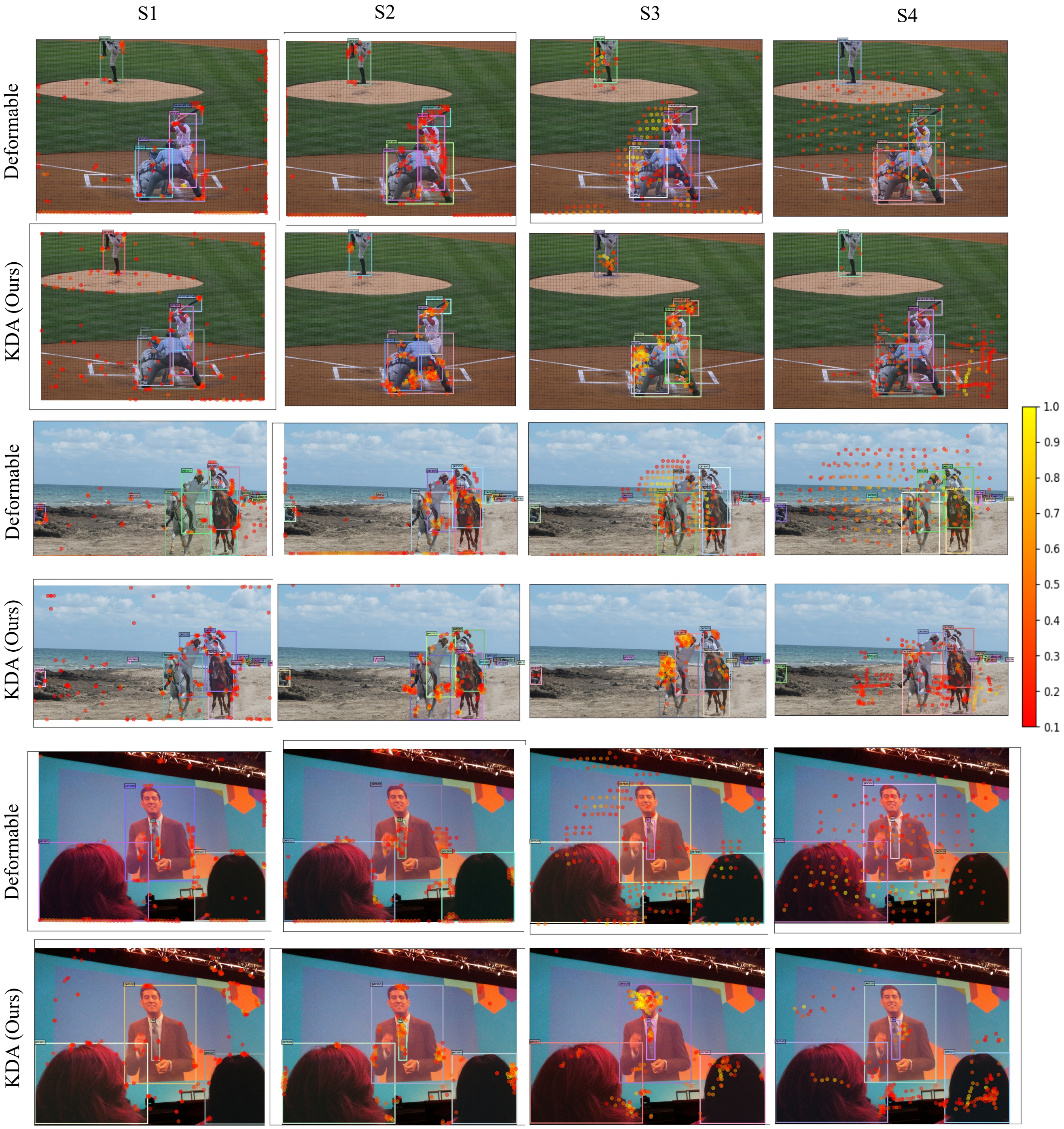


Figure 4. Visualization comparison of deformable and KDA attention in our Lite DINO from all feature scales. In high-level feature maps (S1, S2), KDA and deformable attention show similar attention regions. However, kDA can sample more meaningful locations on low-level maps (S3, S4), which indicates KDA can better extract local features from low-level maps and improve average precision in small object detection.

