Attentional Masking for Pre-trained Deep Networks

Marcus Wallenberg and Per-Erik Forssén[1]

***Abstract*— The ability to direct visual attention is a fundamental skill for seeing robots. Attention comes in two flavours: the gaze direction (overt attention) and attention to a specific part of the current field of view (covert attention), of which the latter is the focus of the present study. Specifically, we study the effects of attentional masking within pre-trained deep neural networks for the purpose of handling ambiguous scenes containing multiple objects. We investigate several variants of attentional masking on partially pre-trained deep neural networks and evaluate the effects on classification performance and sensitivity to attention mask errors in multi-object scenes. We find that a combined scheme consisting of multi-level masking and blending provides the best trade-off between classification accuracy and insensitivity to masking errors. This proposed approach is denoted *multilayer continuous-valued convolutional feature masking* (MC-CFM). For reasonably accurate masks it can suppress the influence of distracting objects and reach comparable classification performance to unmasked recognition in cases without distractors.**

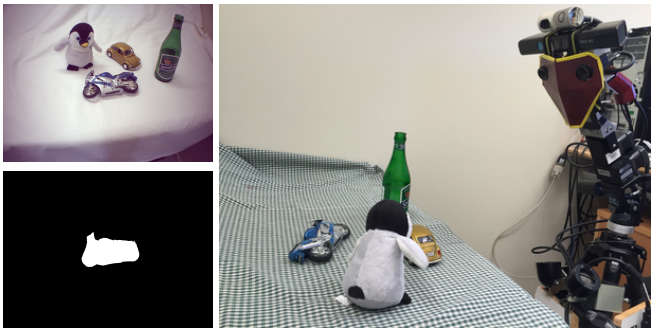

Fig. 1. Robot attention experiment setup. Even for simple scenes such as this one, a deep classification network (DCN) is practically useless when multiple objects are present. By integrating attentional masking to sequentially focus on regions of interest, the DCN can be used to interpret the scene. Left: left camera view (top) and attention mask from [9] (bottom). Right: overview of experiment setup, showing table with objects to recognize and the robot to the right.

I. INTRODUCTION

Deep neural networks, while computationally expensive to train, are often lightweight enough to be used on much less powerful platforms at runtime. Networks trained on huge image databases, such as ImageNet [1], are therefore a common tool for solving tasks other than those they were originally trained for.

One potential application task is object recognition on robot platforms, see figure 1. Here a pre-trained deep neural network that outputs class probabilities can be used for recognition, if combined with an attention system. Classification networks give ambiguous outputs when multiple known objects are present in a scene. Such an output is difficult to use, as it is also obtained e.g. for a single object from a category not present during training. An attention system can resolve some of these ambiguities by using a *saccade-and-fixate* strategy, as is done both by animals [2] and by many seeing robots [3] [4] [5].

We want to use pre-trained deep neural networks on robot vision platforms, and thus seek an answer to the question of how attention masking should be applied to a pre-trained deep network. Specifically, given an image and an object mask, we desire a classification result that accurately corresponds to the indicated object and that is robust to errors in the mask. To this end, we evaluate a number of methods proposed in the literature, and propose to combine their strengths in a method which we name *multilayer continuous-valued convolutional feature masking* (MC-CFM).

In order to perform controlled evaluation of attentional masking for sequential attention, we use pre-trained layers from the `VGG-F` network [6] (trained on ImageNet [1]). These are used to classify ambiguous scenes from the PASCAL-Context semantic segmentation dataset [7], as well as scenes from our robot platform, see figure 1. As a controlled proxy for an attention system, we use the salient object masks provided for a subset of these images in the PASCAL-S dataset [8]. This experiment allows us to evaluate controlled deviations from a best case scenario of perfect attention masks. For recognition on the robot platform we have instead integrated mask generation from *active segmentation* [9] into the pipeline. This method is selected due to its simplicity and because it is specifically designed for extracting a region around an attended fixation point.

