

# An Empirical Evaluation of Current Convolutional Architectures’ Ability to Manage Nuisance Location and Scale Variability

Nikolaos Karianakis  
nikarianakis@ucla.edu

Jingming Dong  
dong@cs.ucla.edu

Stefano Soatto  
soatto@ucla.edu

UCLA Vision Lab, University of California, Los Angeles, CA 90095

## Abstract

We conduct an empirical study to test the ability of convolutional neural networks (CNNs) to reduce the effects of nuisance transformations of the input data, such as location, scale and aspect ratio. We isolate factors by adopting a common convolutional architecture either deployed globally on the image to compute class posterior distributions, or restricted locally to compute class conditional distributions given location, scale and aspect ratios of bounding boxes determined by proposal heuristics. In theory, averaging the latter should yield inferior performance compared to proper marginalization. Yet empirical evidence suggests the converse, leading us to conclude that – at the current level of complexity of convolutional architectures and scale of the data sets used to train them – CNNs are not very effective at marginalizing nuisance variability. We also quantify the effects of context on the overall classification task and its impact on the performance of CNNs, and propose improved sampling techniques for heuristic proposal schemes that improve end-to-end performance to state-of-the-art levels. We test our hypothesis on a classification task using the ImageNet Challenge benchmark and on a wide-baseline matching task using the Oxford and Fischer’s datasets.

## 1. Introduction

Convolutional neural networks (CNNs) are the de-facto paragon for detecting the presence of objects in a scene, as portrayed by an image. CNNs are described as being “approximately invariant” to nuisance transformations such as planar translation, both by virtue of their architecture (the same operation is repeated at every location akin to a “sliding window” and is followed by local pooling) and by virtue of their approximation properties that, given sufficient parameters and transformed training data, could in principle yield discriminants that are insensitive to nuisance transformations of the data represented in the training set. In ad-

dition to planar translation, an object detector must manage variability due to scaling (possibly anisotropic along the coordinate axes, yielding different aspect ratios) and (partial) occlusion. Some nuisances are elements of a transformation group, e.g., the (anisotropic) location-scale group for the case of position, scale and aspect ratio of the object’s support.<sup>1</sup> The fact that convolutional architectures appear effective in classifying images as containing a given object regardless of its position, scale, and aspect ratio [28, 40] suggests that the network can effectively manage such nuisance variability.

However, the quest for top performance in benchmark datasets has led researchers away from letting the CNN manage all nuisance variability. Instead, the image is first pre-processed to yield *proposals*, which are subsets of the image domain (bounding boxes) to be tested for the presence of a given class (Regions-with-CNN [19]). Proposal mechanisms aim to remove nuisance variability due to position, scale and aspect ratio, leaving a “Category CNN” to classify the resulting bounding box as one of a number of classes it is trained with. Put differently, rather than computing the *posterior* distribution<sup>2</sup> with nuisance transformations automatically marginalized, the CNN is used to compute the *conditional* distribution of classes given the data and a sample element that approximates the nuisance trans-

<sup>1</sup>The region of the image the objects projects onto, often approximated by a bounding box.

<sup>2</sup>One can think of the conditional distribution of a class  $c$  given an image  $x$ ,  $p(c|x)$ , as defined by a CNN, as the class posterior  $\int_G p(c|x, g) dP(g|x)$  marginalized with respect to the nuisance group  $G$ . If the nuisances are known, one can use the class-conditionals  $p(c|x, g_r)$  at each nuisance  $g_r \in G$  in order to approximate  $p(c|x)$  with a weighted average of conditionals, i.e.,  $p(c|x) \simeq \sum_r p(c|x, g_r) p(g_r|x)$ .

When a CNN is tested on a proposal  $r \subseteq x$  determined by a reference frame  $x_r$ , it computes  $p(c|x_{|r})$  ( $x$  restricted to  $r$ ), which is an approximation of  $p(c|x, g_r)$ . Then, explicit marginalization (assuming uniform weights) computes  $\frac{1}{|r|} \sum_r p(c|x_{|r})$  which is different from  $\frac{1}{|r|} \sum_r p(c|x, g_r)$  which in turn is different from  $\sum_r p(c|x, g_r) p(g_r|x)$ . This approach is therefore, on average, a lower bound on proper marginalization, and the fact that it would outperform the direct computation of  $p(c|x)$  is worth investigating empirically.

formation, represented by a bounding box. If the goal is the nuisance itself (object support, as in *detection* [10]) it can be found via maximum-likelihood (*max-out*) by selecting the bounding box that yields the highest probability of any class [19, 22]. If the goal is the class regardless of the transformation (as in *categorization* [10]), the nuisance can be approximately *marginalized out* by averaging the conditional distributions with respect to an estimation of the nuisance transformations<sup>2</sup>.

Now, if a CNN was an effective way of computing the marginals with respect to nuisance variability, there would be no benefit in conditioning and averaging with respect to (inferred) nuisance samples. This is a direct corollary of the Data Processing Inequality (DPI, Theorem 2.8.1 in [9]). Proposals are subsets of the whole image, so in theory less informative even after accounting for resolution/sampling artifacts (Fig. 1). *A fortiori*, performance should further decrease if the conditioning mechanism is not very representative of the nuisance distribution, as is the case for most proposal schemes that produce bounding boxes based on adaptively downsampling a coarse discretization of the location-scale group [24]. Class posteriors conditioned on such bounding boxes discard the image outside it, further limiting the ability of the network to leverage on side information, or “context”. Should the converse be true, *i.e.*, should averaging conditional distributions restricted to proposal regions outperform a CNN operating on the entire image, that would bring into question the ability of a CNN to marginalize nuisances such as translation and scaling or else go against the DPI. In this paper we test this hypothesis, aiming to answer to the question: *How effective are current CNNs to reduce the effects of nuisance transformations of the input data, such as location and scaling?*

To the best of our knowledge, this has never been done in the literature, despite the keen interest in understanding the properties of CNNs [20, 21, 34, 39, 43, 46, 47] following their empirical success. We are cognizant of the dangers of drawing sure conclusions from empirical evaluations, especially when they involve a myriad of parameters and exploit training sets that can exhibit biases. To this end, in Sect. 2 we describe a testing protocol that uses recognized existing modules, and keep all factors constant while testing each hypothesis.

