Hierarchical Context Embedding for Region-based Object Detection

Zhao-Min Chen[1,2], Xin Jin[2], Borui Zhao[2], Xiu-Shen Wei[2,*], and Yanwen Guo[1,*]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] Megvii Research Nanjing, Megvii Technology, China
{chenzhaomin123, weixs.gm}@gmail.com, {jinxin, zhaoborui}@megvii.com, ywguo@nju.edu.cn

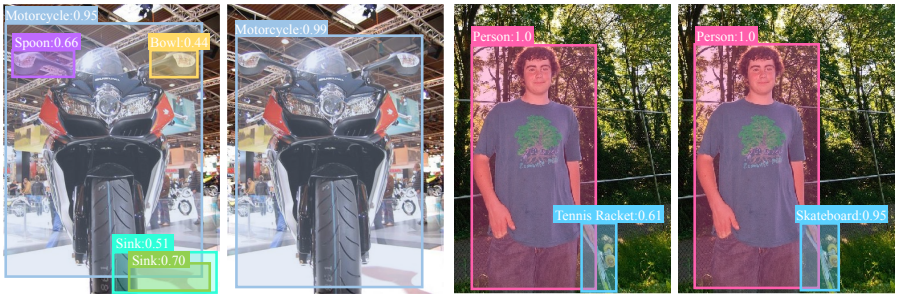
Abstract State-of-the-art two-stage object detectors apply a classifier to a sparse set of object proposals, relying on region-wise features extracted by RoIPool or RoIAlign as inputs. The region-wise features, in spite of aligning well with the proposal locations, may still lack the crucial context information which is necessary for filtering out noisy background detections, as well as recognizing objects possessing no distinctive appearances. To address this issue, we present a simple but effective Hierarchical Context Embedding (HCE) framework, which can be applied as a plug-and-play component, to facilitate the classification ability of a series of region-based detectors by mining contextual cues. Specifically, to advance the recognition of context-dependent object categories, we propose an image-level categorical embedding module which leverages the holistic image-level context to learn object-level concepts. Then, novel RoI features are generated by exploiting hierarchically embedded context information beneath both whole images and interested regions, which are also complementary to conventional RoI features. Moreover, to make full use of our hierarchical contextual RoI features, we propose the early-and-late fusion strategies (*i.e.*, feature fusion and confidence fusion), which can be combined to boost the classification accuracy of region-based detectors. Comprehensive experiments demonstrate that our HCE framework is flexible and generalizable, leading to significant and consistent improvements upon various region-based detectors, including FPN, Cascade R-CNN and Mask R-CNN.

Keywords: Object Detection; Context Embedding; Region-based CNNs.

1 Introduction

The region-based object detectors [14,13,33,23,4,39] popularized by R-CNN framework [14] are conceptually intuitive and flexible, and have achieved top accuracies on challenging benchmarks like MS-COCO [25]. Region-based detectors first generate a sparse set of object proposals, and then refine the proposal locations and classify them as one of the foreground classes or as background using a

* Corresponding authors



(a) Filtering out noisy detections. (b) Recognizing indistinctive objects.

Figure 1. Motivation and example results of our Hierarchical Context Embedding (HCE) framework. By incorporating discriminative context information, our framework can effectively filter out the noisy false positive background detections, and correctly classify objects (*e.g.*, "`skateboard`") which possess no distinctive appearances.

detection head. One crucial module in such a proposal-driven pipeline is the RoIPool [13] or RoIAlign [18] operator, which is responsible for extracting RoI (Region of Interests) features aligned with the proposal locations for the detection head.

In this paper, we revisit the RoI features in region-based detectors from the perspective of context information embedding. Our key motivation relies on the fact that while each RoI in very deep CNNs may have a very large theoretical receptive field that often spans the entire input image [13]. However, the effective receptive field [28] may only occupy a fraction of the full theoretical receptive field, making the RoI features insufficient for characterizing objects that are highly dependent on context information, such as "`bowl`", "`skateboard`" etc. Here, the contextual information means any auxiliary information that can assist in suppressing the false positive detections in noisy backgrounds, or recognizing objects that have no distinctive appearances themselves. For example, as shown in Fig. 1 (a), the semantic features of "`motorcycle`" are strong evidences for filtering out the activations of irrelevant object categories like "`spoon`", "`bowl`", and "`sink`". On the other hand, as shown by Fig. 1 (b), the scene and even the human pose are useful clues for correctly classifying a proposal as "`skateboard`", rather than "`tennis racket`".

Recently, several works exploited the region-level context information to improve the localization ability of two-stage detectors. Chen *et al.* [6] demonstrated that rich contextual information from neighboring regions can better refine the proposal locations for two-stage detectors. Kantorov *et al.* [20] leveraged the surrounding context regions to improve weakly supervised object localization. However, to the best of our knowledge, currently there is no enabling framework which is systematically designed for embedding context information to improve the *classification ability* of region-based detectors.

