Look-into-Object: Self-supervised Structure Modeling for Object Recognition

Mohan Zhou[1,2*], Yalong Bai[2*], Wei Zhang[2†], Tiejun Zhao[1], and Tao Mei[2]

[1]Harbin Institute of Technology
[2]JD AI Research, Beijing, China
{mhzhou99, ylbai}@outlook.com wzhang.cu@gmail.com tjzhao@hit.edu.cn tmei@jd.com

Abstract

Most object recognition approaches predominantly focus on learning discriminative visual patterns while overlooking the holistic object structure. Though important, structure modeling usually requires significant manual annotations and therefore is labor-intensive. In this paper, we propose to "look into object" (explicitly yet intrinsically model the object structure) through incorporating self-supervisions into the traditional framework. We show the recognition backbone can be substantially enhanced for more robust representation learning, without any cost of extra annotation and inference speed. Specifically, we first propose an object-extent learning module for localizing the object according to the visual patterns shared among the instances in the same category. We then design a spatial context learning module for modeling the internal structures of the object, through predicting the relative positions within the extent. These two modules can be easily plugged into any backbone networks during training and detached at inference time. Extensive experiments show that our look-into-object approach (LIO) achieves large performance gain on a number of benchmarks, including generic object recognition (ImageNet) and fine-grained object recognition tasks (CUB, Cars, Aircraft). We also show that this learning paradigm is highly generalizable to other tasks such as object detection and segmentation (MS COCO). Project page: https://github.com/JDAI-CV/LIO.

1. Introduction

Object recognition is one of the most fundamental tasks in computer vision, which has achieved steady progress with the efforts from deep neural network design and abundant data annotations. However, recognizing visually similar objects is still challenging in practical applications, especially when there exist diverse visual appearances, poses, background clutter, and so on.

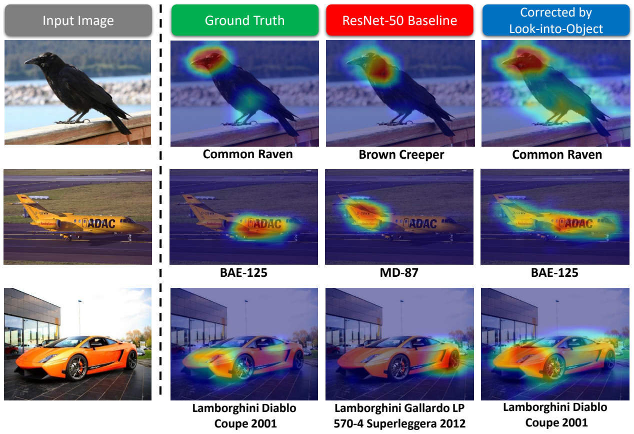
Input Image
Ground Truth
ResNet-50 Baseline
Corrected by Look-into-Object
Common Raven
Brown Creeper
Common Raven
BAE-125
MD-87
BAE-125
Lamborghini Diablo Coupe 2001
Lamborghini Gallardo LP 570-4 Superleggera 2012
Lamborghini Diablo Coupe 2001

Figure 1. Feature map visualization based on the last convolutional layer of ResNet-50 backbone. The first column shows the original images, while the second and the third columns show the maximally responding feature maps from the ground-truth and the predicted labels, respectively. The last column shows the feature maps by plugging our proposed LIO on ResNet-50. Object extend and discriminative regions are all correctly localized owing to the holistic structure modeling. (Best viewed in color).

Suffering from complex visual appearance, it is not always reliable to correctly recognize objects purely based on discriminative regions, even with a large-scale human-labeled dataset. As shown in Fig. 1, a well-trained ResNet-50 (the third column) can still misclassify objects by looking at the wrong parts.

Existing object recognition approaches can be roughly grouped into two groups. One group optimizes the network architecture to learn high-quality representations [29, 21, 15, 7], while the other line of research introduces extra modules to highlight the salient parts explicitly (by bounding-box [2, 16, 18]) or implicitly (by attention [11, 36]). Apparently, the latter one costs more on either annotation (*e.g.* bounding boxes / part locations) or calculation (attentions /

[*] Equal contribution. This work was done at JD AI research.
[†] Corresponding author.

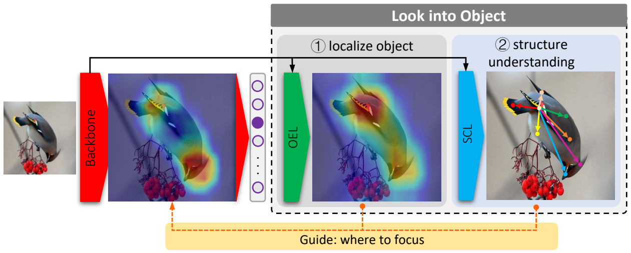

Figure 2. Our proposed Look-into-Object (LIO) approach. Object Extent Learning (OEL) and Spatial Context Learning (SCL) enforce the backbone to learn object extent and internal structure respectively.

detection modules). However, all these methods predominantly focus on learning salient patterns while ignoring the holistic structural composition.