II. RELATED WORK

1) Computational focus of attention: A classical way to direct focus of attention is to use a bounding-box or region-of-interest (ROI) and disregard the rest of the image. This approach has been tried on convolutional neural networks, e.g. in the R-CNN (Region-proposal Convolutional Neural Network) architecture [10]. In our experiments we test two variants of this approach, using either the mask contour or a corresponding bounding box. These are henceforth denoted *input masking* and *input bounding box masking*, respectively. Another recent approach to direct focus is *convolutional feature masking* (CFM) [11]. Here a low resolution spatial

This research is funded by the The Swedish Research Council through a grant for the project Learnable Camera Motion Models (2014-5928) and by Linköping University.

[1]Both authors are at the Department of Electrical Engineering, Linköping University, Sweden {`marcus.wallenberg`, `per-erik.forssen`}`@liu.se`

mask is applied to the finest grid in a *spatial pyramid pooling* (SPP) layer [12]. In this way, coarse-level structures bypass the masking, while only details specific to the masked object are left unsuppressed at fine scale. As it is described in [11], CFM is applied also during training. However, this would make it impossible to compare masking techniques on the same network. Therefore, we apply all masking techniques (including CFM) only at runtime.

2) Models of biological visual attention: Spatial visual attention in primates is believed to use spatial salience maps, that can be computed *bottom-up* from the feature layout of a scene [13]. Maps that appear to have such functions have been found in the lateral intraparietal area (LIP) [14] and in the superior colliculus (SC) [15]. In this paper we focus on spatial masking, and thus implicitly exclude top-down attention which also includes non-spatial components [16] [17].

Walther and Koch [18] point to studies of biological vision that indicate that attentional modulation occurs in areas V1, V4 as well as LGN. This motivates the proposed computational approach where attentional modulation is applied at all convolutional levels of a deep neural network.

III. PAPER OVERVIEW

This paper is organized as follows: In section IV we describe the different approaches to attentional modulation that we test. In section V we describe the experimental results. Methods are tested both on the PASCAL-S dataset and on sequences collected on our robot platform. The paper ends with a concluding discussion and outlook in section VI.

IV. METHODS AND EXPERIMENTAL SETUP

A. Mask generation

On a system level, attentional masking generates several different mask proposals and tests these in sequence. The quality of the individual mask proposals determines the benefits of applying modulation. In [18] a saliency map is adaptively thresholded to obtain attention masks. In CFM a method called *selective search* is used to generate mask proposals [19]. In [9], an *active segmentation* approach is proposed. Here mask proposals are generated with a fixation point as prior for the region location. This method is used in our robot validation experiments.

Initially, we wish to evaluate attentional masking methods without committing to a specific attention or mask generation method. Therefore we use the attention masks (manually annotated from gaze-tracking data) in the PASCAL-S dataset [8]. To simulate errors in these, we gradually expand and shrink them in order to obtain a continuous transition from a mask that is much too restrictive to one that is not at all restrictive enough. We posit that this is a typical case in foreground/background segmentation or detection, as most errors will affect the mask boundary.

The masks are calculated by computing a signed distance transform of the target object. We then normalise this such that the object boundary has a value of 1 and the innermost object pixel has a value of 0. This means that thresholds $0 \leq t_d < 1$ produce a mask that does not contain the entire object, a threshold $t_d = 1$ produces a mask that matches the object contour, and thresholds $t_d > 1$ produce masks that contain both object and background. Examples of the masks generated can be seen in figure 2.

Fig. 2. Mask distortion for robustness evaluation. Left: distance map of target object (the cat). Middle: $t_d = 0.75$ (white), $t_d = 1$ (light grey) and $t_d = 1.25$ (dark grey). Right: mask contours for the three thresholds. These cases represent masks that are too small, correct and too large, respectively.

