Leveraging Visual Attention for out-of-distribution Detection

Luca Cultrera Lorenzo Seidenari Alberto Del Bimbo
University of Florence, MICC
name.surname@unifi.it

Abstract

Out-of-Distribution (OOD) detection is a crucial challenge in computer vision, especially when deploying machine learning models in the real world. In this paper, we propose a novel OOD detection method leveraging Visual Attention Heatmaps from a Vision Transformer (ViT) classifier. Our approach involves training a Convolutional Autoencoder to reconstruct attention heatmaps produced by a ViT classifier, enabling accurate image reconstruction and effective OOD detection. Moreover, our method does not require additional labels during training, ensuring efficiency and ease of implementation. We validate our approach on a standard OOD benchmark using CIFAR10 and CIFAR100. To test OOD in a real-world setting we also collected a novel dataset: WildCapture. Our new dataset comprises more than 60k wild animal shots, from 15 different wildlife species, taken via phototraps in varying lighting conditions. The dataset is fully annotated with animal bounding boxes and species.

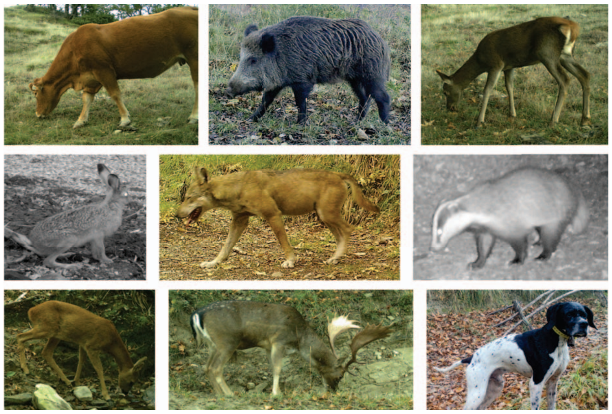


Figure 1: WildCapture: Annotated Wildlife Image Collection. Examples from the proposed dataset showcasing different species and lighting conditions. Top to bottom: Row 1 - Domestic Cattle, Wild Boar, Red Deer; Row 2 - European Hare, Grey Wolf, Eurasian Badger; Row 3 - European Roe Deer, Persian Fallow Deer, Domestic Dog

1. Introduction

Understanding the reliability of machine learning models is paramount when such models are deployed for real-world tasks. One of the main issues of deep learning based classifiers, which is due to the softmax operator, is that they tend to output high scores even for random inputs[37, 19]. Unfortunately, this behavior hinders the reliability of neural network based systems. A classical use case, is to obtain feedback for users regarding the decision of an automatic recognition system. This feature allows to efficiently involve humans in the loop and improve results on samples for which classifiers have weak understanding. We study a specific use case in our work, automatic wildlife recognition[1]. The monitoring of natural environments requires system to be deployed in the wild. In some situations we may not have enough data to train models for all species present in an environment, or simply decide to monitor only certain species (e.g. mammals only). For this reason it is important to be able to estimate when a class has never been observed at training time and, in general, when visual recognition reliability is low. This allows the experts involved to double check results, but only on a small subset of the processed data, making it feasible and sustainable even for large and long-term monitoring efforts. These issues can be effectively addressed through out-of-distribution (OOD) detection[9]. Such algorithms should be as lightweight as possible not requiring too much additional computation on top of possibly already demanding visual classifiers. OOD should be available as a plug-in component for existing classifiers and require as little supervision as possible, with the best case scenario only relying on the very same training data classifiers have been fine-tuned on.

In recent years, attention-based models, such as Vision Transformers (ViT) [14], have achieved remarkable success in various computer vision tasks [22]. However, effectively leveraging attention mechanisms for OOD detection remains an open question.

In this paper, we present a novel and powerful approach for OOD detection that harnesses the power of visual attention heatmaps extracted from a Vision Transformer classifier. Our method centers on training a Convolutional Autoencoder (CAE) to reconstruct the attention heatmaps produced by the ViT classifier.

By utilizing attention heatmaps as the training data for the CAE, our model learns to encode meaningful and salient features of the input images, enabling accurate image reconstruction. We then utilize the error reconstruction mechanism as a powerful signal for OOD detection.

One of the key contributions of this work is the introduction of the WildCapture dataset, a comprehensive collection of images capturing 15 different animal species in real-world scenarios using CameraTrap technology. Each image in the dataset is carefully annotated by experts, including bounding box annotations for each animal, making it ideal for OOD detection research.

Notably, the WildCapture dataset provides a unique challenge due to the presence of visually similar animal classes, resulting in ambiguous decision boundaries for classifiers. In literature there are some datasets that tackle wildlife recognition, such as: Snapshot Serengeti dataset [43]: A vast camera trap dataset with over 3 million images from 225 locations in the African savanna. Annotations based on time thresholds may result in multiple frames receiving the same species label, making it less suitable for controlled experiments. [3] presents a dataset and benchmark designed to measure recognition generalization to novel environments, focusing on camera trap images capturing wild animal populations. The dataset contains images from 20 camera trap locations, with annotations for animal species. However, the uniqueness of our dataset lies in its focus on European species, setting it apart from other existing datasets that mainly cover non-European fauna. Furthermore, our dataset is carefully curated to exclude outlier classes, such as humans, cars, and blank images, ensuring a more targeted and relevant evaluation of recognition algorithms for wildlife environments. This distinctive combination makes our dataset a valuable resource for studying and addressing the challenges of recognition and OOD detection in wildlife environments.

