Compositional Convolutional Neural Networks:
A Deep Architecture with Innate Robustness to Partial Occlusion

Adam Kortylewski   Ju He   Qing Liu   Alan Yuille
Johns Hopkins University

Abstract

Recent findings show that deep convolutional neural networks (DCNNs) do not generalize well under partial occlusion. Inspired by the success of compositional models at classifying partially occluded objects, we propose to integrate compositional models and DCNNs into a unified deep model with innate robustness to partial occlusion. We term this architecture Compositional Convolutional Neural Network. In particular, we propose to replace the fully connected classification head of a DCNN with a differentiable compositional model. The generative nature of the compositional model enables it to localize occluders and subsequently focus on the non-occluded parts of the object. We conduct classification experiments on artificially occluded images as well as real images of partially occluded objects from the MS-COCO dataset. The results show that DCNNs do not classify occluded objects robustly, even when trained with data that is strongly augmented with partial occlusions. Our proposed model outperforms standard DCNNs by a large margin at classifying partially occluded objects, even when it has not been exposed to occluded objects during training. Additional experiments demonstrate that CompositionalNets can also localize the occluders accurately, despite being trained with class labels only. The code used in this work is publicly available [1].

1. Introduction

Advances in the architecture design of deep convolutional neural networks (DCNNs) [17, 22, 11] increased the performance of computer vision systems at image classification enormously. However, recent works [38, 14] showed that such deep models are significantly less robust at classifying artificially occluded objects compared to Humans. Furthermore, our experiments show that DCNNs do not classify real images of partially occluded objects robustly. Thus, our findings and those of related works [38, 14] point out a fundamental limitation of DCNNs in terms of generalization under partial occlusion which needs to be addressed.

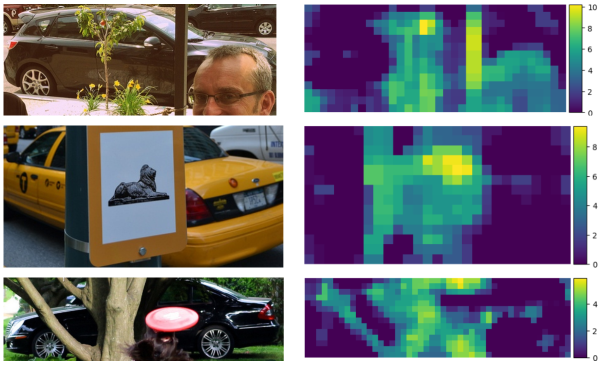

Figure 1: Partially occluded cars from the MS-COCO dataset [20] that are misclassified by a standard DCNN but correctly classified by the proposed CompositionalNet. Intuitively, a CompositionalNet can localize the occluders (occlusion scores on the right) and subsequently focus on the non-occluded parts of the object to classify the image.

One approach to overcome this limitation is to use data augmentation in terms of partial occlusion [6, 35]. However, our experimental results show that after training with augmented data the performance of DCNNs at classifying partially occluded objects still remains substantially worse compared to the classification of non-occluded objects.

Compositionality is a fundamental aspect of human cognition [2, 28, 9, 3] that is also reflected in the hierarchical compositional structure of the ventral stream in visual cortex [34, 27, 21]. A number of works in computer vision showed that compositional models can robustly classify partially occluded 2D patterns [10, 13, 29, 37]. Kortylewski et al. [14] proposed dictionary-based compositional models, a generative model of neural feature activations that can classify images of partially occluded 3D objects more robustly than DCNNs. However, their results also showed that their model is significantly less discriminative at classifying non-occluded objects compared to DCNNs.

In this work, we propose to *integrate* compositional models and DCNNs into a unified deep model with innate robustness to partial occlusion. In particular, we propose to

[1] https://github.com/AdamKortylewski/CompositionalNets

replace the fully-connected classification head of a DCNN with a compositional layer that is regularized to be fully generative in terms of the neural feature activations of the last convolutional layer. The generative property of the compositional layer enables the network to localize occluders in an image and subsequently focus on the non-occluded parts of the object in order to classify the image robustly. We term this novel deep architecture Compositional Convolutional Neural Network (CompositionalNet). Figure 1 illustrates the robustness of CompositionalNets at classifying partially occluded objects, while also being able to localize occluders in an image. In particular, it shows several images of cars that are occluded by other objects. Next to these images, we show occlusion scores that illustrate the position of occluders as estimated by the CompositionalNet. Note how the occluders are accurately localized despite having highly complex shapes and appearances.

Our extensive experiments demonstrate that the proposed CompositionalNet outperforms related approaches by a large margin at classifying partially occluded objects, even when it has not been exposed to occluded objects during training. When trained with data augmentation in terms of partial occlusion the performance increases further. In addition, we perform qualitative and quantitative experiments that demonstrate the ability of CompositionalNets to localize occluders accurately, despite being trained with class labels only. We make several important contributions in this paper:

1. We propose **a differentiable compositional model** that is generative in terms of the feature activations of a DCNN . This enables us to integrate compositional models and deep networks into **compositional convolutional neural networks**, a unified deep model with innate robustness to partial occlusion.

2. While previous works [37, 14, 33, 38] evaluate robustness to partial occlusion on artificially occluded images only, we also **evaluate on real images of partially occluded objects** from the MS-COCO dataset. We demonstrate that CompositionalNets achieve **state-of-the-art results at classifying partially occluded objects** under occlusion.

3. To the best of our knowledge we are the first to **study the task of localizing occluders** in an image and show that CompositionalNets outperform dictionary-based compositional models [14] substantially.

