

Supplementary Materials

Anonymous Authors

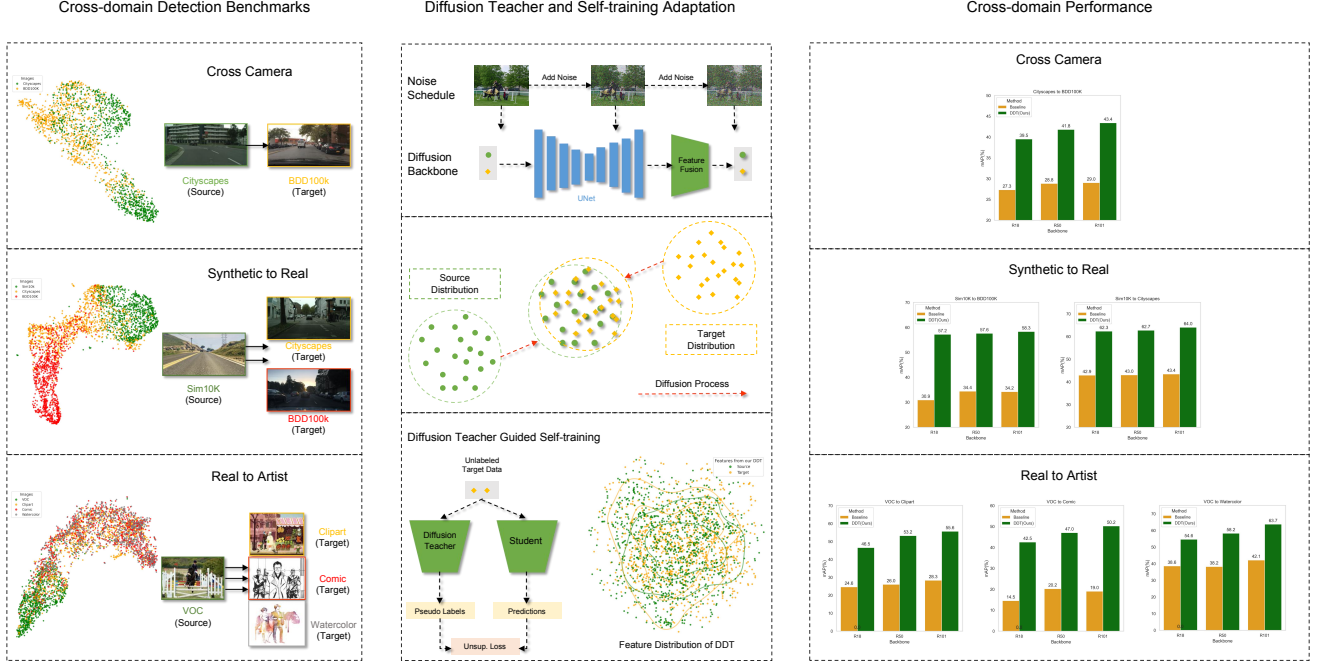


Figure 1: Main content of our paper. Left: We present three cross-domain detection benchmarks and visualize the image distributions from source and target domains using the UMAP method. It is evident that there is a large gap between different domains. Middle: We utilize a frozen-weight diffusion model as the backbone to extract features, and employ a detector with the diffusion backbone as the teacher to guide the learning of the student model on the target domain with the self-training framework. Based on the visualization results of the feature distributions, our method significantly reduces the domain gap. Right: Our method substantially enhances the performance of cross-domain object detection without increasing the inference cost.

1 ADDITIONAL ABLATION STUDIES

In this section, we present additional ablation studies and comparisons to further investigate the effectiveness of our proposed Diffusion Domain Teacher (DDT).

Ablation Study on Unsup Loss Weight λ . We investigate the impact of different unsupervised loss weight λ on cross-domain detection in Tab.1, including Cityscapes [8] to BDD100K [47] (Cs→B), Sim10K [21] to Cityscapes [8] (S→Cs), and VOC [12] to Clipart [19] (V→Ca). Excessively high or low weight λ will degrade the performance of cross-domain detection, so we simply set the parameter λ to 1.

Ablation Study on the Threshold σ of Pseudo Labels. In Tab. 2, we show the impact of different pseudo label threshold σ on the final results. The threshold plays a crucial role in determining the quality of the generated pseudo labels during self-training, and inappropriate threshold will affect cross-domain performance. Ultimately, we choose 0.5 as the default threshold σ setting in all our experiments.

Ablation Study on Data Augmentation of Sup. and Unsup. Branches. We follow previous work [31] and utilize *Strong Augmentation* and *Weak Augmentation* during the self-training process. Tab. 3 presents the impact of *Strong Augmentation* on both supervised and unsupervised data, highlighting the importance of data augmentation in the self-training process.

Ablation Study on Different Versions of Stable Diffusion. In Tab. 4, we present the results of using the popular Stable Diffusion V1.5 (SD-1.5) and the latest version (SD-2.1) respectively. Overall, the results of SD-1.5 are superior to those using SD-2.1, except for the cross-domain detection from Sim10K to BDD100K.

Comparison of Backbone Efficiencies. In Tab. 5, we present a comparative analysis of these models with respect to their architectural sizes, training cost, and inference latency. Diffusion model, with its substantial parameter count and protracted inference time, emerges as an impractical choice for deployment in routine detection tasks within operational settings. Nevertheless, the employment of this model as a strong teacher in self-training framework offers a strategic avenue to leverage its exceptional cross-domain prowess, achieving an average absolute improvement

Table 1: Ablation Result of Unsup. Loss Weight λ .

λ	Cs→B	S→Cs	V→Ca
0.33	42.7	63.5	54.7
0.50	42.4	63.7	55.3
<u>1.0</u>	43.4	64.0	55.6
2.0	43.0	64.0	53.3
3.0	42.9	63.9	50.0

Table 2: Ablation Result of Threshold σ for Pseudo Labels.

σ	Cs→B	S→Cs	V→Ca
0.3	41.9	63.7	52.4
0.4	42.3	64.2	53.9
<u>0.5</u>	43.4	64.0	55.6
0.6	43.2	62.1	56.1
0.7	43.0	60.3	53.2

Table 3: Ablation Study of Data Augmentation.

Settings of data Aug.	Cs→B	S→Cs	V→Ca
DDT(Ours)	43.4	64.0	55.6
w/o Strong Aug. on Sup.	42.0-1.4	62.0-2.0	54.3-1.3
w/o Strong Aug. on Unsup	42.5-0.9	62.2-1.8	52.3-3.3
w/o All Aug.	42.0-1.4	61.8-2.2	52.1-3.5

Table 4: Ablation Result of Different Stable Diffusion Versions.

Detector	Version	Cs→B	S→Cs	S→B	V→Ca	V→Co	V→W
Diffusion Detector	SD-1.5	32.7	58.2	50.1	47.4	44.4	58.7
	SD-2.1	34.6	58.9	56.4	45.4	42.8	55.2
DDT	<u>SD-1.5</u>	43.4	64.0	58.3	55.6	50.2	63.7
	SD-2.1	42.3	63.4	61.6	53.7	48.9	63.3

of 21.2 mAP and a relative improvement of 39.7% compared to the baseline, without introducing any additional inference overhead.

