

Robust Object Detection under Occlusion with Context-Aware CompositionalNets

Angtian Wang* Yihong Sun* Adam Kortylewski† Alan Yuille†
Johns Hopkins University

Abstract

Detecting partially occluded objects is a difficult task. Our experimental results show that deep learning approaches, such as Faster R-CNN, are not robust at object detection under occlusion. Compositional convolutional neural networks (CompositionalNets) have been shown to be robust at classifying occluded objects by explicitly representing the object as a composition of parts. In this work, we propose to overcome two limitations of CompositionalNets which will enable them to detect partially occluded objects: 1) CompositionalNets, as well as other DCNN architectures, do not explicitly separate the representation of the context from the object itself. Under strong object occlusion, the influence of the context is amplified which can have severe negative effects for detection at test time. In order to overcome this, we propose to segment the context during training via bounding box annotations. We then use the segmentation to learn a context-aware CompositionalNet that disentangles the representation of the context and the object. 2) We extend the part-based voting scheme in CompositionalNets to vote for the corners of the object's bounding box, which enables the model to reliably estimate bounding boxes for partially occluded objects. Our extensive experiments show that our proposed model can detect objects robustly, increasing the detection performance of strongly occluded vehicles from PASCAL3D+ and MS-COCO by 41% and 35% respectively in absolute performance relative to Faster R-CNN.

1. Introduction

In natural images, objects are surrounded and partially occluded by other objects. Recognizing partially occluded objects is a difficult task since the appearances and shapes of occluders are highly variable. Recent work [42, 21] has shown that deep learning approaches are significantly less robust than humans at classifying partially occluded ob-

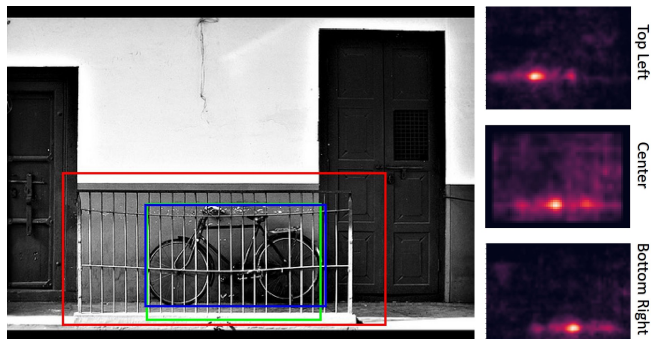


Figure 1: Bicycle detection result for an image of the MS-COCO dataset. Blue box: ground truth; red box: detection result by Faster R-CNN; green box: detection result by context-aware CompositionalNet. Probability maps of three-point detection are to the right. The proposed context-aware CompositionalNet are able to detect the partially occluded object robustly.

jects. Our experimental results show that this limitation of deep learning approaches is even amplified in object detection. In particular, we find that Faster R-CNN is not robust under partial occlusion, even when it is trained with strong data augmentation with partial occlusion. Our experiments show that this is caused by two factors: 1) The proposal network does not localize objects accurately under strong occlusion. 2) The classification network does not classify partially occluded objects robustly. Thus, our work highlights key limitations of deep learning approaches to object detection under partial occlusion that need to be addressed.

In contrast to deep convolutional neural networks (DCNNs), compositional models can robustly classify partially occluded objects from a fixed viewpoint [11, 19] and detect semantic parts of partially occluded object [34, 40]. These models are inspired by the compositionality of human cognition [2, 33, 10, 3] and share similar characteristics with biological vision systems, such as bottom-up sparse compositional encoding and top-down attentional modulations found in the ventral stream [30, 29, 5]. Recent work [20] proposed the Compositional Convolutional Neural Network

*Joint first authors

†Joint senior authors

(CompositionalNet), a generative compositional model of neural feature activations that can robustly classify images of partially occluded objects. This model explicitly represents objects as compositions of parts, which are combined with a voting scheme that enables a robust classification based on the spatial configuration of a few visible parts. However, we find that CompositionalNets as proposed in [20] are not suitable for object detection because of two major limitations: 1) CompositionalNets, as well as other DCNN architectures, do not explicitly disentangle the representation of the context from that of the object. Our experiments show that this has negative effects on the detection performance since context is often biased in the training data (e.g. airplanes are often found in blue background). If objects are strongly occluded, the detection thresholds must be lowered. This in turn increases the influence of the objects' context and leads to false-positive detections in regions with no object (e.g. if a strongly occluded car must be detected, a false airplane might be detected in the sky, seen in Figure 4). 2) CompositionalNets lack mechanisms for robustly estimating the bounding box of the object. Furthermore, our experiments show that region proposal networks do not estimate the bounding boxes robustly when objects are partially occluded.

In this work, we propose to build on and significantly extend CompositionalNets in order to enable them to detect partially occluded objects robustly. In particular, we introduce a detection layer and propose to decompose the image representation as a mixture of context and object representation. We obtain such decomposition by generalizing contextual features in the training data via bounding box annotations. This context-aware image representation enables us to control the influence of the context on the detection result. Furthermore, we introduce a robust voting mechanism to estimate the bounding box of the object. In particular, we extend the part-based voting scheme in CompositionalNets to also vote for two opposite corners of the bounding box in addition to the object center.

Our extensive experiments show that the proposed context-aware CompositionalNets with robust bounding box estimation detect objects robustly even under severe occlusion (Figure 1), increasing the detection performance on strongly occluded vehicles from PASCAL3D+ [38] and MS-COCO [26] by 41% and 35% respectively in absolute performance relative to Faster R-CNN. In summary, we make several important contributions in this work:

1. We propose to decompose the image representation in CompositionalNets as a mixture model of context and object representation. We demonstrate that such **context-aware CompositionalNets** allow for precise control of the influence of the object's context on the detection result, hence, increasing the robustness when classifying strongly occluded objects.