2. Related Work

Classification under partial occlusion. Recent work [38, 14] has shown that current deep architectures are significantly less robust to partial occlusion compared to Humans. Fawzi and Frossard [7] showed that DCNNs are vulnerable to partial occlusion simulated by masking small patches of the input image. Related works [6, 35], have proposed to augment the training data with partial occlusion by masking out patches from the image during training. However, our experimental results in Section 4 show that such data augmentation approaches only have limited effects on the robustness of a DCNN to partial occlusion. A possible explanation is the difficulty of simulating occlusion due to the large variability of occluders in terms of appearance and shape. Xiao et al. [33] proposed TDAPNet a deep network with an attention mechanism that masks out occluded features in lower layers to increase the robustness of the classification against occlusion. Our results show that this model does not perform well on images with real occlusion. In contrast to deep learning approaches, generative compositional models [12, 39, 8, 4, 16] have been shown to be inherently robust to partial occlusion when augmented with a robust occlusion model [13]. Such models have been successfully applied for detecting partially occluded object parts [29, 37] and for recognizing 2D patterns under partial occlusion [10, 15].

Combining compositional models and DCNNs. Liao et al. [19] proposed to integrate compositionality into DCNNs by regularizing the feature representations of DCNNs to cluster during learning. Their qualitative results show that the resulting feature clusters resemble part-like detectors. Zhang et al. [36] demonstrated that part detectors emerge in DCNNs by restricting the activations in feature maps to have a localized distribution. However, these approaches have not been shown to enhance the robustness of deep models to partial occlusion. Related works proposed to regularize the convolution kernels to be sparse [24], or to force feature activations to be disentangled for different objects [23]. As the compositional model is not explicit but rather implicitly encoded within the parameters of the DCNNs, the resulting models remain black-box DCNNs that are not robust to partial occlusion. A number of works [18, 25, 26] use differentiable graphical models to integrate part-whole compositions into DCNNs. However, these models are purely discriminative and thus also are deep networks with no internal mechanism to account for partial occlusion. Kortylewski et al. [14] proposed learn a generative dictionary-based compositional models from the features of a DCNN. They use their compositional model as "backup" to an independently trained DCNN, if the DCNNs classification score falls below a certain threshold.

In this work, we propose to integrate generative compositional models and DCNNs into a unified model that is inherently robust to partial occlusion. In particular, we propose to replace the fully connected classification head with a differentiable compositional model. We train the model parameters with backpropagation, while regularizing the compositional model to be generative in terms of the neural feature activations of the last convolution layer. Our proposed model significantly outperforms related approaches at classifying partially occluded objects while also being able to localize occluders accurately.

3. Compositional Convolutional Neural Nets

In Section 3.1, we introduce a fully generative compositional model and discuss how it can be integrated with DCNNs in an end-to-end system in Section 3.2.

3.1. Fully Generative Compositional Models

We denote a feature map $F^l \in \mathbb{R}^{H\times W\times D}$ as the output of a layer l in a DCNN, with D being the number of channels. A feature vector $f_p^l \in \mathbb{R}^D$ is the vector of features in F^l at position p on the 2D lattice $\mathcal{P}$ of the feature map. In the remainder of this section we omit the superscript l for notational clarity because this is fixed a-priori.

We propose a differentiable generative compositional model of the feature activations $p(F|y)$ for an object class y. This is different from dictionary-based compositional models [14] which learn a model $p(B|y)$, where B is a non-differentiable binary approximation of F. In contrast, we model the real-valued feature activations $p(F|y)$ as a mixture of von-Mises-Fisher (vMF) distributions:

$$p(F|\theta_y) = \prod_p p(f_p|\mathcal{A}_{p,y}, \Lambda) \tag{1}$$

$$p(f_p|\mathcal{A}_{p,y}, \Lambda) = \sum_k \alpha_{p,k,y} p(f_p|\lambda_k), \tag{2}$$

where $\theta_y = \{\mathcal{A}_y, \Lambda\}$ are the model parameters and $\mathcal{A}_y = \{\mathcal{A}_{p,y}\}$ are the parameters of the mixture models at every position $p \in \mathcal{P}$ on the 2D lattice of the feature map F. In particular, $\mathcal{A}_{p,y} = \{\alpha_{p,0,y}, \dots, \alpha_{p,K,y} | \sum_{k=0}^{K} \alpha_{p,k,y} = 1\}$ are the 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 distribution:

$$p(f_p|\lambda_k) = \frac{e^{\sigma_k \mu_k^T f_p}}{Z(\sigma_k)}, \|f_p\| = 1, \|\mu_k\| = 1, \tag{3}$$

where $Z(\sigma_k)$ is the normalization constant. The parameters of the vMF distribution Λ can be learned by iterating between vMF clustering of the feature vectors of all training images and maximum likelihood parameter estimation [1] until convergence. After training, the vMF cluster centers $\{\mu_k\}$ will resemble feature activation patterns that frequently occur in the training data. Interestingly, feature vectors that are similar to one of the vMF cluster centers, are often induced by image patches that are similar in appearance and often even share semantic meanings (see Supplementary A). This property was also observed in a number of related works that used clustering in the neural feature space [30, 19, 29].

The mixture coefficients $\alpha_{p,k,y}$ can also be learned with maximum likelihood estimation from the training images. They describe the expected activation of a cluster center μ_k at a position p in a feature map F for a class y. Note that the spatial information from the image is preserved in the feature maps. Hence, our proposed vMF model (Equation 1) intuitively describes the expected spatial activation pattern of parts in an image for a given class y - e.g. where the tires of a car are expected to be located in an image. In Section 3.2, we discuss how the maximum likelihood estimation of the parameters θ_y can be integrated into a loss function and optimized with backpropagation.