In this paper, we present a novel Hierarchical Context Embedding (HCE) framework for region-based object detectors. Our framework consists of three modules. Firstly, we consider that the simplest way to break the contextual limit in object detection, is to partially cast the object-level feature learning as an image-level multi-label classification task. Building upon this realization, we design an image-level categorical embedding module, which in essence is a multi-label classifier upon the detection backbone, in parallel with the existing region-based detection head. It enables the backbone to exploit the whole image context to learn discriminative features for context-dependent object categories. Even as a standalone enhancement, our image-level categorical embedding module can lead to improvements over existing region-based detectors.

Upon the image-level categorical embedding module, at the instance-level, we design a simple but effective process to generate hierarchical contextual RoI features which can be directly utilized by the region-wise detection head. Because our contextual RoI features are enhanced by image-level categorical supervisions and exploit larger contexts, they are by nature complementary to conventional RoI features, which is trained by region-based detectors and only exploits limited context. Later, the early-and-late strategies, *i.e.*, feature fusion and confidence fusion, are designed to make full use of our contextual RoI features. By quantitative experiments, we demonstrate that they can be combined to further boost the classification accuracy of the detection head.

In general, our proposed HCE framework is easy to implement and is end-to-end trainable. We conduct extensive experiments on MS-COCO 2017 [25] to validate the effectiveness of our HCE framework. Without bells and whistles, our HCE framework delivers consistent accuracy improvements for almost all existing mainstream region-based detectors, including FPN [23], Mask R-CNN [18] and Cascade R-CNN [4]. We also conduct ablation studies to verify the effectiveness of each module involved in our HCE framework. Fig. 1 gives the example images of the baseline method and our method, which demonstrates that our framework can effectively filter out the noisy background detections and correctly classify indistinctive objects by leveraging the context information it exploited.

2 Related Work

2.1 Region-based Object Detection

Convolutional neural networks have lead to a paradigm shift of object detection in the past decades [26]. Among a large number of approaches, the two-stage R-CNN series [14,13,33,23,4] have become the leading detection framework. The pioneer work, *i.e.*, R-CNN [14], extracts region proposals from image with selective search [36], and applies a convolutional network to classify each region of interests independently. Fast R-CNN [13] improves R-CNN by sharing convolutional features among RoIs, which enables fast training and inference. Then, Faster R-CNN [33] advances the region proposal generation with a Region Proposal Network (RPN). RPN shares the feature extraction backbone with the detection

head, which in essence is a Fast R-CNN [33]. Faster R-CNN is a famous two-stage detection framework, and is the foundation for many follow-up works [8,23].

Over very recent years, several algorithms have been proposed to further improve the two-stage Faster R-CNN framework. For example, Feature Pyramid Networks (FPN) [23] constructed inherent CNN feature pyramids, which can largely improve the detection performance of small objects. Mask R-CNN [18] extended Faster RCNN by constructing the mask branch, and boosted the performance of both object detection and instance segmentation. Cascade R-CNN [4] utilized multi-stage training strategy to progressively improve the quality of region proposals, and demonstrated significant gains for high quality (measured by higher IoUs) object detection. Complementary to these works, in this paper, we focus on developing a Hierarchical Context Embedding (HCE) framework to improve the *classification ability* of all region-based detectors. Thanks to the simplicity and generalization ability of our HCE framework, it brings consistent and significant improvements over aforementioned leading region-based detectors, *e.g.*, FPN, Mask R-CNN and Cascade R-CNN.

2.2 Context Information for Object Detection

In object detection, both global context [12] and local context [30] are widely exploited for improving performance, especially when object appearances are insufficient due to small object size, occlusion, or poor image quality. Our work is inspired by some of previous works, but the key motivation or implementation significantly differ with these works. Next, we review several topics in object detection, which are closely related to our work.

Combined Localization and Classification. Before the era of deep learning, Harzallah *et al.* [16] proposed to combine two closely related tasks, *i.e.*, object localization and image classification. They demonstrated that classification can improve detection by a contextual combination and vice versa. Similar in spirit, we utilize the fully image-level context to learn object-level concepts. But differently, we utilize global context to learn CNN features rather than hand-crafted features adopted in [16]. Furthermore, we integrate hierarchical contextual clues beneath both whole images and interested regions to modern region-based CNN detectors, rather than the traditional sliding window detector used by [16].

Region Proposal Refinement. Recently, Chen *et al.* [6] explored the rich contextual information to refine the region proposals for object detection. The neighboring regions with useful contexts can benefit the localization quality of region proposals, which further lead to better detection performance. Instead of refining proposals, we focus on improving the *classification ability* of region-based detectors by embedding hierarchical contextual clues.

Weakly-Supervised Object Detection. In weakly supervised object detection, the bounding box annotations are not provided, and only image-level

categorical labels are available. The common practice [7,2,1,20] in this area is to first generate a set of noisy object proposals, and then learn from these noisy proposals with specially designed robust algorithms. Among them, Kantorov *et al.* [20] proposed a context-aware deep network which leverages the surrounding context regions to improve localization. Unlike the usage of region-level context information [20] for weakly supervised detection, we focus on the task of fully-supervised object detection, and particularly exploit global image-level context to advance the recognition of context-dependent object categories.

