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

Anonymous CVPR submission

Paper ID 1120

Abstract

Recent work has shown that deep convolutional neural networks (DCNNs) do not generalize well under partial oc-clusion. Inspired by the success of compositional models at classifying partially occluded objects, we propose to inte-grate compositional models and DCNNs into a unified deep model with innate robustness to partial occlusion. We term this architecture Compositional Convolutional Neural Net-work. In particular, we propose to replace the fully con-nected classification head of a DCNN with a differentiable compositional model. The generative nature of the compo-sitional model enables it to localize occluders and subse-quently 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 DC-NNs do not classify occluded objects robustly, even when trained with data that is strongly augmented with partial occlusions. Our proposed model outperforms standard DC-NNs by a large margin at classifying partially occluded ob-jects, even when it has not been exposed to occluded ob-jects during training. Additional experiments demonstrate that CompositionalNets can also localize the occluders ac-curately, despite being trained with class labels only.

1. Introduction

Advances in the architecture design of deep convolu-tional neural networks (DCNNs) [14, 18, 8] increased the performance of computer vision systems at image classification enormously. However, recent findings [31, 11] have shown that such deep models are significantly less robust at classifying artificially occluded objects compared to Hu-mans. Furthermore, our experiments show that DCNNs do not classify real images of partially occluded objects robustly. Thus, our findings and those of related works [31, 11] 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 [17] 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 [4, 28]. 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.

A number of works showed that compositional models can robustly classify partially occluded 2D patterns [7, 10, 23, 30]. Kortylewski et al. [11] proposed dictionarybased 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 replace the fully-connected classification head of a DCNN with a compositional layer that is regularized to be fully

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

108 generative in terms of the neural feature activations of the 109 last convolutional layer. The generative property of the 110 compositional layer enables the network to localize occlud-111 ers in an image and subsequently focus on the non-occluded 112 parts of the object in order to classify the image robustly. 113 We term this novel deep architecture Compositional Con-114 volutional Neural Network (CompositionalNet). Figure 1 115 illustrates the robustness of CompositionalNets at classify-116 ing partially occluded objects, while also being able to lo-117 calize occluders in an image. In particular, it shows several 118 images of cars that are occluded by other objects. Next to 119 these images, we show occlusion scores that illustrate the 120 position of occluders as estimated by the Compositional-121 Net. Note how the occluders are accurately localized in the 122 occlusion maps despite having highly complex shapes and 123 appearances. 124

Our extensive experiments demonstrate that the pro-125 posed CompositionalNet outperforms related approaches 126 by a large margin at classifying partially occluded objects, 127 even when it has not been exposed to occluded objects dur-128 ing training. When trained with data augmentation in terms 129 of partial occlusion the performance increases further. In 130 addition, we perform qualitative and quantitative experi-131 ments that demonstrate the ability of CompositionalNets 132 to localize occluders accurately, despite being trained with 133 class labels only. We make several important contributions 134 in this paper: 135

- 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 [30, 11, 27, 31] 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 artificial and real 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 [11] substantially.

2. Related Work

Classification under partial occlusion. Recent work [31, 11] has shown that current deep architectures are significantly less robust to partial occlusion compared to Humans. Fawzi and Frossard [5] showed that DCNNs are vulnerable

to partial occlusion simulated by masking small patches of the input image. Related works [4, 28], 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. [27] 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 [9, 32, 6, 2, 13] have been shown to be inherently robust to partial occlusion when augmented with a robust occlusion model [10]. Such models have been successfully applied for detecting partially occluded object parts [23, 30] and for recognizing 2D patterns under partial occlusion [7, 12].

Combining compositional models and DCNNs. Liao et al. [16] proposed to integrate compositionality into DC-NNs 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. [29] 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 [20], or to force feature activations to be disentangled for different objects [19]. As the compositional model is not explicit but rather implicitly encoded within the parameters of the DCNNs, the resulting models remain black-box DC-NNs that are not robust to partial occlusion. A number of works [15, 21, 22] 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. [11] 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 fea-

ture 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 [11] which learn a model p(B|y), where B is a nondifferentiable 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:

