Foveated BagNet: Inherent Interpretability Does Not Exclude Global Context

Holger Heidrich[1]
holger.heidrich@uni-tuebingen.de

Sarah Müller[2]
sar.mueller@uni-tuebingen.de

Andreas Schilling[1]*
andreas.schilling@uni-tuebingen.de

[1]Department of Computer Science, University of Tübingen, Tübingen, Germany
[2]Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany

Abstract

Deep learning in high-stakes settings requires models that are accurate and interpretable. BagNet addresses this by applying two restrictions to its architecture: (1) multiple output units perceive only a small spatial patch of the input image according to their position, and (2) the overall prediction is the sum of local predictors, preventing any nonlinear interactions between them. As a result, each predictor's impact is immediately readable from its activation, and its position directly points to the corresponding image patch. The model's inherent interpretability comes at the cost of drastically reduced complexity and an inability to capture large-scale image features, resulting in substantially lower accuracy. We propose *FovBagNet*, a foveated extension of BagNet inspired by the mammalian retina. Unlike BagNet, each local predictor depends on a stack of multiscale patches centered at its location which provides high-resolution detail at the center and progressively coarser context toward the periphery. From a scale space perspective, FovBagNet retains the inherent interpretability properties of BagNet, resulting in spatially accurate and faithful saliency maps, which we also demonstrate quantitatively. FovBagNet achieves a top-1 ImageNet accuracy of 0.74, closing most of the 11-point gap between BagNet-33 (0.65) and ResNet-50 (0.76). Additionally, we identify a conceptual limitation of the original logits-based BagNet saliency maps and empirically assess its impact.

1 Introduction

Over the last decade, deep learning methods, especially deep convolutional neural networks, have enabled machines to surpass humans in several image understanding tasks, including image classification, object detection, and image segmentation [27, 37, 41]. These models are nonlinear functions with millions of parameters, estimated from vast amounts of data. The emerging decision processes within such models are so complex and intertwined that they are practically treated as a black box, a concern that has prompted growing scrutiny from the research community [13, 28, 38]. The opacity of these models is not merely a theoretical concern: in high-stakes domains such as medical imaging and autonomous driving, deploying uninterpretable models can have serious consequences [38]. For such use cases, interpretability is a hard requirement.

*Currently visiting The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

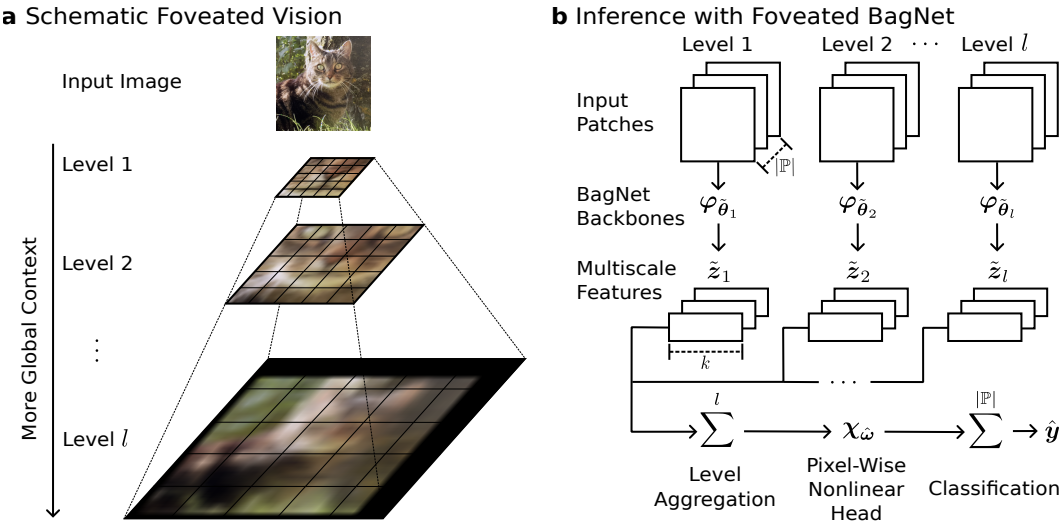

Figure 1: (a) Illustration of the foveated vision strategy, where a selected patch is represented across multiple levels with progressively larger global context while maintaining the same spatial resolution; the final level corresponds to a low-resolution (blurry) representation of the entire image. (b) Overview of the model architecture and inference pipeline, in which multiscale patches from all levels are jointly processed and aggregated to produce the final image classification output.

Methods exist to explain which part of the input image causes a model's decision by assigning relevance values to input pixels, providing so-called saliency or attribution maps. These fall into three groups: (1) post-hoc methods, which often rely on output gradients with respect to input pixels [33, 40, 43, 45, 52] or input perturbations [30, 36], (2) simplified explanations (e.g. Shapley values) of the underlying black box dynamics in the forward pass [9, 17], and (3) interpretable models by design through functional restrictions that allow only linear interactions between certain input components [2, 5, 7, 8].

A prominent example of the latter is BagNet [7], where pixel-wise class predictors depend only on small local patches (receptive fields) of the input image. The image-wide class predictor is simply the average across all patch-wise predictions, revealing how much each patch contributes to a given class prediction, making it inherently interpretable. However, the space of functions a BagNet can represent becomes drastically restricted. Consequently, a BagNet cannot capture complex image features beyond its receptive field and fails to incorporate global context into its local predictors. This limitation typically reflects in decreased image classification performance. In practice, when choosing the receptive field size, a tradeoff between the locality of the decision-explanation and the model performance has to be made: larger receptive fields allow the model to capture more complex features but also increase the amount of pixels that may have caused a certain local class prediction, resulting in less accurately localized saliency maps.

In this work, we relax BagNet's strict locality constraint to enable the incorporation of larger, even global, image features, while preserving a weaker concept of locality inspired by the biological principle of *foveated vision*. In mammalian vision, the density of photoreceptors (cones) is highest at the fovea and falls off steeply with eccentricity [10, 47], giving high-resolution processing at the fixation point and coarser, context-rich perception of the periphery. Analogously, rather than a single patch, each local predictor in our model depends on a stack of patches at different scales, all centered at the predictor's location. Therefore, each predictor covers increasingly large regions of the input image but are blurred and downsampled to the same resolution (Figure 1 a), so fine-grained features remain visible only in the innermost patch, while coarser features gradually become accessible in the outer patches. Just as the fovea captures sharp detail while peripheral vision supplies global structure, our architecture bridges local precision and global context in a principled, interpretable way—realized by processing each scale through a dedicated BagNet backbone (Figure 1 b). The resulting multiscale features are aggregated across scales into a single feature vector per location, then pooled across all image locations into the final class prediction, following the same aggregation scheme as BagNet.

Therefore, this work contributes *FovBagNet*, a foveated version of BagNet, which maintains the inherent interpretability of BagNet by keeping the strict dependence of latent units on local subsets of the input and by preserving the linear aggregation of these units into the overall prediction. At the same time, FovBagNet relaxes BagNet's locality constraint by generalizing it from image space to scale space, so that faithfulness can no longer be judged in image space alone, but has to be evaluated in scale space. In BagNet, zeroing a local neighborhood the size of its patch guarantees erasing that unit's activation; in FovBagNet, the same guarantee holds only once the corresponding pixels are zeroed at every level of the pyramid. In short, FovBagNet's locality guarantee is tied to the highest frequency that must be present for a feature to be detected, and adapts to the scale of larger objects, whereas BagNet's is fixed and simply fails once a feature exceeds its patch size.

Here, we show that *FovBagNet*, with its extended capabilities, (1) outperforms BagNet on the ImageNet classification task, achieving accuracies slightly below the performance of a ResNet-50, (2) retains a scale-space analogue of BagNet's inherent interpretability, and (3) provides spatially accurate saliency activations that are comparable to the ones generated by BagNet. In addition, we identify a conceptual limitation of the original BagNet logits-based saliency maps and evaluate its impact empirically.

2 Related Work

Post-Hoc Interpretability Post-hoc interpretability methods can be categorized by their technical approach into activation-based, gradient-based, and game-theoretic frameworks. Activation-based methods, such as CAM [52], replace global average pooling with a classification layer to provide spatial explanations. Grad-CAM [40] generalizes this approach to a broader set of architectures by weighting feature maps with gradient signals. However, both share structural limitations: their saliency maps remain coarse because their resolution equals the spatial dimensionality of the last convolutional layer's output; moreover, they become less specific as models with larger receptive fields are used.

In contrast, gradient-based attribution, such as Integrated Gradients (IG) [45], provides mathematically rigorous, pixel-level attribution by satisfying axioms such as sensitivity and completeness. IG computes attributions by integrating gradients along a straight path from a baseline input to the actual input, thereby avoiding the saturation artifacts common to simple gradient methods. While IG addresses the heuristic limitations of CAM, it is computationally intensive due to its requirement for path integration and often produces noisy visualizations in practice. Alternatively, SHAP [30] employs a game-theoretic framework to ensure "fair" attribution through the consistency axiom, offering a model-agnostic approach that unifies various prior methods, albeit at a high sampling cost. More recent work solves this by computing [9] or approximating [17] the Shapley values in the forward pass. Another drawback of Shapley values is that they simplify a nonlinear function to a single number per parameter, causing misleading interpretations for certain functions like XOR.

Despite their widespread utility, these post-hoc proxies remain fundamentally unreliable. Saliency methods are reported to fail basic sanity checks [1]. Since post-hoc methods approximate rather than expose the model's actual reasoning, their explanations may be misleading about the underlying computation. These limitations necessitate a shift toward inherently interpretable models [38].

Inherent Interpretability To overcome the unreliability of post-hoc proxies, recent research focuses on architectures that are interpretable by construction [32, 38, 44, 53]. These models can be broadly categorized into functionally constrained and part-based approaches.

Functionally constrained models ensure interpretability through mathematical restrictions on the hypothesis class. Generalized Additive Models (GAMs) achieve transparency by decomposing predictions into a sum of univariate shape functions, one per feature, making each feature's contribution directly readable from its learned curve [23]. Neural Additive Models (NAMs) [2] extend this classical framework by parameterizing each shape function with a dedicated neural subnetwork, recovering the flexibility needed to fit complex feature responses while retaining the additive structure's interpretability. B-cos networks [5] enforce weight-input alignment throughout the network, such that the learned transformations can be linearized from input to output. While this improves explanation stability over vanilla gradients, the resulting map remains a holistic, input-dependent summary, rather than a modular decomposition of which spatial regions contributed to the decision.

