Diversify and Match: A Domain Adaptive Representation Learning Paradigm for Object Detection (Supplementary Material)

Taekyung Kim Minki Jeong Seunghyeon Kim Seokeon Choi Changick Kim
Korea Advanced Institute of Science and Technology, Daejeon, Korea
{tkkim93, rhm033, seunghyeonkim, seokeon, changick}@kaist.ac.kr

1. Appendix

1.1. Universality of Our Framework for Object Detection Networks

While most existing object detectors are structured by single-staged or two-staged architecture, we have been verified our framework only on standard two-staged architecture. To validate the universality of our framework for object detection network, we additionally evaluated our method on a single-staged architecture. Since SSD [6] is one of the standard networks in single-staged object detectors, we conducted the experiments on SSD. We conducted the experiments for the Real-world→Artistic Media Datasets (AMDs) [5] as study cases.

The comparison results are reported in Table 1, 2, and 3. Similar to the results on Faster R-CNN [7], our methods achieved a significant gain of approximately 7 % ∼ 17% compared to the SSD baseline. Furthermore, our method also outperformed other domain adaptation methods regardless of the difference in the object detector. These results prove the universal property of our framework. For Real-world→ Clipart1k case, unlike the Faster R-CNN baseline, our method showed comparable but not the highest class-wise performance for some classes in the animal category. Especially for the sheep class, our method had lower AP than the baseline. However, since these classes showed further lower performance on the Faster R-CNN backbone, these comparable performances can be seen as an advantage of the SSD architecture.

1.2. Performance Comparison with Fully Supervised Models

To check the potential of our method, we compared with the fully supervised models on various adaptation cases. The fully supervised models are implemented by finetuning the Faster R-CNN [7] baseline with the train set in the target domain. The model was fine-tuned for one epoch for the Real-world→Artistic Media Datasets (AMDs) cases and 30k iterations for the urban scene adaptation case.

As shown in Table 4, our method achieved comparable or higher performance compared to the fully supervised models. Especially for the Real-world→Watercolor2k [5], These results can be interpreted as that even though the feature extractor of the models is enriched by abundant source domain data, supervision by insufficient annotations or supervision without domain adaptation methods cannot encourage the model enough to infer discriminatively in the target domain distribution. However, our method sufficiently exploited the rich source domain data and achieved impressive performance without annotations of the target domain, which verifies the potential of our method.

1.3. Universality of Domain Diversification for Architectures of the Domain Shifters

To verify the universality of Domain Diversification (DD) for the domain shift architecture, we conducted additional experiments on the domain shifter from CartoonGAN [1]. Unlike the configurations of the constraint factors in the manuscript, we trained the domain shifters with no constraint, color preservation constraint, and reconstruction constraint factor each. We present qualitative results of the shifted domains in Sec 1.4.

Table 5 shows the result of the ablation study on numbers of shifted domains. Similar to the results on the manuscript, the overall results of each method were improved as the number of shifted domains increases. The performance gain by MRL also amplified as the number of shifted domains increases. However, DD with two domains does not significantly improve compared to each result with the single domain. It seems like the shifted domain with no constraint and with color preservation constraint does not diversified enough so that the efficacy of DD and MRL were degraded than expected. However, the DD with three domains significantly enhances the performance compare to the results with two domains. These results validate the effectiveness of the DD and MRL on newly generated three domains. Thus, we show the universality of the DD for domain shifter architecture.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Baseline [5]	19.8	49.5	20.1	23.0	11.3	38.6	34.2	2.5	39.1	21.6	27.3	10.8	32.5	54.1	45.3	31.2	19.0	19.5	19.1	17.9	26.8
DT [5]	23.3	60.1	24.9	**41.5**	**26.4**	53.0	44.0	**4.1**	45.3	51.5	**39.5**	11.6	**40.4**	62.2	61.1	37.1	**20.9**	39.6	38.4	36.0	38.0
Ours (n=3)	**39.5**	**64.7**	**29.8**	35.6	25.7	**56.3**	**49.7**	3.9	**53.3**	**57.6**	35.4	**35.8**	38.3	**70.3**	**64.6**	**42.7**	17.1	**45.9**	**52.9**	**66.5**	**44.3**

Table 1. Quantitative results for object detection of Clipart1k [5] by adapting from PASCAL VOC [4] on SSD [6].