2 ADDITIONAL EXPERIMENTS

In this section, we showcase further experiments, including: (1) Cityscapes [8] to FoggyCityscapes [39], aiming to validate the results of adverse weather adaptation, and (2) the outcomes of our DDT method employing the FCOS [42] detector, aiming to assess the performance of our approach across different detectors.

Adverse Weather Adaptation from Cityscapes to FoggyCityscapes. We present the result of adverse weather adaptation in Tab. 6. Compared to the baseline, our method with R18, R50, and R101 show improvements of 11.8, 18.0, and 21.6 mAP, respectively. However, our method (50.0 mAP) still falls short of surpassing CMT [1] (50.3 mAP) and HT [10] (50.4 mAP). FoggyCityscapes [39] is a dataset where foggy weather conditions are added to Cityscapes,

with similar images and identical labels from Cityscapes. We observe that our DDT method does not demonstrate superior performance in inter-domain as train and test on Cityscapes, which might be the reason for our weaker performance on the FoggyCityscapes compared to the current state-of-the-art results.

Results of Adaptation with FCOS Detector. Previous domain adaptation methods for detection primarily employ the Faster RCNN and FCOS detectors. In main text, we report the results of our method using the Faster RCNN [6] detector. To further validate the effectiveness of our approach, we also present the performance using the FCOS [42] detector in Tab. 7, 8, 9, 10, 11, 12. Overall, the FCOS detector yielded results comparable to the Faster RCNN and outperforms the previous SOTA results in five out of the six datasets, with the exception of Sim10K to Cityscapes where it falls short of HT [10].

3 ADDITIONAL VISUALIZATION RESULTS

In Fig. 2, 3, 4, 5, 6, 7, we present additional visualization results for Cityscapes to BDD100K (Cs→B), Sim10K to BDD100K (S→B), Sim10K to Cityscapes (S→Cs), VOC to Clipart (V→Ca), VOC to Comic (V→Co), and VOC to Watercolor (V→W).

Table 5: Results of the efficiency comparison of detectors with different backbones.

Methods	Params (M)	Flops (G)	Train Time (s/iter)	Inference Time (ms/image)	Average Gain (mAP)	Average Rel. Improve (%)
ConvNext-Base [36]	105	401	0.672	73.2	/	/
Swin-Base [35]	104	413	0.693	79.4	/	/
VIT-Base [11]	107	605	0.634	116.5	/	/
MAE (VIT-Base) [11, 15]	107	605	0.634	116.5	/	/
GLIP (Swin-Tiny) [24, 35]	45	198	0.486	84.1	/	/
Diff. (Ours)	991	8,256	1.56	780.4	/	/
DDT (R18)	28	137	0.82	24.6	19.0	38.9
DDT (R50)	41	184	0.90	33.6	21.7	40.1
DDT (R101)	60	262	1.04	40.9	22.9	40.2

Table 6: Quantitative results on adaptation from Cityscapes to FoggyCityscapes. The bold indicates the best results.

Method	Reference	Detector	bus	bicycle	car	mcycle	person	rider	train	truck	mAP
UMT [9]	<i>CVPR'21</i>	FRCNN-R101	56.5	37.3	48.6	30.4	33.0	46.7	46.8	34.1	41.7
IIOB [43]	<i>TPAMI'21</i>	FRCNN-V16	46.1	35.3	49.6	29.9	32.8	44.4	38.0	33.0	38.6
SADA [7]	<i>IJCV'21</i>	FRCNN-R50	50.3	45.4	62.1	32.4	48.5	52.6	31.5	29.5	44.0
CDG [25]	<i>AAAI'21</i>	FRCNN-V16	47.5	38.9	53.1	38.3	38.0	47.4	41.1	34.2	42.3
UaDAN [14]	<i>TMM'21</i>	FRCNN-R50	49.4	38.9	53.6	32.3	36.5	46.1	42.7	28.9	41.1
VDD [44]	<i>ICCV'21</i>	FRCNN-V16	52.0	36.8	51.7	34.2	33.4	44.0	34.7	33.9	40.0
O2Net [13]	<i>ACMMM'22</i>	DETR-R50	47.6	45.9	63.6	38.0	48.7	51.5	47.8	31.1	46.8
SSAL [37]	<i>NeurIPS'22</i>	FCOS	50.0	38.7	59.4	26.0	45.1	47.4	25.7	24.5	39.6
DDF [32]	<i>TMM'22</i>	FRCNN-R50	50.4	39.8	56.1	31.1	37.6	45.5	47.0	30.7	42.3
D-ADAPT [20]	<i>ICLR'22</i>	FRCNN-R50	36.3	46.1	61.7	37.3	44.9	54.2	24.7	25.6	42.2
SCAN [27]	<i>AAAI'22</i>	FCOS-R50	48.6	37.3	57.3	31.0	41.7	43.9	48.7	28.7	42.1
SIGMA [28]	<i>CVPR'22</i>	FCOS-R50	50.4	40.6	60.3	31.7	44.0	43.9	51.5	31.6	44.2
TIA [49]	<i>CVPR'22</i>	FRCNN-V16	52.1	38.1	49.7	37.7	34.8	46.3	48.6	31.1	42.3
TDD [16]	<i>CVPR'22</i>	FRCNN-V16	53.0	49.1	68.2	38.9	50.7	53.7	45.1	35.1	49.2
NLTE [33]	<i>CVPR'22</i>	FRCNN-R50	56.7	43.3	58.7	33.7	43.1	50.7	42.7	33.6	45.4
LODS [26]	<i>CVPR'22</i>	FRCNN-V16	39.7	37.8	48.8	33.2	34.0	45.7	19.6	27.3	35.8
PSN [41]	<i>CVPR'22</i>	FRCNN-V16	48.7	39.2	53.0	33.1	37.4	45.2	38.8	31.1	40.9
MGA [52]	<i>CVPR'22</i>	FCOS-R101	53.2	36.9	61.5	27.9	43.1	47.3	50.3	30.2	43.8
MTTrans [48]	<i>ECCV'22</i>	DETR-R50	45.9	46.5	65.2	32.6	47.7	49.9	33.8	25.8	43.4
OADA [46]	<i>ECCV'22</i>	FCOS-V16	48.5	39.8	62.9	34.3	47.8	46.5	50.9	32.1	45.4
SCAN++ [27]	<i>TMM'22</i>	FCOS-R101	48.1	39.5	57.9	30.1	44.2	43.9	51.2	28.2	42.8
MIC [17]	<i>CVPR'23</i>	FRCNN-R101	52.4	47.5	67.0	40.6	50.9	55.3	33.7	33.9	47.6
SIGMA++ [29]	<i>TPAMI'23</i>	FRCNN-V16	52.2	39.9	61.0	34.8	46.4	45.1	44.6	32.1	44.5
CIGAR [34]	<i>CVPR'23</i>	FCOS-V16	56.6	41.3	62.1	33.7	46.1	47.3	44.3	27.8	44.9
CMT [1]	<i>CVPR'23</i>	FRCNN-V16	66.0	51.2	63.7	41.4	45.9	55.7	38.8	39.6	50.3
HT [10]	<i>CVPR'23</i>	FCOS-V16	55.9	50.3	67.5	40.1	52.1	55.8	49.1	32.7	50.4
Baseline	/	FRCNN-R18	38.6	31.3	45.6	26.1	37.6	45.6	13.9	17.6	32.0
DDT(Ours)	/	FRCNN-R18	49.4	44.0	59.0	36.3	47.9	56.5	27.8	30.0	43.8+ 11.8
Baseline	/	FRCNN-R50	39.1	32.0	42.2	23.8	36.4	44.6	14.7	19.7	31.6
DDT(Ours)	/	FRCNN-R50	53.2	51.5	63.8	44.1	50.3	59.3	41.7	33.1	49.6+ 18.0
Baseline	/	FRCNN-R101	35.7	31.9	41.6	23.8	34.9	41.9	5.7	19.7	29.4
DDT(Ours)	/	FRCNN-R101	53.5	52.2	64.2	43.5	50.9	60.0	42.4	33.6	50.0+ 21.6

