

# Localizing Occluders with Compositional Convolutional Networks

Adam Kortylewski   Qing Liu   Huiyu Wang   Zhishuai Zhang   Alan Yuille  
 Johns Hopkins University  
 Baltimore  
 akortyl1@jhu.edu

## Abstract

Compositional convolutional networks are generative models of neural network features, that achieve state of the art results when classifying partially occluded objects [2], even when they have not been exposed to occluded objects during training. While previous results showed the potential of CompositionalNets at localizing occluders, this remains to be confirmed quantitatively. In this work, we study the performance of CompositionalNets at localizing occluders in an image. We propose to extend the original model with a mixture of von-Mises-Fisher distributions. We show that this extension increases the model’s ability to localize occluders in an image while retaining an exceptional performance at classifying partially occluded objects.

## 1. Introduction

In natural images, objects are surrounded and partially occluded by other objects. Current deep models are significantly less robust to partial occlusion compared to Humans [5, 2]. Kortylewski et al. recently proposed the compositional convolutional network (CompositionalNet) [2], a generative model of neural feature activations that can classify partially occluded objects with exceptional performance, even when it has not been exposed to occluded objects during training.

While robustness to partial occlusion is a desirable property of vision systems, they should also be able to localize which parts of the object are occluded. The ability to localize occluders in an image is important because it improves explainability of the classification process and enables important future research for parsing scenes with mutually occluding objects.

Preliminary results in [2] suggest that CompositionalNets might be able to localize the occluder in the image, however, this property was not confirmed quantitatively. In this work, we study the ability of CompositionalNets at localizing occluders in an image. Our experiments show that CompositionalNets are not able to localize the occluder well

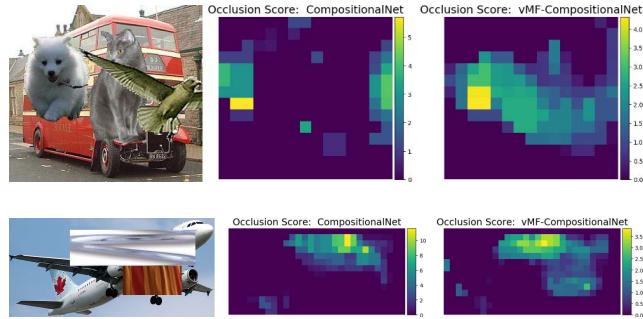


Figure 1: Occluder localization with CompositionalNets. Left: Input image. Middle & Right: Occlusion score based on feature activations in the pool4 layer of standard CompositionalNets [2] and our proposed extension respectively. We can clearly observe that our proposed model can localize occluders better across different objects and occluder types. Note that the CompositionalNets have not been exposed to partially occluded objects at training time.

(see experiments in Section 3), despite being able to classify partially occluded images robustly. We find that the reason for this is a simplified assumption about the distribution of neural feature activations in CompositionalNets (feature activations are binarized and their distribution were modeled using a Bernoulli distribution [2]). We propose to overcome this limitation by modeling the real-valued feature activations as a mixture of von Mises Fisher distributions. We demonstrate that this enhances the ability of CompositionalNets at localizing occluders significantly while retaining its exceptional performance at classifying partially occluded objects.

Our work demonstrates that neural networks are capable of localizing occluders in images when the architecture is compositional and the model is trained to be generative in terms of its low-level feature activations.

## 2. Compositional Convolutional Networks

We first review CompositionalNets as introduced in [2] (Section 2.1) and then propose an extension that enables

them to better localize occluders in images (Section 2.2).

## 2.1. CompositionalNets with Dictionary Encodings

A feature map  $F^l$  is the output of a layer  $l$  in a CNN. A feature vector  $f_p^l \in \mathbb{R}^C$  is the vector of features in  $F^l$  at position  $p$ , where  $p$  is defined on the 2D lattice of the feature map and  $C$  is the number of channels in the layer. Note that the spatial information from the image is preserved in the feature maps. We omit the subscript  $l$  in the remainder of this section because it is fixed in our model (e.g.  $l = 4$  for the layer  $conv_4$ ).