In this paper, we argue that correctly identifying discriminative regions largely depends on the holistic structure of objects. Traditional deep learning based methods can be easily fooled in many cases, e.g., distinguishing front and rear tires of a car, localizing legs of a bird among twigs. It is mainly due to the lack of cognitive ability for structures of objects. Therefore, it is crucial to learn the structure of objects beyond simple visual patterns. Though important, it still remains challenging to systematically learn the object structural composition, especially without additional annotation and extra inference time cost.

In this work, we propose to model the holistic object structure without additional annotation and extra inference time. Specifically, we propose to "look-into-objects" (in short "LIO") to understand the object structure in images by automatically modeling the context information among regions. From the psychological point of view, recognizing an object can be naturally regarded into two stages: 1) roughly localizing the object extent (the whole extent of the object rather than object part) in the image, and 2) parsing the structure among parts within the object.

Accordingly, we design two modules to mimic such a psychological process of object recognition. We propose a novel and generic scheme for object recognition by embedding two additional modules into a traditional backbone network, as shown in Fig. 2. The first one is *Object-Extent Learning Module* (OEL) for object extent localization, while the second is *Spatial Context Learning Module* (SCL) for structure learning within the object.

Naturally, a prerequisite for object structure modeling is that the object extent can be localized. The OEL module enforces the backbone to learn object extent using a pseudo mask. We first measure the region-level correlation between the target image and other positive images in the same category. The regions belonging to the main object would have high correlations, owing to the commonality among images from the same category. As a result, a pseudo mask of object extent can be constructed according to the correlation scores without additional annotation besides the original image labels. Then, the backbone network is trained to regress the pseudo mask for localizing the object. With the end-to-end training, the capacity of object-extent localization for backbone network can be further reinforced.

The SCL module predicts the spatial relationships among regions within the object extent in a self-supervised manner. Given the localized extent learned by the OEL module, the SCL mainly focuses on the internal structure among regions. Specifically, we enforce the backbone network to predict the relative polar coordinates among pairs of regions, as shown in Fig. 2. In this way, the structural composition of object parts can be modeled. This self-supervised signal can benefit the classification network for the object structure understanding by end-to-end training. Obviously, localize the discriminative regions in a well-parsed structure is much easier than in the raw feature maps.

Note that all these modules take the feature representations generated by the classification backbone network as input and operate at a regional level, which leads to a delicate *Look-into-Object* (LIO) framework. Training with such objectives enforces the feature learning of the backbone network by the end-to-end back-propagation. Ideally, both object extent and structure information can be injected into the backbone network to improve object recognition without additional annotations. Furthermore, both modules can be disabled during inference time.

The main contributions can be summarized as follows:

1. A generic LIO paradigm with two novel modules: object-extent learning for object-extent localization, and self-supervised spatial context learning module for modeling object structural compositions.

2. Experimental results on generic object recognition, fine-grained recognition, object detection, and semantic segmentation tasks demonstrate the effectiveness and generalization ability of LIO.

3. From the perspective of practical application, our proposed methods do not need additional annotation and introduce no computational overhead at inference time. Moreover, the proposed modules can be plugged into any CNN based recognition models.

2. Related Work

Generic Object Recognition: General image classification was popularized by the appearance of ILSVRC [27]. With the extraordinary improvement achieved by AlexNet [32], deep learning wave started in the field of computer vision. Since then, a series works, e.g. VGGNet [30], GoogLeNet [33], ResNet [13], Inception Net [33, 35], SENet [15], etc. are proposed to learn better representation for image recognition.

However, general object recognition models still suffer from easy confusion among visually similar objects [1, 8]. The class confusion patterns usually follow a hierarchical structure over the classes. General object recognition networks usually can well separate high-level groups of classes, but it is quite costly to learn specialized feature detectors that separate individual classes. The reason is that the global geometry and appearances of the classes in the same hierarchy can be very similar. As a result, how to identify their subtle differences in the discriminative regions is of vital importance.

Fine-Grained Object Recognition: Different from general object recognition, delicate feature representation of object parts play a more critical role in fine-grained object recognition. Existing fine-grained image classification methods can be concluded in two directions. The first one is to enhance the detailed feature representation ability of the backbone network [34, 31, 37]. The second one is to introduce part locations or object bounding box annotations as an additional optimization objective or supervision besides basic classification network [43, 44, 11, 18].

Similar to general object recognition, deep learning based feature representations achieved great success on fine-grained image recognition [9, 28]. After that, second-order bilinear feature representation learning methods [21] and a series of extensions [39, 17, 42] were proposed for learning local pairwise feature interactions in a translation invariant manner.

However, recognizing objects from a fine-grained category requires the neural network to focus more on the discriminative parts [40]. To address this problem, a large amount of part localization based fine-grained recognition methods are proposed. Most of these methods applied attention mechanism to obtain discriminative regions [11, 25]. Zheng *et al*. [44] tried to generate multiple parts by clustering, then classified these parts to predict the category. Compared with earlier part based methods, some recent works tend to use weak supervisions or even no annotation of parts or key areas [26, 41]. In particular, Peng *et al*. [26] proposed a part spatial constraint to make sure that the model could select discriminative regions, and a specialized clustering algorithm is used to integrate the features of these regions. Yang *et al*. [41] introduced a method to detect informative regions and then scrutinizes them for final predictions. These previous works aim to search for key regions from pixel-level images directly. However, to correctly detect discriminative parts, the deep understanding of the structures of objects and the spatial contextual information of key regions are essential. In turn, the location information of regions in images can enhance the visual representation of neural networks [24], which has been demonstrated on unsupervised feature learning.

