# Deep Learning Does Not Generalize Well to Recognizing Cats and Dogs in Chinese Paintings

Qianqian Gu[1(✉)] and Ross King[2]

[1] The University of Manchester, Manchester M13 9PL, UK
qianqian.gu@manchester.ac.uk
[2] The Alan Turing Institute, London NW1 2DB, UK

**Abstract.** Although Deep Learning (DL) image analysis has made recent rapid advances, it still has limitations that indicate that its approach differs significantly from human vision, e.g. the requirement for large training sets, and adversarial attacks. Here we show that DL also differs in failing to generalize well to Traditional Chinese Paintings (TCPs). We developed a new DL object detection method A-RPN (Assembled Region Proposal Network), which concatenates low-level visual information, and high-level semantic knowledge to reduce coarseness in region-based object detection. A-RPN significantly outperforms YOLO2 and Faster R-CNN on natural images ($P < 0.02$). We applied YOLO2, Faster R-CNN and A-RPN to TCPs with a 12.9%, 13.2% and 13.4% drop in mAP compared to natural images. There was little or no difference in recognizing humans, but a large drop in mAP for cats and dogs (27% & 31%), and very large drop for horses (35.9%). The abstract nature of TCPs may be responsible for DL poor performance.

**Keywords:** Traditional Chinese Paintings · Computational aesthetics · Deep Learning · Object recognition · Machine learning

## 1 Introduction

One of the greatest mysteries in cognitive science is the architecture of human visual system and its virtuosity in object recognition. Human have the impressive ability to recognize visually presented objects with both high speed and accuracy. Attempts to replicate this ability have been made since the start of Artificial Intelligence (AI) research, but these have met with only limited success. Object recognition is traditionally one of the most intractable problems in AI [1].

Recently, with the advent of very large annotated image databases, and advances in Deep Learning (DL), have greatly improved AI object recognition [2]. However, despite many claims to the opposite, DL object recognition is still not nearly as good as in humans. Example of the weaknesses of DL are: the requirement for large training sets, and their susceptibility to adversarial attacks that do not confuse humans. It is hypothesized that the reason for this


P. Kralj Novak et al. (Eds.): DS 2019, LNAI 11828, pp. 166–175, 2019.
https://doi.org/10.1007/978-3-030-33778-0_14

is compiled background knowledge about the world encoded in the human visual system architecture [3,4]. This knowledge was learnt through millions of years of Darwinian evolution.

Visual art is present in all human societies, and dates back as long ago as the Paleolithic. Computational visual aesthetics investigates the relationship between art and computational science. There is a close relationship between the human visual system and the appreciation of art (Gombrich (2000) Art and Illusion, Princeton University Press;) ***Computational Aesthetics*** is the computational investigation of aesthetics. Computational aesthetics is a large subject [5–7], but relatively little work has been done at the interface between computational aesthetics and DL object recognition [8,9]. The limited evidence, based on Western art, points to significant differences in performance between natural image trained classifiers and painting-trained classifiers. To the best of our knowledge the relationship and differences between object recognition in natural images and art has not been investigated.

Traditional Chinese Painting (TCP) is one of the great art traditions of the World (206 BC - now) [10]. Its unique style, which dates back over 2000 years, is instantly recognizable to humans. Here, to explore the differences between human and DL object recognition, we investigate DL object recognition in natural images and TCPs.

## 2 Object Detection in Chinese Painting

Object detection is one of the most widely studied topics in image processing. It differs from the classical image classification problems where models classify images into a single category corresponding to their most salient object. Here we investigate transferring Deep Learning (DL) object detection models from natural image to TCPs.

There are two potential challenges in applying DL to detecting objects in Chinese Paintings. First, there does not exist an official well-structured Chinese Painting Image Database and associated Ontology. And existing work in this area [11] is limited. Second, the characteristics of TCP images are different from natural images as they are produced using specialized tools, materials and techniques.