To address the inherent complexities of OOD detection in wild animal classification, our proposed solution harnesses the distinct advantages of Vision Transformers and autoencoder-based techniques. The Vision Transformer captures fine-grained features and global context through multi-head attention, enabling more robust representations to distinguish visually similar classes. Additionally, the autoencoder facilitates the learning of intricate features specific to each class, leading to enhanced reconstruction performance and more accurate OOD detection.

We extensively evaluate our proposed method on the WildCapture dataset, by splitting it into in-distribution (ID) and OOD subsets, showcasing its superior performance compared to conventional approaches. Moreover, we compare our method with a deterministic baseline and demonstrate its effectiveness in detecting OOD samples in various real-world scenarios.

Futhermore, we conduct comprehensive evaluations of our proposed method also on widely used benchmark datasets, such as CIFAR-10 [25] and CIFAR-100 [25]. We adopt the approach suggested by [47] to differentiate between near-OOD and far-OOD tasks, where the difficulty of OOD detection is contingent upon the semantic proximity of the outliers to the inlier classes. Specifically, near-OOD tasks entail detecting outliers that bear a high degree of similarity to the in-distribution classes, making them inherently more challenging. On the other hand, far-OOD tasks involve distinguishing outliers that are more dissimilar to the in-distribution classes, thus being comparatively easier to detect. By conducting experiments on both near-OOD and far-OOD scenarios, we gain deeper insights into the robustness and generalization capabilities of our proposed approach in differentiating between OOD and in-distribution samples in varying degrees of difficulty. For near-OOD tasks, we set CIFAR-10 as the in-distribution dataset and CIFAR-100 as the OOD dataset, as well as vice versa, where CIFAR-100 serves as the in-distribution dataset and CIFAR-10 as the OOD dataset. These near-OOD tasks pose a greater challenge due to the close semantic similarity between the outliers and in-distribution classes.

In contrast, for the far-OOD tasks, we designate CIFAR-10 and CIFAR-100 as the in-distribution datasets, while using the SVHN [36] dataset as the OOD dataset.

In summary, this paper proposes a novel and efficient OOD detection method based on Visual Attention Heatmaps extracted from ViT classifiers. The introduction of the WildCapture dataset enriches the field of OOD research by providing a challenging real-world dataset for evaluation. Our experimental evaluations demonstrate the superiority of our approach in handling OOD samples and improving overall classification performance, making it a promising solution for various computer vision tasks where accurate OOD detection is crucial.

The dataset, along with the pre-trained models and code for our experiments, will be accessible at https://github.com/lcultrera/WildCapture.

2. Related Works

In this section, we present an overview of various approaches used for OOD detection and discuss recent advancements in attention mechanisms and autoencoder-based techniques.

Out-of-distribution detection In recent years, OOD detection has garnered significant interest in the computer vision community due to its critical role in ensuring the ro-

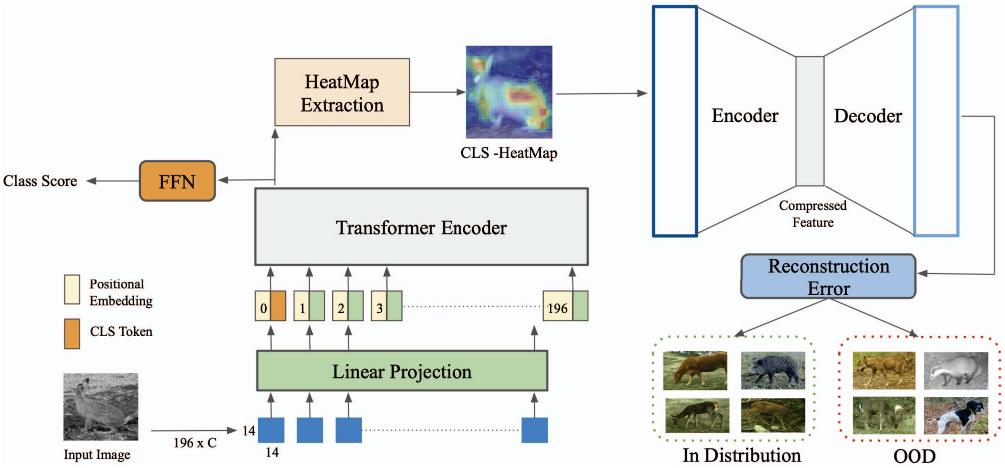


Figure 2: Architecture of our approach: a ViT encoder is used to classify samples, without additional processing, we feed the CLS token heatmap to our Convolutional Autoencoder, then the reconstruction error can be used to perform OOD detection.