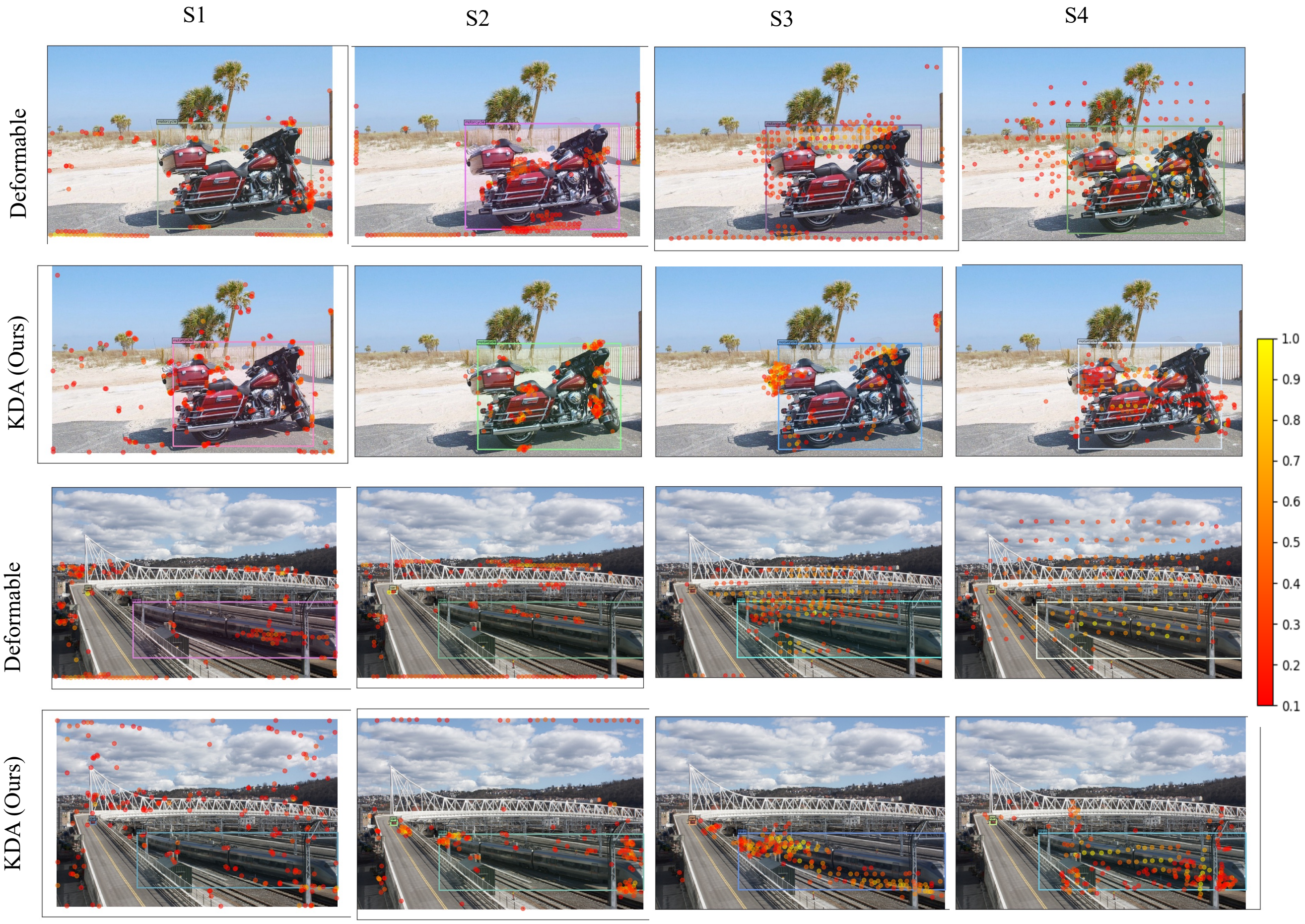


Figure 5. Visualization comparison of deformable and KDA attention in our Lite DINO from all feature scales.

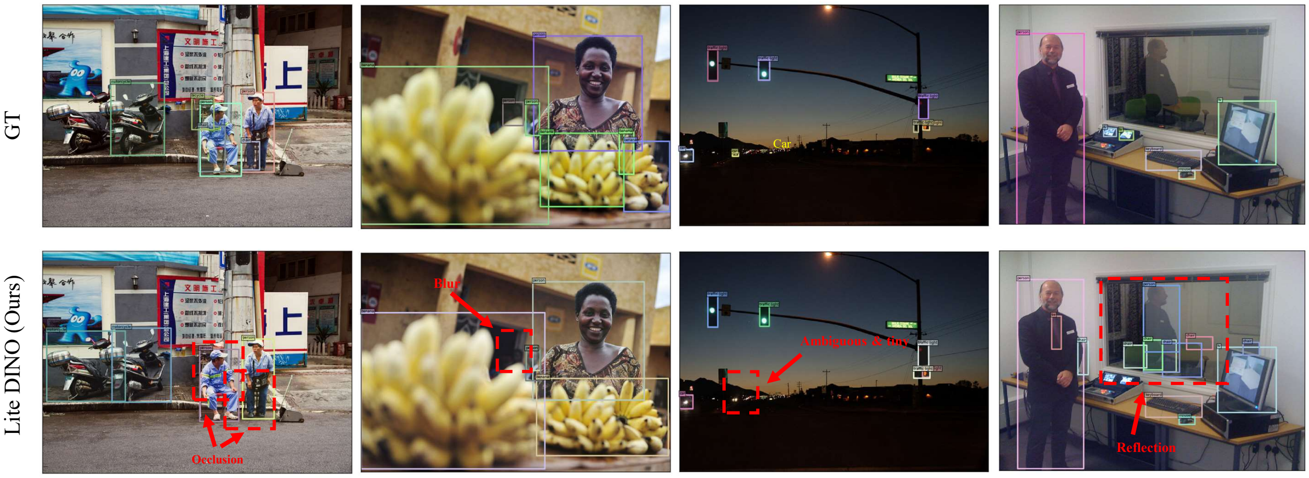


Figure 6. Visualization of failure cases in our Lite DINO.