### 2.1 Chinese Painting and TCPs Data

Chinese painting can be categorized into two major schools of styles: XieYi, or GongBi [12]. XieYi (freehand strokes, Fig. 2) paintings are characterized by exaggerated forms, with the painted objects abstract and sometimes without fully connected edges. GongBi (skilled brush or meticulous approach, Fig. 1) paintings are characterized by close attention to detail and fine brushwork, and the painted objects are more realistic. TCPs can be further classified as: figure painting, landscapes painting, or flower-and-bird painting; only figure paintings and flower-and-bird paintings depict objects as their main subject.

To form our TCPs object detection dataset we integrated two datasets: a self-collected dataset obtained from online open sources, and a high-resolution image

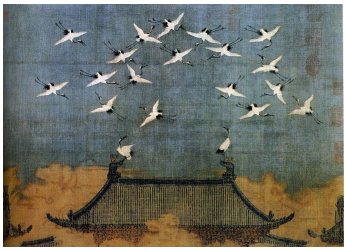

**Fig. 1.** Emperor Huizong Song's skilled brush flower-and-bird painting from Song Dynasty

**Fig. 2.** Beihong Xu's freehand strokes flower-and-bird painting

dataset provided by the Chinese Painting and Calligraphy Community. The sizes of these datasets are: 3,000 images, and 10,000 images respectively. The TCP objects dataset we used in this experiment contains 1,400 images in seven classes: human, horse, cow, bird, plant, cat, and dog. We manually annotated the data with class labels and segmentation knowledge (bounding box), as ground truth. (List of paintings we have used could be found: https://github.com/hiris1228/TCP_object_detect.git).

### 2.2 Object Detection

Current DL object detection models generally identify various objects and location within one single image. A naive approach to solving this problem is to take different regions of interest from the image, and to classify the presence of the object within that region – usually using a ConvNets [13,14]. The R-CNN (Regions + CNN) method was developed to avoid the use of excessive number of regions in object detection. This relies on an external region proposal system based on a selective search algorithm [15–17]. The cost-reduced computational shared model R-FCN [18] attempts to balance translation-invariance in image classification, and translation translation-variance in object detection. The Mask R-CNN approach [19] applies FCN (Fully Convolutional Networks) [20] to each Region of Interest (RoI), and then applies classification and bounding box regression in parallel. Single Shot detectors like YOLO [21] and SSD (Single Shot MultiBox Detector [22]) and are able to obtain relatively good results.

## 3 Chinese Painting Object Detection Approach

The architecture of our TCPs Object Detector A-RPN is a VGG-16 [23] backbone base model as shown in Fig. 3. We adopted the popular two-stage object detection strategy of Faster R-CNN [17]. We generate a set of RoIs, and a classification model independent of selected RoIs. Comparing to ResNet [23], VGG-16 is not very 'deep' (ResNet has 152 layers), but its depth is appropriate given the limited TCP data.

The assembled Region Proposal Network (A-RPN) has three components: a general RPN that return RoIs and Intersection of Union (IoU) scores; a Chinese-Kitten RPN (CN-Kitten RPN), which concatenate both lower-level and high-level features to generate small object sensitive proposals, and refines the RoIs; and a Detection Network R-FCN that takes proposed RoIs, and classifies them into categories and background. All the sub-networks within the A-RPN are fully convolutional with respect to RoIs. All three components are initialized by a hierarchical feature map *M0* through shared convolutional layers to form a pre-trained resembling VGG-16.

The region proposals (RoIs) processed in our A-RPN are divided into two parts. One comes from the generated RPN (RoIs-set1), the other is the small-scale region boxes generated by sliding window run on convCNK (RoIs-set2). The same NMS operation, with threshold 0.7, was applied to both general RPN and CN-Kitten RPN to reduce redundancy. This process returned two set of 2k proposals. Each of these regions has an Intersection of Union (IoU) score, which estimates the chance that the current RoI contains an object.