2.3 Context Information for Other Vision Tasks

Beyond object detection, context information has also been utilized to improve other vision tasks. For example, Wang *et al.* [37] leveraged attention mechanisms and LSTMs to discover semantic-aware regions and capture the long-range contextual dependencies for multi-label image recognition. He *et al.* [17] proposed an adaptive context module to generate multi-scale context representations for semantic segmentation. Qu *et al.* [29] embedded multi-context information (the appearance of the input image and semantic understanding) to obtain the shadow matte. Byeon *et al.* [3] leveraged the LSTM units to capture the entire available past context on video prediction. Li *et al.* [21] adopted the dilated convolution to acquire more contextual information for single image deraining.

3 Approach

3.1 Framework Overview

We begin by briefly describing our Hierarchical Context Embedding (HCE) framework (see Fig. 2) for region-based object detection. Firstly, an image-level categorical embedding module is employed to advance the feature learning of the objects that are highly dependent on larger context clues. Then, hierarchical contextual RoI features are generated by fusing both instance-level and global-level information derived from the image-level categorical embedding module. Finally, early-and-late fusion modules are designed to make full use of the contextual RoI features to improve the classification performance. Our HCE framework is flexible and generalizable, as it can be applied as a plug-and-play component for almost all mainstream region-based object detectors.

3.2 Image-Level Categorical Embedding

As aforementioned, conventional RoI-based training for region-based detectors may lack the context information, which is crucial for learning discriminative filters for context-dependent objects. To break this limitation, in parallel with the RoI-based branch, we exploit image-level categorical embedding upon the detection backbone, enabling the backbone to learn object-level concepts adaptively from *global-level* context. Our image-level categorical embedding module does not

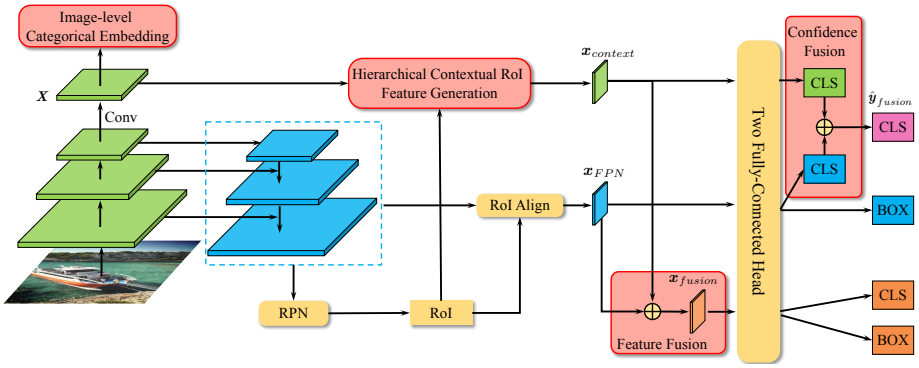


Figure 2. Overview of our Hierarchical Context Embedding (HCE) framework. At the image-level, we design an *image-level categorical embedding* module upon the detection backbone, which enables the network to learn object-level concepts from global-level context. At the instance-level, we generate *hierarchical contextual RoI features* that are complementary to conventional RoI features, and design the early-and-late fusion strategies (*i.e.*, *feature fusion* and *confidence fusion*) to make full use of the contextual RoI features for improving the classification accuracy of the detection head.

require additional annotations, as the image-level labels can be conveniently obtained by collecting all instance-level categories in an image.

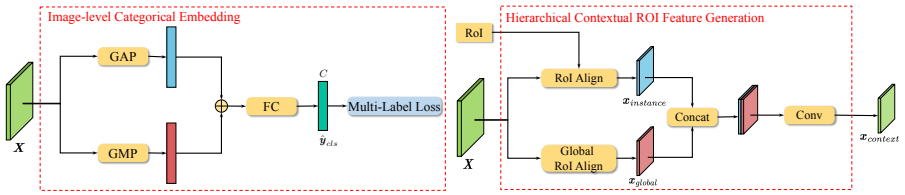
Essentially, our image-level categorical embedding module is based on a multi-label classifier. As shown in Fig. 2 and Fig. 3 (a), we first apply a 3×3 convolution layer on the output of ResNet conv_5 to obtain the input feature map, and then employ both global max-pooling (GMP) and global average-pooling (GAP) for feature aggregation (as in [38]). Here, the additional 3×3 convolution layer aims to alleviate the possible slide effects over the original detection backbone.

We refer to the input feature map to our image-level embedding module as *context-embedded image feature*. This is because the input feature map conveys whole image context for learning all object categories that appear in the image, and in turn, each location of the feature map is supervised by all object categories. By contrast, conventional RoI-based trained by region-based detectors only exploits limited context for learning each object category.

Formally, let $\boldsymbol{X} \in \mathbb{R}^{d \times h \times w}$ denote the input feature map, where d is the channel dimensionality, h and w are the height and width, respectively. Then, the multi-label classifier is constructed by C binary classifiers for all categories:

$$\hat{\boldsymbol{y}}_{cls} = f_{cls}((f_{gmp}(\boldsymbol{X}) + f_{gap}(\boldsymbol{X}))) \in \mathbb{R}^C \,, \tag{1}$$

where C denotes the number of categories, each element of $\hat{\boldsymbol{y}}_{cls}$ is a confidence score (logits), and f_{cls} is binary classifier modeled as one fully-connected layer. We assume that the ground truth label of an image is $\boldsymbol{y} \in \mathbb{R}^C$, where $y^i = \{0, 1\}$ denotes whether object of category i appears in the image or not. The multi-label loss can be formulated as follows



(a) Image-level categorical embedding. (b) Hierarchical contextual RoI feature.

