computers

MDPI

Article

Semantic-Aware Adaptive Binary Search for Hard-Label Black-Box Attack

Yiqing Ma [1], Kyle Lucke [2], Min Xian [2,*] and Aleksandar Vakanski [2,3]

[1] Huntsman Cancer Institute, University of Utah, Salt Lake City, UT 84112, USA; yiqing.ma@hci.utah.edu
[2] Department of Computer Science, University of Idaho, Idaho Falls, ID 83402, USA; klucke@uidaho.edu (K.L.); vakanski@uidaho.edu (A.V.)
[3] Department of Nuclear Engineering and Industrial Management, University of Idaho, Idaho Falls, ID 83402, USA
* Correspondence: mxian@uidaho.edu

Abstract: Despite the widely reported potential of deep neural networks for automated breast tumor classification and detection, these models are vulnerable to adversarial attacks, which leads to significant performance degradation on different datasets. In this paper, we introduce a novel adversarial attack approach under the decision-based black-box setting, where the attack does not have access to the model parameters, and the returned information from querying the target model consists of only the final class label prediction (i.e., hard-label attack). The proposed attack approach has two major components: adaptive binary search and semantic-aware search. The adaptive binary search utilizes a coarse-to-fine strategy that applies adaptive tolerance values in different searching stages to reduce unnecessary queries. The proposed semantic mask-aware search crops the search space by using breast anatomy, which significantly avoids invalid searches. We validate the proposed approach using a dataset of 3378 breast ultrasound images and compare it with another state-of-the-art method by attacking five deep learning models. The results demonstrate that the proposed approach generates imperceptible adversarial samples at a high success rate (between 99.52% and 100%), and dramatically reduces the average and median queries by 23.96% and 31.79%, respectively, compared with the state-of-the-art approach.

Keywords: adversarial attack; hard-label black-box attack; adaptive binary search; breast ultrasound; semantic-aware search

Citation: Ma, Y.; Lucke, K.; Xian, M.; Vakanski, A. Semantic-Aware Adaptive Binary Search for Hard-Label Black-Box Attack. *Computers* **2024**, *13*, 203. https://doi.org/10.3390/computers13080203

Academic Editors: Robertas Damaševičius and Leandros Maglaras

Received: 21 June 2024
Revised: 9 August 2024
Accepted: 16 August 2024
Published: 18 August 2024

1. Introduction

Breast cancer has emerged as one of the most prevalent types of cancer globally, contributing to nearly 12% of all newly diagnosed cancer cases, and is estimated to affect around 30% of female cancer cases in the U.S. in 2024 [1]. Although Deep Neural Networks (DNNs) demonstrated unprecedented performance in medical image classification, recent research [2,3] has indicated that DNNs as well as conventional machine learning (ML) models can be compromised by adversarial samples. That is, adversarial samples can be synthesized by adding imperceptible perturbations to clean inputs and cause the target DNNs to misclassify such samples. Adversarial attacks [4–6] have been realized to achieve high attack success rates by introducing low levels of perturbations. Adversarial attacks can be categorized into white-box attacks and black-box attacks. In a white-box adversarial setting, attackers are assumed to have complete knowledge of the targeted model, including a knowledge of the model architecture, parameters, gradients, objective function, etc. Similarly, prior works [7] have demonstrated that adversarial samples from one ML model can be transferred to other models in a black-box setting. The black-box setting is more challenging because adversaries do not have access to the model structure or parameters. It is also more realistic since most model developers do not provide such access to users. In both white-box and black-box attacks, adversarial attacks can be categorized as targeted

and untargeted. Targeted attacks fool a model into falsely predicting a specific label for the adversarial image. Untargeted attacks generate an adversarial sample that is classified as some arbitrary incorrect label. Adversarial attacks can also occur under two scenarios: (1) evasion attack and (2) poisoning attack. In an evasion attack, only malicious samples are produced to evade detection and induce a misclassification. In a poisoning attack, the attacker injects fake training data intending to corrupt the training. Poisoning attacks, while potentially more successful, can be harder to accomplish since the attacker must have access to the training dataset before the model is trained. This makes black-box attacks have more practical significance.

Most existing adversarial attacks are designed for evading DNNs designed for the classification of natural image datasets (e.g., CIFAR-10 [8] and ImageNet [9]). However, some modalities of medical images (e.g., breast ultrasound images) possess domain-specific characteristics distinct from natural images. As a result, black-box attacks are more challenging with medical images in terms of the significant number of model queries to achieve a successful attack. As shown in the second row of Figure 1, Rays [10] needs extremely large numbers of queries to generate adversarial images for five breast ultrasound images. To overcome the challenges in existing attacks, we propose a new coarse-to-fine approach to reduce the number of queries in extreme examples using a combination of adaptive binary search and semantic-aware search. The method is motivated by declining the swing search of the decision boundary at an early stage and narrowing down search regions according to the search depth. Furthermore, as shown in five adversarial images (the third row of Figure 1), the proposed method achieves similar peak signal-to-noise ratio (PSNR) scores as RaySS but with significantly fewer numbers of queries.

