engineering proceedings

MDPI

*Proceeding Paper*

# Enhancing Explainability in Convolutional Neural Networks Using Entropy-Based Class Activation Maps [†]

**Eric Boketsu-Boulu and Ghazal Rouhafzay**

Department of Computer Science, University of Moncton, NB, Canada; email1@email.com

* Correspondence: ghazal.rouhafzay@umoncton.ca

[†] Presented at The 11th International Electronic Conference on Sensors and Applications (ECSA-11), 26–28 November 2024; Available online: https://sciforum.net/event/ecsa-11.

**Abstract:** With the emergence of visual sensors and their widespread application in intelligent systems, precise and interpretable visual explanations have become essential for ensuring the reliability and effectiveness of these systems. Sensor data, such as that from cameras operating in different spectra, LiDAR, or other imaging modalities, is often processed using complex deep learning methods, whose decision-making processes can be unclear. Accurate interpretation of network decisions is particularly critical in domains such as autonomous vehicles, medical imaging, and security systems. Moreover, during the development and deployment of deep learning architectures, the ability to accurately interpret results is crucial for identifying and mitigating any sources of bias in the training data, thereby ensuring fairness and robustness in the model's performance. Explainable AI (XAI) techniques have garnered significant interest for their ability to reveal the rationale behind network decisions. In this work, we propose leveraging entropy information to enhance Class Activation Maps (CAMs). We explore two novel approaches: the first replaces the traditional gradient averaging scheme with entropy values to generate feature map weights, while the second directly utilizes entropy to weigh and sum feature maps, thereby reducing reliance on gradient-based methods, which can sometimes be unreliable. Our results demonstrate that entropy-based CAMs offer significant improvements in highlighting relevant regions of the input across various scenarios.

**Keywords:** Explainable AI; deep learning; Class Activation Maps

**Citation:** Boketsu-Boulu, E.; Rouhafzay, G. Enhancing Explainability in Convolutional Neural Networks Using Entropy-Based Class Activation Maps. *Eng. Proc.* **2024**, *5*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s):

Published: 26 November 2024

## 1. Introduction

The rapid advancement of deep learning has led to the development of highly complex models, often characterized by their difficult-to-interpret inner workings. As these models are increasingly deployed in many intelligent systems with critical applications, the need for transparency and interpretability has become crucial. This necessity has sparked significant interest in the field of Explainable AI (XAI), which seeks to clarify the reasoning behind a neural network's decisions and outputs. Enhancing the explainability of AI models not only improves trust but also serves as a crucial tool for identifying and mitigating biases present in training datasets.

In particular, Convolutional Neural Networks (CNNs), widely used in image classification tasks, often operate as "black boxes" due to the complexity of their decision-making processes. To address this challenge, a growing body of research has focused on developing methods that reveal the underlying factors driving CNNs' classification outcomes. They aim to provide insights into the features and regions of images that significantly influence the network's decisions [1].

From another perspective, a promising area within AI research is the development of weakly supervised object detection approaches, which focus on identifying regions of interest in an image containing an object class without requiring exhaustive pixel-level annotations. Thus, with two different objectives, many methods developed for

Explainable AI (XAI) share similarities with weakly supervised object detection approaches. The techniques used to interpret and explain the decisions of neural networks—by highlighting the key features or regions driving classification—can also be effectively leveraged as a weakly supervised approach for object detection [2,3]. Despite the progress made, many of the proposed XAI methods encounter challenges in specific scenarios, such as when dealing with multiple objects of the same class in an image. These limitations underscore the need for further research to identify failure cases and refine existing techniques. Addressing these challenges is essential for developing more robust and reliable methods that can effectively explain CNNs' decisions, thereby enhancing the practical utility and trustworthiness of AI systems.

## 2. Literature Review

In a broad categorization, the methods developed to interpret the decisions of complex machine learning models can be divided into gradient-free and gradient-based approaches. Gradient-free methods do not rely on backpropagated gradients within the CNN. For instance, Local Interpretable Model-agnostic Explanations (LIME) [4] approximates a black-box model with a locally interpretable linear model by perturbing the input and observing the resulting changes in prediction. This method is model-agnostic and provides valuable insights across all types of machine learning models. Similarly, SHapley Additive exPlanations (SHAP) [5] leverages Shapley values from cooperative game theory [6] to fairly distribute the model's prediction among input features, offering consistent explanations by considering the contribution of each feature across all possible subsets. When used for a Convolutional Neural Network (CNN) model, both LIME and SHAP can be computationally very expensive. The Occlusion Sensitivity by Zeiler and Fergus [7] is another gradient-free method that identifies critical regions in an input by systematically masking parts of the input and observing the effect on the model's output. This straightforward approach effectively highlights which areas of the input are most influential. Finally, Counterfactual methods such as CX-TOM [8] further explain predictions of a model by illustrating the minimal changes needed in the input to alter the model's decision. Such a technique is inspired by the way humans understand and explain phenomena through "Theory of Mind".