Part-based and localized architectures achieve interpretability by imposing spatial constraints on which information each part of the model can access. ProtoPNet [8], and extensions of it [50, 51], classify images by comparing latent feature patches against a learned dictionary of prototypical parts, providing an explicit "this looks like that" reasoning, but its explanations are sparse and exemplar-based rather than dense and spatially continuous. BagNet [7] achieves localized interpretability by restricting receptive fields to small image patches, aggregating patch-level logits into a final prediction. This produces dense, spatially aligned saliency maps by construction, without any post-hoc attribution step. Proto-BagNet [12] combines prototype learning with BagNets.

However, these architectural constraints introduce a persistent accuracy-interpretability trade-off: the structural limitations required for interpretability—whether additive decompositions, small receptive fields, weight-input alignment, or prototype matching—can substantially reduce classification performance relative to unconstrained black-box models. Our work directly targets this trade-off in the context of BagNet-style localized architectures.

Foveated Vision and Multiscale Foveated vision provides a reasonable inductive bias for computer vision: high-resolution processing at a fixation point paired with coarser, context-rich perception in the periphery. This asymmetry allows a visual system to be simultaneously detail-sensitive and globally aware without processing every region at full resolution.

Recent deep learning approaches to foveated processing explored non-grid graph-based convolutional architectures [26], foveated convolutional layer [46], and transformer-based models with biologically-inspired radial-polar pooling regions [25] or peripheral attention [31]. Beyond classification, foveated sampling has been shown to boost accuracy in multimodal tasks such as visual question answering and object detection under tight pixel budgets [21], and training on foveated images has been shown to improve robustness to adversarial attacks [42]. Additionally, there are works combining post-hoc interpretability with foveated vision [4, 33]. Multiscale architectures represent a closely related direction. Res2Net [19] introduces hierarchical residual-like connections within a single residual block to represent multiscale features at a granular level, enabling richer feature hierarchies without a significant parameter overhead. Saccader [15] builds a more complex sequential attention pipeline on top of BagNet, returning a sparse set of localized glimpses rather than a dense saliency map. While interpretability is a stated design goal of Saccader, the sparse glimpse representation lacks the dense, spatially aligned bag-of-features saliency that makes BagNet directly interpretable.

Crucially, none of these works yield the dense, spatially aligned multiscale saliency that this work focuses on. To the best of our knowledge, we propose the first model that combines inherent interpretable saliency maps with foveated vision.

3 Methods

Our method builds on the BagNet architecture [7], which itself is derived from a ResNet-50 architecture [24], where nearly all 3×3 convolution layers are replaced by 1×1 convolution layers. The number of the remaining 3×3 convolutions determines the BagNet's receptive field size $r = q$. The three variants of BagNet–q with $q \in \{9, 17, 33\}$ are commonly known. However, it is possible to create BagNets with various different depths and receptive field sizes. To formally introduce our method, we define the forward pass of the *dilated* BagNet backbone as

$$\boldsymbol{\varphi}_{\boldsymbol{\theta}}^{(q,s,d)} \colon \mathbb{R}^{3\times h_{\text{in}}\times w_{\text{in}}} \to \mathbb{R}^{k\times h_{\text{out}}\times w_{\text{out}}}, \tag{1}$$

$$\boldsymbol{x} \mapsto \boldsymbol{z}, \tag{2}$$

with learnable model parameters $\boldsymbol{\theta}$ that map an RGB input image $\boldsymbol{x}$ of height h_{in} and width w_{in} to a latent representation $\boldsymbol{z}$, containing a k dimensional feature vector per output pixel. For the sake of completeness, we include the global stride s which affects the output height $h_{\text{out}} = \lfloor h_{\text{in}}/s \rfloor$ and width $w_{\text{out}} = \lfloor w_{\text{in}}/s \rfloor$ and is typically set to $s = 8$ during training and $s = 1$ for evaluation. We also extend the model by a *dilation factor* d that we simply pass to all its internal convolution layers. Setting $d = 1$ results in the standard BagNet backbone $\boldsymbol{\varphi}_{\boldsymbol{\theta}}^{(q,s,1)}$. Note that the dilation parameter influences the resulting receptive field size according to $r = d(q-1)+1$. In the BagNet approach, a linear head

$$\boldsymbol{\psi}_{\boldsymbol{\omega}} \colon \mathbb{R}^{k\times h_{\text{out}}\times w_{\text{out}}} \to \mathbb{R}^{c\times h_{\text{out}}\times w_{\text{out}}}, \tag{3}$$

$$\boldsymbol{z} \mapsto \boldsymbol{S}, \tag{4}$$

transforms the features $\boldsymbol{z}$ to $\boldsymbol{S} = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_c)$ containing one saliency map $\boldsymbol{S}_j$ for each of the c occurring object classes (according to the weights $\boldsymbol{\omega}$). $\boldsymbol{S}$ is then spatially averaged to obtain the image-wide class logits $\overline{\boldsymbol{S}} = \left(\overline{\boldsymbol{S}}_1, \ldots, \overline{\boldsymbol{S}}_c\right) \in \mathbb{R}^c$ and the corresponding probabilities $\hat{\boldsymbol{y}} = \text{softmax}(\overline{\boldsymbol{S}}) \in \mathbb{R}^c$ subsequently.

Next, we define the discrete scale-space representation $\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_l) \in \mathbb{R}^{l \times 3 \times h_{\text{in}} \times w_{\text{in}}}$ of the input image $\boldsymbol{x}$ recursively as an l-level Gauss-stack with $\tilde{\boldsymbol{x}}_1 = \boldsymbol{x}$ and $\tilde{\boldsymbol{x}}_{i+1} = \mathcal{G}_\sigma^{2^{i-1}}(\tilde{\boldsymbol{x}}_i)$ for $i \in \{1, \ldots, l-1\}$ where $\mathcal{G}_\sigma^d$ denotes the d-dilated convolution with a Gaussian kernel of scale σ (Figure 1 a). Note that we apply reflection padding to keep the blurred images' resolutions equal for all levels. Finally, we generate pixel-wise aligned multiscale image features $\tilde{\boldsymbol{z}}$ from one dilated BagNet backbone per scale level independently (Figure 1 b):

$$\boldsymbol{\rho}_{\hat{\boldsymbol{\theta}}}^{(q,s)} \colon \mathbb{R}^{l \times 3 \times h_{\text{in}} \times w_{\text{in}}} \to \mathbb{R}^{l \times k \times h_{\text{out}} \times w_{\text{out}}}, \tag{5}$$

$$\tilde{\boldsymbol{x}} \mapsto \tilde{\boldsymbol{z}} = \left(\boldsymbol{\varphi}_{\hat{\boldsymbol{\theta}}_i}^{\left(q,s,2^{i-1}\right)}(\tilde{\boldsymbol{x}}_i) \right)_{i=1}^{l}. \tag{6}$$

The first scale level is designed to match BagNet's receptive field size: the receptive field size $r_i = 2^{i-1}(q-1)+1$ of $\tilde{\boldsymbol{z}}_i$ (asymptotically) doubles at each level, so that $\tilde{\boldsymbol{z}}_1$ perceives a $q \times q$ neighborhood in the original image with full detail, and the model behaves identically to BagNet at this stage. By incorporating further levels $i = 2, \ldots, l$, the model accesses increasingly larger (but coarser) patches, with the final level potentially depending on a heavily blurred version of the entire image. This is achieved by applying Gaussian blurring at each level, progressively removing high-frequency detail—equivalent to a low-pass filter whose cutoff decreases with scale—so that higher levels retain only coarse, low-frequency features. This introduces a clear trade-off: lower levels retain fine-grained, high-frequency information over small spatial extents, while higher levels capture broader context at the cost of detail. Crucially, this trade-off preserves inherent interpretability: any structure requiring fine-grained features for recognition can only be perceived from locations sufficiently close to the relevant image region, maintaining the model's tendency to localize. The resulting pattern of increasingly larger but progressively blurred receptive fields closely mirrors the biological principle of foveated vision, in which the density of visual receptors diminishes with distance from the fovea [47].

From here on, all following steps until the final spatial aggregation operate independently on each pixel and therefore do not incorporate further spatial information, keeping the receptive fields untouched. Subsequently, we use a deep *nonlinear* head $\boldsymbol{\chi}_{\tilde{\boldsymbol{\omega}}}^{(m)} \colon \mathbb{R}^{k \times h_{\text{out}} \times w_{\text{out}}} \to \mathbb{R}^{c \times h_{\text{out}} \times w_{\text{out}}}$ that consists of m sequentially concatenated 1×1 residual bottleneck blocks [24] to aggregate the previously defined features to pixel-wise logits $\tilde{\boldsymbol{S}} = \boldsymbol{\chi}_{\tilde{\boldsymbol{\omega}}}^{(m)}\left(\sum_{i=1}^{l} \tilde{\boldsymbol{z}}_i\right)$. The spatial accumulation of these logits and their conversion to probabilities follows the same procedure as described above.

FovBagNetLin

Using a linear head, such as $\boldsymbol{\psi}_{\boldsymbol{\omega}}$, instead, results in our *linear* baseline method *FovBagNetLin*. In this case, we can compute the level-wise saliency maps $\left(\tilde{\boldsymbol{S}}_{\text{lin}}^{(1)}, \ldots, \tilde{\boldsymbol{S}}_{\text{lin}}^{(l)}\right) \in \mathbb{R}^{l \times c \times h_{\text{out}} \times w_{\text{out}}}$ before the aggregation of levels to obtain the saliency map $\tilde{\boldsymbol{S}}_{\text{lin}} = \sum_{i=1}^{l} \tilde{\boldsymbol{S}}_{\text{lin}}^{(i)}$. Thus, we get one saliency map $\tilde{\boldsymbol{S}}_{\text{lin}}^{(i)} = \boldsymbol{\psi}_{\boldsymbol{\omega}}(\tilde{\boldsymbol{z}}_i)$, for each level i, revealing not only where, but also in which scale the object class of interest is detected. FovBagNet sacrifices this scale-wise interpretability for improved classification performance while leaving the spatial interpretability still intact. I.e., when the saliency for a given class and position is high, FovBagNet cannot tell the exact scale, or even which nonlinear combination of scales is responsible. However, the ability to learn nonlinear relationships between different scales allows the model to be spatially interpretable while incorporating global context.

Baselines

FovBagNet does not share parameters across scales, each branch of the backbone adds parameters to the model. Therefore, FovBagNet has more model parameters than BagNet. To bypass this difference, we constructed an ablation using multiple branches BagNet models that share a nonlinear head,

architecturally matched to FovBagNet in parameter count and structure, but without blurring or dilation (i.e., without the foveated design itself).

