Deep LAC: Deep Localization, Alignment and Classification for Fine-grained Recognition

Di Lin†, Xiaoyong Shen†, Cewu Lu‡, Jiaya Jia†
† The Chinese University of Hong Kong. ‡ The Hong Kong University of Science and Technology.

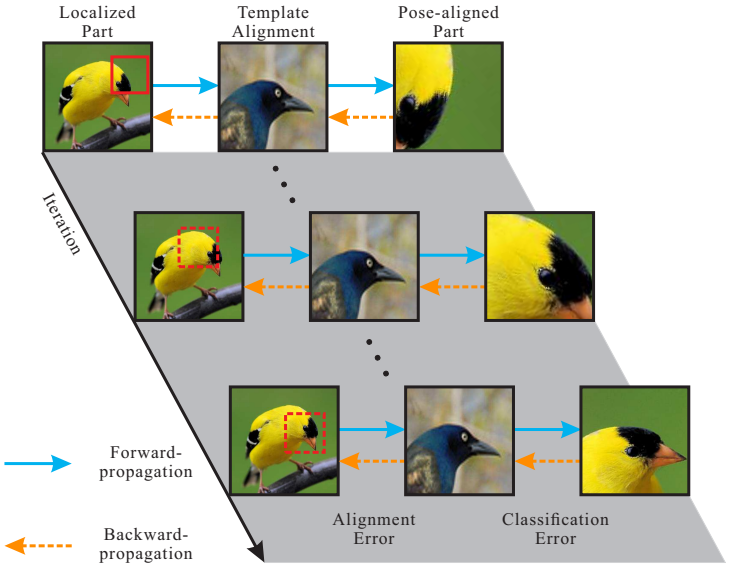

Figure 1: The one-way procedure from localization to template alignment makes each module rely on results from the previous one. Contrarily, back-propagation highlighted by dashed arrow makes it possible to refine localization according to the classification and alignment results. It forms a bi-directional refinement process.

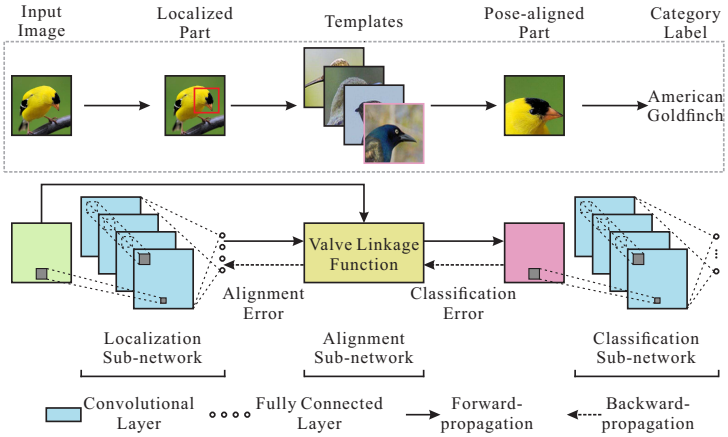

Figure 2: Deep LAC. It consists of localization, alignment and classification sub-networks. With the help of VLF, alignment sub-network outputs pose-aligned part image for classification sub-one in FP stage, while classification and alignment errors can be propagated back to localization sub-one in the BP stage.

Fine-grained object recognition aims to identify sub-category object classes, which includes finding subtle difference among species of animals, product brands, and even architectural styles. Thanks to recent success of convolutional neural networks (CNN) [5], good performance was achieved on fine-grained tasks [9]. However, existing solutions perform localization, alignment, and classification independently and consecutively. This procedure is illustrated in Figure 1 using solid-line arrows where parts are localized, aligned according to templates, and then fed into the classification neural network. Obviously, any error arising during localization could influence alignment and classification, which makes the fine-grained recognition still have much room to improve.

In this paper, we propose a feedback-control framework to back-propagate alignment and classification errors to localization, in order to optimally update all states in each iteration. This process is highlighted by dashed arrows in Figure 1. A valve linkage function (VLF) is proposed in the template alignment sub-network to optimally connect the localization and classification modules in our deep localization, alignment and classification (LAC) system. The architecture of our deep LAC system is shown in Figure 2. In FP, VLF outputs a pose-aligned part image to classification. In BP, it should be a function containing necessary parameters for updating localization sub-network. Therefore, our VLF not only connects all sub-networks, but also functions as information valve to compromise classification and alignment errors. If alignment is good enough in the FP stage, VLF guarantees corresponding accurate classification. Otherwise, errors propagated from classification finely tune the previous modules. These effects make the whole network quickly reach a stable state.

We apply our method to the widely employed CUB-200-2011 datasets [7] for automatic classification. The results are listed in Table 1.

Methods	Accuracy
Lee et al. [6]	41.01%
Berg et al. [1]	56.89%
Goering et al. [4]	57.84%
Chai et al. [2]	59.40%
Gravves et al. [3]	62.70%
Zhang et al. [8]	64.96%
Zhang et al. [9]	76.37%
Ours (head)	72.00%
Ours (body)	52.65%
Ours (head+body)	78.12%
Whole image	65.00%
Ours (head+body) + whole image	**80.26%**

Table 1: Comparison with state-of-the-arts on the CUB-200-2011 dataset.

[1] Thomas Berg and Peter N Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013.

[2] Yuning Chai, Victor Lempitsky, and Andrew Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *ICCV*, 2013.

[3] Efstratios Gavves, Basura Fernando, CGM Snoek, AWM Smeulders, and Tinne Tuytelaars. Fine-grained categorization by alignments. In *ICCV*, 2013.

[4] Christoph Goering, Erik Rodner, Alexander Freytag, and Joachim Denzler. Nonparametric part transfer for fine-grained recognition. In *CVPR*, 2014.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[6] Yong Jae Lee, Alexei A Efros, and Martial Hebert. Style-aware mid-level representation for discovering visual connections in space and time. In *ICCV*, 2013.

[7] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[8] Ning Zhang, Ryan Farrell, Forrest Iandola, and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 2013.

[9] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.