Detecting People in Artwork with CNNs

Nicholas Westlake[1], Hongping Cai[2], and Peter Hall[1]

[1] Department of Computer Science, Unversity of Bath, Bath, UK
{n.westlake,p.m.hall}@bath.ac.uk
[2] Department of Computer Science, University of Bristol, Bristol, UK
hongping.cai@bristol.ac.uk

Abstract. CNNs have massively improved performance in object detection in photographs. However research into object detection in artwork remains limited. We show state-of-the-art performance on a challenging dataset, *People-Art*, which contains people from photos, cartoons and 41 different artwork movements. We achieve this high performance by fine-tuning a CNN for this task, thus also demonstrating that training CNNs on photos results in overfitting for photos: only the first three or four layers transfer from photos to artwork. Although the CNN's performance is the highest yet, it remains less than 60% AP, suggesting further work is needed for the cross-depiction problem.

Keywords: CNNs, cross-depiction problem, object recognition

1 Introduction

Object detection has improved significantly in recent years, especially as a result of the resurgence of convolutional neural networks (CNNs) and the increase in performance and memory of GPUs. However, in spite of the successes in photo-based recognition and detection, research into recognition within styles of images other than natural images (photos) remains limited [1]. We refer to this as the *cross-depiction problem*: detecting objects regardless of how they are depicted (photographed, painted, drawn, etc.).

We believe that cross-depiction recognition is an interesting and open problem. It is interesting because it forces researchers to look beyond the surface appearance of object classes. By analogy, just as a person retains their identity

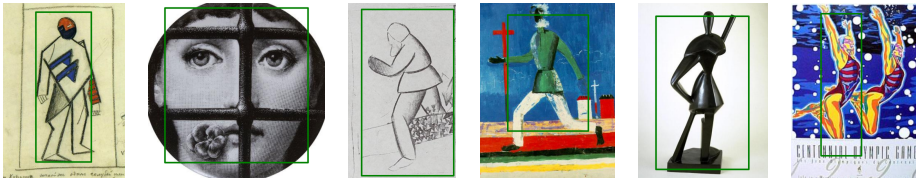

Fig. 1. Detecting people across different depictive styles a challenge: here we show some successful detections.

no matter what clothes they wear, so an object retains its class identity no matter how it is depicted: a dog is a dog whether photographed, painted in oils, or drawn with a stick in the sand.

Cross-depiction is a practical problem too: an example is an image search. The world contains images in all sorts of depictions. Any recognition solution that does not generalise across these depictions is of limited power. Yet most current computer vision methods tacitly assume a photographic input, either by design or training. Any model premised on a single depictive style e.g. photos will lack sufficient descriptive power for cross-depiction recognition. Therefore, an image search using methods will limit its results to photos and photo-like depictions.

In our paper, we talk about natural images (photos) and non-natural images (artwork) as a linguistic convenience. We would argue that this is a false dichotomy: the universe of all images includes images in all possible depictive styles, and there is no particular reason to privilege any one style. Nevertheless, we acknowledge that the distribution of styles is not uniform: photos may be more abundant and certainly are in computer vision datasets such as ImageNet [2]. This creates problems for generalisation: training a detector on photos alone constrains it not only in terms its ability to handle denotational varieties, but projective and pose varieties too, as we discuss later.

We present a new dataset, *People-Art*, which contains photos, cartoons and images from 41 different artwork movements. Unlike the *Photo-Art* dataset [3], which had 50 classes, this dataset has a single class: people. We labelled people since we observe that people occur far more frequently across the wide spectrum of depictive styles than other classes, thus allowing a far greater variety. Detecting people within this dataset is a challenging task because of the huge range of ways artists depict people: from Picasso's cubism to Disney's Sleeping Beauty. The best performance on a pre-release of the dataset is 45% average precision (AP), from a CNN that was neither trained nor fine-tuned for this task. By fine-tuning a state-of-the-art CNN for this task [4], we achieved 58% AP, a substantial improvement.

As well as achieving state-of-art performance on our *People-Art* dataset, we make the following contributions, in order of strength:

1. We show that a simple tweak for the "Fast Region-based Convolutional Network" method (Fast R-CNN) [4], changing the criteria for negative training exemplars compared to default configuration, is key to higher performance on artwork.
2. We show the extent to which fine-tuning a CNN on artwork improves performance when detecting people in artwork on our dataset (Section 5.2) and the *Picasso* dataset [5] (Section 5.4). We show that this alone is not a solution: the performance is still less than 60% AP after fine tuning, suggesting the need for futher work.
3. Consistent with earlier work [6], we show that the lower convolutional layers of a CNN generalise to artwork: others benefit from fine-tuning (Section 5.1).

We begin by presenting related work and our *People-Art* dataset.