Table 7: Quantitative results on adaptation from Cityscapes to BDD100K (Cs→B) with FCOS. The bold indicates the best results.

Method	Reference	Detector	bicycle	bus	car	mcycle	person	rider	truck	mAP
DA-Faster [6]	<i>CVPR'18</i>	FRCNN-V16	22.4	18.0	44.2	14.2	28.9	27.4	19.1	24.9
SWDA [38]	<i>CVPR'19</i>	FRCNN-V16	23.1	20.7	44.8	15.2	29.5	29.9	20.2	26.2
SCDA [54]	<i>CVPR'19</i>	FRCNN-V16	23.2	19.6	44.4	14.8	29.3	29.2	20.3	25.8
CRDA [45]	<i>CVPR'20</i>	FRCNN-R101	25.5	20.6	45.8	14.9	32.8	29.3	22.7	27.4
SED [30]	<i>AAAI'21</i>	FRCNN-V16	25.0	23.4	50.4	18.9	32.4	32.6	20.6	29.0
TDD [16]	<i>CVPR'22</i>	FRCNN-V16	28.8	25.5	53.9	24.5	39.6	38.9	24.1	33.6
PT [5]	<i>ICML'22</i>	FRCNN-V16	28.8	33.8	52.7	23.0	40.5	39.9	25.8	34.9
EPM [18]	<i>ECCV'20</i>	FCOS-R101	20.1	19.1	55.8	14.5	39.6	26.8	18.8	27.8
SIGMA [28]	<i>CVPR'22</i>	FCOS-R50	26.3	23.6	64.1	17.9	46.9	29.6	20.2	32.7
SIGMA++ [29]	<i>TPAMI'23</i>	FRCNN-V16	27.1	26.3	65.6	17.8	47.5	30.4	21.1	33.7
NSA [53]	<i>ICCV'23</i>	FRCNN-V16	/	/	/	/	/	/	/	35.5
HT [10]	<i>CVPR'23</i>	FCOS-V16	38.0	30.6	63.5	28.2	53.4	40.4	27.4	40.2
Baseline	/	FCOS-R18	18.7	12.6	49.0	11.0	40.1	23.4	14.5	24.2
DDT(Ours)	/	FCOS-R18	35.7	26.2	63.2	24.3	53.5	35.7	27.0	37.9+ 13.7
Baseline	/	FCOS-R50	21.7	15.9	49.1	13.7	40.4	26.6	14.6	26.0
DDT(Ours)	/	FCOS-R50	38.0	32.0	64.0	25.9	55.9	36.8	29.3	40.3+ 14.3
Baseline	/	FCOS-R101	27.0	16.4	51.4	14.7	44.0	28.8	21.2	29.1
DDT(Ours)	/	FCOS-R101	37.9	36.1	64.5	30.8	56.9	38.7	31.8	42.4+ 13.3

Table 8: Quantitative results on adaptation from Sim10K to BDD100K (S→B) with FCOS. The bold indicates the best results.

Method	Reference	Detector	mAP(car)
SWDA [38]	<i>CVPR'19</i>	FRCNN-V16	42.9
CDN [40]	<i>ECCV'20</i>	FRCNN-V16	45.3
Baseline	/	FCOS-R18	36.5
DDT(Ours)			56.0+19.5
Baseline		FCOS-R50	38.7
DDT(Ours)			56.2+17.5
Baseline	/	FCOS-R101	36.5
DDT(Ours)			57.4+20.9

Table 9: Quantitative results on adaptation from Sim10K to Cityscapes (S→Cs) with FCOS. The bold indicates the best results.

Method	Reference	Detector	mAP(car)
DA-Faster [6]	<i>CVPR'18</i>	FRCNN-V16	39.0
SWDA [38]	<i>CVPR'19</i>	FRCNN-V16	40.7
HTCN [3]	<i>CVPR'20</i>	FRCNN-R101	42.5
UMT [9]	<i>CVPR'21</i>	FRCNN-R101	43.1
SSAL [37]	<i>NeurIPS'22</i>	FCOS-R50	51.8
O2NET [13]	<i>ACMMM'22</i>	DDETR-R50	54.1
DDF [32]	<i>TMM'22</i>	FRCNN-R50	44.3
D-ADAPT [20]	<i>ICLR'22</i>	FRCNN-R50	51.9
SCAN [27]	<i>AAAI'22</i>	FCOS-V16	52.6
MTTrans [48]	<i>ECCV'22</i>	DDETR-R50	57.9
SIGMA [28]	<i>CVPR'22</i>	FCOS-R50	53.7
TDD [29]	<i>CVPR'22</i>	FRCNN-V16	53.4
MGA [52]	<i>CVPR'22</i>	FCOS-R101	54.1
OADA [46]	<i>ECCV'22</i>	FCOS-V16	59.2
SIGMA++ [29]	<i>TPAMI'23</i>	FCOS-V16	53.7
CIGAR [34]	<i>CVPR'23</i>	FCOS-V16	58.5
NSA [53]	<i>ICCV'23</i>	FRCNN-V16	56.3
HT [10]	<i>CVPR'23</i>	FRCNN-V16	65.5
Baseline	/	FCOS-R18	47.0
DDT(Ours)			61.4+14.4
Baseline		FCOS-R50	48.4
DDT(Ours)			62.5+14.1
Baseline	/	FCOS-R101	51.5
DDT(Ours)			63.5+12.0

Table 10: Quantitative results on adaptation from VOC to Comic (V→Co) with FCOS. The bold indicates the best results.