Method	bike	bird	car	cat	dog	person	mAP
Baseline [5]	79.8	49.5	38.1	35.1	30.4	65.1	49.6
DT [5]	82.8	47.0	40.2	34.6	35.3	62.5	50.4
Ours (n=3)	**86.3**	**51.8**	**45.1**	**42.6**	**42.4**	**70.7**	**56.5**

Table 2. Quantitative results for object detection of Watercolor2k [5] by adapting from PASCAL VOC [4] on SSD [6].

Method	bike	bird	car	cat	dog	person	mAP
Baseline [5]	43.9	10.0	19.4	12.9	20.3	42.6	24.9
DT [5]	43.6	13.6	30.2	16.0	26.9	48.3	29.8
Ours (n=3)	**51.9**	**21.8**	**39.7**	**26.6**	**37.1**	**61.3**	**41.2**

Table 3. Quantitative results for object detection of Comic2k [5] by adapting from PASCAL VOC [4] on SSD [6].

1.4. Additional Qualitative Comparisons

Figure 1, 2, and 3 show additional qualitative results for the object detection of Clipart1k, Watercolor2k, and Comic2k [5] by adapting from PASCAL VOC [4], respectively. Figure 4, 5, 6, and 7 show additional shifted results in the shifted domains with various configurations of constraint factors (i.e., color preservation, reconstruction, both of them, or none of them). Specifically, Fig. 4 shows shifted images generated for Real-world→Clipart1k [5] through the domain shifter architecture of CartoonGAN [1].

References

[1] Y. Chen, Y.-K. Lai, and Y.-J. Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool. Domain adaptive faster R-CNN for object detection in the wild. *CoRR*, abs/1803.03243, 2018.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016.

[4] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010.

[5] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Method	V→ Cl	V→ Wa	V→ Co	C→ FC
Baseline	24.9	39.8	21.4	17.9
Ours (DD+MRL)	41.8	52.0	34.5	34.6
Oracle	50.0	48.8	31.6	36.0

Table 4. Comparison results with the fully supervised models on various adaptation cases. Backbone network is Faster R-CNN [7]. V, Cl, Wa, Co, C, FC denote PASCAL VOC [4], Clipart1k [5], Watercolor2k [5], Comic2k [5], Cityscapes [3], and Foggy Cityscapes [8]. Oracle denotes full supervision with the train set of the target domain.

