Compositional Generative Networks and Robustness to Perceptible Image Changes

Adam Kortylewski
Dept. Computer Science
Johns Hopkins University
Baltimore, USA
akortyl1@jhu.edu

Ju He
Dept. Computer Science
Johns Hopkins University
Baltimore, USA
jhe47@jhu.edu

Qing Liu
Dept. Computer Science
Johns Hopkins University
Baltimore, USA
qingliu@jhu.edu

Christian Cosgrove
Dept. Computer Science
Johns Hopkins University
Baltimore, USA
ccosgro2@jhu.edu

Chenglin Yang
Dept. Computer Science
Johns Hopkins University
Baltimore, USA
cyang76@jhu.edu

Alan L. Yuille
Dept. Computer Science
Johns Hopkins University
Baltimore, USA
ayuille1@jhu.edu

***Abstract*—Current Computer Vision algorithms for classifying objects, such as Deep Nets, lack robustness to image changes which, although perceptible, would not fool a human observer. We quantify this by showing how performances of Deep Nets degrades badly on images where the objects are partially occluded and degrades even worse on more challenging and adversarial situations where, for example, patches are introduced in the images to target the weak points of the algorithm. To address this problem we develop a novel architecture, called Compositional Generative Networks (Compositional Nets) which is innately robust to these types of image changes. This architecture replaces the fully connected classification head of the deep network by a generative compositional model which includes an outlier process. This enables it, for example, to localize occluders and subsequently focus on the non-occluded parts of the object. We conduct classification experiments in a variety of situations including artificially occluded images, real images of partially occluded objects from the MS-COCO dataset, and adversarial patch attacks on PASCAL3D+ and the German Traffic Sign Recognition Benchmark. Our results show that Compositional Nets are much more robust to occlusion and adversarial attacks, like patch attacks, compared to standard Deep Nets, even those which use data augmentation and adversarial training. Compositional Nets can also accurately localize these image changes, despite being trained only with class labels. We argue that testing vision algorithms in an adversarial manner which probes for the weakness of the algorithms, e.g., by patch attacks, is a more challenging way to evaluate them compared to standard methods, which simply test them on a random set of samples, and that Compositional Nets have the potential to overcome such challenges.**

***Index Terms*—Deep Networks, Compositional Networks, Robustness**

I. Introduction

Datasets have been pivotal for the advancement of the computer vision field over the last two decades. Early datasets such as Caltech-101 [12] defined standardized benchmarks that enabled the comparison of computer vision algorithms with standard metrics. These developments sparked the current paradigm of measuring research progress in computer vision in terms of performance improvements on well-known datasets for large-scale image classification [8], segmentation [10], [25], pose estimation [40], and part detection [4].

However, the focus on dataset performance encourages researchers to develop computer vision models that work well on a particular dataset, but do not transfer well to other datasets. We argue that this *lack of robustness* is caused by the paradigm of evaluating computer vision algorithms on balanced annotated datasets (BAD). This can be criticized as being problematic due to combinatorial complexity of visual scenes. Tougher performance measures are needed [45] that evaluate algorithms on data that differs in statistical properties from the training data [33] or by adversarial examiners [29]. These tougher performance tests can prune out algorithms (whose performance on BAD looks good) and encourage the community to develop algorithms that are reliable and which can lead to assured autonomy.

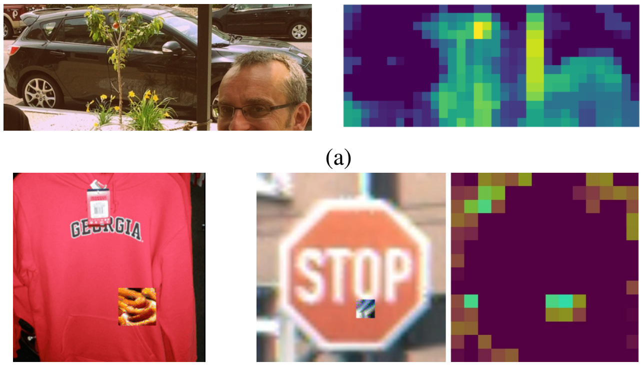
In this paper, we study a particularly important special case where objects are partially occluded, either randomly placed or selected by an adversarial mechanism (e.g., patch-based attacks [44]). The algorithms are trained on data without occlusion or patch attacks and hence the statistics of the test dataset differ from those in the training set. Occlusions are important because they happen frequently in real world conditions while patch attacks can be thought of as attacks where occluders are placed so as to confuse the algorithms. Our experiments show that the performance of deep networks degrades badly under these conditions, almost dropping to zero for black-box targeted patch attacks, while an alternative algorithms known as compositional networks [20] are much more robust (by an order of magnitude to patch attacks). We also show that standard deep network defenses [26], e.g., training using occluded or attacked images, only improves performance slightly.