bustness and reliability of machine learning models. Several approaches have been proposed to tackle this challenging problem [48]. One common strategy involves using uncertainty estimation methods, such as Bayesian modeling [31, 35], to measure the uncertainty of predictions and identify OOD samples. Some baseline techniques in this category include: Maximum over Softmax Probabilities (MSP) [19] in which the authors propose to use the maximum softmax probability as the confidence score, Mahalanobis distance [27], Outliers Exposure [20], which leverage on a large set of known outliers, and Monte Carlo Markov Chain [2] that allow sampling from high-dimensional distributions. Other methods using posterior approximations are: Monte Carlo DropOut [16], a dropout-based technique that applies dropout during inference, Stochastic Weight Averaging Gradient descent (SWAG) [32], by sampling different weight paths SWAG estimates the uncertainty in model predictions, Laplacian Approximation [39] which leverages the Laplace approximation to estimate the uncertainty in model predictions. Ensemble models [26] are widely recognized as effective tools for improving the robustness and performance of machine learning systems, including OOD detection. In the context of OOD detection, ensemble methods involve combining multiple base models to make predictions, and they can fall under both the probabilistic and uncertainty-based categories. [29] propose a method for OOD detection that uses temperature scaling and input perturbations to enhance model sensitivity to out-of-distribution samples. [6] propose to use deep hybrid models for OOD. [44, 30] propose to impose ”distance-preserving” constraints on the model, with the goal of enhancing its performance in out-of-distribution (OOD) detection tasks, using Jacobian Penality [17] or Spectral Normalization (SN) [33]. Both models suggest employing ”distance aware” output layers, which utilize RBF kernels and Gaussian processes. These layers enable the models to capture and understand the relationships and dissimilarities between different data points, aiding in effective OOD detection.

While some previous studies demonstrate that supervised methods can partially mitigate the issue of incorrectly high-confidence predictions on OOD inputs [13, 29], they still suffer from the limitations of relying on labeled OOD data for training. Our proposed unsupervised method, on the other hand, overcomes this drawback by utilizing attention heatmaps and autoencoder-based image reconstruction, effectively detecting OOD samples without the need for additional labels. This makes our approach more practical and applicable to real-world scenarios where labeled OOD data may be scarce or unavailable.

Some common unsupervised approaches include Density Estimation: [23, 7, 34, 40]. Other methods leverage anomaly detection techniques [45], such as autoencoders [38], to learn a compact representation of the input data and detect outliers. Recent studies confirm that using augumentation, adversarial perturbation [8, 41, 49] helps in OOD detection task. One of the key strengths of our proposed OOD detection method is that it does not rely on data augmentation, adversarial learning, or any prior knowledge about the out-of-distribution samples. While other approaches may require extensive augmentation techniques or the knowledge of specific OOD characteristics, our method solely relies on the autoencoder-based error reconstruction mechanism, making it simpler and more practical to deploy. This advantage not only simplifies the implementation process but also eliminates the risk of introducing bias or artifacts associated with augmentation or adversarial learning.

Attention Mechanism in Computer Vision: Attention based approaches have gained significant traction in com-

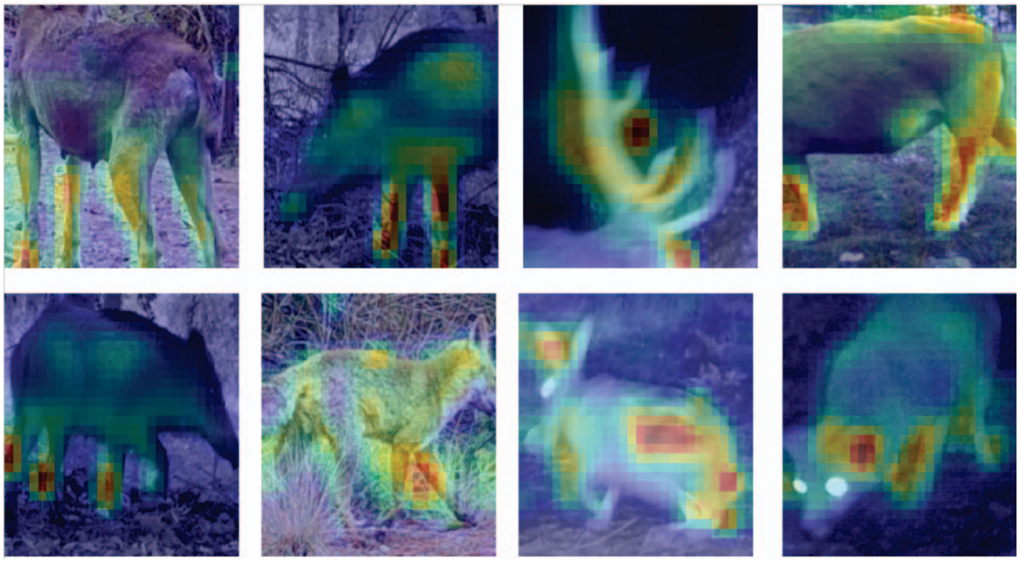


Figure 3: Attention heatmaps examples overlaid on source samples. Interestingly, the visual classifier attend to legs and antlers.

puter vision tasks due to their ability to focus on relevant features and regions within an image.