## 1.1. Contributions

We first show that a baseline (AlexNet [28]) with single-model top-5 error of 19.96% on ImageNet 2014 Classification slightly *decreases* in performance (to 20.41%) when constrained to the ground-truth bounding boxes (Table 1). This may seem surprising at first, as it would appear to violate Theorem 2.6.5 of [9] (on average, conditioning on the true value of the nuisance transformation must reduce uncertainty in the classifier). However, note that the re-

striction to bounding boxes does not just condition on the location-scale group, but also on *visibility*, as the image outside the bounding box is ignored. Thus, *the slight decrease in performance measures the loss from discarding context by ignoring the image beyond the bounding box*. When we pad the true bounding boxes with a 10-pixel rim, we show that, conditioned on such “ground-truth-with-context” indeed does decrease the error as expected, to 17.65%. In Fig. 1 we show the classification performance as a function of the rim size all the way to the whole image for AlexNet and VGG16 [40]. A 25% rim yields the lowest top-5 errors on the ImageNet validation set for both models. This also indicates that the context effectively leveraged by current CNN architectures is limited to a relatively small neighborhood of the object of interest.

The second contribution concerns the *proper sampling* of the nuisance group. If we interpret the CNN restricted to a bounding box as a function that maps samples of the location-scale group to class-conditional distributions, where the proposal mechanism *down-samples* the group, then classical sampling theory [38] teaches that we should retain *not* the value of the function at the samples, but its *local average*, a process known as *anti-aliasing*. Also in Table 1, we show that simple uniform averaging of 4 and 8 samples of the isotropic *scale* group (leaving location and aspect ratio constant) reduces the error to 15.96% and 14.43% respectively. This is again unintuitive, as one expects that averaging conditional densities would produce less discriminative classifiers, but in line with recent developments concerning “domain-size pooling” [12].

To test the effect of such anti-aliasing on a CNN absent the knowledge of ground truth object location, we follow the methodology and evaluation protocol of [16] to develop a domain-size pooled CNN and test it in their benchmark classification of wide-baseline correspondence of regions selected by a generic low-level detector (MSER [32]). Our third contribution is to show that this procedure improves the baseline CNN by 5–15% mean AP on standard benchmark datasets (Table 3 and Fig. 5 in Sect. 2.2).

Our fourth contribution goes towards answering the question set forth in the preamble: We consider two popular baselines (AlexNet and VGG16) that perform at the state-of-the-art in the ImageNet Classification challenge and introduce novel sampling and pruning methods, as well as an adaptively weighted marginalization based on the inverse Rényi entropy. Now, if *averaging* the conditional class posteriors obtained with various sampling schemes should improve overall performance, that would imply that the *implicit* “marginalization” performed by the CNN is inferior to that obtained by sampling the group, and averaging the resulting class conditionals.<sup>2</sup> This is indeed our observation, *e.g.*, for VGG16, as we achieve an overall performance of 8.02%, compared to 13.24% when using the whole image

Method	AlexNet		VGG16	
Whole image	19.96		13.24	
Ground-Truth Bounding Box (GT)	20.41		12.44	
	Isotropically	Anisotropically	Isotropically	Anisotropically
GT padded with 10 px	17.66	17.65	10.91	10.30
Ave-GT, 4 domain sizes (padded with [0,30] px)	15.96	16.00	9.65	8.90
Ave-GT, 8 domain sizes (padded with [0,70] px)	14.43	14.22	8.66	7.84

Table 1. AlexNet’s and VGG16’s top-5 error on the ImageNet 2014 classification challenge when the ground-truth localization is provided, compared to applying the model on the entire image. We pad the ground truth with various rim sizes both isotropically and anisotropically. Then we show how averaging the class posteriors performs when applying the network on concentric domain sizes around the ground truth.

(Table 2). There are, however, caveats to this answer, which we discuss in Sect. 3.

Our fifth contribution is to actually provide a method that performs at the state of the art in the ImageNet Classification challenge when using a single model. In Table 2 we provide various results and time complexity. We achieve a top-5 classification error of 15.82% and 8.02% for AlexNet and VGG16, compared to 17.55% and 8.85% error when they are tested with 150 regularly sampled crops [40], which corresponds to 9.9% and 9.4% relative error reduction, respectively. Data augmentation techniques such as scale jittering and an ensemble of several models [23, 40, 42] could be deployed along with our method.

The source code implementing our method and the scripts necessary to reproduce the evaluation are available at [http://vision.ucla.edu/~nick/proj/cnn\\_nuisances/](http://vision.ucla.edu/~nick/proj/cnn_nuisances/).

## 1.2. Related work

The literature on CNNs and their role in Computer Vision is rapidly evolving. Attempts to understand the inner workings of CNNs are being conducted [6, 20, 21, 29, 34, 39, 43, 46, 47], along with theoretical analysis [2, 4, 8, 41] aimed at characterizing their representational properties. Such intense interest was sparked by the surprising performance of CNNs [6, 11, 19, 23, 28, 36, 37, 40, 42] in Computer Vision benchmarks [10, 15], where many couple a proposal scheme [1, 5, 7, 14, 24, 25, 27, 31, 35, 44, 48] with a CNN. As our work relates to a vast body of work, we refer the reader to references in the papers that describe the benchmarks we adopt, namely [6], [28] and [40].