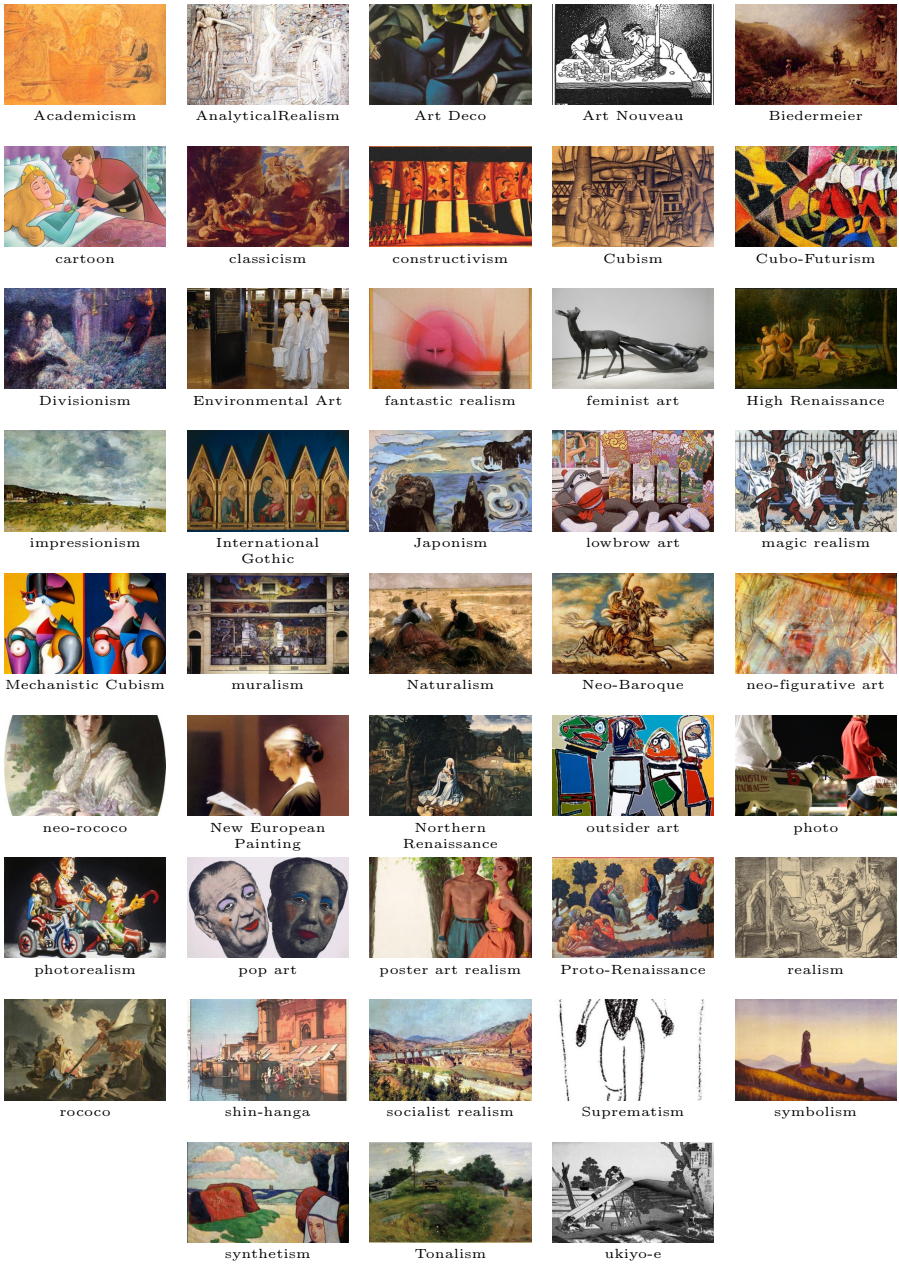

Fig. 2. Our *People-Art* dataset contain images from 43 different styles of depiction: here we show one example for depiction style.

2 Related Work

We use a state-of-the-art CNN to improve performance on a cross-depiction dataset, thereby contributing towards cross-depiction object recognition. We first explore related work on deep learning for object detection and localisation (largely in photos), followed by previous work on the cross-depiction problem.

2.1 Deep Learning for Object Detection and Localisation

Deep learning has been around for a few decades [7–9]. After a period of limited use within computer vision, Krizhevsky et al. (2012) [10] demonstrated a vast performance improvement for image classification over previous state-of-the-art methods, using a deep CNN. As a result, the use of CNNs surged within computer vision.

Early CNN based approaches for object localisation [11–14] used the same sliding-window approach used by previous state-of-the-art detection systems [15, 16]. As CNNs became larger, and with an increased number of layers, this approach became intractable. However, Sermanet et al. (2014) [17] demonstrated that few windows are required, provided the CNN is fully convolutional. Furthermore, as the size of their receptive fields increased, CNNs either became or were trained to be less sensitive to precise location and scale the input. As a result, obtaining a precise bounding box using sliding window and non-maximal suppression became difficult. One early approach attempted to solve this issue by training a separate CNN for precise localisation [18].

Szegedy et al. (2013) [19] modified the architecture of Krizhevsky et al. (2012) [10] for localisation by replacing the final layer of the CNN with a regression layer. This layer produces a binary mask indicating whether a given pixel lies within the bounding box of an object. Schulz and Behnke (2011) [20] previously used a similar approach with a much smaller network for object segmentation.