As another baseline, we modify ResNet-50 to a high resolution version. We remove the first two pooling operations to match BagNet's spatial feature dimensions, except we use zero-padding in all convolutions. To maintain the same receptive field characteristics, we compensate for the missing subsampling steps by multiplying the dilation of each convolution layer by 2^n, where n denotes the number of pooling operations removed before the corresponding layer.

Logits are not Inherently Interpretable Saliencies

Adebayo et al. show that even if a saliency map is visually appealing and plausible, it does not necessarily faithfully explain the model's decision. The logits that BagNet (as well as our method) returns can also be misleading when directly treated as saliencies: They are unnormalized, so the value $S_{j,\boldsymbol{p}}$ at a pixel coordinate $\boldsymbol{p} \in \mathbb{P} = \{1, \ldots, h\} \times \{1, \ldots, w\}$ does not in isolation reflect its contribution to the predicted probability $\hat{y}_j$: the values $S_{a,\boldsymbol{p}}$ of all competing classes $a \neq j$ are also required. For instance, if $S_{a,\boldsymbol{p}} = S_{j,\boldsymbol{p}}$ for all $a \neq j$, the softmax output $\hat{\boldsymbol{y}}$ is unaffected due to shift invariance. We therefore define the following normalized alternatives. First, the pixel-wise probabilities $\pi_{j,\boldsymbol{p}} = \mathrm{softmax}_j(\boldsymbol{S}_{\cdot,\boldsymbol{p}})$ provide a natural normalization over all classes. The remaining two saliency maps are Δ-local and Δ-global, defined as $\Delta^{\text{loc}}_{j,\boldsymbol{p}} = S_{j,\boldsymbol{p}} - \max_{a \neq j} S_{a,\boldsymbol{p}}$, and $\Delta^{\text{glob}}_{j,\boldsymbol{p}} = S_{j,\boldsymbol{p}} - S_{a^*,\boldsymbol{p}}$ with $a^* = \mathrm{argmax}_{a \neq j}\, \hat{y}_a$. Both measure the pixel-wise logit margin between the class of interest j and its strongest competitor. Δ-local selects the strongest competitor independently for each pixel, whereas Δ-global uses the single strongest competing class at image level, a^*.

Evaluation of Locality

We evaluate locality using binary object masks. Interpretability in image classification fundamentally differs from semantic segmentation, focusing on explaining model decisions rather than assigning pixel-wise labels. Nevertheless, object masks are commonly used as a practical proxy to assess whether model explanations are spatially aligned with relevant image regions [3, 40, 49, 52].

To quantify spatial alignment between a model's saliency map and the ground-truth object mask, we derive a binary classifier from each model's saliency map. Given an input image $\boldsymbol{x}$, its class label y and the corresponding binary object mask $\boldsymbol{o}$, which equals one at all pixel locations where the object class is present and zero anywhere else. Applying the model to the input image yields the saliency map $\mathbf{S}$ leading to the binary object mask prediction $\hat{\boldsymbol{o}} \in \{0, 1\}^{h \times w}$ with

$$\hat{o}_{\boldsymbol{p}} = \begin{cases} 1 & \text{if } y = \mathrm{argmax}^c_{j=1}\, S_{j,\boldsymbol{p}}, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

at the pixel $\boldsymbol{p}$. For each image in the dataset we compute a binary confusion matrix with labels $\boldsymbol{o}$ and predictions $\hat{\boldsymbol{o}}$. The sum of all these confusion matrices is the accumulated binary confusion matrix for the entire dataset, from which we derive accuracy, precision, and recall.

4 Results

In this section we compare FovBagNet against several baselines including its *linear* variant FovBagNetLin, BagNet-9,17, and 33, as well as the previously defined baseline architectures multi-branch BagNet and high resolution ResNet-50. For the following experiments, we trained all models for 100 epochs on the ImageNet [11] dataset's classification task using the categorical cross-entropy loss. For more details on our implementation and compute resources, see Appendix A and Appendix B, respectively.

As already mentioned, the model strides are set to $s = 1$ in this section, hence we can simplify the image height $h = h_{\text{in}} = h_{\text{out}} = 224$ and width $w = w_{\text{in}} = w_{\text{out}} = 224$. The evaluated version of FovBagNet operates on $l = 4$ scale levels with $\sigma = \sqrt{2}$ and patch size $q = 33$. The FovBagNet's (and the Multi-Branch BagNet's) nonlinear head is built of $m = 8$ residual bottleneck blocks. To maintain the same overall depth in all models, the BagNet backbones in the FovBagNet and the Multi-Branch BagNet are made from only 2,2,2, and 2 residual bottleneck blocks instead of 3,4,6,

and 3 like the other models including FovBagNetLin (Table 3). Since we compare the multi-branch BagNet to a FovBagNet operating on $l = 4$ scale levels, setting the number of branches to four is a natural choice, resulting in a total parameter count of 65.7M for both models. In the tables within this section, we refer to this baseline as *BagNet-33-4b*. For simplicity, we did not apply any advanced training schemes as those proposed by Wightman et al..

4.1 FovBagNet Outperforms BagNet on ImageNet

For our first experiment, we evaluated each model's classification performance on the ImageNet validation dataset. In terms of top-1 accuracy, FovBagNet outperformed BagNet-9, 17, and 33 by 30, 17, and 9 (percentage) points, respectively, while remaining within 2 points of ResNet-50 (Table 1, column 1). Most importantly BagNet-33-4b achieves 69 points and improves over plain BagNet-33 by 4 points, still falling well short of FovBagNet's 74 points. This indicates that the foveated design, not merely additional parameters or backbone count, is responsible for a significant part of the improvement. Similarly, in top-5 accuracy, our method exceeds the BagNet variants by 23, 12, and 5 points, respectively, and falls short of ResNet-50 by less than 1 point (Table 1, column 2). These results support our initial claim, that incorporating global context substantially improves over the classification performance of the BagNet models while approaching the accuracy of an unconstrained convolutional architecture.

Table 1: Image classification performance on the validation splits of ImageNet and ImageNet-S as well as mask prediction performance on the ImageNet-S validation set. All values in %.

Method	Accuracy ImageNet		Accuracy ImageNet-S		Mask Prediction		
	Top-1	Top-5	Top-1	Top-5	Acc.	Prec.	Rec.
BagNet-9	44.12	68.65	48.61	71.90	56.13	70.74	4.74
BagNet-17	56.63	79.71	62.39	83.28	57.96	**74.39**	10.31
BagNet-33	65.11	86.30	71.38	88.95	61.54	73.83	22.76
BagNet-33-4b	69.45	88.91	75.04	91.00	62.12	74.07	24.93
ResNet-50	**75.33**	**92.24**	**80.19**	**93.37**	57.31	52.30	**73.20**
FovBagNetLin	68.55	88.55	74.61	90.77	63.32	63.99	42.29
FovBagNet	74.03	91.50	78.87	92.90	**66.94**	68.54	49.43

The values reported in Table 1 are means over three random seeds per configuration; corresponding standard deviations are reported in subsection C.1. While classification accuracy is stable across all methods, ResNet-50's mask prediction scores vary an order of magnitude more across seeds than those of BagNet and FovBagNet, which we attribute to its lack of architectural constraints tying predictions to localized input regions. This performance analysis, however, does not provide insight into the models' interpretability properties, which we analyze in subsequent experiments.

4.2 FovBagNet Provides Spatially Aligned Saliencies

Next, we evaluated the saliency maps of our models trained on ImageNet using the object mask labels provided by ImageNet-S [18], a subset of the ImageNet validation set comprising 12k of 50k images across 919 of 1000 classes. For evaluation, we applied a 192×192 center-cropping on both masks and saliencies. Images with no object present inside this crop were discarded. How we derive metrics to quantify the spatial alignment between saliency maps and object masks as a proxy for interpretability is described in section 3.

Before analyzing interpretability, we examine classification performance on ImageNet-S. FovBagNet again outperformed BagNet-9, 17, and 33 by 30, 18, and 8 points in top-1 accuracy, while matching ResNet-50 within one point (Table 1, column 3). Top-5 accuracy shows the same trend (Table 1, column 4). Overall, the ImageNet-S classification results are similar to the ones on ImageNet, except that all models performed slightly worse on the full ImageNet validation set. We attribute this phenomenon to a selection bias: images that are less ambiguous to label for humans—but also for artificial neural networks—were more likely to become a part of ImageNet-S. Consequently, many difficult samples were filtered out.

Turning to interpretability, we assess spatial alignment between saliency maps and object masks using accuracy, precision, and recall as described in section 3 (Table 1, columns 5–7). FovBagNet achieved the highest accuracy, followed by FovBagNetLin. Across models, we observe a clear precision-recall tradeoff: BagNet models are highly precise but achieve very low recall (5–22 points), whereas ResNet-50 shows the inverse pattern with the highest recall (76 points) but the lowest precision. FovBagNet occupies a middle ground, balancing precision and recall more evenly than either baseline. The large variances in precision and recall suggest that the classifiers strongly differ in terms of sensitivity. To gain more nuanced and comparable insights, we analyzed the receiver operating characteristic (ROC) and precision-recall (PR) curves. We used the logit activation map $\boldsymbol{S}_y$ of the true class y as thresholds and observed that FovBagNet achieved the best performance across both ROC-AUC and PR-AUC, outperforming all BagNet variants and ResNet-50 (Table 2, first two columns), surpassing the strongest baseline, BagNet-33, by four points in ROC-AUC and one point in PR-AUC. The ROC curve further shows FovBagNet consistently on top, with only slightly lower true positive rates than the BagNets at very low false positive rates (Figure 2, first column). The PR curve shows that FovBagNet maintains higher precision at mid-to-high recall, indicating an improved ranking of hard positives.

As a complementary, more direct faithfulness check, we perturbed each model's most salient pixels and measured the resulting drop in predicted-class confidence (AOPC; subsection C.3, Table 9). Raw AOPC favors the highest-performing classifiers—ResNet-50 and FovBagNet ($q = 33$) lead at 66.3% and 65.2%—but conflates saliency quality with general occlusion sensitivity. Normalizing against a random-perturbation baseline removes this confound: ResNet-50's advantage disappears entirely, while FovBagNetLin and FovBagNet ($q = 33$) show the largest gaps (up to 6.8 pp), confirming genuinely informative saliency maps. FovBagNet ($q = 17$) and FovBagNet ($q = 9$) retain smaller but clearly positive gaps (4.64 and 2.85 pp), whereas all BagNet variants score negatively, indicating random perturbation disturbs their predictions more than saliency-guided perturbation does.