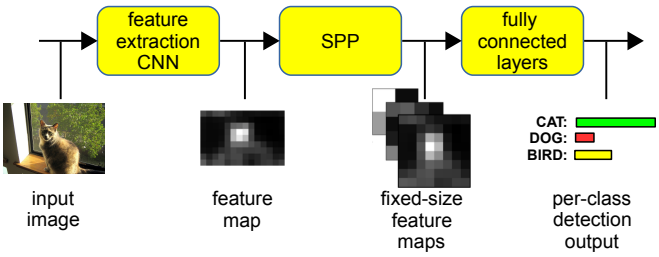

Fig. 3. Network structure used for masking evaluation. The feature extraction CNN consists of the first 13 layers (`conv1` to `pool5`) of the `VGG-F` network [6].

B. Network structure and mask propagation

The network we use to compare masking methods on is a slightly modified version of the `VGG-F` network [6], which is similar in structure and performance to the popular AlexNet [20]. The reason for this choice is that this network structure has become a de facto standard model for comparison. While its classification performance is no longer state-of-the-art, it is fast, of manageable size and complexity and freely available in many different formats. The only internal modification consists of the insertion of an SPP layer between the convolutional and fully-connected layers. This is done in order to use images of arbitrary size. We also change the number of output nodes to correspond to the number of classes in our experiments. The result is that we have 13 pre-trained feature extraction layers, and two trainable fully-connected layers. This network structure is illustrated in figure 3.

To trace the path of input pixels through the network, we perform forward propagation of the input mask. In each convolutional layer, the mask value of an input element (pixel) is added with equal weights to each output affected by that element. In each MAX-pooling layer, the mask value is only propagated if it belongs to the maximum (since otherwise it will not affect the result). Other layer types do not affect the spatial footprint of the mask. An example of the mask propagation is shown in figure 4.

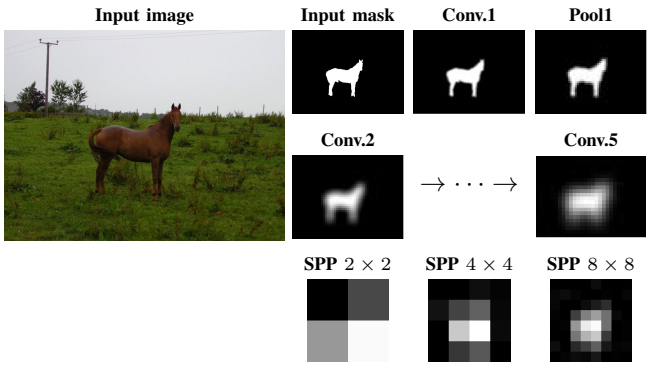

Fig. 4. Mask propagation for selected layers. Left: input image. Right: Masks propagated through the network. Shown here is the sum over all feature dimensions. The masks for layers 3 and 4 (normalisation and ReLU) are not shown, as they do not affect the shape of the mask. The 1×1 SPP layer mask sum is also omitted, since it is a single scalar without spatial structure.

C. Object class and data selection

In order to focus on the ambiguous multi-object cases, we select a subset of the most commonly co-occurring out of the 459 classes annotated in PASCAL-Context[1]. Specifically, we choose all classes that co-occur in at least ten images, resulting in a total of 11 classes appearing in 528 images with an average co-occurrence of 1.77 objects per image. The selected classes are *horse, cup, chair, dog, bird, tree, bottle, motorbike, bicycle, car* and *person*. Note that for co-occurrence of N objects, the best-case misclassification rate of an unmasked whole-image classification approach is $(N-1)/N$, since the system has no way of dealing with multiple objects. For training, we use all images in PASCAL-Context not present in PASCAL-S. For evaluation, we use all instances of the trained classes found in the 528 images from PASCAL-S (a total of 892 object instances), see figure 5 for examples.

D. Input masking

The simplest form of masking considered is to apply the attention mask directly onto the image data before computation of features. The masking is thus applied only once, before any feature extraction is performed. This approach, although simple, is likely to result in false contours that cause artifacts in feature map calculation, and can thus impede recognition[2]. We evaluate two versions of this, one using the object contour mask obtained at a particular threshold and one using an axis-aligned bounding box fitted to contain this contour.