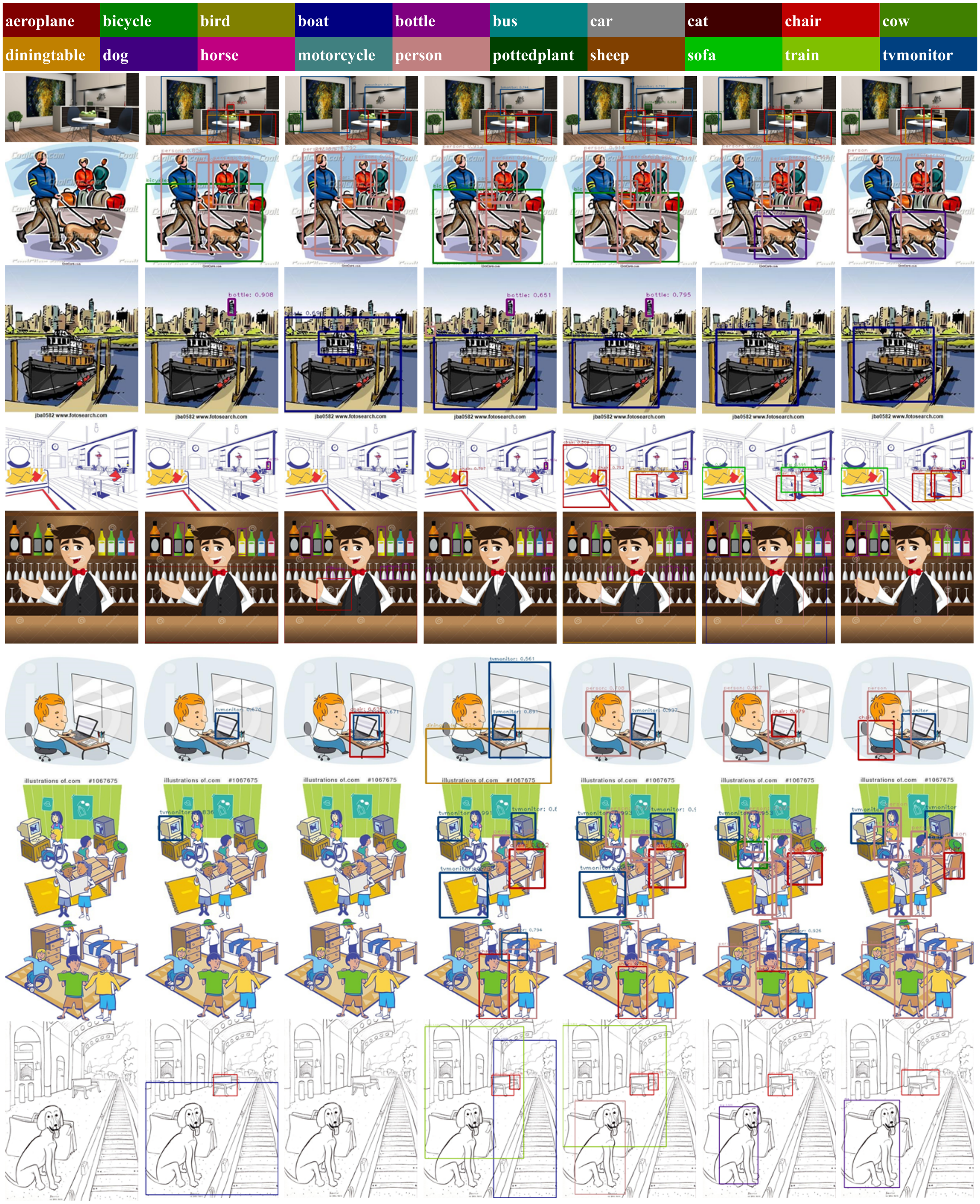
DD Configuration				DD	DD+MRL	offset
#SD	None	CP	R	mAP		
0				24.9	-	-
1	✓			29.7	33.8	+4.1
2	✓	✓		31.9	36.1	+4.2
3	✓	✓	✓	33.7	40.1	+6.4

Table 5. Results of the ablation study on the configuration of the shifted domains. DD, DD+MRL denote Domain Diversification and Multi-domain-invariant Representation Learning, respectively. The offset denotes the performance improvement through MRL. None, CP, and R denote the shifted domains trained with no constraint, color preservation constraint, and reconstruction constraint, respectively.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