Method	Reference	Detector	bicycle	bird	car	cat	dog	person	mAP
DA-Faster [6]	<i>CVPR'18</i>	FRCNN-V16	31.1	10.3	15.5	12.4	19.3	39.0	21.2
SWDA [38]	<i>CVPR'19</i>	FRCNN-V16	36.4	21.8	29.8	15.1	23.5	49.6	29.4
STABR [22]	<i>CVPR'19</i>	SSD-V16	50.6	13.6	31.0	7.5	16.4	41.4	26.8
MCRA [50]	<i>ECCV'20</i>	FRCNN-V16	47.9	20.5	37.4	20.6	24.5	50.2	33.5
I3Net [4]	<i>CVPR'21</i>	SSD-V16	47.5	19.9	33.2	11.4	19.4	49.1	30.1
DBGL [2]	<i>ICCV'21</i>	FRCNN-R101	35.6	20.3	33.9	16.4	26.6	45.3	29.7
D-ADAPT [20]	<i>ICLR'22</i>	FRCNN-R101	52.4	25.4	42.3	43.7	25.7	53.5	40.5
Baseline	/	FCOS-R18	17.7	7.1	8.3	2.4	5.7	25.3	11.1
DDT(Ours)			54.2	26.6	45.5	29.8	34.4	72.2	43.8+32.7
Baseline	/	FCOS-R50	20.9	7.2	11.3	4.7	7.9	27.0	13.2
DDT(Ours)			55.1	32.2	51.4	33.8	38.7	74.6	47.6+34.4
Baseline	/	FCOS-R101	26.0	9.5	15.4	7.3	8.0	29.3	15.9
DDT(Ours)			55.6	38.2	55.6	36.6	48.1	75.9	51.6+35.7

Table 11: Quantitative results on adaptation from VOC to Watercolor (V→W) with FCOS. The bold indicates the best results.

Method	Reference	Detector	bicycle	bird	car	cat	dog	person	mAP
SWDA [6]	<i>CVPR'19</i>	FRCNN-V16	82.3	55.9	46.5	32.7	35.5	66.7	53.3
MCRA [51]	<i>ECCV'20</i>	FRCNN-V16	87.9	52.1	51.8	41.6	33.8	68.8	56.0
UMT [9]	<i>CVPR'21</i>	FRCNN-R101	88.2	55.3	51.7	39.8	43.6	69.9	58.1
I10D [43]	<i>TPAMI'21</i>	FRCNN-V16	95.8	54.3	48.3	42.4	35.1	65.8	56.9
I3Net [4]	<i>CVPR'21</i>	SSD-V16	81.1	49.3	46.2	35.0	31.9	65.7	51.5
SADA [7]	<i>IJCV'21</i>	FRCNN-R50	82.9	54.6	52.3	40.5	37.7	68.2	56.0
CDG [25]	<i>AAAI'21</i>	FRCNN-V16	97.7	53.1	52.1	47.3	38.7	68.9	59.7
VDD [44]	<i>ICCV'21</i>	FRCNN-V16	90.0	56.6	49.2	39.5	38.8	65.3	56.6
DBGL [2]	<i>ICCV'21</i>	FRCNN-R101	83.1	49.3	50.6	39.8	38.7	61.3	53.8
AT [31]	<i>CVPR'22</i>	FRCNN-V16	93.6	56.1	58.9	37.3	39.6	73.8	59.9
LODS [26]	<i>CVPR'22</i>	FRCNN-R101	95.2	53.1	46.9	37.2	47.6	69.3	58.2
Baseline	/	FCOS-R18	69.8	34.9	37.8	23.3	16.0	48.7	38.4
DDT(Ours)			80.3	60.1	52.5	42.4	34.3	75.8	57.6+19.2
Baseline	/	FCOS-R50	66.9	42.4	44.6	21.5	13.7	48.3	39.6
DDT(Ours)			94.4	63.1	51.8	40.8	34.3	75.9	60.1+20.5
Baseline	/	FCOS-R101	64.0	44.3	41.8	25.7	21.5	53.6	41.8
DDT(Ours)			96.9	65.6	55.4	49.6	40.5	77.2	64.2+22.4

Table 12: Quantitative results on adaptation from VOC to Clipart ($V \rightarrow Ca$) with FCOS. The bold indicates the best results.

Method	Reference	Detector	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	psn	plant	sheep	sofa	train	tv	mAP
SWDA [38]	<i>CVPR'19</i>	FRCNN-V16	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	17.0	12.5	33.8	65.5	61.6	52.0	9.3	24.9	54.1	49.1	38.1
CRDA [45]	<i>CVPR'20</i>	FRCNN-R101	28.7	55.3	31.8	26.0	40.1	63.6	36.6	9.4	38.7	49.3	17.6	14.1	33.3	74.3	61.3	46.3	22.3	24.3	49.1	44.3	38.3
HTCN [3]	<i>CVPR'20</i>	FRCNN-R101	33.6	58.9	34.0	23.4	45.6	57.0	39.8	12.0	39.7	51.3	21.1	20.1	39.1	72.8	63.0	43.1	19.3	30.1	50.2	51.8	40.3
SAPNet [23]	<i>ECCV'20</i>	FRCNN-R101	27.4	70.8	32.0	27.9	42.4	63.5	47.5	14.3	48.2	46.1	31.8	17.9	43.8	68.0	68.1	49.0	18.7	20.4	55.8	51.3	42.2
UMT [9]	<i>CVPR'21</i>	FRCNN-R101	39.6	59.1	32.4	35.0	45.1	61.9	48.4	7.5	46.0	67.6	21.4	29.5	48.2	75.9	70.5	56.7	25.9	28.9	39.4	43.6	44.1
IIOD [43]	<i>TPAMI'21</i>	FRCNN-V16	41.5	52.7	34.5	28.1	43.7	58.5	41.8	15.3	40.1	54.4	26.7	28.5	37.7	75.4	63.7	48.7	16.5	30.8	54.5	48.7	42.1
SADA [7]	<i>IJCV'21</i>	FRCNN-R50	29.4	56.8	30.6	34.0	49.5	50.5	47.7	18.7	48.5	64.4	20.3	29.0	42.3	84.1	73.4	37.4	20.5	39.8	41.2	48.0	43.3
UaDAN [14]	<i>TMM'21</i>	FRCNN-R50	35.0	72.7	41.0	24.4	21.3	69.8	53.5	2.3	34.2	61.2	31.0	29.5	47.9	63.6	62.2	61.3	13.9	7.6	48.6	23.9	40.2
DBGL [2]	<i>ICCV'21</i>	FRCNN-R50	28.5	52.3	34.3	32.8	38.6	66.4	38.2	25.3	39.9	47.4	23.9	17.9	38.9	78.3	61.2	51.7	26.2	28.9	56.8	44.5	41.6
AT [31]	<i>CVPR'22</i>	FRCNN-V16	33.8	60.9	38.6	49.4	52.4	53.9	56.7	7.5	52.8	63.5	34.0	25.0	62.2	72.1	77.2	57.7	27.2	52.0	55.7	54.1	49.3
D-ADAPT [20]	<i>ICLR'22</i>	FRCNN-R50	56.4	63.2	42.3	40.9	45.3	77.0	48.7	25.4	44.3	58.4	31.4	24.5	47.1	75.3	69.3	43.5	27.9	34.1	60.7	64.0	49.0
TIA [49]	<i>CVPR'22</i>	FRCNN-R101	42.2	66.0	36.9	37.3	43.7	71.8	49.7	18.2	44.9	58.9	18.2	29.1	40.7	87.8	67.4	49.7	27.4	27.8	57.1	50.6	46.3
LODS [26]	<i>CVPR'22</i>	FRCNN-R101	43.1	61.4	40.1	36.8	48.2	45.8	48.3	20.4	44.8	53.3	32.5	26.1	40.6	86.3	68.5	48.9	25.4	33.2	44.0	56.5	45.2
CIGAR [34]	<i>CVPR'23</i>	FCOS-R101	35.2	55.0	39.2	30.7	60.1	58.1	46.9	31.8	47.0	61.0	21.8	26.7	44.6	52.4	68.5	54.4	31.3	38.8	56.5	63.5	46.2
CMT [1]	<i>CVPR'23</i>	FRCNN-V16	39.8	56.3	38.7	39.7	60.4	35.0	56.0	7.1	60.1	60.4	35.8	28.1	67.8	84.5	80.1	55.5	20.3	32.8	42.3	38.2	47.0
Baseline	/	FCOS-R18	18.7	26.0	15.0	10.1	19.5	65.6	30.6	1.8	24.3	4.2	24.1	7.9	24.9	42.1	33.5	26.1	0.2	17.2	23.0	11.2	21.3
DDT(Ours)			48.9	58.9	32.3	30.0	42.4	72.4	54.5	11.2	48.6	38.9	30.2	27.5	40.1	87.7	76.0	53.2	33.5	38.8	49.5	47.1	46.1+ 24.8
Baseline	/	FCOS-R50	40.0	26.7	17.8	21.0	31.9	32.2	28.8	12.2	36.3	35.7	28.3	6.1	25.5	43.1	37.2	33.5	5.1	25.6	24.5	26.3	26.9
DDT(Ours)			50.7	53.1	34.1	41.5	57.0	86.3	57.1	9.3	49.5	52.8	33.6	32.4	49.0	93.1	82.1	57.8	37.1	42.6	54.1	60.8	51.7+ 24.8
Baseline	/	FCOS-R101	33.6	42.3	21.2	20.1	32.9	62.0	30.0	14.5	41.1	17.9	33.0	9.1	30.4	46.5	39.1	37.4	8.8	22.6	27.3	16.1	29.3
DDT(Ours)			58.6	73.2	42.0	48.0	54.7	84.7	65.2	17.0	55.9	49.4	35.5	40.5	58.6	84.8	82.9	58.0	39.1	41.7	54.7	61.3	55.3+ 26.0