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

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

where $\theta_y = \{A_y, \Lambda\}$ are the model parameters and $A_y = \{A_{p,y}\}$ 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, $A_{p,y} = \{\alpha_{p,0,y}, \ldots, \alpha_{p,K,y} | \sum_{k=0}^{K} \alpha_{p,k,y} = 1\}$ are the mixture coefficients, K is the number of mixture components and $\Lambda = \{\lambda_k = \{\sigma_k, \mu_k\} | k = 1, \ldots, K\}$ are the parameters of the vMF distribution:

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

where $Z(\sigma_k)$ is the normalization constant. The parame-ters of the vMF distribution Λ can be learned by iterating between vMF clustering of the feature vectors of all train-ing images and maximum likelihood parameter estimation [1] until convergence. After training, the vMF cluster cen-ters $\{\mu_k\}$ will resemble feature activation patterns that fre-quently occur in the training data. Interestingly, feature vec-tors that are similar to one of the vMF cluster centers, are often induced by image patches that are similar in appear-ance and often even share semantic meanings (see Supple-mentary A). This property was also observed in a number of related works that used clustering in the neural feature space [24, 16, 23].

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 [11]:

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

with $\mathcal{V}=\{\nu^m \in \{0,1\}, \sum_m \nu^m=1\}$ and $\Theta_y = \{\theta_y^m, m = 1, \ldots, M\}$. 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 [10], 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:

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

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

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

The binary variables $Z^m = \{z_p^m \in \{0,1\} | p \in P\}$ 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 [10, 11] use a single occluder model. We instead use a mixture of several occluder models that are learned in an unsupervised manner:

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

$$=\prod_{n}\left(\sum_{k}\beta_{n,k}p(f_{p}|\sigma_{k},\mu_{k})\right)^{\tau_{n}},\qquad(9)$$



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 $\{E_y^m\}$ using the mixture models $\{A_y^m\}$. O and $\{E_y^m\}$ compete in explaining L (red box) and are combined to compute an occlusion robust score $\{s_y^m\}$. The binary occlusion maps $\{Z_y^m\}$ indicate which positions in L are occluded. The final class score s_y is computed as $s_y = \max_m s_y^m$ 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 proposed fully generative compositional model is a directed acyclic graph. 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:

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

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

$$E_y^m = \{ l_p^T \mathcal{A}_{p,y}^m | \forall p \in \mathcal{P} \} \in \mathbb{R}^{H \times W},$$
(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 $\{E_y^m\}$ are used to estimate the overall likelihood of the individual mixtures as $s_y^m =$ $p(F|\theta_y^m, \beta) = \sum_p \max(E_{p,y}^m, O_p)$. The final model likelihood is computed as $s_y = p(F|\Theta_y) = \max_m s_y^m$ and the final occlusion map is selected accordingly as $\mathcal{Z}_y = \mathcal{Z}_y^{\overline{m}} \in$ $\mathbb{R}^{H \times W}$ where $\overline{m} = \operatorname{argmax}_m s_y^m$.

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. The trainable parameters of a CompositionalNet are $T = \{\omega, \Lambda, A_y\}$. We optimize those parameters jointly using stochastic gradient descent. The loss function is composed of four terms:

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

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

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 [25] is:

$$\mathcal{L}_{vmf}(F,\Lambda) = -\sum_{p} \max_{k} \log p(f_p|\mu_k)$$
(14)

$$= C \sum_{p} \min_{k} \mu_{k}^{T} f_{p}, \qquad (15)$$

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:

$$\mathcal{L}_{mix}(F, \mathcal{A}_y) = -\sum_p (1 - z_p^{\uparrow}) \log \left[\sum_k \alpha_{p,k,y}^{m^{\uparrow}} p(f_p | \lambda_k) \right]$$
(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 [24] and extended in [11]. The dataset consists of images and corresponding segmentations of vehicles from the PASCAL3D+ dataset [26] 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).

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 OccludedVehicle dataset. The images were generated by cropping



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).

out objects from the MS-COCO [17] 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 [18] model that is pretrained on ImageNet[3]. We initialize the compositional model parameters $\{\mu_k\}, \{\mathcal{A}_u\}$ using clustering as described in Section 3.1 and set the vMF variance to $\sigma_k = 30, \forall k \in$ $\{1, \ldots, K\}$. We train the three sets of model parameters $\{\omega, \{\mu_k\}, \{A_u\}\}$ using backpropagation. In order to reduce the memory consumption during training, we perform coordinate gradient descent and hence learn one set of variables for one epoch while keeping the other two fixed. We learn the parameters of n = 5 occluder models $\{\beta_1, \ldots, \beta_n\}$ in an unsupervised manner as described in Section 3.1 and keep them fixed throughout the experiments. We set the number of mixture components 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. The code, models and data used for out experiments are publicly available¹.