Figure 3. The design of our image-level categorical embedding module and hierarchical contextual RoI feature generation module.

$$\mathcal{L}_{mll} = -\sum_{c=1}^{C} y^c \log(\sigma(\hat{y}^c_{cls})) + (1 - y^c) \log(1 - \sigma(\hat{y}^c_{cls})) \,, \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function.

Because the global feature learning strategy is complementary to RoI-based training, our image-level categorical embedding module standalone can boost the performance of existing region-based detectors (demonstrated later by experiments, cf. Table 2). However, one limitation of image-level categorical embedding might be that the derived context-embedded image feature can not be directly used by the detection head.

3.3 Hierarchical Contextual RoI Feature Generation

To further benefit region-wise classification, we generate hierarchical contextual RoI features by combining the instance-level and global-level information from the context-embedded image features. The hierarchical contextual RoI feature generation process is shown in Fig. 3 (b).

Context-Embedded Instance-Level Feature We apply RoIAlign [18] with proposals generated by RPN on the context-embedded feature map $\boldsymbol{X}$ to obtain RoI features $\boldsymbol{x}_{instance}$:

$$\boldsymbol{x}_{instance} = f_{RoIAlign}(\boldsymbol{X}; h', w') \in \mathbb{R}^{d \times 7 \times 7} \,, \tag{3}$$

where $f_{RoIAlign}(\cdot)$ is the RoIAlign operation and h' and w' are the height and width of the RoI, respectively. As $\boldsymbol{x}_{instance}$ is extracted from the context-embedded image feature $\boldsymbol{X}$, we term it as *context-embedded instance-level feature*.

Context-Aggregated Global-Level Feature To leverage larger context, we exploit RoIAlign on the context-embedded image feature $\boldsymbol{X}$ to aggregate the global-level context. We refer to the derived RoI feature as *context-aggregated global-level feature* $\boldsymbol{x}_{global}$:

$$\boldsymbol{x}_{global} = f_{RoIAlign}(\boldsymbol{X}; H, W) \in \mathbb{R}^{d \times 7 \times 7} \,, \tag{4}$$

where H and W are the height and width of the input image, respectively.

Once context-embedded instance-level feature $\boldsymbol{x}_{instance}$ and context-aggregated global-level feature $\boldsymbol{x}_{global}$ obtained, we concatenate these two RoI features and apply a 1×1 convolution layer to obtain our hierarchical contextual RoI feature $\boldsymbol{x}_{context}$:

$$\boldsymbol{x}_{context} = f_{conv}([\boldsymbol{x}_{instance} : \boldsymbol{x}_{global}]) \in \mathbb{R}^{d \times 7 \times 7}, \tag{5}$$

where $f_{conv}(\cdot)$ denotes the 1×1 convolution operation, [:] refers to concatenation and the ReLU nonlinearity operations are performed following the convolution layer. As the resulting hierarchical contextual RoI feature $\boldsymbol{x}_{context}$ absorbs rich context information from the context-embedded image feature $\boldsymbol{X}$, it is by nature complementary to the conventional RoI feature extracted from the feature pyramid network (FPN) [23].

3.4 Early-and-Late Fusion and Inference

To make full use of our contextual RoI feature $\boldsymbol{x}_{context}$, we design the early-and-late fusion strategies, *i.e.*, feature fusion and confidence fusion, which has been proven effective in many applications [15,11]. We show that early-and-late fusion is also well suited to improve region-wise detectors, as it can fully absorb hierarchically embedded information from different levels.

Feature Fusion To incorporate our contextual RoI features $\boldsymbol{x}_{context}$ into region-based detection pipeline, the simplest way is fusing them with the original RoI features extracted from the feature pyramid network (FPN) [23] with element addition. Formally, let $\boldsymbol{x}_{FPN}$ denote the original RoI feature extracted from FPN, and $\boldsymbol{x}_{fusion}$ denote the fused RoI feature, then we have:

$$\boldsymbol{x}_{fusion} = \boldsymbol{x}_{context} + \boldsymbol{x}_{FPN} \in \mathbb{R}^{d \times 7 \times 7}. \tag{6}$$

As shown in Fig. 2, the fused feature map $\boldsymbol{x}_{fusion}$ is then fed into the $2fc$ detection head to produce refined bounding boxes and classification scores.

Confidence Fusion We also consider a simple confidence fusion strategy which is complementary to feature fusion. We apply the $2fc$ head on our hierarchical contextual RoI feature $\boldsymbol{x}_{context}$ to produce a classification confidence (logits), and then fuse it with that from the corresponding FPN RoI feature $\boldsymbol{x}_{FPN}$ by addition. Formally, let $\hat{\boldsymbol{y}}_{fusion}$ denote the fused the confidence:

$$\hat{\boldsymbol{y}}_{fusion} = f_{2fc}(\boldsymbol{x}_{context}) + f_{2fc}(\boldsymbol{x}_{FPN}) \in \mathbb{R}^{C}. \tag{7}$$

The fused confidence is transformed by a soft-max layer to produce a novel classification score.