Attention mechanisms have been widely adopted in various computer vision tasks, including image classification, object detection, and segmentation [18]. They empower models to concentrate on specific regions of an image, enhancing their understanding of the content. In the context of OOD detection, these mechanisms become instrumental in identifying the image regions that are most likely indicative of an OOD sample. Vision Transformers [14] have proven the efficacy of self-attention combined with large-scale pretraining for vision tasks. Their remarkable capability to efficiently adapt to smaller downstream tasks and generalize even in few-shot scenarios makes them an appealing choice for tasks like out-of-distribution (OOD) detection. Indeed, [15] propose an ensemble of ViT models to perform OOD detection. [24] use a transformer-based architecture modelling the OOD task as an object-attribute-based semantic representation learning.

AutoEncoder In OOD Autoencoders have proven to be effective in various domains, including image denoising, dimensionality reduction, and anomaly detection[28], making them a valuable tool for tasks like out-of-distribution (OOD) detection . [11, 21] use supervised autoencoders for uncertainty estimation in OOD scenarios.

[10] combine the visual attention heatmaps with convolutional autoencoders to retrive anomalies in autonomous driving tasks.

In conclusion our proposed method for out-of-distribution (OOD) detection is superior due to several key advantages. We achieve excellent performance without the need for additional labels or data during training, making it efficient and easy to implement. Unlike other approaches, we do not rely on data augmentation or complex training techniques, simplifying the process while maintaining effectiveness. Additionally, our model's ability to handle visually similar classes, coupled with the power of Visual Attention Heatmaps and Convolutional Autoencoders, further enhances OOD detection accuracy. These features make our method a robust and practical solution for OOD detection in various computer vision tasks.

3. Method

In this, we present the details of our proposed out-of-distribution (OOD) detection model, which can be summarized in four key steps:

- Train the Vision Transformer Classifier: We begin by training a state-of-the-art Vision Transformer classifier using large-scale pre-training.
- Extract Visual Attention Heatmaps: From the trained ViT classifier, we extract Visual Attention Heatmaps, highlighting the most relevant regions within each input image. These heatmaps serve as valuable guides for focusing on critical areas during the OOD detection process.
- Convolutional Autoencoder Training: We proceed to train a Convolutional Autoencoder using the extracted attention heatmaps as training data. The autoencoder learns to encode the meaningful and distinctive representations of the attention maps, facilitating precise image reconstruction.

- Image Reconstruction Error as Discriminatory Feature for OOD Detection: The core of our OOD detection model lies in the image reconstruction process. By comparing the reconstructed attention heatmaps with the original ones, we can effectively identify OOD samples based on their deviations from the learned in-distribution patterns.

Figure 2 shows a high level view of our approach.

3.1. ViT Backbone

The Vision Transformer is a state-of-the-art approach for various tasks, including classification [14]. Unlike a traditional convolutional approach, ViT relies on a Multi-Head architecture. Specifically, an image is divided into a sequence of patches, which are linearly projected and fed into an Encoder [14]. The core of the Encoder is the Multi-Head Attention. This Multi-Head Attention enables the model to capture global dependencies and contextual information, allowing the system to model both long-range interactions and small details present in the image.

In this work, we leverage ViT's strengths to train a classifier, exploiting its ability to learn from patch-level features and capture intricate relationships among different parts of an image. Interestingly, attention heatmaps, after fine-tuning, encode a semantic representation of input samples. We exploit this rich and at the same time light representation of input images to learn a representation for OOD detection. Figure 3 showcases examples of attention heatmaps generated by the proposed approach.

To perform classification, we fine-tune a pre-trained ViT [46] on ImageNet21k [12]. The pre-training procedure adheres to the guidelines outlined in [42], ensuring consistency with the suggested approach. The model takes input images of size 224×224 and divides them into patches of size 16×16. This way, each image is split into a grid of 14×14 patches, resulting in a total of 196 patches. We use the CrossEntropy as Loss function.

The results of our experiments and evaluations are presented in section 5.

3.2. Using Visual Attention to Train an Autoencoder

The attention map provided by Vision Transformer can be highly beneficial in discriminating between different species in wild animal classification. The attention map is a visual representation that highlights the regions in the image that the model considers most relevant for making its predictions. It allows us to gain insights into what parts of the image the ViT focuses on when making classification decisions.

In the context of wild animal classification, where species might exhibit visual similarities, the attention map can serve as a valuable tool to understand how the model distinguishes between different animals. By analyzing the attention map, we can identify the key features or distinctive patterns that the model relies on to make accurate classifications. Furthermore, using visual attention map can lead to improved model interpretability and explainability.

According to this, after training the Vision Transformer classifier, we proceed to extract the Visual Attention Heatmaps for each image in both the training and test sets. To facilitate efficient storage and analysis, we resize the attention heatmaps to a standardized size of $128 \times 128 \times 1$.

We train a Convolutional Autoencoder for the task of out-of-distribution (OOD) detection, leveraging visual attention extracted from a pre-trained Vision Transformer. The Convolutional AutoEncoder is designed to reconstruct input images, then we use the reconstruction error to generate precision-recall curves for OOD detection. The architecture, as in [5], consists of an encoder and decoder, each comprising several convolutional layers, with Leaky ReLU activation functions to introduce a regularization effect. The encoder takes grayscale input images of size $128 \times 128 \times 1$ and progressively reduces the spatial dimensions while increasing the number of channels. It culminates in a bottleneck layer of size $512 \times 1 \times 1$. The decoder then upscales and progressively reconstructs the original input image through transposed convolutions and activations. Details about the model are shown in figure 2.

