Journal of *Imaging*

MDPI

Article

Explainable Image Similarity: Integrating Siamese Networks and Grad-CAM

Ioannis E. Livieris [1,*], Emmanuel Pintelas [2], Niki Kiriakidou [3] and Panagiotis Pintelas [2]

[1] Department of Statistics & Insurance, University of Piraeus, GR 185-34 Piraeus, Greece
[2] Department of Mathematics, University of Patras, GR 265-00 Patras, Greece; e.pintelas@upatras.gr (E.P.); ppintelas@gmail.com (P.P.)
[3] Department of Informatics and Telematics, Harokopio University of Athens, GR 177-78 Athens, Greece; kiriakidou@hua.gr
* Correspondence: livieris@unipi.com

Abstract: With the proliferation of image-based applications in various domains, the need for accurate and interpretable image similarity measures has become increasingly critical. Existing image similarity models often lack transparency, making it challenging to understand the reasons why two images are considered similar. In this paper, we propose the concept of explainable image similarity, where the goal is the development of an approach, which is capable of providing similarity scores along with visual factual and counterfactual explanations. Along this line, we present a new framework, which integrates Siamese Networks and Grad-CAM for providing explainable image similarity and discuss the potential benefits and challenges of adopting this approach. In addition, we provide a comprehensive discussion about factual and counterfactual explanations provided by the proposed framework for assisting decision making. The proposed approach has the potential to enhance the interpretability, trustworthiness and user acceptance of image-based systems in real-world image similarity applications.

Keywords: explainability; siamese networks; Grad-CAM; recommendations.

Citation: Livieris, I.E.; Pintelas, E.; Kiriakidou, N.; Pintelas P. Explainable Image Similarity: Integrating Siamese Networks and Grad-CAM. *J. Imaging* **2023**, *9*, 224. https://doi.org/10.3390/jimaging9100224

Academic Editor: Christos Chrysoulas

Received: 6 September 2023
Revised: 3 October 2023
Accepted: 12 October 2023
Published: 14 October 2023

1. Introduction

In many real-world scenarios, the ability to measure image similarity is crucial for decision-making processes, intelligent systems as well as user interactions; therefore, image similarity models constitute a vital role in various computer vision tasks [1–5]. For example, in image retrieval systems, users often search for similar images based on a reference image or specific visual features [1]. Image similarity models allow these systems to find relevant images quickly and accurately. In content-based image analysis, large image databases are categorized and organized using image similarity models; hence, enabling the efficient automatic identification of similar images [2]. In copyright infringement detection or multimedia management, image similarity assists in identifying duplicate or visually similar images [3]. Furthermore, in medical imaging, comparing and matching medical images can aid in the diagnosis and identification of diseases or abnormalities [4]. Finally, image similarity can also assist in visual search engines, where users are able to visually find similar images without relying on text-based queries [5].

Siamese neural networks [6] probably constitute the most efficient and widely utilized class of image similarity models. During the last decade, they have been successfully applied for addressing image similarity tasks by quantifying the similarity between images through numerical values [7–9]. The backbone of this class of neural networks is convolutional layers, which are characterized by their remarkable abilities for image processing. Nevertheless, due to their architectural design, Siamese networks are not able to provide the users with human-understandable explanations about why two images are deemed similar. As the adoption of image-based technologies continues to grow in

diverse applications such as medical imaging, e-commerce, social media and security, the need for explainability in image similarity becomes paramount [10]. Explainability is a critical aspect of deep learning (DL), especially when dealing with complex models composed of convolutional layers. Although convolutional-based neural network models, such as Siamese networks, are highly effective in several image processing tasks, they lack transparency and explainability; thus, they are considered "black boxes" [11]. Notice that many traditional machine learning models, such as decision trees and linear models, often have the advantage of being interpretable, since their decision-making process is based on understandable features. Nevertheless, Siamese networks learn intricate and abstract features through layers of convolutions, making it challenging to directly interpret their decisions.