2. We propose a robust **part-based voting mechanism for bounding box estimation** that enables the accurate estimation of an object's bounding box even under severe occlusion.
3. Our experiments demonstrate that context-aware CompositionalNets combined with a part-based bounding box estimation **outperform Faster R-CNN networks at object detection under partial occlusion** by a significant margin.

2. Related Work

Region selection under occlusion. The detection of an object involves the estimation of its location, class and bounding box. While a search over the image can be implemented efficiently, e.g. using a scanning window [24], the number of potential bounding boxes is combinatorial with the number of pixels. The most widely applied approach for solving this problem is to use Region Proposal Networks (RPNs) [13] which enable the learning of fast approaches to object detection [12, 28, 4]. However, our experiments demonstrate that RPNs do not estimate the bounding box of an object correctly under occlusion.

Image classification under occlusion. The classification network in deep object detection approaches is typically chosen to be a DCNN, such as ResNet [14] or VGG [32]. However, recent work [42, 21] has shown that standard DCNNs are significantly less robust to partial occlusion compared to humans. A potential approach to overcome this limitation of DCNNs is to use data augmentation with partial occlusion [8, 39] or top-down cues [36]. However, our experiments demonstrate that data augmentation approaches have only a limited impact on the generalization of DCNNs under occlusion. In contrast to deep learning approaches, generative compositional models [17, 43, 9, 6, 23] have proven to be robust to partial occlusion in the context of detecting object parts [34, 19, 40] and recognizing objects from a fixed viewpoint [11, 22]. Additionally, CompositionalNets [20], which integrate compositional models with DCNN architecture, were shown to be significantly more robust for image classification under occlusion.

Object Detection under occlusion. Sheng [37] *et al.* propose a boosted cascade framework for detecting partially visible objects. However, their approach uses handcrafted features and can only be applied to images where objects are artificially occluded by cutting out image patches. Additionally, a number of deep learning approaches have been proposed for detecting occluded objects [31, 27]; however, these methods require detailed part-level annotations to reconstruct the occluded objects. Xiang and Savarese [35] propose to use 3D models and to treat occlusion as a multi-label classification task. However, in a real-world scenario, the classes of occluders can be difficult to model in 3D and

are often not known a priori (e.g. the particular type of fence in Figure 1). Also, other approaches are based on videos or stereo images [25, 16], however, we focus on object detection in still images. Most related to our work is part-based voting approaches [41, 15] that have proven to work reliably for semantic part detection under occlusion. However, these methods assume a fixed size bounding box which limits their applicability in the context of object detection.

In this work, we extend CompositionalNets to context-aware object detectors with a part-based voting mechanism that can robustly estimate the object’s bounding box even under very strong partial occlusion.

3. Object Detection with CompositionalNets

In Section 3.1 we discuss prior work on CompositionalNets. We propose a generalization of CompositionalNets to detection in Section 3.2, introducing a detection layer and a robust bounding box estimation mechanism. Finally, we introduce context-aware CompositionalNets in Section 3.3, enabling the model to separate the context from the object representation, making it robust to contextual biases in the training data, while still being able to leverage contextual information under strong occlusion.

Notation. The output of a layer l in a DCNN is referenced as *feature map* $F^l = \psi(I, \Omega) \in \mathbb{R}^{H \times W \times D}$, where I is the input image and Ω are the parameters of the feature extractor. *Feature vectors* are vectors in the feature map $f_p^l \in \mathbb{R}^D$ at position p , where p is defined on the 2D lattice of F^l with D being the number of channels in the layer. We omit subscript l in the following for convenience because this layer is fixed a priori in our experiments.

3.1. Prior work: CompositionalNets

CompositionalNets [20] are DCNNs with an inherent robustness to partial occlusion. Their architecture resembles that of a VGG-16 network [32], where the fully connected head is replaced with a differentiable generative compositional model of the feature activations $p(F|y)$ and y is the category of the object. The compositional model is defined as a mixture of von-Mises-Fisher (vMF) distributions:

$$p(F|\Theta_y) = \sum_m \nu_m p(F|\theta_y^m), \quad (1)$$

$$p(F|\theta_y^m) = \prod_p p(f_p|\mathcal{A}_{p,y}, \Lambda), \quad (2)$$

$$p(f_p|\mathcal{A}_{p,y}, \Lambda) = \sum_k \alpha_{p,k,y} p(f_p|\lambda_k), \quad (3)$$

with $\{\nu_m \in \{0, 1\}, \sum_{m=1}^M \nu_m = 1\}$. Here M is the number of mixtures of compositional models and ν_m is a binary assignment variable that indicates which mixture component is active. $\Theta_y = \{\theta_y^m = \{\mathcal{A}_{p,y}^m, \Lambda\} | m = 1, \dots, M\}$ are the

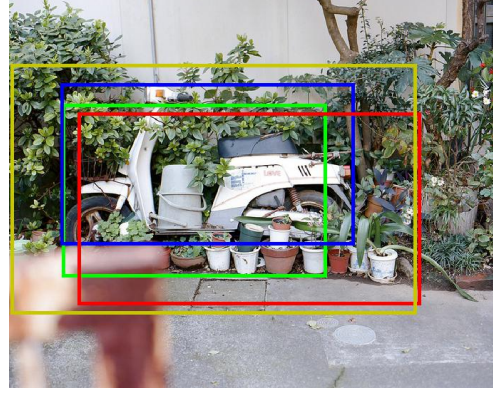


Figure 2: Object detection under occlusion with RPNs and proposed robust bounding box voting. Blue box: ground truth; red box: Faster R-CNN (RPN+VGG); yellow box: RPN+CompositionalNet; green box: context-aware CompositionalNet with robust bounding box voting. Note how the RPN-based approaches fail to localize the object, while our proposed approach can accurately localize the object.

