Supplementary Material for Exploring Resolution and Degradation Clues as Self-supervised Signal for Low Quality Object Detection

Ziteng Cui[1], Yingying Zhu[2], Lin Gu[3,4⋆], Guo-Jun Qi[5], Xiaoxiao Li[6], Renrui Zhang[7], Zenghui Zhang[1], and Tatsuya Harada[4,3]

[1] Shanghai Jiao Tong University
[2] University of Texas at Arlington
[3] RIKEN AIP
[4] The University of Tokyo
[5] Laboratory for Machine Perception and Learning
[6] The University of British Columbia
[7] Shanghai AI Laboratory

A Results on FPN structure.

We evaluate AERIS's effectiveness on CenterNet [8] in the main text, where the CenterNet structure only take one level feature map for detection decoder D_o, also single-level feature structure could better illustrate how our self-supervised signal directly enhances the performance. Further more, more recent object detectors take FPN [5] and FPN series structures for detection decoding, take advantage of the multi-branch design, different level feature maps would outputs multi object detection results.

We evaluate our method on RetinaNet-FPN with Swin-T backbone for the multi-level scenario, name as AERIS-FPN. The experiments are done on the **COCO-d** dataset, we trained the model for 24 epochs with SGD optimizer, the initial learning rate is 0.01 and decay to one-tenth at 16 and 22 epoch, data augmentation is same as `mmdetection` [2], other setting is same as Table 1 in main text. For evaluating different places of FPN to add the ARRD decoder D_r, Different level output features of FPN (1, 2, 3, 4, 5) are used for the reconstruction (1 is highest resolution feature map, 2 is second high resolution feature map and so on), we implement ARRD decoder on single level feature map (1) and mixer of multi-level feature map (1,2) and (1,2,3) and so on. The experiment results has been shown in Table A1, we are glad our framework achieves a consistent improvement also on multi-level scenario. This strength our belief that the proposed strategy could leverage the recent advances in vision architectures.

Table A1. Experiments on RetinaNet-FPN structure on **COCO-d** dataset, we compare the ARRD with different level inputs from FPN.

Metrics	w/o ARRD	ARRD (1,2,3,4)	ARRD (1,2,3)	ARRD (1, 2)	ARRD (1)
AP	21.5	**22.1**	22.0	21.7	22.0
AP_s	2.5	2.8	2.8	2.8	**3.0**
AP_m	19.0	**19.7**	19.4	19.3	19.4
AP_l	46.0	**47.1**	46.5	46.3	46.8

Table B1. Experiments compare with fine-tuned restoration methods on **COCO-d** dataset.

	SRGAN [4]	DBPN [3]	Real-SR [1]	BSRGAN [7]	Restormer [6]	Ours
w/o FT	14.8	15.0	14.2	16.8	11.4	18.4
w FT	15.5	16.4	15.4	16.9	12.3	

B Restoration Network Fine-tune.

Avoiding fixed degradation parameter is actually the unique advantage of our method, since our AERIS framework directly finds intrinsic equivariant representation against various resolutions and degradation. To compare with existing image restoration methods as fair as possible, we further fine-tune the pre-train image restoration network [4, 3, 1, 7, 6] on the multi-degradation settings, same as the dataset generation of **COCO-d** except the resolution factor s, since the up-sampling resolution is often fixed in super-resolution network. We then make experiments on **COCO-d** dataset and results are shown in Table. B1. The fine-tune process would further improve the image restoration networks' performance, and our AERIS also gain best performance among different restoration methods.

C Different Scale in Training.

In AERIS training, we choose the down-sampling ratio s from a uniform distribution $s \sim (1, 4)$, for ablation study we further make the experiments with the different down-sampling scale range in training stage, the experimental results on **COCO-d** dataset are shown in Table C1, $s = 1$ means keep the original resolution and $s \sim (1, x)$ means to choose s from an uniform distribution $U(1, x)$.

References

1. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: Proceedings of the IEEE International Conference on Computer Vision (2019)

* Corresponding author.

Table C1. Ablation study about different training scale on **COCO-d** dataset.

s	∼(1, 4)	∼(1, 3)	∼(1, 2)	1
AP	18.4	**18.5**	17.8	17.6
AP_s	**2.7**	2.5	2.3	2.0
AP_m	**16.4**	16.0	15.0	14.8
AP_l	42.5	**44.0**	43.2	41.9

2. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
3. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
4. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 105–114 (2017). https://doi.org/10.1109/CVPR.2017.19
5. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944 (2017). https://doi.org/10.1109/CVPR.2017.106
6. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022)
7. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: IEEE International Conference on Computer Vision (2021)
8. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. CoRR **abs/1904.07850** (2019), `http://arxiv.org/abs/1904.07850`