Bilen et. al. [3] also explore the idea of introducing proposals in classification. However, their approach leverages on a significantly larger number of candidates and focuses on sophisticated classifiers and post-normalization of class posteriors. Our investigation targets selecting a very small subset of the most discriminative candidates among generic object proposals, while building on popular CNN models.

## 2. Experiments

### 2.1. Large-scale Image Classification

**What if we trivialize location and scaling?** First, we test the hypothesis that eliminating the nuisances of location and scaling by providing a bounding box for the object of interest will improve the classification accuracy. This is not a given, for restricting the network to operate on a bounding box prevents it from leveraging on context outside it. We use the AlexNet and VGG16 pretrained models, which are provided with the MatConvNet open source library [45], and test their top-1 and top-5 classification errors on the ImageNet 2014 classification challenge [10]. The validation set consists of 50,000 images, where at each of them one “salient” class is annotated a priori by a human. However, other ImageNet classes appear in many of the images, which can confound any classifier.

We test the classifier in various settings (Table 1); first, by feeding the entire image to it and letting the classifier manage the nuisances. Then we test the ground-truth annotated bounding box and concentric regions that include it. We try both isotropic and anisotropic expansion of the ground-truth region. We observe similar behavior, which is also consistent for both models.

Only for AlexNet at Table 1 using the object’s ground-truth support performs slightly worse than using the whole image. After we pad the object region with a 10-pixel rim, the top-5 classification error decreases fast. However, there is a trade-off between context and clutter. Providing too much context has diminishing returns. In Fig. 1 we show how the errors vary as a function of the rim size around the object of interest. Performance starts dropping down when we add more than 25% rim size. This padding gives 15.08% and 8.37% top-5 error for AlexNet and VGG16, as opposed to 19.96% and 13.24% respectively, when classifying the whole image.

To ensure that this improvement is not due to downsampling, we repeat the experiment with fixed resolution for the whole image and every subregion. We achieve this by shrinking each region with the same downsampling factor

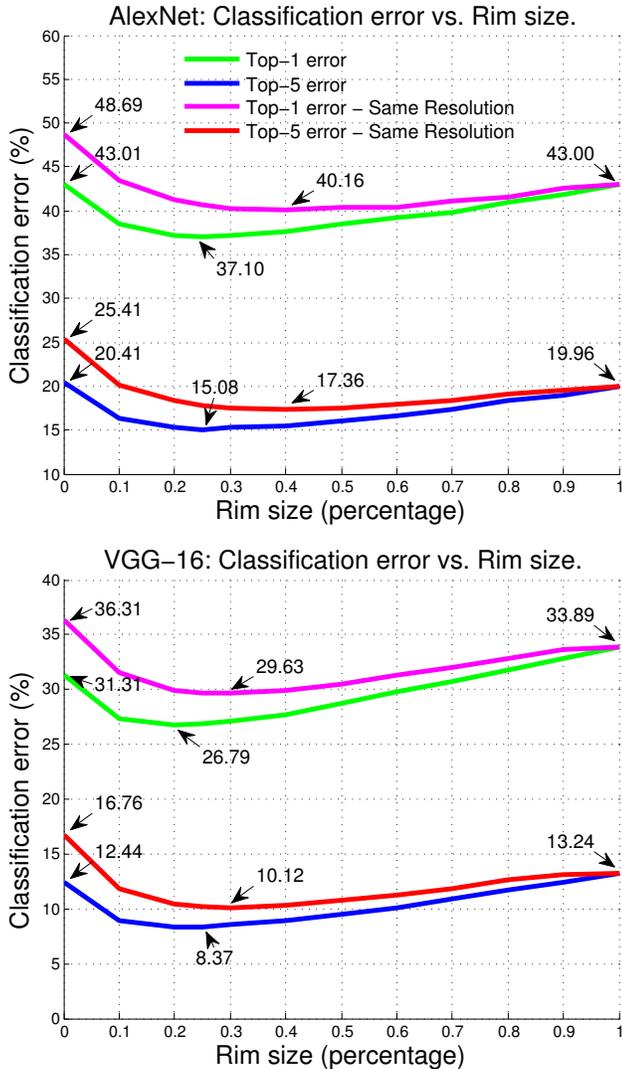


Figure 1. The top-1 and top-5 classification errors in ImageNet 2014 as a function of the rim size for AlexNet (above) and VGG16 (below) architecture. A 0 rim size corresponds to the ground-truth bounding box, while 1 refers to the whole image. A relatively small rim around the ground truth provides the best trade-off between informative context and clutter.

that we apply to the whole image to pass to the CNN. Finally we rescale the downsampled region to the CNN input. These results appear with the label “same resolution” in Fig. 1.

Finally, we apply domain size average pooling on the class posterior (*i.e.*, the network’s softmax output layer) with 4 and 8 domain sizes that are concentric with the ground truth. The added rim has the declared size either at both dimensions (for the anisotropic case) or only along the minimum dimension (for the isotropic case), and it is uniformly sampled in the range  $[0, 30]$  and  $[0, 70]$ , respectively.

The latter one further reduces the top-5 error to 14.22% for AlexNet, which is lower than any single domain size (*c.f.* Fig. 1). This suggests that explicitly marginalizing samples can be beneficial. Next we test whether the improvement stands when using object proposals.