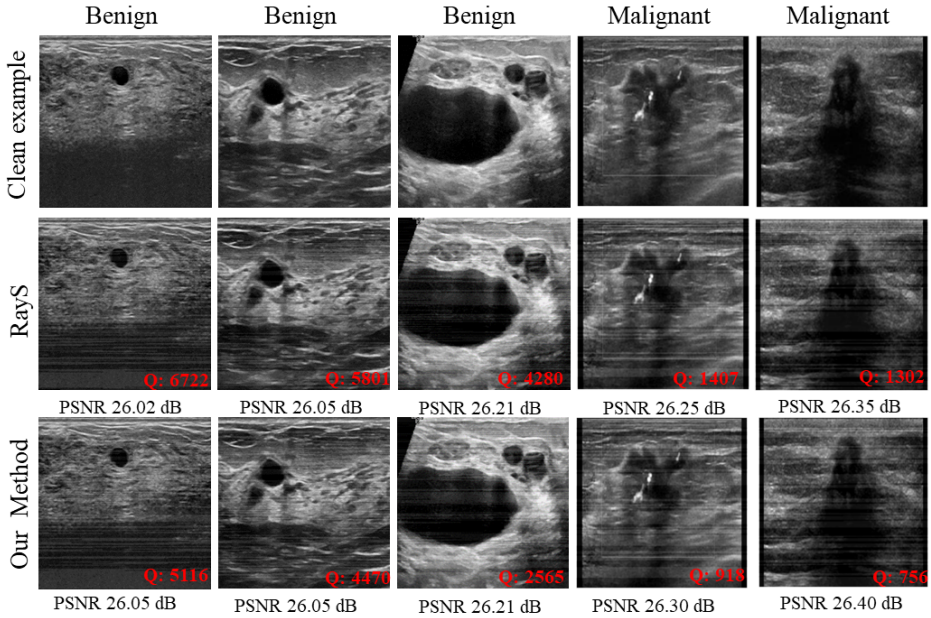

Figure 1. Adversarial images generated using RaySS and our method. Q: number of queries, PNSR: peak signal-to-noise ratio.

Adversarial attack research focuses on studying and understanding the vulnerabilities of ML models to malicious attacks. The purpose of this research is to identify weaknesses in ML models and to develop techniques to defend against adversarial attacks to ensure the reliability and trustworthiness of ML-based systems. The primary contributions of our study are summarized below.

- The proposed adaptive binary search algorithm effectively reduces unnecessary queries by searching for adversarial samples in a coarse-to-fine manner.
- The proposed semantic-aware search algorithm avoids invalid searches by cropping the search space using semantic masks from breast anatomy.

- The combination of the two above algorithms leads to a novel hard-label black-box attack approach. It significantly reduces the number of queries for searching adversaries for extreme samples (Figure 1).

The rest of the paper is organized as follows. In Section 2 we discuss relevant background information and related work. In Section 3 we present the general formulation of a hard-label black-box adversarial attack, followed by our proposed methods. Section 4 outlines our experimental setup, relevant parameter settings, and results. Finally, in Section 5 we summarize our proposed work and experimental results.

2. Related Works

2.1. White-Box Adversarial Attack

L-BFGS [2] was the first article that demonstrated that adding a small perturbation to an image can evade a target ML classifier and produce a misclassification. FGSM attack [3] used the sign of the gradient of a neural network loss with respect to the input image to find adversarial perturbations. Ref. [11] proposed an FGSM-like attack that used an attention map to help determine salient image regions that, when perturbed, are more likely to induce a misclassification. PGD attack [4] applied FGSM iteratively with a smaller distortion in each step to minimize the overall perturbation. CW attack [5] formulated the adversarial example generation as a constrained optimization problem that approximates the minimal perturbation for misclassifying an input sample. In [12], the authors develop a novel attack framework that uses trust regions [13] to more efficiently generate adversarial samples with smaller perturbations when compared to other white-box methods like CW.

2.2. Black-Box Adversarial Attack

A query-based (soft-label) black-box attack obtains information about a target ML model by querying the model. In a score-based attack, an adversary can acquire probability confidence scores from the targeted model; e.g., zeroth order optimization approaches [14] can employ returned confidence scores to estimate the gradient and generate adversarial samples. SimBA [15] used confidence scores to create an adversarial sample that moves away from clean samples and toward the decision boundary. LeBA [16] generates adversarial samples using a surrogate model. Surrogate models are trained to mimic the behavior of the model being attacked (sometimes referred to as the victim model). LeBA trains the surrogate model using information about the soft labels predicted using the victim model, thereby encouraging the surrogate model to more closely match the victim model. In [17], the authors develop a framework for training a generalized surrogate model by performing meta-training across several different models and datasets. This generalized model can then be fine-tuned using a sm all amount of query-based information from the target model. SignHuner [18] estimated the gradient sign bits [19] based on the gradient from the previous step to reach faster convergence.

Decision-based (hard-label) attacks only have access to the top-1 class prediction from the target model [20]. HJSA [21] utilizes binary information at the decision boundary to calculate the directional gradient. SignOPT [22] extended the work in [23], and specified a single query oracle for computing the zeroth-order gradient direction. RayS [10] built upon the research in SignHuner and SignOPT and proposed a hyperparameter-free decision-based attack without zeroth-order gradient estimation. It significantly reduced the number of queries for attacking DNNs trained to perform natural image classification to several hundred. But, as shown in Figure 1, it can fail to use this small number of queries to find adversaries for many extreme BUS images.

2.3. Adversarial Defense Strategies

In addition to developing adversarial attacks, adversarial machine learning also involves finding ways to defend against such attacks. These defense strategies can be broadly categorized as detection methods or adversarial defense methods. Detection methods aim to develop methods to identify adversarial images, while adversarial defense methods aim

to improve the adversarial robustness of a model. Models with good adversarial robustness are less likely to misclassify adversarial images. In [24], the authors propose a novel adversarial detection framework. An image is processed via smoothing and quantization. If the predicted label of the image before and after processing differs, it is likely that the image has been adversarially tampered with. Another way to detect such attacks is to embed imperceptible watermarks into images that are known to be clean [25]. If the extracted watermark is different than the original watermark, it is likely the image has been tampered with. In [26], the authors propose a novel defense strategy that combines features extracted from a deep convolutional network with feature vectors created using more traditional image processing techniques like HOG [27].

2.4. Breast Ultrasound Image (BUS) Classification

Recent studies showed that DNNs can enhance the classification of breast ultrasound images. Ref. [28] indicated deep learning outcomes in biomedical applications could be significantly enhanced through the development of a pre-trained and fine-tuned convolutional neural network (CNN) architecture using medical imaging data. Ref. [29] designed a dual-sampling convolutional neural network (DSCNN) with residual networks for the diagnosis of breast tumor images. Their network could prevent gradient disappearance and degradation, leading to improved accuracy by using a parallel DSCNN. ESTAN [30] presented a new architecture that utilizes two encoders to extract global and local information and integrate image context information at varying scales. The authors introduced row–column-wise kernels that conform to the horizontal arrangement of the breast anatomy tissue layers in BUS images. By employing this approach, their network demonstrated enhanced segmentation performance for tumors of varying sizes and surpassed the performance of existing state-of-the-art methods for BUS segmentation. MT-ESTAN [31] is a multi-task neural network that combines two separate tasks, (1) tumor classification (primary task) and (2) tumor segmentation (secondary task). The backbone for this multi-task model is based on ESTAN. The network learns from both tasks and mitigates low generalization issues caused by small training datasets. The learned shared features between object segmentation and classification improve the robustness and generalizability of the model.