Different from previous works, our proposed method focuses on modeling spatial connections among object parts for understanding object structure and localizing discriminative regions. Inspired by the studies that contextual information among objects influences the accuracy and efficiency of object recognition [14], the spatial information among regions within objects also benefits the localization of discriminative regions. Thus we introduce two modules in our proposed method; the first one aims to detect the main objects, and the second one inferences the spatial dependency among regions in objects. The experimental results show that our method can improve the performance of both general object recognition and fine-grained object recognition. Moreover, our method has no additional overhead except the backbone network feedforward during inference.

3. Approach

In this section, we introduce our proposed LIO approach. As shown in Fig. 3, our network is mainly organized by three modules:

- **Classification Module** (CM): the backbone classification network that extracts basic image representations and produces the final object category.
- **Object-Extent Learning Module** (OEL): a module for localizing the main object in a given image.
- **Spatial Context Learning Module** (SCL): a self-supervised module to strengthen the connections among regions through interactions among feature cells in CM.

Given an image I and its ground truth one-hot label $\boldsymbol{l}$, we can get the feature maps $\boldsymbol{f}(I)$ of size $N \times N \times C$ from one of the convolutional layers, and the probability vector $\boldsymbol{y}(I)$ from the classification network. C is the channel size of that layer, and $N \times N$ is the size of each feature map in $\boldsymbol{f}(I)$. The loss function of the classification module (CM) $\mathcal{L}_{cls}$ can be written as:

$$\mathcal{L}_{cls} = -\sum_{I \in \mathcal{I}} \boldsymbol{l} \cdot \log \boldsymbol{y}(I), \tag{1}$$

where $\mathcal{I}$ is the image set for training.

The object-extent learning module and spatial context learning module are designed to help our backbone classification network learn representations beneficial to structure understanding and object localization. These two modules are light-weighted, and only a few learnable parameters are introduced. Furthermore, OEL and SCL are disabled at inference time, and only the classification module is needed for computational efficiency.

3.1. Object-Extent Learning (OEL)

Localizing the extent of the object in an image is a prerequisite for understanding the object structure. A typical approach is to introduce bounding boxes or segmentation

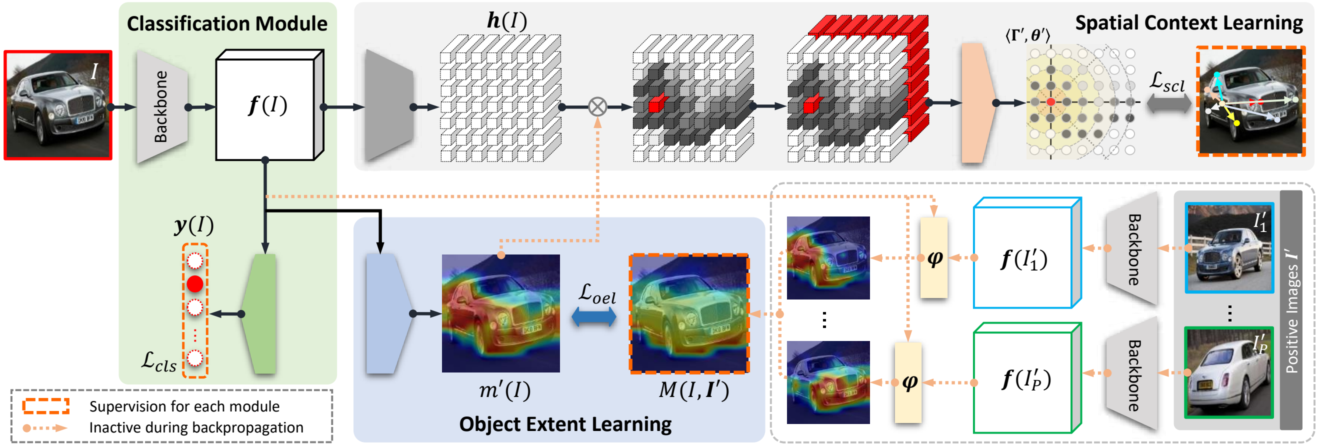

Figure 3. The overall pipeline of our Look-into-object (LIO) framework. The feature maps $\boldsymbol{f}(I)$ extracted from the classification module are further fed into spatial context learning module and object-extent learning module. After end-to-end training, the backpropagation signals from spatial context learning module and object-extent learning module can jointly optimize the representation learning of the backbone network in classification module. Only the classification module (in the green box) is activated during inference.

annotations, which cost much on data collection. For typical image recognition task that lacks localization or segmentation annotations, we propose a new module called *Object-Extent Learning* to help the backbone network distinguish the foreground and background.

We can partition the feature maps $\boldsymbol{f}(I)$ into $N \times N$ feature vector $\boldsymbol{f}(I)_{i,j} \in \mathbb{R}^{1\times C}$, where i and j are the horizontal and vertical indices respectively ($1 \leq i, j \leq N$). Each feature vector centrally responds to a certain region in input image I.