Girshick et al. (2014) [21] introduced "regions with CNN features" (R-CNN), which surpassed previous approaches. The authors used selective search [22], a hierarchical segmentation method, to generate region proposals: possible object locations within an image. Next, a CNN obtains features from each region and a support vector machine (SVM) classifies each region. In addition, they used a regression model to improve the accuracy of the bounding box output by learning bounding box adjustments for each class-agnostic region proposal. He et al. (2015) [23] improved the run-time performance by introducing SPP-net, which uses a spatial pyramid pooling (SPP) [24, 25] layer after the final convolutional layer. The convolutional layers operate on the whole image, while the SPP layer pools based on the region proposal to obtain a fixed length feature vector for the fully connected layers.

Girshick (2015) [4] later introduced Fast R-CNN which improves upon R-CNN and SPP-net and allows the CNN to output a location of the bounding box (relative to the region proposal) directly, along with class detection score, thus replacing the SVM. Furthermore, this work enables end-to-end training of the

whole CNN for both detection and bounding box regression. We use this approach to achieve state-of-the-art performance on our *People-Art* dataset and detail the method in Section 4.

To make Fast R-CNN even faster and less dependent on selective search [22], Lenc and Vedaldi (2015) [26] used a static set of region proposals. Ren et al. (2015) [27] instead used the output of the existing convolutional layers plus additional convolutional layers to predict regions, resulting in a further increase in accuracy and efficiency.

Redmon et al. (2015) [28] proposed "You Only Look Once" (YOLO), which operates quicker though with less accuracy than other state-of-art approaches. A single CNN operates on an entire image, divided in a grid of rectangular cells, without region proposals. Each cell outputs bounding box predictions and class probabilities; unlike previous work, this occurs simultaneously. Huang et al. (2015) [29] proposed a similar system, introducing up-sampling layers to ensure the model performs better with very small and overlapping objects.

2.2 Cross-Depiction Detection and Matching

Early work relating to non-photographic images focused on matching hand-drawn sketches. Jacobs et al. (1995) [30] used wavelet decomposition of image colour channels to allow matching between a rough colour image sketch and a more detailed colour image. Funkhouser et al. (2003) [31] used a distance transform of a binary line drawing, followed by fourier analysis of the distance transforms at fixed radii from the centre of the drawing, to match 2D sketches and 3D projections, with limited performance. Hu and Collomosse (2013) [32] used a modified version of Histograms of Oriented Gradients (HOG) [15] to extract descriptors at interest-points in the image: for photographs, these are at Canny edges [33] pixels; for sketches, these are sketch strokes. Wang et al. (2015) [34] used a siamese CNN configuration to match sketches and 3D model projections, optimising the CNN to minimise the distances between sketches and 3D model projections of the same class.

Another cross-depiction matching approach, by Crowley et al. (2015) [35], uses CNN generated features to match faces between photos and artwork. This relies on the success of a general face detector [36], which succeeds on artwork which is "largely photo-realistic in nature" but has not been verified on more abstract artwork styles such as cubism.

Other work has sought to use self-similarity to detect patterns across different depictions such as Shechtman and Irani (2007) [37] and Chatfield et al. (2009) [38] who used self-similarity descriptors formed by convolving small regions within in image over a larger region. This approach is not suitable for identifying (most) objects as a whole: for example, the results show effective matching of people forming a very specific pose, not of matching people as an object class in general.

Recent work has focused on cross-depiction object classification and detection. Wu et al. (2014) [3] improved upon Felzenszwalb et al.'s Deformable

Part-based Model (DPM) [16] to perform cross-depiction matching between photographs and "artwork", (including "clip-art", cartoons and paintings). Instead of using root and part-based filters and a latent SVM, the authors learnt a fully connected graph to better model object structure between depictions, using the structured support vector machine (SSVM) formulation of Cho et al. (2013) [39]. In addition, each model has separate "attributes" for photographs and "artwork": at test-time, the detector uses the maximum response from either of "attribute" set, to achieve depiction invariance. This work improved performance for detecting objects in artwork, but depended on a high performing DPM to bootstrap the model. Our dataset is more challenging than the one used, leading to a low accuracy using DPM and hence this is approach is also not suitable.

Zissermann et al. (2014) [40] evaluate the performance of CNNs learnt on photos for classifying objects in paintings, showing strong performance in spite of the different domain. Their evaluation excludes people as a class, as people appear frequently in their paintings without labels. Our *People-Art* dataset addresses this issue: all people are labelled and hence we provide a new benchmark. We also believe our dataset contains more variety in terms of artwork styles and presents a more challenging problem. Furthermore, we advance their findings: we show the performance improvement when a CNN is fine-tuned for this task rather than simply fine-tuned on photos.

3 The *People-Art* Dataset and its Challenges

Our *People-Art* dataset[3] contains images divided into 43 depiction styles. Images from 41 of these styles came from *WikiArt.org* while the photos came from PASCAL VOC 2012 [41] and the cartoons from google searches. We labelled people since, according to our empirical observations, people are drawn or painted more often than other objects. Consequently, this increases the total number of individual instances and thus the range of depictive styles represented. Figure 2 shows one painting from each style represented in our *People-Art* dataset.

The 41 depictive styles from *WikiArt.org* are catagorised based on art movements. These depiction styles cover the full range of projective and denotational styles, as defined by Willats [42]. In addition, we propose that these styles cover many poses, a factor which Willats did not consider.