Compositionality is a fundamental aspect of human cognition [2], [3], [14], [35] that is also reflected in the hierarchical compositional structure of the ventral stream in visual cortex [28], [34], [43]. A number of works in computer vision showed that compositional models can robustly classify partially occluded 2D patterns [15], [19], [37], [47]. Kortylewski et al. [20] proposed to *integrate* compositional models and DCNNs into a unified deep model with innate robustness to partial occlusion. In particular, they 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. This novel deep architecture is called Compositional Convolutional Neural Network (CompositionalNet). Figure 1 illustrates the robustness of CompositionalNets at classifying partially occluded objects and at defending very powerful patch-based attacks. In particular, it shows an image of a car that is occluded by other objects (Fig. 1a). Next to the image, 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. Moreover, Figure 1b shows a successful patch-based attack on a standard deep network [44]. Only a small modification of the image induces a misclassification of the t-shirt as pretzel with high confidence. In Figure 1c, we show a patch-based attack on a CompositionalNet [20]. The CompositionalNet can defend the attack, by localizing the adversarial patch and discarding it during classification.

Our experiments demonstrate that CompositionalNets outperform related approaches by a large margin at classifying partially occluded objects, even when they have not been exposed to occluded objects during training. They also show that CompositionalNets can defend against patch-based attacks with very large success compared to standard deep networks.

II. Related Work

Classification under partial occlusion. Recent work [22], [48] has shown that current deep architectures are significantly less robust to partial occlusion compared to Humans. Fawzi and Frossard [11] showed that DCNNs are vulnerable to partial occlusion simulated by masking small patches of the input image. Related works [9], [46], 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 IV show that such data augmentation approaches only have limited effects on the robustness of a DCNN to partial occlusion. Xiao et al. [41] 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. In contrast to deep learning approaches, generative compositional models [7], [13], [17], [24], [49] have been shown to be inherently robust to partial occlusion when augmented with a robust occlusion model [19]. Such models have been applied for detecting partially occluded object parts [37], [47] and for recognizing 2D patterns under partial occlusion [15], [23].



(a)

(b)

(c)

Fig. 1: Localizing occluder and patch-based attacks with CompNets. (a) CompositionalNets can localize the occluders (occlusion scores on the right) and subsequently focus on the non-occluded parts of the object to classify the image. (b) A patch-based attack induces a state-of-the-art model to misclassify an image of a t-shirt as pretzel with very high confidence. (c) Similarly to natural occluders, CompositionalNets can also localize and discard patch-based attacks.

Combining compositional models and DCNNs. Related works proposed to regularize the convolution kernels to be sparse [32], or to force feature activations to be disentangled for different objects [31]. 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. Kortylewski et al. [22] proposed to 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. This idea was subsequently extended [20], [21], [36] such that the compositional model was integrated into an end-to-end trainable deep network architecture, and was shown to have enhanced robustness to random natural occluders.

Adversarial patch attacks and defenses. Deep networks' fragility under occlusion is not limited to "natural" examples: their accuracy drops to near zero when presented with carefully crafted adversarial patches. Black-box adversarial patch attacks adapt the texture of an adversarial patch to achieve a high attack success rate with small area [6], [44], and refine the location search strategy using reinforcement learning rather than random search [44].

Many defenses against perturbation-based adversarial attacks have been proposed [16], [18], [26], [42]; however, defenses against patch-based attacks are less well studied. Two recent works have adapted adversarial training to the patch attack setting [5], [27]. In contrast, we are the first to show that black-box patch-based adversarial attacks can be defended against *without adversarial training* by *using an architecture that is innately robust to occlusion.*

III. Compositional Convolutional Neural Nets

We now present CompositionalNets [20], a fully generative compositional model that is integrated with DCNNs in an end-to-end image classification system.

A. 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 $f^l_p \in \mathbb{R}^D$ 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 define a differentiable generative compositional model of the feature activations $p(F|y)$ for an object class y. We model $p(F|y)$ 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.

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.

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 [22]:

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

with $\mathcal{V}{=}\{\nu^m \in \{0,1\}, \sum_m \nu^m{=}1\}$ and $\Theta_y = \{\theta^m_y, m = 1,\dots,M\}$. 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.