Gradient-based methods use the gradients of the classification score with respect to either the input or the features extracted by intermediate layers, such as the final convolutional layer, to understand and explain the model's decisions. These gradients are then used to weigh the extracted feature maps, effectively highlighting the most salient parts of the input that contribute to the model's prediction.

Grad-CAM (Gradient-weighted Class Activation Mapping) proposed by Selvaraju et al. [9] is a widely used method in explainable AI computing the gradients of the classification score with respect to feature maps in the last convolutional layer. Other variants of Grad-CAM have also been successfully applied to networks dedicated to processing 3D data, such as PointNet, to identify important regions on 3D objects. [10]. HiResCAM (High-Resolution Class Activation Mapping) [11] addresses the resolution limitations of Grad-CAM by focusing on high-resolution feature maps, providing finer-grained and more detailed visual explanations. Respond-CAM [12] advances the interpretability of convolutional neural networks (CNNs), particularly in the domain of 3D biomedical imaging, by addressing limitations in existing visualization methods like Grad-CAM. Respond-CAM incorporates a "sum-to-score" property, which ensures that the generated heatmaps accurately highlight the regions crucial for predictions in 3D images. This enhancement leads to more precise and interpretable visual explanations of CNNs.

Other algorithms, such as Eigen-CAM [13], argue that gradient information could be erratic or deceptive especially in the deeper segments of a deep CNN where gradients might be weak or disappear. Eigen-CAM eliminates reliance on gradient information, by applying principal component analysis (PCA) directly to the output of convolutional maps. This approach uses the principal components to generate clearer and more stable heatmaps, thereby improving the quality and interpretability of visual explanations.

## 3. Entropy CAM

Equations (1) and (2) describe the computation of gradcam map; The gradients of the classification score with respect to the final convolutional layer are global-average-pooled to obtain the weights of each extracted feature.

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

$$L_{Grad-CAM}^c = ReLU \sum_k \alpha_k^c \, A^k \tag{2}$$

A significant drawback of Grad-CAM as discussed in HiResCAM [8] and can be perceived in Equation (1) is its approach of averaging the gradient maps, which can lead to the neglect of important pixel-level information. By taking an average of the gradients across the spatial dimensions, Grad-CAM generates a course heatmap that highlights the general areas of the input that are influential to the model's decision. However, this averaging process smooths out the fine-grained details, potentially masking critical pixel-level contributions that could be essential for understanding the model's behavior. As a result, the resulting visual explanations may lack precision, providing a broad overview of relevant regions rather than a detailed, pixel-specific understanding of the model's focus.

In this work we suggest exploring the entropy information both in gradient maps and in feature maps extracted by the last convolution layer to create a feature map. The amount of disorder in a feature map obtained by a CNN can provide insights into various aspects of network performance and characteristics. To begin with it showcases the amount of information in the feature map; higher disorder indicates a more intricate range of values, which means that the feature map captures a wide array of features or patterns from the input data. On the hand lower disorder suggests a more uniform distribution potentially indicating that the feature map contains less information. Additionally, the level of disorder can shed light on how effective the feature map is, at distinguishing different elements. Maps with higher entropy could potentially better differentiate between various input patterns by capturing richer and more diverse information within them. Conversely lower entropy may indicate a decreased ability to discriminate between inputs effectively. Lastly entropy provides clues about the network's focus and attentiveness. In this scenario high entropy could suggest that the network is not narrowing its attention to special features but rather reacting broadly to the input at hand. Conversely a decrease in entropy may indicate that the system is giving attention to specific attributes while possibly neglecting others.