For each proposal, the classification score $\hat{\boldsymbol{y}}_{fusion}$, paired with the refined bounding box predicted the FPN RoI feature, forms another prediction in parallel with the prediction from the feature fusion branch. It is worth mentioning that the weights of the $2fc$ head applied on different RoI features are shared.

Inferences Our early-and-late fusion strategy produces two different predictions for a single object proposal. To obtain the final result, as shown by the pipeline in Fig. 2, we firstly collect all the boxes and confidences from two prediction branches (*i.e.*, feature fusion and confidence fusion), and then perform NMS over all these boxes. Furthermore, as demonstrated later in experiments, while our two fusion strategies are complementary during training, using only one prediction branch during inference will not cause obvious performance drop but reduce computational cost. However, the performance by only using one fusion strategy for training is inferior to that by using two fusion strategies together.

Loss Function The whole network is trained end-to-end, and the overall loss is computed as follows:

$$\mathcal{L} = \mathcal{L}_{feat} + \mathcal{L}_{conf} + \mathcal{L}_{mll} + \mathcal{L}_{rpn} \,, \tag{8}$$

where $\mathcal{L}_{feat}$ and $\mathcal{L}_{conf}$ are the losses for the feature fusion and confidence fusion branches, respectively. All loss terms are considered equally important, without extra hyper-parameters to characterize the trade-off between them, which reveals HCE is generalized and not tricky.

4 Experiments

We conduct extensive experiments on the MS-COCO 2017 dataset [25] to demonstrate the effectiveness and generalization ability of our hierarchical context embedding framework. MS-COCO 2017 is the most popular benchmark for general object detection, which contains 80 object categories, 118K images for training, 5K images for validation (`val`) and 20K for testing (`test-dev`). We report the standard COCO-style Average Precision (AP) with different IoU thresholds from 0.5 to 0.95 with an interval of 0.05 as metric. All models are trained on COCO training set and evaluate on the `val` set. For fair comparisons with the state-of-the-art, we also report the results on the `test-dev` set.

4.1 Implementation Details

We implement our method and re-implement all baseline methods based on MMDetection codebase [5]. The re-implementations of the baselines strictly follow the default settings of MMDetection. Images are resized such that the short edge has 800 pixels while the long edge has less than 1333 pixels. We use no data augmentation except horizontal flipping for training. The ResNet is exploited as backbone, which is pre-trained on ImageNet [10]. Models are trained in a batch size of 16 on 8 GPUs. We train all models with SGD optimizer for 12 epochs in the total, with the initial learning rate as 0.02 and decreased by a factor of 0.1 at 8th epoch and 11th epoch. Weight decay and momentum are set as 0.0001 and 0.9, respectively. We also adopt the linear warming up strategy to begin the training of our model.

Table 1. Compared with baselines (FPN [23], Mask R-CNN [18] and Cascade R-CNN [4]) on MS-COCO 2017 `val`. "HCE" denotes that the models are trained and inferred on both feature fusion and confidence fusion. Clearly, our HCE framework achieves consistent accuracy gains overall all baseline detectors on all evaluation metrics.

Backbone	Method	HCE	AP	AP^{50}	AP^{75}	AP^S	AP^M	AP^L
ResNet-50-FPN	FPN		36.3	58.3	39.1	21.6	40.2	46.9
		✓	**38.4**	**61.0**	**41.8**	**22.9**	**42.5**	**49.1**
	Mask R-CNN		37.3	59.1	40.3	22.2	41.1	48.3
		✓	**38.8**	**61.3**	**42.1**	**23.2**	**42.8**	**49.7**
	Cascade R-CNN		40.5	58.7	44.1	22.3	43.6	53.8
		✓	**41.7**	**60.5**	**45.0**	**23.4**	**44.9**	**55.2**
ResNet-101-FPN	FPN		38.3	60.1	41.7	22.8	42.8	49.8
		✓	**40.0**	**62.3**	**43.4**	**24.0**	**44.1**	**51.9**
	Mask R-CNN		39.4	60.9	43.0	23.3	43.7	51.5
		✓	**40.5**	**62.6**	**44.0**	**24.4**	**44.5**	**53.4**
	Cascade R-CNN		41.9	60.1	45.7	23.2	45.9	56.2
		✓	**43.0**	**61.6**	**46.9**	**24.6**	**46.6**	**57.4**

4.2 Comparisons with Baselines

To demonstrate the generality of our HCE framework, we consider three well-known region-based object detectors as our baseline systems, including Feature Pyramid Network (FPN) [23], Mask R-CNN [18] and Cascade R-CNN [4]. All detectors are instantiated with two different backbones, *i.e.*, ResNet-50 and ResNet-101 with FPN. Integrating our framework with Mask R-CNN and Cascade R-CNN is as straightforward as with FPN. For example, we apply our framework within each training stage of Cascade R-CNN.

Comparison results on MS-COCO 2017 `val` are shown in Table 1. Our HCE framework achieves consistent accuracy gains overall all baseline detectors on all evaluation metrics. Specifically, without the bells and whistles, our method improves 2.1% and 1.7% AP for FPN with ResNet-50 and ResNet-101 backbones, respectively. While for more advanced Mask R-CNN and Cascade R-CNN, our method also brings more than 1% AP improvement on both ResNet-50 and ResNet-101 backbones, *e.g.,* improving the AP for Mask R-CNN with ResNet-50-FPN from 37.3% to 38.8%.