We believe that our dataset is challenging for the following reasons:

range of denotational styles This is the style with which primitive marks are made (brush strokes, pencil lines, etc.) [42]. We consider photos to be a depictive style in its own right.

range of projective style This includes linear camera projection, orthogonal projection, inverse perspective, and in fact a range of ad-hoc projections [42]. An extreme form is shown in cubism, in which it is common for the view of a person from many different viewpoints to be drawn or painted on the 2D canvas [5].

[3] https://github.com/BathVisArtData/PeopleArt

range of poses Though pose is handled by previous computer vision algorithms [16], we have observed that artwork, in general, exhibits a wider variety of poses than photos.

overlapping, occluded and truncated people This occurs in artwork as in photos, and perhaps to a greater extent.

4 CNN architecture

We use the same architecture as Fast R-CNN [4], which is built around a modified version of the Caffe library [43]. The CNN has two inputs: an image and a set of class-agnostic rectangular region proposals. Many algorithms exist for generating region proposals; we use selective search [22] with the default configuration.

The first stage of the CNN operates on the entire image (having been resized to a fixed dimension while preserving aspect ratio). This stage consists of convolutional layers, rectified linear units (ReLUs) [10, 44], max-pooling layers and, in some cases, local response normalisation layers [10]. The final layer is a region of interest (ROI) pooling layer which is novel to Fast R-CNN: as well as the input from the previous convolutional or ReLU layer, this layer receives another input, a region proposal or ROI; the output is a fixed-length feature vector formed by

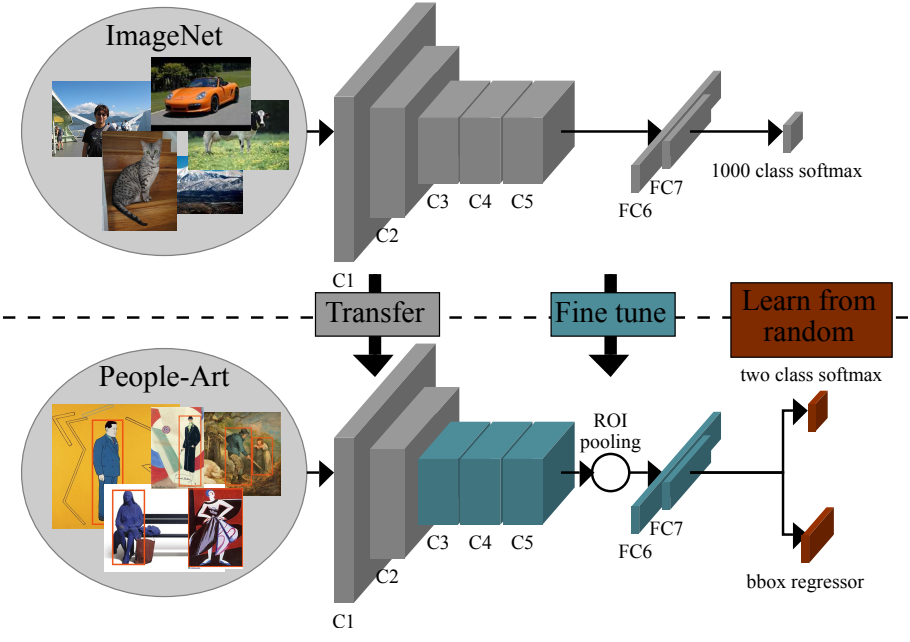

Fig. 3. We use a network pre-trained on ImageNet and fine-tuned on our *People-Art* dataset (training and validation sets): we fix the weights for the first F layers, selected by validation.

max-pooling of the convolution features. In order to preserve information about the global structure of the ROI, i.e. at a scale within an order of magnitude of the ROI size, the max-pooling happens over a uniformly spaced rectangular grid, size $H \times W$. As a result, the layer outputs feature vector with CHW dimensions where C is the number of channels of the previous convolutional layer.

This feature vector is the input to the second stage of the CNN, which is fully connected. It consists of inner product and ReLU layers, as well as dropout layers (training only) aimed at preventing overfitting [45]. The output for each class is a score and a set of four co-ordinate which indicate the bounding box co-ordinates relative to the ROI. We modified the final layer to output a score and bounding box prediction for only one class: person.

We use the same approach for training as Fast R-CNN, which uses stochastic gradient descent (SGD) with momentum [10], initialising the network with weights from the pre-trained models, in our case, trained on ImageNet [2, 10]. We fix the weights of the first F convolutional layers to those in the pre-trained model; this parameter is selected by validation. We experiment with different criteria for the region proposals to use as training ROI, as detailed in Section 5.1. Since the final inner product layers have a different size output as we only detect one class, we use random (Gaussian) initialisation. Figure 3 shows our network architecture in detail.

We fine-tune the models (pre-trained on ImageNet) using our *People-Art* dataset (training and validation sets). We test three different models: CaffeNet, which is a reproduction of AlexNet [10] with some minor changes, Oxford VGG's "CNN M 1024" (VGG1024) [46] and Oxford VGG's "Net D" (VGG16) [47]. Both CaffeNet and VGG1024 have five convolutional layers and local response normalisation layers and vary slightly: in particular VGG1024 has more weights and channels. VGG16 is much a larger network, with thirteen convolutional layers and no local response normalisation. Except for the number of dimensions, all three networks have the same ROI pooling layer and fully connected network structure: each CNN's fully connected network structure consists of two inner product layers, each followed by ReLU and dropout layers (training only).