Inspired by the principle that objects in the image from the same category always share some commonality, and the commonality, in turn, help the model recognize objects, we sample a positive image set $\boldsymbol{I}' = \{I'_1, I'_2, \cdots, I'_P\}$ with the same label $\boldsymbol{l}$ of image I, and then measure the region-level correlations between $\boldsymbol{f}(I)_{i,j}$ and each image $I' \in \boldsymbol{I}'$ by

$$\varphi_{i,j}(I, I') = \frac{1}{C} \max_{1 \leq i', j' \leq N} \langle \boldsymbol{f}(I)_{i,j}, \boldsymbol{f}(I')_{i',j'} \rangle, \quad (2)$$

where $\langle \cdot, \cdot \rangle$ denotes dot product.

Jointly trained with the classification objective $\mathcal{L}_{cls}$, the correlation score $\varphi_{i,j}$ is usually positively correlated with the semantic relevance to $\boldsymbol{l}$.

After that, we can construct a $N \times N$ semantic mask matrix $\varphi(I, I')$ for the object extent in I.

Therefore, the commonality of images from the same category can be well captured by this semantic correlation mask φ, and the values in φ distinguish the main object area and background naturally, as shown in Fig. 4.

Taking the impact of viewpoint variation and deformation into account, we use multiple positive images to localize the main area of an object. Therefore, we get a weakly supervisory pseudo label to mimic the object localization masks:

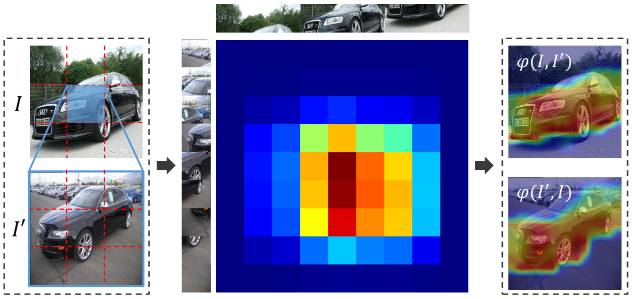

Figure 4. Correlation calculation helps to localize object extent.

$$M(I, \boldsymbol{I}') = \frac{1}{P} \sum_{p=1}^{P} \varphi(I, I'_p). \quad (3)$$

Also, $M(I, \boldsymbol{I}')$ can be regarded as the representations of the commonality shared among images from the same category.

The primary purpose of the OEL module is to enrich the classification network from the commonality and infer the semantic mask of the object extent. Thus we equip a simple stream after $\boldsymbol{f}(I)$ to fuse all feature maps in $\boldsymbol{f}(I)$ with weights. The features are processed by a 1×1 convolution to obtain outputs with one channel $m'(I)$. Different from traditional attention that aims to detect some specific parts or regions, our OEL module is trained for gathering all regions within the object and neglect the background or other irrelevant objects.

The loss of OEL module $\mathcal{L}_{oel}$ can be defined as the distance between pseudo mask $M(I, \boldsymbol{I}')$ of the object extent and $m'(I)$, which can be expressed as:

$$\mathcal{L}_{oel} = \sum_{I \in \mathcal{I}} \text{MSE}\big(m'(I), M(I, \boldsymbol{I}')\big), \quad (4)$$

where MSE is defined as a mean-square-error loss function.

$\mathcal{L}_{oel}$ is helpful to learn a better representation of the object extent according to visual commonality among images in the same category. By end-to-end training, the object-extent learning module can enrich the backbone network by detecting the main object extent.

3.2. Spatial Context Learning (SCL)

Structural information plays a significant role in image comprehension. Classical general convolutional neural networks use convolutional kernels to extract structural information in the image, and fuse the multi-level information by stacking layers. We propose a self-supervised module called *Spatial Context Learning* to strengthen the structural information for the backbone network by learning the spatial context information in objects.

Given an image I, our SCL module also acts on the feature maps $\boldsymbol{f}(I)$ and aims to learn the structural relationships among regions. Firstly, the feature map is processed by a 1×1 convolution plus a ReLU such that we get the new map $\boldsymbol{h}(I) \in \mathbb{R}^{N \times N \times C_1}$, describing the spatial information of different feature cells. Each cell in $\boldsymbol{h}(I)$ centrally represents the semantic information of an area of the image I. The structural relationships among different parts of an object can be easily modeled by building spatial connections among different regions.

In this paper, we apply polar coordinates for measuring the spatial connections among different regions. Given a reference region $R_o = R_{x,y}$ whose indices are (x, y) in $N \times N$ plane, and a reference horizontal direction, the polar coordinates of region $R_{i,j}$ can be written as $(\Gamma_{i,j}, \theta_{i,j})$:

$$\begin{aligned} \Gamma_{i,j} &= \sqrt{(x-i)^2 + (y-j)^2}/\sqrt{2}N \\ \theta_{i,j} &= (\mathbf{atan2}(y-j, x-i) + \pi)/2\pi, \end{aligned} \tag{5}$$

where $0 < \Gamma_{i,j} \leq 1$ measures the relative distance between R_o and $R_{i,j}$, $\mathbf{atan2}(\cdot)$ returns a unambiguous value in range of $(-\pi, \pi]$ for the angle converting from Cartesian coordinates to polar coordinates, and $\theta_{i,j}$ measures the polar angle of $R_{i,j}$ corresponding to the horizontal direction. It is worth noting that, to ensure a wide range of the distribution of the values of θ, ideally, the region within the object extent should be selected as the reference region. In this paper, the region who respond to the maximum value in $m(I)$ is selected:

$$R_o = R_{x,y}, \text{where}\, (x, y) = \arg \max_{1 \leq x,y \leq N} m'(I)_{i,j} \tag{6}$$