However, Ref. [32] implemented FGSM, BIM [33], PGD, and CW on medical images and proved that medical deep learning systems are vulnerable to small carefully engineered perturbations. They explained the reasons are the complex biological textures in medical images that cause higher gradient regions, and those regions are sensitive to small adversarial perturbations. Moreover, DNNs that are currently used for large-scale natural image classification may not be optimized for medical imaging tasks due to overparameterization. This can lead to a sharp loss landscape and an increased susceptibility to adversarial attacks.

3. Materials and Methods

3.1. Preliminary

Let f be a DNN model, x_0 the input (clean image), and y the true class label associated with x_0. A general untargeted hard-label black-box attack can be formulated as

$$x^* = \arg\min_x \ \mathrm{D}(x, x_0)_\infty \text{ such that } f(x) \neq y \tag{1}$$

where $\mathrm{D}(x, x_0)_\infty$ is the distance between x and x_0 according to the L_∞ norm. The goal is to generate an adversarial sample x nearest to the x_0. OPT and SignOPT [22,23] re-formulated the hard-label black-box attack as a two-stage optimization

$$\theta^* = \arg\min_\theta \ g(\theta) \tag{2}$$

$$g(\theta) = \text{minimize } \lambda \quad \text{s.t. } f(x_0 + \lambda \frac{\theta}{||\theta||}) \neq y. \tag{3}$$

Instead of directly searching an adversarial sample, SignOPT searches the direction θ to minimize the distortion $g(\theta)$ in Equation (2). In Equations (2) and (3), $g(\theta)$ is the distance from x_0 to the closest adversarial sample along the search direction θ and λ is the decision boundary radius in the corresponding search direction. $g(\theta)$ is calculated using a binary search algorithm, and SignOPT uses the gradient descent method to solve this optimization problem. They estimate the gradient by taking the average of a random Gaussian vector (descent direction) through the sign of 200 random directions.

RayS [10] significantly improved the optimization of the above-formulated problem by bounding the search space to a discrete set of ray directions, i.e., $\theta \in \{-1, 1\}^d$ where d is the number of dimensions of the space. It also modified the optimization framework of Equation (3) to a non-zeroth-order gradient estimation on the L_∞ norm ball. RayS reduced the possible searching directions from $\mathbb{R}^d$ [22,23] to 2^d. RayS used a greedy search algorithm without estimating any gradients, which only stored the best search direction established on the previous search direction. Different from OPT and SignOPT [21–23], taking multiple or an average of a few queries to determine the next search direction, RayS queried $f(x)$ once before doing a binary search for λ. The search direction is selected to perform a binary search, only if an adversarial sample x can be found.

Nevertheless, RayS lacked a reliable way to look for an adversarial example when the block partitions rise. The reason is the search directions get closer to each other. Their small constant binary search tolerance causes more unnecessary queries from hierarchical search. We proposed a novel approach that applies adaptive binary search and semantic-aware search to reduce the search queries on extreme examples. Our proposed method follows RayS (Equation (3)).

3.2. Adaptive Binary Search

The binary search has been widely adopted in hard-label black-box attack [10,15,21–23] to efficiently search adversaries along a direction (Equation (2)). It aims to minimize the distortion to generate adversarial examples close to the decision boundary. The efficiency of the binary search depends on the number of queries required to locate the best adversarial samples. The number of queries in binary search is directly determined by a tolerance value that defines the acceptable precision of the target adversary. For a specific search direction in the binary search algorithm, a qualified adversary should meet the following condition defined by a tolerance τ

$$||x_e - x_s||_\infty \leq \tau \text{ s.t. } f(x_s) = y \text{ and } f(x_e) \neq y, \tag{4}$$

where x_s is a clean sample, and x_e is its adversary along a search direction. In RayS [10], τ was defined as a fixed small positive value. Using a small τ allows the algorithm to find adversaries close to a decision boundary but needs more queries to search for a valid match. On the other hand, a large τ requires fewer queries but generates adversaries that are not close to a decision boundary. RayS used a hierarchical searching strategy and a fixed small τ to search a sequence of adversaries to approach the searching goal defined by the distance upper bound ϵ. Because of the fixed small τ, all adversaries are searched using the same high precision, resulting in many unnecessary queries. This may lead to an enormous number of queries, especially for a deep search hierarchy.

Inspired by the trade-off between tolerance and the number of queries, this work proposes the adaptive binary search (AdptBS) method detailed in Algorithm 1, which reduces the number of queries by using adaptive tolerance values. In the input of AdptBS, r_{best} is the distance from the current best adversary to the clean sample; ϵ is the upper bound of the distance and defines our searching goal. In line 1, θ_u is the unit search direction, and $x_0 + \theta$ is an adversary of x_0 along θ_u. In lines 2–3, the algorithm examines if an adversary can be generated using the best radius along the current search direction; if not, a better adversary with a shorter distance cannot be generated using θ_u, and the binary search is ended with only one query. Line 4 initializes the binary search by setting x_0 to the clean image and the initial adversarial sample to $x_0 + min(r_{best}, ||\theta||_2) \cdot \theta_u$. Lines 5 and 6 define

a soft margin ($[\mu \cdot \tau + \epsilon, \tau + \epsilon]$) and reduce the tolerance using the decay rate (c) if the distance is in the soft margin. The tolerance is used as the stopping condition of the binary search in lines 7–12. This adaptive change in the tolerance applies large τ when adversarial samples are not close to the decision boundary, which leads to a coarse search. When the distance between the adversary and the clean sample is in the soft margin, the reduced τ results in a fine search that produces adversaries closer to the decision boundary. The design reduces unnecessary fine searches in all directions and could reduce the number of queries significantly. As shown in Figure 2, for a direction with an initial faraway adversary, a coarse search with a large τ is applied; and for a direction with a close adversary, a fine search with a small τ is used. The blue bar defines the soft margin.