5 Experiments

For both validation and testing, our benchmark is average precision (AP): we calculate this using the same method as PASCAL Visual Object Classes (VOC) detection task [48]. A positive detection is one whose intersection over union (IoU) overlap with a ground-truth bounding box is greater than 50%; duplicate detections are considered false. Annotations marked as difficult are excluded.

5.1 ROI Selection and Layer Fixing for CNN Fine-Tuning

Although we used the default selective search settings to generate region proposals, we experimented with different criteria to specify which region proposals to use in training. The default configuration of Fast-RCNN [4] defines positive

ROI be region proposals whose IoU overlap with a ground-truth bounding box is at least 0.5, and defines negative ROI to be those whose overlap lies in the interval $[0.1, 0.5)$. The cutoff between positive and negative ROI matches the definition of positive detection according the VOC detection task [48]. Girshick (2015) states that the lower cut-off (0.1) for negative ROI appears to act as a heuristic to mine hard examples [4, 16].

We experimented with two alternative configurations for fine tuning:

gap We discarded ROI whose IoU overlap with a ground-truth bounding box lies in the interval $[0.4, 0.6)$: we hypothesised that ROI lying in this interval are ambiguous and hamper training performance.

all-neg We removed the lower bound for negative ROI. We hypothesised that this would improve performance on our *People-Art* dataset for two reasons:

1. This results in the inclusion of ROI containing classes which appear similar to people, for example animals with faces.
2. This permits the inclusion of more artwork examples, for example images without any people present. We hypothesised that this would make the CNN better able to discern between features caused by the presence of people and features resulting from a particular depiction style.

We fixed all other hyper-parameters of the CNN except for F, the number of convolutional layers whose weights we fix to those learnt from ImageNet, which we select based validation performance.

Table 1 shows the validation performance for the different criteria, i.e. from testing on the validation set after fine-tuning on the *People-Art* training set. Removing the lower bound on negative ROI (all-neg) results in a significant increase in performance, around a 9 percentage point increase in average precision

Table 1. Validation performance using different criteria for positive and negative ROI: we use CNNs pre-trained on ImageNet, fine-tune on the training set and then test on the validation set; we select the best configuration for each CNN (bold).

		ROI IoU			
CNN	**configuration**	**negative**	**positive**	**fixed layers (F)**	**AP**
CaffeNet	default	$[0.1, 0.5)$	≥ 0.5	2	33.7%
CaffeNet	gap	$[0.1, 0.4)$	≥ 0.6	2	33.5%
CaffeNet	**all-neg**	$\mathbf{[0.0, 0.5)}$	$\mathbf{\geq 0.5}$	**0**	**42.5%**
CaffeNet	gap + all-neg	$[0.0, 0.4)$	≥ 0.6	1	42.2%
VGG1024	default	$[0.1, 0.5)$	≥ 0.5	1	38.4%
VGG1024	gap	$[0.1, 0.4)$	≥ 0.6	3	35.8%
VGG1024	**all-neg**	$\mathbf{[0.0, 0.5)}$	$\mathbf{\geq 0.5}$	**1**	**42.6%**
VGG1024	gap + all-neg	$[0.0, 0.4)$	≥ 0.6	1	42.0%
VGG16	default	$[0.1, 0.5)$	≥ 0.5	1	43.9%
VGG16	gap	$[0.1, 0.4)$	≥ 0.6	2	39.0%
VGG16	all-neg	$[0.0, 0.5)$	≥ 0.5	3	50.0%
VGG16	**gap + all-neg**	$\mathbf{[0.0, 0.4)}$	$\mathbf{\geq 0.6}$	**3**	**50.1%**

in the best performing case. Indeed, it appears that what is *not* a person is as important as what *is* a person for training. Discarding ROI with an IoU overlap in the interval $[0.4, 0.6)$ yields mixed results: it was marginally beneficial in one case, and detrimental in all others.

We note that the optimal number of convolutional layers for which to fix weights to the pre-trained model, F, varies across the different training configurations, even for the same CNN. The variation in performance could be explained by stochastic variation caused by the use of SGD. The performance falls rapidly for $F \geq 5$; we therefore conclude that the first three or four convolutional layers transfer well from photos to artwork. Fine-tuning these layers yields no significant improvement nor detriment in performance. In this respect, we show similar results to Yosinski et al. (2014) [6] for our task: i.e. the first three or four convolutional layers are more transferable than later layers, in our case from photos to artwork.

For all later experiments, including the performance benchmarks, we select the configuration which maximises performance on the *validation set* (bold in Table 1) and re-train (fine-tune) using the combined *train and validation sets*.

5.2 Performance Benchmarks on the People-Art Dataset

Table 2 shows how each CNN model and other methods perform on the *People-Art test set*. The best performing CNN, VGG16, scores 58% AP, an improvement of 13 percentage points on the best previous result 45% [28]. The results demonstrate the benefits of fine-tuning the CNN (on the *training and validation sets* of *People-Art*) for the task. We also conclude that training and fine-tuning a CNN on photos yields a model which overfits to photographic images.