This ground-truth polar coordinates is regarded as supervision for guiding the SCL module training. Specifically, the SCL module is designed for predicting the polar coordinates of region $R_{i,j}$ by jointly considering the representations of target region $R_{i,j}$ and reference region R_o from $\boldsymbol{h}(I)$. We first apply channel-wise concatenation for $h(I)_{i,j}$ and $h(I)_{x,y}$, then the outputs are handled by a fully-connected layer with ReLU to get the predicted polar coordinates $(\Gamma'_{i,j}, \theta_{i,j})'$. Since our proposed modules mainly focus on modeling the spatial structures of different parts within the object, the object-extent mask $m'(I)$ learned from the OEL module is also adapted in the SCL module.

There are two objectives in the SCL module. The first one measures the relative distance differences of all regions with object:

$$\mathcal{L}_{dis} = \sum_{I \in \mathcal{I}} \sqrt{\frac{\sum_{1 \leq i,j \leq N} m'(I)_{i,j} (\Gamma'_{i,j} - \Gamma_{i,j})^2}{\sum m'(I)}}. \tag{7}$$

The other one measures the polar angle differences of regions inside the object. Considering the structural information for an object should be rotation invariant, and robust to various appearances and poses of the object, we measure the polar angle difference $\mathcal{L}_{\angle}$ according to the *standard deviation* of gaps between predicted polar angles and ground-truth polar angles:

$$\begin{aligned} \mathcal{L}_{\angle} &= \sum_{I \in \mathcal{I}} \sqrt{\frac{\sum_{1 \leq i,j \leq N} m'(I)_{i,j} \left(\theta_{\Delta_{i,j}} - \overline{\theta}_{\Delta}\right)^2}{\sum m'(I)}}, \\ \theta_{\Delta_{i,j}} &= \begin{cases} \theta'_{i,j} - \theta_{i,j}, & \text{if } \theta'_{i,j} - \theta_{i,j} \geq 0 \\ 1 + \theta'_{i,j} - \theta_{i,j}, & \text{otherwise,} \end{cases} \end{aligned} \tag{8}$$

where $\overline{\theta}_{\Delta} = \frac{1}{\sum m'(I)} \sum_{1 \leq i,j \leq N} m'(I)_{i,j} \theta_{\Delta_{i,j}}$ is the mean of the gaps between predicted polar angles and ground-truth polar angles. In this way, our SCL could focus on modeling the relative structure among parts of the object rather than the absolute position of regions that is sensitive to the reference direction. Moreover, owing to the usage of predicted semantic mask $m'(I)$, other visual information except for the main object, e.g., background, is ignored during regressing polar coordinates.

Overall, the loss function of the *Spatial Context Learning Module* can be written as:

$$\mathcal{L}_{scl} = \mathcal{L}_{dis} + \mathcal{L}_{\angle}. \tag{9}$$

With $\mathcal{L}_{scl}$, the backbone network can recognize the pattern structures, i.e., the composition of the object. By end-to-end training, the spatial context learning module can empower the backbone network to model the spatial dependence among parts of the object.

3.3. Joint Structure Learning

In our framework, the classification, object-extent learning and spatial context learning modules are trained in an end-to-end manner, in which the network can leverage both enhanced object localization and object structural information. The whole framework is trained by minimizing the following objective:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{oel} + \beta\mathcal{L}_{scl}. \tag{10}$$

We set $\alpha = \beta = 0.1$ for all experimental results reported in this paper.

During inference, both the SCL and OEL are removed, and only the Classification Module is kept. Thus, the framework does not introduce additional computational overhead at inference time and runs faster for practical product deployment.

Moreover, the object-extent learning module and spatial context learning module can be attached to different stages of feature maps generated from different convolutional layers of the Classification Module. Thus we can model the structural information of the object in different granularity levels. Together, the overall training method is named as **multi-stage LIO**. For example, we can jointly optimize our framework by the combination of $\mathcal{L}_{7\times7}$ (extracting feature maps from the last convolutional layer with $N = 7$) and $\mathcal{L}_{14\times14}$ (from the penultimate convolutional layer with $N = 14$) for ResNet-50.

4. Experiments

To show the superiority of our proposed look-into-object framework, we evaluate the performance on two object recognition settings: fine-grained object recognition and generic image classification. Furthermore, we also explore our LIO framework in other tasks, such as object detection and segmentation, to study its generalization ability.

Unless specially mentioned, the spatial context learning module and object-extent learning module are applied on the feature map of the last stage in the backbone classification network, and three positive images are used for training procedure by default. For all of these tasks, we did not use any additional annotations.

4.1. Fine-grained Object Recognition

For fine-grained object recognition, we test LIO on three different standard benchmarks: CUB-200-2011 (CUB) [4], Stanford Cars (CAR) [19] and FGVC-Aircraft (AIR) [23].