4.3 FovBagNet Shows Robust Interpretability Performance Across Saliencies

Next, we evaluated all models using the normalized saliency maps introduced in section 3, which yield more spatially meaningful explanations than raw logit saliencies. Compared to logit saliencies, ROC and PR-AUCs decreased for all models and methods, except the Probabilities-PR-AUC of ResNet-50, which stayed equal (Table 2). This indicates that these tasks are more difficult because a pixel must be attributed to a single class and cannot be considered universally important anymore. BagNet models showed the strongest drop, whereas ResNet-50 remained relatively stable across saliency maps but stayed consistently behind FovBagNet and FovBagNetLin. Notably, this drop is pronounced even for BagNet-33-4b, which matches FovBagNet's parameter count: while the two differ by only 5 pp in Top-1 classification accuracy (Table 1), BagNet-33-4b trails FovBagNet by 3–11 pp across the four saliency maps in Table 2 (e.g., 56.31% vs. 66.82% Probabilities-ROC, 54.69% vs. 64.32% Δ-Local-ROC), indicating that the foveated design contributes to saliency alignment beyond what matched capacity alone provides. Note that BagNet-9 and BagNet-17 have lower AUCs than random guessing for probabilities and Δ-local saliencies at recalls above 0.25 (Figure 2). We investigated this phenomenon and found that it is due to the presence of strong shortcuts within certain object classes [20] (examples in the Appendix: Figure 6, Figure 5). As visible in the PR curves, BagNet models achieved higher precision at low recall, while FovBagNet maintained higher precision at mid-to-high recall across all saliency maps (Figure 2). Therefore, FovBagNet demonstrates a better ranking of hard positives and a more robust interpretability performance across all saliency maps, resulting in higher areas under the curves (Table 2).

Figure 3 visualizes saliency maps generated by different attribution methods for all models on two representative input images. The qualitative patterns reflect the quantitative findings from the preceding analysis. For the BagNet models, saliency maps derived from logits appeared most interpretable, whereas more valid normalized attribution methods provided limited insight. Overall, these maps were highly localized and predominantly focused on texture-like features rather than global object structure. In contrast, the saliency maps of ResNet-50 were diffuse and poorly localized, which was consistent with its large receptive field and reduced spatial specificity.

Our FovBagNet models, however, produced informative and consistent saliency maps across all attribution methods. The highlighted regions extended beyond purely local textures and captured more semantically meaningful object parts, indicating a balance between local and global feature

Table 2: Saliency map alignment with object masks on the ImageNet-S validation set. Areas under ROC and PR-Curves for various saliency maps derived from different scores. All values in %.

Method	Logits		Probabilities		Δ-Local		Δ-Global	
	ROC	PR	ROC	PR	ROC	PR	ROC	PR
BagNet-9	63.84	62.74	44.11	44.66	43.17	43.64	56.18	55.73
BagNet-17	69.78	67.66	47.48	49.87	46.35	48.44	58.90	59.16
BagNet-33	74.69	69.79	57.48	58.79	55.76	56.92	62.40	61.77
BagNet-33-4b	74.86	70.11	56.31	58.25	54.69	56.64	63.02	62.61
ResNet-50	67.68	58.14	64.87	57.80	63.04	56.27	62.23	55.60
FovBagNetLin	74.44	66.93	**66.89**	61.74	**64.95**	60.50	63.68	59.69
FovBagNet	**78.31**	**71.04**	66.82	**65.62**	64.32	**63.54**	**67.39**	**65.14**

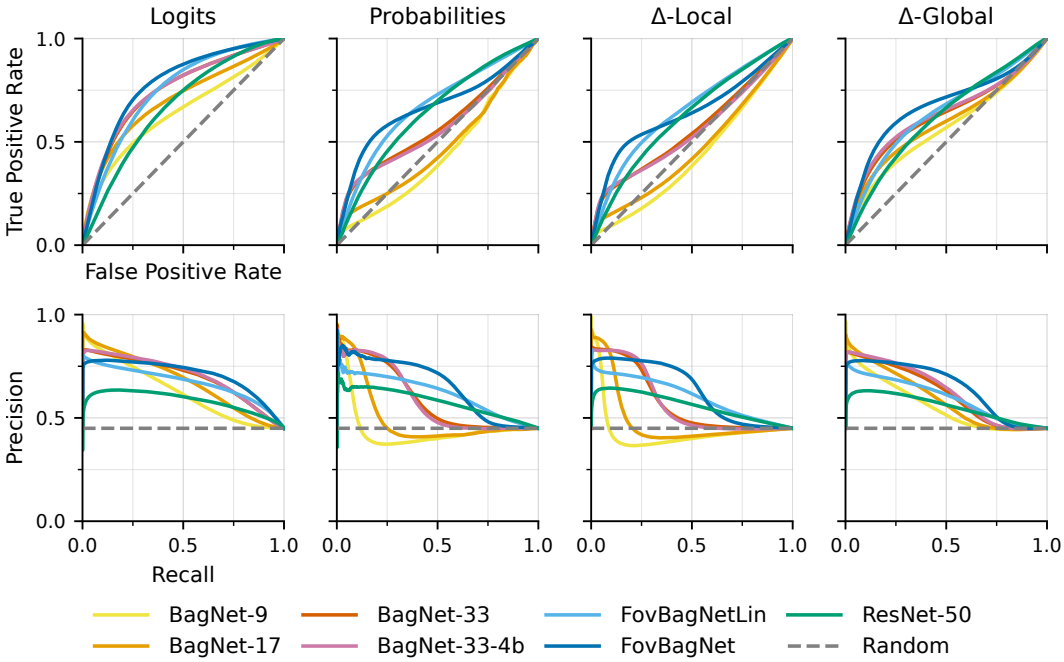

Figure 2: ROC- and PR-curves for different saliency scores.

utilization. The qualitative saliencies also indicate how FovBagNet improves upon its linear version. FovBagNet demonstrates enhanced mask alignment with more defined object borders and fewer attribution holes in areas where objects are visible. More examples are shown in the Appendix (Figure 4, Figure 5, Figure 6).

5 Conclusion and Discussion

Our results suggest that the accuracy gap between interpretable bag-of-features models and standard deep networks can largely be closed by incorporating foveated, multiscale structure inspired by the human visual system. FovBagNet achieves 74% top-1 accuracy on ImageNet, closing most of the 11-point gap between BagNet-33 and ResNet-50, while retaining spatially faithful saliency maps, demonstrating that interpretability and competitive performance are not mutually exclusive. By narrowing this gap, we expect to lower the barrier to adopting inherently interpretable models in practice—a shift that has the potential to expose unfair or spurious behavior in vision models.

The current design uses an independent BagNet for each scale level, increasing the parameter count, memory, and compute costs compared to baselines. These branches are independent and therefore parallelizable, and our full-resolution evaluation runs approximately 40× faster than the official BagNet implementation. Although overfitting was not observed, the performance gap between

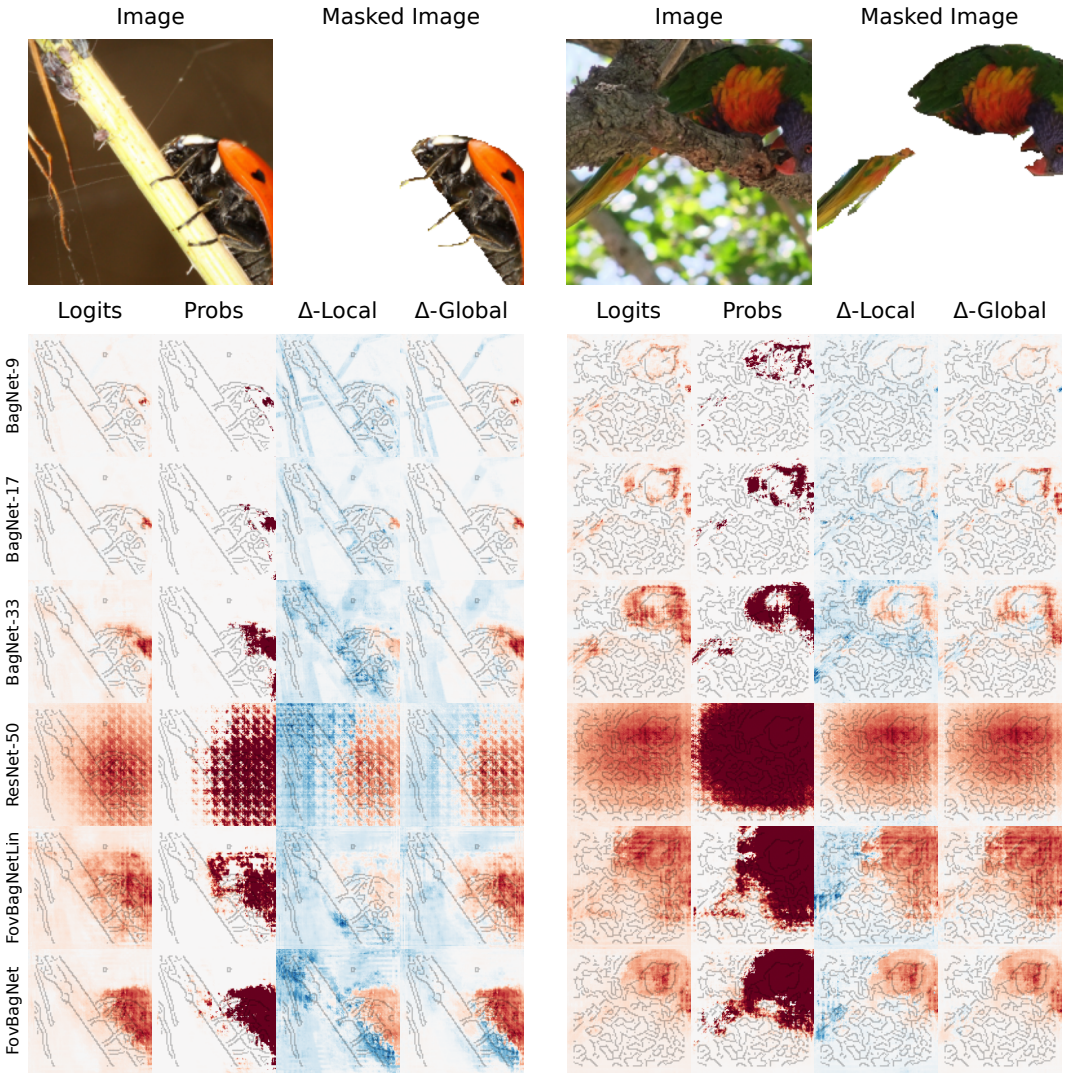

Figure 3: The four saliency maps logits, probabilities, Δ-global, and Δ-local (columns) generated by different models (rows) from two input images. The masked image shows the object of interest.