Mixture of compositional models. The model in Equation 1 assumes that the 3D pose of an object is approximately constant in images. This is a common assumption of generative models that represent objects in image space. We can represent 3D objects with a generalized model using mixtures of compositional models as proposed in [14]:

$$p(F|\Theta_y) = \sum_m \nu^m p(F|\theta_y^m), \tag{4}$$

with $\mathcal{V}$=$\{\nu^m \in \{0,1\}, \sum_m \nu^m$=$1\}$ and $\Theta_y = \{\theta_y^m, m = 1, \dots, M\}$. 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. Intuitively, each mixture component m will represent a different viewpoint of an object (see Supplementary B). The parameters of the mixture components $\{\mathcal{A}_y^m\}$ need to be learned in an EM-type manner by iterating between estimating the assignment variables $\mathcal{V}$ and maximum likelihood estimation of $\{\mathcal{A}_y^m\}$. We discuss how this process can be performed in a neural network in Section 3.2.

Occlusion modeling. Following the approach presented in [13], compositional models can be augmented with an occlusion model. The intuition behind an occlusion model is that 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:

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

$$p(f_p, z_p^m{=}1) = p(f_p|\beta, \Lambda)\; p(z_p^m{=}1), \tag{6}$$

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

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 m. The occlusion prior $p(z_p^m{=}1)$ is fixed a-priori. Related works [13, 14] use a single occluder model. We instead use a mixture of several occluder models that are learned in an unsupervised manner:

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

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

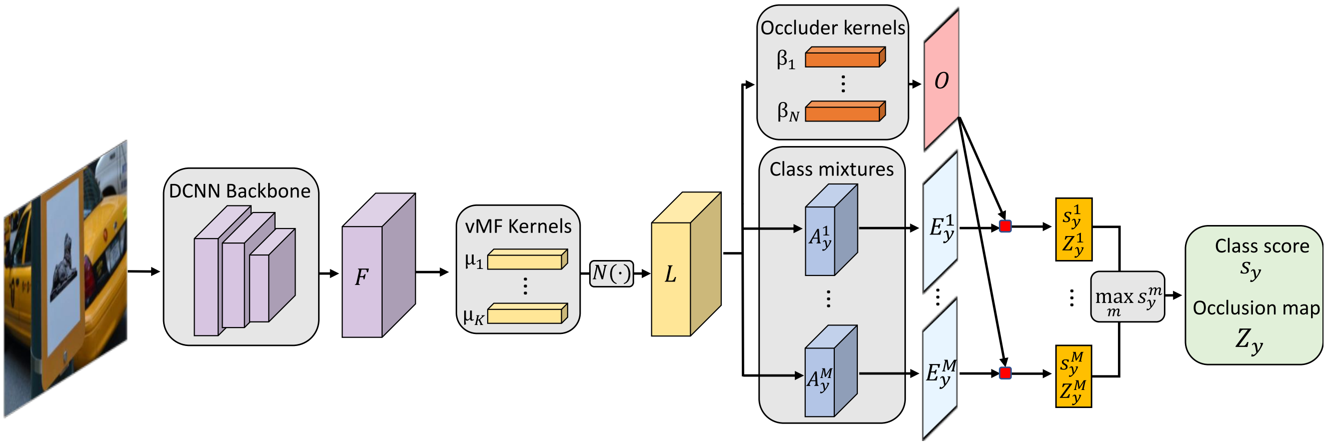

Figure 2: Feed-forward inference with a CompositionalNet. A DCNN backbone is used to extract the feature map F, followed by a convolution with the vMF kernels $\{\mu_k\}$ and a non-linear vMF activation function $\mathcal{N}(\cdot)$. The resulting vMF likelihood L is used to compute the occlusion likelihood O using the occluder kernels $\{\beta_n\}$. Furthermore, L is used to compute the mixture likelihoods $\{E^m_y\}$ using the mixture models $\{A^m_y\}$. O and $\{E^m_y\}$ compete in explaining L (red box) and are combined to compute an occlusion robust score $\{s^m_y\}$. The binary occlusion maps $\{Z^m_y\}$ indicate which positions in L are occluded. The final class score s_y is computed as $s_y = \max_m s^m_y$ and the occlusion map Z_y is selected accordingly.

where $\{\tau_n \in \{0,1\}, \sum_n \tau_n = 1\}$ indicates which occluder model explains the data best. The parameters of the occluder models β_n are learned from clustered features of random natural images that do not contain any object of interest (see Supplementary C). Note that the model parameters β are independent of the position p in the feature map and thus the model has no spatial structure. Hence, the mixture coefficients $\beta_{n,k}$ intuitively describe the expected activation of μ_k anywhere in natural images.

Inference as feed-forward neural network. The computational graph of our fully generative compositional model is directed and acyclic. Hence, we can perform inference in a single forward pass as illustrated in Figure 2.

We use a standard DCNN backbone to extract a feature representation $F = \psi(I, \omega) \in \mathbb{R}^{H\times W\times D}$ from the input image I, where ω are the parameters of the feature extractor. The vMF likelihood function $p(f_p|\lambda_k)$ (Equation 3) is composed of two operations: An inner product $i_{p,k} = \mu_k^T f_p$ and a non-linear transformation $\mathcal{N} = \exp(\sigma_k i_{p,k})/Z(\sigma_k)$. Since μ_k is independent of the position p, computing $i_{p,k}$ is equivalent to a 1×1 convolution of F with μ_k. Hence, the vMF likelihood can be computed by:

$$L = \{\mathcal{N}(F * \mu_k) | k = 1, \dots, K\} \in \mathbb{R}^{H\times W\times K} \tag{10}$$

(Figure 2 yellow tensor). The mixture likelihoods $p(f_p|\mathcal{A}^m_{p,y}, \Lambda)$ (Equation 2) are computed for every position p as a dot-product between the mixture coefficients $\mathcal{A}^m_{p,y}$ and the corresponding vector $l_p \in \mathbb{R}^K$ from the likelihood tensor:

$$E^m_y = \{l_p^T \mathcal{A}^m_{p,y} | \forall p \in \mathcal{P}\} \in \mathbb{R}^{H\times W}, \tag{11}$$