overall compositional model parameters and $\mathcal{A}_y^m = \{\mathcal{A}_{p,y}^m\}$ are the parameters of the mixture components at every position $p \in \mathcal{P}$ on the 2D lattice of the feature map F . In particular, $\mathcal{A}_{p,y}^m = \{\alpha_{p,0,y}^m, \dots, \alpha_{p,K,y}^m | \sum_{k=0}^K \alpha_{p,k,y}^m = 1\}$ are the vMF mixture coefficients, K is the number of mixture components and $\Lambda = \{\lambda_k = \{\sigma_k, \mu_k\} | k = 1, \dots, K\}$ are the parameters of the vMF mixture distributions:

$$p(f_p|\lambda_k) = \frac{e^{\sigma_k \mu_k^T f_p}}{Z(\sigma_k)}, \|f_p\| = 1, \|\mu_k\| = 1, \quad (4)$$

where $Z(\sigma_k)$ is the normalization constant. The model parameters $\{\Omega, \{\Theta_y\}\}$ can be trained end-to-end as described in [20].

Occlusion modeling. Following the approach presented in [19], CompositionalNets can be augmented with an occlusion model. Intuitively, an occlusion model defines a robust likelihood, where at each position p in the image either the object model $p(f_p|\mathcal{A}_{p,y}^m, \Lambda)$ or an occluder model $p(f_p|\beta, \Lambda)$ is active:

$$p(F|\Theta_y^m, \beta) = \prod_p p(f_p, z_p^m = 0)^{1-z_p^m} p(f_p, z_p^m = 1)^{z_p^m}, \quad (5)$$

$$p(f_p, z_p^m = 1) = p(f_p|\beta, \Lambda) p(z_p^m = 1), \quad (6)$$

$$p(f_p, z_p^m = 0) = p(f_p|\mathcal{A}_{p,y}^m, \Lambda) (1 - p(z_p^m = 1)). \quad (7)$$

The binary variables $\mathcal{Z}^m = \{z_p^m \in \{0, 1\} | p \in \mathcal{P}\}$ indicate if the object is occluded at position p for mixture component



Figure 3: Example of robust bounding box voting results. Blue box: ground truth; red box: bounding box by Faster R-CNN; green box: bounding box generated by robustly combining voting results. Our proposed part-based voting mechanism generates probability maps (right) for the object center (cyan point), the top left corner (purple point) and the bottom right corner (yellow point) of the bounding box.

m . The occluder model is defined as a mixture model:

$$p(f_p|\beta, \Lambda) = \prod_n p(f_p|\beta_n, \Lambda)^{\tau_n} \quad (8)$$

$$= \prod_n \left(\sum_k \beta_{n,k} p(f_p|\sigma_k, \mu_k) \right)^{\tau_n}, \quad (9)$$

where $\{\tau_n \in \{0, 1\}, \sum_n \tau_n = 1\}$ indicates which component of the occluder model best explains the data. The parameters of the occluder model β_n can be learned in an unsupervised manner from clustered features of random natural images that do not contain any object of interest.

3.2. Detection with Robust Bounding Box Voting

A natural way of generalizing CompositionalNets to object detection is to combine them with RPNs. However, our experiments in Section 4.1 show that RPNs cannot reliably localize strongly occluded objects. Figure 2 illustrates this limitation by depicting the detection results of Faster R-CNN trained with CutOut [8] (red box) and a combination of RPN+CompositionalNet (yellow box). We propose to address this limitation by introducing a robust part-based voting mechanism to predict the bounding box of an object based on the visible object parts (green box).

CompositionalNets with detection layer. CompositionalNets as introduced in [20] are part-based object representations. In particular, the object model $p(F|\Theta_y)$ is decomposed into a mixture of compositional models $p(F|\theta_y^m)$, where each mixture component represents the object class y from a different pose [20]. During inference, each mixture component accumulates votes from part models $p(f_p|A_{p,y})$ across different spatial positions p of the feature map F . Note that CompositionalNets are learned from images that are cropped based on the bounding box of the object [20]. By making the object centered in the image (see Figure 5), each mixture component $p(F|\theta_y^m)$ can be thought



Figure 4: Influence of context in aeroplane detection under occlusion. Blue box: ground truth; orange box: bounding box by CompositionalNets ($\omega = 0.5$); green box: bounding box by Context-Aware CompositionalNets ($\omega = 0.2$). Probability maps of the object center are on the right. Note how reducing the influence of the context improves the localization response.

of as accumulating votes from the part models for the object being in the center of the feature map.

Based on this intuition, we generalize CompositionalNets to object detection by introducing a detection layer that accumulates votes for the object center over all positions p in the feature map F . In order to achieve this, we propose to compute the object likelihood by scanning. Thus, we shift the feature map w.r.t. the object model along all points p from the 2D lattice of the feature map. This process will generate a spatial likelihood map:

$$R = \{p(F_p|\Theta_y)|p \in \mathcal{P}\}, \quad (10)$$

where F_p denotes the feature map centered at the position p . Using this generalization we can perform object localization by selecting all maxima in R above a threshold t after non-maximum suppression. Our proposed detection layer can be implemented efficiently with modern hardware using convolution-like operations.