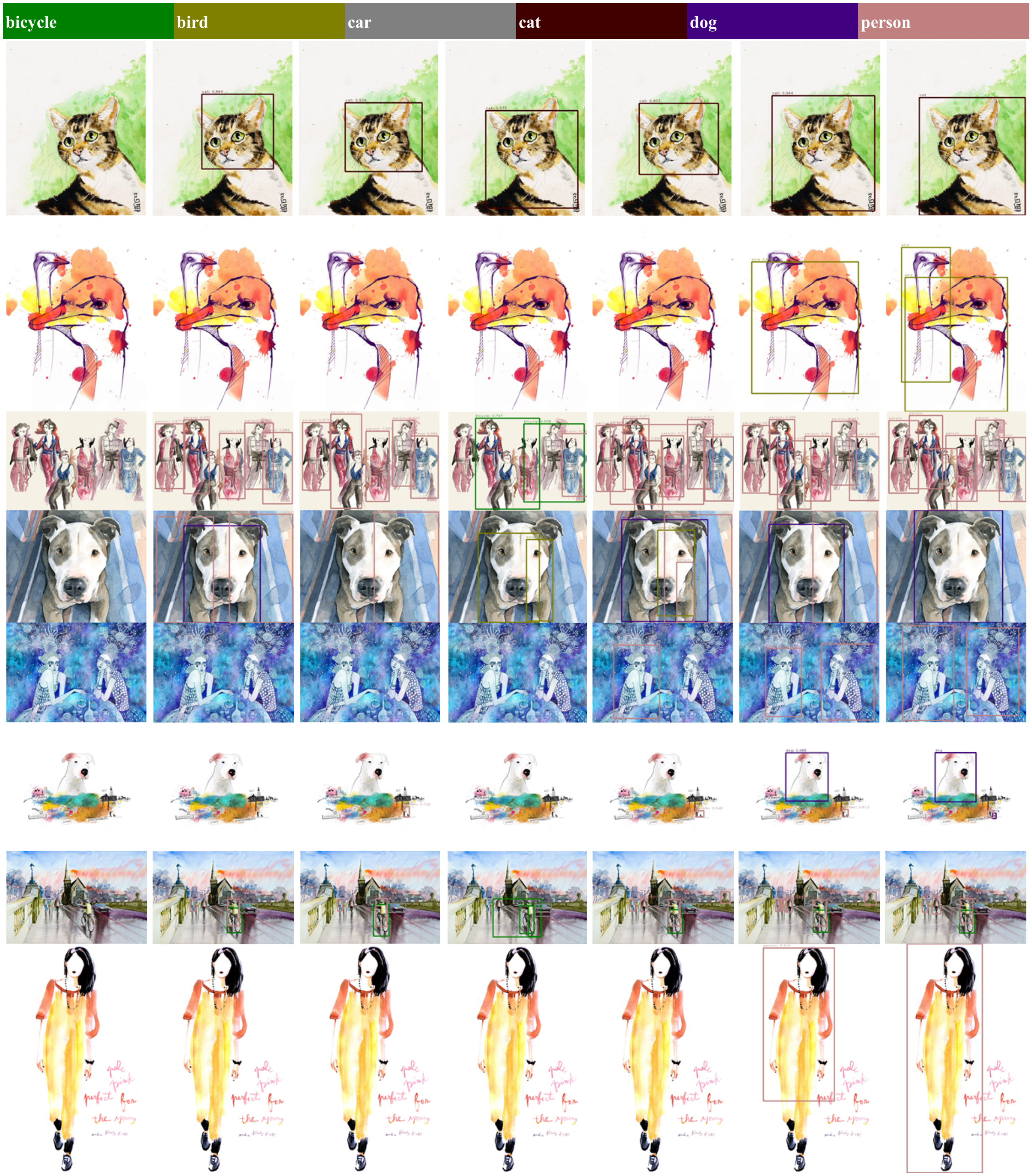
[7] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 91–99, Cambridge, MA, USA, 2015. MIT Press.

[8] C. Sakaridis, D. Dai, and L. V. Gool. Semantic foggy scene understanding with synthetic data. *CoRR*, abs/1708.07819, 2017.



