AttZoom: Attention Zoom for Better Visual Features

Daniel DeAlcala, Aythami Morales, Julian Fierrez, Ruben Tolosana
Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain
daniel.dealcala@uam.es aythami.morales@uam.es julian.fierrez@uam.es ruben.tolosana@uam.es

Abstract

We present Attention Zoom, a modular and model-agnostic spatial attention mechanism designed to improve feature extraction in convolutional neural networks (CNNs). Unlike traditional attention approaches that require architecture-specific integration, our method introduces a standalone layer that spatially emphasizes high-importance regions in the input. We evaluated Attention Zoom on multiple CNN backbones using CIFAR-100 and TinyImageNet, showing consistent improvements in Top-1 and Top-5 classification accuracy. Visual analyses using Grad-CAM and spatial warping reveal that our method encourages fine-grained and diverse attention patterns. Our results confirm the effectiveness and generality of the proposed layer for improving CCNs with minimal architectural overhead.

1. Introduction

The attention mechanisms in AI allow models to focus their capacity on the most relevant information, much as the human brain allocates cognitive resources to different aspects of sensory input through attention [2]. This ability is crucial for efficient information processing in many human activities such as learning [11, 40], as not all data are equally important. A key trend in AI has been to improve performance by drawing inspiration from brain function, leading to the incorporation of attention mechanisms into AI models in what is called brain-inspired artificial intelligence (BIAI) [43, 55]. These mechanisms help systems filter out irrelevant details and focus on critical aspects, improving tasks such as pattern recognition and decision-making [33].

Deep convolutional neural networks (CNNs) have become the backbone of many state-of-the-art image recognition systems [31]. However, standard CNNs process all spatial regions in an input image uniformly, which can lead to suboptimal performance when only certain regions are relevant for a given task [52]. This is common in many image-based recognition problems, e.g., in biometrics for person recognition [20, 47–49]. To address this non-uniformity in the relevance of different parts of the data, some researchers have developed weight-based fusion schemes [15, 16], combining the different pieces of information with fixed or adaptive weights [18, 19]. In a different line of action, attention mechanisms have been proposed to guide the focus of trained networks toward the most informative parts of the image [38].

In the literature, the development of attention mechanisms has played a crucial role in computer vision. Most attention methods rely on dedicated attention blocks integrated throughout the network architecture, leading to models such as [23, 28, 36, 52, 53]. In contrast, our approach introduces a standalone attention layer that helps the model focus on different regions of the image. Its independent nature makes it easy to implement and apply, while allowing seamless integration into any model architecture. This wrapping of the existing model is inspired by our previous work [39], where new layers were introduced not to emphasize certain regions, but to remove certain sensitive data.

We evaluated our approach by inserting the proposed attention layer into several popular CNN backbones and performing classification experiments on CIFAR-100 [32] and TinyImageNet [8] datasets. The results show consistent improvements across all models tested.

Our contributions are threefold:

- We propose an attention mechanism that enhances feature extraction from images, easily applicable without architectural changes to subsequent processing via existing neural models and new ones.
- We integrate the attention module into multiple CNN backbones without architecture-specific modifications.
- We demonstrate the effectiveness of our method through extensive experiments on benchmark datasets.

The remainder of this paper is organized as follows. Section 2 reviews related work on attention mechanisms. Section 3 introduces our proposed attention method, AttZoom, including its mathematical formulation. Section 4 describes the integration process into existing architectures, along with pseudocode. Section 5 outlines our experimental setup, covering datasets, models, and training details. The results are presented in Section 6, followed by the conclusions in Section 7.

2. Related Works

Attention mechanisms have been extensively studied in the literature, especially in recent years, leading to significant advances in AI. These methods can be categorized in various ways. In this work, we focus on the mechanisms of attention applied to images and classify them into three main groups: channel attention, spatial attention, and self-attention, following the taxonomy presented in [22, 23].

2.1. Channel Attention

Channel attention mechanisms learn to recalibrate the importance of each channel within a convolutional feature map. In convolutional networks, each channel represents a specific set of features detected by different filters. Channel attention enables the network to selectively emphasize or suppress certain channels based on the image context, improving the representation of relevant features [17].

One of the most fundamental works in channel attention is Squeeze & Excitation (SE) [28]. The SE block compresses each channel into a single value using global average pooling (Squeeze) and then adjusts its importance through two fully connected layers (Excitation): one with ReLU and another with Sigmoid. Several works have aimed to improve SE efficiency, such as Efficient Channel Attention (ECA) [51], which replaces fully connected layers with 1D convolutions to eliminate the need for dimensionality reduction. Similarly, [56] applies L2 normalization per channel, scales with a learnable vector, uses tanh normalization [19] for attention, and integrates the output with the input via a residual connection. Another notable work is ResNeSt [57], a ResNet variant that incorporates split-attention blocks, leveraging global pooling, convolutions, and softmax to adjust channel importance. Furthermore, CBAM [52] and BAM [41] combine channel and spatial attention by extracting descriptors through maximum and average pooling, processing them with a multilayer perceptron, and applying a sigmoid activation to generate an attention map.

A wide range of other studies have explored channel attention. For example, [9] employs covariance matrices decomposed into eigenvalues to construct its attention mechanism, while [42] uses global average pooling in the frequency domain. For a more comprehensive overview of channel attention methods, we refer readers to extensive surveys such as [4, 22, 23].