Robust bounding box voting. While CompositionalNets can be generalized to localize partially occluded objects using our proposed detection layer, estimating the bounding box of an object under occlusion is more difficult because a significant amount of the object might not be visible (Figure 3). We propose to solve this problem by generalizing the part-based voting mechanism in CompositionalNets to vote for the bounding box corners in addition to the object center. In particular, we learn additional mixture components that model the expected feature activations F around bounding box corners $p(F_p|\Theta_y^c)$, where $c = \{ct, bl, tr\}$ are the object center ct and two opposite bounding box corners $\{bl, tr\}$. Figure 3 illustrates the spatial likelihood maps R^c of all three models. We generate a bounding box using the two points that have maximal likelihood. Note how the bounding boxes can be localized accurately despite large parts of the object being occluded.

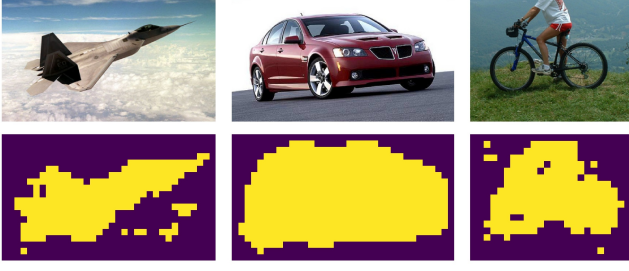


Figure 5: Context segmentation results. A standard CompositionalNet learns a joint representation of the image including the context. Our context-aware CompositionalNet will disentangle the representation of the context from that of the object based on the illustrated segmentation masks.

We discuss how the parameters of all models can be learned jointly in an end-to-end manner in Section 3.4.

3.3. Context-aware CompositionalNets

CompositionalNets, as well as standard DCNNs, do not separate the representation of the context from the object. The context can be useful for recognizing objects due to biases, e.g. aeroplanes are often surrounded by blue sky. Relying too strongly on context can be misleading when objects are strongly occluded (Figure 4), since the detection thresholds must be lowered under strong occlusion. This in turn increases the influence of the objects' context and leads to false-positive detection in regions with no object. Hence, it is important to have control over the influence of contextual cues on the detection result.

In order to gain control over the influence of context, we propose a Context-aware CompositionalNets (CA-CompositionalNets), which separates the representation of the context from the object in the original CompositionalNets by representing the feature map F as a mixture of two models:

$$p(f_p | \mathcal{A}_{p,y}^m, \chi_{p,y}^m, \Lambda) = \omega p(f_p | \chi_{p,y}^m, \Lambda) + \quad (11)$$

$$(1 - \omega) p(f_p | \mathcal{A}_{p,y}^m, \Lambda). \quad (12)$$

Here, $\chi_{p,y}^m$ are the parameters of the context model that is defined to be a mixture of vMF likelihoods (Equation 3). The parameter ω is a prior that controls the trade-off between context and object, which is fixed a priori at test time. Note that setting $\omega = 0.5$ retains the original CompositionalNet as proposed in [20]. Figure 4 illustrates the benefits of reducing the influence of the context on the detection result under partial occlusion. The context parameters $\chi_{p,y}^m$ and object parameters $\mathcal{A}_{p,y}^m$ can be learned from the training data using maximum likelihood estimation. However, this presumes an assignment of the feature vectors f_p in the training data to either the context or the object.

Context segmentation. Therefore, we propose to segment the training images into context and object based on the available bounding box annotation. Here, our assumption is that any feature that has a receptive field outside of the scope of the bounding boxes would be considered as a part of the context. We first randomly extract features that are considered to be context into a population during training. Then, we cluster the population using K-means++ algorithm[1] and receive a dictionary of context feature centers $E = \{e_q \in \mathbb{R}^D | q = 1, \dots, Q\}$. We apply a threshold on the cosine similarity $s(E, f_p) = \max_q [(e_q^T f_p) / (\|e_q\| \|f_p\|)]$ to segment the context and the object in any given training image (Figure 5).

3.4. Training Context-Aware CompositionalNets

We train our proposed CA-CompositionalNet including the robust bounding box voting mechanism jointly end-to-end using backpropagation. Overall, the trainable parameters of our models are $T^c = \{\Omega, \Lambda, \{\Theta_y^c\}, \{\chi_y^c\}\}$ where $c \in \{ct, bl, tr\}$. The loss function has three main objectives: optimizing the parameters of the generative compositional model such that it can explain the data with maximal likelihood (\mathcal{L}_g), while also localizing (\mathcal{L}_{detect}) and classifying (\mathcal{L}_{cls}) the object accurately in the training images. While \mathcal{L}_g is learned from images \tilde{I}^c with feature maps F^c that are centered at $c \in \{c, bl, tr\}$, the other losses are learned from unaligned training images I with feature maps F .

Training Classification with Regularization. We optimize the parameters jointly using SGD:

$$\mathcal{L}_{cls}(y, y') = \mathcal{L}_{class}(y, y') + \mathcal{L}_{weight}(\Omega) \quad (13)$$

where $\mathcal{L}_{class}(y, y')$ is the cross-entropy loss between the network output $y' = \psi(I, \Omega)$ and the true class label y . We use a temperature \mathcal{T} in the softmax classifier: $f(y)_i = \frac{e^{y_i \cdot \mathcal{T}}}{\sum_i e^{y_i \cdot \mathcal{T}}}$. $\mathcal{L}_{weight} = \|\Omega\|_2^2$ is a weight regularization on the DCNN parameters.