The initial τ and μ is set to 0.1 and 0.9, respectively, by experiments. The value of ϵ (0.3) is adopted from RayS [10]. The decay rate, c is set to 0.1 by experiments.

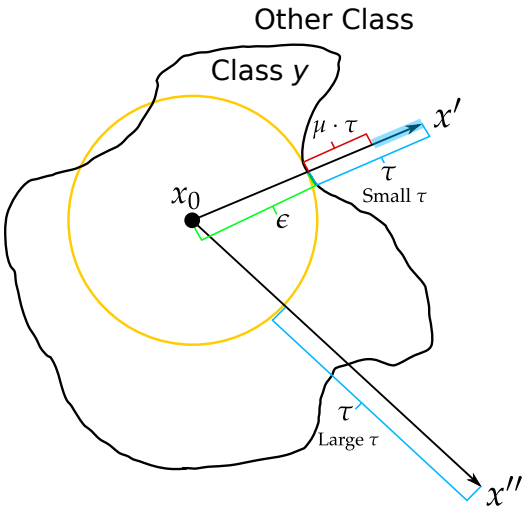

Figure 2. Adaptive tolerance range in Algorithm 1. x_0 is clean image. x' and x'' are two adversarial samples. The tolerance τ changes adaptively along different search directions.

Algorithm 1 AdptBS

Input: Model f, clean image x_0, label y, distance upper bound ϵ, search direction θ, best radius r_{best}, and tolerance τ

1: $\theta_u = \frac{\theta}{||\theta||_2}$ ▷ normalization
2: **if** $f(x_0 + r_{best} \cdot \theta_u) == y$ **then** ▷ Invalid search direction
3: return τ, ∞
4: $x_s = x_0$, $x_e = x_0 + min(r_{best}, ||\theta||_2) \cdot \theta_u$ ▷ x_s is the start point, x_e is the endpoint
5: **if** $\mu \cdot \tau + \epsilon < ||x_e - x_s||_\infty < \tau + \epsilon$ **then**
6: $\tau = \tau \cdot c$ ▷ c is the decay rate
7: **while** $||x_e - x_s||_\infty > \tau$ **do**
8: $x_m = (x_s + x_e)/2$
9: **if** $f(x_m) \neq y$ **then**
10: $x_e = x_m$
11: **else**
12: $x_s = x_m$
13: return τ, $||x_e - x_0||_\infty$

3.3. Semantic-Aware Search

RayS [10] determines the search directions using the raw image space with high dimensions. Guessing Smart [34] proposed, using a regional mask in their attack, to limit the perturbation to specific regions, thereby reducing the dimensionality of the search

directions and avoiding queries that were unlikely to change a model's results. In BUS images, tumor classes are mainly determined by image features in image regions of breast tumors and mammary tissues, and it is more efficient to search adversaries by adding perturbations only to these regions.

Inspired by the regional masking approach [34] and the nature of BUS images, this work proposes a semantic-aware search approach to reduce the dimensionality of search directions. In Algorithm 2, the semantic mask M is generated by using A2DMN [35] and has non-zero values for tumor and mammary regions (Figure 3), and Q is the query budget. In lines 3–5, the vector of the search direction is split into 2^k blocks, and all pixels in a block change directions together. Lines 10–11 adaptively increase the number of blocks to crop when fine-splitting ($k \geq K$) is applied. Line 14 limits the search to blocks in the union of the semantic mask M and cropping mask, which lets the search focus on regions of interest. The cropping mask trims blocks at the top and bottom of an image. Lines 15–17 attempt to find an adversary along the current direction; if this fails, the current direction is skipped. Lines 18–19 apply the proposed AdaptBS algorithm to get the best adversary along the current direction and update the best search direction (θ_{best}) and best distance.

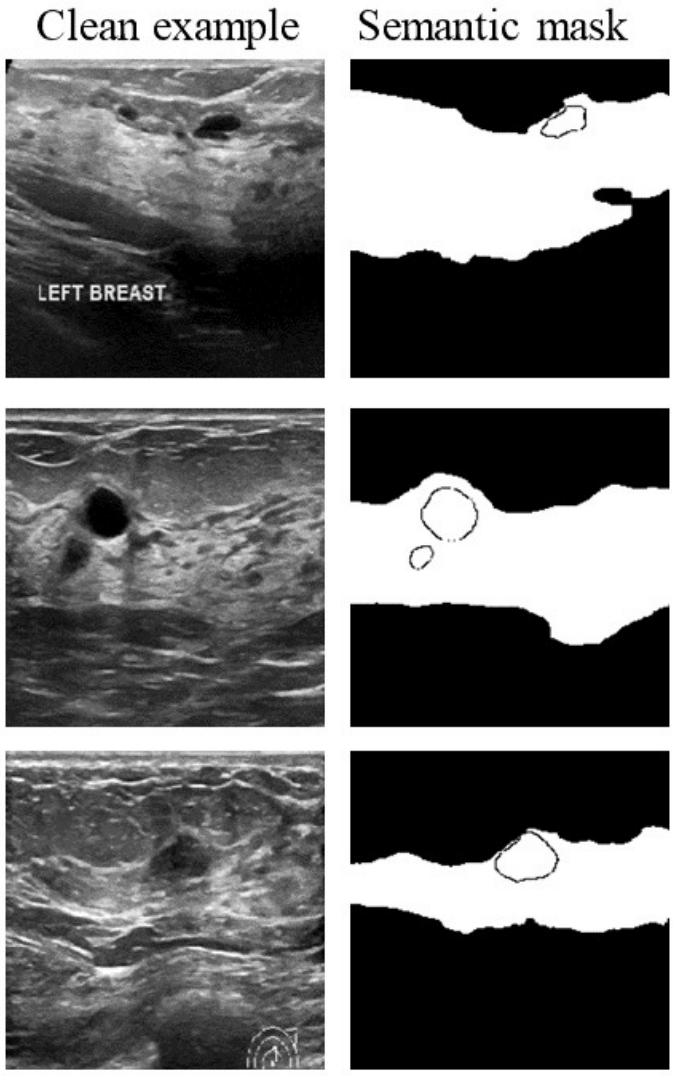

Figure 3. BUS images (**left column**) and their respective semantic masks for the mammary and tumor regions (**right column**).

