Domain Generalization by Solving Jigsaw Puzzles - Supplementary Material

Fabio M. Carlucci[1*] Antonio D'Innocente[2,3] Silvia Bucci[3]
Barbara Caputo[3,4] Tatiana Tommasi[4]

[1]Huawei, London [2]University of Rome Sapienza, Italy
[3]Italian Institute of Technology [4]Politecnico di Torino, Italy

fabio.maria.carlucci@huawei.com {antonio.dinnocente, silvia.bucci}@iit.it
{barbara.caputo, tatiana.tommasi}@polito.it

We provide here some further analysis and experimental results on using jigsaw puzzle and other self-supervised tasks as auxiliary objectives to improve generalization across visual domains.

Visual explanation and Failure cases The relative position of each image patch with respect to the others captures visual regularities which are at the same time shared among domains and discriminative with respect to the object classes. Thus, by solving jigsaw puzzles we encourage the network to localize and re-join relevant object sub-parts regardless of the visual domain. This helps to focus on the most informative image areas. For an in-depth analysis of the learned model we adopted the Class Activation Mapping (CAM, [2]) method on ResNet-18, with which we produced the activation maps in Figure 1 for the PACS dataset. The first two rows show that JiGen is better at localizing the object class with respect to Deep All. The last row indicates that the mistakes are related to some flaw in data interpretation, while the localization remains correct.

Self-supervision by predicting image rotations Reordering image patches to solve jigsaw puzzle is not the only self-supervised approach that can be combined with supervised learning for domain generalization. We ran experiments by using as auxiliary self-supervised task the rotation classifier (four classes $[0°, 90°, 180°, 270°]$) proposed in [1]. We focused on the PACS dataset with the Alexnet-based architecture, following the same protocol used for JiGen. The obtained accuracy (Table 1) is higher than the Deep All baseline, but still lower than what obtained with our method. Indeed object 2d orientation provides useful semantic information when dealing with real photos, but it becomes less critical for cartoons and sketches.

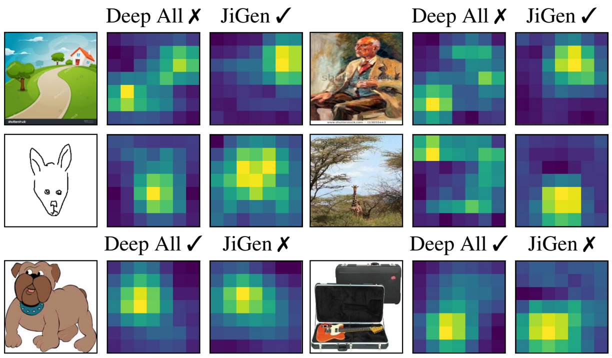

Figure 1. CAM activation maps: yellow corresponds to high values, while dark blue corresponds to low values. JiGen is able to localize the most informative part of the image, useful for object class prediction regardless of the visual domain.

PACS	art_paint.	cartoon	sketches	photo	Avg.
		Alexnet			
Deep All	66.68	69.41	60.02	**89.98**	71.52
Rotation	**67.67**	69.83	61.04	**89.98**	72.13
JiGen	67.63	**71.71**	**65.18**	89.00	**73.38**

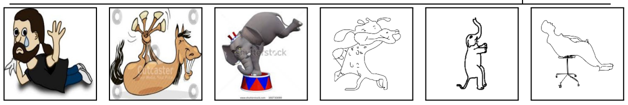

Table 1. *Top*: results obtained by using Rotation recognition as auxiliary self-supervised task. *Bottom*: three cartoons and three sketches that show objects with odd orientations.

References

[1] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

[2] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[*] This work was done while at University of Rome Sapienza, Italy