During the training process, the model is optimized to minimize the Mean Squared Error (MSE) loss between the reconstructed heatmaps and the original input.

3.3. Training details

In this section, we provide a comprehensive overview of the training details for both the Classifier and the Convolutional Autoencoder models.

Vision Transformer classifier We finetuned the proposed classifier using a pretrained Vision Transformer model on the ImageNet-21K dataset. The model was initialized with a patch size of 16x16, and the input images were resized to $224 \times 224 \times 3$ during training. We conducted the finetuning process for 50 epochs, utilizing the Cross Entropy loss function to optimize the model's performance. To optimize the model's parameters, we employed the Adam optimizer with an initial learning rate of 0.0001. Additionally, we incorporated a learning rate scheduler to dynamically adjust the learning rate during training. Specifically, we employed the $StepLR$ scheduler with a step$-$size of 7 epochs and a multiplicative factor of 0.1. This setup allowed us to gradually reduce the learning rate every 7 epochs by multiplying it with the specified gamma factor, which effectively aided in stabilizing and enhancing the convergence of the model during the fine-tuning process.

Convolutional Auto-Encoder During Convolutional Autoencoder training, we utilized grayscale visual attention heatmaps extracted from the Vision Transformer, as explained in Section 3.2, with an input size of $128 \times 128 \times 1$. The Autoencoder architecture comprises encoder and decoder blocks, where specific details for each layer can be found in Table 1. To regularize each convolutional layer, we employed Leaky ReLU activation with a negative slope of 0.2. For optimization, we utilized the Adam optimizer with an initial learning rate of 0.0001. To enhance convergence stability and overall model performance, we implemented a linear learning rate scheduler. During the first 40 epochs, the learning rate halved every 10 epochs, after which it remained constant. This schedule ensured efficient training while preserving the fine-tuned model's performance. The Autoencoder's primary objective during training was to minimize the Mean Squared Error (MSE) loss between the reconstructed output and the input images. This training setup empowered the Autoencoder to learn meaningful representations of the input data, facilitating precise image reconstruction and substantially contributing to the subsequent out-of-distribution Detection process.

Encoder					
Layer	Input Shape	Output Shape	Kernel	Stride	Padding
Conv2d (1 → 32)	3×128×128	32×64×64	4×4	2	1
Conv2d (32 → 32)	32×64×64	32×32×32	4×4	2	1
Conv2d (32 → 32)	32×32×32	32×32×32	3×3	1	1
Conv2d (32 → 64)	32×32×32	64×16×16	4×4	2	1
Conv2d (64 → 64)	64×16×16	64×16×16	3×3	1	1
Conv2d (64 → 128)	64×16×16	128×8×8	4×4	2	1
Conv2d (128 → 64)	128×8×8	64×8×8	3×3	1	1
Conv2d (64 → 32)	64×8×8	32×8×8	3×3	1	1
Conv2d (32 → 512)	32×8×8	512×1×1	8×8	1	0
Decoder					
ConvTranspose2d (512 → 32)	512×1×1	32×8×8	8×8	1	0
Conv2d (32 → 64)	32×8×8	64×8×8	3×3	1	1
Conv2d (64 → 128)	64×8×8	128×8×8	3×3	1	1
ConvTranspose2d (128 → 64)	128×8×8	64×16×16	4×4	2	1
Conv2d (64 → 64)	64×16×16	64×16×16	3×3	1	1
ConvTranspose2d (64 → 32)	64×16×16	32×32×32	4×4	2	1
Conv2d (32 → 32)	32×32×32	32×32×32	3×3	1	1
ConvTranspose2d (32 → 32)	32×32×32	32×64×64	4×4	2	1
ConvTranspose2d (32 → 1)	32×64×64	1×128×128	4×4	2	1

Table 1: Convolutional Autoencoder layers details

4. WildCapture: Annotated Wildlife Image Collection

The dataset used in this study represents a valuable resource for the task of animal species classification in a real-world scenario. Comprising more than $60k$ images from 15 distinct classes, each representing different animal species. These species include: Beech Marten, Crested Porcupine, Domestic Cattle, Domestic Dog, Domestic Horse, Eurasian Badger, European Hare, European Roe Deer, Grey Wolf, Persian Fallow Deer, Red Deer, Red Fox, Western Polecat, Wild Boar, and Wild Cat. The dataset was collected using CameraTrap technology, providing authentic and diverse wildlife imagery. To ensure accurate annotations, each image underwent handcrafted annotation by domain experts, guaranteeing high-quality ground truth labels. Moreover, it encompasses a diverse range of lighting conditions, including both daylight and nocturnal images. In Figure 1, we present a selection of examples from our dataset, highlighting various species and lighting conditions. This broad spectrum of lighting scenarios allows for a more thorough evaluation of model generalization and adaptability in real-world scenarios. In figure 4 is shown the dataset class distribution.