(Figure 2 blue planes). Similarly, the occlusion likelihood can be computed as $O = \{\max_n l_p^T \beta_n | \forall p \in \mathcal{P}\} \in \mathbb{R}^{H\times W}$ (Figure 2 red plane). Together, the occlusion likelihood O and the mixture likelihoods $\{E^m_y\}$ are used to estimate the overall likelihood of the individual mixtures as $s^m_y = p(F|\theta^m_y, \beta) = \sum_p \max(E^m_{p,y}, O_p)$. The final model likelihood is computed as $s_y = p(F|\Theta_y) = \max_m s^m_y$ and the final occlusion map is selected accordingly as $\mathcal{Z}_y = \mathcal{Z}^{\bar{m}}_y \in \mathbb{R}^{H\times W}$ where $\bar{m} = \text{argmax}_m\, s^m_y$.

3.2. End-to-end Training of CompositionalNets

We integrate our compositional model with DCNNs into *Compositional Convolutional Neural Networks* (CompositionalNets) by replacing the classical fully connected classification head with a compositional model head as illustrated in Figure 2. The model is fully differentiable and can be trained end-to-end using backpropagation. Algorithm 1 shows the initialization and training of our CompositionalNets as pseudo code. The trainable parameters of a CompositionalNet are $T = \{\omega, \Lambda, \mathcal{A}_y\}$. We optimize those parameters jointly using stochastic gradient descent. The loss function is composed of four terms:

$$\mathcal{L}(y, y', F, T) = \mathcal{L}_{class}(y, y') + \gamma_1 \mathcal{L}_{weight}(\omega) + \tag{12}$$

$$\gamma_2 \mathcal{L}_{vmf}(F, \Lambda) + \gamma_3 \mathcal{L}_{mix}(F, \mathcal{A}_y). \tag{13}$$

$\mathcal{L}_{class}(y, y')$ is the cross-entropy loss between the network output y' and the true class label y. $\mathcal{L}_{weight} = \|\omega\|_2^2$ is a weight regularization on the DCNN parameters. $\mathcal{L}_{vmf}$ and $\mathcal{L}_{mix}$ regularize the parameters of the compositional model to have maximal likelihood for the features in F. $\{\gamma_1, \gamma_2, \gamma_3\}$ control the trade-off between the loss terms.

Algorithm 1 Training of CompositionalNets

```
Input: Set of training images $I = \{I_1, \dots, I_H\}$,
labels $y = \{y_1, \dots, y_H\}$, VGG backbone $\psi(\cdot, \omega)$,
background images $B = \{B_1, \dots, B_R\}$.
Output: Model parameters $T = \{\omega, \{\mu_k\}, \{\mathcal{A}_y^m\}\}$,$\{\beta_n\}$.
//extract features
$\{F_h\} \leftarrow \psi(\{I_h\}, \omega)$
//initialize vMF kernels by ML
$\{\mu_k\}$←cluster_and_ML($\{f_{h,p}|h$=$\{1,..., H\}, p \in \mathcal{P}\}$)
$\{L_h\}$ ← compute_vMF_likelihood($\{F_h\}, \{\mu_k\}$) [Eq. 10]
//initialize mixture models by ML
$\{\mathcal{A}_y^m\}$ ← cluster_and_ML($\{L_h\}, y$)
$\{\beta_n\}$ ← learn_background_models($B$,$\psi(\cdot, \omega)$,$\{\mu_k\}$)
for #epochs do
    for each image $I_h$ do
        $\{y_h', m^\uparrow, \{z_p^\uparrow\}\}$ ← inference($I_h, T, \{\beta_n\}$)
        $T$ ← optimize($y_h$,$y_h'$,$\omega$,$\{\mu_k\}$,$\mathcal{A}_y^{m^\uparrow}$,$\{z_p^\uparrow\}$) [Sec. 3.2]
```

The vMF cluster centers μ_k are learned by maximizing the vMF-likelihoods (Equation 3) for the feature vectors f_p in the training images. We keep the vMF variance σ_k constant, which also reduces the normalization term $Z(\sigma_k)$ to a constant. We assume a hard assignment of the feature vectors f_p to the vMF clusters during training. Hence, the free energy to be minimized for maximizing the vMF likelihood [31] is:

$$\mathcal{L}_{vmf}(F, \Lambda) = -\sum_p \max_k \log p(f_p|\mu_k) \tag{14}$$

$$= C \sum_p \min_k \mu_k^T f_p, \tag{15}$$

where C is a constant. Intuitively, this loss encourages the cluster centers μ_k to be similar to the feature vectors f_p.

In order to learn the mixture coefficients $\mathcal{A}_y^m$ we need to maximize the model likelihood (Equation 4). We can avoid an iterative EM-type learning procedure by making use of the fact that the the mixture assignment ν_m and the occlusion variables z_p have been inferred in the forward inference process. Furthermore, the parameters of the occluder model are learned a-priori and then fixed. Hence the energy to be minimized for learning the mixture coefficients is:

$$\mathcal{L}_{mix}(F, \mathcal{A}_y) = -\sum_p (1-z_p^\uparrow) \log\Big[\sum_k \alpha_{p,k,y}^{m^\uparrow} p(f_p|\lambda_k)\Big] \tag{16}$$

Here, $z_p^\uparrow$ and $m^\uparrow$ denote the variables that were inferred in the forward process (Figure 2).

4. Experiments

We perform experiments at the tasks of classifying partially occluded objects and at occluder localization.