We first initialize LIO with ResNet-50 backbone pretrained on ImageNet classification task, and then finetune our framework on the datasets above-mentioned. The input images are resized to a fixed size of 512×512 and randomly cropped into 448×448 for scale normalization. We adopt random rotation and horizontal flip for data augmentation. All above transformations are standard in the literature. Both ResNet-50 baseline and LIO/ResNet-50 are trained for 240 epochs to ensure complete convergence. SGD is used to optimize the training loss as defined in Equation 10.

Method	Accuracy (%)		
	CUB	CAR	AIR
CoSeq (+BBox) [18]	82.8	92.8	-
FCAN (+BBox) [22]	84.7	93.1	-
B-CNN [21]	84.1	91.3	84.1
HIHCA [3]	85.3	91.7	88.3
RA-CNN [11]	85.3	92.5	88.2
OPAM [26]	85.8	92.2	-
Kernel-Pooling [7]	84.7	91.1	85.7
MA-CNN [45]	86.5	92.8	89.9
DeepKSPD-rootm [10]	86.5	93.2	91.0
MAMC [25]	86.5	93.0	-
HBP [42]	87.1	93.7	90.3
DFL-CNN [38]	87.4	93.1	91.7
NTS-Net [41]	87.5	93.9	91.4
DCL [6]	87.8	**94.5**	**93.0**
ResNet-50 Baseline	85.5	92.7	90.3
LIO/ResNet-50 (7×7)	87.3	93.9	92.4
LIO/ResNet-50 (14×14)	87.3	94.2	92.3
LIO/ResNet-50 (28×28)	87.6	94.0	92.4
LIO/ResNet-50 (multi-stage)	**88.0**	**94.5**	92.7

Table 1. Comparison results on three different fine-grained object recognition benchmarks.

During testing, only the backbone network is applied for classification. The input images are centrally cropped and then fed into the backbone classification network for final predictions.

Detailed results are summarized in Table 1. Besides plugging the OEL and SCL to the last stage feature map of size 7×7, we also tested these two modules on the penultimate stage 14×14 output, and the antepenultimate stage 28×28 output. Then these three different stages of models are combined into a multi-stage LIO. As in Table 1, the LIO embedded ResNet-50 can achieve significantly better accuracy than baseline ResNet-50. Moreover, the multi-stage LIO achieves significant performance improvements on all three benchmarks, which proves the effectiveness of the proposed region-level structure learning framework.

It worthy note that LIO and our previous work DCL [6] target at different research lines in the fine-grained recognition task. DCL aims to learn *discriminative local regions*, while LIO tries to understand the *structure of the whole object*. Both of these two kinds of methods can benefit the fine-grained object recognition, while LIO works better on recognition of flexible objects (CUB), and can be further expanded into generic object recognition (Sec. 4.2), object detection and segmentation (Sec. 4.3) since object structure information plays an essential role in those tasks.

4.2. Generic Object Recognition on ImageNet

We also evaluate the performance of our proposed LIO on large-scale general object recognition dataset ImageNet-

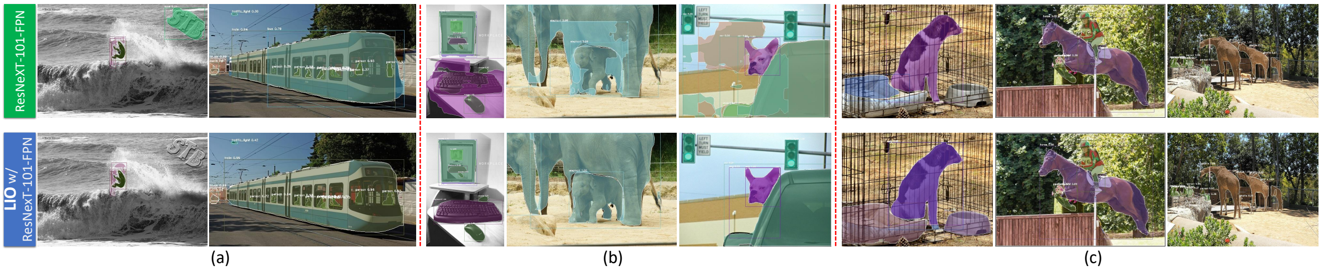

Figure 5. Qualitative examples for COCO object detection and instance segmentation. Our LIO based method can help improve the performance according to object structure information in three aspects: (a) reducing incorrect object label prediction. (b) neglecting noisy segmentation mask. (c) completing fragmentary segmentation mask. Best viewed in electronic version.

Method	Top-1 err.	Top-5 err.
ResNet-50 [13]	24.80	7.48
LIO/ResNet-50 (7 × 7)	23.63	7.12
LIO/ResNet-50 (14 × 14)	23.60	7.10
LIO/ResNet-50 (multi-stage)	**22.87**	**6.64**

Table 2. Single-crop error rates (%) of single model on the ImageNet-1K validation set.

1K (ILSVRC-2012), which including 1.28 million images in 1000 classes. For compatibility test, we evaluate our method on the commonly used backbone network ResNet-50. Following standard practices, we perform data augmentation with random cropping to a size of 224 × 224 pixels and perform random horizontal flipping. The optimization is performed using SGD with momentum 0.9 and a mini-batch size of 256.

The experimental results are reported in Table 2. We can find that LIO boosts the performance of three different backbone networks on the ImageNet-1K validation set, which further demonstrates the generality ability of our proposed object recognition framework. With a lightweight LIO plugin, the performance of typical ResNet-50 can even achieve the performance of SE-ResNet-50 [15].