Explainability techniques aim to shed light by providing insights into how and why a convolutional-based model makes certain predictions by understanding the features and patterns, which the model learns from the data. These techniques not only enhance our understanding about the model's decision process but also play a vital role in building trust and accountability in artificial intelligence systems. More specifically, they enable us to verify the reasoning behind the predictions [12], identify potential biases, errors, or misinterpretations in model predictions and provide a means to improve their performance [13]. In addition, in some domains, there are strict regulations that require models to be interpretable. For instance, the General Data Protection Regulation in Europe includes the "*right to explanation*", which mandates that individuals should be provided with an explanation for automated decisions [14]. Finally, in certain contexts, there may be legal or ethical requirements to explain model predictions to end-users or stake-holders, making interpretability a crucial aspect of the deployment [10,15].

In the literature, several research directions have focused on enhancing the interpretability of deep learning models, particularly in the fields of computer vision [16,17]. Explainable artificial intelligence (XAI) techniques, such as attention mechanisms [18] and the Gradient-weighted Class Activation Mapping (Grad-CAM) technique [19], have been successfully applied to image classification, object detection and semantic segmentation tasks. However, the application of XAI to image similarity remains underexplored. In light of the increasing adoption of image-based technologies across various domains, the demand for explainable image similarity is considered crucial. Users and decision makers seek transparency in understanding why certain images are considered similar, especially in critical applications such as medical diagnosis or security surveillance. Therefore, exploring the integration of new or existing XAI techniques [20] with image similarity models [21] provides insights into the underlying similarities between images. Moreover, exploring the notion of similarity from a human-centric perspective may lead to novel contributions in image understanding and user-friendly applications.

In this work, we propose a new concept, named "*explainable image similarity*". Our primary aim is to bridge the gap between numerical similarity scores and human-understandable explanations. Along this line, we propose a new algorithmic framework, which integrates Siamese networks and Grad-CAM for providing explainability in image similarity tasks. The former are utilized for calculating the similarity between two input images while the latter is used for visualizing and interpreting the decisions made by a convolutional-based Siamese network. An attractive advantage of the proposed framework is that it is able to provide an image similarity score along with visual intuitive explanations for its decisions (factual explanations) together with explanations based on its ability regarding "what if" scenarios (counterfactual explanations). Finally, we provide a comprehensive discussion about factual and counterfactual explanations as well as the valuable insights and recommendations which can be made from the application of the proposed framework on three real-world use case scenarios.

At this point it is worth mentioning that although the Grad-CAM technique has been widely used and studied in a variety of domains, to the best of our knowledge it has never been utilized for image similarity tasks.

Summarizing, the main contributions of this work are described as follows:

- We propose the concept "*explainable image similarity*", highlighting the need for providing human-understandable explanations for image similarity tasks.
- We propose a new conceptual framework for explainable image similarity, which integrates Siamese networks along with the Grad-CAM technique, which is able to provide reliable, transparent and interpretable decisions on image similarity tasks.
- The proposed framework produces factual and counterfactual explanations, which are able to provide valuable insights and be used for making useful recommendations.

The rest of this paper is organized as follows: Section 2 presents the state-of-the-art works relative to the Grad-Cam technique and image similarity applications. Section 3 presents the concept of "*explainable image similarity*" as well as a detailed discussion about the proposed framework while Section 4 presents three use cases scenarios from its application. Finally, Section 5 discusses the proposed research, summarizes its conclusions and provides some interesting ideas for future work.

2. Related Work

Convolutional-based Neural Networks (CNNs) revolutionized modern computer vision and are widely regarded as the cornerstone choice for addressing image processing tasks [4,5,22,23]. The core element of CNNs are convolutional layers, which exploit a set of learnable filters (kernels) for generating feature maps. The aim is to highlight distinct attributes such as edges, textures and shapes, allowing subsequent layers to recognize higher-level representations.

Nowadays, explainability and interpretability play a significant role in bridging the gap between the advanced capabilities of DL models and the need for transparency and accountability in their decision-making processes. However, as CNNs become deeper and more complex, understanding how and why they make particular predictions becomes challenging. Grad-CAM [19] is a novel technique, which enhances the interpretability of CNNs focusing on highlighting the regions of an input image that significantly contribute to a specific prediction; thus, it has been applied in various applications. Hsiao et al. [24] exploited the flexibility of the Grad-CAM technique towards accurate visualization and interpretable explanation of CNNs. In particular, the authors utilized Grad-CAM to provide reliable and accurate analysis results for fingerprint recognition. Generally, fingerprints are difficult to analyze manually; hence, this study contributed to the assistance of criminal investigation cases. In similar research, Sang-Ho et al. [25] provided another application of the Grad-CAM technique in which they focused on providing a trading strategy for simultaneously achieving higher returns compared to benchmark strategies. Along this line, the authors used the Grad-CAM technique in conjunction with a CNN model aiming to develop a trustworthy method for meeting explainability as well as profitability in finance, therefore, fulfilling the challenging investors' needs.