E. Spatial pyramid pooling and convolutional feature masking

Spatial pyramid pooling (SPP) and Convolutional feature masking (CFM) are implemented as described in [12] and [11], respectively. In the SSP layer, the feature map produced by the convolutional layers of the feature extraction network is spatially MAX-pooled into a set of fixed-size grids. We use grid sizes of 1×1, 2×2, 4×4 and 8×8 spatial cells. In CFM, masking is applied at the finest level of the SPP layer. The masking is done such that if at least half of the pixels contributing to the feature vector at a location are from within the attention mask, this feature vector is retained, otherwise it is set to zero. In our implementation, this is done by thresholding the propagated mask at a value of 0.5, after which it is used as a mask on the convolutional feature map. The resulting fixed-size feature vectors are then concatenated before entering the fully-connected layers.

[1]We omit class 431, "unknown" since, although it commonly co-occurs with other classes, it does not represent an actual object type.

[2]A problem that is mentioned in [11] and that is further aggravated by the typically Gabor-like shape of first-layer filters.

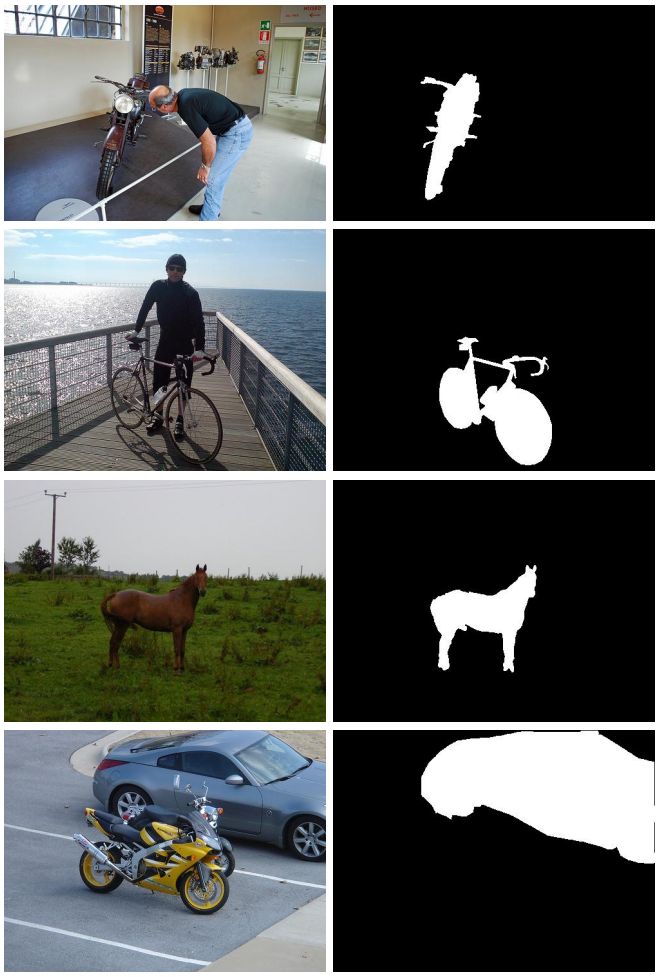

Fig. 5. Example of evaluation data. Left: images from PASCAL-VOC 2010. Right: corresponding masks from PASCAL-S. Each example has a co-occurrence of at least two of the used classes.

F. Walther and Koch saliency masking

Walther and Koch [18] propose an early-layer feature masking scheme where a saliency map is used to linearly blend a masked and an unmasked feature map. We apply this masking at the second layer, with a strength of 0.4 (see equation (1)), as suggested in their publication.

Method	RME (PASCAL-S)	RME (robot)
Input, contour	0.65	0.17
Input, bounding box	0.65	0.21
Walther & Koch [18]	0.90	0.58
CFM [11]	0.91	0.88
MC-CFM (ours)	**0.54**	**0.04**

TABLE I

RATIO OF MINIMAL ERRORS (RME) VS. UNMASKED BASELINE.