4.3. Object Detection and Segmentation on COCO

Meanwhile, considering the object structure information would be helpful for object detection and segmentation tasks, we also investigate our proposed LIO on the object detection/segmentation task on MS COCO dataset [20]. We adopt the basic Mask R-CNN [12] and plug the LIO behind the Region Proposal Network, such that the structural information of each object can be well modeled. The SCL module can directly act on the object features after ROI pooling, thus the OEL module is disabled. We implemented the novel detection/segmentation network based on *mmdetection* [5] toolbox and keep all hyper-parameters as default.

We apply the LIO module on the basic baseline of ResNet-50-C4 and a higher baseline of ResNeXt-101-FPN. The models are trained on COCO `train2017` set and evaluated in the COCO `val2017` set. We report the standard COCO metrics including AP, AP_{50}, AP_{75} (averaged precision over multiple IoU thresholds), and AP_S, AP_M, AP_L (AP across scales). Experimental results described in Table 3 show that modeling structural compositions benefit object understanding and lead to better results on semantic segmentation. This demonstrated the effectiveness and generalization ability of our LIO for object structural compositions learning. Some examples of results by our basic ResNeXt-101-FPN and our approach are given in Fig. 5.

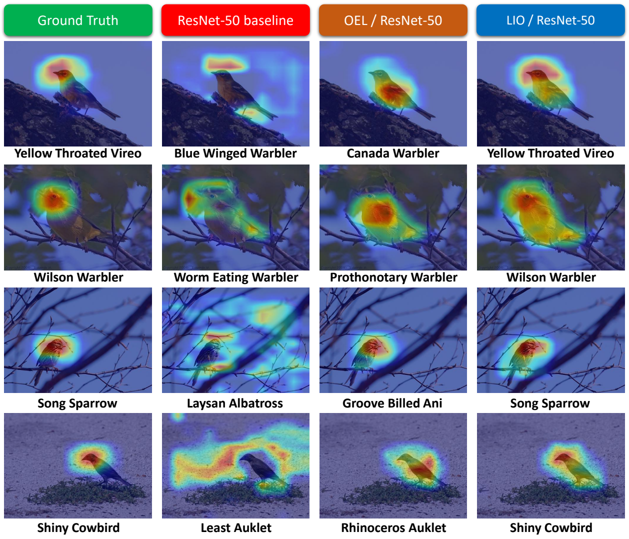

Figure 6. Visualization of feature maps by using OEL and SCL respectively. OEL enforce the backbone focus on object extent. SCL is helpful for not only searching discriminative region in object extent, but also completing the object extent localized by OEL.

4.4. Ablation Studies

To demonstrate the effects of the OEL module and SCL module, we perform the module separation experiments on CUB [4] and CAR [19]. Both OEL and SCL act on the last stage feature map from the ResNet-50 backbone. The results are shown in Table 4. We can find that both modules improve performance significantly. In detail, as we show

Method	Object Detection						Semantic Segmentation					
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50-C4	35.9	56.1	38.9	18.0	40.1	49.7	31.5	52.8	33.0	12.1	34.7	49.3
LIO/ResNet-50-C4	**37.6**	**57.5**	**41.0**	**21.0**	**41.8**	**52.0**	**32.6**	**54.1**	**34.7**	**14.3**	**35.7**	**51.3**
ResNeXT-101-FPN	41.1	62.8	45.0	24.0	45.4	52.6	37.1	59.4	39.7	17.7	40.5	53.8
LIO/ResNeXT-101-FPN	**42.0**	**63.3**	**46.0**	**24.7**	**46.1**	**54.3**	**37.9**	**60.0**	**40.6**	**18.1**	**41.1**	**54.8**

Table 3. Object detection and segmentation results on COCO `val2017` set.

Method	Accuracy (%)	
	CUB	CAR
ResNet-50 [13]	85.50	92.73
SCL	86.74	93.82
OEL	86.99	93.83
LIO	**87.31**	**93.89**
LIO w/ GM	87.37	-

Table 4. Ablation studies conducted on the proposed framework. ResNet-50: Basic ResNet-50 neural network trained by $\mathcal{L}_{cls}$. OEL: Model trained by $\mathcal{L}_{cls} + \alpha\mathcal{L}_{oel}$. SCL: Model trained by $\mathcal{L}_{cls} + \beta\mathcal{L}_{scl}$. LIO: Model trained by $\mathcal{L}$. GM: Ground truth semantic segmentation annotations.

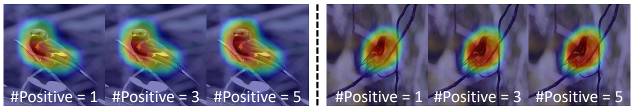

Figure 7. Visualization of the changes of pseudo segmentation masks given different number of positive images.

Dataset	# Positive Images		
	1	3	5
CUB	86.83	87.31	87.30
CAR	93.81	93.89	93.89

Table 5. The effect of the number of positive images on accuracy.

in Fig. 6, the SCL provides a principled way to learn the spatial structure, which is helpful for mining discriminative regions in an object. Moreover, the object extent can be localized by the OEL module according to the in-class region correlations and further defeats the negative influences from the diverse poses, appearance and background clutter. Together, the overall performance can be further improved owing to the complementary of nature SCL and OEL.