CN-Kitten RPN is an enhanced RPN that combines multi-layer features knowledge and RoI pooling layer to refine RoI proposals. This approach is designed to increase the model's sensitivity to small-scale object detection from knowledge of Feature Pyramid Networks [24]. With differing low-level visual information, the high-level convolutional features may be too coarse when we project our RoIs from the feature map to the original image. The aim is to better fit small objects, and to better utilize fine-grained features due to TCPs characteristics. We observed no significant drop in recall with small objects.

RoIs-set2 is generated from a $1\times1\times256$ fully-connected layer which is reduced from fine-grained feature map convCNK. convCNK is the multi-layer feature concatenation of conv5 and conv4 which is similar FCN-16s [20]. We apply an unsampling filter on conv5 to obtain transposed convolutional layer [25] conv5'. Then L2-normalize each layer per spatial location, and re-scales it with the same resolution as conv4 to obtain convCNK. Simultaneously, we applied a bilinear interpretation on convCNK, and obtained a semantic segmentation heat-map of the entire image for later refinement.

In order to obtain the final top 300 proposals as input of Detection Network R-FCN. A top rank voting model was applied to both RoI-set1 and RoI-set2. This ranks all anchors with their IoU scores from the highest to the lowest. It then projects all the selected anchors to the semantic segmentation heat-map, and computes an extra IoU score with the overlapped percentage. Then it looks to find whether there are existing anchors in its opposite set that have an overlapped rate higher than 0.7, if yes, it merges these two anchors by averaging their coordinates. The method repeats this procedure until it obtains proposals with the highest 300 pairs of IoU scores.

The detection network identifies and regresses the bounding box of regions likely to contain classes base on each proposed region. Unlike in the original Faster R-CNN [17], we apply Fast R-CNN [16] with R-FCN [18] as the detection network. As a member of FCN [20] family, it construct a set of position-sensitive

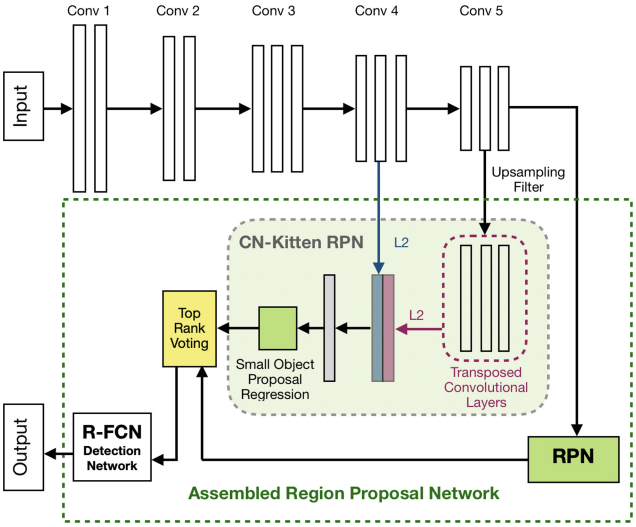


**Fig. 3.** Assembled Region Proposal Network (A-RPN) architecture.

score maps to incorporate translation variance. As there are no learning layers after the RoI layer, only a shared fully convolutional architecture, R-FCN is nearly cost-free, and even faster than the other pioneering DL image classification architectures [13,15–17,23] with a competitive mAP (mean Average Precision).

R-FCN uses a bank of specialized convolutional layers to encode as score maps position information with respect to a relative spatial position [18]. Our Detection network removes all FC layers, and computes all learn-able weight layers on the entire image to create a bank of position-sensitive score maps for C + 1 categories (C object categories + 1 background). Each set of score maps for one particular category represents a $k \times k$ spatial grid describing relative position information. Selective pooling only returns one score out of $k \times k$ on class prediction after performing average voting on these shared sets of score maps.