2.2. Spatial Attention

Spatial attention methods highlight specific regions within an image by assigning weights to different spatial locations in the feature map. Unlike channel attention, which modulates entire feature channels, spatial attention operates directly on the spatial distribution of features. Our approach falls into this category as it emphasizes the importance of particular image regions to enhance recognition performance.

CBAM [52] combines channel and spatial attention by taking advantage of both average-pooled and max-pooled features to generate a 2D spatial attention map. The Recurrent Attention Model (RAM) [38] employs recurrent neural networks (RNNs) and reinforcement learning (RL) to train the model in selecting relevant regions for attention. Several other works also build spatial attention mechanisms based on RNNs [3, 21, 54]. Co-attention & Co-excitation [27] utilizes non-local networks to capture long-range dependencies.

A significant portion of spatial attention research is based on pyramids [29, 34, 58], where attention maps are generated from features extracted at different levels of the model. For a more detailed discussion of spatial attention methods, we refer the reader to comprehensive surveys such as [4, 22, 23].

2.3. Self Attention

Self attention enables each position in a data sequence (image, text, etc.) to interact with all other positions, assigning weights based on their relevance. This allows for more effective long-range dependency modeling compared to traditional convolutions, which have a more limited receptive field. Self-attention is the core mechanism behind Transformers, which have not only revolutionized NLP [7, 10, 14, 37, 50] but also achieved significant advances in computer vision [5, 6, 44, 59] and pattern recognition on temporal functions [12, 13].

Self attention can be traced back to the LSTMs [25], where an attention vector was generated at each hidden state to indicate which parts of the sequence should interact [46]. However, Self-Attention gained prominence due to Transformers [50]. Transformers follow an encoder-decoder architecture, where the encoder applies a self-attention mechanism to capture global dependencies in the input, followed by a feedforward layer. The decoder has a similar structure but incorporates a cross-attention layer, allowing it to focus on the relevant information from the encoder. Numerous Transformer variants have been developed for specialized tasks. For example, Swin Transformers [35] introduce sliding window attention and hierarchical structures to enhance performance in computer vision tasks.

3. Proposal: Attention Zoom (AttZoom)

In this work, we introduce our method, Attention Zoom, a spatial attention mechanism that incorporates Attention-Zoom layers. Intuitively, these layers help the model focus on the most important regions of the image by amplifying relevant information for classification while suppressing less informative areas that the network deems irrelevant. In

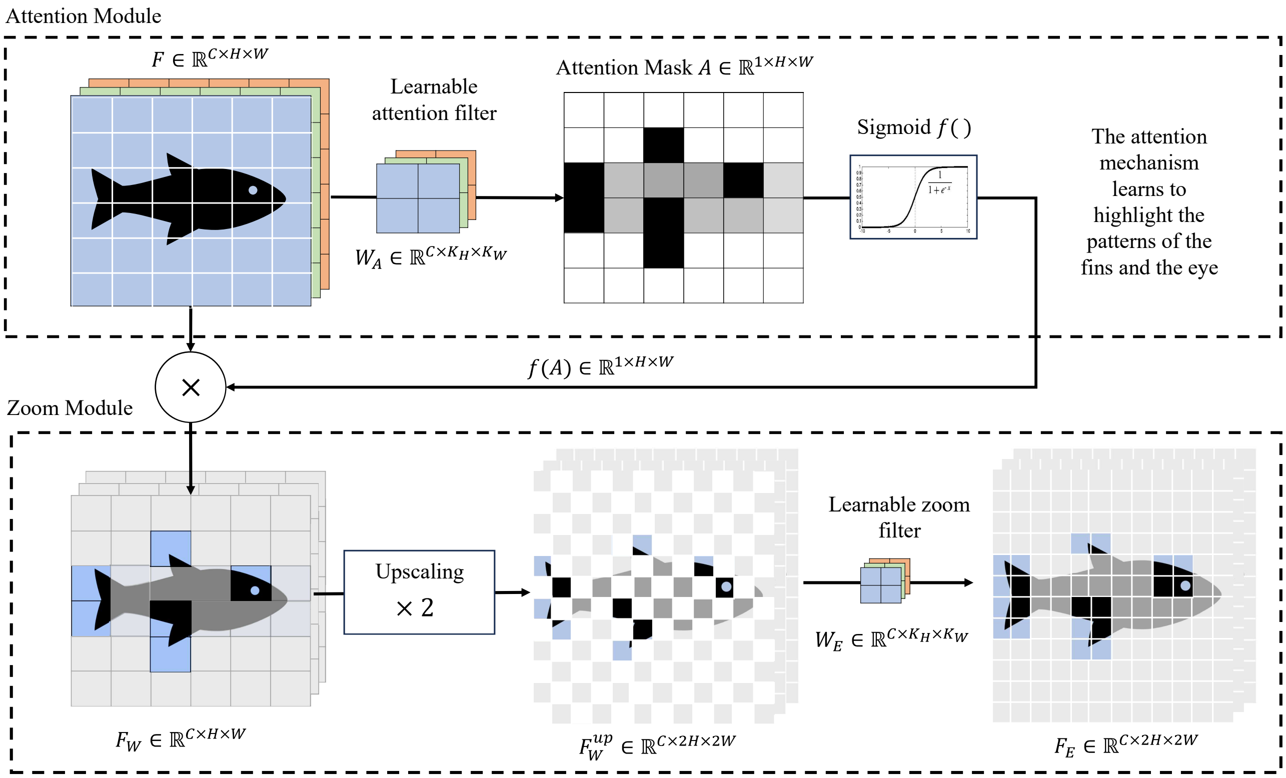

Figure 1. Block diagram of the proposed attention mechanism. Note that this is a conceptual representation, and the intermediate feature F does not necessarily correspond to an image, but rather to a learned representation. In this example, the most relevant features are located around the fins and the eye. The attention module learns to mask the most relevant regions, while the zoom module enhances the patterns within these regions.