**Introducing object proposals.** We deploy a proposal algorithm to generate “object” regions within the image. We use Edge Boxes [48], which provide a good trade-off between recall and speed [24].

First, we decide the number of proposals which will provide a satisfactory cover for the majority of objects present in the dataset. In a single image we search for the highest Intersection over Union (IoU) overlap between the ground-truth region and any proposed sample and in turn we evaluate the network’s performance on the most overlapping sample. We repeat this process for various number of proposals  $N$  in a small subset of validation set and finally choose  $N = 80$ , which provides a satisfactory trade-off between classification performance and computational cost.

Among the extracted proposals, we choose the most informative subset for our task, based on pruning criteria that we introduce later. Next we discuss what other samples we use, which are also drawn in Fig. 2.

**Domain-size pooling and regular crops.** We investigate the influence of domain-size pooling at test time both as stand-alone technique and as additional proposals for the final method which is described in Algorithm 1. We deploy domain-size aggregation of the network’s class posterior over  $D$  sizes that are uniformly sampled in the range  $[r, 1]$ , where 1 is the normalized size of the original image. After parameter search, we choose  $D = 5$  and  $r = 0.6$ . We use both the original and the horizontally flipped area, which gives 10 samples in total.

Finally, we use standard data augmentation techniques from the literature. As customary, the image is isotropically rescaled to a predefined size, and then a predetermined selection of crops is extracted [28, 40, 42].

**Pruning samples.** Continuing to sample patches within the image has diminishing return in terms of discriminability, while including more background patches with noisy class posterior distribution. We adopt an information-theoretic criterion to filter the samples that we use for the subsequent approximate marginalization.

For each proposal  $n \in N$  we evaluate the network and take the normalized softmax output  $v^n \in \mathbb{R}^{\mathcal{C}}$ , where  $v_i^n \in [0, 1], i = \{1, \dots, \mathcal{C}\}$  and  $\mathcal{C} = 1,000$  on ILSVRC classification. The output is a set of non-negative numbers which sum up to 1. We can interpret the vector  $v^n$  as a probability distribution on the discrete space of classes  $\{1, \dots, \mathcal{C}\}$  and compute the Rényi entropy as  $\mathbb{H}_\alpha(v^n) = \frac{1}{1-\alpha} \log(\sum_{i=1}^{\mathcal{C}} (v_i^n)^\alpha)$ .

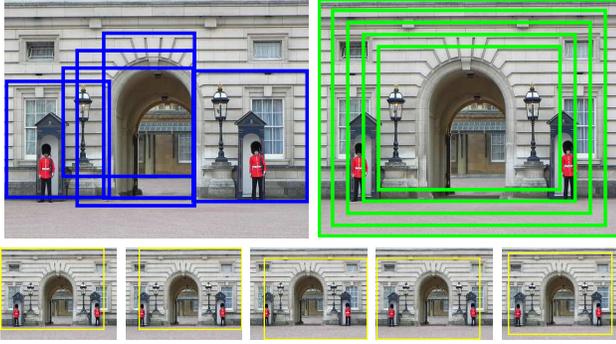


Figure 2. Visualizing different sampling strategies. Upper left: Object proposals. Generic proposals using Edge Boxes [48]. Upper right: Concentric domain sizes are centered at the center of the image. Below: Regular crops [28, 40, 42].

Our conjecture is that more discriminative class distributions tend to be more peaky with less ambiguity among the classes, and therefore lower entropy. In Fig. 3 we show how selecting a subset of image patches whose class posterior has lower entropy improves classification performance.

We extract  $N$  candidate object proposals<sup>3</sup> [48] and evaluate the network for both the original candidates and their horizontal flips. Then we keep a small subset  $E$ , whose posterior distribution has the lowest entropy. We use Rényi entropy with relatively small powers ( $\alpha = 0.35$ ), as we found that it encourages selecting regions with more than one highly-confident candidate object. While the parameter  $\alpha$  increases, the entropy is increasingly determined by the events of highest probability. Larger  $\alpha$  would be more effective for images with a single object, which is not the case in most images in ILSVRC.

Finally we introduce a weighted average of the selected posteriors as  $\sum_r p(c|x_{|r})p(x_{|r})$ , where  $x_{|r}$  is the support of sample  $r$  and  $p(x_{|r})$  is the weight of its posterior<sup>2</sup>. We try both uniform weights and weights proportional to the inverse entropy of the posterior  $p(c|x_{|r})$ . The latter is expected to perform better, as it naturally gives higher weight to the most discriminative samples.

**Comparisons.** To compare various sampling and inference strategies, we use the AlexNet and VGG16 models. All classification results in Table 2 refer to the validation set of the ILSVRC 2014 [10], except for the last row which demonstrates results on the test set. On the rows 2–5 we show the performance of popular multi-crop methods [28, 40, 42]. Then we compare them with strategies that involve concentric domain sizes (rows 6–8) and object proposals (rows 9–14).

<sup>3</sup>We introduce a prior encouraging the largest proposals among the ones that the standard setting in [48] would give. To this end, instead of directly extracting, for example,  $N = 80$  proposals, we generate 200 and keep the 80 largest ones (Algorithm 1).