### A generative model of binary dictionary encodings.

The authors in [2] proposed to encode the feature maps  $F$  with a dictionary  $D = \{d_1, \dots, d_K\}$ . The dictionary is learned by clustering the vectors from the feature maps of all training images  $\{F^n | n = 1, \dots, N\}$ . The real valued feature vectors  $f_p$  are encoded with a sparse binary vector  $b_p$  by detecting the nearest neighbors of  $f_p$  in the learned part dictionary  $D$  using the cosine distance  $g(\cdot)$ . Hence, the element  $b_{p,k} = 1$  if  $g(f_p, d_k) > \delta$ . Intuitively,  $b_p$  encodes which parts of the dictionary  $D$  are detected at position  $p$  in the feature map  $F$ . A generative model of the binary activation matrix  $B$  is defined as Bernoulli distribution:

$$p(B|\mathcal{A}_y) = \prod_p p(b_p|\alpha_{p,y}) \quad (1)$$

$$= \prod_{p,k} \alpha_{p,k,y}^{b_{p,k}} (1 - \alpha_{p,k,y})^{1-b_{p,k}}. \quad (2)$$

Where  $\alpha_{p,k,y}$  is the probability that the part  $d_k$  is active at position  $p$  for the object class  $y$ , and thus  $b_{p,k} = 1$ . Note that parts are assumed to be independently distributed which makes the model in spirit similar to bag of words models. However, the important difference is that the spatial position of the part detections is preserved, hence capturing the spatial structure of the object.

**Mixture of compositional models.** The independence assumption between parts in Equation 1, the model assumes that the spatial distribution of parts in  $B$  is approximately the same. This assumption does not hold for 3D objects, because e.g. by changing the 3D pose of an object the relative spatial distribution of parts changes strongly (e.g. the location of the tires of a car in the image change between the side view and a frontal view). This problem can be resolved by using mixtures of compositional models:

$$p(B|\mathcal{A}_y, \mathcal{V}) = \prod_m p(B|\mathcal{A}_y^m)^{\nu_m}, \quad (3)$$

$$\sum_m \nu_m = 1, \nu_m \in \{0, 1\}. \quad (4)$$

The intuition is that each mixture component  $m$  will represent images of an object that have approximately the same spatial part distribution (i.e. similar viewpoint and 3D structure). The parameters of the individual mixtures  $\mathcal{A}_y^m$  as well

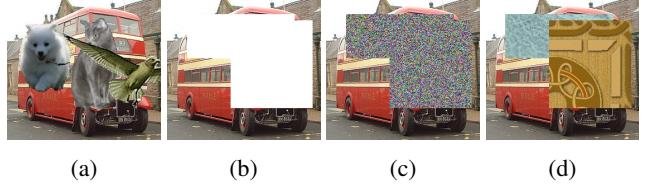


Figure 2: Visualization of synthetic partial occlusions with 40-60% occlusion. We simulate different types of occluder appearances: (a) natural objects, (b) white color, (c) random noise, (d) textures.

as the mixture assignment variables  $\mathcal{V}$  can be learned using maximum likelihood estimation while alternating between estimating  $\mathcal{A}_y^m$  and  $\mathcal{V}$ .

**Occlusion modeling.** Partial occlusion of an object will change the part activation patterns in  $B$  such that parts may be missing and other parts might be active at a previously unseen location. The intuition behind an occlusion model is that at each position  $p$  in the image either the object model  $\mathcal{A}_y$  or a background model  $\beta$  is active:

$$p(B|\Gamma) = \prod_p p(b_p|FG)^{z_p} p(b_p|BG)^{1-z_p}, \quad (5)$$

$$z_p \in \{0, 1\}, \quad (6)$$

$$p(b_p|FG) = p(b_p|\alpha_{p,y})p(z_p), \quad (7)$$

$$p(b_p|BG) = p(b_p|\beta)(1 - p(z_p)). \quad (8)$$

$\Gamma = \{\mathcal{A}_y; \beta; \mathcal{Z}\}$  The binary variable  $z_p$  indicates if the object is visible at position  $p$ . The occlusion prior  $p(z_p)$  can be learned or alternatively be set manually (see Section 3). The background model is defined as:  $p(b|\beta) = \prod_k \beta_k^{b_k} (1 - \beta_k)^{1-b_k}$ . Here the background model is assumed to be independent of the position in the image and thus has no spatial structure. The background model can be estimated by  $\beta = \frac{1}{J} \sum_{j=1}^J b_j$ , where  $J$  part detection vectors  $b_j$  are randomly sampled on a set of background images that do not contain one of the objects of interest. The maximum likelihood estimate of the occlusion variables  $z_p$  can be computed efficiently due to the independence assumption in the occlusion model (Equation 5).

## 2.2. Fully Generative CompositionalNets

The model as presented in the previous section can recognize partially occluded objects with high-performance [2], however, it is not able to discriminate well between the occluder and the object (see occluder localization experiments in Section 3). One reason is that instead of modeling the distribution of the real-valued features  $p(F|y)$ , the authors in [2] binarize the features with a heuristic threshold and just model the distribution of binary activations  $p(B|y)$  (Equation 1). However, the binarization induces a loss of information that is useful to discriminate between the object

Classification under Occlusion														
Occ. Area	0%	Level-1: 20-40%				Level-2: 40-60%				Level-3: 60-80%				Mean
Occ. Type	-	w	n	t	o	w	n	t	o	w	n	t	o	-
VGG	99.2	97.9	97.9	97.6	90.3	91.6	90.5	89.7	68.8	54.7	52.3	48.1	47.5	78.9
CompMixOcc-Dictionary	93.1	93.9	93.5	93.5	93.4	92.4	92.4	92.0	91.1	87.6	84.9	84.3	87.0	90.7
CompMixOcc-Full	92.4	92.2	92.1	92.2	92.7	90.9	90.2	89.4	90.4	85.8	84.7	79.0	87.8	89.2
CompositionalNet-Dictionary	98.8	97.8	97.5	97.2	95.9	94.7	94.5	94.0	91.2	87.4	85.4	83.8	85.4	92.6
CompositionalNet-Full	98.9	97.7	97.4	97.2	96.3	94.2	93.5	93.0	91.3	85.9	85.2	79.1	86.5	92.0
Human	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.3	98.3	98.3	98.3	99.5

Table 1: Object classification under occlusion. The proposed fully generative model performs on par with the dictionary-based model proposed in [2], while it performs much better at occlusion localization (see Figure 3).

and the occluder.

In this paper, we propose a generative model of the real valued features  $p(F|y)$ . In particular, we propose to replace the Bernoulli distribution over binary features (see Equation 1) with a mixture of von Mises Fisher (vMF) distributions:

$$p(F|\Theta_y) = \prod_p p(f_p|\mathcal{A}_{p,y}, \theta) \quad (9)$$

$$= \prod_p \sum_k \alpha_{p,k,y} p(f_p|S_k, \mu_k), \quad (10)$$

where  $\Theta_y = \{\mathcal{A}_{0,y}, \dots, \mathcal{A}_{\mathcal{P},y}, \theta\}$  are the model parameters at every position  $p \in \mathcal{P}$  on the lattice of the feature map  $F$ ,  $\mathcal{A}_{p,y} = \{\alpha_{p,0,y}, \dots, \alpha_{p,K,y} | \sum_{k=0}^K \alpha_{p,k,y} = 1\}$  are the mixture coefficients,  $K$  is the number of mixture components,  $\theta = \{\theta_k = \{S_k, \mu_k\} | k = 1, \dots, K\}$  are the parameters of the vMF mixture distributions:

$$p(f_p|S_k, \mu_k) = \frac{e^{S_k \mu_k^T f_p}}{Z(S_k)}, \|f_p\| = 1, \|\mu_k\| = 1, \quad (11)$$