Fig. 1, we show the block diagram of the proposed attention/enhancement mechanism.

We start with an intermediate feature map $F \in \mathbb{R}^{C\times H\times W}$. The first step involves learning a Spatial Attention Map $A \in \mathbb{R}^{1\times H\times W}$. To obtain this Spatial Attention Map, the model learns a convolutional kernel that processes the intermediate feature map F to identify the most relevant regions. We denote this convolutional kernel by $W_A \in \mathbb{R}^{C\times K_H\times K_W}$. By convolving this kernel with the feature map, we obtain our attention map:

$$A = W_A * F$$

To preserve spatial dimensions $H \times W$, we use a stride of 1 and zero padding. Once the Attention Map is obtained, we apply the following function:

$$f(A) = \begin{cases} 1, & \text{if } \sigma(A) \geq \text{threshold} \\ \sigma(A), & \text{if } \sigma(A) < \text{threshold} \end{cases}$$

$$\sigma(A) = \frac{1}{1+e^{-A}}$$

This step is crucial because it allows the network to preserve the regions it considers important while gradually suppressing irrelevant areas. We obtain the weighted feature map by performing an element-wise multiplication $\odot$ between the input feature map (each channel independently) and the learned attention map:

$$F_W = F \odot f(A), \quad \text{with } f(A) \in \mathbb{R}^{1\times H\times W}$$

After the weighted feature map F_W we proceed to the 'zoom' operation. This step amplifies the information in the regions deemed important by the attention map while diluting or reducing the details in less relevant areas.

To achieve this, we apply two transformations. The first involves inserting a zero into every spatial position of F_W effectively increasing the resolution. This produces the Upsampled Weighted Feature Map:

$$F_W^{Up} = \text{Upsample}(F_W) \in \mathbb{R}^{C\times H'\times W'}$$

where the new dimensions are defined as:

$$H' = 2H, W' = 2W$$

(Note that, without loss of generality, the multiplier introduced here 2 can be increased for a higher zoom.)

This upsampling operation is illustrated in Figure 1. Mathematically, it can be expressed as follows:

$$F_W^{Up}(c, 2h, 2w) = F_W(c, h, w), \quad F_W^{Up}(c, 2h{+}1, w) = 0,$$

$$F_W^{Up}(c, h, 2w + 1) = 0, \quad F_W^{Up}(c, 2h + 1, 2w + 1) = 0$$

for $c = 0, \dots, C-1; h = 0, \dots, H-1; w = 0, \dots, W-1$.

We then apply a convolution, obtaining the Enhanced Feature Map:

$$F_E = W_E * F_W^{Up}$$

This convolution modulates the information density, either amplifying or suppressing features depending on their importance as determined by the attention map.

- In regions where $f(A)$ reduced values (i.e., where the attention map multiplied the pixels to a value below a threshold), inserting zeros and applying a convolution further diminishes this information, pushing values closer to zero.
- In contrast, in regions where $f(A)$ kept values unchanged (ie, where A had values above a threshold), inserting zeros followed by convolution performs a local upsampling. This effectively expands the feature representation in those areas, increasing the amount of retained information and enlarging its receptive field.

This process ensures that important features are emphasized while irrelevant regions fade out, allowing the model to focus on the most discriminative parts of the image.

4. Method Integration

Before proceeding with our experiments, we highlight several key aspects of the Attention-Zoom layers and how they differ from existing attention-based methods.

Most attention mechanisms introduce an attention block at multiple points within a model, effectively defining a new architecture (e.g., ResNext [53], SE [28], CBAM [52]). These architectures are typically modifications of a backbone such as ResNet [24], incorporating attention-based enhancements to improve performance. Other approaches leverage attention blocks to design entirely new architectures, as seen in Transformer-based models [50].

In contrast, our work introduces the Attention-Zoom layer, which expands the information and increases the receptive field in certain image regions while reducing the significance of others. Unlike traditional attention modules, Attention-Zoom is not structured as a block and does not modify the network architecture. Instead, it is a modular layer that can be seamlessly integrated at any stage of a model.

This design offers several advantages. First, it maintains the original network architecture while adding only a few extra parameters. Second, it is highly adaptable and can be incorporated into any model without requiring structural modifications (simply by adding the layer in the early stages). Moreover, it can be combined with other attention-based approaches such as ResNext [53], SE [28], CBAM [52], etc. The pseudocode for the Attention-Zoom layer is presented in Algorithm 1.