Training the generative context-aware CompositionalNet. The overall loss function for training the parameters of the generative context-aware model is composed of two terms:

$$\mathcal{L}_g(F^c, T) = \mathcal{L}_{vmf}(F^c, \Lambda) \quad (14)$$

$$+ \sum_c \sum_p \mathcal{L}_{con}(f_p^c, \mathcal{A}_y^c, \chi_y^c) \quad (15)$$

In order to avoid the computation of the normalization constants $\{Z[\sigma_k]\}$, we assume that the vMF variances $\{\sigma_k\}$ are constant. Under this assumption, the vMF parameters $\{\mu_k\}$ can be optimized with the loss $\mathcal{L}_{vmf}(F, \Lambda) = \mathcal{C} \sum_p \min_k \mu_k^T f_p$, where \mathcal{C} is a constant factor [20]. The parameters of the context-aware model \mathcal{A}_y^c and χ_y^c are learned by optimizing the context loss:

$$\mathcal{L}_{con}(f_p, \mathcal{A}_y^c, \chi_y^c) = \pi_p \mathcal{L}_{mix}(f_p, \mathcal{A}_{p,y}^c) \quad (16)$$

where $\pi_p \in \{0, 1\}$ is a context assignment variable that indicates if a feature vector f_p belongs to the context or to the object model. We estimate the context assignments a priori using segmentation as described in Section 3.3. Given the assignments we can optimize the model parameters $\mathcal{A}_{p,y}^c$ by minimizing [21]:

$$\mathcal{L}_{mix}(F, \mathcal{A}_y^c) = -\sum_p (1 - z_p^\uparrow) \log \left[\sum_k \alpha_{p,k,y}^{m^\uparrow, c} p(f_p | \lambda_k) \right] \quad (17)$$

The context parameters $\chi_{p,y}^c$ can be learned accordingly. Here, z_p^\uparrow and m^\uparrow denote the variables that were inferred in the forward process. Note that the parameters of the occluder model are learned a priori and then fixed.

Training for localization and bounding box localization. We denote the normalized response map of the ground truth class as $X^c \in \mathbb{R}^{H \times W}$ and the ground truth annotation as $\bar{X}^c \in \mathbb{R}^{H \times W}$. The elements of the response map are computed as:

$$x_p^c = \frac{x_{p,\hat{m}}}{\sum_p x_{p,\hat{m}}}, \hat{m} = \underset{m}{\operatorname{argmax}} \max_p p(f_p | \mathcal{A}_{p,y}^m, \chi_{p,y}^m, \Lambda). \quad (18)$$

The ground truth map \bar{X}^c is a binary map where the ground truth position is set to $X^c(c) = 1$ and all other entries are set to zero. The detection loss is then defined as:

$$\mathcal{L}_{detect}(X^c, \bar{X}^c, F, T^c) = 1 - \frac{2 \cdot \sum_p (x_p^c \cdot \bar{x}_p^c)}{\sum_p x_p^c + \sum_p \bar{x}_p^c} \quad (19)$$

End-to-end training. We train all parameters of our model end-to-end with backpropagation. The overall loss function is:

$$\mathcal{L} = \mathcal{L}_{cls}(y, y') + \sum_c (\epsilon_1 \mathcal{L}_g(F^c, T^c) \quad (20)$$

$$+ \epsilon_2 \mathcal{L}_{detect}(X^c, \bar{X}^c, F, T^c)) \quad (21)$$

ϵ_1, ϵ_2 control the trade-off between the loss terms. The optimization process is discussed in more detail in Section 4.

4. Experiments

We perform experiments on object detection under artificially-generated and real-world occlusion.

Datasets. While it is important to evaluate algorithms on real images of partially occluded objects, simulating occlusion enables us to quantify the effects of partial occlusion more accurately. Inspired by the success of datasets with artificially-generated occlusion in image classification [15], we propose to generate an analogous dataset for object detection. In particular, we build on the PASCAL3D+ dataset, which contains 12 classes of unoccluded objects. We synthesize an *OccludedVehiclesDetection* dataset similar to the



Figure 6: Example of images in OccludedVehiclesDetection dataset. Each row shows increasing amounts of context occlusion, whereas each column shows increasing amounts of object occlusion.

dataset proposed in [15] for classification, which contains 6 classes of vehicles at a fixed scale (224 pixels) and various levels of occlusion. The occluders, which include humans, animals and plants, are cropped from the MS-COCO dataset [26]. In an effort to accurately depict real-world occlusions, we superimpose the occluders onto the object, such that the occluders are placed not only inside the bounding box of the objects, but also on the background. We generate the dataset in a total of 9 occlusion levels along two dimensions. We define three levels of object occlusion: FG-L1: 20-40%, FG-L2: 40-60% and FG-L3: 60-80% of the object area occluded. Furthermore, we define three levels of context occlusion around the object: BG-L1: 0-20%, BG-L2: 20-40% and BG-L3: 40-60% of the context area occluded. An example of occlusion levels are showed in Figure 6.

In order to evaluate the tested models on real-world occlusions, we test them on a subset of the MS-COCO dataset. In particular, we extract the same classes of objects and scale as in the OccludedVehiclesDetection dataset from the MS-COCO dataset. We select occluded images and manually separate them into two groups: light occlusion (2 sub-levels) and heavy occlusions (3 sub-levels), with increasing occlusion levels. This dataset is built from images in both Training2017 and Val2017 set of MS-COCO due to a limited amount of heavily occluded objects in MS-COCO Dataset. The light occlusion set contains 2890 images, and the heavy occlusion set contains 788 images. We term this dataset *OccludedCOCO*.