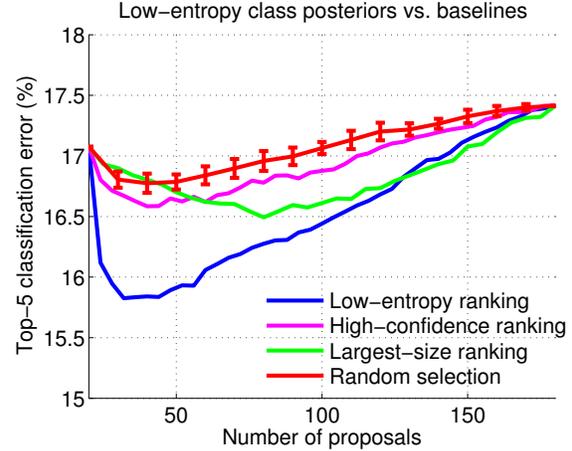


Figure 3. We show the top-5 error as a function of the number of proposals we average to produce the final posterior. Samples are generated with Algorithm 1 and classified with AlexNet. The blue curve corresponds to selecting samples with the lowest-entropy posteriors. We compare our method with simple strategies such as random selection, ranking by largest-size or highest confidence of proposals. The random sample selection was run 10 times and we visualize the estimated 99.7% confidence intervals as error-bars. Empirically, the discriminative power of the classifier increases when the samples are selected with the least entropy criterion.

Before extracting the crops and in order to preserve the aspect ratio of each single image, we rescale it so that its minimum dimension is 256. The proposals are extracted at the original image resolution and then they are rescaled anisotropically to fit the model’s receptive field. Additionally, some multi-crop algorithms resize the image in  $S$  different scales and then sample  $C$  patches of fixed size  $224 \times 224$  densely over the image. Szegedy et al. [42] use  $S = 4$  scales and  $C = 36$  crops per scale, which yields 144

**Algorithm 1** Regular & adaptive sampling in classification.

- *Object proposals.* We extract several object proposals from the image  $x$  (e.g., 200 Edge Boxes [48] and keep the  $N$  largest ones). Among them we choose  $E$  proposals whose class posterior has the lowest Rényi entropy with parameter  $\alpha$ . After hyper-parameter search, we choose  $N = 80$ ,  $E = 12$  and  $\alpha = 0.35$ .
- *$D$  concentric domain sizes* around the center of  $x$  (including their horizontal flip). We use 5 sizes that are uniformly extracted in the normalized range  $[0.6, 1]$ , where 1 corresponds to the whole image ( $D = 10$ ).
- *$C$  crops.* Regular crops; e.g.,  $C = 10$  or  $C = 50$  in 1 or 3 scales, as in [28, 40, 42].
- The class conditionals are approximated as  $\sum_r p(c|x_{|r})p(x_{|r})$ , where  $p(x_{|r})$  is either uniform or equals to the inverse entropy of the posterior  $p(c|x_{|r})$ .

Method			AlexNet			VGG16			#eval	#ave
# crops	# sizes	# proposals	top-1	top-5	t (s/im)	top-1	top-5	t (s/im)		
–	$D = 1$	–	43.00	19.96	0.01	33.89	13.24	0.06	1	1
$C = 10$	–	–	41.50	18.69	0.06	27.55	9.29	0.48	10	10
$C = 50$	–	–	41.01	18.05	0.66	27.44	9.12	1.34	50	50
$C = 10 \times 3$	–	–	40.58	17.97	0.16	27.23	8.88	1.26	30	30
$C = 50 \times 3$	–	–	<b>40.41</b>	<b>17.55</b>	<b>0.82</b>	<b>27.14</b>	<b>8.85</b>	<b>3.48</b>	150	150
–	$D = 10$	–	40.00	17.86	0.08	28.16	9.46	0.60	10	10
$C = 10$	$D = 10$	–	39.38	17.08	0.22	26.94	8.83	1.08	20	20
$C = 10 \times 3$	$D = 10$	–	39.36	17.07	0.46	26.76	8.68	1.88	40	40
–	–	$E = 40$	40.18	17.53	1.26	25.60	8.24	3.02	160	40
$C = 10$	–	$E = 20$	38.91	16.63		25.28	7.91		170	30
–	$D = 10$	$E = 12$	38.05	16.19	1.34	25.19	8.11	4.38	170	22
$C = 10$	$D = 10$	$E = 12$	37.69	15.83		25.11	8.01		180	32
$C = 10$	$D = 10$	$E = 12$ (fast)	37.71	15.88	0.94	25.12	8.08	3.70	180	32
$C = 10$	$D = 10$	$E = 12$ (W, fast)	<b>37.57</b>	<b>15.82</b>	<b>1.28</b>	<b>25.11</b>	<b>8.02</b>	<b>3.80</b>	180	32
$C = 10$	$D = 10$	$E = 12$ (test set)	37.417	16.018	–	25.117	7.909	–	180	32

Table 2. Top-1 and top-5 errors on the ImageNet 2014 classification challenge. The rows 2–5 include the common data augmentation strategies in the literature [28, 40, 42] (*i.e.*, regular sampling). The next three rows use concentric domain sizes that are uniformly sampled in the range [0.6, 1] with 1 being the normalized size of the original image (*c.f.* Fig. 2). Finally, in the last seven rows, we introduce adaptive sampling, which consists of a data-driven object proposal algorithm [48] and an entropy criterion to select the most discriminative samples on the fly based on the extracted class posterior distribution. The last row shows results on the test set. #eval stands for the number of samples that are evaluated for each method, while #ave is the number of samples that are eventually element-wise averaged to produce one single vector with class confidences. The previous top-reported with regular sampling and our results are shown in bold.

patches in all. Following the methodology from Simonyan et al. [40], it is comparable to deploy  $S = 3$  scales and extract  $C = 50$  crops per scale ( $5 \times 5$  regular grid with flips), for a total of 150 crops over 3 scales (row 5 in Table 2).

The results, presented in Table 2, indicate as expected that scale jittering at test time improves the classification performance for both 10-crop and 50-crop strategies. Additionally, the 50-crop strategy is better than the 10-crop strategy for both models. The results on row 5 in bold are the lowest errors that can be achieved with these specific single models<sup>4</sup> using only regular crops.