## 4 Experiment and Results

To investigate the differences between DL and human image analysis we applied DL to images that humans can easily interpret: natural image and Traditional Chinese Paintings (TCPs). We compared our A-RPN's outputs with two popular DL object detection models. We ran six experiments: Single Shot detectors on natural images, YOLO(N), and on Chinese paintings, YOLO(P); Faster R-CNN on natural images, Faster R-CNN(N), and on Chinese paintings, Faster R-CNN(P); A-RPN on natural images, A-RPN(N), and on Chinese paintings, A-RPN(P). The natural image dataset that we used was PASCAL VOC2007 trainval. Both datasets contains the same seven classes. We kept one-third of the data as test set, and divided the remaining data into 90% training set and 10% validation set. The results are shown in Table 1.

**Table 1.** The mAP Comparison for Flower-and-Bird and Figure Classification. And A-RPN has proved that it outperforms the other two models.

| Methods | YOLO(N) | YOLO(P) | Faster RCNN(N) | Faster RCNN(P) | ARPN(N) | ARPN(P) |
|---|---|---|---|---|---|---|
| Human | 82.1 | 76.3 | 81.1 | 79.5 | 83.2 | 82.4 |
| Horse | 71.8 | 28.3 | 75.9 | 39.7 | 78.3 | 42.4 |
| Cow | 75.4 | 69.1 | 76.4 | 74.3 | 78.7 | 80.5 |
| Bird | 81.1 | 75.2 | 79.3 | 77.7 | 76.8 | 70.0 |
| Plant | 32.0 | 33.8 | 36.2 | 42.9 | 43.5 | 50.0 |
| Cat | 78.7 | 65.1 | 81.3 | 52.0 | 82.7 | 55.5 |
| Dog | 79.2 | 61.6 | 82.4 | 53.8 | 83.6 | 52.2 |
| Overall | 71.37 | 58.48 | 73.22 | 59.98 | **75.25** | **61.85** |

Table 1 shows that all the DL models can achieve more than 70% mAP on natural image object detection. A-RPN achieved the best performance in classification, and has significantly higher mAP ($P < 0.02$) than YOLO2 and Faster R-CNN. (Using the McNemar test, the Z-model statistic against YOLO2 and Faster R-CNN are respectively $-3.7905$ and $-2.4188$). Object recognition performance significantly drops when applied to TCPs. The YOLO2 model has a mAP performance drop of 12.9% while Faster R-CNN drops 13.2% and A-RPN drops 13.4%. The statistical difference in performance between the three methods on TCP object recognition is not significant.

The success of object recognition varies greatly between object class. The performance on classes 'Human' and 'Bird' is stable and accurate for all methods. Performance on the class 'Plant' is stable but has low mAP in all models. Class 'Cow' works better in natural image dataset than TCPs, but the drop is relatively small, while A-RPN increases with 1.8% mAP. All three models confuse the 'Cat' and 'Dog' classes in TCPs. For the class 'Cat' the decreases in performance were 13.6%, 29.3%, and 27.2% for A-RPN. For the class 'Dog' the decreases in performance were 17.6%, 28.6%, and 31.4%. The largest drop in performance was for the class 'Horse' in TCPs. Only A-RPN could achieve more than 40% mAP, while their mAPs on Nature Image data are all above 70%.

Table 2 and Fig. 4 show that our region proposals network has limitations when allocating object bounding boxes. There are 5 classes (out of 7 in total) that failed to achieve 85% detecting rate of one particular type of object. Intersections of Unions (IoU) is the region of interest union with the ground truth bounding box. Our threshold was set as 0.5. But after verifying, we noticed that only 63% of the ground truth bounding boxes were detected with ratios above 0.5. 22% of the objects were not detected, especially in the plant class, and 15% of all bounding box which does not cover objects.