Evaluation. In order to exclusively observe the effects of foreground and background occlusions on various models, we only consider the occluded object in the image for evaluation. Evidently, for the majority of the dataset, there is often only one object of a particular class that is present in the image. This enables us to quantify the effects of levels of occlusions in the foreground and background on the

method	FG L0	FG L1			FG L2			FG L3			Mean
	BG L0	BG L1	BG L2	BG L3	BG L1	BG L2	BG L3	BG L1	BG L2	BG L3	–
Faster R-CNN	98.0	88.8	85.8	83.6	72.9	66.0	60.7	46.3	36.1	27.0	66.5
Faster R-CNN with reg.	97.4	89.5	86.3	89.2	76.7	70.6	67.8	54.2	45.0	37.5	71.1
CA-CompNet via RPN $\omega=0.5$	74.2	68.2	67.6	67.2	61.4	60.3	59.6	46.2	48.0	46.9	60.0
CA-CompNet via RPN $\omega=0$	73.1	67.0	66.3	66.1	59.4	60.6	58.6	47.9	49.9	46.5	59.6
CA-CompNet via BBV $\omega=0.5$	91.7	85.8	86.5	86.5	78.0	77.2	77.9	61.8	61.2	59.8	76.6
CA-CompNet via BBV $\omega=0.2$	92.6	87.9	88.5	88.6	82.2	82.2	81.1	71.5	69.9	68.2	81.3
CA-CompNet via BBV $\omega=0$	94.0	89.2	89.0	88.4	82.5	81.6	80.7	72.0	69.8	66.8	81.4

Table 1: Detection results on the OccludedVehiclesDetection dataset under different levels of occlusions (BBV as in Bounding Box Voting). All models trained on PASCAL3D+ unoccluded dataset except Faster R-CNN with reg. was trained with CutOut. The results are measured by correct AP(%) @IoU0.5, which means only corrected classified images with $IoU > 0.5$ of first predicted bounding box are treated as true-positive. Notice with $\omega = 0.5$, context-aware model reduces to a CompositionalNet as proposed in [20].

method	light occ.		heavy occ.		
	L0	L1	L2	L3	L4
Faster R-CNN	81.7	66.1	59.0	40.8	24.6
Faster R-CNN with reg.	84.3	71.8	63.3	45.0	33.3
Faster R-CNN with occ.	85.1	76.1	66.0	50.7	45.6
CA-CompNet via RPN $\omega=0$	62.0	55.0	49.7	45.4	38.6
CA-CompNet via BBV $\omega=0.5$	83.5	77.1	70.8	51.7	40.4
CA-CompNet via BBV $\omega=0.2$	88.7	82.2	77.8	65.4	59.6
CA-CompNet via BBV $\omega=0$	91.8	83.6	76.2	61.1	54.4

Table 2: Detection results on OccludedCOCO Dataset, measured by AP(%) @IoU0.5. All models are trained on PASCAL3D+ dataset, Faster R-CNN with reg. is trained with CutOut and Faster R-CNN with occ. is trained with images in same dataset but occluded by all levels of occlusion with the same set of occluders.

accuracy of the model predictions. Thus, the means of object detection evaluation must be altered for our proposed occlusion dataset. Given any model, we only evaluate the bounding box proposal with the highest confidence given by the classifier via IoU at 50%.

Runtime. The convolution-like detection layer has an inference time of 0.3s per image.

Training setup. We implement the end-to-end training of our CA-CompositionalNet with the following parameter settings: training minimizes loss described in Equation 20, with $\epsilon_1 = 0.2$ and $\epsilon_2 = 0.4$. We applied the Adam Optimizer [18] with various learning rates of $lr_{vgg} = 2 \cdot 10^{-6}$, $lr_{vc} = 2 \cdot 10^{-5}$, $lr_{mixture\ model} = 5 \cdot 10^{-5}$ and $lr_{corner\ model} = 5 \cdot 10^{-5}$ on different parts of CompositionalNets. The model is trained for a total of 2 epochs with 10600 iteration per epoch. The training costs in total of 3 hours on a machine with 4 NVIDIA TITAN Xp GPUs.

Faster R-CNN is trained for 30000 iterations, with a learning rate, $lr = 1 \cdot 10^{-3}$, and a learning rate decay,

$lr_{decay} = 0.1$. Specifically, the pretrained VGG-16 [32] on the ImageNet dataset [7] was modified in its fully-connected layer to accommodate the experimental settings. In the experiment on OccludedCOCO, we set the threshold of Faster R-CNN to 0, preventing the occluded targets to be ignored due to low confidence and guarantees at least one proposal in the required class.

4.1. Object Detection under Simulated Occlusion

Table 1 shows the results of the tested models on the OccludedVehiclesDetection dataset (see Figure 7 for qualitative results). The models are trained on the images from the original PASCAL3D+ dataset with unoccluded objects.

Faster R-CNN. As we evaluate the performance of the Faster R-CNN, we observe that under low levels of occlusion, the neural network performs well. In mid to high levels of occlusions, however, the neural network fails to detect the objects robustly. When trained with strong data augmentation in terms of partial occlusion using CutOut [8], the detection performance increases under strong occlusion. However, the model still suffers from a 59.9% drop in performance on strong occlusion, compared to the non-occlusion setup. We suspect that the inaccurate prediction is due to two major factors: 1) The Region Proposal Network (RPN) in the Faster R-CNN is not able to predict accurate proposals of objects that are heavily occluded. 2) The VGG-16 classifier cannot successfully classify valid object regions under heavy occlusion.