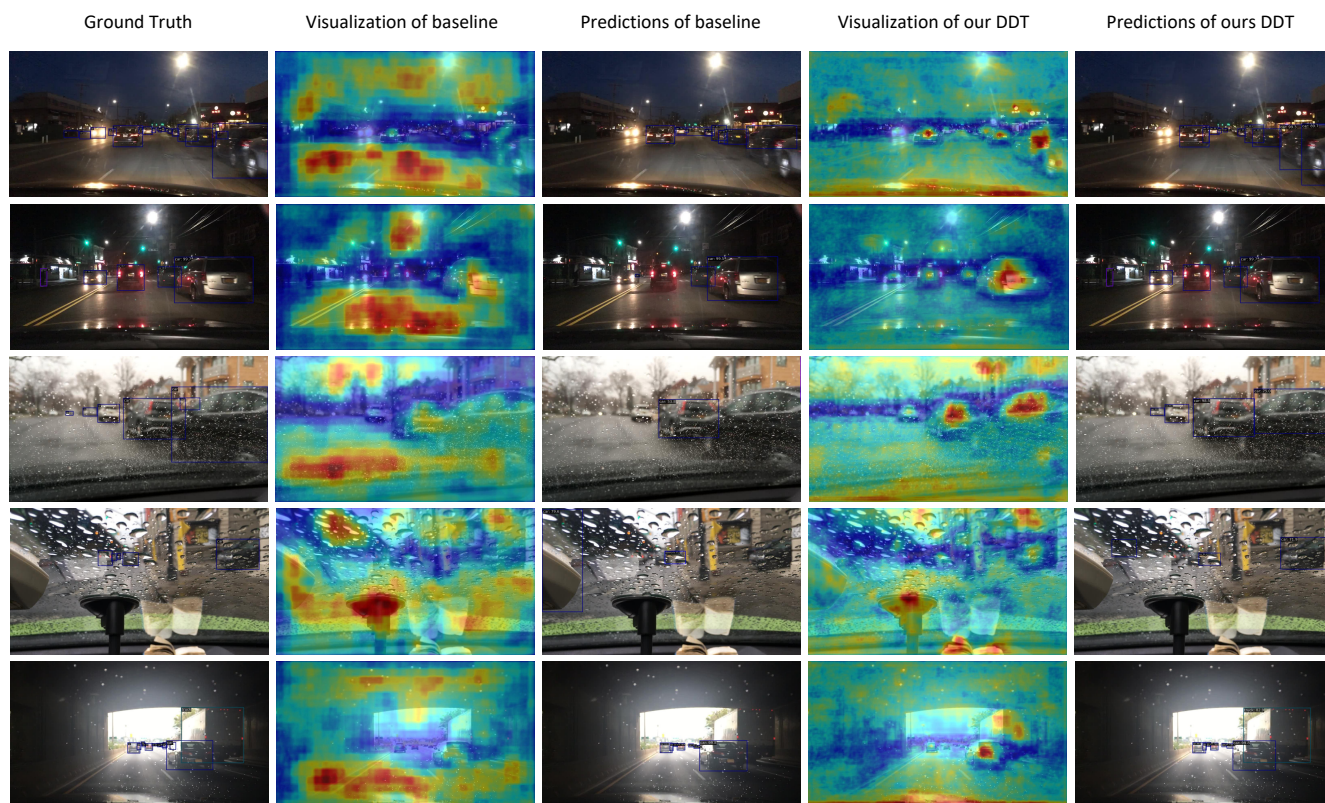


Figure 2: Qualitative prediction results and feature visualization of baseline and our DDT from Cityscapes to BDD100K.