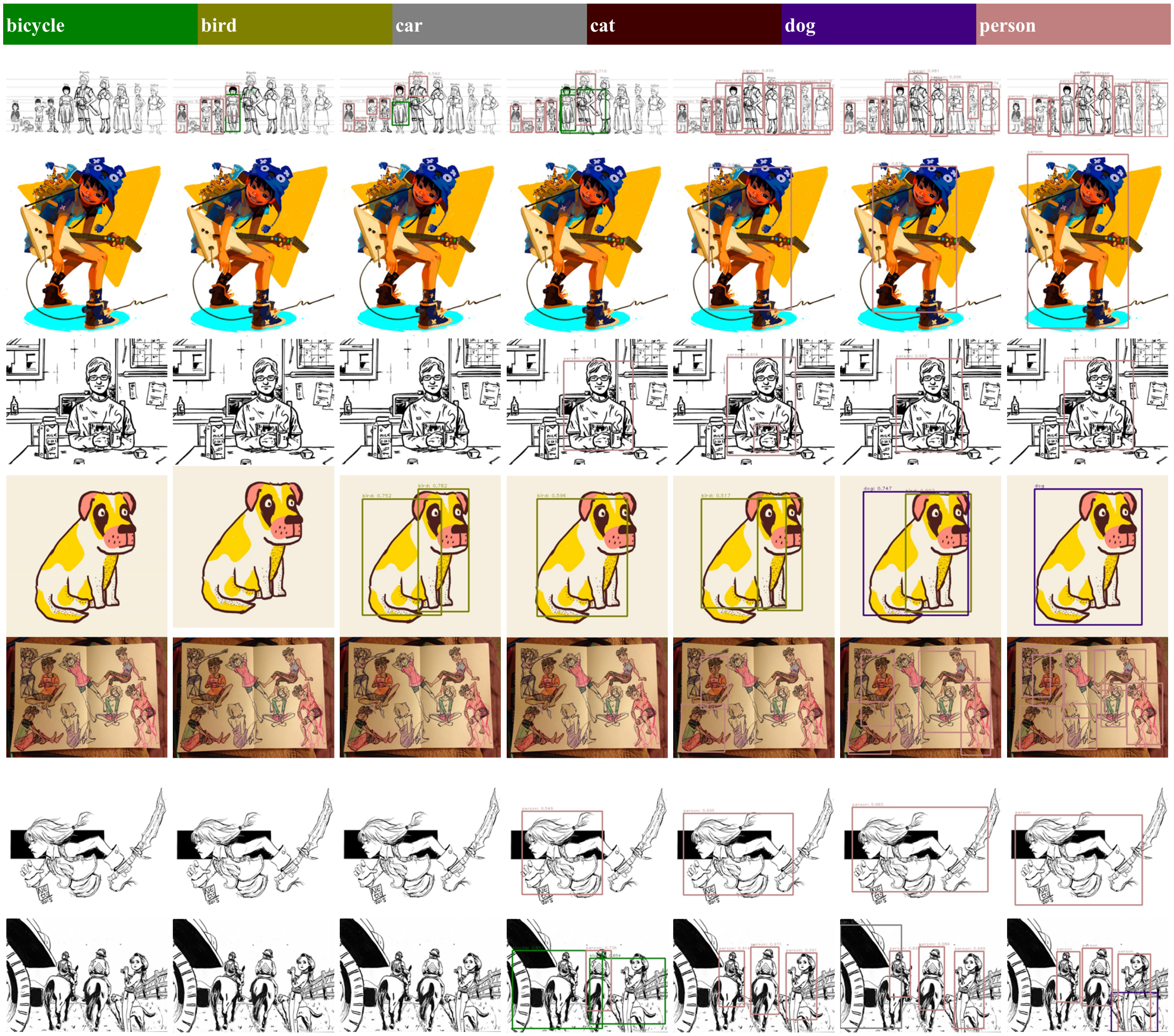
(a) Input image (b) Baseline (c) DAF (Img) [2] (d) DT [5] (e) Ours (DD) (f) Ours (DD+MRL) (g) Ground-truth

Figure 1. Qualitative results for object detection of Clipart1k [5] by adapting from PASCAL VOC [4].



(a) Input image (b) Baseline (c) DAF (Img) [2] (d) DT [5] (e) Ours (DD) (f) Ours (DD+MRL) (g) Ground-truth

Figure 2. Qualitative results for object detection of Watercolor2k [5] by adapting from PASCAL VOC [4].



(a) Input image (b) Baseline (c) DAF (Img) [2] (d) DT [5] (e) Ours (DD) (f) Ours (DD+MRL) (g) Ground-truth

Figure 3. Qualitative results for object detection of Comic2k [5] by adapting from PASCAL VOC [4].