G. Multilayer continuous-valued convolutional feature masking

In order to combine the strengths of the masking techniques used by Walther and Koch [18] and Dai et al. [11], we propose an combined scheme in which masking is applied at all network levels. We also remove the binarisation used in CFM, and apply a continuous-valued mask proportional to the amount of foreground present at each feature map location. We denote this scheme, *multilayer continuous-valued convolutional feature masking* (MC-CFM). The purpose of this is to apply a suppression directly proportional to the amount of background present in the receptive field at all scales and locations, while avoiding artifacts such as false contours produced by masking of the input image. Also, as in [18], we use linear blending to control the masking strength. The masked output $R_M(x, y)$ is linearly blended with the unmasked output $R_0(x, y)$ to produce the final result

$$R(x, y) = \mu R_M(x, y) + (1 - \mu) R_0(x, y), \qquad (1)$$

where $\mu \in [0, 1]$ varies the amount of masking from none ($\mu = 0$) to complete ($\mu = 1$). We optimise blend weights for each layer on the top 1 % of training images containing the highest rate of class co-occurrence in PASCAL-Context. We do this by maximising the *classification margin*, which we define as the difference in activation strength between the correct output and the strongest competitor. During optimisation, we sample this at 9 equidistant mask thresholds t_d in the range $[0.0, 2.0]$. The resultant blend weights are shown in figure 6, bottom.

V. RESULTS

We have compared the masking methods in two different experiments: (1) recognition on a validation subset of PASCAL-S, with the ideal attention masks provided (as discussed in section IV-C), (2) attentional masking on the robot platform, using masks from the *active segmentation* algorithm [9]. Whereas the first experiment is a controlled evaluation of mask distortion when training and test situations are similar, the second experiment uses both *as-is* masks, and has a significant domain shift (training on real objects in PASCAL-S, but recognition on toys seen by our robot).

A. Results on PASCAL-S

The masking methods were evaluated at 141 equidistant thresholds, t_d, in the range $[0, 3.5]$. Note that this range is chosen to be larger than the tuning range ($[0.0, 2.0]$) to verify

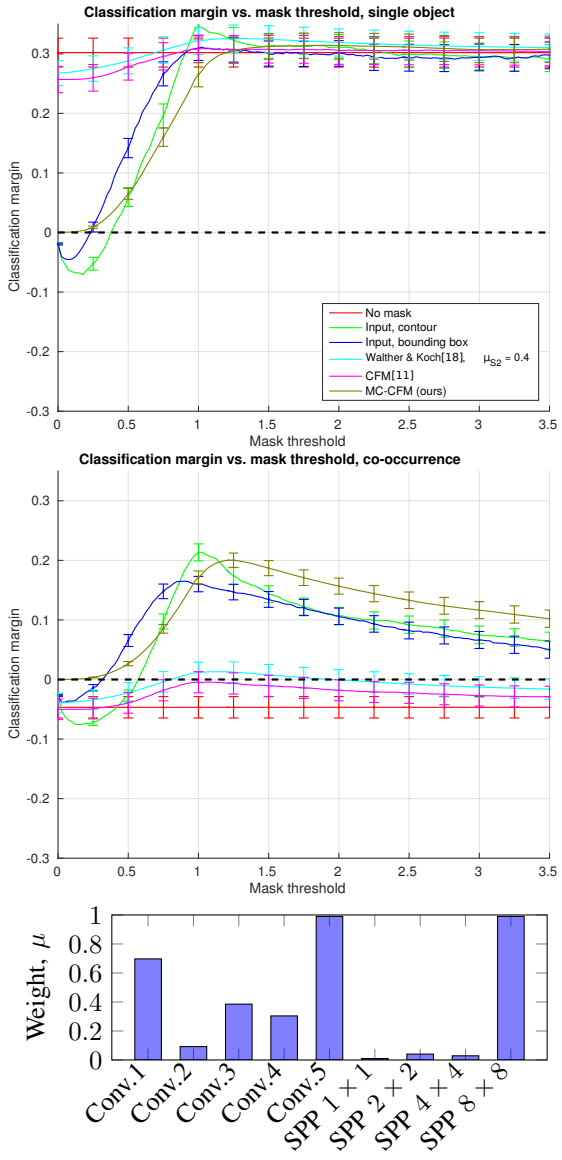

Fig. 6. Average classification margin for single object (top) and co-occurrence images (middle) on PASCAL-S. Error bars are shown for every 10th threshold value. See top plot for legend. Bottom: weights used in MC-CFM. Here, a value of 0 corresponds to no masking at all, while a value of 1 corresponds to complete masking.

that no artifacts occur outside the tuning range. The results of this evaluation are shown in figures 6 and 7. Table I shows the ratio of minimal errors (RME) compared to the unmasked baseline under co-occurence.