We proceed to investigate the performance of the region proposals on occluded images. In particular, we replace the VGG-16 classifier in the Faster R-CNN with a standard CompositionalNet classifier [20], which is expected to be more robust to occlusion. From the results in Table 1, we observe two phenomena: 1) In high levels of occlusion, the performance is better than Faster R-CNN. Thus, the CompositionalNet generalizes to heavy occlusions better than the VGG-16 classifier. 2) In low levels of occlu-

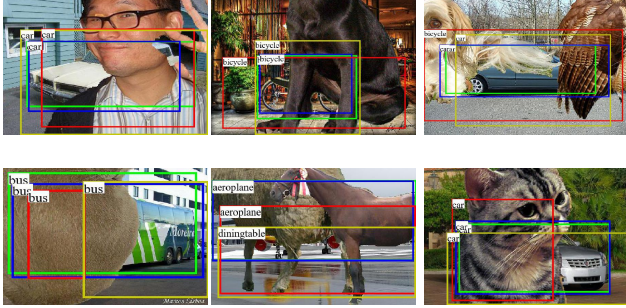


Figure 7: Selected examples of detection results on the OccludedVehiclesDetection dataset. All of these 6 images are the heaviest occluded images (foreground level 3, background level 3). Blue box: ground truth; green box: proposals of CA-CompositionalNet via BB Voting; yellow box: proposals of CA-CompositionalNet via RPN; red box: proposals of Faster R-CNN.

sion, the performance is worse than Faster R-CNN. The proposals generated by the RPN seem to be not accurate enough to be correctly classified, as CompositionalNets are high-precision models and require a precise alignment of the bounding box to the object center.

Effect of robust bounding box voting. Our approach of estimating corners of the bounding box substantially improves the performance of the CompositionalNets, in comparison to the RPN. This further validates our conclusion that the CompositionalNet classifier requires precise proposals to classify objects correctly with partial occlusions.

Effect of context-aware representation. With $\omega = 0.5$, we observe that the precision of the detection decreases. Furthermore, the performance between $\omega = 0.5$ and $\omega = 0$ follows a similar trend over all three levels of foreground occlusions: the performance decreases as the level of background occlusion increases from BG-L1 to BG-L3. This further confirms our understanding of the effects of the context as a valuable source of information in object detection.

4.2. Object Detection under Realistic Occlusion

In the following, we evaluate our model on the OccludedCOCO dataset. As shown in Table 2 and Figure 8, our CA-CompositionalNet with robust bounding box voting outperforms Faster R-CNN and CompNet+RPN significantly. In particular, fully deactivating the context ($\omega = 0$) increases the performance compared to the original model ($\omega = 0.5$), indicating that too much weight is put on the contextual information in the standard CompNets. Furthermore, controlling the prior of the context model to $\omega = 0.2$ reaches an optimal performance under strong occlusion where the context is helpful, but does slightly decrease the performance under low occlusion.

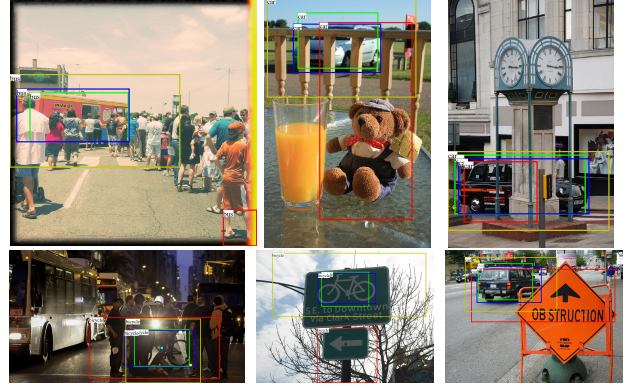


Figure 8: Selected examples of detection results on OccludedCOCO Dataset. Blue box: ground truth; green box: proposals of CA-CompositionalNet via BB Voting; yellow box: proposals of CA-CompositionalNet via RPN; red box: proposals of Faster R-CNN.

5. Conclusion

In this work, we studied the problem of detecting partially occluded objects under occlusion. We found that standard deep learning approaches that combine proposal networks with classification networks do not detect partially occluded objects robustly. Our experimental results demonstrate that this problem has two causes: 1) Proposal networks are more strongly misguided the more context is occupied by the occluders. 2) Classification networks do not classify partially occluded objects robustly. We made the following contributions to resolve these problems:

CompositionalNets for object detection. CompositionalNets have proven to classify partially occluded objects robustly. We generalize CompositionalNets to object detection by extending their architecture with a detection layer.

Robust bounding box voting. We proposed a robust part-based voting mechanism for bounding box estimation by leveraging the unoccluded parts of the object, which enabled the accurate estimation of an object’s bounding box even under severe occlusion.

Context-aware CompositionalNets. CompositionalNets, and other DCNN-based classifiers, do not separate the representation of the context from that of the object. We proposed to segment the object from its context using bounding box annotations and showed how the segmentation can be used to learn a representation in an end-to-end manner that disentangles the context from the object.

Acknowledgement. This work was partially supported by the Swiss National Science Foundation (P2BSP2.181713) and the Office of Naval Research (N00014-18-1-2119).