Additionally, it can be observed that our improvements for Mask R-CNN and Cascade R-CNN baselines are not as significant as FPN. We conjecture that this is because Mask R-CNN and Cascade R-CNN themselves integrate mechanisms for better feature learning, which might overlap with the performance gains with our method. Specifically, Mask R-CNN benefits from extra accurate instance-level mask supervisions, while Cascade R-CNN enjoys IoU-specific multi-stage training to progressively refine object proposals and learn discriminative features for IoU-specific proposals. However, even in these cases, our method can also obtain +1% AP improvement over these competing baseline methods.

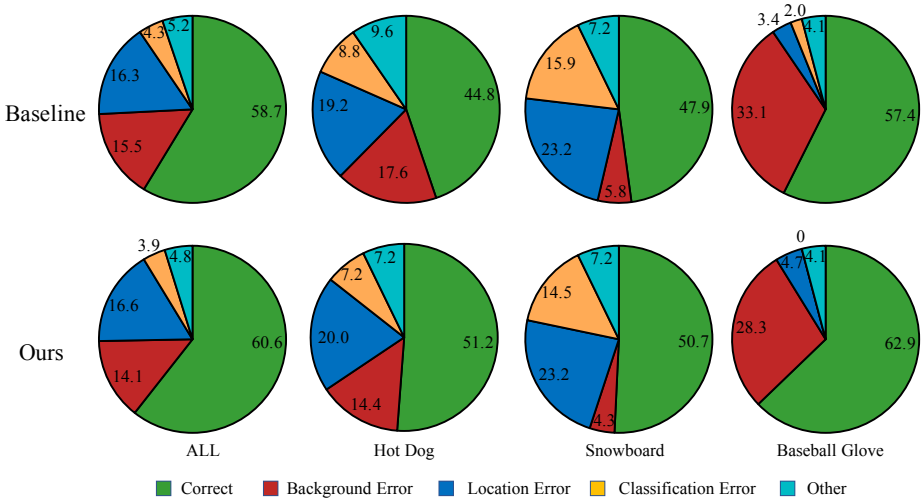


Figure 4. Error Analyses: These illustrations show the percentage of different error types in the top N detections (N = # objects in that category).

4.3 Error Analyses

In the following, we perform error analyses to further understand in what aspects our HCE framework improves the region-based object detectors. Following the settings of [31], we choose the top N predictions for each category during inference time. Each prediction is classified based on the type of error:

- Correct: correct class and IOU > 0.5
- Location Error: correct class and $0.1 <$ IOU < 0.5
- Background Error: IOU < 0.1 for any object
- Classification Error: class is wrong and IOU > 0.5
- Other: class is wrong and $0.1 <$ IOU < 0.5

We compare different error types between the FPN baseline and our method with ResNet-50 as backbone on MS-COCO 2017 `val`. Fig. 4 shows the results of each error type averaged across all 80 categories, and each error type for "`hot dog`", "`snowboard`" and "`baseball glove`" which are highly dependent on context information. Obviously, our method can effectively improve the classification ability of region-based detector and reduce the background errors to a large extent, without compromising the localization performance or increasing other type of errors. Our improvements are particularly noticeable for context-dependent object categories. For example, the (normalized) correctly recognized instances of "`hot dog`" increase from 44.8% to 51.2%, while the background false positive detections reduce from 17.6% to 14.4%. These observations validate that our HCE framework can indeed improve the classification ability.

Table 2. Impacts of different context embedding operations on MS-COCO 2017 `val`. "MLL" means we leverage the image-level categorical embedding module to advance the learning of context-dependent categories. "Instance" and "Global" denotes that we utilize instance-level (cf. Eq (3)) or global-level (cf. Eq (4)) contextual features to further improve the region-wise detection head.

Method	MLL	Instance	Global	AP	AP^{50}	AP^{75}	AP^S	AP^M	AP^L
FPN				36.3	58.3	39.1	21.6	40.2	46.9
	✓			36.8	58.9	39.7	21.9	40.5	47.2
	✓	✓		37.8	59.9	40.9	22.2	41.4	48.9
	✓	✓	✓	**38.4**	**61.0**	**41.8**	**22.9**	**42.5**	**49.1**

Table 3. Effects of different fusion strategies during *training*, evaluated by detection performance on MS-COCO 2017 `val`. The models share the same backbone network ResNet50-FPN. "FF Train" means that we apply feature fusion (FF) for training, while "CF Train" means confidence fusion (CF) are applied for training.

Method	FF Train	CF Train	AP	AP^{50}	AP^{75}	AP^S	AP^M	AP^L
FPN			36.8	58.9	39.7	21.9	40.5	47.2
	✓		37.6	60.3	40.7	22.5	41.4	48.2
		✓	37.4	60.2	40.1	23.0	41.1	47.6
	✓	✓	**38.4**	**61.0**	**41.8**	**22.9**	**42.5**	**49.1**

4.4 Ablation Studies

In this section, we conduct three series of ablation experiments to analyze the proposed method, using ResNet-50 as backbone on MS-COCO 2017 `val`.