FovBagNetLin and FovBagNet suggests that the multiscale structure itself, rather than the additional parameters, drives the improvement. If only a single scale were beneficial, the gap between the two variants would be smaller. Future work should nevertheless explore cross-scale weight sharing to improve parameter efficiency and achieve scale invariance.

Moreover, a hyperparameter ablation of blur strength σ, the number of scale levels l, and the receptive field size q (subsection C.2) clarified how FovBagNet's design choices interact: scale levels and receptive field size can partially substitute for one another, additional levels yield diminishing returns, and the model remains robust to the blur strength within the range tested.

Beyond interpretability of image classification, we see particular promise in self-, semi-, and unsupervised medical imaging tasks. Histopathological whole slide imaging is a natural fit because tissue is examined at multiple magnifications, which mirrors FovBagNet's multiscale structure. Additionally, the bag-of-features formulation naturally aligns with multiple instance learning. In high-stakes clinical settings, the model's spatially grounded predictions are more than a theoretical convenience—interpretability is a practical prerequisite for adoption.

Acknowledgments and Disclosure of Funding

Acknowledgments

The authors acknowledge the support of the Training Center for Machine Learning (TCML) Tübingen, funded by the German Federal Ministry of Education and Research (BMBF) with grant number 01IS17054. This research also utilized compute resources at the Tübingen Machine Learning Cloud, funded by the German Research Foundation (DFG) under grant number DFG FKZ INST 37/1057-1 FUGG. This work was further supported by the Hertie Foundation through its funding of the Hertie Institute for AI in Brain Health. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting SM.

References

[1] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

[2] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton. Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, 34:4699–4711, 2021.

[3] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3319–3327, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.354. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.354.

[4] N. M. Blauch, G. A. Alvarez, and T. Konkle. Fovi: A biologically-inspired foveated interface for deep vision models. *arXiv preprint arXiv:2602.03766*, 2026.

[5] M. Böhle, N. Singh, M. Fritz, and B. Schiele. B-cos alignment for inherently interpretable cnns and vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (6):4504–4518, 2024.

[6] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

[7] W. Brendel and M. Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *International Conference on Learning Representations*, 2019. URL https://openreview.net/pdf?id=SkfMWhAqYQ.

[8] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[9] L. Chen, S. Lou, K. Zhang, J. Huang, and Q. Zhang. Harsanyinet: Computing accurate shapley values in a single forward propagation. *arXiv preprint arXiv:2304.01811*, 2023.

[10] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson. Human photoreceptor topography. *Journal of comparative neurology*, 292(4):497–523, 1990.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[12] K. Djoumessi, B. Bah, L. Kühlewein, P. Berens, and L. Koch. This actually looks like that: Proto-bagnets for local and global interpretability-by-design. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 718–728. Springer, 2024.

[13] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[14] J. Edin, A. G. Motzfeldt, C. L. Christensen, T. Ruotsalo, L. Maaløe, and M. Maistro. Normalized aopc: fixing misleading faithfulness metrics for feature attributions explainability. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1730, 2025.

[15] G. Elsayed, S. Kornblith, and Q. V. Le. Saccader: Improving accuracy of hard attention models for vision. *Advances in neural information processing systems*, 32, 2019.

[16] W. Falcon and The PyTorch Lightning team. Pytorch lightning, Mar. 2019. URL `https://doi.org`.

[17] K. Fan, Y. Yang, and C. Ma. Enhancing interpretability for vision models via shapley value optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 3786–3794, 2026.

[18] S. Gao, Z.-Y. Li, M.-H. Yang, M.-M. Cheng, J. Han, and P. Torr. Large-scale unsupervised semantic segmentation. 2022.

[19] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019.

[20] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[21] A. Gizdov, S. Ullman, and D. Harari. Seeing more with less: human-like representations in vision models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4408–4417, 2025.

[22] S. Hanson and L. Pratt. Comparing biases for minimal network construction with back-propagation. *Advances in neural information processing systems*, 1, 1988.

[23] T. Hastie and R. Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[25] A. Jonnalagadda, W. Y. Wang, B. Manjunath, and M. P. Eckstein. Foveater: Foveated transformer for image classification. *arXiv preprint arXiv:2105.14173*, 2021.

[26] G. W. Killick, P. Henderson, J. P. Siebert, and G. Aragon-Camarasa. Foveation in the era of deep learning. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. URL `https://papers.bmvc2023.org/0703.pdf`.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[28] Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[29] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[30] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[31] J. Min, Y. Zhao, C. Luo, and M. Cho. Peripheral vision transformer. *Advances in Neural Information Processing Systems*, 35:32097–32111, 2022.

[32] F. Mumuni and A. Mumuni. Explainable artificial intelligence (xai): from inherent explainability to large language models. *arXiv preprint arXiv:2501.09967*, 2025.

[33] M. P. Panda, M. Tiezzi, M. Vilas, G. Roig, B. M. Eskofier, and D. Zanca. Fovex: Human-inspired explanations for vision transformers and convolutional neural networks. *Int. J. Comput. Vision*, 133(10):7437–7459, July 2025. ISSN 0920-5691. doi: 10.1007/s11263-025-02543-y. URL https://doi.org/10.1007/s11263-025-02543-y.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[35] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(64)90137-5. URL https://www.sciencedirect.com/science/article/pii/0041555364901375.

[36] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

[37] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[38] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[39] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

[41] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[42] M. Shah, A. Kashaf, and B. Raj. Training on foveated images improves robustness to adversarial attacks. *Advances in Neural Information Processing Systems*, 36:71008–71028, 2023.

[43] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[44] A. Sudjianto and A. Zhang. Designing inherently interpretable machine learning models. *arXiv preprint arXiv:2111.01743*, 2021.

[45] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[46] M. Tiezzi, S. Marullo, A. Betti, E. Meloni, L. Faggi, M. Gori, and S. Melacci. Foveated neural computation. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 19–35. Springer, 2022.

[47] E. Wells-Gray, S. Choi, A. Bries, and N. Doble. Variation in rod and cone density from the fovea to the mid-periphery in healthy human retinas using adaptive optics scanning laser ophthalmoscopy. *Eye*, 30(8):1135–1143, 2016.

[48] R. Wightman, H. Touvron, and H. Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021.

[49] J. Wilfling, M. Valdenegro-Toro, and M. Zullich. Evaluating the quality of saliency maps for distilled convolutional neural networks. In *ESANN*, 2024.

[50] M. Xue, Q. Huang, H. Zhang, J. Hu, J. Song, M. Song, and C. Jin. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. *arXiv preprint arXiv:2208.10431*, 2022.

[51] A. Yu and Y.-B. Yang. Prism: Prototype-based reasoning with inter-modal semantic mining for interpretable image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2853–2863, 2026.

[52] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[53] P. Zschech, S. Weinzierl, and M. Kraus. Inherently interpretable machine learning: A contrasting paradigm to post-hoc explainable ai: P. zschech et al. *Business & Information Systems Engineering*, pages 1–19, 2025.

A Implementation Details

The code repository is available under `https://github.com/hheidrich/FovBagNet`. We implemented all our experiments using the libraries PyTorch [34] and PyTorch Lightning [16]. For our experiments, we trained all models for $T = 100$ epochs on ImageNet's train split using the categorical cross entropy loss. As an optimizer we applied stochastic gradient descent with an overall batch-size of 256, an initial learning rate $\eta = 0.1$, momentum $\alpha = 0.9$, and weight decay $\lambda = 10^{-4}$ [6, 22, 35]. We used cosine annealing [29] to reduce the learning rate $\eta_t = \frac{\eta}{2}\left(1 + \cos\left(\frac{t}{T}\pi\right)\right)$ after each epoch $t \in \{1, \dots, T\}$. The versions of FovBagNet and FovBagNetLin were both trained and evaluated on $l = 4$ scale levels with patch size $q = 33$. All models are trained—like the original BagNets—on a total stride factor $s = 8$ as shown in Table 3. During the training, we augmented the input images using random resized cropping (with scale from 0.08 to 1 and ratio between $3/4$ and $4/3$), horizontal flipping (with $p = 0.5$), and color jittering (with brightness 0.4, contrast 0.4, saturation 0.4).

For the evaluation, we increased the saliency map resolution to match the input resolution $h = w = 224$. Therefore, we reduced the total stride factor to $s = 1$. Internally, our implementation removes a stride in the model architecture and introduces dilations in all subsequent layers to increase the output resolution while keeping the model topology unchanged. This approach is much faster than computing all patches individually because it reuses the intermediate results shared by adjacent patches.

Table 3: Architectures of the models we evaluated in this work's experiments. A stride of two is applied once within the layers conv2_x, conv3_x, and conv4_x, respectively. This causes for the reductions of resolution. Note that layer conv5_x does not apply any strides larger than one—the output resolution remains unchanged.

Layer	$h \times w$ out	BagNet-9	BagNet-17	BagNet-33	ResNet-50 (hr.)	FovBagNetLin	FovBagNet
conv1_1	224×224	3×3, 64, stride 1					
conv1_2	224×224	3×3, 64, stride 1					
conv2_x	112×112	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 64 \\ 1\times1, 64 \\ 1\times1, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 64 \\ 1\times1, 64 \\ 1\times1, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 64 \\ 1\times1, 64 \\ 1\times1, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 64 \\ 1\times1, 64 \\ 1\times1, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 64 \\ 1\times1, 64 \\ 1\times1, 256 \end{bmatrix} \times 1$
conv3_x	56×56	$\begin{bmatrix} 1\times1, 128 \\ 1\times1, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 128 \\ 1\times1, 128 \\ 1\times1, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 128 \\ 1\times1, 128 \\ 1\times1, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 128 \\ 1\times1, 128 \\ 1\times1, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 128 \\ 1\times1, 128 \\ 1\times1, 512 \end{bmatrix} \times 1$
conv4_x	28×28	$\begin{bmatrix} 1\times1, 256 \\ 1\times1, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, 256 \\ 1\times1, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 256 \\ 1\times1, 256 \\ 1\times1, 1024 \end{bmatrix} \times 5$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 256 \\ 1\times1, 256 \\ 1\times1, 1024 \end{bmatrix} \times 5$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 1$ $\begin{bmatrix} 1\times1, 256 \\ 1\times1, 256 \\ 1\times1, 1024 \end{bmatrix} \times 1$
conv5_x	28×28	$\begin{bmatrix} 1\times1, 512 \\ 1\times1, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 1\times1, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 1\times1, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 1\times1, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 1\times1, 512 \\ 1\times1, 2048 \end{bmatrix} \times 2$
		—				sum of all levels	
head	28×28						$\begin{bmatrix} 1\times1, 512 \\ 1\times1, 512 \\ 1\times1, 2048 \end{bmatrix} \times 8$
	1×1	average pool, 1000-d fc, softmax					

B Compute Resources

The training of a FovBagNet model took approximately 80 hours and 140 hours for FovBagNetLin on an internal compute cluster node. Both models were trained parallelly across 4 Nvidia A100-GPUs with 40 GB of VRAM each. On a single consumer-level GPU (Nvidia GeForce RTX 3060) FovBagNet's inference time is 0.9 seconds per image with a consumption of approximately 6.2 GB

of VRAM (1.1 seconds and 6.3 GB for FovBagNetLin).