As such, we explore two different versions of using entropy information to create CAM. In the first approach, we simply replace the averaging scheme used to produce the weights for each feature map in Grad-CAM with the entropy of the map. Equations (3) and (4) describe the CAM calculations.

$$E_k^c = entropy(\frac{\partial y^c}{\partial A_{ij}^k}) \tag{3}$$

$$L_{Grad\_Entripy-CAM}^c = ReLU \sum_k E_k^c \, A^k \tag{4}$$

In a second attempt to reduce the dependence on gradient information, which could be unreliable, we weigh and sum the feature maps based on the entropy values of the feature maps from the last convolutional layer.

$$L^{c}_{Frature\_Entripy-CAM} = ReLU \sum_{k} entropy(A^{k})\, A^{k} \tag{5}$$

## 4. Experiment Setup

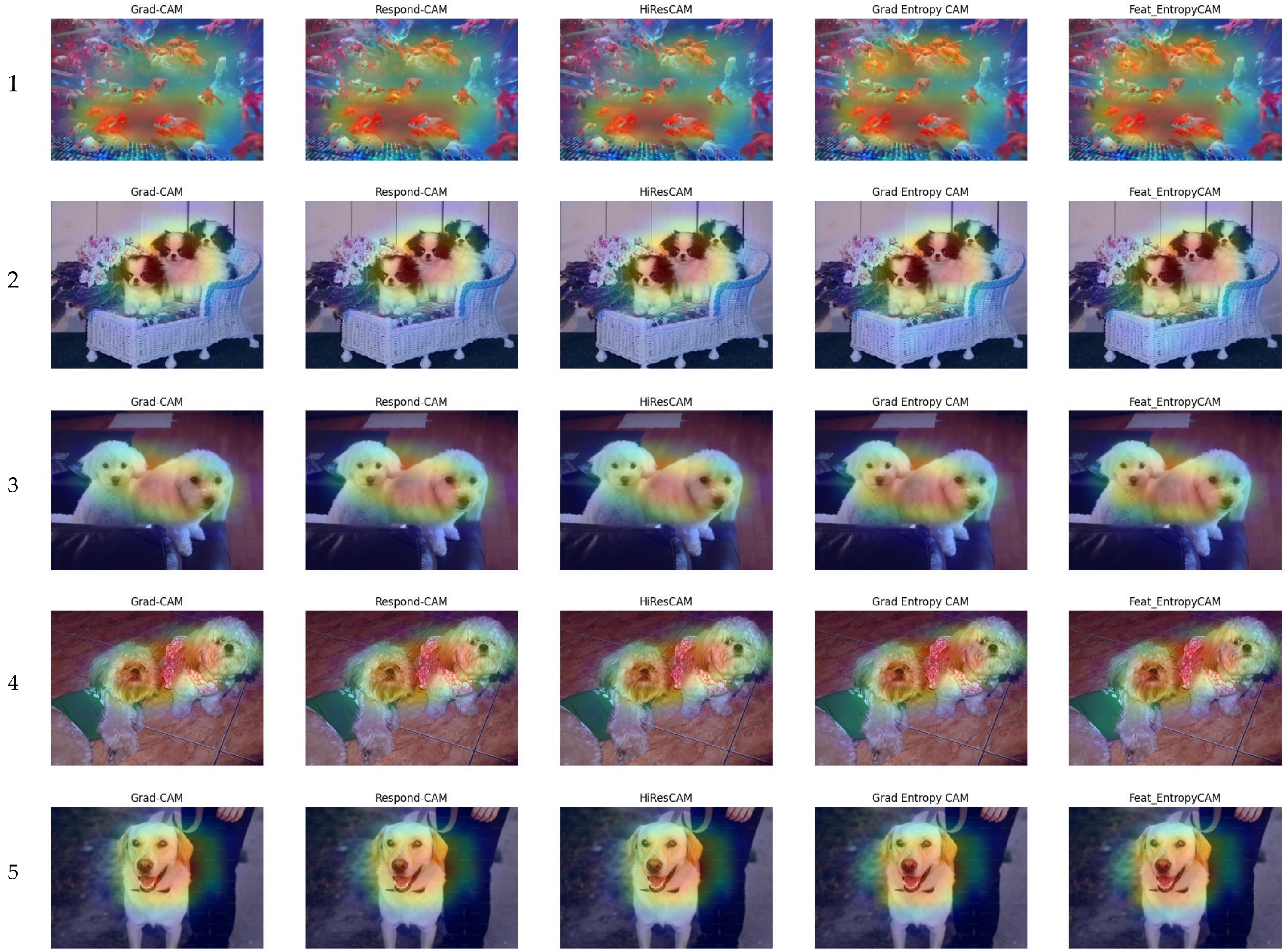
To create activation maps using the proposed Feature Entropy CAM and Grad Entropy CAM methods, we leverage the Xception [14] network pretrained on ImageNet [15]. We collected a dataset composed of challenging images from multiple sources, including the Stanford Dogs dataset [16], Fruit Classification dataset [17], and Goldfish dataset [18], all containing classes within the 1000 ImageNet categories. These datasets were chosen for their open access and their challenging nature with multiple objects. The Xception network was fine-tuned on 70% of the images and tested on the remaining 30%. The goal was to evaluate how effectively each method visualizes all object instances and more precisely locates them.