Algorithm 1 Attention-Zoom Layer

Input: Feature Map $F \in \mathbb{R}^{C\times H\times W}$
Output: Enhanced Feature Map $F_E \in \mathbb{R}^{C\times H'\times W'}$
Step 1: Compute Spatial Attention Map
$A = \sigma(W_A * F)$ *(Conv + Sigmoid)*
Step 2: Apply Attention Mask
$A = \mathbf{1}(A \geq threshold) + A \cdot \mathbf{1}(A < threshold)$
$F_W = F \odot A$ *(Element-wise multiplication)*
Step 3: Apply Upsampling with Zeros
$F_W^{Up} = \text{UpsampleZeros}(F_W)$
Step 4: Apply Enhancement Convolution
$F_E = W_E * F_W^{Up}$ *(Final convolution)*
Return F_E

5. Experimental Protocol

In this work, we integrate the Attention-Zoom layers into various architectures trained on different datasets under a well-defined experimental protocol that ensures fair comparisons.

5.1. Data and Models

We evaluated our method on two widely used benchmark datasets. Specifically, we consider CIFAR-100 [32] and Tiny-ImageNet [8], as they provide a suitable balance between complexity and scalability for image classification tasks.

We conducted experiments on two categories of architectures. First, we evaluate Attention-Zoom layers on standard state-of-the-art models widely used in the literature, including ResNet-50 [24], DenseNet-121 [30] and MobileNet [26].

Second, we evaluate the integration of our method into architectures that already incorporate attention mechanisms, specifically Squeeze-and-Excitation (SE) [28], Convolutional Block Attention Module (CBAM) [52], and ResNeXt [53]. As these attention mechanisms have been integrated into ResNet-50, they provide particularly meaningful comparisons for our study.

For each architecture, we follow a two-step evaluation:

(1) training the model without Attention-Zoom layers, and (2) retraining it with Attention-Zoom layers, which allows us to directly compare their impact on performance.

5.2. Collaboration, Not Competition

The primary goal of our study is to evaluate whether the inclusion of Attention-Zoom layers effectively enhances model performance. As discussed in Sect. 4, a key advantage of our method is its modularity: it consists of a single layer that can be seamlessly integrated into any model.

Therefore, while direct comparisons with attention-based methods are relevant, our approach extends beyond competition. Instead, we demonstrate how Attention-Zoom layers can be incorporated into existing attention-based architectures to further improve their performance. This highlights the flexibility and complementary nature of our method.

5.3. Model Training

To ensure a rigorous and fair evaluation, we adopt a robust hyperparameter optimization strategy. Specifically, we utilize Optuna [1], a Bayesian optimization framework, to perform a hyperparameter search.

For each result reported in Sec. 6, we conduct an Optuna Bayesian search over 30 trials. This means that each model undergoes 30 training runs with different hyperparameter configurations, following a guided search strategy. The best configuration is then selected for comparison.

This approach ensures that our reported results are not biased by stochastic factors such as random hyperparameter selection. Instead, they reflect the best achievable performance for each model, making our comparisons statistically meaningful.

The following parameters are optimized by Bayesian search:

- **Batch Size:** 32, 64, 128, or 256.
- **Learning Rate:** Sampled logarithmically between 10^{-5} and 10^{-1}.
- **Weight Decay:** Sampled logarithmically between 10^{-6} and 10^{-2}.
- **Optimizer:** Adam or SGD. For SGD, we tuned the momentum between 0.7 and 1 in linear scale.
- **Scheduler:** One of CosineAnnealingLR, StepLR, ReduceLROnPlateau, or OneCycleLR. For StepLR, we also tune the step size between 10 and 30 (linear scale).

Models are trained from scratch for 50 epochs with early stopping (patience of 5 epochs). The data augmentation strategy includes RandomCrop, RandomHorizontalFlip, and AutoAugment.

While more complex augmentation, optimization strategies, or pre-trained models could be employed, our setup is sufficiently robust to achieve competitive results under realistic training conditions. We intentionally avoid overly specialized training techniques or pre-trained models, as the goal of this work is not to extract marginal gains for each model but to establish a strong and fair experimental framework for comparing models trained with the Attention-Zoom Layer.

6. Results

We first report in Table 1 the Top-1 and Top-5 accuracy scores for the models presented in Sec. 5.1, all trained on CIFAR-100. In addition to the results, we include the best hyperparameters identified by Optuna [1] using Bayesian search to achieve these performances. We observe that our method consistently improves the performance of all evaluated models. Although some architectures such as DenseNet121 and SE ResNet exhibit modest gains of less than 1%, others such as MobileNet and ResNet50 achieve improvements of around 10%. The remaining models fall somewhere in between.

Table 2 reports again the best results obtained through Bayesian optimization, this time on the Tiny-ImageNet dataset. The trends are consistent with those observed in the previous table: some models benefit substantially more from the introduction of the Attention-Zoom layers than others. In particular, SE ResNet and ResNeXt show comparable performance with and without our layer. However, models such as ResNet50 and MobileNet exhibit more significant performance improvements.

An interesting observation arises when comparing models that already incorporate attention mechanisms with those that do not. For models with existing spatial attention modules (last 6 rows in Tables 1 2), the performance gains tend to be smaller. However, our method still yields improvements, reaching 5% in some cases. These results, along with the ease of implementation of our method (which consists of a single, easily integrable layer), suggest that this module can benefit a wide range of models, regardless of whether they already include attention mechanisms, but with even greater benefits observed in those that do not.