In Figure 4, the cropping mask can be considered as a pruned branch to remove unnecessary parts of the search space, compared with RayS. The union of semantic masks and cropping masks skips checking a region that does not contain crucial features.

Algorithm 2 Semantic-Aware Search

Input: : Model f, clean image x_0, label y, distance upper bound ϵ, query limit Q, semantic mask M

Initialize the search direction $\theta_{best} = (1, \cdots, 1)$, best radius $r_{best} = \infty$, and block level $k = 0$
Initialize binary search tolerance $\tau = 0.1$, block level cut point K, and cropping block size $crop_k$
function BLOCKSPLIT(k)
    cut θ_{best} into 2^k blocks of equal size and save the splitting blocks into a list
    return the block list
remaining queries = Q
while remaining queries > 0 **do**
    $\theta_{tmp} = \theta_{best}$
    blocks = BlockSplit(k)
    if $k > K$ **then**
        $crop_k = crop_k \cdot 2$ ▷ skip more blocks for fine splitting
    for i in blocks **do**
        $\theta_{tmp}[ind_i] = -1 \cdot \theta_{tmp}[ind_i]$ ▷ ind_i contains the indices of all pixels in block i
        if $i \subseteq$ (blocks[$crop_k$:$-crop_k$] $\cup M$) **then**
            if An adversary x_{adv} along θ_{tmp} can be found **then**
                $\theta_{tmp} = x_{adv} - x_0$
            else: continue ▷ skip invalid search direction
            $\tau, r_{tmp} = \text{AdptBS}(f, x_0, y, \epsilon, \theta_{tmp}, r_{best}, \tau)$
            if $r_{tmp} < r_{best}$ **then**
                $r_{best} = r_{tmp}, \theta_{best} = \theta_{tmp}$
    $k \mathrel{+}= 1$
    if $r_{best} < \epsilon$ **then** ▷ early stopping
        break
return $x_0 + \frac{\theta_{best}}{||\theta_{best}||_2} \cdot r_{best}$

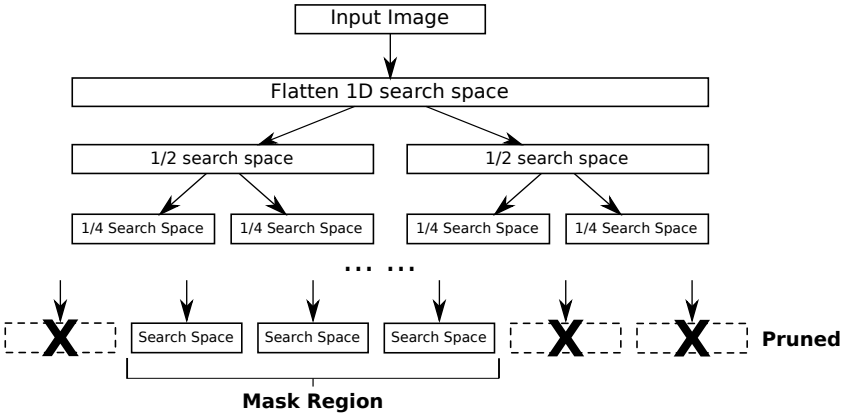

Figure 4. Block splitting in semantic-aware search.

4. Results

4.1. Experiment Setup

4.1.1. Datasets and Metrics

We validate the performance of attack algorithms using four publicly available BUS datasets: BUSI [36], BUSIS [37], HMSS [38], and Dataset B [39]. The combined dataset

contains a total of 3378 images (1698 benign and 1680 malignant). Most images in BUS datasets are rectangular. However, DNNs need a square input shape, so images must be resized to a uniform square shape. If rectangular images are directly resized to a square shape, the morphology of the tumor region will be distorted. To avoid these changes, all images and their corresponding semantic masks are zero-padded to be square following the procedure used in MT-ESTAN [31]. After padding, images and masks are resized to be 224 × 224 pixels. We randomly select 700 images from the dataset for testing, and the rest of the images are used for training. The number of average (AVG) and median (MED) queries, the attack success (SR) rate, and the peak signal-to-noise ratio (PSNR) is used to evaluate the performance of different adversarial attacks quantitatively. The number of average and median queries is counted from successfully attacked images on the test set. The success rate is the ratio between successful attacks and total attacks, and it is calculated using only images that were correctly classified via the targeted model on the test set. An attack is considered successful if the L_∞ norm between the adversarial sample and the clean image is less than the given ϵ. PSNR is used to measure the quality of the produced adversarial images. Higher PSNR values correspond to higher-quality adversarial images.

4.1.2. Experiment Environment

All neural network models and adversarial attacks are implemented using Python 3.7.0, Keras 2.3.1 [40], TensorFlow 1.13.1 [41], and Pytorch 1.9.1 [42]. All experiments were conducted with NIVIDA Quadro RTX 8000 GPUs, equipped with CUDA Toolkit 10.2.

4.1.3. Target Model Settings

We test the effectiveness of the proposed approach by attacking five well-known image classification networks, ResNet50 [43], VGG16 [44], DenseNet121 [45], MobileNetV2 [46], and InceptionV3 [47]. These models are selected because they are widely used and achieve reasonably good performance, are applied as target models in other black-box attack approaches, and have publicly available source codes. Each model is pre-trained using the ImageNet dataset and fine-tuned on the combined BUS dataset. All models were fine-tuned for 100 epochs with the Adam optimizer, a learning rate of 0.0001, and a batch size of 4. Their respective accuracy on the clean BUS dataset is 82.94%, 80.7%, 87.83%, 83.36%, and 83.64%. ResNet50 is used to validate the effectiveness of the adaptive binary search and semantic-aware search, as well as to find the best parameter settings for the proposed attacks. We compare the proposed approach with three other hard-label black-box attacks (OPT, Sign-OPT, and RayS) across the five aforementioned image classifiers. We select the three approaches because they achieve state-of-the-art performance in the setting of black-box hard-label attacks. It is an interesting topic to study the effectiveness of the proposed method against different defense mechanisms, and we will explore this in the future. The proposed experiments using five different network architectures are sufficient to validate the effectiveness of the proposed method to improve query efficiency.