**Table 2.** Heat-Map for A-RPN classification. Miss-classification scenarios with error more than 15% have been highlighted in the table. The N/A column represents the scenario that – there is no bounding box detected over the current ground truth bounding box.

| Truth | Prediction | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Human | Horse | Cow | Bird | Plant | Cat | Dog | N/A |
| Human | 0.82 | 0 | 0 | 0 | 0 | 0 | 0 | **0.18** |
| Horse | 0 | 0.35 | 0.03 | **0.18** | 0.02 | 0 | **0.27** | **0.15** |
| Cow | 0 | 0.09 | 0.72 | 0 | 0 | 0 | 0.01 | **0.18** |
| Bird | 0 | 0 | 0 | 0.78 | 0.01 | 0.03 | 0 | **0.18** |
| Plant | 0 | 0 | 0 | 0.04 | 0.38 | 0.01 | 0.01 | **0.56** |
| Cat | 0 | 0 | 0 | 0.02 | 0.01 | 0.51 | **0.38** | 0.08 |
| Dog | 0 | 0.01 | 0.09 | 0 | 0 | **0.29** | 0.51 | 0.10 |

## 5 Discussion and Future Works

Putting to one side the differences between human and DL image recognition, it is interesting to consider whether in principle images in TCPs are harder to recognize using DL. We have shown A-RPN performs slightly better than YOLO and Faster R-CNN, but its success in natural image cannot be transferred to TCPs. There are a number of possible reasons for the drop in performance of DL in TCPs. First, DL models require large training sets, but the number of TCPs is quite limited, and their usage generally involve licenses. Therefore the limited size of the data set we used potentially prevented the DL models being fully trained.

Another reason may be the initialization of the layer M0 feature map. When the A-RPN is initialized, we use a pre-trained CNN layer trained on natural images. Natural images differ in a number of ways from images in TCPs. They inherently include perspective, while most Chinese paintings do not. In TCPs objects are often depicted in a highly abstract manner, easily comprehended by the human, but quite different from natural images.

Ineffective feature formation may also be a factor. CNN models have their own texture representations [23,26]. Whether the work was 'linear' (contour-led) or 'painterly' (reliant more on brushstrokes denoting light and shadow) [27] has been proven to have effect on prediction. The majority of Chinese paintings are only black and white. In CNN models white color is often treated as 'no color', with a probability of being an object of zero, and the 'black ink' textures are similar in every single class.

We hypothesize that the abstract nature of TCPs may fundamentally restrict the ability of DL systems to recognize objects in TCPs. Objects in TCPs do not have fully connected edges. When CNN edge detection filters are applied on non-edge pixels in low-level layers, the resulting matrices will be filled with really small numbers or even zeros. In CNN texture representation strategy, the

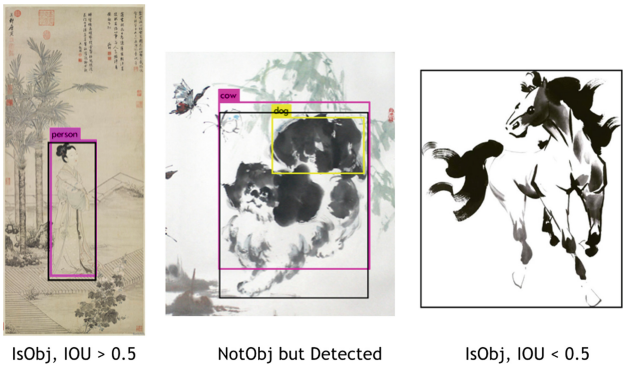


**Fig. 4.** Bounding Box Detection Cases: Left, object detected with IoU score greater than 0.5; Middle, object detected with IoU less than threshold; Right, object not detected.

smaller the edge detection matrix is, the lower the convolution value will be. And this situation gets worse during strided convolutions. Finally, CNNs weakens the contribution from ambiguous edge pixels, and further separates edges that should be treated as connected in TCPs. This complicates object detection task in TCPs, and is also the reason why we absorbed the concept from FPN [24] to assemble features from different convolutional layers to try to maximize the sensitivity on learning edge information.