C Additional Quantitative Evaluation

C.1 Multiple Runs

To report error bars of the experimental results of Table 1, we ran three models (with different random seeds) per configuration. Results are reported in Table 4 (classification) and Table 5 (mask prediction).

Classification accuracy is stable across all methods, including ResNet-50, with standard deviations below 1 percentage point throughout Table 4. For mask prediction, however, BagNet and FovBagNet remain tightly clustered (std ≤ 0.6 pp across accuracy, precision, and recall), while ResNet-50's std is an order of magnitude larger (± 4.51, ± 4.05, ± 5.51, respectively). Since this instability is specific to mask prediction, we hypothesize it stems from ResNet-50 lacking the architectural constraints that tie patch-based predictions to fixed, localized input regions: different seeds may reach comparably accurate classifiers that rely on different, less spatially consistent evidence. This suggests the architectural constraints underlying BagNet and FovBagNet make the spatial basis of their predictions–and thus their saliency and mask predictions–substantially more reproducible across training runs.

Table 4: Average image classification performance with standard deviations across three seeds on the validation splits of ImageNet and ImageNet-S. All values in %.

Method	Accuracy ImageNet		Accuracy ImageNet-S	
	Top-1	Top-5	Top-1	Top-5
BagNet-9	44.12 ± 0.07	68.65 ± 0.15	48.61 ± 0.11	71.90 ± 0.03
BagNet-17	56.63 ± 0.70	79.71 ± 0.56	62.39 ± 0.28	83.28 ± 0.08
BagNet-33	65.11 ± 0.34	86.30 ± 0.31	71.38 ± 0.18	88.95 ± 0.16
BagNet-33-4b	69.45 ± 0.26	88.91 ± 0.16	75.04 ± 0.25	91.00 ± 0.10
ResNet-50	**75.33** ± 0.29	**92.24** ± 0.11	**80.19** ± 0.17	**93.37** ± 0.08
FovBagNetLin ($q = 33$)	68.55 ± 0.23	88.55 ± 0.23	74.61 ± 0.32	90.77 ± 0.13
FovBagNet ($q = 9$)	63.48 ± 0.36	84.99 ± 0.27	69.43 ± 0.39	87.50 ± 0.24
FovBagNet ($q = 17$)	70.71 ± 0.18	89.63 ± 0.05	76.20 ± 0.08	91.45 ± 0.07
FovBagNet ($q = 33$)	74.03 ± 0.15	91.50 ± 0.13	78.87 ± 0.13	92.90 ± 0.13

Table 5: Average mask prediction performance with standard deviations across three seeds on the ImageNet-S validation set. All values in %.

Method	Mask Prediction		
	Accuracy	Precision	Recall
BagNet-9	56.13 ± 0.02	70.74 ± 0.12	4.74 ± 0.10
BagNet-17	57.96 ± 0.02	**74.39** ± 0.11	10.31 ± 0.10
BagNet-33	61.54 ± 0.12	73.83 ± 0.22	22.76 ± 0.50
BagNet-33-4b	62.12 ± 0.15	74.07 ± 0.01	24.93 ± 0.56
ResNet-50	57.31 ± 4.51	52.30 ± 4.05	**73.20** ± 5.51
FovBagNetLin ($q = 33$)	63.32 ± 0.25	63.99 ± 0.10	42.29 ± 0.58
FovBagNet ($q = 9$)	60.05 ± 0.08	74.09 ± 0.05	17.89 ± 0.28
FovBagNet ($q = 17$)	64.54 ± 0.07	73.56 ± 0.15	33.99 ± 0.31
FovBagNet ($q = 33$)	**66.94** ± 0.14	68.54 ± 0.35	49.43 ± 0.60

C.2 Hyperparameter Ablation Study

FovBagNet is configurable through certain hyperparameters:

- Gaussian kernel blur strength σ: All experiments reported in this work use $\sigma = \sqrt{2}$. Note that none ($\sigma = 0$) or too weak blurring can introduce aliasing. Blurring too strongly ($\sigma \to \infty$) would eventually eliminate all information except that of the first scale (which is not blurred and therefore not affected by σ). We ablate $\sigma \in \{1, 2\}$ below.
- Number of scale levels l: A higher number gives more global context. Our default is $l = 4$. We ablate $l \in \{2, 4, 5, 6\}$ below.
- Patch size q: Receptive field size of a BagNet architecture.
- Depth of the nonlinear head m: Our main results use a 26-layer BagNet backbone (2-2-2-2 blocks) with 8 additional blocks in the nonlinear head to match the depth of a ResNet-50.

For our ablation study, we ablated σ, the number of scale levels l, and the patch size q, as we expected the highest effect of them. We trained and evaluated additional versions of FovBagNet with patch sizes $q = 9$ and $q = 17$ at the same number of scale levels as our main configuration $l = 4$, similar to the original BagNet paper's patch-size sweep. To compensate for the smaller patch sizes, we also trained FovBagNet versions with more scale levels as the configurations ($q = 17, l = 5$) and ($q = 9, l = 6$).

Table 6: Ablations of the patch size q and the number of scale levels l with their parameter counts and resulting ImageNet validation accuracies in %.

Method	q	l	σ	Parameter Count	ImageNet Val Accuracy
FovBagNet	33	4	$\sqrt{2}$	65.7M	74.0
FovBagNet	33	4	1	65.7M	74.0
FovBagNet	33	4	2	65.7M	73.0
FovBagNet	33	2	$\sqrt{2}$	43.3M	72.9
FovBagNet	17	4	$\sqrt{2}$	57.3M	70.7
FovBagNet	17	5	$\sqrt{2}$	66.4M	70.7
FovBagNet	9	4	$\sqrt{2}$	55.2M	63.5
FovBagNet	9	6	$\sqrt{2}$	72.4M	65.3
BagNet-33-4b	33	-	-	65.7M	69.5

First, we compared the ablations on the ImageNet classification task. The main takeaways from Table 6 are:

- Blur strength has a small but still measurable effect: fixing $q = 33, l = 4$, accuracy is stable between $\sigma = 1$ and $\sigma = \sqrt{2}$ (74.0% both), but degrades at $\sigma = 2$ (73.0%), a drop consistent with the hypothesis above.
- Patch size drives accuracy: as with BagNet, larger q gives higher accuracy; this holds for FovBagNet too ($63.5\% \to 70.7\% \to 74\%$ as $q = 9 \to 17 \to 33$, $l = 4$ fixed).
- More levels help, but with steeply diminishing returns: fixing $q = 33$, accuracy rises from 72.9% at $l = 2$ to 74% at $l = 4$ (a 1.1pp gain for 22.4M additional parameters), whereas at $q = 17$ going from $l = 4$ to $l = 5$ there are no significant gains (70.7% both) for 9.1M additional parameters. The earliest scale levels thus contribute disproportionately more than later ones, which is consistent with most of the useful global context being captured within the first few scales. Notably, even the leanest configuration tested ($q = 33, l = 2$, 43.3M params) still outperforms the BagNet-33-4b baseline (69.5%) while using over 20M fewer parameters.
- Levels can trade off against patch size, but only to a point: ($q = 17, l = 4, 70.7\%$) already beats the BagNet-33-4b ($q = 33, 69.5\%$), despite using patches half the size. But this compensation has limits: even with two extra levels, ($q = 9, l = 6, 65.3\%$) still falls short of the same 69.5% baseline.

Second, we isolated the effect of different patch sizes for FovBagNet on the task of saliency map alignment with object masks from ImageNet-S, keeping the number of levels constant at $l = 4$ (Table 7, Table 8).

Table 7: Saliency map alignment with object masks on the ImageNet-S validation set. Areas under ROC and PR-Curves for various saliency maps derived from logits and probabilities. All values in %.

Method	Logits		Probabilities	
	ROC-AuC	PR-AuC	ROC-AuC	PR-AuC
BagNet-9	63.84 ± 0.12	62.74 ± 0.12	44.11 ± 0.12	44.66 ± 0.04
BagNet-17	69.78 ± 0.10	67.66 ± 0.15	47.48 ± 0.00	49.87 ± 0.04
BagNet-33	74.69 ± 0.03	69.79 ± 0.09	57.48 ± 0.19	58.79 ± 0.22
BagNet-33-4b	74.86 ± 0.12	70.11 ± 0.07	56.31 ± 0.33	58.25 ± 0.29
ResNet-50	67.68 ± 1.21	58.14 ± 0.25	64.87 ± 1.06	57.80 ± 0.32
FovBagNetLin ($q = 33$)	74.44 ± 0.68	66.93 ± 0.74	**66.89** ± 0.47	61.74 ± 0.24
FovBagNet ($q = 9$)	72.81 ± 0.14	70.66 ± 0.15	51.36 ± 0.14	54.29 ± 0.15
FovBagNet ($q = 17$)	77.75 ± 0.16	**73.35** ± 0.22	59.62 ± 0.24	61.86 ± 0.24
FovBagNet ($q = 33$)	**78.31** ± 0.21	71.04 ± 0.19	66.82 ± 0.24	**65.62** ± 0.21

Table 8: Saliency map alignment with object masks on the ImageNet-S validation set. Areas under ROC and PR-Curves for various saliency maps derived from Δ-Local and Δ-Global. All values in %.

