Artistic Object Recognition by Unsupervised Style Adaptation

Christopher Thomas and Adriana Kovashka

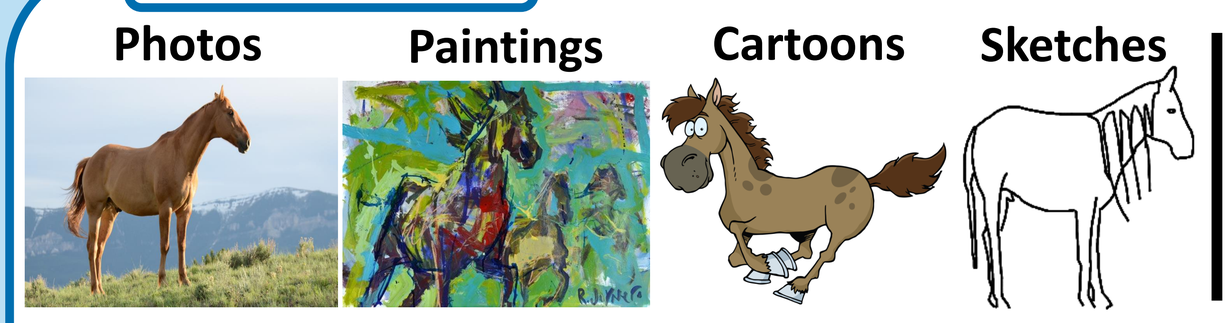
University of Pittsburgh

Department of Computer Science

ACCV2018 2 – 6 December 2018 Perth Western Australia

14th Asian Conference on Computer Vision

Introduction



Artistic domains such as paintings, cartoons, and sketches exhibit **wider variation in object appearance across domains than within a single domain**. A confounding problem is that **sufficient labeled target domain training data may not be available** (or may not even exist). We present a **method for recognition in artistic domains using style transfer techniques**. Our method **transforms labeled photos into labeled images of the domain of interest for free.** We then use our synthetically created data and original images to train classifiers. Our method **outperforms state-of-the-art domain adaptation baselines** on artistic recognition, while **requiring orders of magnitude less target data**.

New Dataset: CASPA

CASPA: Cartoons, Sketches, Paintings

We create and release for download a new dataset of **10** animal categories (**bear**, **bird**, **cat**, **cow**, **dog**, **elephant**, **giraffe**, **horse**, **sheep**, and **zebra**) in **3 artistic domains**: cartoons, sketches, and paintings.

Our dataset contains **18,446 artistic images**, almost twice the number of the PACS (Li et al., 2017) dataset. Our dataset also includes nearly **70,000** photos of the animals of interest (also sig. more than PACS).

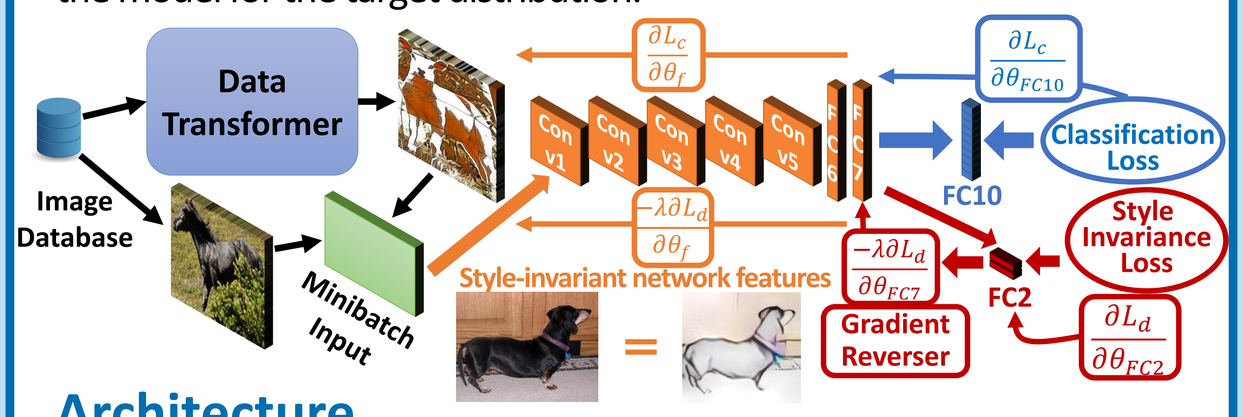
Download our dataset from: **www.cs.pitt.edu/~chris/artistic_objects**

Method Overview

Our method trains classifiers which explicitly **control for the stylistic gap between source and target domains** using synthetic training data produced using style transfer techniques.

Motivation

By training on photos alone, our model is not prepared to recognize in the domain of interest. **Transformed images are *more* like the target**, but still slightly differ. Invariance to ***both*** source domains best prepares the model for the target distribution.



Architecture

We illustrate our model's architecture and training procedure above. We provide photos and style-transformed photos to our model and learn **style-invariant CNN features**. Our model is trained to minimize two losses: a **classification loss** (to recognize the categories of interest) and a **style-invariance loss** (which ensures that the CNN features of our two input types are indistinguishable.

Problem Formulation

Let $\boldsymbol{I_s} = \{I_s^i, y_s^i\}_{i=1}^{N_s}$ represent our source dataset s of $|N_s|$ labeled images and $\boldsymbol{I_t}$ our unlabeled target dataset. We use style transfer networks $\boldsymbol{\Psi}(I_s^i) \rightarrow \widehat{I_s^i}$ to ***transform each labeled source photo into a labeled source image of the same style as our target*** to form $\widehat{\boldsymbol{I_s}}$.