Then we present our methods and observe that using the AlexNet network with  $D = 10$  concentric domain sizes outperforms most multi-crop algorithms even if it only evaluates and averages 10 patches. Furthermore, combining it with 10 common crops achieves the best results for both networks, even without using 3-scale jittering. One interpretation for these improvements is that the concentric samples serve a natural prior for the majority of ILSVRC images,

<sup>4</sup>Specifically, we use the VGG16 model which is trained without scale jittering at training and appears on the first row of D area in Table 3 in [40]. Pre-trained models for both AlexNet and VGG16 are publicly available with the MatConvNet toolbox [45]. Simonyan et al. in their evaluation with 50 crops and 3 scales report 8.6% top-5 error on ImageNet 2014 validation. In contrast our implementation produces 8.85%, which can be attributed to using a different pre-trained model, as the initial weights are sampled from a zero-mean Gaussian distribution with standard deviation 0.01 and there might also be minor differences in the training process.

*i.e.*, the object of interest lies most probably at the center than at the image boundaries. This is a common assumption in the literature that also appears in large-scale video segmentation [26].

Following, we introduce the adaptive sampling mechanism with Algorithm 1 and reduce the top-5 error to 15.83% and 8.01% for AlexNet and VGG16 respectively. To set this in perspective, Krizhevsky et al. [28] report 16.4% top-5 error when they combine 5 models. We improve this performance with one single model. The relative improvement for the deployed instances of AlexNet and VGG16, compared to the data-augmentation methods used in [40, 42], is 9.9% and 9.4%, respectively. Row 14 shows results where the marginalization is weighted based on the entropy (notated as  $W$ ), while the methods in rows 9–13 use uniform weights (*c.f.* Algorithm 1). At the last row we show results from the ILSVRC test server for our top-performing method (row 13).

Regular and concentric crops assume that objects occupy most of the image or appear near the center. This is a known bias in the ImageNet dataset. To analyze the effect of adaptive sampling, we calculate the intersection over union error between the objects and the regular and concentric crops, and show in Fig. 4 the performance of various methods as a function of the IoU error. The improvement of using adaptive sampling (via proposals) over only regular and concen-

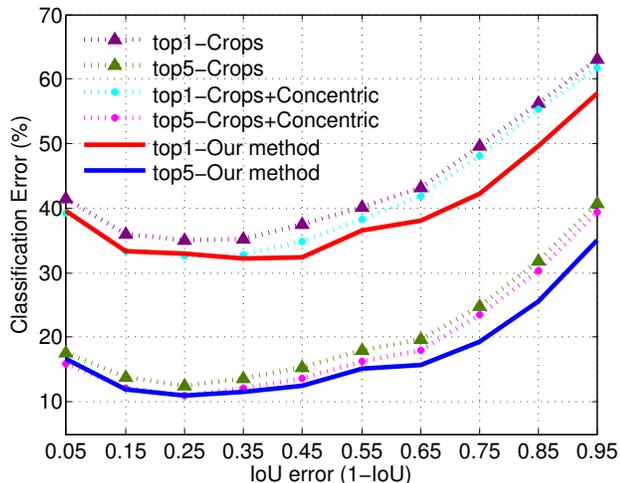


Figure 4. Classification error as a function of the IoU error between the objects and the regular and concentric crops.

tric crops is increased as IoU error grows, indicating that objects occupy less domain or are far away from the center.

**Time complexity.** In Table 2 we show the number of evaluated samples ( $\#eval$ ) and the subset that is actually averaged ( $\#ave$ ) to extract a single class posterior vector. The sequential time needed for each method is linear to the number of evaluated patches  $\#eval$ . We run the experiments with the MatConvNet library and parallelize the load for VGG16 so that the testing is done in batches of  $B = 20$  patches. We report the time profile<sup>5</sup> for each method in Table 2. A few entries cover two boxes, as their methods are evaluated together. Extracting the proposals is not a major bottleneck if using an efficient algorithm [24], such as Edge Boxes [48]. In rows 13–14 we report results of our faster version, where the Edge Boxes do not leverage edge sharpening and use one decision tree. Overall, compared to the 150-crop strategy, the object proposal scheme introduces marginal computational overhead.

## 2.2. Wide-Baseline Correspondence

We test the effect of domain-size pooling in correspondence tasks with a convolutional architecture, as done by [12] for SIFT [30], using the datasets and protocols of [16]. This is illustrated in Fig. 2 (upper right), but here the domain sizes are centered around the detector. We expect that such averaging will increase the discriminability of detected regions and in turn the matching ability, similar to the benefits that we see on the last rows of Table 1.

We use maximally-stable extremal regions (MSER) [32] to detect candidate regions, affine-normalize them, align them to the dominant orientation, and re-scale them for

<sup>5</sup>We use a machine equipped with a NVIDIA Tesla K80 GPU, 24 Intel Xeon E5 cores and 64G RAM memory.

head-to-head comparisons. For a detected scale  $\sigma$  at each MSER, the DSP-CNN samples  $D$  domain sizes within a neighborhood  $[\lambda_1\sigma, \lambda_2\sigma]$  around it, computes the CNN responses on these samples and averages the posteriors. The deployed deep network is the unsupervised convolutional network proposed by [16], which is trained with surrogate labels from an unlabeled dataset (see the methodology in [13]), with the objective of being invariant to several transformations that are commonly observed in images captured from different viewpoints. As opposed to network-classifiers, here the task is correspondence and the network is purely a region descriptor, whose last two layers (3 and 4) are the representations.