Given the simplicity of implementing our method, the goal is not to compete with existing spatial attention mechanisms, but rather to complement them (Sec. 5.2). However, since the three evaluated attention-based architectures (S&E ResNet, CBAM ResNet, and ResNeXt) are all built on the ResNet backbone, we compare these architectures (without our AttZoom) to a standard ResNet with AttZoom. This allows us to evaluate whether simply adding our layer to a standard ResNet can approach the performance of more complex architectures like the ones mentioned above.

Our results show that simply adding our AttZoom layer brings the performance of standard ResNet close to—or even above—that of more complex attention-based models, as observed in cases like CBAM on TinyImageNet and CIFAR-100, and SE on TinyImagenet.

Architecture	Top-1 Acc (%)	Top-5 Acc (%)	Hyper-params
ResNet50	58.50	87.75	128, 1^{-3}, 3^{-4}, Adam, OneCycleLR
ResNet50 +AttZoom	67.66	90.15	128, 7^{-4}, 1^{-4}, Adam, OneCycleLR
MobileNet	63.62	88.61	64, 3^{-3}, 2^{-5}, Adam, OneCycleLR
MobileNet +AttZoom	76.43	95.01	256, 7^{-3}, 2^{-5}, Adam, OneCycleLR
DenseNet121	75.43	93.69	64, 3^{-4}, 1^{-6}, Adam, CosineAnnealingLR
DenseNet121 + AttZoom	75.82	94.41	128, 2^{-4}, 4^{-6}, Adam, CosineAnnealingLR
Attention Architecture	**Top-1 Acc (%)**	**Top-5 Acc (%)**	**Hyper-params**
SE ResNet	76.95	94.19	64, 4^{-4}, 2^{-5}, Adam, CosineAnnealingLR
SE ResNet + AttZoom	77.06	94.38	64, 2^{-3}, 6^{-4}, SGD(0.99), CosineAnnealingLR
CBAM ResNet	59.72	85.58	64, 5^{-4}, 2^{-4}, Adam, OneCycleLR
CBAM ResNet + AttZoom	64.32	87.70	128, 5^{-4}, 4^{-5}, Adam, StepLR(29)
ResNext	72.83	92.59	64, 7^{-4}, 1^{-5}, Adam, ReduceLROnPlateau
ResNext + AttZoom	77.53	95.34	64, 1^{-4}, 4^{-3}, Adam, CosineAnnealingLR

Table 1. Top-1 and Top-5 accuracy on CIFAR-100 for all models introduced in Sec. 5.1, with and without AttZoom layers. The table also reports the best hyperparameters found via Bayesian optimization using Optuna [1].

Architecture	Top-1 Acc (%)	Top-5 Acc (%)	Hyper-params
ResNet50	48.58	74.33	64, 2^{-3}, 8^{-4}, SGD(0.96), ReduceLROnPlateau
ResNet50 +AttZoom	63.23	84.12	32, 5^{-2}, 2^{-4}, SGD(0.9), CosineAnnealingLR
MobileNet	55.26	78.26	64, 1^{-2}, 1^{-6}, Adam, OneCycleLR
MobileNet +AttZoom	64.89	85.30	64, 1−1, 1−4, SGD(0.8), CosineAnnealingLR
DenseNet121	64.19	85.44	32, 6^{-5}, 2^{-4}, Adam, StepLR(30)
DenseNet121 + AttZoom	66.29	86.45	64, 1−3, 1−3, SGD(0.9), CosineAnnealingLR
Attention Architecture	**Top-1 Acc (%)**	**Top-5 Acc (%)**	**Hyper-params**
SE ResNet	65.92	86.01	64, 6^{-4}, 2^{-4}, Adam, CosineAnnealingLR
SE ResNet + AttZoom	66.06	86.36	64, 1^{-3}, 5^{-2}, Adam, CosineAnnealingLR
CBAM ResNet	52.38	76.82	64, 5^{-3}, 6^{-4}, SGD(0.9), StepLR(30)
CBAM ResNet + AttZoom	57.88	81.08	128, 2^{-2}, 3^{-3}, SGD(0.7), CosineAnnealingLR
ResNext	63.20	84.08	64, 2^{-5}, 9^{-5}, Adam, StepLR(30)
ResNext + AttZoom	63.77	84.23	32, 2^{-3}, 2^{-5}, SGD(0.9), CosineAnnealingLR

Table 2. Top-1 and Top-5 accuracy on Tiny-ImageNet for all models, with and without AttZoom layers. We also report the best hyperparameters found via Bayesian optimization using Optuna [1].

6.1. Interpretability: GradCam Visualization

To qualitatively analyze our Attention Zoom Layers, we employ the well-known Grad-CAM method [45]. Grad-CAM uses gradients to compute the importance of spatial locations within convolutional layers. This allows us to visualize and interpret the differences in attention between models.

In this section, we present Grad-CAM maps under different conditions to better understand the effect of AttZoom:

- We compare models with and without AttZoom layers. To avoid overloading the paper with visualizations, we focus on two models that show a clear performance gain with AttZoom (e.g., ResNet50 and ResNet50 CBAM) and two others with more modest improvements (e.g., DenseNet121 and ResNeXt), across different datasets.
- We include two types of scenario: images where both models correctly classify the input (Scenario 1) and images where both models fail (Scenario 2). This avoids cherry-picking favorable examples.