Context Embedding Operations We first investigate the impacts of different context embedding operations in our HCE framework. Specifically, there are three context embedding operations involved in our framework. Firstly, the image-level categorical embedding module employs multi-label learning (denoted as "MLL") to embed global-level context to advance the learning of context-dependent categories. Then, for further improving region-based classification, both the context-embedded instance-level feature (denoted as "Instance") and the context-aggregated global-level feature (denoted as "Global") are combined to generate hierarchical contextual RoI feature.

Table 2 shows the performance improvements by progressively integrating more context embedding operations. Solely applying MLL on the detection backbone gives 0.5% AP improvement. This verifies that image-level categorical embedding advances the feature learning for context-dependent object categories. Then, the context-embedded instance-level feature which can be directly utilized by the detection head brings another 1.0% AP improvement. Finally, global-level context embedding for contextual RoI feature improves 0.6% AP. These results suggest that the context embedding operations in our framework are complementary with each other.

Table 4. Effects of different fusion strategies in testing, which are evaluated by the inference time and detection performance on MS-COCO 2017 `val`. Note that all models are trained with both fusion strategies. "FF Test" denotes that we evaluate the feature fusion (FF) strategy during inference, while "CF Test" means the results are evaluated by confidence fusion (CF) strategy. Inference speed is evaluated on a single 1080ti GPU.

Method	FF Test	CF Test	Speed	AP	AP^{50}	AP^{75}	AP^S	AP^M	AP^L
FPN			0.087s	36.3	58.3	39.1	21.6	40.2	46.9
	✓		0.090s	38.2	60.8	41.5	22.6	42.2	49.0
		✓	0.094s	38.3	60.8	41.6	22.8	42.3	49.0
	✓	✓	0.100s	**38.4**	**61.0**	**41.8**	**22.9**	**42.5**	**49.1**

Fusion Strategies in Training We consider the proposed two fusion strategies, feature fusion and confidence fusion, are complementary to each other. To verify this, we evaluate the performance by training the model with feature fusion and confidence fusion individually, as well as both of them. Table 3 shows the results of different fusion strategies. Specifically, "FF Train" means that we apply feature fusion (FF) for training, while "CF Train" means confidence fusion (CF) are applied for training. Utilizing feature fusion and confidence fusion individually for training can outperform the baseline (FPN with MLL) by 0.8% and 0.6% AP, respectively. Training with both fusion strategies achieves the best result, and is clearly better than using each individual fusion strategy separately.

Fusion Strategies in Testing We also evaluate each fusion strategy independently during inference, with all HCE models trained with both fusion strategies. Table 4 shows the results of each fusion strategy and the combined fusion strategies. "FF Test" denotes that we evaluate the feature fusion (FF) strategy during inference, while "CF Test" means that the results are evaluated by confidence fusion (CF) strategy. We can see that once the model is trained with both fusion strategies, using only one fusion branch for inference will not cause obvious accuracy drop, but brings computational economy. For example, using the feature fusion branch for inference adds very minimal time cost (0.003s) to the baseline, but increases the AP from 36.3% to 38.2%. These results also prove the complementarity of the proposed two fusion strategies.

4.5 Comparisons with State-of-the-art

We compare our proposed method with state-of-the-art on MS-COCO 2017 `test-dev`. For fair comparisons, we report the performance of all methods with single-model inference. Specifically, we apply our method on FPN, Mask R-CNN and Cascade R-CNN in 2× training scheme without bells and whistles. Table 5 shows all comparison results.

Our hierarchical context embedding framework, when integrated with FPN, Mask R-CNN and Cascade R-CNN object detectors, consistently outperforms state-of-the-art object detectors using the same backbone network. For fairly

Table 5. Comparisons with the state-of-the-art single-model detectors on MSCOCO 2017 `test-dev`. "*" denotes using tricks (with bells and whistles) during inference.

Method	Backbone	AP	AP^{50}	AP^{75}	AP^S	AP^M	AP^L
YOLOv3 [32]	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9
SSD513 [27]	Res101	31.2	50.4	33.3	10.2	34.5	49.8
RetinaNet [24]	Res101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
FCOS [35]	Res101-FPN	41.5	60.7	45.0	24.4	44.8	51.6
FPN [23]	Res101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Mask R-CNN [18]	Res101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
Cascade R-CNN [4]	Res101-FPN	42.8	62.1	46.3	23.7	45.5	55.2
Deformable R-FCN* [9]	Aligned-Inception-ResNet	37.5	58.0	40.8	19.4	40.1	52.5
DCNv2* [40]	Res101-DeformableV2	46.0	**67.9**	**50.8**	**27.8**	49.1	**59.5**
IoU-Net [19]	Res101-FPN	40.6	59.0	–	–	–	–
TridentNet [22]	Res101	42.7	63.6	46.5	23.9	46.6	56.6
Cascade +Rank-NMS [34]	Res101-FPN	43.2	61.8	47.0	24.6	46.2	55.4
HCE FPN	Res101-FPN	41.0	63.5	44.7	23.4	44.2	52.2
HCE Mask R-CNN	Res101-FPN	41.6	63.9	45.4	23.7	44.7	53.1
HCE Cascade R-CNN	Res101-FPN	44.1	63.2	47.9	25.2	46.9	57.0
HCE Cascade R-CNN*	Res101-FPN	**46.5**	65.6	50.6	27.4	**49.9**	59.4

comparisons with Deformable R-FCN* and DCNv2* which adopt multi-scale 3x training scheme and multi-scale testing, we follow the same experimental setting to train our HCE Cascade R-CNN*. It gives an AP of 46.5%, which surpasses R-FCN* and DCNv2*. These results demonstrate the superior performance of the proposed context embedding framework.