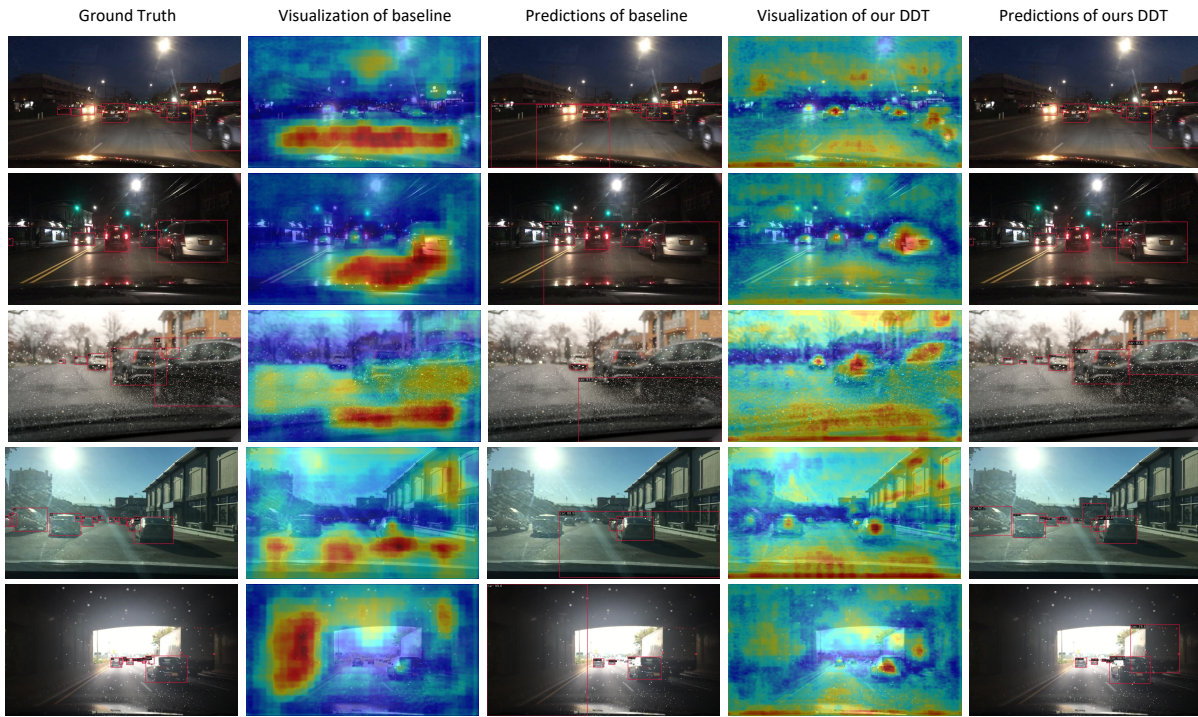


Figure 3: Qualitative prediction results and feature visualization of baseline and our DDT from Sim10K to BDD100K.

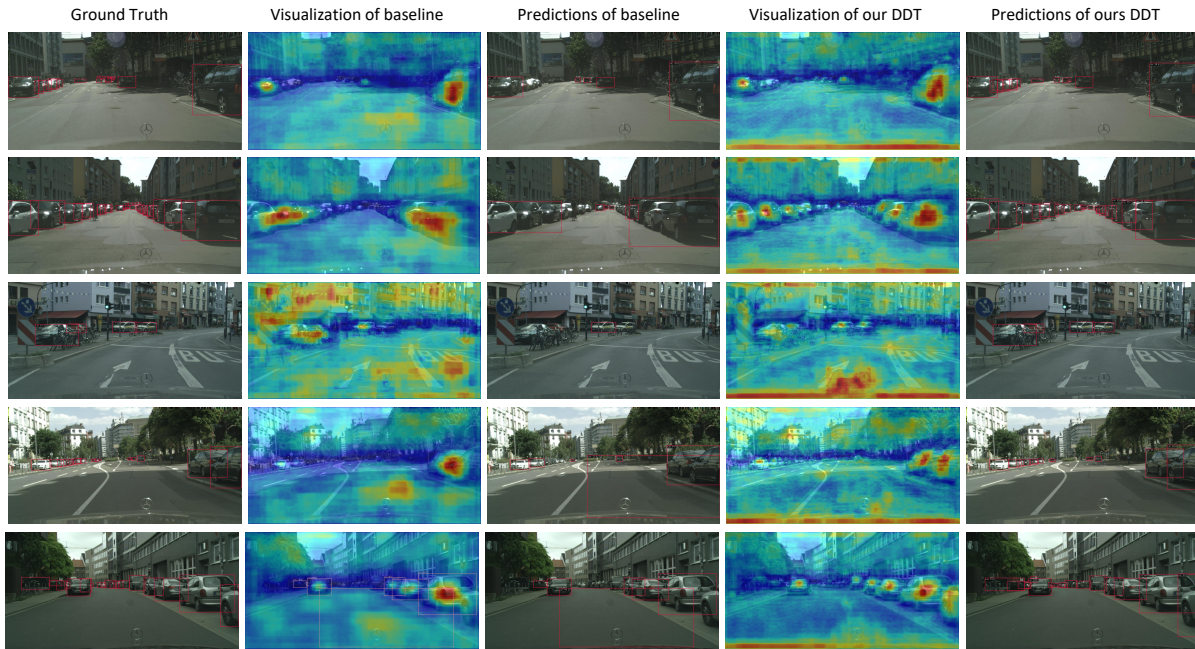


Figure 4: Qualitative prediction results and feature visualization of baseline and our DDT from Sim10K to Cityscapes.