Many principles of Chinese Painting are derived from Daoism [28]. For example in Daoism empty space is an important concept, and a symbol of the void or nothingness. The most important text in Daoism states: 'Having and not having arise together' (Laozi 2). Chinese Painting has also been influenced by Buddhism, which emphasizes that 'What is form that is emptiness, what is emptiness that is form' (Paramita Hridaya Sutra). These beliefs have led Chinese paintings to stress the concept of Designing White Space — if one's mind can reach there, there is no need for the touch of any brush and 'formless is the image grand' (Laozi 4). Similarly, an important canon of Chinese painting describes its rhythmic vitality as Qi, a metaphysical concept of cosmic power: with Qi empty space is not blank, it is alive, like air. This prominent characteristic of Chinese painting turns its treatment of empty space as solid space. Taking the horse in XieYi style in Fig. 2 as an example, the absence of clear edges depicting the outline of a horse feed the human imagination.

## 6 Conclusion

We have demonstrated that state-of-the-art object DL recognition methods perform substantially worse on TCPs compared to natural images. We argue that these results point to interesting and important differences between DL systems and the human visual system that require further investigation. For example:

what features of the human visual system that enable it to effortlessly recognize cats and horses in Chinese paintings, deal with the white space, etc.; What is the relationship between aesthetics and image recognition? We hope, and expect, that further research will investigate these important questions, and lead to computational systems that can recognize objects in Chinese paintings as well as humans, and perhaps also appreciate their beauty as greatly as humans.

## References


1. Grill-Spector, K., Kourtzi, Z., Kanwisher, N.: The lateral occipital complex and its role in object recognition. Vis. Res. **41**(10–11), 1409–1422 (2001)
2. Schrimpf, M., et al.: Brain-score: which artificial neural network for object recognition is most brain-like? BioRxiv, p. 407007 (2018)
3. Grill-Spector, K., Weiner, K.S.: The functional architecture of the ventral temporal cortex and its role in categorization. Nat. Rev. Neurosci. **15**(8), 536 (2014)
4. Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A.: Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci. Rep. **6**, 27755 (2016)
5. Birkhoff, G.D.: Aesthetic Measure. Harvard University Press, Cambridge (1933)
6. Greenfield, G.: On the origins of the term computational aesthetics (2005)
7. Neumann, L., Sbert, M., Gooch, B., Purgathofer, W., et al.: Defining computational aesthetics. In: Computational Aesthetics in Graphics, Visualization and Imaging, pp. 13–18 (2005)
8. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint: arXiv:1508.06576 (2015)
9. Crowley, E.J., Zisserman, A.: The art of detection. In: Hua, G., Jégou, H. (eds.) ECCV 2016, Part I. LNCS, vol. 9913, pp. 721–737. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46604-0_50
10. Lanciotti, L.: The British museum book of Chinese art (1993)
11. Wu, X., Li, G., Liang, Y.: Modeling Chinese painting images based on ontology. In: 2013 International Conference on Information Technology and Applications, pp. 113–116. IEEE (2013)
12. Jiang, S., Huang, T.: Categorizing traditional Chinese painting images. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004, Part I. LNCS, vol. 3331, pp. 1–8. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30541-5_1
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
14. LeCun, Y., et al.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **1**(4), 541–551 (1989)
15. 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 (CVPR), pp. 580–587 (2014)
16. Girshick, R.: Fast R-CNN. In: The IEEE International Conference on Computer Vision (ICCV), December 2015
17. 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, pp. 91–99 (2015)

18. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, pp. 379–387 (2016)
19. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2961–2969 (2017)
20. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015)
21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
22. Liu, W., et al.: SSD: Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part I. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint: arXiv:1409.1556 (2014)
24. Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
25. Zeiler, M.D., Taylor, G.W., Fergus, R., et al.: Adaptive deconvolutional networks for mid and high level feature learning. In: The IEEE International Conference on Computer Vision (ICCV), vol. 1, p. 6 (2011)
26. Lin, T.-Y., Maji, S.: Visualizing and understanding deep texture representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2791–2799 (2016)
27. Elgammal, A., Liu, B., Kim, D., Elhoseiny, M., Mazzone, M.: The shape of art history in the eyes of the machine. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
28. Shaw, M.: Buddhist and taoist influences on Chinese landscape painting. J. Hist. Ideas **49**(2), 183–206 (1988)