In Fig. 5 (left) we show the comparison between CNN and DSP-CNN on Oxford dataset [33]. CNN’s layer 4 is the representation for each MSER and DSP-CNN simply averages this layer’s responses for all  $D$  domain sizes. We use  $\lambda_1 = 0.7$ ,  $\lambda_2 = 1.5$  and  $D = 6$  sizes that are uniformly sampled in this neighborhood. There is a 15.1% improvement based on the matching mean average precision.

Fischer’s dataset [16] includes 400 pairs of images, some of them with more extreme transformations than those in the Oxford dataset. The types of transformations include zooming, blurring, lighting change, rotation, perspective and nonlinear transformations. In Fig. 5 (center) and Table 3 we show comparisons between CNN and DSP-CNN for layer-3 and layer-4 representations and demonstrate 7.7% and 5.0% relative improvement. We use  $\lambda_1 = 0.5$ ,  $\lambda_2 = 1.4$  and  $D = 10$  domain sizes. These parameters are selected with cross-validation. In Table 3 we show comparisons with baselines, such as using the raw data and DSP-SIFT [12]. After fine parameter search ( $\lambda_1 = 0.5$ ,  $\lambda_2 = 1.24$ ) and concatenating the layers 3 and 4, we achieve state of the art performance as shown in Fig. 5 (right), observing though the high dimensionality of this method compared to local descriptors.

Method	Dim	mAP
Raw patch	4,761	34.79
SIFT [30]	128	45.32
DSP-SIFT [12]	128	<b>53.72</b>
CNN-L3 [16]	9,216	48.99
CNN-L4 [16]	8,192	50.55
DSP-CNN-L3	9,216	52.76
DSP-CNN-L4	8,192	53.07
DSP-CNN-L3-L4	17,408	<b>53.74</b>
DSP-CNN-L3 (PCA128)	128	51.45
DSP-CNN-L4 (PCA128)	128	52.33
DSP-CNN-L34 (concat. PCA128)	256	<b>52.69</b>

Table 3. Matching mean average precision for different approaches on Fischer’s dataset [16].

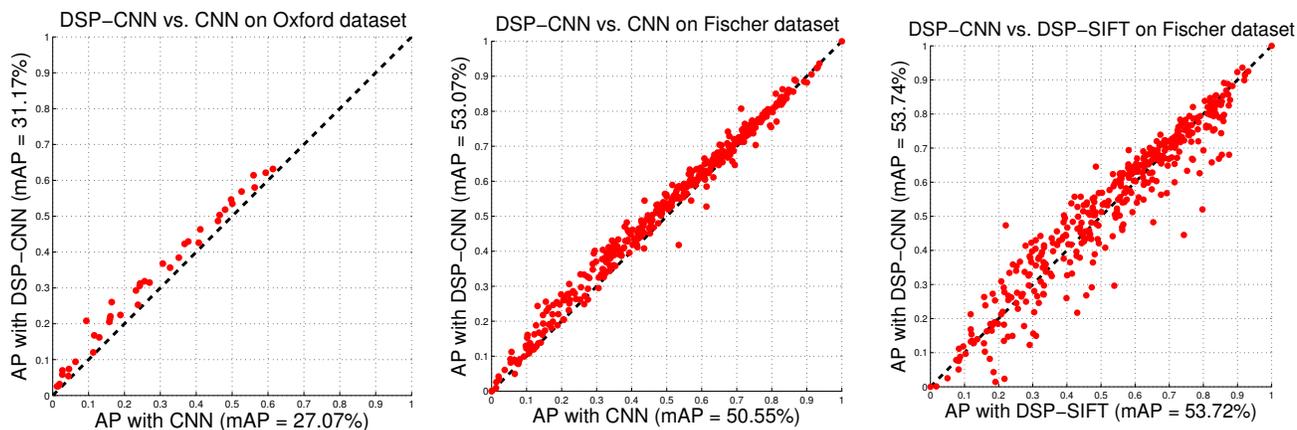


Figure 5. Head to head comparison between CNN and DSP-CNN on the Oxford [33] (left) and Fischer’s [16] (center) datasets. The layer-4 features of the unsupervised network from [16] are used as descriptors. The DSP-CNN outperforms its CNN counterpart in terms of matching mAP by 15.1% and 5.0%, respectively. Right: DSP-CNN performs comparably to the state-of-the-art DSP-SIFT descriptor [12].

Given the inherent high-dimensionality of CNN layers, we perform dimensionality reduction with principal component analysis to investigate how this affects the matching performance. In Table 3 we show the performance for compressed layer-3 and layer-4 representations with PCA to 128 dimensions and their concatenation. There is a modest performance loss, yet the compressed features outperform the single-scale features by a large margin.

### 3. Discussion

Our empirical analysis indicates that CNNs, that are designed to be invariant to nuisance variability due to small planar translations – by virtue of their convolutional architecture and local spatial pooling – and learned to manage global translation, distance (scale) and shape (aspect ratio) variability by means of large annotated datasets, in practice are less effective than a naive and in theory counterproductive practice of sampling and averaging the conditionals based on an ad-hoc choice of bounding boxes and their corresponding planar translation, scale and aspect ratio.

This has to be taken with the due caveats: First, we have shown the statement empirically for *few* choices of network architectures (AlexNet and VGG), trained on *particular* datasets that are unlikely to be representative of the complexity of visual scenes (although they may be representative of the same scenes as portrayed in the test set), and with a specific choice of *parameters* made by their respective authors, both for the classifier and for the evaluation protocol. To test the hypothesis in the fairest possible setting, we have kept all these choices constant while comparing a CNN trained, in theory, to “marginalize” the nuisances thus described, with the same applied to bounding boxes provided by a proposal mechanism. To address the

arbitrary choice of proposals, we have employed those used in the current state-of-the-art methods, but we have found the results representative of other choices of proposals.