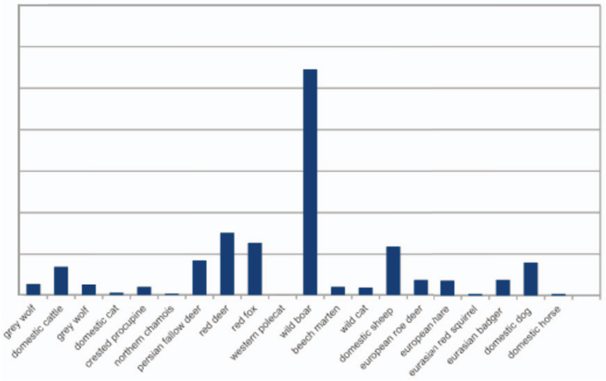


Figure 4: Class distribution

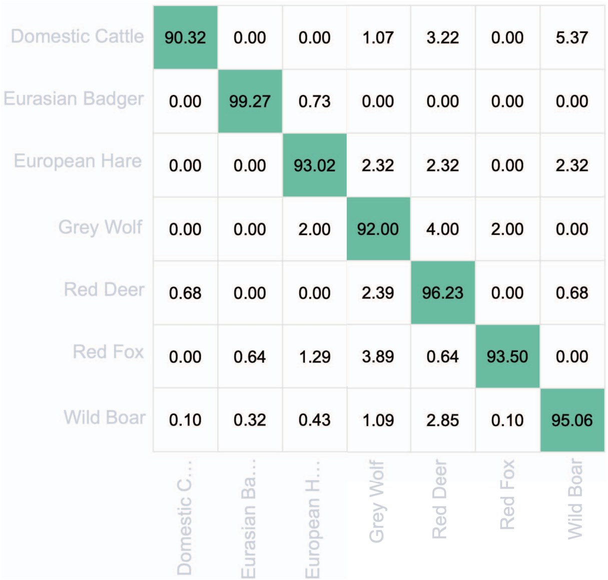
To enhance the dataset's utility for both classification and out-of-distribution (OOD) detection, we employed the MegaDetector framework [3] to annotate bounding boxes (bbox) around each animal depicted in the images. This step facilitated the cropping of individual animal instances during the training process, leading to improved model performance in recognizing fine-grained details. This dataset allowed us to explore the effectiveness of our OOD detection approach, as we randomly split the dataset to create distinct OOD classes not encountered during the training phase.

By leveraging this dataset, we provide a robust evaluation platform for our novel approach. The dataset's comprehensive representation of various animal species, combined with expert annotations and bbox information, ensures the reliability and realism of our experimental results, highlighting the potential of our OOD detection technique in real-world wildlife applications.

5. Experiments

Firstly, we present the results of the Vision Transformer classifier on the proposed WildCapture Dataset, achieving a validation accuracy of 94.30% on the test set. To provide a comprehensive view of the classifier's performance, we present a detailed confusion matrix (Figure 5), highlighting correct and misclassified predictions across different classes.

The high validation accuracy and visually informative



	Domestic C...	Eurasian Ba...	European H...	Grey Wolf	Red Deer	Red Fox	Wild Boar
Domestic Cattle	90.32	0.00	0.00	1.07	3.22	0.00	5.37
Eurasian Badger	0.00	99.27	0.73	0.00	0.00	0.00	0.00
European Hare	0.00	0.00	93.02	2.32	2.32	0.00	2.32
Grey Wolf	0.00	0.00	2.00	92.00	4.00	2.00	0.00
Red Deer	0.68	0.00	0.00	2.39	96.23	0.00	0.68
Red Fox	0.00	0.64	1.29	3.89	0.64	93.50	0.00
Wild Boar	0.10	0.32	0.43	1.09	2.85	0.10	95.06

Figure 5: Confusion matrix from the ViT classifier on WildCapture Datset.

confusion matrix underscore the VIT classifier's robustness and generalization capabilities in handling complex and diverse visual data. The model's ability to discern fine-grained differences between various wild animal species further solidifies its effectiveness in real-world scenarios.

As we delve deeper into the experimental results, we proceed to evaluate the performance of our proposed OOD detection process.

First of all we compare our proposed model against a deterministic baseline and an ensemble baseline (Deep Ensembles with 2 independent DNNs) with MC dropout . The deterministic baseline is implemented using a ViT classifier, and we apply the softmax function to the probability scores for each class. The threshold for distinguishing between in-distribution and out-of-distribution samples is set as follows:

$$\text{Threshold} = 1 - \max(\text{Softmax}(\text{prob})) \qquad (1)$$

In this baseline approach, any sample with a maximum softmax probability score below the threshold is classified as out-of-distribution, while samples above the threshold are considered in-distribution.

To demonstrate the effectiveness of the proposed method we rely on several datasets. As a real-world scenario we use WildCapture. We pick randomly split of classes to obtain in-distribution and out-of-distribution sets. We use the in-distribution set to train the Vision Transformer as described in section 3.1 and the AE. Then we use the out-of-distribution split as a test set.

In order to prove the efficacy of our method we use also the Caltech CameraTrap dataset [4] as out-of-distribution set. We avoid overlap between classes in our WildCapture in-distribution set and Caltech CameraTrap.

The results of this experiment are summarized in Table 2, which clearly showcases the superiority of our method in detecting out-of-distribution samples compared to the baselines and alternative approaches. In fact our method outperform the baselines in both AUPR and AUROC metrics.