Datasets. For evaluation we use the *Occluded-Vehicles* dataset as proposed in [30] and extended in [14]. The dataset consists of images and corresponding segmentations of vehicles from the PASCAL3D+ dataset [32] that were synthetically occluded with four different types of occluders: segmented *objects* as well as patches with *constant white color*, *random noise* and *textures* (see Figure 5 for examples). The amount of partial occlusion of the object varies in four different levels: 0% (L0), 20-40% (L1), 40-60% (L2), 60-80% (L3).

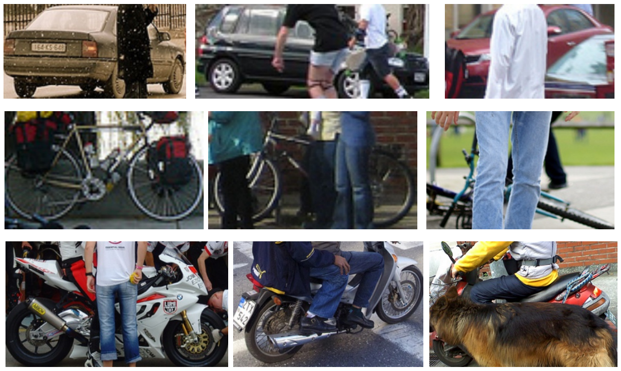

Figure 3: Images from the Occluded-COCO-Vehicles dataset. Each row shows samples of one object class with increasing amount of partial occlusion: 20-40% (Level-1), 40-60% (Level-2), 60-80% (Level-3).

While it is reasonable to evaluate occlusion robustness by testing on artificially generated occlusions, it is necessary to study the performance of algorithms under realistic occlusion as well. Therefore, we introduce a dataset with images of real occlusions which we term *Occluded-COCO-Vehicles*. It consists of the same classes as the Occluded-Vehicle dataset. The images were generated by cropping out objects from the MS-COCO [20] dataset based on their bounding box. The objects are categorized into the four occlusion levels defined by the Occluded-Vehicles dataset based on the amount of the object that is visible in the image (using the segmentation masks available in both datasets). The number of test images per occlusion level are: 2036 (L0), 768 (L1), 306 (L2), 73 (L3). For training purpose, we define a separate training dataset of 2036 images from level L0. Figure 3 illustrates some example images from this dataset.

Training setup. CompositionalNets are trained from the feature activations of a VGG-16 [22] model that is pretrained on ImageNet[5]. We initialize the compositional model parameters $\{\mu_k\}, \{\mathcal{A}_y\}$ using clustering as described in Section 3.1 and set the vMF variance to $\sigma_k = 30, \forall k \in \{1, \dots, K\}$. We train the model parameters $\{\{\mu_k\}, \{\mathcal{A}_y\}\}$ using backpropagation. We learn the parameters of $n = 5$ occluder models $\{\beta_1, \dots, \beta_n\}$ in an unsupervised manner as described in Section 3.1 and keep them fixed throughout the experiments. We set the number of mixture components

PASCAL3D+ Vehicles Classification under Occlusion

Occ. Area	**L0: 0%**	**L1: 20-40%**				**L2: 40-60%**				**L3: 60-80%**				Mean
Occ. Type	-	w	n	t	o	w	n	t	o	w	n	t	o	
VGG	99.2	96.9	97.0	96.5	93.8	92.0	90.3	89.9	79.6	67.9	62.1	59.5	62.2	83.6
CoD[14]	92.1	92.7	92.3	91.7	92.3	87.4	89.5	88.7	90.6	70.2	80.3	76.9	87.1	87.1
VGG+CoD [14]	98.3	96.8	95.9	96.2	94.4	91.2	91.8	91.3	91.4	71.6	80.7	77.3	87.2	89.5
TDAPNet [33]	**99.3**	98.4	98.6	98.5	97.4	96.1	97.5	96.6	91.6	82.1	88.1	82.7	79.8	92.8
CompNet-p4	97.4	96.7	96.0	95.9	95.5	95.8	94.3	93.8	92.5	86.3	84.4	82.1	88.1	92.2
CompNet-p5	**99.3**	98.4	**98.6**	98.4	96.9	98.2	98.3	97.3	88.1	90.1	89.1	83.0	72.8	93.0
CompNet-Multi	**99.3**	**98.6**	**98.6**	**98.8**	**97.9**	**98.4**	**98.4**	**97.8**	**94.6**	**91.7**	**90.7**	**86.7**	**88.4**	**95.4**

Table 1: Classification results for vehicles of PASCAL3D+ with different levels of artificial occlusion (0%,20-40%,40-60%,60-80% of the object are occluded) and different types of occlusion (w=white boxes, n=noise boxes, t=textured boxes, o=natural objects). CompositionalNets outperform related approaches significantly.

MS-COCO Vehicles Classification under Occlusion

Train Data	**PASCAL3D+**					**MS-COCO**					**MS-COCO + CutOut**					**MS-COCO + CutPaste**				
Occ. Area	**L0**	**L1**	**L2**	**L3**	Avg	**L0**	**L1**	**L2**	**L3**	Avg	**L0**	**L1**	**L2**	**L3**	Avg	**L0**	**L1**	**L2**	**L3**	Avg
VGG	97.8	86.8	79.1	60.3	81.0	99.1	88.7	78.8	63.0	82.4	99.3	90.9	87.5	75.3	88.3	99.3	92.3	89.9	80.8	90.6
CoD	91.8	82.7	83.3	76.7	83.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
VGG+CoD	98.0	88.7	80.7	69.9	84.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TDAPNet	98.0	88.5	85.0	74.0	86.4	**99.4**	88.8	87.9	69.9	86.5	99.3	90.1	88.9	71.2	87.4	98.1	89.2	90.5	79.5	89.3
CompNet-p4	96.6	91.8	85.6	76.7	87.7	97.7	92.2	86.6	82.2	89.7	97.8	91.9	87.6	79.5	89.2	98.3	93.8	88.6	84.9	91.4
CompNet-p5	98.2	89.1	84.3	78.1	87.5	99.1	92.5	87.3	82.2	90.3	99.3	93.2	87.6	84.9	91.3	**99.4**	93.9	90.6	**90.4**	93.5
CompNet-Mul	**98.5**	**93.8**	**87.6**	**79.5**	**89.9**	**99.4**	**95.3**	**90.9**	**86.3**	**93.0**	**99.4**	**95.2**	**90.5**	**86.3**	**92.9**	**99.4**	**95.8**	**91.8**	**90.4**	**94.4**