4.1.4. Adversarial Attack Settings

Following the adversarial attack settings in RayS [10], we set the distance upper bound ϵ to 0.05, and the maximum number of queries to 10,000 for all attacks. We use the original paper's parameter settings for OPT [23] and Sign-OPT [22].

4.2. The Effectiveness of Adaptive Binary Search

In this section, different binary search strategies in the RayS attack framework are validated using ResNet50. The original RayS set a fixed binary search tolerance τ to 0.001. The results of RayS with other τ values and AdptBS are reported in Table 1. The results in the first three rows show that the number of queries (AVG and MED) is sensitive to τ. Small τ values need more queries to find adversaries. On the other hand, large τ values can significantly reduce the queries but may also lead to a decreased success rate. Therefore, the attacker would need to perform the attack several times to find the best value of τ, which is

not efficient. Thus, an adaptive τ can automatically obtain the best value for searching the initial attack starting point and adjust tolerance based on the current distance between the adversarial sample and the decision boundary. The proposed AdaptBS algorithm is used to replace the fixed tolerance binary search algorithm in the RayS attack, and its results are shown in the last three rows of Table 1. The AdptBS method preserves the high success rate (99.83%) of the original RayS with fixed small τ and reduces the AVG and MED queries by 21.47% and 29.37%. μ decides when to change τ; the best result is $\mu = 0.9$ and the performance is reduced when using smaller τ. $\mu = 0.9$ restricts τ to quickly reduce to a very small value, which uses large τ for immense perturbations and small τ when perturbations are near ϵ.

Table 1. The Results of the attacks with different binary search methods. The percentage values in the parenthesis show the reduction of queries compared with the baseline method (first row).

Attack Method	μ	τ	Queries (AVG) ↓	Queries (MED) ↓	SR (%) ↑
RayS [10]	-	0.001(original)	411.94	248.5	99.83
	-	0.1	299.06 (−27.40%)	159.0 (−36.01%)	99.15 (−0.68%)
	-	0.01	346.07 (−15.99%)	206.0 (−20.63%)	99.83
AdptBS	0.9	Adaptive	**323.47 (−21.47%)**	**175.5 (−29.37%)**	99.83
	0.8		325.91 (−20.88%)	178.5 (−28.16%)	99.83
	0.7		330.26 (−19.82%)	181.0 (−27.16%)	99.83

4.3. The Effectiveness of Semantic-Aware Search

The semantic-aware search in RayS and Semantic-Aware AdaptBS are compared using ResNet50. The semantic-aware search aims to reduce the search space. It is integrated into RayS and the proposed AdptBS algorithm and the results are shown in Table 2. The original RayS with the proposed semantic-aware search reduces the AVG and MED queries by 3.87% and 4.22%, respectively. The parameters used for the attack are $\tau = 0.001$, $crop_k = 2$ and $K = 5$. The semantic-aware search with the proposed AdptBS algorithm can reduce the AVG queries by 23.96% and the MED queries by 31.79% with $\mu = 0.9$, $crop_k = 2$, and $K = 5$. The impressive results demonstrate that adding small perturbations only to breast tumors and mammary regions can find adversarial samples more efficiently. Table 2 demonstrates that the best $crop_k$ is 2, and the SR is reduced when $crop_k$ is increased to a larger value. Table 2 shows MED queries in Semantic-Aware AdptBS for $K = 5$ and $K = 6$ are both 169.5. Since $K = 5$ performed slightly better in terms of AVG queries, we set $K = 5$ for the rest of the experiments.

Table 2. Results of attacks with semantic-aware search.

Attack Method	μ	$crop_k$	K	Queries (AVG)↓	Queries (MED)↓	SR (%)↑
RayS [10]	-	-	-	411.94	248.5	99.83
RayS + Semantic Mask	-	2	4	402.29 (−0.23%)	236.5 (−4.82%)	99.83
	-	2	5	395.98 (−3.87%)	238.0 (−4.22%)	99.83
	-	2	6	399.68 (−2.97%)	238.5 (−4.02%)	99.83
	-	3	5	404.70 (−1.75%)	243.5 (−2.01%)	99.83
	-	4	5	402.29 (−2.34%)	236.5 (−5.07%)	99.83
Semantic-Aware AdptBS	0.9	2	4	308.62 (-25.07%)	164 (−34.00%)	99.66 (−0.17%)
		2	5	**313.22 (−23.96%)**	**169.5 (−31.79%)**	99.83
		2	6	318.21 (−22.75%)	169.5 (−31.79%)	99.83
		3	5	309.18 (−24.94%)	170.0 (−31.58%)	99.49 (−0.34%)
		4	5	308.62 (−25.08%)	164.0 (−34.00%)	99.66 (−0.17%)

The proposed method is also effective at reducing the number of queries for searching adversaries for extreme samples. When using RayS, 29.51% and 15% of test images use more than 400 and 600 queries, respectively. However, for the proposed approach, only 15.68% and 9.61% of test images use more than 400 and 600 queries, respectively. Instances of extreme cases in both approaches are shown in Figure 1.