Method	**ID: WildCapture**	**OOD:WildCapture**		**OOD: CCT[4]**	
	Accuracy	**AUROC**	**AUPR**	**AUROC**	**AUPR**
Deterministic	94.30	57.44	61.46	57.43	68.06
Ensemble	81.89	41.25	90.94	53.36	84.65
Ours	**94.30**	**92.63**	**92.25**	**99.29**	**97.17**

Table 2: Results on WildCapture as in-distribution dataset

In table 3 and in table 4 we compare our method with some state-of-the-art methods using respectively CIFAR10 and CIFAR100 as in-distribution sets. In both experiments we use respectively CIFAR100 and CIFAR10 as near OOD task and SVHN as far OOD task. The proposed approach achieves state-of-the-art performance, demonstrating remarkable accuracy and outperforming existing techniques in accurately detecting out-of-distribution samples with 100% AUPR and AUROC in both benchmarks. This compelling result underscores the efficacy and versatility of our method in handling diverse and challenging datasets, making it a promising solution for out-of-distribution detection tasks.

Method	**ID: CIFAR10**	**OOD:CIFAR100**		**OOD:SVHN**	
	Accuracy	**AUROC**	**AUPR**	**AUROC**	**AUPR**
DUQ [44]	95.50	90.80	88.80	97.20	96.90
SNPG [30]	96.00	91.60	91.10	97.80	97.50
Vit Ensemble [15]	**98.70**	98.52	98.70	98.58	99.82
DHM [6]	96.30	**100.00**	**100.00**	**100.00**	**100.00**
Ours	97.80	**100.00**	**100.00**	**100.00**	**100.00**

Table 3: Results on Cifar10 as in-distribution dataset

Method	**ID: CIFAR100**	**OOD:CIFAR10**		**OOD:SVHN**	
	Accuracy	**AUROC**	**AUPR**	**AUROC**	**AUPR**
DUQ [44]	79.90	83.90	87.20	89.70	90.80
SNPG [30]	80.50	86.30	87.50	92.80	93.50
Vit Ensemble [15]	**91.71**	96.23	96.32	97.80	98.87
DHM [6]	81.30	**100.00**	**100.00**	**100.00**	**100.00**
Ours	89.80	**100.00**	**100.00**	**100.00**	**100.00**

Table 4: Results on Cifar100 as in-distribution dataset

6. Ablation Study

On the Importance of Fine-Tuning In this ablation study, we investigate the effectiveness of pre-training and fine-tuning approaches on the proposed out-of-distribution detection mechanism. We compare pre-trained classifiers'

OOD detection capabilities without fine-tuning to those with fine-tuning on the in-distribution dataset. AUROC and AUPR metrics (as used in Section 5) assess the models' OOD detection performance. Our results demonstrate the crucial role of fine-tuning of the ViT classifier in enhancing the quality of attention heat-maps generated for the autoencoder, revealing the importance of adapting the models to the specific in-distribution data (Table 5).

The fine-tuning process guides the classifier to learn more discriminative and accurate representations, resulting in more meaningful attention heat-maps used for image reconstruction. The considerable difference in AUPR and AUROC scores between the fine-tuned model (92.25% and 92.63%) and the model without fine-tuning (62.50% and 47.82%) highlights the significance of fine-tuning for generating more informative attention heat-maps. By doing so, the overall OOD detection performance of the autoencoder significantly improves without the need for additional labeled data or complex modifications to the model architecture.

Method	**ID: WildCapture**	**OOD:WildCapture**	
	Accuracy	**AUROC**	**AUPR**
Not FineTuned	14.35	47.82	62.50
FineTuned	**94.30**	**92.63**	**92.25**

Table 5: Results on WildCapture as in-distribution dataset with and w/o fine-tuning

On the Importance of Pre-Training Table 6 presents an investigation into the impact of pre-training strategies on the performance of Vision Transformer models for out-of-distribution detection on the CIFAR100 dataset. We explore two scenarios: fine-tuning a ViT model pretrained on ImageNet-21K and training a ViT model from scratch.

In the first scenario (Table 5), we fine-tune the ViT model on our in-distribution dataset to enhance its sensitivity to class-specific features. Fine-tuning aims to produce precise and informative attention heat-maps, improving the autoencoder's ability to accurately reconstruct input images.

For the second scenario, we train the ViT model from scratch on our in-distribution dataset, examining its performance without external pre-training.

Remarkably, the model trained from scratch achieves outstanding performance (100% in AUPR and AUROC with SVHN as OOD), indicating its effective learning of class-specific features and discriminative patterns.

However, the fine-tuned model exhibits the best overall performance. Fine-tuning allows the ViT model to adapt and specialize for our OOD detection task, capturing intricate details and nuances present in the data. By leveraging pre-training on ImageNet-21K and refining on our dataset, the fine-tuned model produces attention heat-maps tailored to our OOD detection problem.

