Research Statement

Kai Zhao, PhD, CS Dept. Nankai University

http://kaizhao.net kaiz.xyz@gmail.com

1 Research Statement

My research aims to build machines to perceive the visual world as humans do. This is nontrivial because visual signals are highly unstructured and present various deformation, occlusion and other disturbance. Data-driven learning is an efficient way of building visual intelligence because we humans set up our visual system by watching a large number of scenes every day. However, there is no such model as the human visual system that can handle various circumstances. We have to design variant models in case of different tasks and data types. Human expertise plays an important role in model designing. For example, convolution and pooling are typically used in CNNs to process spatially structured data, because they are shift-invariant and can capture spatial correspondence. While the recurrent operator is used to handle time series like human language and voice because these signals present temporal structures.

2 Past Research

During the past few years, my research mainly focuses on **embedding human priors into statistical models by designing novel CNN architectures and optimization objectives**.

2.1 Embed human prior through optimization objectives

Learning discriminative representations is critical to accurate face recognition. That is, representations of facial images from the same identity (person) should be close in the embedding space; while images of different identities should be far away. Based on this heuristic, I proposed a regularization term, named *exclusive regularization* [8], to explicitly enlarge the distance between features of different identities (Fig.1).

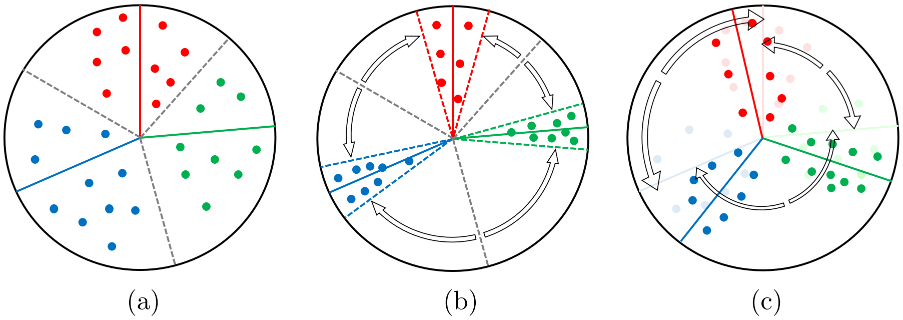

Figure 1: (a) Separateable representations and (b) representations with intra-class compactness. (c) RegularFace [8] (right) explicitly enlarges the distance between representations of different identities (distinguished by colors), leading to discriminative feature embeddings.

Salient object detection (SOD) aims to segment the outstanding area of an image. This task can be regarded as a binary pixel-wise classification problem. Most CNN-based SOD methods use the cross-entropy loss as loss function. And the detection results are evaluated in terms of F-measure. The gradients of cross-entropy loss get smaller when predictions are approaching the target, leading to blurry results especially on the edge of the salient area. I propose to use a unified criterion in both training and evaluation for SOD [6]. By relaxing the indifferentiable F-measure formulation into a continuous and differentiable function, [6] directly maximizes F-measure during training. Meanwhile, the proposed optimization objective holds considerable gradients even in the saturated area, leading to high-contrast predictions.

In the above two representative works [8, 6] I embedded human priors (I, representations of different identities should stand far away for discriminative feature learning and; II, the salient area in an image should be contrastive with others) into the model by designing novel and task-associated optimization objectives.

2.2 Embed human prior through network architectures

Skeleton, aka the medial axis, is a compact description of object topology and geometry. Efficient skeleton detection requires the detector to capture various sizes of context because the skeleton presents various scales, as shown in Fig. 2. Convolutional layers are stacked sequentially in CNNs. Shallower layers with smaller receptive fields capture low-level textures, while deeper layers with larger receptive fields capture high-level semantics. In [4] I propose to use convolutional features from different levels as skeleton detectors at different scales. Shallow layers are responsible to detect thin skeleton while deeper layers detect thick skeleton. Then in [3] I extended the method of [4]. The extended version can simultaneously predict object skeleton and skeleton scales. We can recover the object segment through the skeleton and skeleton scales.

Later in [7] I proposed to extract richer convolutional features through hierarchical feature integration (Hi-Fi), with further performance improvements. The Hi-Fi architecture enriches feature representations by gradually fusing convolutional features of adjacent layers. According to our experiments, this feature integration mechanism also improves the performance of edge detection.