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.

¹Website will be written here.

CVPR 2020 Submission #1120. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

PASCAL3D+ Vehicles Classification under Occlusion

W

92.0

87.4

91.2

96.1

95.8

98.2

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,

L2: 40-60%

t

89.9

88.7

91.3

96.6

93.8

97.3

0

79.6

90.6

91.4

91.6

92.5

88.1

94.6

W

67.9

70.2

71.6

82.1

86.3

90.1

91.7

n

90.3

89.5

91.8

97.5

94.3

98.3

98.4 98.4 97.8

L1: 20-40%

t

96.5

91.7

96.2

98.5

95.9

98.4

98.8

0

93.8

92.3

94.4

97.4

95.5

96.9

97.9

n

97.0

92.3

95.9

98.9

96.0

98.6

98.6

540 541

Occ. Area

Occ. Type

CoD[11]

VGG+CoD [11]

TDAPNet [27]

CompNet-p4

CompNet-p5

CompNet-Multi

VGG

L0: 0%

99.2

92.1

98.3

99.3

97.4

99.3

99.3

W

96.9

92.7

96.8

98.4

96.7

98.4

98.6

o=natural objects). CompositionalNets outperform related approaches significantly.

5	4	2
5	4	3

- 544 545
- 546
- 547 548
- 549 **550** 551
- 554
- 552 553
- 555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

571

575

578

579

581

582

MS-COCO Vehicles Classification under Occlusion

Train Data		PAS	SCAL:	3D+		MS-COCO					MS-COCO + CutOut					MS-COCO + CutPaste				
Occ. Area	LO	L1	L2	L3	Avg	LO	L1	L2	L3	Avg	LO	L1	L2	L3	Avg	LO	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.

570 Furthermore, we compare to a dictionary-based compositional model (CoD) and a combination of both models 572 (VGG+CoD) as reported in [11]. We also list the results 573 of TDAPNet as reported in [27]. We report results of 574 CompositionalNets learned from the pool4 and pool5 layer of the VGG-16 network respectively (CompNet-p4 & 576 CompNet-p5), as well as as a multi-layer CompositionalNet 577 (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 580 the models are exposed to images with different amount of partial occlusion (L0-L3).

583 We observe that CompNet-p4 and CompNet-p5 outperform VGG-16, CoD as well as the combination of both sig-584 585 nificantly. Note how the CompositionalNets are much more discriminative at level L0 compared to dictionary-based 586 587 compositional models. While CompNet-p4 and CompNet-588 p5 perform on par with the TDAPNet, CompNet-Multi outperforms TDAPNet significantly. We also observe that 589 CompNet-p5 outperforms CompNet-p4 for low occlusions 590 (L0 & L1) and for stronger occlusions if the occluders 591 592 are rectangular masks. However, CompNet-p4 outperforms 593 CompNet-p5 at strong occlusion (L2 & L3) when the occluders are objects. As argued by Xiao et al. [27] this could be attributed to the fact that occluders with more finegrained shapes disturb the features in higher layers more severely.

L3: 60-80%

t

59.5

76.9

77.3

82.7

82.1

83.0

0

62.2

87.1

87.2

79.8

88.1

72.8

88.4

n

62.1

80.3

80.7

88.1

84.4

89.1

90.7 86.7

Mean

83.6

87.1

89.5

92.8

92.2

93.0

95.4

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 PAS-CAL3D+ data and evaluated on the MS-COCO data. While the performance drops for all models in this transfer learning setting, CompositionalNets still outperform the other approaches significantly. Note that the combination of the DCNN with the dictionary-based compositional model (VGG+CoD) has a high performance 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 594 595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

with non-occluded images. The performance of the CompositionalNet increases substantially (p4: 3%, p5: 2.8%, multi: 3.1%) after fine-tuning.
The third performance of The D 2 (MS COCCO C 40).