Table 2: Classification results for vehicles of MS-COCO with different levels of real occlusion (L0: 0%,L1: 20-40%,L2 40-60%, L3:60-80% of the object are occluded). The training data consists of images from: PASCAL3D+, MS-COCO as well as data from MS-COCO that was augmented with CutOut and CutPaste. CompositionalNets outperform related approaches in all test cases.

to $M = 4$. The mixing weights of the loss are chosen to be: $\gamma_1 = 0.1$, $\gamma_2 = 5$, $\gamma_3 = 1$. We train for 60 epochs using stochastic gradient descent with momentum $r = 0.9$ and a learning rate of $lr = 0.01$.

4.1. Classification under Partial Occlusion

PASCAL3D+. In Table 1 we compare our CompositionalNets to a VGG-16 network that was pre-trained on ImageNet and fine-tuned with the respective training data. Furthermore, we compare to a dictionary-based compositional model (CoD) and a combination of both models (VGG+CoD) as reported in [14]. We also list the results of TDAPNet as reported in [33]. We report results of CompositionalNets learned from the pool4 and pool5 layer of the VGG-16 network respectively (CompNet-p4 & CompNet-p5), as well as as a multi-layer CompositionalNet (CompNet-Multi) that is trained by combining the output of CompNet-p4 and CompNet-p5. In this setup, all models are trained with non-occluded images ($L0$), while at test time the models are exposed to images with different amount of partial occlusion ($L0$-$L3$).

We observe that CompNet-p4 and CompNet-p5 outperform VGG-16, CoD as well as the combination of both significantly. Note how the CompositionalNets are much more discriminative at level $L0$ compared to dictionary-based compositional models. While CompNet-p4 and CompNet-p5 perform on par with the TDAPNet, CompNet-Multi outperforms TDAPNet significantly. We also observe that CompNet-p5 outperforms CompNet-p4 for low occlusions ($L0$ & $L1$) and for stronger occlusions if the occluders are rectangular masks. However, CompNet-p4 outperforms CompNet-p5 at strong occlusion ($L2$ & $L3$) when the occluders are objects. As argued by Xiao et al. [33] this could be attributed to the fact that occluders with more fine-grained shapes disturb the features in higher layers more severely.

MS-COCO. Table 2 shows classification results under a realistic occlusion scenario by testing on the Occluded-COCO-Vehicles dataset. The models in the first part of the Table are trained on non-occluded images of the PASCAL3D+ data and evaluated on the MS-COCO data. While the performance drops for all models in this transfer learning setting, CompositionalNets outperform the other approaches significantly. Note that combining a DCNN with the dictionary-based compositional model (VGG+CoD) performs well at low occlusion $L0$&$L1$ but lower performance for $L2\&L3$ compared to CoD only.

The second part of the table (MS-COCO) shows the classification performance after fine-tuning on the $L0$ training set of the Occluded-COCO-Vehicles dataset. VGG-16 achieves a similar performance as for the artificial object occluders in Table 1. After fine-tuning, TDAPNet improves at level $L0$ and decreases on average for the levels $L1-3$. Overall it does not significantly benefit from fine-tuning with non-occluded images. The performance of the CompositionalNet increases substantially (p4: 3%, p5: 2.8%, multi: 3.1%) after fine-tuning.

The third and fourth parts of Table 2 (MS-COCO-CutOut & MS-COCO-CutPaste) show classification results after training with strong data augmentation in terms of partial occlusion. In particular, we use CutOut [6] regularization by masking out random square patches of size 70 pixels. Furthermore, we propose a stronger data augmentation method *CutPaste* which artificially occludes the training images in the Occluded-COCO-Vehicles dataset with all four types of artificial occluders used in the Occluded-Vehicles dataset. While data augmentation increases the performance of the VGG network, the model still suffers from strong occlusions and falls below the CompNet-Multi model that was only trained on non-occluded images. TDAPNet does not benefit from data augmentation as much as the VGG network. For CompositionalNets the performance increases further when trained with augmented data. Overall, CutOut augmentation does not have a large effect on the generalization performance of CompositionalNets, while the proposed CutPaste augmentation proves to be stronger. In particular, the CompNet-p5 architecture benefits strongly, possibly because the network learns to extract more reliable higher level features under occlusion.

In summary, the classification experiments clearly highlight the robustness of CompositionalNets at classifying partially occluded objects, while also being highly discriminative when objects are not occluded. Overall CompositionalNets significantly outperform dictionary-based compositional models and other neural network architectures at image classification under artificial as well as real occlusion in all three tested setups - at transfer learning between datasets, when trained with non-occluded images and when trained with strongly augmented data.

4.2. Occlusion Localization

While it is important to classify partially occluded robustly, models should also be able to localize occluders in images accurately. This improves the explainability of the classification process and enables future research e.g. for parsing scenes with mutually occluding objects. Therefore, we propose to test the ability of CompositionalNets and dictionary-based compositional models at occluder localization. We compute the occlusion score as the log-likelihood ratio between the occluder model and the object model: $\log p(f_p|z_p^m{=}1)/p(f_p|z_p^m{=}0)$, where $m = \text{argmax}_m\, p(F|\theta_y^m)$ is the model that fits the data the best.