In computer vision, the concept of image similarity consists of a fundamental building block for various real-world applications, ranging from image retrieval [26] and pattern recognition [25] to anomaly detection [27]. Siamese networks [6] have been established as state-of-the-art models for tackling image similarity tasks, especially where the available labeled data are limited. Their special architectural design enables them to learn and capture intricate relationships between pairs of images, allowing for the precise quantification of similarity and/or dissimilarity.

Appalaraju and Chaoji [7] proposed a new approach for identifying similar images using a deep Siamese network, named SimNet. In more detail, SimNet is trained on pairs of positive and negative images using a novel online pair mining strategy (OPMS). OPMS has been inspired by curriculum learning, a methodology for training DL models, aiming to ensure consistently increasing difficulty of input image pairs during the training process. Furthermore, another characteristic of SimNet is that it is composed of a multi-scale CNN, which is able to learn a joint image embedding of top and lower layers. For evaluating the model's performance, they utilized the widely used computer-vision object recognition

dataset, named CIFAR10. The experimental analysis and use case examples showed that the proposed SimNet model is able to better capture fine-grained similarities between images, compared to traditional CNNs. Additionally, the authors stated that the adopted curriculum learning strategy led to faster model training.

Melekhov et al. [8] proposed a novel methodology for exploiting Siamese networks for dealing with image similarity and classification problems. For detecting the matching and non-matching image pairs, the authors suggested representing them as feature vectors and distinguish the similarity between the input images using the Euclidean distance of these calculated feature vectors. In particular, those feature vectors are obtained through convolutional layers while the model training was based on contrastive loss. In their research, the authors used a large set of images from five different landmarks for evaluating the performance of the proposed Siamese model for image matching against widely used models such as AlexNet, HybridNet and sHybridNet. Based on their experimental analysis, the authors concluded that the proposed model reported promising performance on image similarity and classification tasks while in contrast to traditional models it is able to efficiently handle datasets with imperfect ground truth labels.

Rossi et al. [9] introduced a novel supervised Siamese deep learning architecture, which is a new Content-Based Image Retrieval system (CBIR) for assisting the process of interpreting a prostate radiological Magnetic Resonance Image (MRI). The rationale behind the architecture of the proposed approach is to integrate all available information in multi-parametric medical imaging tasks for predicting diagnostically similar images. Additionally, for handling multi-modal and multi-view MRIs, the authors considered the diagnostic severity of the lesion, assessed by the PI-RADS score [28], as the similarity criterion. It is worth mentioning that despite its initial purpose of development, this approach can be utilized for several diagnostic medical imaging retrievals due to its general design. As regards the experimental analysis, the authors presented that the performance of Siamese-based CBIRs was superior to that of the most widely used autoencoder-based CBIRs, for both diagnostic and information retrieval metrics.

In this research, we introduce the concept of explainable image similarity for providing useful, interpretable and transparent insights into the underlying factors driving image relationships and comparisons. In addition, we propose a new framework which integrates Siamese networks together with the Grad-CAM technique. The former are used for calculating the similarity between input images while the latter is used for visualizing and interpreting the decisions made by convolutional-based neural networks. In contrast to previous presented state-of-the-art approaches, the proposed framework is able to provide an image similarity score along with visual intuitive explanations for its decisions. The presented use case scenarios demonstrate the applicability of the proposed framework as well as a path for providing insights and useful recommendations from factual and counterfactual explanations.

3. Explainable Image Similarity

In this section, we present the proposed framework which is able to provide similarity scores along with visual transparent and understandable explanations for its decisions. We recall that our primary goal is to propose the concept of explainable image similarity for bridging the gap between numerical similarity scores and human-understandable explanations. By offering interpretable explanations, explainable image similarity not only enhances the usability of similarity-based applications but also empowers users to comprehend the reasoning behind the model's decisions, ultimately fostering informed and confident decision making.

