Artistic Object Recognition by Unsupervised Style Adaptation (Supplementary Material)

Christopher Thomas and Adriana Kovashka

University of Pittsburgh, Pittsburgh PA 15260, USA
{chris,kovashka}@cs.pitt.edu

1 Introduction

This document presents supplementary results which complement our main text.

We first include an ablation of our method in Section 2. There, we explore the importance of selecting "representative styles" as opposed to random style images, whether style transfer matters for this task, and whether our style invariance loss improves performance.

In Section 3, we test the most competitive methods from our main text using a 152-layer residual network instead of AlexNet to see if our conclusions continue to hold for more recent architectures. We find that they do.

In Section 4, we show the result of training our networks on a *single* modality, as opposed to two modalities at once. We find that training on a single modality works *worse* overall than training on both photos and style-transferred photos.

Finally, in Section 5 we show the difference between Johnson's and Huang's style transfer methods, and explain how these differences lead to different recognition results on different datasets. These observations help decide which style transfer method is better for a given target domain, depending on whether fine-grained recognition is necessary.

	CASPA			
Method	**Paintings**	**Cartoons**	**Sketches**	**AVG**
Photo-AlexNet	0.663	0.222	0.398	0.428
Ours-Style Modification	0.649	0.385	*0.594*	*0.543*
Ours-Style Modification (-L)	0.646	0.300	0.514	0.486
Ours-Style Selection (shown in main)	*0.677*	*0.406*	**0.625**	**0.569**
Ours-Style Selection (-L)	**0.702**	**0.433**	0.485	0.540
Ours-Edge Maps	0.635	0.360	0.556	0.517

Table 1. We explore the impact of different synthetic modalities when training our method and the impact of removing the style invariance (domain confusion) loss. The best-performing method per row is shown in **bold**, and the second-best in *italics*.

2 Method Ablation

In this section we present an experiment testing different ways of constructing the style-transferred images as our auxiliary modality.

We first explore the importance of selecting "representative styles" from our target dataset vs. randomly choosing target styles. Rows with Style Selection use our style selection procedure described in the main text to choose ten representative styles for transfer, while Style Modification use ten randomly chosen style images. Both use Johnson et al. [4] for performing style transfer. We observe that for most modalities style selection tends to improve performance. We observe that there is a fairly consistent improvement when "representative styles" are chosen rather than random styles.

We next explore the importance of using style transfer as opposed to a more generic, target-domain-agnostic synthetic modality. We extract edge maps on the COCO images using the technique of Xie et al. [8]. Aside from using *edge maps instead of style-transferred images*, we train our AlexNet classification network as described in our main text, i.e. on two modalities of data, using a style invariance loss and a classification loss. Our experiments show that this technique does improve performance over just using COCO images on average, though not as much as performing style transfer, demonstrating the importance of controlling for style in artistic domains. We note that edge maps work most competitively on sketches. This makes sense because edge maps consist of coarse outlines of objects, much like sketches.

Finally, we explore the importance of our style invariance (domain confusion) loss. Rows indicated by (-L) use a network trained *without* the style invariance loss. We find that for our most photorealistic modality, paintings, the loss is of little benefit. However, for sketches, our most non-photorealistic domain, we obtain a substantial improvement. Interestingly, we find that for Style Selection on cartoons, the loss slightly hurts performance, but for Style Modification, it significantly helps. This may be because for Style Selection, the styles have been carefully chosen and already represent the target domain well, but for Style Modification, because the styles are randomly chosen, explicitly encouraging style invariance with the loss improves performance.

ResNet Results				
	CASPA			
Method	**Paintings**	**Cartoons**	**Sketches**	**AVG**
PHOTO-RESNET	*0.733*	0.487	0.539	0.586
LONG [6]	0.709	0.512	0.600	0.607
BOUSMALIS [1]	0.719	*0.529*	*0.639*	*0.629*
OURS-JOHNSON	**0.775**	**0.555**	**0.719**	**0.683**
UPPER BOUND	0.919	0.837	0.945	0.900

Table 2. Our best performing methods from the main text, using the ResNet-152 [2] architecture. The best-performing method (excluding UPPER BOUND) per row is shown in **bold**, and the second-best in *italics*.

3 ResNet Results

In this section we experiment with the 152-layer residual network architecture of [2] fine-tuned on our ten animal categories, which has been shown to perform significantly better than AlexNet on many problems. We test the most promising methods from Table 1 from our main text.

We see that our conclusions from Table 1 also hold for ResNets. In particular, our method of using style transfer, style selection and a style invariance loss outperforms all other methods across all categories on our dataset. Similar to AlexNet, the least improvement is obtained on paintings, given their similarity to photos.

	CASPA			
Method	**Paintings**	**Cartoons**	**Sketches**	**AVG**
Photo-AlexNet	0.663	0.222	0.398	0.428
Only Synthetic	0.611	*0.429*	0.418	0.486
Ours-Style Selection (shown in main)	*0.677*	0.406	**0.625**	**0.569**
Ours-Style Selection (-L)	**0.702**	**0.433**	*0.485*	*0.540*
Upper Bound	0.842	0.741	0.917	0.833

Table 3. Training on a single modality instead of multiple modalities. We find that our method works significantly better when trained with a second modality.