4.4. Attack on Other Deep Classifiers

RayS and the proposed method are used to attack five deep learning models, ResNet50, DenseNet121, VGG16, MobileNetV2, and InceptionV3. The five models are pre-trained on the ImageNet dataset and fine-tuned on the training set of the combined BUS dataset based on the settings used in [31]. The tolerance τ of RayS is 0.001 for all target models. The parameters for Semantic-Aware AdptBS are $\mu = 0.9$, $crop_k = 2$, and $K = 5$ for all models. As shown in Table 3, the proposed method outperforms RayS in terms of AVG and MED queries when attacking all models. For VGG16, the MED queries of the proposed method are 33.67% less than that of the RayS. For DeseNet121, the MED queries of the proposed method are 32.03% less than that of the RayS. In addition, DenseNet121 required more queries during attacks, which indicates that the model is more robust than the other four models. Since we set the distance upper bound ϵ to 0.05, the PSNR score among all five models except MobileNetV2 is close to 27 dB. The proposed attack achieved a similar PSNR score with fewer queries.

Table 3. Results of attacking five different models.

Model	Method	Queries (AVG)↓	Queries (MED)↓	SR (%)↑	AVG PSNR (dB)↑
ResNet50	RayS	411.94	248.5	99.83	27.90
	Ours	313.22	169.5	99.83	27.91
DenseNet121	RayS	618.67	384.0	99.52	27.57
	Ours	509.28	261.0	99.52	27.55
VGG16	RayS	456.06	297.0	100	27.64
	Ours	368.68	197.0	100	27.58
MobileNetv2	RayS	417.27	256.5	100	28.16
	Ours	317.61	177	100	28.18
Inceptionv3	RayS	483.88	296	99.83	27.87
	Ours	370.35	196	99.83	27.85

4.5. Comparison with State-of-the-Art Attacks

Three state-of-the-art hard-label black-box attack approaches (i.e., OPT [23], SignOPT [22], and RayS) are compared with the proposed method. The three attacks use a fixed tolerance for binary search. ResNet50 is used as the baseline classifier. As shown in Table 4, Sign-OPT and OPT randomly initialize a starting point with Gaussian noise or uniform noise and require many queries when using binary search to find the direction with the shortest distance to the decision boundary and to calculate the directional derivative. Each search direction generated using random noise needs a binary search to find the closest distance to the decision boundary. The binary search for each direction causes a massive number of queries. RayS significantly improves the success rate and reduces the queries from several thousand to only several hundred due to its novel search strategy in a discrete space. These three attacks all use the same tolerance for binary search, which produces more queries to find an adversarial sample close to the decision boundary. Moreover, all three attacks search the entire image for each iteration, which is a large search space. The local semantic-aware search shrinks the search space to only include significant features. The proposed method achieves the same success rate as RayS but outperforms it in terms of

AVG and MED queries; e.g., the proposed method's AVG queries are 21.5% less than Rays, and its MED queries are 29.4% less than Rays. Since the proposed method outperforms the state-of-the-art by a large margin, a statistical analysis is not necessary to prove the significance of the results. Both OPT and SignOPT have higher PSNR, but the success rate is lower than our methods.

Table 4. The performance of the state-of-the-art hard-label black-box attack approaches on ResNet50.

Method	Queries (AVG)↓	Queries (MED)↓	SR (%)↑	AVG PSNR (dB)↑
OPT [23]	3218.36	2120.5	33.72	41.45
Sign-OPT [22]	7066.05	7137.0	24.78	42.63
RayS [10]	411.94	248.5	99.83	27.90
Ours	**323.47**	**175.5**	99.83	27.91

5. Conclusions

Developing adversarial attacks is critical for investigating and mitigating potential vulnerabilities in deep learning-based models, especially in the context of high-risk applications like computer-aided diagnostic systems and self-driving vehicles. A failure to mitigate these vulnerabilities could lead to improper medical diagnoses, significant property damage, or loss of life. This work introduces a novel black-box adversarial attack approach against deep learning classifiers for breast ultrasound images. It only requires hard-label predicted outputs created via the target model for the generation of adversarial samples. The proposed attack method integrates the semantic-aware search and adaptive binary search and outperforms state-of-the-art approaches in terms of average and median queries. The adaptive binary search component allows the automatic selection of an appropriate tolerance for binary search in different search stages. Using a semantic mask reduces the attack search space, which is critical due to the tremendous impact on model prediction. Experimental results on a large dataset of BUS images demonstrate the query efficiency and the effectiveness of the proposed black-box attack.

Despite the good performance of the proposed attacks, there are several limitations and shortcomings. The Semantic AdptBS algorithm is dependent on the availability and quality of semantic masks. In general, image classification datasets do not have semantic masks available. Moreover, it is unclear whether the performance of the attack on different learning tasks (e.g., image segmentation, video classification, etc.) will be affected. Moreover, in this manuscript, we only have the attacks on BUS images. Therefore, there is a chance our method is slightly biased towards this image modality, but given that the algorithms make no explicit assumptions about image modality, we do not believe this is likely. It is also unclear how well these attacks will perform on natural images, images from other medical modalities, or in the presence of adversarial defense strategies. Future work will be focused on investigating and mitigating these potential shortcomings.