In the following, we briefly present the main components of the proposed framework which is based on the integration of the Grad-CAM technique to Siamese networks as well as a detailed description, paying special attention to its capabilities and advantages.

3.1. Background

Siamese neural networks [6] constitute a special class of deep learning architectures, which are used in tasks involving similarity comparison, such as image or text matching [7,9,29]. These networks are characterized by their robustness to data which exhibit variations, distortions or noise as well as their requirement of significantly less labeled training data compared to neural networks; therefore, they have been well established for real-world scenarios [25–27,30,31]. A traditional Siamese network is composed of two identical sub-networks with shared weights (backbone network), allowing them to extract and encode into fixed-size feature vectors (embeddings) from input pairs. Then, the similarity of the input images (similarity score) is provided by computing the distance between the calculated embeddings.

Gradient-weighted Class Activation Mapping (Grad-CAM) [19] is a powerful and model-agnostic technique in the field of computer vision, which enhances the interpretability of deep neural networks. Grad-CAM provides a way to visualize and localize the regions of an input image, which contribute most to the model's decision. For obtaining the class-discriminative localization map, denoted by $L_{Grad-CAM}$, we initially calculate the neuron importance weights α_k using the gradient of the model's output y with respect to the k-th map activations A^k of a selected convolutional layer, which are flowed back and are global-average-pooled over the width (index i) and height (index j) dimensions, that is,

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

where Z is the total number of spatial locations in the feature maps. Then, we perform a weighted combination of forward activation maps, and follow it by ReLU activation function for calculating $L_{Grad-CAM}$, namely

$$L_{Grad-CAM} = \text{ReLU}\left(\sum_k a_k A^k \right). \tag{2}$$

By utilizing the gradients with respect to the model's internal feature maps, Grad-CAM generates an activation map, which highlights the discriminative regions responsible for the model's decision.

3.2. Proposed Framework

Next, we provide a detailed description of the proposed framework while a high-level presentation of its architecture is highlighted in Figure 1. Initially, two images are considered an input in a Siamese network, which are processed by the backbone network for encoding them into fixed-size feature vectors (embeddings). Then, the image embeddings are used for discerning similarities and differences between the input images and ultimately calculating their similarity score. Independently, the Grad-CAM technique is applied to the last convolutional layer of the backbone network for the development of the Grad-CAM heatmaps and visualizing the features, which significantly impact the Siamese model's decisions (factual explanations).

In addition, the proposed framework is able to provide counterfactual explanations. Actually, a counterfactual explanation provides a description of "*what would have not happened when a certain decision was taken*" [19]. This transparency not only enhances the model's interpretability but also empowers stakeholders to identify potential biases, assess model fairness and build trust in AI-driven systems, leading to more accountable and reliable artificial intelligence solutions. The counterfactual explanations can be easily developed by a slight modification to the Grad-CAM technique, namely, by simply replacing y with $1 - y$ in Equation (1). To summarize, the advantages of the proposed framework are:

- *Counterfactual explanations:* The identification of regions, which would make the network change its prediction, could highlight concepts that confuse the model. Therefore, by removing those concepts, the model's decisions may be more accurate or more confident.
- *Bias evaluation of model's decisions:* In case the Siamese model is performing well on both training and testing data (not-biased model), Grad-CAM heatmaps may be used to visualize the features, which significantly impact the model's decisions. In contrast, in case the Siamese model is performing well on the training data but it is not able to generalize well (biased model), Grad-CAM heatmaps can be efficiently used to identify unwanted features in which the model focuses on.