4 Single Modality Results

Our method in Table 1 (in both the main text and supplementary) uses both photos and style-transferred photos, together with the style invariance loss. We wanted to see whether we could avoid the style invariance loss altogether by training only on our style-transferred modality. In Table 3 above, we see that Only Synthetic performs significantly worse overall than training on multiple modalities.

The one exception is for cartoons, where we find that training on a single modality works better. However, we find that when we remove the style invariance loss, Ours-Style Selection (-L) outperforms Only Synthetic. This indicates that even though neither method has the invariance loss, adding photos as an additional modality still performs better than just training on style-transferred photos.

One explanation for this is that while the style transferred photos bear the style of the target domain, there are still domain differences unaccounted for by style transfer. Thus, the style-transferred photos still have some domain gap with the target modality. Thus, retaining the photos as a training modality along with the style transferred images enables us to learn a style-invariant representation (when we use our style invariance loss), while also preserving some image features which may be missing in our style-transferred photos, such as textures. Additionally, training on multiple modalities forces our networks to learn a more general representation, rather than a domain-specific one. The generality obtained by training a multi-modality network proves to be useful at bridging some of the gap between the style-transferred images and the actual target domain.

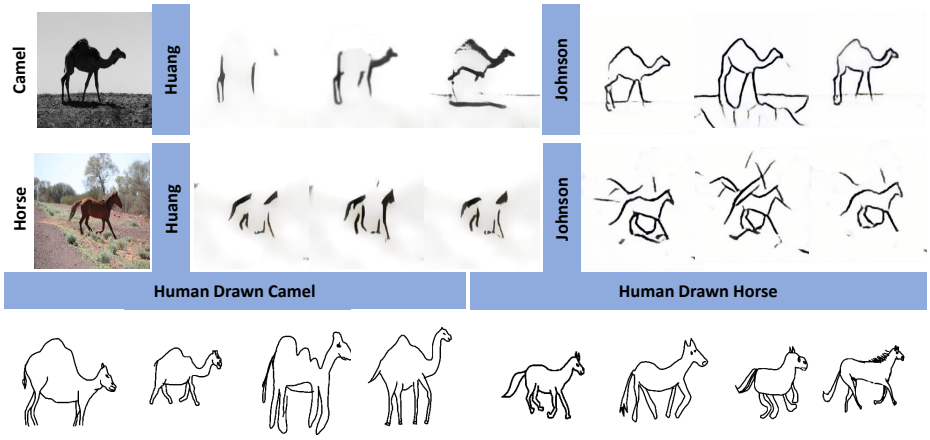

Fig. 1. Comparison of Huang's vs. Johnson's outputs on sketches. We observe that Huang's method often omits important details (i.e. the hump on the camel, which is necessary for distinguishing between the categories).

5 Johnson [4] vs. Huang [3] on Sketches

In Table 1 of our main text, we obtain a somewhat contradictory result. On the full **Sketchy** [7] and **CASPA** datasets, using Johnson's style transfer technique significantly outperforms Huang's, while on the **PACS** [5] dataset, using Huang's method slightly outperforms using Johnson's. This is somewhat puzzling since the **Sketch** modality in both **PACS** and **CASPA** is a subset of the categories from **Sketchy**. Upon closer inspection of the style-transferred data, however, *this result makes sense*. The **PACS** dataset contains seven categories: dog, elephant, giraffe, guitar, horse, house, and person. Importantly, most of these objects have significantly different overall shapes. The **CASPA** dataset contains ten categories, all of which are animals: bear, bird, cat, cow, dog, elephant, giraffe, horse, sheep, and zebra. Differentiating among these categories is more challenging when only looking at coarse shape (i.e. horse could be mistaken for dog if only looking at an outline). The full **Sketchy** dataset contains 125 categories, which require even finer-grained distinctions (i.e. wading bird vs. parrot).

When we look at the style-transferred images produced by Johnson's method compared to Huang's on the sketch domain, we find that Huang's method tends to produce sketches missing important details, while Johnson's usually does not. This is important because when using Huang's method the network may learn to only look for coarse shapes of an object and is thus not able to distinguish between objects with similar shapes. Thus, a network trained with sketches from Huang's method cannot distinguish between finer-grained categories as well as one trained with sketches from Johnson's method. We illustrate the difference between Huang's sketches vs Johnson's in Fig. 1. Notice that Huang omits the camel's hump and the horse's body, while Johnson includes these details.

References

1. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
3. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
4. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
5. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Computer Vision (ICCV), 2017 IEEE International Conference on. pp. 5543–5551. IEEE (2017)
6. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. In: Advances in Neural Information Processing Systems. pp. 136–144 (2016)
7. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM Transactions on Graphics (TOG) **35**(4), 119 (2016)
8. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1395–1403 (2015)