and  $Z(S_k)$  is the normalization constant. Accordingly, we define the background model as  $p(f|\beta) = \sum_k \beta_k p(f|S_k, \mu_k)$ . The parameters of the vMF mixture model  $\Theta_y$  and the background model  $\beta$  can be learned with maximum likelihood estimation, as proposed in [2]. In the following experiments, we show that replacing the dictionary-based encoding of the features with the proposed fully generative model of the features (while keeping the remaining model the same) significantly enhances the ability of CompositionalNets to localize occluders in an image, while retaining an exceptional performance at classifying partially occluded objects.

### 3. Experiments

In this section, we compare our proposed fully generative CompositionalNet with the dictionary-based model as described in Section 2.1 at object classification and occluder localization. The experiments are performed on the OccludedVehicles dataset proposed in [3] and extended in [2]. The

dataset consists of vehicles from the PASCAL3D+ dataset [4] that were synthetically occluded by four different types of occluders: *objects* (Figure 2a) and patches with *constant white color* (Figure 2b), *random noise* (Figure 2c) and *textures* (Figure 2d). At training time all models are trained on non-occluded images, while at test time the models are exposed to images with different amount of partial occlusion. The CompositionalNets are trained from the feature activations of the pool4-layer of a VGG model that was pretrained on ImageNet [1]. The detailed training setup and parameter settings are chosen as specified in [2].

**Localizing occluders with CompositionalNets.** Figure 4 illustrates occlusion scores at different positions  $p$  of the corresponding feature map  $F$ . We compute the occlusion score as the log-likelihood ratio  $\log \frac{p(f_p|BG)}{p(f_p|FG)}$  of the background and foreground model. Note that we visualize only positive occlusion scores to highlight the localization of the occluder. We can observe that the fully generative model can localize occluders significantly better than in the dictionary-based model for all types of occluders. Figure 3 shows the ROC curves of both models when using the occlusion score for classifying each pixel as being occluder or not. The dictionary-based model (dotted lines) performs poorly for any type of occluder except the plain white ones. In contrast, our fully generative model significantly improves the quality of the occluder classification for all types of occluders. Note how textured occluders (objects and textured masks) are more difficult to localize.

In summary, our experimental results show that the fully generative CompositionalNet outperforms the dictionary-based model presented in [2] at occlusion localization significantly (Figure 3), while it performs on par at classifying partially occluded objects (Table 1).

### 4. Conclusion

In this work, we considered the problem of classifying partially occluded objects and localizing the occluders under the constraint that partially occluded objects are not part of the training data. We showed that dictionary-based Com-

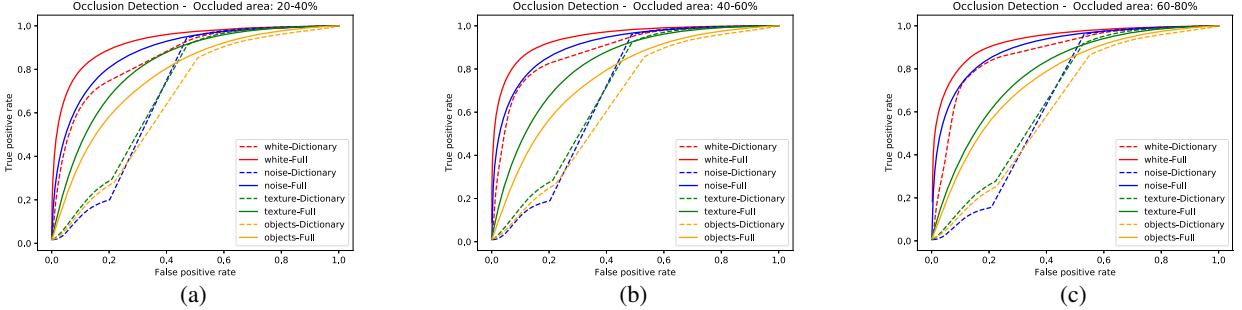


Figure 3: Occluder localization with dictionary-based CompositionalNets and our proposed fully generative CompositionalNet for different levels of partial occlusion: (a) 20-40%, (b) 40-60% and (c) 60-80% of the object is occluded. Our model significantly outperforms dictionary-based CompositionalNets.