Figure 2 shows the results. Scenario 1, where both models are correct, is the most informative. The key observation is that models with AttZoom layers are more fine-grained and have a diverse collection of details. A clear case is observed with fish images in TinyImageNet: while models without AttZoom focus on a single feature or a broad region, models with AttZoom attend to multiple smaller regions, such as the dorsal and lateral fins and the mouth.

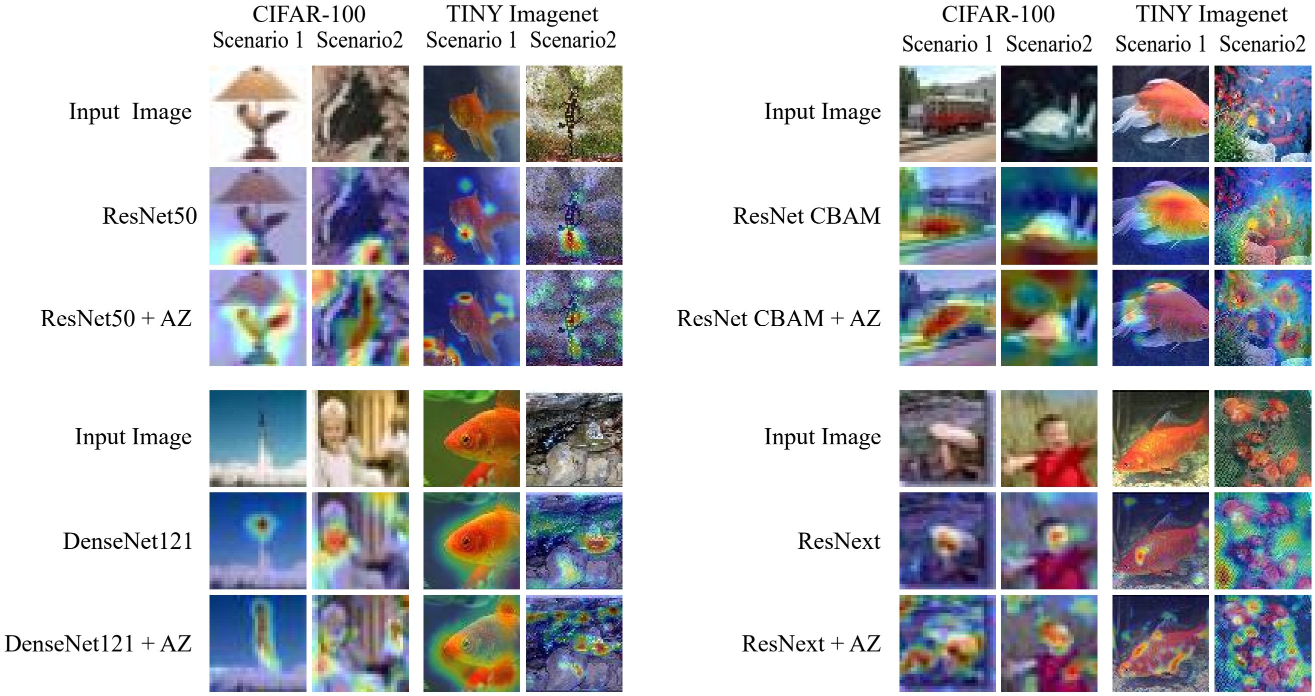

Figure 2. Grad-CAM visualizations for models with and without AttZoom layers. We show examples from four representative architectures (ResNet50, CBAM ResNet, DenseNet121, ResNeXt) on different datasets. Each row corresponds to either Scenario 1 (both models classify correctly) or Scenario 2 (both models fail) and both datasets.

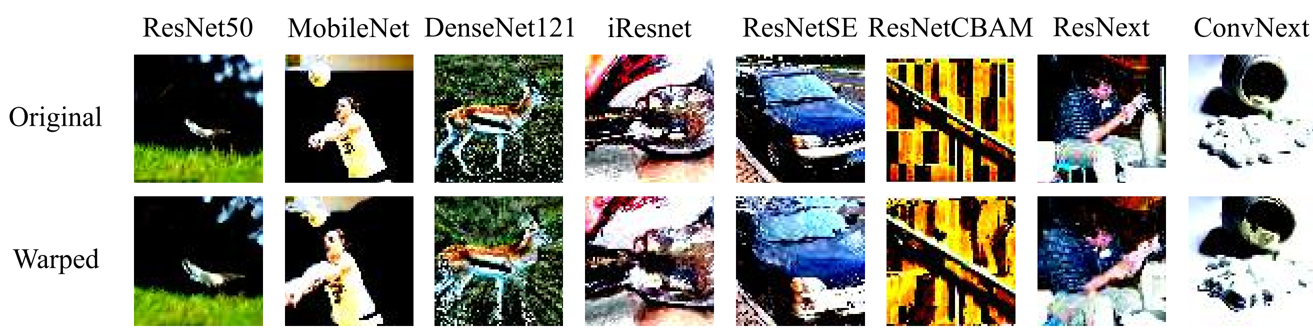

Figure 3. Warped image visualizations based on the attention produced by the AttZoom Layer. We show one correctly classified image per model, using the Tiny ImageNet dataset. The warping expands regions with high attention and contracts those with low attention.