5 Conclusions

In this paper, we investigated the limitation of context information on conventional region-based detectors, and proposed a novel and effective Hierarchical Context Embedding (HCE) framework to facilitate the classification ability of current region-based detectors. Comprehensive experiments demonstrated the consistent outperforming accuracy on almost all existing mainstream region-based detectors, include FPN, Mask R-CNN and Cascade R-CNN. In the future, we will concentrate in extending the usage scope of our HCE framework and adapting it to one-stage detection paradigm.

Acknowledgements: Z.-M. Chen's contribution was made when he was an intern in Megvii Research Nanjing. This research was supported by the National Key Research and Development Program of China under Grant 2017YFA0700800, the National Natural Science Foundation of China under Grants 61772257 and the Fundamental Research Funds for the Central Universities 020914380080.

References

1. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with convex clustering. In: CVPR. pp. 1081–1089 (2015)
2. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: CVPR. pp. 2846–2854 (2016)
3. Byeon, W., Wang, Q., Kumar Srivastava, R., Koumoutsakos, P.: Contextvp: Fully context-aware video prediction. In: ECCV. pp. 753–769 (2018)
4. Cai, Z., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. In: CVPR. pp. 6154–6162 (2018)
5. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
6. Chen, Z., Huang, S., Tao, D.: Context refinement for object detection. In: ECCV. pp. 71–86 (2018)
7. Cinbis, R.G., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. TPAMI **39**(1), 189–203 (2016)
8. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: Object detection via region-based fully convolutional networks. In: NIPS. pp. 379–387 (2016)
9. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: ICCV. pp. 764–773 (2017)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
11. Ebersbach, M., Herms, R., Eibl, M.: Fusion methods for icd10 code classification of death certificates in multilingual corpora. In: CLEF (2017)
12. Galleguillos, C., Belongie, S.: Context based object categorization: A critical survey. CVIU **114**(6), 712–722 (2010)
13. Girshick, R.: Fast R-CNN. In: ICCV. pp. 1440–1448 (2015)
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. pp. 580–587 (2014)
15. Gunes, H., Piccardi, M.: Affect recognition from face and body: early fusion vs. late fusion. In: SMC. vol. 4, pp. 3437–3443. IEEE (2005)
16. Harzallah, H., Jurie, F., Schmid, C.: Combining efficient object localization and image classification. In: CVPR. pp. 237–244 (2009)
17. He, J., Deng, Z., Zhou, L., Wang, Y., Qiao, Y.: Adaptive pyramid context network for semantic segmentation. In: CVPR. pp. 7519–7528 (2019)
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV. pp. 2961–2969 (2017)
19. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: ECCV. pp. 784–799 (2018)
20. Kantorov, V., Oquab, M., Cho, M., Laptev, I.: Contextlocnet: Context-aware deep network models for weakly supervised localization. In: ECCV. pp. 350–365 (2016)
21. Li, X., Wu, J., Lin, Z., Liu, H., Zha, H.: Recurrent squeeze-and-excitation context aggregation net for single image deraining. In: ECCV. pp. 254–269 (2018)
22. Li, Y., Chen, Y., Wang, N., Zhang, Z.: Scale-aware trident networks for object detection. In: ICCV. pp. 6054–6063 (2019)
23. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)

24. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755 (2014)
26. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. IJCV **128**(2), 1–30 (2019)
27. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: ECCV. pp. 21–37. Springer (2016)
28. Luo, W.: Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In: NIPS (2016)
29. Qu, L., Tian, J., He, S., Tang, Y., Lau, R.W.: Deshadownet: A multi-context embedding deep network for shadow removal. In: CVPR. pp. 4067–4075 (2017)
30. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: ICCV. pp. 1–8 (2007)
31. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016)
32. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 pp. 1–6 (2018)
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
34. Tan, Z., Nie, X., Qian, Q., Li, N., Li, H.: Learning to rank proposals for object detection. In: ICCV. pp. 8273–8281 (2019)
35. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. In: ICCV. pp. 9627–9636 (2019)
36. Uijlings, J., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV **104**(2), 154–171 (2013)
37. Wang, Z., Chen, T., Li, G., Xu, R., Lin, L.: Multi-label image recognition by recurrently discovering attentional regions. In: ICCV. pp. 464–472 (2017)
38. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: CBAM: Convolutional block attention module. In: ECCV. pp. 3–19 (2018)
39. Xu, C.D., Zhao, X.R., Jin, X., Wei, X.S.: Exploring categorical regularization for domain adaptive object detection. In: CVPR. pp. 11724–11733 (2020)
40. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: CVPR. pp. 9308–9316 (2019)