Author Contributions: Conceptualization, Y.M. and M.X.; methodology, Y.M. and M.X.; software, Y.M. and K.L.; validation, Y.M., K.L. and M.X.; formal analysis, Y.M. and A.V.; investigation, Y.M. and A.V.; resources, M.X. and A.V.; data curation, Y.M. and K.L.; writing—original draft preparation, Y.M.; writing—review and editing, Y.M., K.L., A.V. and M.X.; visualization, Y.M. and K.L.; supervision, M.X. and A.V.; project administration, M.X.; funding acquisition, M.X. and A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. American Cancer Society. Cancer Facts & Figures. Available online: https://cancerstatisticscenter.cancer.org/#!/ (accessed on 12 June 2024).
2. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
3. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
4. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
5. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (sp), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
6. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
7. Papernot, N.; McDaniel, P.; Goodfellow, I. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. *arXiv* **2016**, arXiv:1605.07277.
8. Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. 2009; pp. 32–33. Available online: https://xueshu.baidu.com/usercenter/paper/show?paperid=1b030ma06t5208m06s6s0ju0e4025736 (accessed on 15 August 2024).
9. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
10. Chen, J.; Gu, Q. Rays: A ray searching method for hard-label adversarial attack. In Proceedings of the Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual, 6–10 July 2020; pp. 1739–1747.
11. Dong, X.; Han, J.; Chen, D.; Liu, J.; Bian, H.; Ma, Z.; Li, H.; Wang, X.; Zhang, W.; Yu, N. Robust superpixel-guided attentional adversarial attack. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12895–12904.
12. Yao, Z.; Gholami, A.; Xu, P.; Keutzer, K.; Mahoney, M.W. Trust region based adversarial attack on neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11350–11359.
13. Steihaug, T. The Conjugate Gradient Method and Trust Regions in Large Scale Optimization. *SIAM J. Numer. Anal.* **1983**, *20*, 626–637. [CrossRef]
14. Chen, P.Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017; pp. 15–26.
15. Guo, C.; Gardner, J.; You, Y.; Wilson, A.G.; Weinberger, K. Simple black-box adversarial attacks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 2484–2493.
16. Yang, J.; Jiang, Y.; Huang, X.; Ni, B.; Zhao, C. Learning black-box attackers with transferable priors and query feedback. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12288–12299.
17. Ma, C.; Chen, L.; Yong, J.H. Simulating unknown target models for query-efficient black-box attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11835–11844.
18. Al-Dujaili, A.; O'Reilly, U. There are No Bit Parts for Sign Bits in Black-Box Attacks. *arXiv* **2019**, arXiv:1902.06894.
19. Bernstein, J.; Wang, Y.X.; Azizzadenesheli, K.; Anandkumar, A. signSGD: Compressed optimisation for non-convex problems. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 560–569.
20. Brendel, W.; Rauber, J.; Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv* **2017**, arXiv:1712.04248.
21. Chen, J.; Jordan, M.I.; Wainwright, M.J. Hopskipjumpattack: A query-efficient decision-based attack. In Proceedings of the 2020 IEEE Symposium on Security and Privacy (sp), Francisco, CA, USA, 18–21 May 2020; pp. 1277–1294.
22. Cheng, M.; Singh, S.; Chen, P.; Chen, P.Y.; Liu, S.; Hsieh, C.J. Sign-opt: A query-efficient hard-label adversarial attack. *arXiv* **2019**, arXiv:1909.10773.
23. Cheng, M.; Le, T.; Chen, P.Y.; Yi, J.; Zhang, H.; Hsieh, C.J. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv* **2018**, arXiv:1807.04457.
24. Liang, B.; Li, H.; Su, M.; Li, X.; Shi, W.; Wang, X. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Trans. Dependable Secur. Comput.* **2018**, *18*, 72–85. [CrossRef]
25. Klington, G.; Ramesh, K.; Kadry, S. Cost-Effective watermarking scheme for authentication of digital fundus images in healthcare data management. *Inf. Technol. Control* **2021**, *50*, 645–655. [CrossRef]
26. Lal, S.; Rehman, S.U.; Shah, J.H.; Meraj, T.; Rauf, H.T.; Damaševičius, R.; Mohammed, M.A.; Abdulkareem, K.H. Adversarial Attack and Defence through Adversarial Training and Feature Fusion for Diabetic Retinopathy Recognition. *Sensors* **2021**, *21*, 3922. [CrossRef]
27. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.

28. Hijab, A.; Rushdi, M.A.; Gomaa, M.M.; Eldeib, A. Breast cancer classification in ultrasound images using transfer learning. In Proceedings of the 2019 Fifth International Conference on Advances in Biomedical Engineering (ICABME), Tripoli, Lebanon, 17–19 October 2019; pp. 1–4.
29. Xie, J.; Song, X.; Zhang, W.; Dong, Q.; Wang, Y.; Li, F.; Wan, C. A novel approach with dual-sampling convolutional neural network for ultrasound image classification of breast tumors. *Phys. Med. Biol.* **2020**, *65*, 245001. [CrossRef] [PubMed]
30. Shareef, B.; Vakanski, A.; Freer, P.E.; Xian, M. Estan: Enhanced small tumor-aware network for breast ultrasound image segmentation. *Healthcare* **2022**, *10*, 2262. [CrossRef] [PubMed]
31. Shareef, B.M.; Xian, M.; Sun, S.; Vakanski, A.; Ding, J.; Ning, C.; Cheng, H.D. A Benchmark for Breast Ultrasound Image Classification. *SSRN Electron. J.* **2023**. [CrossRef]
32. Ma, X.; Niu, Y.; Gu, L.; Wang, Y.; Zhao, Y.; Bailey, J.; Lu, F. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognit.* **2021**, *110*, 107332. [CrossRef]
33. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; pp. 99–112.
34. Brunner, T.; Diehl, F.; Le, M.T.; Knoll, A. Guessing Smart: Biased sampling for efficient black-box adversarial attacks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4958–4966.
35. Lucke, K.; Vakanski, A.; Xian, M. A2DMN: Anatomy-Aware Dilated Multiscale Network for Breast Ultrasound Semantic Segmentation. In Proceedings of the 2024 IEEE ISBI, Athens, Greece, 27–30 May 2024; pp. 1–5.
36. Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; Fahmy, A. Dataset of breast ultrasound images. *Data Brief* **2020**, *28*, 104863. [CrossRef] [PubMed]
37. Zhang, Y.; Xian, M.; Cheng, H.D.; Shareef, B.; Ding, J.; Xu, F.; Huang, K.; Zhang, B.; Ning, C.; Wang, Y. BUSIS: A benchmark for breast ultrasound image segmentation. *Healthcare* **2022**, *10*, 729. [CrossRef]
38. Geertsma, T. Ultrasoundcases.info, FujiFilm. Available online: https://www.ultrasoundcases.info/ (accessed on 1 September 2022).
39. Yap, M.H.; Pons, G.; Marti, J.; Ganau, S.; Sentis, M.; Zwiggelaar, R.; Davison, A.K.; Marti, R. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 1218–1226. [CrossRef]
40. Chollet, F. Keras. 2015. Available online: https://github.com/fchollet/keras (accessed on 1 January 2022).
41. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. *arXiv* **2016**, arXiv:1603.04467.
42. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, *1–12*.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
45. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
46. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
47. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *arXiv* **2015**, arXiv:512.00567.