As noted in Section 4, Fast R-CNN (unlike YOLO) relies on an external algorithm, here selective search [22], to generate region proposals. We used the default settings, which are tuned to photos. Selective Search achieves a recall

Table 2. Performance of different methods on the test set of our *People-Art* dataset: the best performance is achieved using a CNN (Fast R-CNN) fine-tuned on *People-Art*

	datasets		
method	**pre-train**	**fine tuning**	**average precision**
Fast R-CNN (CaffeNet)	ImageNet	People-Art (train+val)	46%
Fast R-CNN (VGG1024)	ImageNet	People-Art (train+val)	51%
Fast R-CNN (VGG16)	ImageNet	People-Art (train+val)	59%
Fast R-CNN (CaffeNet)	ImageNet	VOC 2007	36%
Fast R-CNN (VGG1024)	ImageNet	VOC 2007	36%
Fast R-CNN (VGG16)	ImageNet	VOC 2007	43%
DPM [16]	People-Art	N/A	33%
YOLO [28]	ImageNet	VOC 2010	45%

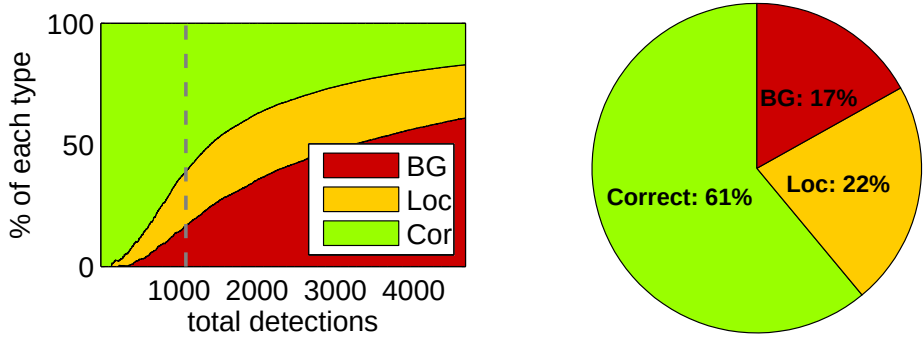

Fig. 4. Left: The proportion of detections by type as the threshold decreases: either correct, a background region (BG) or poor localisation (LOC); Right: the proportion for D=1088, the actual number of people, marked as a grey dashed line on the left plot

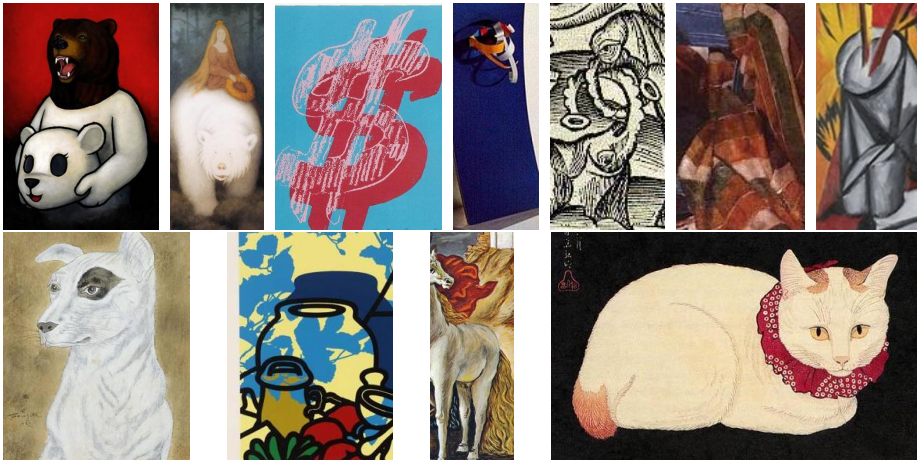

Fig. 5. False positive detections on background regions from the best performing CNN

Fig. 6. False positive detections due to poor localisation from the best performing CNN

rate of 98% on the *People-Art test set*. As such, this does not appear to be a limiting factor for the performance.

We attempted to fine-tune YOLO [28] on *People-Art*. The default configuration results in an exploding gradient, perhaps due to the sparsity of regions containing objects (only people in this case) compared to other datasets. We expect that a brute-force search over the parameters or heuristic may solve this problem in future work.

5.3 Detection Performance on People-Art

We used the tools of Hoiem et al. (2012) [49] to analyse the detection performance of the best performing CNN. Since we only have a single class (person), detections have three types based on their IoU with a ground truth labelling:

Cor correct i.e. $IoU \geq 0.5$
Loc false positive caused by poor localisation, $0.1 \leq IoU < 0.5$
BG a background region, $IoU < 0.1$

Figure 4 shows the detection trend: the proportion of detection types as the number of detections increases, i.e. from reducing the threshold. At higher thresholds, the majority of incorrect detections are caused by poor localisation; at lower thresholds, background regions dominate. In total, there are 1088 people labelled in the test set, and that are not labelled difficult. The graph in Figure 4 shows a grey dashed line corresponding to this number detections and Figure 4 shows a separate pie chart for this threshold. This threshold corresponding to this number of detections is significant: with perfect detection, there would be no false positives or false negatives. This shows that poor localisation is the bigger cause of false positives, though only slightly more so than background regions.