In summary, while training the ViT model from scratch shows promising results, fine-tuning the pretrained ViT model on our in-distribution dataset yields the best OOD detection performance. This emphasizes the importance of leveraging pre-training along with fine-tuning for state-of-the-art OOD detection. Moreover, attention heat-maps derived from fine-tuned models are more informative for the autoencoder's reconstruction process, leading to improved precision and recall in detecting out-of-distribution samples

Method	**ID: CIFAR10**	**OOD:CIFAR100**		**OOD:SVHN**	
	Accuracy	**AUROC**	**AUPR**	**AUROC**	**AUPR**
From Scratch	89.40	81.86	93.39	**100.00**	**100.00**
Fine-Tuned	**97.80**	**100.00**	**100.00**	**100.00**	**100.00**

Table 6: Results on Cifar10 as in-distribution dataset

7. Conclusion

This paper introduces a novel and powerful approach to address the critical challenge of out-of-distribution detection in computer vision. By leveraging attention-based mechanisms and autoencoder-based techniques, our model captures fine-grained features and class-specific patterns, significantly enhancing OOD detection performance.

A notable contribution of our work is the introduction of the WildCapture dataset, a comprehensive collection of real-world wildlife images meticulously annotated by experts. This dataset serves as a valuable resource for evaluating and advancing OOD detection in wildlife environments.

In the context of automatic wildlife recognition [1], where systems operate in real-world, wild environments, data collection for all wildlife species can be challenging or limited, leading to potential uncertainties in model predictions. Our approach effectively addresses these challenges by providing accurate out-of-distribution detection, enabling experts to confidently evaluate and validate model predictions for better decision-making.

The experimental evaluations on the WildCapture dataset and widely-used benchmarks, such as CIFAR-10/100, demonstrate the superiority of the proposed model in handling OOD samples. Furthermore, the unsupervised nature of our data-driven approach makes it an appealing solution for situations where obtaining additional labels may prove challenging or even infeasible.

Acknowledgements

This work was supported by the European Commission under European Horizon 2020 Programme, grant number 951911—AI4Media.

This work was supported by Tuscany Region FSC project REWIRE.

References

[1] Jorge A Ahumada, Eric Fegraus, Tanya Birch, Nicole Flores, Roland Kays, Timothy G O'Brien, Jonathan Palmer, Stephanie Schuttler, Jennifer Y Zhao, Walter Jetz, et al. Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environmental Conservation*, 47(1):1–6, 2020.

[2] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50:5–43, 2003.

[3] S Beery, D Morris, and S Yang. Efficient pipeline for camera trap image review. arxiv. *arXiv preprint arXiv:1907.06772*, 2019.

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[5] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

[6] Senqi Cao and Zhongfei Zhang. Deep hybrid models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4743, 2022.

[7] Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

[8] Sungik Choi and Sae-Young Chung. Novelty detection via blurring. *arXiv preprint arXiv:1911.11943*, 2019.

[9] Peng Cui and Jinjia Wang. Out-of-distribution (ood) detection based on deep learning: A review. *Electronics*, 11(21):3500, 2022.

[10] Luca Cultrera, Federico Becattini, Lorenzo Seidenari, Pietro Pala, and Alberto Del Bimbo. Explaining autonomous driving with visual attention and end-to-end trainable region proposals. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–13, 2023.

[11] Steve Dias Da Cruz, Bertram Taetz, Thomas Stifter, and Didier Stricker. Autoencoder attractors for uncertainty estimation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2553–2560. IEEE, 2022.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[18] Mohammed Hassanin, Saeed Anwar, Ibrahim Radwan, Fahad S Khan, and Ajmal Mian. Visual attention methods in deep learning: An in-depth survey. *arXiv preprint arXiv:2204.07756*, 2022.

[19] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[20] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[21] Philipp Joppich, Sebastian Dorn, Oliver De Candido, Jakob Knollmüller, and Wolfgang Utschick. Classification and uncertainty quantification of corrupted data using supervised autoencoders. In *Physical Sciences Forum*, volume 5, page 12. MDPI, 2022.

[22] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[23] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[24] Rajat Koner, Poulami Sinhamahapatra, Karsten Roscher, Stephan Günnemann, and Volker Tresp. Oodformer: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976*, 2021.

[25] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[26] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[27] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[28] Pengzhi Li, Yan Pei, and Jianqiang Li. A comprehensive survey on design and application of autoencoder in deep learning. *Applied Soft Computing*, page 110176, 2023.

[29] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[30] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

[31] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[32] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

[33] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[34] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

[35] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[36] Yuval Netzer, Tao Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[37] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[38] Stanislav Pidhorskyi, Ranya Almohsen, Donald A. Adjeroh, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *Neural Information Processing Systems*, 2018.

[39] Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, volume 6. International Conference on Representation Learning, 2018.

[40] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

[41] Gabi Shalev, Yossi Adi, and Joseph Keshet. Out-of-distribution detection using multiple semantic label representations. *Advances in Neural Information Processing Systems*, 31, 2018.

[42] Andreas Steiner, Alexander Kolesnikov, , Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[43] Alexandra Swanson, Margaret Kosmala, Chris J. Lintott, Robert Simpson, Arfon M. Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, 2, 2015.

[44] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020.

[45] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L. Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *European Conference on Computer Vision*, 2018.

[46] Ross Wightman. Pytorch image models. `https://github.com/huggingface/pytorch-image-models`, 2019.

[47] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon A. A. Kohl, taylan. cemgil, S. M. Ali Eslami, and Olaf Ronneberger. Contrastive training for improved out-of-distribution detection. *ArXiv*, abs/2007.05566, 2020.

[48] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

[49] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.