The mask weights used by MC-CFM are shown in figure 6. As can be seen, the weight distribution represents a combination of both early-layer masking and SPP masking, indicating that the best results in multi-object cases are obtained using a combination of both approaches. The total amount of masking obtained on the convolutional layers is higher than that found in [18]. The results also indicate that SPP masking is most useful at fine scale, as is suggested in [11].

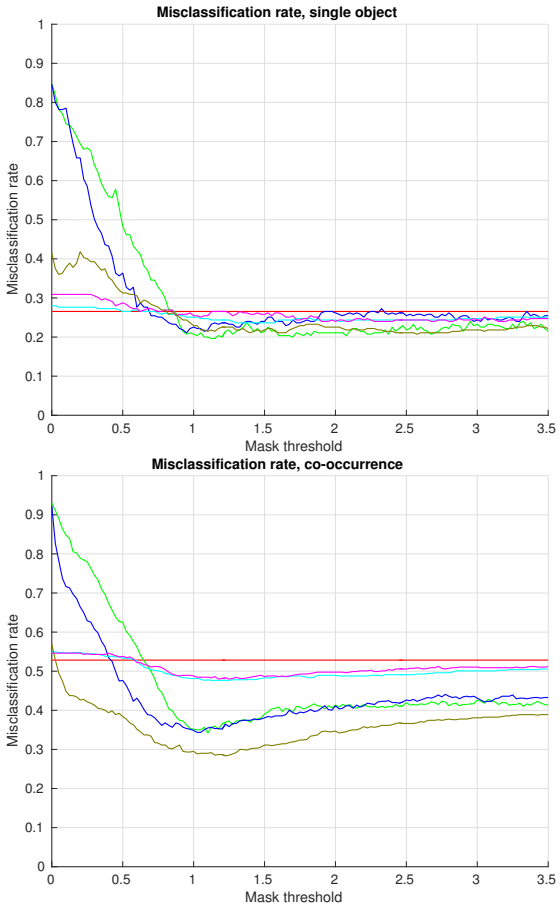

Fig. 7. Misclassification rates for single object (top) and co-occurrence images (bottom) on the PASCAL-S dataset. See figure 6, top for legend.

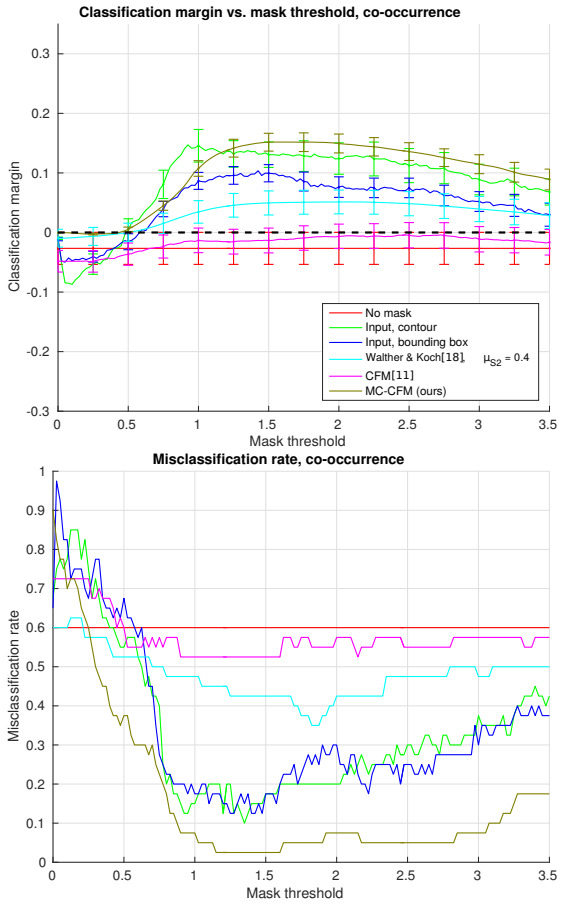

Fig. 8. Classification margin (top) and misclassification rate (bottom) for the robot dataset. Each evaluation image contains one target object and three distractors.