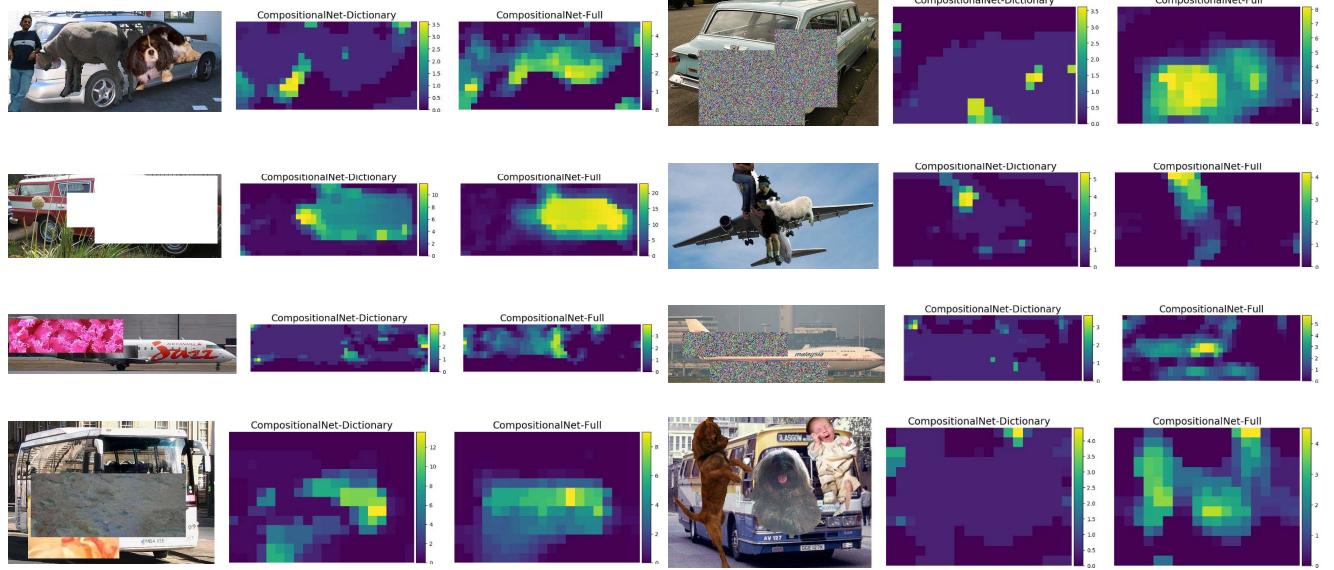


Figure 4: Visualization of occlusion localization results (not cherry picked). Each result consists of three images: The input image, the occlusion scores of a dictionary-based CompositionalNet and our proposed fully generative CompositionalNet. Note how our model can localize occluders with higher certainty across objects and occluder types.

positionalNets are not able to localize occluders well, although they can classify partially occluded objects well. We proposed an extension to CompositionalNets that enables them to accurately localize occluders in images while retaining an exceptional performance at classifying partially occluded objects. Our work shows that neural networks are capable of localizing occluders in images when their architecture is compositional and they are trained to be fully generative in terms of their high-level features. An accurate occluder localization enhances the interpretability of the network’s prediction and enables important future work on parsing scenes with mutually occluding objects.

## References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image

database. In *CVPR*, 2009. 3

[2] Adam Kortylewski, Qing Liu, Huiyu Wang, Zhishuai Zhang, and Alan Yuille. Compositional convolutional networks for robust object classification under occlusion. *arXiv preprint arXiv:1905.11826*, 2019. 1, 2, 3

[3] 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

[4] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 3

[5] 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