Occlusion modeling. Following the approach presented in [19], 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 $p(f_p|\mathcal{A}^m_{p,y},\Lambda)$ or an occluder model $p(f_p|\beta,\Lambda)$ is active:

$$p(F|\theta^m_y,\beta) {=} \prod_p p(f_p, z^m_p{=}0)^{1-z^m_p} p(f_p, z^m_p{=}1)^{z^m_p}, \tag{5}$$

$$p(f_p, z^m_p{=}1) = p(f_p|\beta,\Lambda)\ p(z^m_p{=}1), \tag{6}$$

$$p(f_p, z^m_p{=}0) = p(f_p|\mathcal{A}^m_{p,y},\Lambda)\ (1{-}p(z^m_p{=}1)). \tag{7}$$

The binary variables $\mathcal{Z}^m = \{z^m_p \in \{0,1\}|p \in \mathcal{P}\}$ indicate if the object is occluded at position p for mixture component m. The occlusion prior $p(z^m_p{=}1)$ is fixed a-priori. Related works [19], [22] use a single occluder model. We instead use a mixture of several occluder models that are learned in an unsupervised manner:

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

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

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. 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 the 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:

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

(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 $s^m_y = p(F|\theta^m_y,\beta) =$

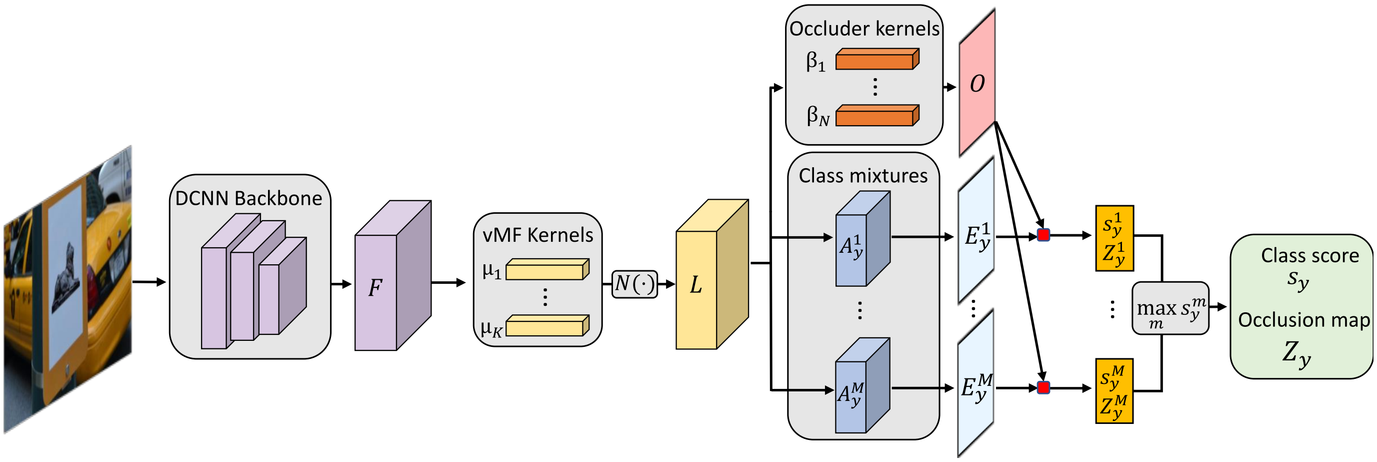


Fig. 2: Feed-forward inference with a CompositionalNet. A DCNN backbone is used to extract the feature map F, followed by a convolution with the vMF kernels $\{\mu_k\}$ and a non-linear vMF activation function $\mathcal{N}(\cdot)$. The resulting vMF likelihood L is used to compute the occlusion likelihood 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.

$\sum_p \max(E^m_{p,y}, O_p)$. 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 $\bar{m} = \text{argmax}_m\, s^m_y$.

Training of CompositionalNets with clustering and backpropagation. The compositional model is integrated 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 using a combination of clustering-based initialization and fine-tuning using end-to-end using backpropagation. 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:

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

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

$\mathcal{L}_{class}(y,y')$ 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. They are initialized by k-means clustering of the feature vectors f_p of the training images. Afterwards, they are fine-tuned by maximizing the vMF likelihood [39]:

$$\mathcal{L}_{vmf}(F,\Lambda) = C \sum_p \min_k \mu_k^T f_p, \quad (14)$$

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 $\mathcal{A}^m_y$ we need to maximize the model likelihood (Equation 4). The mixture coefficients are also initialized using clustering. In particular, we perform spectral clustering of the training images based on their vmf cluster center activations. For fine-tuning, 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:

$$\mathcal{L}_{mix}(F,\mathcal{A}_y) = -\sum_p (1-z^{\uparrow}_p) \log\Big[\sum_k \alpha^{m^{\uparrow}}_{p,k,y} p(f_p|\lambda_k)\Big] \quad (15)$$

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

IV. Experiments

Datasets. For evaluation we use the *Occluded-Vehicles* dataset as proposed in [38] and extended in [22]. The dataset consists of images and corresponding segmentations of vehicles from the PASCAL3D+ dataset [40] that were synthetically occluded with four different types of occluders: segmented *objects* as well as patches with *constant white color*, *random noise* and *textures*. The amount of partial occlusion of the object varies in four different levels: 0% (L0), 20-40% (L1), 40-60% (L2), 60-80% (L3).