Method	Δ-Local		Δ-Global	
	ROC-AuC	PR-AuC	ROC-AuC	PR-AuC
BagNet-9	43.17 ± 0.03	43.64 ± 0.04	56.18 ± 0.03	55.73 ± 0.02
BagNet-17	46.35 ± 0.03	48.44 ± 0.02	58.90 ± 0.04	59.16 ± 0.07
BagNet-33	55.76 ± 0.17	56.92 ± 0.12	62.40 ± 0.09	61.77 ± 0.06
BagNet-33-4b	54.69 ± 0.34	56.64 ± 0.26	63.02 ± 0.19	62.61 ± 0.11
ResNet-50	63.04 ± 1.58	56.27 ± 0.54	62.23 ± 1.21	55.60 ± 0.31
FovBagNetLin ($q = 33$)	**64.95** ± 0.47	60.50 ± 0.27	63.68 ± 0.19	59.69 ± 0.20
FovBagNet ($q = 9$)	50.23 ± 0.17	53.34 ± 0.18	60.78 ± 0.10	61.67 ± 0.15
FovBagNet ($q = 17$)	57.96 ± 0.26	60.60 ± 0.14	64.98 ± 0.06	**65.27** ± 0.11
FovBagNet ($q = 33$)	64.32 ± 0.22	**63.54** ± 0.19	**67.39** ± 0.12	65.14 ± 0.15

Remarkably, the FovBagNet with $q = 17$ achieves higher scores than the BagNet-33-4b in all categories using fewer parameters (57.3M vs. 65.7M) while producing comparable classification results on the ImageNet validation set (70.7% vs. 69%). We interpret this as evidence of the effectiveness of foveation pixel localization. This is further emphasized by the fact that the BagNet-33-4b baseline only slightly improves upon the original BagNet-33, despite an improvement in image classification from 64% to 69%.

C.3 Empirical Faithfulness Evaluation via Pixel Perturbation

To empirically evaluate the faithfulness of FovBagNet, we perturb the pixels with the highest impact, as defined by our model's predicted saliency maps, and approximate the area above the perturbation curve (AOPC) [39]:

$$\text{AOPC}(\varphi, \boldsymbol{x}) = \int_0^1 \varphi(\boldsymbol{x})_{j^*} - \varphi(\boldsymbol{x}_{\varphi,\alpha})_{j^*} \, d\alpha, \tag{8}$$

for model φ on image $\boldsymbol{x}$. Here, α is the strength of the perturbation: it determines which quantile of the most salient pixels (according to φ) is set to zero in the perturbed image $\boldsymbol{x}_{\varphi,\alpha}$. By definition,

$\boldsymbol{x}_{\varphi,0} = \boldsymbol{x}$ holds for any φ. The subscript $j^* = \operatorname{argmax}_{j=1}^{c} \varphi(\boldsymbol{x})_j$ denotes the index of the predicted class, given the original image $\boldsymbol{x}$.

In practice, we approximate the integral by computing the difference $\varphi(\boldsymbol{x})_{j^*} - \varphi(\boldsymbol{x}_{\varphi,\alpha})_{j^*}$ over a finite set of interval boundaries $\alpha \in \mathbb{A}$, and numerically integrate using the trapezoidal method. We distribute the boundaries $\mathbb{A}$ geometrically rather than linearly to increase sensitivity to earlier stages of perturbation

$$\mathbb{A} = \left\{0, \sqrt{2}^{(-15)}, \sqrt{2}^{(-14)}, \ldots, \sqrt{2}^{(-1)}, 1\right\}. \tag{9}$$

This gives $|\mathbb{A}| = 17$ perturbation strengths at which the difference is evaluated. Table 9 lists the resulting AOPC values.

Table 9: Faithfulness evaluation with pixel perturbations on the ImageNet-S validation set. All values in %.

Method	AOPC (saliency-guided)	AOPC (random)	Normalized AOPC (saliency−random)
BagNet-9	34.63 ± 0.22	35.04 ± 0.24	-0.42 ± 0.02
BagNet-17	48.11 ± 0.09	49.13 ± 0.10	-1.01 ± 0.03
BagNet-33	56.31 ± 0.30	58.42 ± 0.22	-2.11 ± 0.08
BagNet-33-4b	62.26 ± 0.19	64.23 ± 0.17	-1.97 ± 0.28
ResNet-50	**66.31** ± 1.65	**66.58** ± 0.13	-0.26 ± 1.65
FovBagNetLin ($q = 33$)	56.64 ± 0.56	49.82 ± 1.67	**6.82** ± 1.33
FovBagNet ($q = 9$)	53.63 ± 0.57	50.78 ± 0.93	2.85 ± 0.36
FovBagNet ($q = 17$)	60.85 ± 0.22	56.21 ± 0.68	4.64 ± 0.64
FovBagNet ($q = 33$)	65.16 ± 0.10	58.77 ± 0.81	6.39 ± 0.71

The plain AOPC score favors models with high classification performance: ResNet-50 and FovBagNet ($q = 33$) lead at 66.3% and 65.2%, respectively, followed by BagNet-33, FovBagNet ($q = 17$), FovBagNetLin, and FovBagNet ($q = 9$) at intermediate levels (62.3%, 60.9%, 56.6%, and 53.6%), while BagNet-17 and BagNet-9 trail at 48.1% and 34.6%.

Normalization against a random-perturbation baseline changes this ranking substantially. ResNet-50's raw AOPC advantage turns out to be an artifact of general occlusion sensitivity rather than saliency quality: its random baseline (66.6%) is actually slightly higher, leaving no measurable saliency-specific gap. By contrast, FovBagNetLin and FovBagNet ($q = 33$) show the largest normalized gaps (6.8 and 6.4 pp, with overlapping standard deviations), despite their lower raw scores, indicating that their saliency maps are comparatively more informative about which pixels actually drive the prediction. FovBagNet ($q = 17$) retains a smaller but still clearly positive gap (4.64 ± 0.64 pp), consistent with some loss of faithfulness at coarser resolution. All BagNet variants show significantly negative gaps, indicating that random perturbation disturbs their predictions more than saliency-guided perturbation.

In short: raw AOPC alone favors ResNet-50, but normalizing against random perturbation (as Edin et al. [14] recommends) reveals FovBagNet as the most faithful model by this metric, while also exposing a genuine weakness in the BagNet baselines.

D Additional Qualitative Evaluation

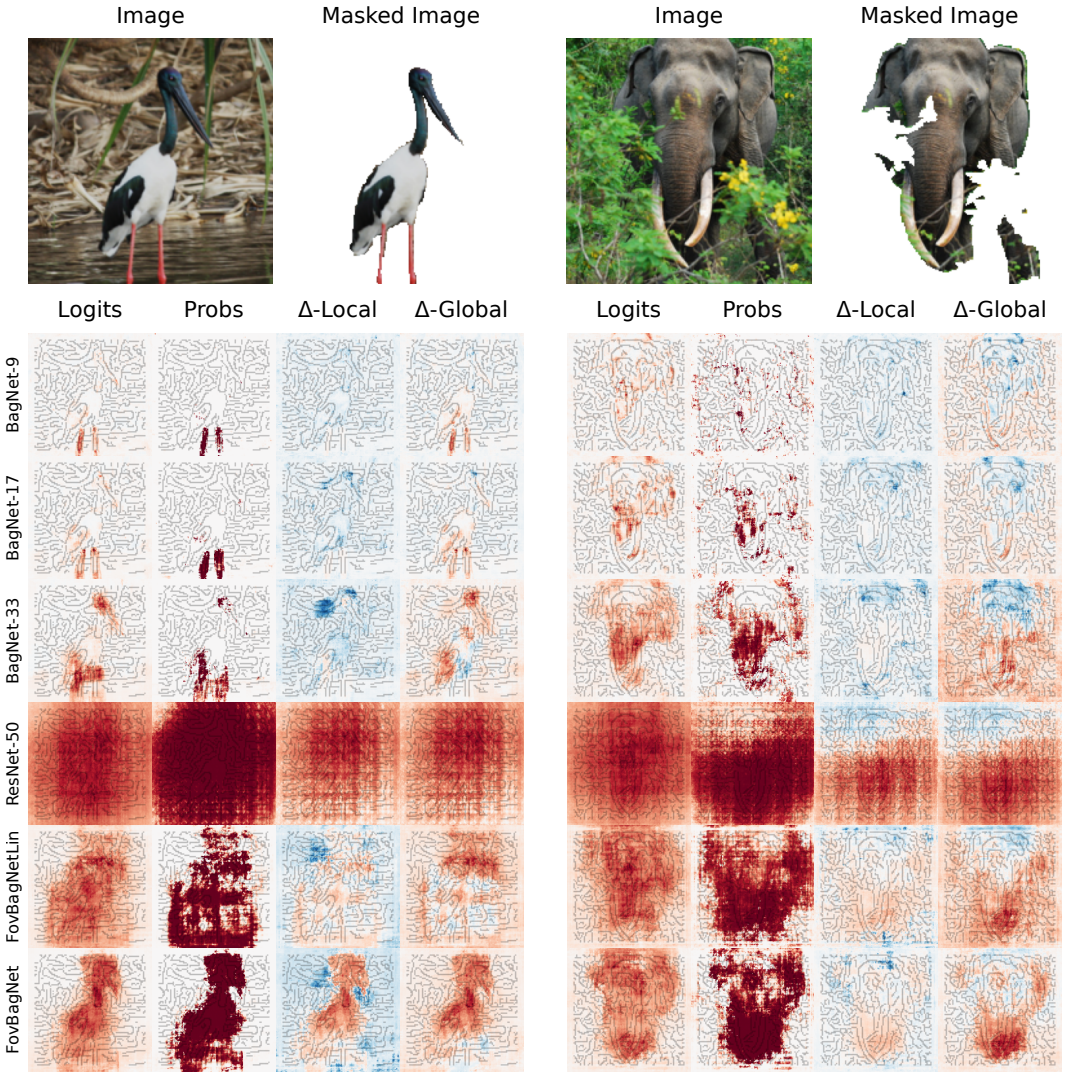

Figure 4: Global context allows FovBagNet to capture objects more holistically compared to the BagNet models. Blurry periphery restricts the activation maps to align with the object.