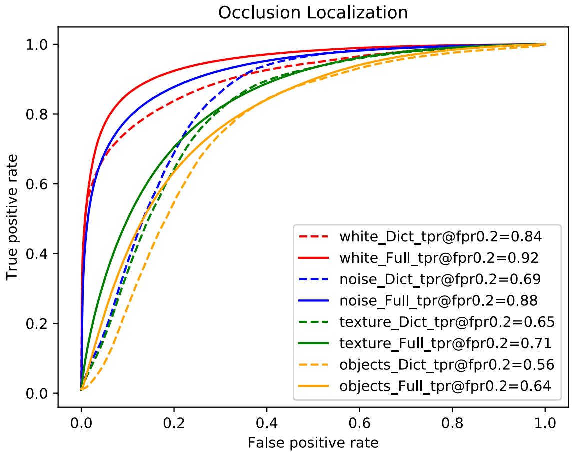

Figure 4: ROC curves for occlusion localization with dictionary-based compositional models and the proposed CompositionalNets averaged over all levels of partial occlusion (L1-L3). CompositionalNets significantly outperform dictionary-based compositional models.

Quantitative results. We study occluder localization quantitatively on the Occluded-Vehicle dataset using the ground truth segmentation masks of the occluders and the objects. Figure 4 shows the ROC curves of CompositionalNets (solid lines) and dictionary-based compositional models (dashed lines) when using the occlusion score to classify each pixel as occluded or non-occluded over all occlusion levels $L1-L3$. We evaluate the localization quality only for images that were correctly classified by each model. The ROC curves show that for both models it is more difficult to localize textured occluders compared to white and noisy occluders. Furthermore, it is more difficult to localize natural object occluders compared to textured boxes, likely because of their fine-grained irregular shape. Overall, CompositionalNets significantly outperform dictionary-based compositional models. At a false acceptance rate of 0.2, the performance gain of CompositionalNets is: 12% (white), 19% (noise), 6% (texture) and 8% (objects).

Qualitative results. Figure 5 qualitatively compares the occluder localization abilities of dictionary-based compositional models and CompositionalNets. We show images of real and artificial occlusions and the corresponding occlusion scores for all positions p of the feature map F. Both models are learned from the `pool4` feature maps of a VGG-16 network. We show more example images in Supplementary D. Note that we visualize the positive values of the occlusion score after median filtering for illustration purposes (see Supplementary E for unfiltered results). We observe that CompositionalNets can localize occluders sig-

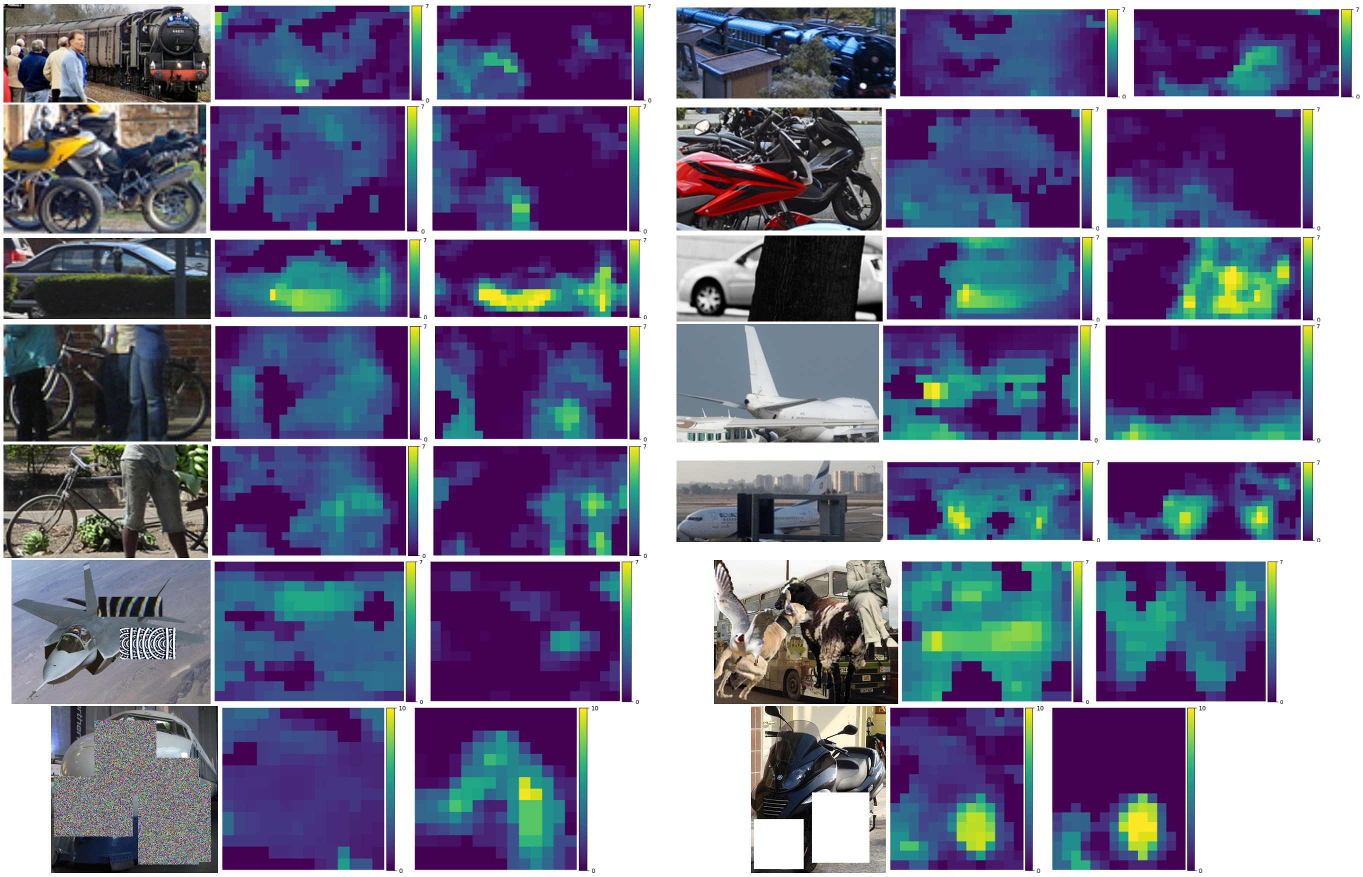

Figure 5: Qualitative occlusion localization results. Each result consists of three images: The input image, and the occlusion scores of a dictionary-based compositional model [14] and our proposed CompositionalNet. Note how our model can localize occluders with higher accuracy across different objects and occluder types for real as well as for artificial occlusion.