Our network above is trained to perform the following two tasks:

$$\boldsymbol{T}(\{\boldsymbol{I_s}, \widehat{\boldsymbol{I_s}}\}; \boldsymbol{\theta_Y}; \boldsymbol{\theta_F}; \boldsymbol{\theta_D}) \rightarrow \{\mathbf{y}, \mathbf{d}\}$$

where $\boldsymbol{\theta_Y}; \boldsymbol{\theta_F}; \boldsymbol{\theta_D}$ are the parameters used for **classification**, the **shared parameters**, and those used to **predict the image's domain.**

We thus seek to ***maximally confuse*** the domain classifier while ***minimizing the number of misclassified objects:***

$$\mathbf{min}_{\boldsymbol{\theta_F};\boldsymbol{\theta_Y}} \mathbf{max}_{\boldsymbol{\theta_D}} \boldsymbol{\alpha} \boldsymbol{L_d}(\{\boldsymbol{I_s}, \widehat{\boldsymbol{I_s}}\}, \mathbf{d}) + \boldsymbol{\beta} \boldsymbol{L_y}(\{\boldsymbol{I_s}, \widehat{\boldsymbol{I_s}}\}, \mathbf{y})$$

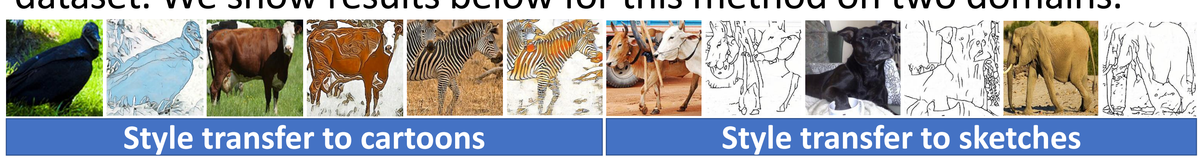
where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are weighting hyperparameters.

Style Transformation

In order to transform images into the style of the target domain, we use two off-the-shelf methods for style transformation:

Johnson et al., 2016

Requires **training a single network for each target style, which is computationally expensive.** We achieved the best results by using 10 "representative styles" from style clusters within our artistic dataset. We show results below for this method on two domains.



Style transfer to cartoons | Style transfer to sketches

Huang et al., 2017

This method performs style transfer in feature space and then generates a transformed image and does not require training separate networks. For our Huang result, **we randomly sample a target image for each photo** and transform it using that style.

Experimental Results

Select Baselines

Long et al., 2016 learn separate classifiers for source / target domains

Bousmalis et al., 2017 use GAN to modify source data to be like target data

Datasets

PACS (Li et al., 2017) – Paintings, **C**artoons, **S**ketches in seven categories

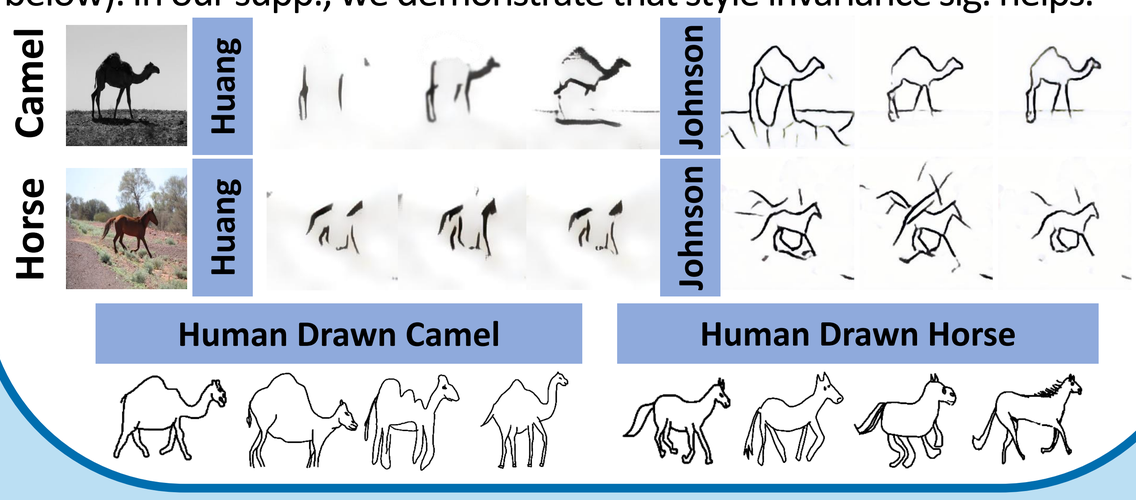
CASPA – Our dataset of paintings, cartoons, and sketches in ten categories

Sketchy (Sangkloy et al., 2017) – 75,471 sketches in 125 categories

Castrejon et al., 2016 – 205 *scene* categories in multiple modalities

METHOD	PACS	CASPA	Sketchy	Castrejon-Clipart	Castrejon-Sketches
Photo-AlexNet	0.388	0.428	0.093	0.0689	0.0213
Long et al.	0.566	0.517	*0.303*	0.0727	0.0402
Bousmalis et al.	0.547	0.519	0.284	0.0839	*0.0464*
Ours-Johnson	*0.587*	**0.569**	**0.326**	*0.0847*	**0.0479**
Ours-Huang	**0.596**	*0.554*	0.234	**0.0885**	0.0456
Upper Bound	0.904	0.833	0.822	0.5937	0.3120

We show a subset of our results above comparing our method to our **top two performing baselines** on four datasets using Alexnet. ***Variations of our method are always best for all datasets.*** We observe that Huang and Johnson often perform comparably, but at times differ significantly (see below). In our supp., we demonstrate that style invariance sig. helps.



Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Number 1566270.