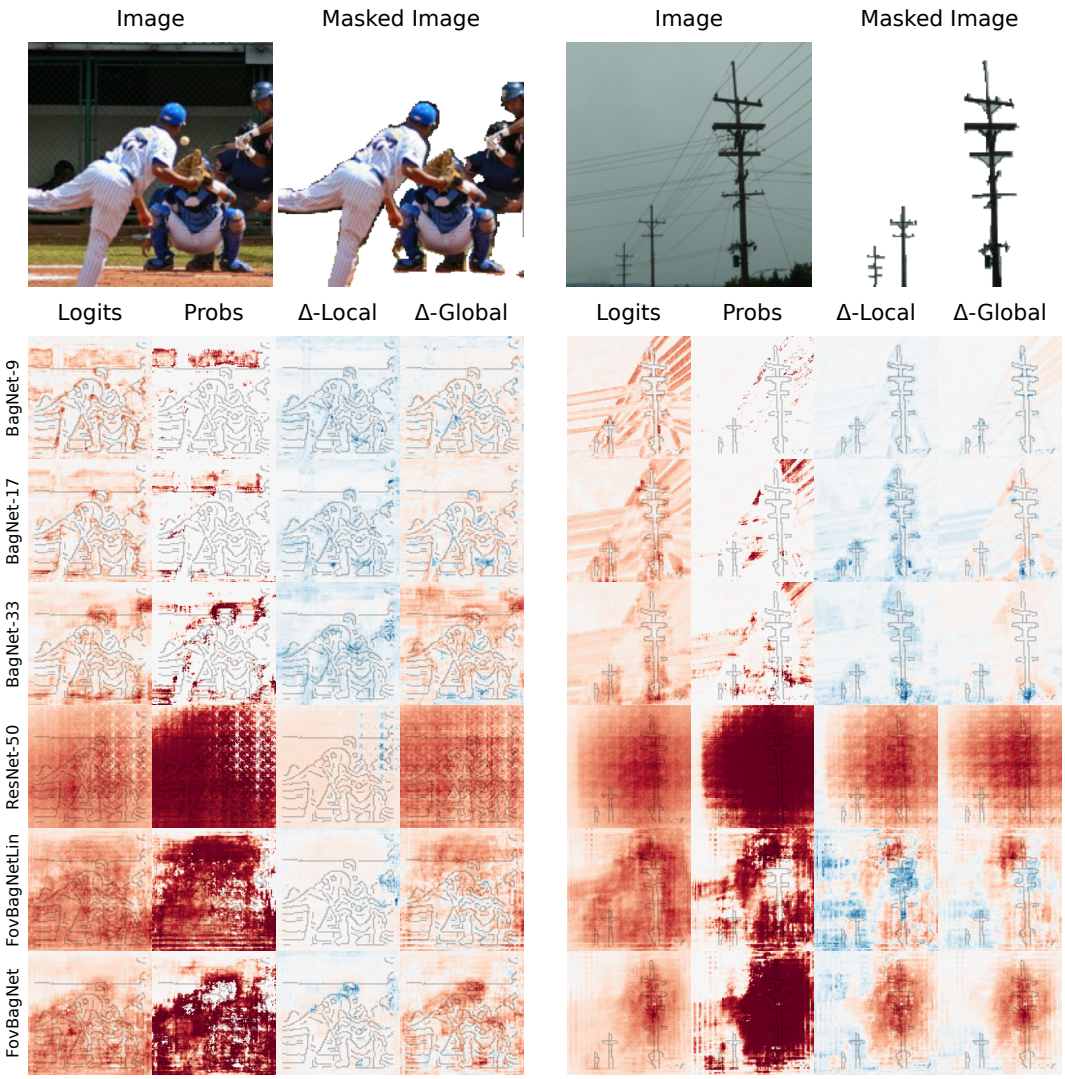

Figure 5: Examples of shortcut exploitation in classes with characteristic backgrounds. Standard BagNet models rely on local background features rather than the target object itself. In the left image, the fence pattern appears to correlate with baseball scenes in ImageNet, while in the right image, overhead cables frequently co-occur with the "pole" (electricity pylon) class. In contrast, FovBagNet mitigates this behavior by incorporating foveation and broader contextual information.

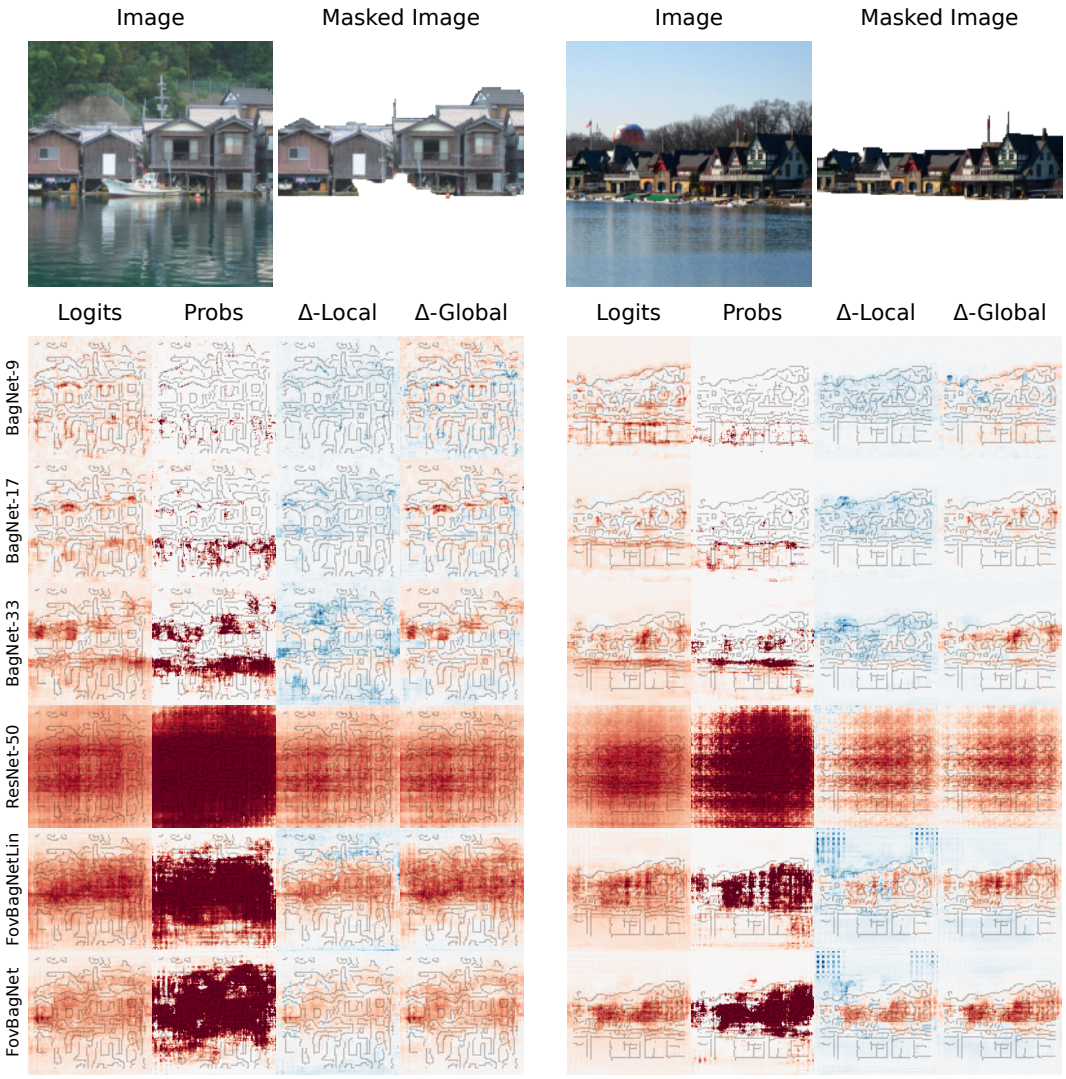

Figure 6: Shortcut exploitation in the "boathouse" class. Standard BagNet models rely primarily on reflections of the boathouse in the water rather than on the boathouse itself, suggesting a shortcut based on local co-occurrence statistics. One possible explanation is that recognizing the "boathouse" class requires jointly reasoning about house and water features; without sufficient global context, the model captures this association only through the reflected pattern. In contrast, FovBagNet resolves this behavior by incorporating foveation and broader contextual information.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions in scope and strength. (1) The claim of improved classification performance is substantiated in section 4, where FovBagNet achieves a top-1 ImageNet accuracy of 0.74, closing most of the gap between BagNet-33 (0.65) and ResNet-50 (0.76), without claiming to match ResNet-50. (2) The claim of sharing a similar but relaxed inherent interpretability property with BagNet is grounded in the architectural choice described in section 3: Each scale level uses its own BagNet backbone, preserving patch-based attribution per level. (3) Saliency quality is evaluated in section 4, using spatial alignment metrics (accuracy, precision, recall, ROC/PR-AUC) against ImageNet-S object masks, and we claim comparability to BagNet rather than superiority. The identified conceptual limitation of logit-based saliency maps and the proposed normalized attribution alternatives are presented in section 3 and evaluated empirically in section 3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our work discusses limitations in section 5. Beyond those, we note that (1) training and evaluation are limited to ImageNet and ImageNet-S, leaving generalization to other datasets unexplored, and (2) our hyperparameter ablation study (subsection C.2) uses a single run per configuration, so we do not report variance estimates there. Computational resources are documented in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: This work is not a theory paper and contains no formal proofs or theorems. However, some claims—such as the argument about logit-based saliency being inherently floored—are argued through intuition rather than formal proof. We believe these are sufficiently supported by the accompanying empirical evidence.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All training and evaluation details necessary to reproduce our results are provided in the Appendix, including model architecture, hyperparameters, training procedure, and evaluation protocol. In addition, we release our full code, along with a lockfile (`uv.lock`) pinning exact package versions, to ensure an identical software environment at reproduction time.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The research code repository is publicly available on `https://github.com/hheidrich/FovBagNet`, including setup, training, and evaluation instructions sufficient to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All training and test details necessary to understand the results are provided in the Appendix A, including model architecture, hyperparameters, optimizer settings, data splits, and evaluation protocol.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The standard deviations for all table entries appearing in the main paper are reported in the appendix. These standard deviations were computed from three models that were trained using different random seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details on compute resources used for training and evaluation, including hardware specifications and runtime estimates, are provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: To the best of our knowledge, our work follows the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Societal impacts are discussed in section 5. Our work proposes an interpretable image classification model, which is broadly a beneficial direction—improved interpretability supports transparency and accountability in ML systems. Potential negative impacts are limited, as this is foundational research without a direct path to harmful applications. We do not identify fairness, privacy, or security concerns specific to this work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets used in this work, ImageNet and ImageNet-S, are properly cited. The original BagNet implementation is cited and credited. Standard libraries (e.g., PyTorch, PyTorch Lightning) are referenced. All assets are used in accordance with their respective terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release our code publicly at `https://github.com/hheidrich/FovBagNet` under the MIT License. The repository includes a README documenting installation (via `uv`), training, checkpoint resumption, and evaluation on both ImageNet-S masks and ImageNet bounding boxes, with expected directory structures for all required data.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.