References

- [1] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007. 5
- [2] Elie Bienenstock and Stuart Geman. Compositionality in neural systems. In *The Handbook of Brain Theory and Neural Networks*, pages 223–226. 1998. 1
- [3] Elie Bienenstock, Stuart Geman, and Daniel Potter. Compositionality, mdl priors, and object recognition. In *Advances in Neural Information Processing Systems*, pages 838–844, 1997. 1
- [4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2
- [5] Rasquinha R.J. Zhang K. Connor C.E. Carlson, E.T. A sparse object coding scheme in area v4. *Current Biology*, 2011. 1
- [6] Jifeng Dai, Yi Hong, Wenze Hu, Song-Chun Zhu, and Ying Nian Wu. Unsupervised learning of dictionaries of hierarchical compositional models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2505–2512, 2014. 2
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 7
- [8] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2, 4, 7
- [9] Sanja Fidler, Marko Boben, and Ales Leonardis. Learning a hierarchical compositional shape vocabulary for multi-class object representation. *arXiv preprint arXiv:1408.5516*, 2014. 2
- [10] Jerry A Fodor, Zenon W Pylyshyn, et al. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. 1
- [11] Dileep George, Wolfgang Lehrach, Ken Kansky, Miguel Lázaro-Gredilla, Christopher Laan, Bhaskara Marthi, Xinghua Lou, Zhaoshi Meng, Yi Liu, Huayan Wang, et al. A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science*, 358(6368):eaag2612, 2017. 1, 2
- [12] Ross Girshick. Fast r-cnn. *IEEE International Conference on Computer Vision*, 2015. 2
- [13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 2
- [14] 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
- [15] Z. Zhang J. Zhu L. Xie J. Wang, C. Xie and A. Yuille. Detecting semantic parts on partially occluded objects. *British Machine Vision Conference*, 2017. 3, 6
- [16] Yin Li Jian Sun and Sing Bing Kang. Symmetric stereo matching for occlusion handling. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3
- [17] Ya Jin and Stuart Geman. Context and hierarchy in a probabilistic image model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2145–2152. IEEE, 2006. 2
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7
- [19] Adam Kortylewski. *Model-based image analysis for forensic shoe print recognition*. PhD thesis, University of Basel, 2017. 1, 2, 3
- [20] Adam Kortylewski, Ju He, Qing Liu, and Alan Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 3, 4, 5, 7
- [21] Adam Kortylewski, Qing Liu, Huiyu Wang, Zhishuai Zhang, and Alan Yuille. Combining compositional models and deep networks for robust object classification under occlusion. *arXiv preprint arXiv:1905.11826*, 2019. 1, 2, 6
- [22] Adam Kortylewski and Thomas Vetter. Probabilistic compositional active basis models for robust pattern recognition. In *British Machine Vision Conference*, 2016. 2
- [23] Adam Kortylewski, Aleksander Wiczeorek, Mario Wieser, Clemens Blumer, Sonali Parbhoo, Andreas Morel-Forster, Volker Roth, and Thomas Vetter. Greedy structure learning of hierarchical compositional models. *arXiv preprint arXiv:1701.06171*, 2017. 2
- [24] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 2
- [25] Ang Li and Zejian Yuan. Symmnet: A symmetric convolutional neural network for occlusion detection. *British Machine Vision Conference*, 2018. 3
- [26] 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. 2, 6
- [27] N. Dinesh Reddy Minh Vo Srinivasa G. Narasimhan. Occlusion-net: 2d/3d occluded keypoint localization using graph networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2
- [28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* 28, 2015. 2
- [29] Chelazzi L. Connor C.E. Conway B.R. Fujita I. Gallant J.L. Lu H. Vanduffel W. Roe, A.W. Toward a unified theory of visual area v4. *Neuron*, 2012. 1
- [30] Emeric E. Stuphorn V. Connor C.E. Sasikumar, D. First-pass processing of value cues in the ventral visual pathway. *Current Biology*, 2018. 1

- [31] Xiao Bian Zhen Lei Shifeng Zhang, Longyin Wen and Stan Z. Li. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. *arXiv preprint arXiv:1807.08407*, 205. 2
- [32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 3, 7
- [33] Ch von der Malsburg. Synaptic plasticity as basis of brain organization. *The Neural and Molecular Bases of Learning*, 411:432, 1987. 1
- [34] Jianyu Wang, Cihang Xie, Zhishuai Zhang, Jun Zhu, Lingxi Xie, and Alan Yuille. Detecting semantic parts on partially occluded objects. *British Machine Vision Conference*, 2017. 1, 2
- [35] Yu Xiang and Silvio Savarese. Object detection by 3d aspectlets and occlusion reasoning. *IEEE International Conference on Computer Vision*, 2013. 2
- [36] Mingqing Xiao, Adam Kortylewski, Ruihai Wu, Siyuan Qiao, Wei Shen, and Alan Yuille. Tdapnet: Prototype network with recurrent top-down attention for robust object classification under partial occlusion. *arXiv preprint arXiv:1909.03879*, 2019. 2
- [37] Shengye Yan and Qingshan Liu. Inferring occluded features for fast object detection. *Signal Processing, Volume 110*, 2015. 2
- [38] Roozbeh Mottaghi Yu Xiang and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. *IEEE Winter Conference on Applications of Computer Vision*, 2014. 2
- [39] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *arXiv preprint arXiv:1905.04899*, 2019. 2
- [40] Zhishuai Zhang, Cihang Xie, Jianyu Wang, Lingxi Xie, and Alan L Yuille. Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1372–1380, 2018. 1, 2
- [41] Jianyu Wang Lingxi Xie Alan L. Yuille Zhishuai Zhang, Cihang Xie. Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3
- [42] Hongru Zhu, Peng Tang, Jeongho Park, Soojin Park, and Alan Yuille. Robustness of object recognition under extreme occlusion in humans and computational models. *CogSci Conference*, 2019. 1, 2
- [43] Long Leo Zhu, Chenxi Lin, Haoda Huang, Yuanhao Chen, and Alan Yuille. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In *European Conference on Computer Vision*, pages 759–773. Springer, 2008. 2