In Scenario 2, where both models fail, neither is able to focus on the relevant regions or on enough details to produce a correct prediction. However, we still observe that models with our proposed AttZoom tend to explore more distinct areas of the image. This broader and finer attention is what underlies the performance improvements reported in Tables 1 and 2.

6.2. Visual Analysis via Attention-Guided Warping

In this section, we qualitatively illustrate the effect of integrating the AttZoom layer by spatially warping the image based on its attention. Specifically, we identify the regions where the AttZoom Layer focuses its attention and then apply a warping function to the input grid that expands high-attention areas and contracts low-attention ones.

Figure 3 shows some of these warped examples. Here, we select correctly classified images and present one example per each of the 8 models introduced in Sect. 5.1, using Tiny ImageNet due to its higher-resolution samples. In the resulting visualizations, regions that are spatially expanded correspond to areas with strong attention responses, indicating that the model relies on those fine-grained details to make its predictions.

7. Conclusions

In this work, we proposed Attention Zoom (AttZoom): a spatial visual attention mechanism that allows networks to focus on finer image details, thereby enhancing performance. Unlike traditional attention mechanisms that rely on complex architectural changes, our method is modular, architecture-agnostic, and easy to integrate. This simplicity enables seamless incorporation into a wide range of architectures, including those that already employ attention mechanisms.

We demonstrated its effectiveness in various CNN backbones and datasets, showing consistent improvements in classification accuracy. Through both quantitative and qualitative analyses (including Grad-CAM visualizations and attention-based warping), we showed that models equipped with AttZoom attend to more informative and fine-grained regions in the input.

Our results suggest that AttZoom can serve as a plug-and-play module to improve visual attention in existing architectures with minimal parameter overhead.

Future work will explore the application of AttZoom to other image processing and recognition tasks such as object detection and segmentation. Extensions of the AttZoom idea as a wrapper of existing AI models for application to text and temporal functions are also in our plans.

Acknowledgement

This work has been supported by projects BBforTAI (PID2021-127641OB-I00 MICINN/FEDER), HumanCAIC (TED2021-131787B-I00 MICINN), M2RAI (PID2024-160053OB-I00 MICIU/FEDER) and Cátedra ENIA UAM-VERIDAS (NextGenerationEU PRTR TSI-100927-2023-2). Work conducted in the ELLIS Unit Madrid. D. DeAlcala is supported by a FPU Fellowship (FPU21/05785). G. Mancera is supported by a FPI Fellowship (PRE2022-104499).

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019. 5, 6

[2] Charles H Anderson, David C Van Essen, and Bruno A Olshausen. Directed Visual Attention and the Dynamic Control of Information Flow. *Neurobiology of Attention*, pages 11–17, 2005. 1

[3] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple Object Recognition with Visual Attention. *arXiv preprint arXiv:1412.7755*, 2014. 2

[4] Gianni Brauwers and Flavius Frasincar. A General Survey on Attention Mechanisms in Deep Learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3279–3298, 2023. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Proceedings of the European conference on computer vision (ECCV)*, pages 213–229. Springer, 2020. 2

[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative Pretraining From Pixels. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1691–1703. PMLR, 2020. 2

[7] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking Attention with Performers. *arXiv preprint arXiv:2009.14794*, 2020. 2

[8] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A Downsampled Variant of ImageNet as an Alternative to the CIFAR datasets. *arXiv preprint arXiv:1707.08819*, 2017. 1, 4

[9] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-Order Attention Network for Single Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11065–11074, 2019. 2

[10] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, 2019. Association for Computational Linguistics. 2

[11] Roberto Daza, Luis F. Gomez, Julian Fierrez, Aythami Morales, Ruben Tolosana, and Javier Ortega-Garcia. DeepFace-Attention: Multimodal face biometrics for attention estimation with application to e-learning. *IEEE Access*, 12:111343–111359, 2024. 1

[12] Paula Delgado-Santos, Ruben Tolosana, Richard Guest, Ruben Vera-Rodriguez, and Julian Fierrez. M-GaitFormer: Mobile biometric gait verification using transformers. *Engineering Applications of Artificial Intelligence*, 125:106682, 2023. 2

[13] Paula Delgado-Santos, Ruben Tolosana, Richard Guest, Parker Lamb, Andrei Khmelnitsky, Colm Coughlan, and Julian Fierrez. SwipeFormer: Transformers for mobile touchscreen biometrics. *Expert Systems with Applications*, 237:121537, 2024. 2

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1*, pages 4171–4186, 2019. 2

[15] Julian Fierrez. *Adapted Fusion Schemes for Multimodal Biometric Authentication*. PhD thesis, Univ. Politecnica de Madrid, 2006. 1

[16] Julian Fierrez, Aythami Morales, Ruben Vera-Rodriguez, and David Camacho. Multiple classifiers in biometrics. Part

1: Fundamentals and review. *Information Fusion*, 44:57–64, 2018. 1

[17] Julian Fierrez, Aythami Morales, Ruben Vera-Rodriguez, and David Camacho. Multiple classifiers in biometrics. Part 2: Trends and challenges. *Information Fusion*, 44:103–112, 2018. 2

[18] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Adapted user-dependent multimodal biometric authentication exploiting general information. *Pattern Recognition Letters*, 26(16):2628–2639, 2005. 1