Moreover, we also try to replace the pseudo semantic mask $M(I, \boldsymbol{I}')$ with the ground-truth mask for LIO. The results show that our learning based method can construct a high-quality semantic mask, which is even very close to the ground-truth mask (87.3% vs. 87.4% accuracy on CUB).

4.5. Discussions

Number of Positive Images: The number P of positive images in a batch is an important parameter for the object-extent learning Module. We visualized the pseudo mask $M(I, \boldsymbol{I}')$ by given different number of positive images P in Fig. 7. We also evaluate our method on CUB and CAR with different numbers of positive images, and the recognition accuracy is shown in Table 5. With more positive images used, the framework gets better in structural learning and result in better performance. Finally, the performance will stop rising or falling and become steady. For a rigid object structure, such as CAR, we only need a few positive images for generating a reasonable pseudo extent mask.

In general, feeding only one positive image may let the backbone learn fragmentary object extent for viewpoint diversity. The increase of P leads to rapidly rising memory usage. Thus we use $P = 3$ for experiments in this paper to trade-off between final performance and computation cost.

Model Efficiency: During training time, our LIO introduced three additional layers besides the backbone network, including one convolutional layer in the OEL module, one convolutional layer and one fully-connected layer in the SCL module. For LIO/ResNet-50 (28x28), there are only 0.26 million new parameters introduced in our LIO, which is 1.01% of #Params of original ResNet-50.

An important property is that both OEL and SCL modules can be disabled during testing. It means that the final classification model size is the same as the original backbone network. The baseline backbone network can be significantly improved without any computation overhead at inference time.

5. Conclusions

In this paper, we proposed a Look-into-Object (LIO) framework to learn structure information for enhancing object recognition. We show that supervised object recognition could largely benefit from "additional but free" self-supervision, where geometric spatial relationship significantly rectifies the localization of discriminative regions and even result in better object detection and segmentation. Structural information, which was overlooked in prior literature, reliably prevents the network from falling into local confusion. Moreover, our plug-in style design can be widely adopted for injecting extra supervision into the backbone network without additional computational overhead for model deployment.

References

[1] Alsallakh Bilal, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Do convolutional neural networks learn class hierarchy? *IEEE transactions on visualization and computer graphics*, 24(1):152–162, 2018. 3

[2] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 1

[3] Sijia Cai, Wangmeng Zuo, and Lei Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 511–520, 2017. 6

[4] W. Catherine, B. Steve, W. Peter, P. Pietro, and B. Serge. The caltech-ucsd birds-200-2011 dataset. (CNS-TR-2011-001), 2011. 6, 7

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 7

[6] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019. 6

[7] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2930, 2017. 1, 6

[8] Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei. What does classifying more than 10,000 image categories tell us? In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 71–84, 2010. 3

[9] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014. 3

[10] Melih Engin, Lei Wang, Luping Zhou, and Xinwang Liu. Deepkspd: Learning kernel-matrix-based spd representation for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 612–627, 2018. 6

[11] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017. 1, 3, 6

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 7, 8

[14] Howard S Hock, Gregory P Gordon, and Robert Whitehurst. Contextual relations: the influence of familiarity, physical plausibility, and belongingness. *Perception & Psychophysics*, 16(1):4–8, 1974. 3

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1, 2, 7

[16] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1173–1182, 2016. 1

[17] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 365–374, 2017. 3

[18] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5546–5555, 2015. 1, 3, 6

[19] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, Dec 2013. 6, 7

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[21] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015. 1, 3, 6

[22] Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016. 6

[23] S. Maji, E. Rahtu, J. Kannala, M.B. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. 6

[24] N. Mehdi and F. Paolo. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision – ECCV 2016*, pages 69–84, Cham, 2016. Springer International Publishing. 3

[25] S. Ming, Y. Yuchen, Z. Feng, and D. Errui. Multi-attention multi-class constraint for fine-grained image recognition. pages 834–850, 2018. 3, 6

[26] Y. Peng, X. He, and J. Zhao. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, 27(3):1487–1500, March 2018. 3, 6

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2

[28] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014. 3

[29] Ya-Fang Shih, Yang-Ming Yeh, Yen-Yu Lin, Ming-Fang Weng, Yi-Chang Lu, and Yung-Yu Chuang. Deep co-occurrence feature learning for visual object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4123–4132, 2017. 1

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3

[32] Ilya Sutskever, Geoffrey E Hinton, and A Krizhevsky. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[34] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, June 2015. 3

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2

[36] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017. 1

[37] Jingwen Wang, Jianlong Fu, Yong Xu, and Tao Mei. Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks. In *IJCAI*, pages 3484–3490, 2016. 3

[38] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4148–4157, 2018. 6

[39] Xing Wei, Yue Zhang, Yihong Gong, Jiawei Zhang, and Nanning Zheng. Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 355–370, 2018. 3

[40] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, June 2015. 3

[41] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 420–435, 2018. 3, 6

[42] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. Hierarchical bilinear pooling for fine-grained visual recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 574–589, 2018. 3, 6

[43] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan. Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia*, 19(6):1245–1256, June 2017. 3

[44] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017. 3

[45] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017. 6