Supplementary for "General Object Pose Transformation Network from Unpaired Data"

Yukun Su[1,2], Guosheng Lin[2*], Ruizhou Sun[1], and Qingyao Wu[1,3*]

[1] School of Software Engineering, South China University of Technology
[2] Nanyang Technological University, Singapore
[3] Pazhou Lab, Guangzhou, China
suyukun666@gmail.com

Abstract. This supplementary material contains more details of the network architecture, additional qualitative results of transferred examples and additional experimental results in our paper "General Object Pose Transformation Network from Unpaired Data".

1 Details of Network Architecture

The overall architecture of our proposed network is shown in Tab 1, which depicts some of the detailed layers and modules. The foreground generator G_{fg} and the background generator G_{bg} both aim to perform reconstruction, thus, they have a similar architecture with encoder and decoder. Since we inject more TPS and dense warping information in G_{fg}, therefore, they do not share weights during training. Specifically, the background generator G_{bg} can be arbitrary Unet-like [6] network for background inpainting. We deploy a discriminator in the form as [8] to discriminate the generated fake images and real samples. We train our network on all datasets using the same proposed loss functions with the same hyper-parameters in an end-to-end manner. In our paper, we set the hyper-parameters empirically. During the training, we do not use any data augmentation strategies, and thus it is easy to reproduce. And we replace the least-squares loss with a hinge function to stabilize the discriminator.

2 More Qualitative Results

We perform more qualitative results on four datasets: *Mammals*, *Birds*, *Human* and *Cars* dataset in Fig 1 and Fig 2 respectively. We can yield more realistic images compared to other methods, and we take the early step to explore the general object pose transformation, which can be beneficial to applications covering wide range of objects.

* Corresponding authors.

	Module	Layers in the module	Output shape (H × W × C)
Correspondence Learning	Feature extractor × 2	Conv2d (3 × 3)	256 × 256 × 64
		Conv2d (4 × 4)	256 × 256 × 128
		Conv2d (3 × 3)	128 × 128 × 256
		Conv2d (4 × 4)	64 × 64 × 256
		Conv2d (3 × 3)	64 × 64 × 256
		Resblock × 3	64 × 64 × 256
	TPS Matching	Resblock × 2	64 × 64 × 256
		Conv2d (4 × 4)	16 × 16 × 256
	Dense Matching	Resblock × 2	64 × 64 × 256
		Conv2d (1 × 1)	64 × 64 × 256
Generating Network	*TSC* encoder	Bilinear Sampler	$h^i \times w^i \times 3$
		Conv2d (3 × 3)	$h^i \times w^i \times c^i$
	SS decoder	Bilinear Interpolation	$h^i \times w^i \times 3$
		Conv2d (3 × 3)	$h^i \times w^i \times 128$
		Conv2d (3 × 3)	$h^i \times w^i \times c^i$
	Generator	Conv2d (3 × 3)	256 × 256 × 64
		Conv2d (3 × 3)	8 × 8 × 1024
		Resblock × 5	128 × 128 × 256
		Conv2d (3 × 3)	128 × 128 × 128
		Resblock × 2	256 × 256 × 64
		Conv2d (3 × 3)	256 × 256 × 3

Table 1. The network architecture of our **UFO-PT**. The i^{th} *TSC* encoder and *SS* decoder outputs features with dimensions matching the i^{th} block in the generator.

3 Future work

As we mentioned in our main paper, our paper still have a large room for further improvement such as the masks we generate by using the off-the-shelf saliency detection method [5] and some of the bad cases of the human faces. Note that since we focus on general object pose transformation, we thereby do not adopt any prior cues like skeleton pose [1] and Face-Loss [4] to refine our model. In the future work, one can adopt more dedicated objective functions to improve the local region patterns [3] for pose transformation. Besides, smoother and finer masks can be obtained by using stronger detectors [7,2] for further improvement, and coarse-to-fine network design can be also considered.