We also test algorithms under realistic occlusion by introducing 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

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 [22]	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 [22]	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 [41]	**99.3**	98.4	98.9	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 I: 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 II: 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.

cropping out objects from the MS-COCO [25] 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.

A. Classification under Partial Occlusion

PASCAL3D+. In Table I 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 [22]. We also list the results of TDAPNet as reported in [41]. 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 CompositionalNet outperforms VGG-16, CoD and the combination of both significantly when occlusion exists. CompositionalNet also achieves better performance at level $L0$. Notice that CompNet-p5 outperforms CompNet-p4 in most situations.

MS-COCO. Table II 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 still outperform the other approaches significantly.

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 gets slightly improvement while TDPANet only improves at level $L0$. By contrast, CompositionalNets increases at all stages.

The third and fourth parts of Table II (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 [9] 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 OccludedVehicles dataset. All methods can benefit from these data augmentation strategies while VGG and TDAPNet still suffer from high occlusion situations.

B. Occlusion Localization

We test the ability of CompositionalNets and dictionary-based compositional models at occluder localization. We compute the occlusion score as the log-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 $m{=}\operatorname{argmax}_m p(F|\theta_y^m)$ is the model that fits the data the best. We study occluder localization quantitatively on the

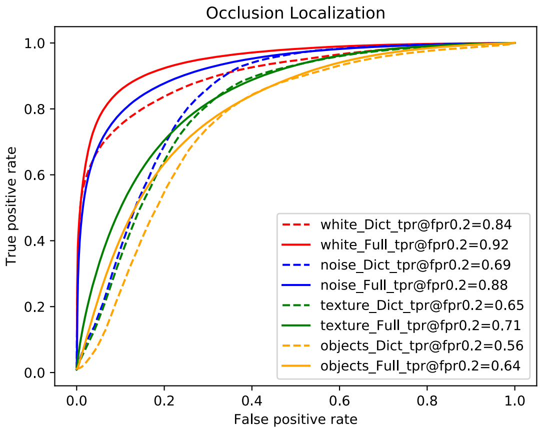


Fig. 3: ROC curves for occlusion localization with dictionary-based compositional models and the proposed Compositional-Nets averaged over all levels of partial occlusion (L1-L3).

Occluded-Vehicle dataset using the ground truth segmentation masks of the occluders and the objects. Figure 3 shows the ROC curves of CompositionalNets (solid lines) and dictionary-based compositional models (dashed lines) when using the occlusion score for classifying each pixel as occluder or object at all occlusion levels $L1 - L3$. Overall, we observe that **CompositionalNets can localize occluders accurately** for real as well as artificial occluders.

C. Defense against Patch-based Attacks

Baselines. We compare against two baselines: an standard CNN and a CNN that was trained on patch-based adversarial examples [27]. For both cases, we use a VGG16 [30] model pretrained on ImageNet [8] (the same as the backbone used for our CompNet), and we fine-tune on the training set. For patch-based adversarial training, we use the best-known state-of-the-art code provided with [27].

Attacks. We study black-box attacks because they are architecture-agnostic and more likely to arise in the real world [44]. In particular, we compare the robustness of these models on two state-of-the-art methods: Texture PatchAttack [44] and the patch attack version of Sparse-RS [6]. Both of these attack methods use patches whose locations and textures are optimized in a black-box fashion (where the objective is to fool the model with the smallest number of queries possible). We evaluate with two different attack methods because they use different methods for determining the patch locations and generating the patch textures.

CompNets are robust to patch attacks. Table III shows that CompNets are robust to patch attacks. When only one patch is used, CompNets are able to defend against more than 90% of TPA attacks and 80% of Sparse-RS attacks on PASCAL3D+. This shows that CompNets are the first architecture that is naturally robust to black-box patch attacks.

CompNets are more robust than adversarially trained architectures. Our results show that CompNets are significantly more robust than normal and adversarially trained CNNs on PASCAL3D+ (Table III). For example, CompNets are up to 4x more robust than a comparable adversarially trained CNN on PASCAL3D+. Remarkably, training the CompNet comes at negligible computational cost compared to adversarial training, and it has superior robustness. This result has never been shown before in prior work.