nificantly better compared to the dictionary-based compositional model for real as well as artificial occluders. In particular, it seems that dictionary-based compositional models often detect false positive occlusions. Note how the artificial occluders with white and noise texture are better localized by both models compared to the other occluder types.

In summary, our qualitative and quantitative occluder localization experiments clearly show that CompositionalNets can localize occluders more accurately compared to dictionary-based compositional models. Furthermore, we observe that localizing occluders with variable texture and shape is highly difficult, which could be addressed by developing advanced occlusion models.

5. Conclusion

In this work, we studied the problem of classifying partially occluded objects and localizing occluders in images. We found that a standard DCNN does not classify real images of partially occluded objects robustly, even when it has been exposed to severe occlusion during training. We proposed to resolve this problem by integrating compositional models and DCNNs into a unified model. In this context, we made the following contributions:

Compositional Convolutional Neural Networks. We introduce CompositionalNets, a novel deep architecture with innate robustness to partial occlusion. In particular we replace the fully connected head in DCNNs with differentiable generative compositional models.

Robustness to partial occlusion. We demonstrate that CompositionalNets can classify partially occluded objects more robustly compared to a standard DCNN and other related approaches, while retaining high discriminative performance for non-occluded images. Furthermore, we show that CompositionalNets can also localize occluders in images accurately, despite being trained with class labels only.

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] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep):1345–1382, 2005. 3

[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] 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

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1, 2, 7

[7] Alhussein Fawzi and Pascal Frossard. Measuring the effect of nuisance variables on classifiers. Technical report, 2016. 2

[8] 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

[9] Jerry A Fodor, Zenon W Pylyshyn, et al. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. 1

[10] 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

[11] 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. 1

[12] Ya Jin and Stuart Geman. Context and hierarchy in a probabilistic image model. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2145–2152. IEEE, 2006. 2

[13] Adam Kortylewski. *Model-based image analysis for forensic shoe print recognition*. PhD thesis, University_of_Basel, 2017. 1, 2, 3

[14] 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, 3, 5, 6, 8

[15] Adam Kortylewski and Thomas Vetter. Probabilistic compositional active basis models for robust pattern recognition. In *British Machine Vision Conference*, 2016. 2

[16] Adam Kortylewski, Aleksander Wieczorek, 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

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[18] Xilai Li, Xi Song, and Tianfu Wu. Aognets: Compositional grammatical architectures for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6220–6230, 2019. 2

[19] Renjie Liao, Alex Schwing, Richard Zemel, and Raquel Urtasun. Learning deep parsimonious representations. In *Advances in Neural Information Processing Systems*, pages 5076–5084, 2016. 2, 3

[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. 1, 5

[21] Dennis Sasikumar, Erik Emeric, Veit Stuphorn, and Charles E Connor. First-pass processing of value cues in the ventral visual pathway. *Current Biology*, 28(4):538–548, 2018. 1

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 5

[23] Austin Stone, Huayan Wang, Michael Stark, Yi Liu, D Scott Phoenix, and Dileep George. Teaching compositionality to cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5058–5067, 2017. 2

[24] Domen Tabernik, Matej Kristan, Jeremy L Wyatt, and Aleš Leonardis. Towards deep compositional networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3470–3475. IEEE, 2016. 2

[25] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 190–206, 2018. 2

[26] Wei Tang, Pei Yu, Jiahuan Zhou, and Ying Wu. Towards a unified compositional model for visual pattern modeling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2784–2793, 2017. 2

[27] Siavash Vaziri, Eric T Carlson, Zhihong Wang, and Charles E Connor. A channel for 3d environmental shape in anterior inferotemporal cortex. *Neuron*, 84(1):55–62, 2014. 1

[28] Ch von der Malsburg. Synaptic plasticity as basis of brain organization. *The neural and molecular bases of learning*, 411:432, 1987. 1

[29] 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, 3

[30] Jianyu Wang, Zhishuai Zhang, Cihang Xie, Vittal Premachandran, and Alan Yuille. Unsupervised learning of object semantic parts from internal states of cnns by population encoding. *arXiv preprint arXiv:1511.06855*, 2015. 3, 5

[31] Jianyu Wang, Zhishuai Zhang, Cihang Xie, Yuyin Zhou, Vittal Premachandran, Jun Zhu, Lingxi Xie, and Alan Yuille. Visual concepts and compositional voting. *arXiv preprint arXiv:1711.04451*, 2017. 5

[32] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82. IEEE, 2014. 5

[33] 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, 6

[34] Yukako Yamane, Eric T Carlson, Katherine C Bowman, Zhihong Wang, and Charles E Connor. A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature neuroscience*, 11(11):1352, 2008. 1

[35] 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. 1, 2

[36] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836, 2018. 2

[37] 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

[38] 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

[39] Long Leo Zhu, Chenxi Lin, Haoda Huang, Yuanhao Chen, and Alan Yuille. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In *Computer vision–eccv 2008*, pages 759–773. Springer, 2008. 2