## 5. Results and Discussion

**I**n Figure 1, we compare explanation heatmaps produced by Grad-CAM, Respond-CAM, HiResCAM, Grad Entropy CAM, and Feature Entropy CAM. Each method highlights the regions of the image that most contributed to the network's decision. Among the visualized examples they are correctly classified by the network as "goldfish", "Japanese_spaniel", "Maltese_dog", "Shih-Tzu", "Labrador_retriever", "strawberry", "banana", and "strawberry" respectively.

These visualizations suggest that in many cases entropy-based methods for generating heatmaps can provide a finer and more nuanced view of feature importance compared to other methods Specially Grad-CAM. Entropy-based maps perform better in expanding the most important regions to cover all object instances, as seen notably in rows 1 and 8. Additionally, these maps often more accurately identify the best locations of other objects, as demonstrated in rows 6 and 7.

Entropy captures the level of disorder or variability within both gradient and feature maps, offering a more precise understanding of the network's focus. High entropy can reflect rich, diverse feature representations, allowing for better discrimination of input patterns.


Grad-CAM
Respond-CAM
HiResCAM
Grad Entropy CAM
Feat_EntropyCAM
1
2
3
4
5

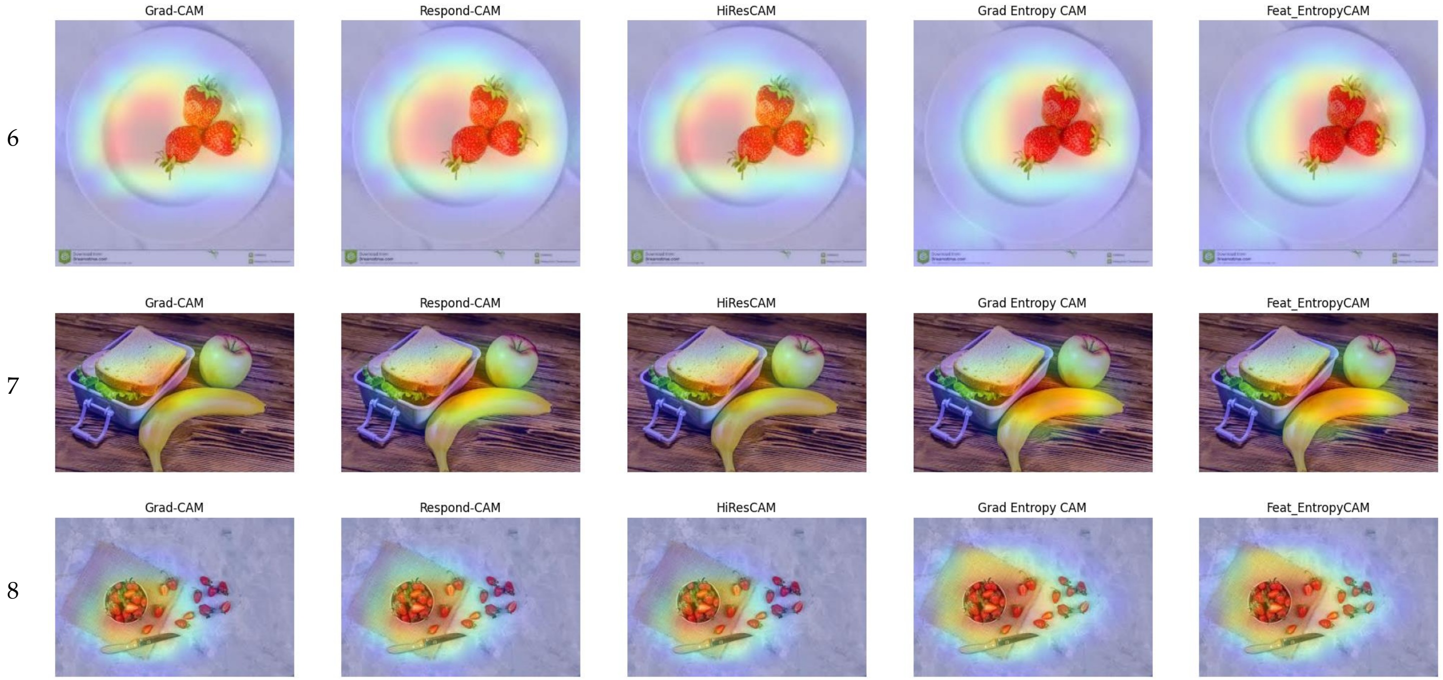


**Figure 1.** Comparison of CAM generated by different methods of the literature and entropy-based CAMs.

## 6. Conclusions

In conclusion, this work introduces two entropy-based CAM visualization techniques that utilize the amount of information contained in gradients and feature maps to generate heatmaps. Comparative visual evaluation indicates that these methods offer enhanced precision in determining the exact importance and localization of relevant regions in the input. By incorporating pixel-level entropy, these techniques provide more detailed and accurate explanations of model behavior, making them particularly beneficial in applications requiring fine-grained analysis of input features.