As can be seen in figure 7 (top), all methods are effective when no co-occurrence is present. Unless the mask is significantly smaller than the true object extent, all masking methods improve over the unmasked baseline. An interesting effect is that the inclusion of significant amounts of background does not seem to provide useful context information in these unambiguous cases.

The results when multiple objects are present are shown in figure 7 (bottom). Here MC-CFM surpasses all the other masking methods. It also retains some of the benefits of CFM for low thresholds, resulting in a smaller drop in performance for an underestimated mask than the input masking techniques[3]. When the mask includes large amounts of non-object areas, the effect of co-occurrence becomes evident. This results in gradually reduced classification performance.

B. Results on Robot Platform

For the second experiment we use images acquired with the platform shown in figure 9, and generate masks using *active segmentation* [9]. Fixation points are set manually on objects of interest in a reference image and then centred and matched between cameras using correlation-based visual

[3]The bounding box variant is inherently less sensitive to these cases, since it represents a consistent overestimation of object size.

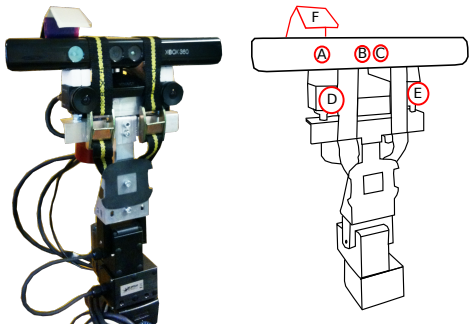

Fig. 9. Robot platform used to capture images in figure 10. The images were captured using wide-angle cameras (D) and (E). Pan-tilt stereo rig with Kinect. (A) - SLP projector, (B) - RGB camera, (C) - NIR camera, (D) - Right wide-angle camera, (E) - Left wide-angle camera, (F) - Manual SLP shutter (open).

servoing. We have collected five series of stereo fixations on four objects, resulting in a total of 40 images. Examples of images obtained are shown in figure 10. In order to isolate the effect of masking from other errors, we use a plain background to simplify segmentation and obtain the "ideal" mask approximately at a threshold of $t_d = 1$. Despite this, the masks from [9] contain significant errors; object parts are

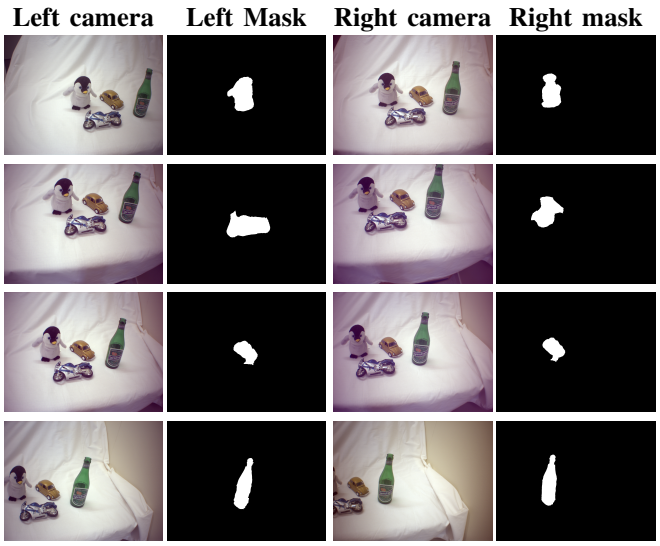

Fig. 10. Example images from the robot evaluation dataset. Images in columns 1 and 3 are left and right camera frames respectively, corresponding masks are shown in columns 2 and 4. Each example has a co-occurrence of four objects, where one is the target and the others are distractors.

occasionally missing, and background is sometimes included in the mask. This means that the robot experiment differs from the PASCAL-S experiment in at least two respects: larger difference between training set and test set objects, and less accurate attention masks.