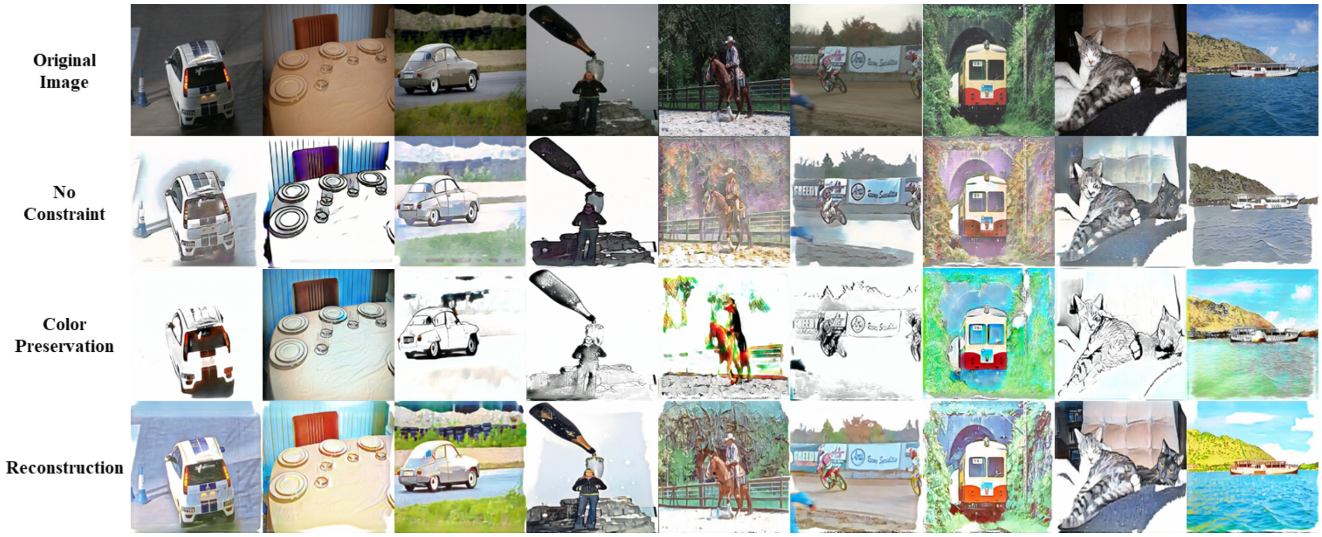


Figure 4. Shifted samples in the shifted domains of the Clipart1k [5] by adapting from PASCAL VOC [4]. The domain shifter architecture is from CartoonGAN [1].

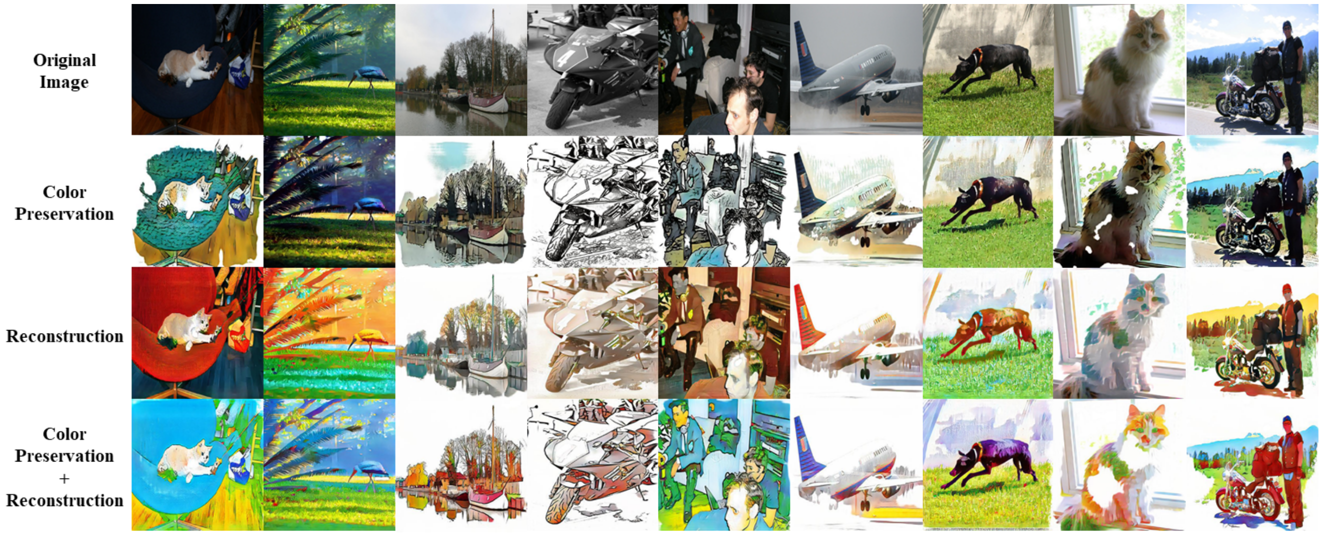


Figure 5. Shifted samples in the shifted domains of Clipart1k [5] by adapting from PASCAL VOC [4].