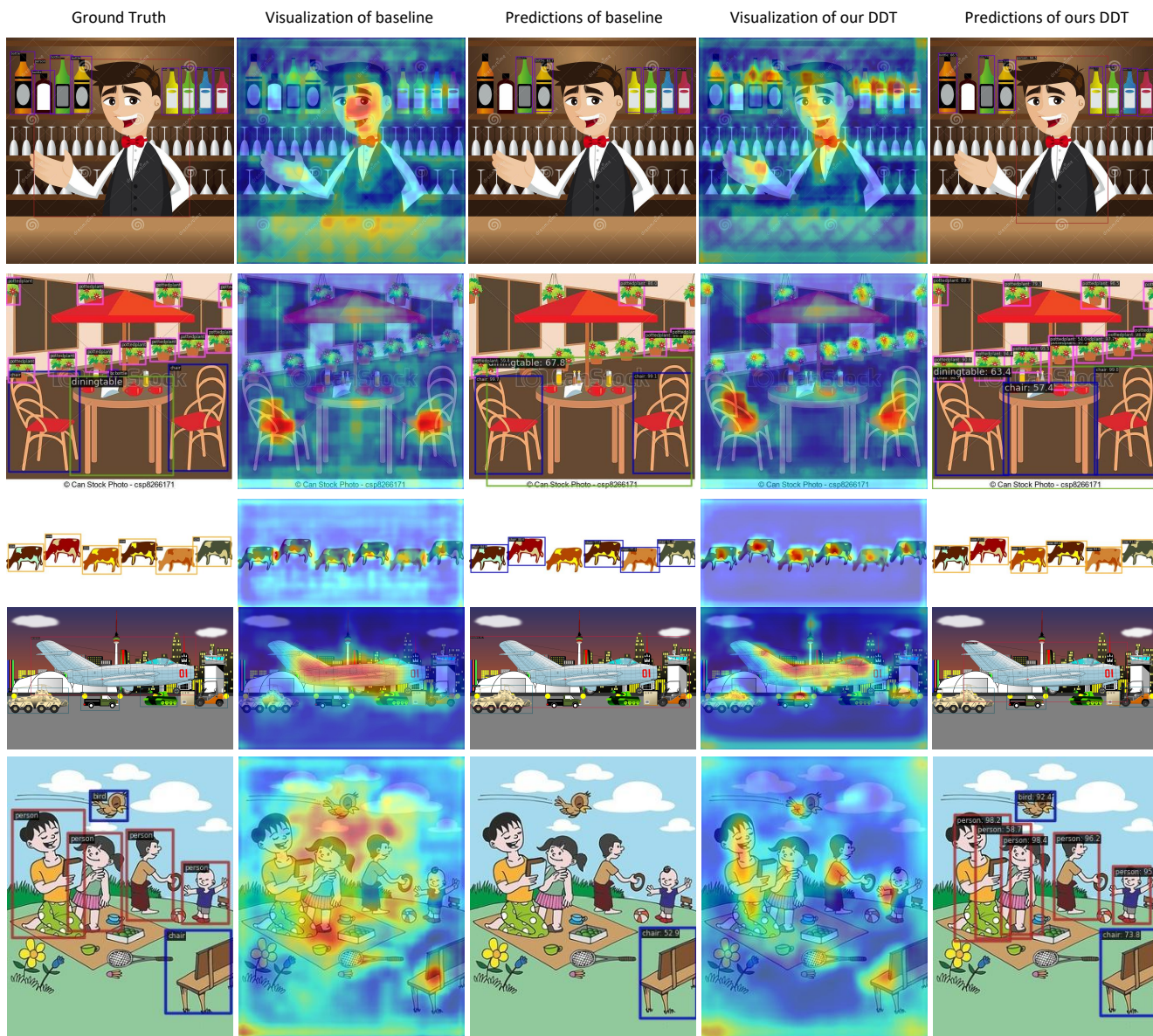


Figure 5: Qualitative prediction results and feature visualization of baseline and our DDT from VOC to Clipart.

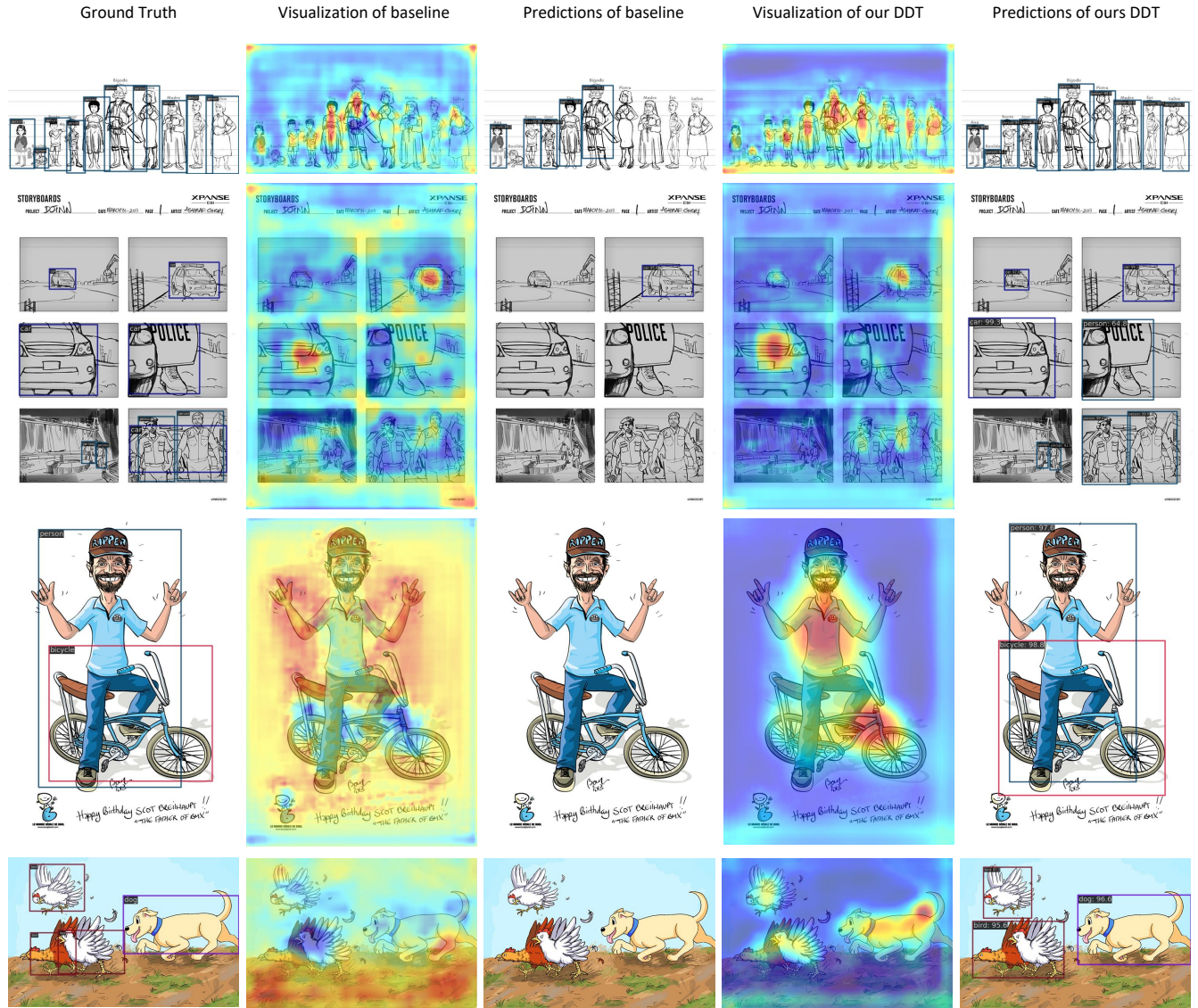


Figure 6: Qualitative prediction results and feature visualization of baseline and our DDT from VOC to Comic.