The results of this evaluation are shown in figure 8, and mirror those obtained during co-occurrence on the PASCAL-S dataset. The classification margin is shown in the upper plot. Input masking and MC-CFM both have high values, with input masking often achieving a better classification margin. In terms of misclassification rate, see figure 8 (bottom), MC-CFM performs consistently better than the other methods. From this we can conclude that MC-CFM is better at classifying the images, but input masking is either more confident when correct or less overconfident when wrong. Table I shows the ratio of minimal errors, compared to the unmasked baseline under co-occurrence.

VI. CONCLUSIONS

We have implemented and compared mechanisms for incorporation of an attention mask into a deep neural network consisting of pre-trained feature extraction layers, with a task-adapted classifier on top. We show that an approach combining early-layer feature map masking and multi-scale spatial pyramid masking is better at disambiguating cases of object co-occurrence than image masking, early-layer masking or convolutional feature masking.

Our experiments are focussed on the attentional masking component. In a real application there are several other aspects that should be studied further to improve performance. The most pressing issue is to improve the mask generation. This includes investigating the use of multiple mask proposals for each gaze point, and a more sophisticated blending approach than the linear method used here. The incorporation of top-down attention components (e.g. [16] [17]) could also further improve performance, especially in cases where the initial segmentation is poor, or a specific object is sought.

The low impact of context on classification performance is also worth investigating further. Whether this is caused by the network topology, or that most network layers are pre-trained is currently an open question. Previous work on the PASCAL-S dataset has found modest improvements when context is utilized [7]. While context may not decrease the level of ambiguity in co-occurrence cases, it should (at least in typical scenes) provide a less noisy set of candidates to to begin with.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.
[2] M. Land and D. Nilsson, *Animal eyes*, ser. Oxford animal biology series. Oxford University Press, Incorporated, 2002.
[3] A. Borji and L. Itti, "Scene classification with a sparse set of salient regions," in *ICRA*, 2011.
[4] A. Y. Ng, G. Bradski, S. Gould, M. Messner, P. Baumstarck, S. Chung, J. Arfvidsson, A. Kaehler, and B. Sapp, "Peripheral-foveal vision for real-time object recognition and tracking in video," *IJCAI*, 2007.
[5] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, and D. G. Lowe, "Curious george: An attentive semantic robot," *Robotics and Autonomous Systems Journal*, 2008.
[6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional networks," in *BMVC*, 2014.
[7] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014.
[8] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *CVPR14*, 2014.
[9] A. Mishra, Y. Aloimonos, and C. L. Fah, "Active segmentation with fixation," in *ICCV*, 2009.
[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.
[11] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *CVPR*, 2015.
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *ECCV*, 2014.
[13] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE TPAMI*, 1998.
[14] M. E. Goldberg, J. W. Bisley, K. D. Powell, and J. Gottlieb, *Visual Perception - Fundamentals of Awareness. Chapter 10: Saccades, salience and attention: the role of the lateral intraparietal area in visual behavior*, ser. Progress in Brain Research. Elsevier B.V., 2006.
[15] R. J. Krauzlis, L. P. Lovejoy, and A. Zénon, "Superior colliculus and visual spatial attention," *Annual Reviews Neuroscience*, 2013.
[16] G. Humphreys, C. Romani, A. Olson, M. Riddoch, and J. Duncan, "Non-spatial extinction following lesions of the parietal lobe in humans." *Nature*, vol. 372, no. 6502, pp. 357–359, 1994.
[17] J. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial intelligence*, vol. 78, no. 1, pp. 507–545, 1995.
[18] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 2006, no. 19, pp. 1395–1407, 2006.
[19] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *IJCV*, 2013.
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.