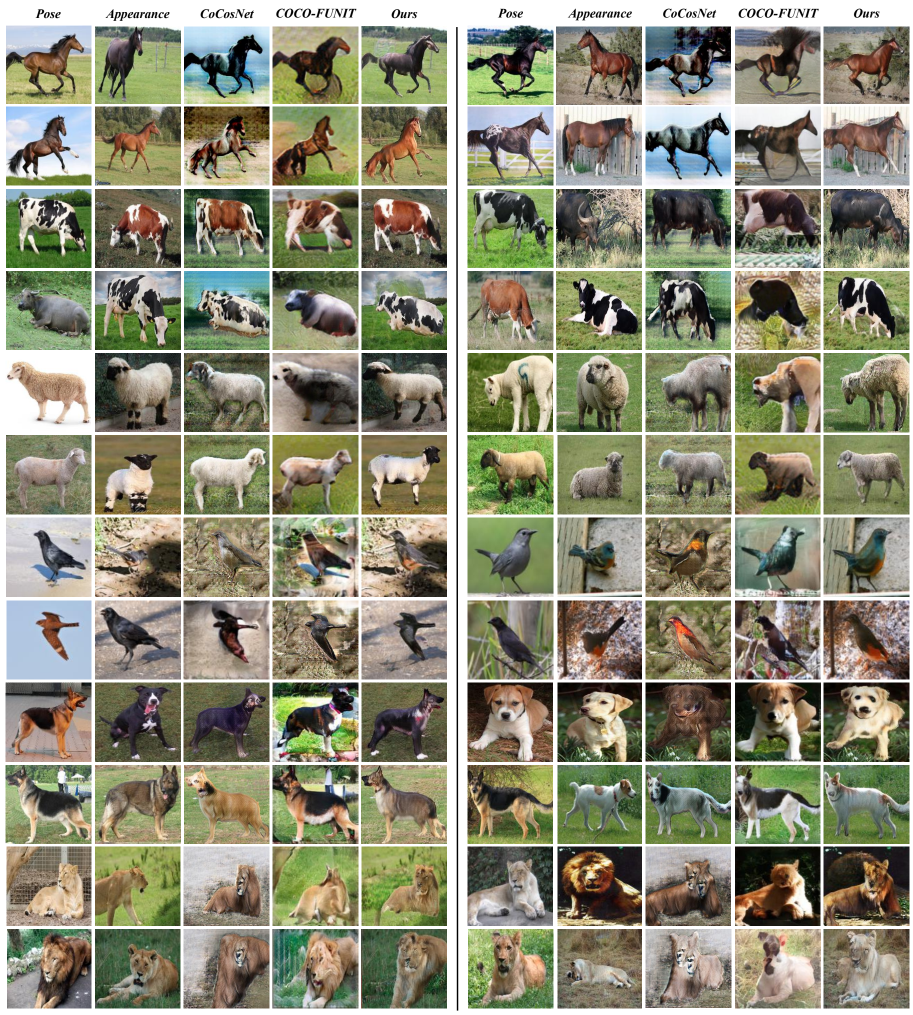

Fig. 1. More qualitative comparison of different methods on *Mammals* and *Birds* dataset.

Fig. 2. More qualitative comparison of different methods on *Human* and *Cars* dataset.

References

1. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019) 2
2. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017) 2
3. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 2
4. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017) 2
5. Nguyen, D.T., Dax, M., Mummadi, C.K., Ngo, T.P.N., Nguyen, T.H.P., Lou, Z., Brox, T.: Deepusps: Deep robust unsupervised saliency prediction with self-supervision. arXiv preprint arXiv:1909.13055 (2019) 2
6. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 1
7. Su, Y., Deng, J., Sun, R., Lin, G., Wu, Q.: A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. arXiv preprint arXiv:2203.04708 (2022) 2
8. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018) 1