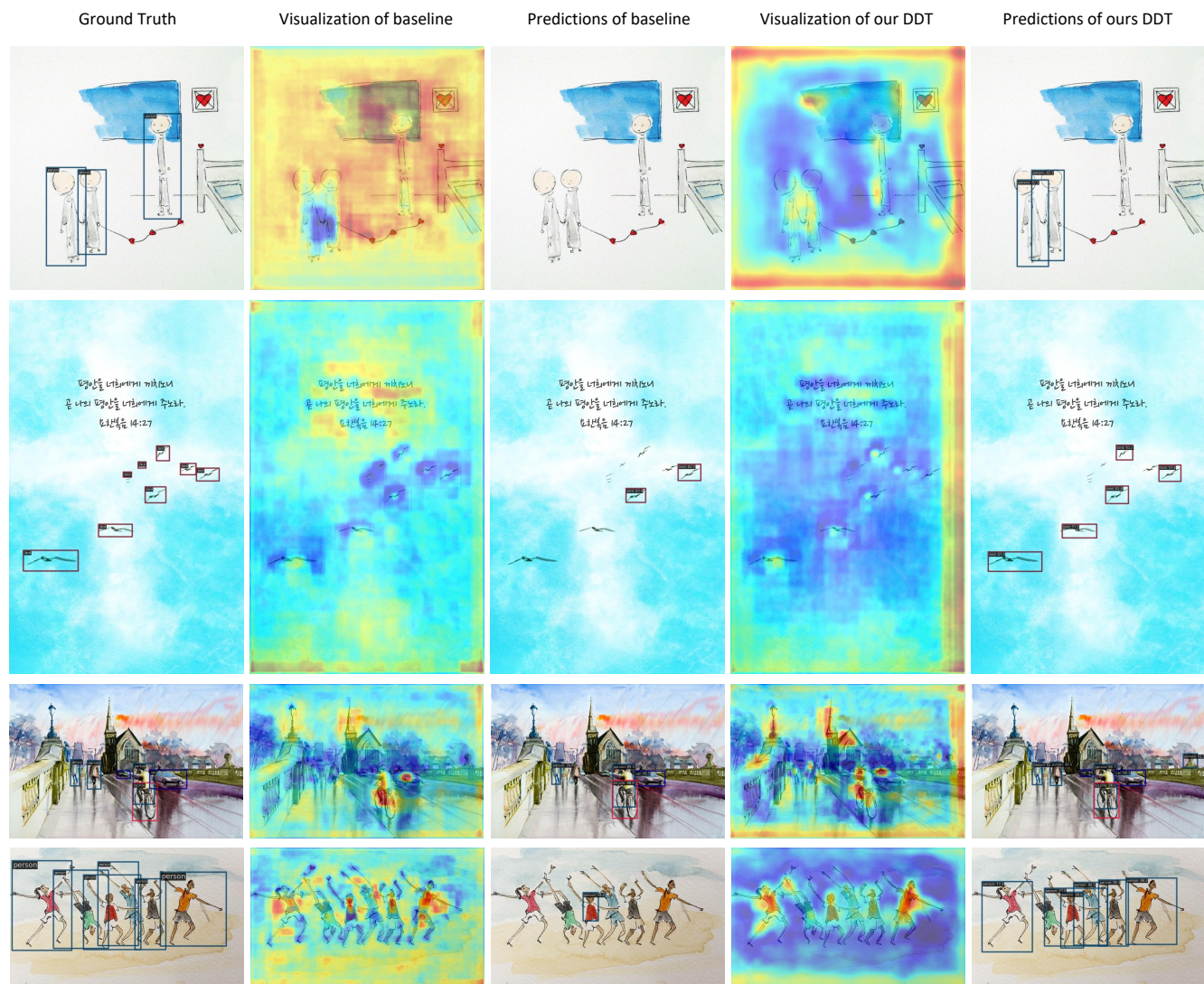


Figure 7: Qualitative prediction results and feature visualization of baseline and our DDT from VOC to Watercolor.

REFERENCES

- [1] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. 2023. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23839–23848.
- [2] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. 2021. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2703–2712.
- [3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. 2020. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8869–8878.
- [4] Chaoqi Chen, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. 2021. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12576–12585.
- [5] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. 2022. Learning Domain Adaptive Object Detection with Probabilistic Teacher. In *International Conference on Machine Learning*. PMLR, 3040–3055.
- [6] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3339–3348.
- [7] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2021. Scale-aware domain adaptive faster r-cnn. *International Journal of Computer Vision* 129, 7 (2021), 2223–2243.
- [8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.
- [9] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. 2021. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4091–4101.
- [10] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. 2023. Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23829–23838.
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiuhua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88 (2010), 303–338.
- [13] Kaixiong Gong, Shuang Li, Shugang Li, Rui Zhang, Chi Harold Liu, and Qiang Chen. 2022. Improving transferability for domain adaptive detection transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1543–1551.
- [14] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. 2021. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia* 24 (2021), 2502–2514.
- [15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.
- [16] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. 2022. Cross domain object detection by target-perceived dual branch distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9570–9580.
- [17] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. 2023. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11721–11732.
- [18] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. 2020. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*. 16. Springer, 733–748.
- [19] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5001–5009.
- [20] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. 2021. Decoupled Adaptation for Cross-Domain Object Detection. In *International Conference on Learning Representations*.
- [21] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. 2017. Driving in the Matrix: Can virtual worlds replace human-generated annotations for real world tasks?. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE.
- [22] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. 2019. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6092–6101.
- [23] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. 2020. Spatial Attention Pyramid Network for Unsupervised Domain Adaptation. In *European Conference on Computer Vision*. 481–497.
- [24] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiyu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10965–10975.
- [25] Shuai Li, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. 2021. Category dictionary guided unsupervised domain adaptation for object detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 1949–1957.
- [26] Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. 2022. Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8014–8023.
- [27] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. 2022. Scan: Cross domain object detection with semantic conditioned adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1421–1428.
- [28] Wuyang Li, Xinyu Liu, and Yixuan Yuan. 2022. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5291–5300.
- [29] Wuyang Li, Xinyu Liu, and Yixuan Yuan. 2023. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [30] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. 2021. A free lunch for unsupervised domain adaptive object detection without source data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 8474–8481.
- [31] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. 2022. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7581–7590.
- [32] Dongnan Liu, Chaoyi Zhang, Yang Song, Heng Huang, Chenyu Wang, Michael Barnett, and Weidong Cai. 2022. Decompose to adapt: Cross-domain object detection via feature disentanglement. *IEEE Transactions on Multimedia* 25 (2022), 1333–1344.
- [33] Xinyu Liu, Wuyang Li, Qiushi Yang, Baopu Li, and Yixuan Yuan. 2022. Towards robust adaptive object detection under noisy annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14207–14216.
- [34] Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, and Yong Xu. 2023. CIGAR: Cross-Modality Graph Reasoning for Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23776–23786.
- [35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.
- [36] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11976–11986.
- [37] Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz, and Mohsen Ali. 2021. Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection. *Advances in Neural Information Processing Systems* 34 (2021), 22770–22782.
- [38] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. 2019. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6956–6965.
- [39] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision* 126 (2018), 973–992.
- [40] Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. 2020. Adapting object detectors with conditional domain normalization. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*. 16. Springer, 403–419.
- [41] Renshuai Tao, Hainan Li, Tianbo Wang, Yanlu Wei, Yifu Ding, Bowei Jin, Hongping Zhi, Xianglong Liu, and Aishan Liu. 2022. Exploring endogenous shift for cross-domain detection: A large-scale benchmark and perturbation suppression network. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 21157–21167.
- [42] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2020. FCOS: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 1922–1933.
- [43] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. 2021. Instance-invariant domain adaptive object detection via progressive disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2021), 4178–4193.
- [44] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. 2021. Vector-decomposed disentanglement for domain-invariant object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9342–9351.

- [45] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. 2020. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11724–11733.
- [46] Jayeon Yoo, Inseop Chung, and Nojun Kwak. 2022. Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In *European Conference on Computer Vision*. Springer, 691–708.
- [47] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2636–2645.
- [48] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. 2022. MT-Trans: Cross-domain object detection with mean teacher transformer. In *European Conference on Computer Vision*. Springer, 629–645.
- [49] Liang Zhao and Limin Wang. 2022. Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14217–14226.
- [50] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. 2020. Adaptive object detection with dual multi-label prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII* 16. Springer, 54–69.
- [51] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. 2020. Adaptive object detection with dual multi-label prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII* 16. Springer, 54–69.
- [52] Wenzhang Zhou, Dawei Du, Libo Zhang, Tiejian Luo, and Yanjun Wu. 2022. Multi-granularity alignment domain adaptation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9581–9590.
- [53] Wenzhang Zhou, Heng Fan, Tiejian Luo, and Libo Zhang. 2023. Unsupervised Domain Adaptive Detection with Network Stability Analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6986–6995.
- [54] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. 2019. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 687–696.