**Author Contributions:** Conceptualization, E.B.-B. and G.R.; methodology, G.R.; software, E.B.-B. and G.R.; validation, E.B.-B., and G.R.; formal analysis, E.B.-B.; investigation, E.B.-B., and G.R.; data curation, E.B.-B., and G.R.; writing—original draft preparation, G.R.; writing—review and editing, G.R.; visualization, E.B.-B.; supervision, G.R.; project administration, G.R. All authors have read and agreed to the published version of the manuscript.

**Funding:**

**Institutional Review Board Statement:**

**Informed Consent Statement:**

**Data Availability Statement:**

**Conflicts of Interest:**

## References

1. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3431–3440. https://doi.org/10.1109/CVPR.2016.371.
2. Cinbis, R.G.; Verbeek, J.; Schmid, C. Weakly Supervised Object Localization with Multi-Fold MIL Training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1687–1694. https://doi.org/10.1109/CVPR.2014.218.
3. Rouhafzay, G.; Tian, H.; Payeur, P. Warm Liquid Spill Detection and Tracking Using Thermal Imaging. In Proceedings of the IEEE 9th International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Chemnitz, Germany, 15–17 June 2022; pp. 92–97. https://doi.org/10.1109/CIVEMSA53371.2022.9853707.
4. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **2016**, *22*, 1135–1144. https://doi.org/10.1145/2939672.2939778.
5. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
6. von Neumann, J.; Morgenstern, O. *Theory of Games and Economic Behavior*; Princeton University Press: Princeton, NJ, USA, 1944.
7. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. *Eur. Conf. Comput. Vis.* **2014**, *8689*, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53.
8. Akula, A.R.; Wang, K.; Liu, C.; Saba-Sadiya, S.; Lu, H.; Todorovic, S.; Chai, J.; Zhu, S.-C. CX-ToM: Counterfactual Explanations with Theory-of-Mind for Enhancing Human Trust in Image Recognition Models. *Adv. Neural Inf. Process. Syst.* **2023**, *34*, 12345–12356.
9. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proc. IEEE Int. Conf. Comput. Vis.* **2017**, *2017*, 618–626. https://doi.org/10.1109/ICCV.2017.74.
10. Souai, Y.; Rouhafzay, G.; Cretu, A.-M. A Deep-Learning-Based Approach for Saliency Determination on Point Clouds. *Eng. Proc.* **2022**, *27*, 17. https://doi.org/10.3390/ecsa-9-13271.
11. Draelos, R.L.; Carin, L. Use HiResCAM Instead of Grad-CAM for Faithful Explanations of Convolutional Neural Networks. *arXiv* 2020, arXiv:2011.11293. Available online: https://arxiv.org/abs/2011.11293 (accessed on).
12. Zhao, G.; Zhou, B.; Wang, K.; Jiang, R.; Xu, M. Respond-CAM: Analyzing Deep Models for 3D Imaging Data by Visualizations. *IEEE Trans. Vis. Comput. Graph.* **2020**, *27*, 16–26.
13. Muhammad, A.; Kolouri, S.; Nasrabadi, N.M. Eigen-CAM: Class Activation Map using Principal Components. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 311–320. https://doi.org/10.1109/CVPR42600.2020.00321.
14. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258. https://doi.org/10.1109/CVPR.2017.195.

15. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
16. Khosla, A.; Jayadevaprakash, N.; Yao, B.; Li, F.-F. Stanford Dogs Dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 17–24. https://doi.org/10.1109/CVPR.2011.5995737.
17. Saxena, U. Fruits Classification. Retrieved from Kaggle. 2020. Available online: https://www.kaggle.com/datasets/utkarsh-saxenadn/fruits-classification (accessed on).
18. Bisht, S. Goldfish. Retrieved from Kaggle. 2020. Available online: https://www.kaggle.com/datasets/sumitbisht27/goldfish (accessed on).

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.