Attack success rates: PASCAL3D+					
		Acc.	TPA $(n = 1)$	TPA $(n = 4)$	Sparse-RS $(n = 1)$
untargeted	VGG16	**98.8**	91.6	95.4	99.6
	VGG16 + adv. train [27]	96.0	34.2	79.5	75.4
	CompNet	98.2	**7.8**	**24.9**	**18.0**
targeted	VGG16	**98.8**	52.6	88.0	84.7
	VGG16 + adv. train [27]	96.0	8.6	53.3	33.5
	CompNet	98.2	**2.4**	**8.2**	**5.8**

TABLE III: CompNets are significantly more robust than normal and adversarially trained CNNs under targeted and untargeted Texture Patch Attacks [44] and Sparse-RS attacks [6].

Untargeted TPA success rate: PASCAL3D+			
# patches:	$n = 1$	$n = 4$	$n = 8$
VGG16	91.6	95.4	94.1
VGG16 (+adv. train)	34.2	79.5	95.7
CompNet	**7.8**	**24.9**	**49.2**

TABLE IV: CompNets are more robust than adversarially trained CNNs, even with more patches.

CompNets are robust under harder attacks. We conduct an ablation study varying the number of patches (Table IV). We observe that increasing the number of patches leads to drops in robustness. However, CompNets are able to handle multiple patches more gracefully than the other models: going from one patch to four patches only results in a 17-point increase in attack success rate, whereas the adversarially trained model suffers a 45-point increase in attack success rate. Overall, CompNets show greater robustness across the board, even against harder attack configurations.

V. Conclusion

We question standard performance measures evaluating computer vision algorithms on balanced annotated datasets (BAD). Tougher performance tests are needed to prune out algorithms (whose performance on BAD look good) and will help develop algorithms that are reliable leading to assured autonomy. We studied this problem for the special case of occluders and patch-based attacks, showing that standard deep nets perform poorly in these scenarios, while deep networks with compositional representations are robust to patch attacks out of the box. Without expensive adversarial training, CompositionalNets can detect, locate, and ignore adversarial patches.

Acknowledgments. We acknowledge funding from the Swiss National Science Foundation with Grant P2BSP2.181713 and from the Institute for Assured Autonomy at JHU with Grant IAA 80052272.

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.

[2] Elie Bienenstock and Stuart Geman. Compositionality in neural systems. In *The handbook of brain theory and neural networks*, pages 223–226. 1998.

[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.

[4] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014.

[5] Ping-yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studor, and Tom Goldstein. Certified defenses for adversarial patches. *arXiv preprint arXiv:2003.06693*, 2020.

[6] Francesco Croce, Maksym Andriushchenko, Naman D Singh, Nicolas Flammarion, and Matthias Hein. Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks. *Eur. Conf. Comput. Vis.*, 2020.

[7] 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.

[8] 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.

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

[10] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

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

[12] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[13] Sanja Fidler, Marko Boben, and Ales Leonardis. Learning a hierarchical compositional shape vocabulary for multi-class object representation. *arXiv preprint arXiv:1408.5516*, 2014.

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

[15] 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.

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[17] 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.

[18] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

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

[20] Adam Kortylewski, Ju He, Qing Liu, and Alan L Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8940–8949, 2020.

[21] Adam Kortylewski, Qing Liu, Angtian Wang, Yihong Sun, and Alan Yuille. Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. *International Journal of Computer Vision*, 2020.

[22] 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.

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

[24] 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.

[25] 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.

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[27] Sukrut Rao, David Stutz, and Bernt Schiele. Adversarial training against location-optimized adversarial patches. *Eur. Conf. Comput. Vis.*, 2020.

[28] 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.

[29] Michelle Shu, Chenxi Liu, Weichao Qiu, and Alan Yuille. Identifying model weakness with adversarial examiner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11998–12006, 2020.

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

[31] 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.

[32] 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.

[33] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *arXiv preprint arXiv:2005.09241*, 2020.

[34] 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.

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

[36] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12645–12654, 2020.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 501–509, 2019.

[43] 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.

[44] Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhi Cao, and Alan Yuille. Patchattack: A black-box texture-based attack with reinforcement learning. *Eur. Conf. Comput. Vis.*, 2020.

[45] Alan L Yuille and Chenxi Liu. Deep nets: What have they ever done for vision? *International Journal of Computer Vision*, pages 1–22, 2020.

[46] 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.

[47] 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.

[48] 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.

[49] 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.