The third and fourth parts of Table 2 (MS-COCO-CutOut 652 & MS-COCO-CutPaste) show classification results after 653 training with strong data augmentation in terms of partial 654 occlusion. In particular, we use CutOut [4] regulariza-655 tion by masking out random square patches of size 70 pix-656 els. Furthermore, we propose a stronger data augmenta-657 tion method CutPaste which artificially occludes the train-658 ing images in the Occluded-COCO-Vehicles dataset with 659 all four types of artificial occluders used in the Occluded-660 Vehicles dataset. While data augmentation increases the 661 performance of the VGG network, the model still suf-662 fers from strong occlusions and falls below the CompNet-663 Multi model that was only trained on non-occluded im-664 ages. TDAPNet does not benefit from data augmentation 665 as much as the VGG network. For CompositionalNets the 666 performance increases further when trained with augmented 667 data. Overall, CutOut augmentation does not have a large 668 effect on the generalization performance of Compositional-669 Nets, while the proposed CutPaste augmentation proves to 670 be stronger. In particular, the CompNet-p5 architecture ben-671 efits strongly, possibly because the network learns to extract 672 more reliable higher level features under occlusion. 673

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 $m = argmax_m p(F | \theta_y^m)$ is the model that fits the data the best.

697 Quantitative results. We study occluder localization
698 quantitatively on the Occluded-Vehicle dataset using the
699 ground truth segmentation masks of the occluders and the
690 objects. Figure 4 shows the ROC curves of Compositional701 Nets (solid lines) and dictionary-based compositional mod-



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.

els (dashed lines) when using the occlusion score for classifying each pixel as occluder or object at all occlusion levels L1 - L3. 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. In particular, it shows 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 significantly better compared to the dictionarybased 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

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863



Figure 5: Visualization of occlusion localization results. Each result consists of three images: The input image, the occlusion scores of a dictionary-based compositional model [11] and the occlusion scores of 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.

localization experiments clearly show that Compositional-Nets 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

789

790

791

792

793

794

795

796

797

798

799

800

801

802

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 propose to replace the fully connected classification head of a DCNN backbone with a 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.

867

868

869

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963 964

965

966

967

968

969

970

971

- 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] 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 Con- ference on Computer Vision and Pattern Recognition*, pages
 2505–2512, 2014. 2
 - [3] 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
 - [4] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017. 1, 2, 7
 - [5] Alhussein Fawzi and Pascal Frossard. Measuring the effect of nuisance variables on classifiers. Technical report, 2016.
 2
 - [6] Sanja Fidler, Marko Boben, and Ales Leonardis. Learning a hierarchical compositional shape vocabulary for multiclass object representation. *arXiv preprint arXiv:1408.5516*, 2014. 2
 - [7] 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
 - [8] 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
 - [9] 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
- 903 [10] Adam Kortylewski. Model-based image analysis for foren 904 sic shoe print recognition. PhD thesis, University_of_Basel,
 905 2017. 1, 2, 3
- 906 [11] 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
- 910 [12] Adam Kortylewski and Thomas Vetter. Probabilistic compositional active basis models for robust pattern recognition. In
 912 British Machine Vision Conference, 2016. 2
- [13] 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

- [14] 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
- [15] 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
- [16] 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
- [17] 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
- [18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 1, 5
- [19] 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
- [20] 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
- [21] 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
- [22] 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
- [23] 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
- [24] 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
- [25] 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
- [26] 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 Vi*sion, pages 75–82. IEEE, 2014. 5

[27] Mingqing Xiao, Adam Kortylewski, Ruihai Wu, Siyuan

arXiv:1909.03879, 2019. 2, 6

Qiao, Wei Shen, and Alan Yuille. Tdapnet: Prototype

network with recurrent top-down attention for robust ob-

ject classification under partial occlusion. arXiv preprint

[28]	Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. <i>arXiv preprint arXiv:1905.04899</i> , 2019. 1, 2	
[29]	Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. In- terpretable convolutional neural networks. In <i>Proceedings</i> of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8827–8836, 2018. 2	
[30]	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 <i>Proceedings of the IEEE Conference on Computer Vision</i> <i>and Pattern Recognition</i> , pages 1372–1380, 2018. 1, 2	
[31]	Hongru Zhu, Peng Tang, Jeongho Park, Soojin Park, and Alan Yuille. Robustness of object recognition under extreme occlusion in humans and computational models. <i>CogSci</i> <i>Conference</i> , 2019. 1, 2	
[32]	Long Leo Zhu, Chenxi Lin, Haoda Huang, Yuanhao Chen, and Alan Yuille. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In <i>Computer vision–eccv 2008</i> , pages 759–773. Springer, 2008. 2	
		10