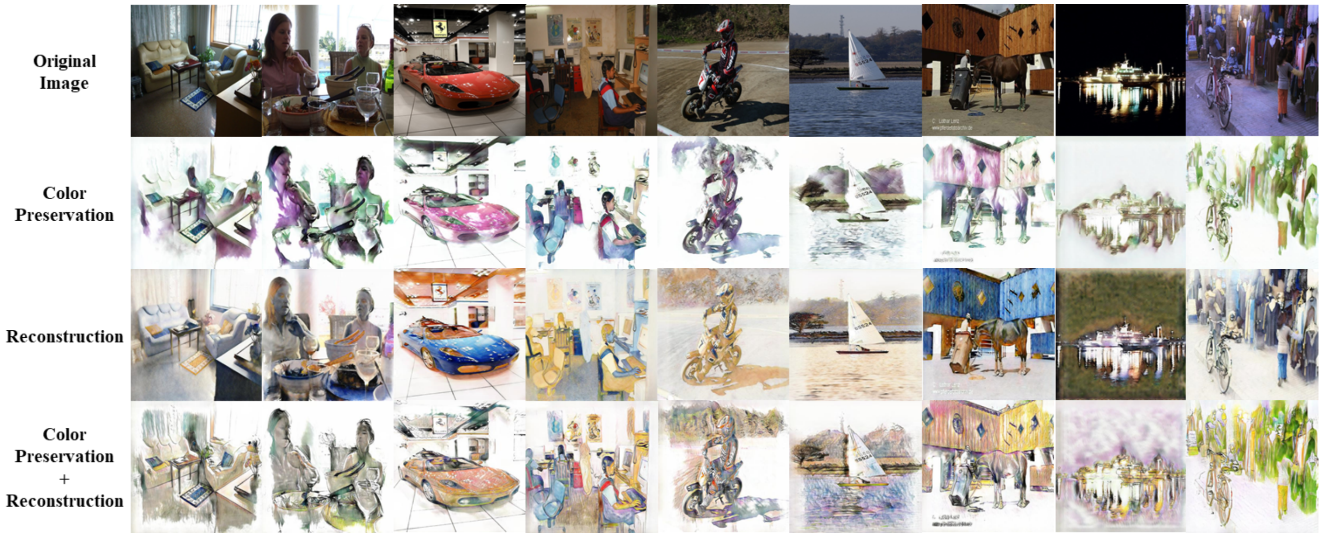


Figure 6. Shifted samples in the shifted domains of Watercolor2k [5] by adapting from PASCAL VOC [4].

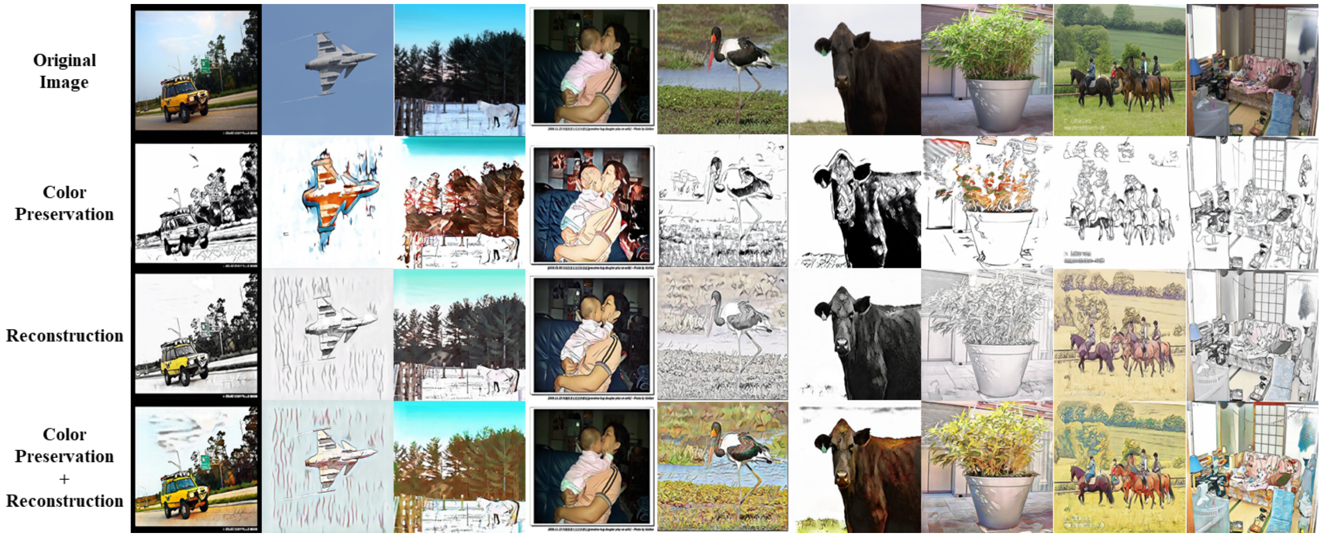


Figure 7. Shifted samples in the shifted domains of Comic2k [5] by adapting from PASCAL VOC [4].