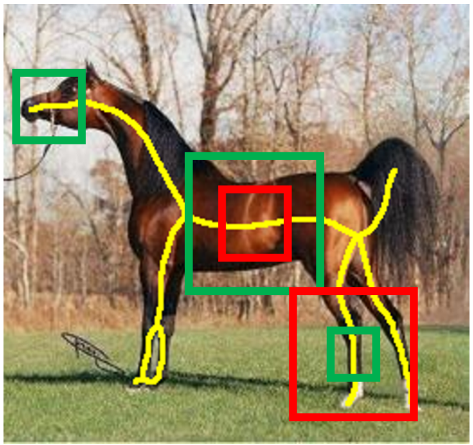

Figure 2: Object skeleton (yellow curves) presents variant scales. Only detecting windows (green boxes) at proper size can capture enough context for successful skeleton detection.

2.3 Combine random forest and DNN

Another previous work [1, 2] tries to combine the feature learning ability of neural networks and the decision-making ability of decision trees. In this work we combine the CNN and random forest in an end-to-end manner: the CNN extracts image features, afterward, the random forest makes decisions using the extracted features. The overall system is trained in an alternative way: we fix CNN parameters W and update the forest parameter Θ, and then vice versa. The proposed method is applied to multiple tasks including gene series classification and facial age estimation. The tree routing part and associated back-propagation part are implemented in pure CUDA. I learned CUDA programming through this project.

3 Future Research Plan

3.1 Semisupervised and unsupervised learning

Humans learn to perceive the visual world mostly by self-supervised learning. While today's state-of-art machine vision system relies on large amount of annotated training images. The need for large amounts of annotated data has been a bottleneck to building large scale machine visual system that performs human-level visual intelligence. Unsupervised or supervised learning is a prospective way of breaking the limitation of human supervision. Sometimes the data contains inherent structures and redundant information. For example, objects in videos always change smoothly along with time. Therefore, dense annotation of object location is unnecessary for training a tracker. And when tracking an object along with time, by reversing the video the tracker should go back to the origin. The use of inherent time-consistency in video data can help unsupervised learning of pixel correspondence [5].

3.2 Light-weight CNN design

Deep CNNs are typically over-parameterized, leading to large computational overhead and memory footprint in inference. In my future research, I would try to design light-weight network architectures with limited inference costs. Network pruning and automatically architecture search may help to find a optimal network architecture with minimal computational cost and neglectable performance drop. Currently, I have an ongoing work to prune less important filters in a pretrained network.

References

[1] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan Loddon Yuille. Deep differentiable random forests for age estimation. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2, 3

[2] Wei Shen, Kai Zhao, Yilu Guo, and Alan Yuille. Label distribution learning forests. In *Proceedings of Advances in neural information processing systems*, 2017. 2, 3

[3] Wei Shen, Kai Zhao, Yuan Jiang, Yan Wang, Xiang Bai, and Alan Yuille. Deepskeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images. *IEEE Transactions on Image Processing*, 26(11):5298–5311, 2017. 2, 3

[4] Wei Shen, Kai Zhao, Yuan Jiang, Yan Wang, Zhijiang Zhang, and Xiang Bai. Object skeleton extraction in natural images by fusing scale-associated deep side outputs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 222–230. IEEE, 2016. 2, 3

[5] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 2

[6] Kai Zhao, Shanghua Gao, Wenguan Wang, and Ming ming Cheng. Optimizing the F-measure for threshold-free salient object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 1, 3

[7] Kai Zhao, Wei Shen, Shanghua Gao, Dandan Li, and Ming-Ming Cheng. Hi-Fi: Hierarchical feature integration for skeleton detection. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1191–1197. International Joint Conferences on Artificial Intelligence Organization, 7 2018. 2, 3

[8] Kai Zhao, Jingyi Xu, and Ming-Ming Cheng. Regularface: Deep face recognition via exclusive regularization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 3

[1, 2, 3, 4] are done when I was a master student at Shanghai University. *Wei Shen*[1] is my master's supervisor. [6, 7, 8] are done during my Ph.D. study at Nankai University. My Ph.D. supervisor is prof Ming-ming Cheng [2].

[1] Dr. Wei. Shen is now a full-time faculty at Johns Hopkins University

[2] https://scholar.google.com/citations?user=huWpVyEAAAAJ&hl=en