In addition to answering the question posed in the introduction, along the way we have shown that by framing the marginalization of nuisance variables as the averaging of a *sub-sampling* of marginal distributions we can leverage of concepts from classical sampling theory to *anti-alias* the overall classifier, which leads to a performance improvement both in categorization, as measured in the ImageNet benchmark, and correspondence, as measured in the Oxford and Fischer’s matching benchmarks.

Of course, like any universal approximator, a CNN can in principle capture the geometry of the discriminant surface by “learning away” nuisance variability, given sufficient resources in terms of layers, number of filters, and number of training samples. So in the abstract sense a CNN *can* indeed marginalize out nuisance variability. The analysis conducted show that, at the level of complexity imposed by current architectures and training set, it does so less effectively than ad-hoc averaging of proposal distributions.

This leaves researchers the choice of investing more effort in the design of proposal mechanisms [18, 36], subtracting duties from the Category CNN downstream, or invest more effort in scaling up the size and efficiency of learning algorithms for general CNNs so as to render the need for a proposal scheme moot.

### Acknowledgments

This research is supported by ARO W911NF-15-1-0564/66731-CS, ONR N00014-13-1-034, and AFOSR FA9550-15-1-0229. We gratefully acknowledge NVIDIA Corporation for donating a K40 GPU that was used in support of some of the experiments.

## References

- [1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 3
- [2] F. Anselmi, L. Rosasco, and T. Poggio. On Invariance and Selectivity in Representation Learning. *arXiv preprint arXiv:1503.05938*, 2015. 3
- [3] H. Bilen, M. Pedersoli and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *British Machine Vision Conference*, 2014. 3
- [4] J. Bruna and S. Mallat. Invariant scattering convolution networks. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 3
- [5] J. Carreira and C. Sminchisescu. CPMC: Automatic object segmentation using constrained parametric min-cuts. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 3
- [6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014. 3
- [7] M. Cheng, Z. Zhang, W. Lin, and P. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 3
- [8] T. Cohen and M. Welling. Learning the irreducible representations of commutative lie groups. In *International Conference on Machine Learning*, 2014. 3
- [9] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012. 2
- [10] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F.-F. Li. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2, 3, 5
- [11] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3
- [12] J. Dong and S. Soatto. Domain-size pooling in local descriptors: DSP-SIFT. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2, 7, 8
- [13] A. Dosovitskiy, J. Springenberg, M. Riedmiller and T. Brox. Unsupervised feature learning by augmenting single images. In *Advances in Neural Information Processing Systems*, 2014. 7
- [14] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 3
- [15] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) challenge. In *International Journal of Computer Vision*, 2010. 3
- [16] P. Fischer, A. Dosovitskiy, and T. Brox. Descriptor matching with convolutional neural networks: a comparison to sift. *arXiv preprint arXiv:1405.5769*, 2014. 2, 7, 8
- [17] R. Gens, and P. Domingos. Deep Symmetry Networks. In *Advances in Neural Information Processing Systems*, 2014.
- [18] R. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, 2015. 8
- [19] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1, 2, 3
- [20] I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, and A. Y. Ng. Measuring invariances in deep networks. In *Advances in Neural Information Processing Systems*, 2009. 2, 3
- [21] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, 2015. 2, 3
- [22] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *IEEE European Conference on Computer Vision*, 2014. 2
- [23] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE International Conference on Computer Vision*, 2015. 3
- [24] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. What makes for effective detection proposals? In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 2, 3, 4, 7
- [25] A. Humayun, F. Li, and J. M. Rehg. RIGOR: Reusing inference in graph cuts for generating object regions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 3
- [26] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 6
- [27] P. Krähenbühl and V. Koltun. Geodesic object proposals. In *IEEE European Conference on Computer Vision*, 2014. 3
- [28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 1, 2, 3, 4, 5, 6
- [29] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *International Conference on Artificial Intelligence and Statistics*, 2015. 3
- [30] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, 2004. 7
- [31] S. Manen, M. Guillaumin, and L. V. Gool. Prime object proposals with randomized Prim’s algorithm. In *IEEE International Conference on Computer Vision*, 2013. 3
- [32] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. In *Image and Vision Computing*, 2004. 2, 7
- [33] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. In *International Journal of Computer Vision*, 2005. 7, 8
- [34] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecog-

- nizable images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2, 3
- [35] E. Rahtu, J. Kannala, and M. Blaschko. Learning a category independent object detection cascade. In *IEEE International Conference on Computer Vision*, 2011. 3
- [36] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 3, 8
- [37] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations*, 2014. 3
- [38] C. E. Shannon. A mathematical theory of communication. In *ACM SIGMOBILE Mobile Computing and Communications Review*, 2001. 2
- [39] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations*, 2014. 2, 3
- [40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1, 2, 3, 4, 5, 6
- [41] S. Soatto and A. Chiuso. Visual Representations: Defining properties and deep approximation. In *International Conference on Learning Representations*, 2016. 3
- [42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3, 4, 5, 6
- [43] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 2, 3
- [44] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013. 3
- [45] A. Vedaldi and K. Lenc. MatConvNet: Convolutional neural networks for MATLAB. In *ACM Conference on Multimedia Conference*, 2015. 3, 6
- [46] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, 2014. 2, 3
- [47] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *IEEE European Conference on Computer Vision*, 2014. 2, 3
- [48] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *IEEE European Conference on Computer Vision*, 2014. 3, 4, 5, 6, 7