Figure 5 shows false positives caused by background regions. Some are caused by mammals which is understandable given these, like people, have faces and bodies. Others detections have less clear causes. Figure 6 show the false positives caused by poor localisation. In some of the cases, the poor localisation is caused by the presence of more than one person, which leads to the bounding box covering multiple people. In other cases, the bounding box does not cover the full extent of the person, i.e. it misses limbs or the lower torso. We believe that this shows the extent to which the range of poses makes detecting people in artwork a challenging problem.

5.4 Performance Benchmarks on the Picasso Dataset

In addition to the results on *People-Art*, we show results on the *Picasso Dataset* [5]. The dataset contains a set of Picasso paintings and labellings for people which are based on the median of the labellings given by multiple human participants. Table 3 shows how each CNN and other methods perform. As before, each CNN performed better if it was fine-tuned on *People-Art* rather than *VOC 2007*; moreover, DPM performs better than CNNs fine-tuned on *VOC 2007* but worse

Table 3. Performance of different methods on the *Picasso* dataset

method	training	fine tuning	average precision
Fast R-CNN (CaffeNet)	ImageNet	People-Art	45%
Fast R-CNN (VGG1024)	ImageNet	People-Art	44%
Fast R-CNN (VGG16)	ImageNet	People-Art	44%
Fast R-CNN (CaffeNet)	ImageNet	VOC 2007	29%
Fast R-CNN (VGG1024)	ImageNet	VOC 2007	37%
Fast R-CNN (VGG16)	ImageNet	VOC 2007	33%
DPM [16]	VOC 2007	N/A	38%
YOLO [28]	ImageNet	VOC 2012	53%

than those fine-tuned on *People-Art*. This confirms our earlier findings: CNNs fine-tuned on photos overfit to photo. In addition, we show that our fine-tuning results in a model which is not just better for *People-Art* but a dataset containing artwork which we did not train on.

Interestingly, the best performing CNN is the smallest (CaffeNet), suggesting that the CNNs may still be overfitting to less abstract artwork. Furthermore, the best performing method is YOLO despite being fine-tuned on photos (*VOC 2012*). Selective Search achieved a recall rate of 99% on the *Picasso Dataset*, so this is unlikely to be the reason that Fast R-CNN performs worse than YOLO. We therefore believe that YOLO's design is more robust to abstract forms of art.

5.5 The Importance of Global Structure

Earlier work [3, 50, 51] suggests that structure is invariant across depictive styles, and therefore useful for cross-depiction detection. As described in Section 4, Fast R-CNN includes an ROI pooling layer, which carries out max-pooling over $H \times W$ uniformly spaced rectangular grid. Therefore, the ROI pooling layer captures the global structure of the person, while earlier convolutional layers only pick up the local structure.

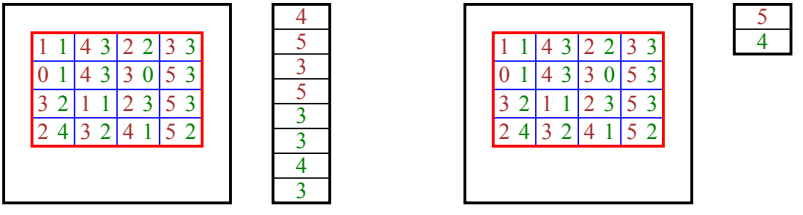

Fig. 7. Two pooling layers and their resulting feature vectors from a two channel input; Left: An ROI pooling layer (red grid) takes the maximum for each channel in each cell of an ROI (blue grid) resulting in an 8 dimensional vector; Right: A global max-pooling layer simply takes the maximum yielding a 2 dimensional vector

Table 4. Replacing the ROI pooling layer (default) with a single cell max-pooling layer yields a performance drop greater than not fine tuning *People-Art*

Fine-Tuning	People-Art		VOC 2007
ROI Pooling	default	single cell	default
CaffeNet	46%	34%	36%
VGG1024	51%	35%	36%
VGG16	59%	40%	43%

To test whether the *global structure* is useful for detecting and localising people in artwork, we replaced the ROI pooling layer replaced with a single cell max-pooling layer. This is equivalent to setting $W = 1$ and $H = 1$ for the ROI pooling layer (see Figure 7). This is similar to "bag of visual word" algorithms: with $W = H = 1$, the fully connected layers have no information about the location the previous layer's output. We fine-tuned as before.

Table 4 shows the results. In all cases, replacing the default ROI pooling layer with a single cell max-pooling layer results in worse performance. On top of this, the performance is worse than when fine-tuned on *VOC 2007* with the default configuration. This supports the claim of earlier work, that structure is invariant across depictive styles.

6 Conclusion

We have demonstrated state-of-the-art cross-depiction detection performance on our challenge dataset, *People-Art*, by fine-tuning a CNN for this task. In doing so, we have shown that a CNN trained on photograph alone overfits to photos, while fine-turning on artwork allows the CNN to better generalise to other styles of artwork. We have also made other observations, including the importance of negative exemplars from artwork.

The performance on our *People-Art* dataset, though the best so far, is still less than 60% AP. We have demonstrated that the CNN often detects other mammals instead of people or makes other spurious detections and often fails to localise people correctly. We propose further work to address these issues.

In addition, the dataset only covers a subset of possible images containing people. Our dataset does not include African, Babylonian, Chinese or Egyptian art, the Bayeux Tapestry, stained glass windows, photos of sculptures and all kinds of other possibilities. Therefore, we are only beginning to examine the cross-depiction problem, which provides a huge scope for further research.