[19] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Target dependent score normalization techniques and their application to signature verification. *IEEE Trans. on Systems, Man Cybernetics - Part C*, 35(3):418–425, 2005. Invited Paper. 1, 2

[20] E. Gonzalez-Sosa, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia. Exploring facial regions in unconstrained scenarios: Experience on ICB-RW. *IEEE Intelligent Systems*, 33(3):60–63, 2018. 1

[21] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1462–1471. PMLR, 2015. 2

[22] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8(3):331–368, 2022. 2

[23] Mohammed Hassanin, Saeed Anwar, Ibrahim Radwan, Fahad Shahbaz Khan, and Ajmal Mian. Visual attention methods in deep learning: An in-depth survey. *Information Fusion*, 108:102417, 2024. 1, 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[26] Andrew G Howard. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. In *Proceedings of the IEEE/CVF Confernce on Computer Vision and Pattern Recognition (CVPR)*, pages 21–26, 2017. 4

[27] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[28] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 1, 2, 4

[29] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. SPAN: Spatial Pyramid Attention Network for Image Manipulation Localization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 312–328. Springer, 2020. 2

[30] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 4

[31] Javier Huertas-Tato, Alejandro Martin, Julian Fierrez, and David Camacho. Fusing CNNs and statistical indicators to improve image classification. *Information Fusion*, 79:174–187, 2022. 1

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. *Toronto, ON, Canada*, 2009. 1, 4

[33] Qiuxia Lai, Salman Khan, Yongwei Nie, Hanqiu Sun, Jianbing Shen, and Ling Shao. Understanding More About Human and Machine Attention in Deep Neural Networks. *IEEE Transactions on Multimedia*, 23:2086–2099, 2020. 1

[34] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial Attention Pyramid Network for Unsupervised Domain Adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 481–497. Springer, 2020. 2

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10012–10022, 2021. 2

[36] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 1

[37] Gonzalo Mancera, Aythami Morales, Julian Fierrez, et al. PBa-LLM: Privacy- and bias-aware NLP using named-entity recognition (NER). In *IAPR Intl. Conf. on Document Analysis and Recognition Workshops (ICDARw)*, 2025. 2

[38] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent Models of Visual Attention. *Advances in Neural Information Processing Systems*, 27, 2014. 1, 2

[39] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. SensitiveNets: Learning agnostic representations with application to face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 43(6): 2158–2164, 2021. 1

[40] Miriam Navarro, Álvaro Becerra, Roberto Daza, Ruth Cobos, Aythami Morales, and Julian Fierrez. VAAD: Visual attention analysis dashboard applied to e-learning. In *European Conf. on Computer Vision Workshops (ECCVw)*, 2024. 1

[41] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck Attention Module. *arXiv preprint arXiv:1807.06514*, 2018. 2

[42] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. FcaNet: Frequency Channel Attention Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 783–792, 2021. 2

[43] Jing Ren and Feng Xia. Brain-inspired Artificial Intelligence: A Comprehensive Review. *arXiv preprint arXiv:2408.14811*, 2024. 1

[44] Sergio Romero-Tapiador, Ruben Tolosana, et al. Are vision-language models ready for dietary assessment? exploring the next frontier in AI-powered food image recognition. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRw)*, 2025. 2

[45] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 618–626, 2017. 6

[46] Ruben Tolosana, Paula Delgado-Santos, Andres Perez-Uribe, Ruben Vera-Rodriguez, Julian Fierrez, and Aythami Morales. DeepWriteSYN: On-line handwriting synthesis via deep short-term representations. In *AAAI Conf. on Artificial Intelligence (AAAI)*, pages 600–608, 2021. 2

[47] Ruben Tolosana, Sergio Romero-Tapiador, Ruben Vera-Rodriguez, Ester Gonzalez-Sosa, and Julian Fierrez. Deepfakes detection across generations: Analysis of facial regions, fusion, and performance evaluation. *Engineering Applications of Artificial Intelligence*, 110:104673, 2022. 1

[48] P. Tome, J. Fierrez, R. Vera-Rodriguez, and D. Ramos. Identification using face regions: Application and assessment in forensic scenarios. *Forensic Science International*, (233): 75–83, 2013.

[49] Pedro Tome, Julian Fierrez, Ruben Vera-Rodriguez, and Javier Ortega-Garcia. Combination of face regions in forensic scenarios. *Journal of Forensic Sciences*, 60(4):1046–1051, 2015. 1

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 4

[51] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[52] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 1, 2, 4

[53] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500, 2017. 1, 4

[54] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2048–2057. PMLR, 2015. 2

[55] Shuangming Yang, Xinyu Hao, Bin Deng, Xile Wei, Huiyan Li, and Jiang Wang. A survey of brain-inspired artificial intelligence and its engineering. *Life Research*, 1(1):23–29, 2018. 1

[56] Zongxin Yang, Linchao Zhu, Yu Wu, and Yi Yang. Gated Channel Transformation for Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11794–11803, 2020. 2

[57] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. ResNeSt: Split-Attention Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRw)*, pages 2736–2746, 2022. 2

[58] Ting Zhao and Xiangqian Wu. Pyramid Feature Attention Network for Saliency Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3085–3094, 2019. 2

[59] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 2