Acknowledgements

This research was funded in part by EPSRC grant reference EP/K015966/1. This research made use of the Balena High Performance Computing Service at the University of Bath.

References

1. Hall, P., Cai, H., Wu, Q., Corradi, T.: Cross-depiction problem: Recognition and synthesis of photographs and artwork. Computational Visual Media **1**(2) (2015) 91–103
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 248–255
3. Wu, Q., Cai, H., Hall, P.: Learning graphs to model visual objects across different depictive styles. In: Computer Vision–ECCV 2014. Springer (2014) 313–328
4. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1440–1448
5. Ginosar, S., Haas, D., Brown, T., Malik, J.: Detecting people in cubist art. In: Computer Vision-ECCV 2014 Workshops, Springer (2014) 101–116
6. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems. (2014) 3320–3328
7. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics **36**(4) (1980) 193–202
8. Giebel, H.: Feature extraction and recognition of handwritten characters by homogeneous layers. In: Zeichenerkennung durch biologische und technische Systeme/Pattern Recognition in Biological and Technical Systems. Springer (1971) 162–169
9. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4) (1989) 541–551
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
11. Matan, O., Baird, H.S., Bromley, J., Burges, C.J., Denker, J.S., Jackel, L.D., Le Cun, Y., Pednault, E.P., Satterfield, W.D., Stenard, C.E., et al.: Reading handwritten digits: A zip code recognition system. Computer **25**(7) (1992) 59–63
12. Nowlan, S.J., Platt, J.C.: A convolutional neural network hand tracker. Advances in Neural Information Processing Systems (1995) 901–908
13. Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on **20**(1) (1998) 23–38
14. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 3626–3633
15. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 1., IEEE (2005) 886–893
16. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(9) (2010) 1627–1645
17. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. In: ICLR. (2014)

18. Vaillant, R., Monrocq, C., Le Cun, Y.: Original approach for the localisation of objects in images. IEE Proceedings-Vision, Image and Signal Processing **141**(4) (1994) 245–250
19. Szegedy, C., Toshev, A., Erhan, D.: Deep neural networks for object detection. In: Advances in Neural Information Processing Systems. (2013) 2553–2561
20. Schulz, H., Behnke, S.: Object-class segmentation using deep convolutional neural networks. In: Proceedings of the DAGM Workshop on New Challenges in Neural Computation, Citeseer (2011) 58–61
21. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 580–587
22. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision **104**(2) (2013) 154–171
23. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **37**(9) (2015) 1904–1916
24. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. Volume 2., IEEE (2005) 1458–1465
25. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 2., IEEE (2006) 2169–2178
26. Lenc, K., Vedaldi, A.: R-cnn minus r. arXiv preprint arXiv:1506.06981 (2015)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. (2015) 91–99
28. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640 (2015)
29. Huang, L., Yang, Y., Deng, Y., Yu, Y.: Densebox: Unifying landmark localization with end to end object detection. arXiv preprint arXiv:1509.04874 (2015)
30. Jacobs, C.E., Finkelstein, A., Salesin, D.H.: Fast multiresolution image querying. In: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, ACM (1995) 277–286
31. Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D., Jacobs, D.: A search engine for 3d models. ACM Transactions on Graphics (TOG) **22**(1) (2003) 83–105
32. Hu, R., Collomosse, J.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Computer Vision and Image Understanding **117**(7) (2013) 790–806
33. Canny, J.: A computational approach to edge detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on (6) (1986) 679–698
34. Wang, F., Kang, L., Li, Y.: Sketch-based 3d shape retrieval using convolutional neural networks. arXiv preprint arXiv:1504.03504 (2015)
35. Crowley, E.J., Parkhi, O.M., Zisserman, A.: Face painting: querying art with photos. In: British Machine Vision Conference. (2015)
36. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference. Volume 1. (2015) 6

37. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE (2007) 1–8
38. Chatfield, K., Philbin, J., Zisserman, A.: Efficient retrieval of deformable shape classes using local self-similarities. In: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, IEEE (2009) 264–271
39. Cho, M., Alahari, K., Ponce, J.: Learning graphs to match. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 25–32
40. Crowley, E.J., Zisserman, A.: In search of art. In: Workshop at the European Conference on Computer Vision, Springer (2014) 54–70
41. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2) (June 2010) 303–338
42. Willats, J.: Art and representation: New principles in the analysis of pictures. Princeton University Press (1997)
43. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
44. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings ofs the 27th International Conference on Machine Learning (ICML-10). (2010) 807–814
45. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1) (2014) 1929–1958
46. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014)
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
48. Everingham, M., Winn, J.: The pascal visual object classes challenge 2007 (voc2007) development kit. University of Leeds, Tech. Rep (2007)
49. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: European conference on computer vision, Springer (2012) 340–353
50. Xiao, B., Song, Y.Z., Balika, A., Hall, P.M.: Structure is a visual class invariant. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer (2008) 329–338
51. Xiao, B., Yi-Zhe, S., Hall, P.: Learning invariant structure for object identification by using graph methods. Computer Vision and Image Understanding **115**(7) (2011) 1023–1031