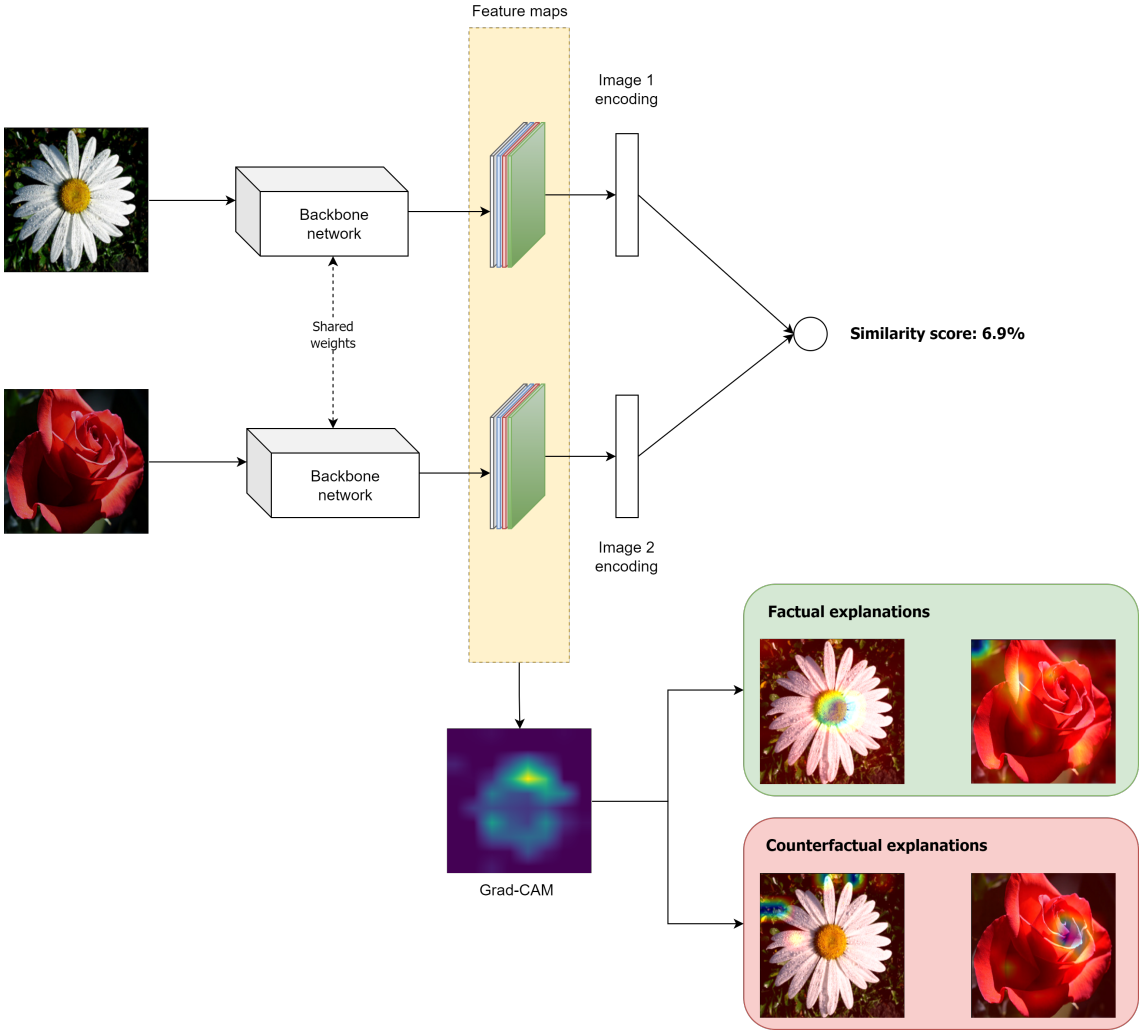

Figure 1. Architecture of the proposed framework.

4. Application of Proposed Framework and Use Case Scenarios

Next, we provide some use case scenarios from the application of the proposed framework to three (3) well-known datasets from different real-world application domains:

- *Flowers*. This dataset contains 4242 images (320 × 240) of flowers, which were categorized into five classes: "chamomile", "tulip", "rose", "sunflower" and "dandelion".
- *Skin cancer*. This dataset concerns images (224 × 224) of 1400 malignant and 1400 benign oncological diseases.
- *AirBnB*. This few-show dataset is composed of 864 interior and exterior house pictures (600 × 400) scraped from AirBnB over 3 cities, which were classified in 12 classes: "backyard", "basement", "bathroom", "bedroom", "decor", "dining-room", "entrance", "house-exterior", "kitchen", "living-room", "outdoor", "staircase" and "TV room".

The presented use cases focus on highlighting how the proposed framework could be used for image similarity tasks, what useful conclusions could be drawn by factual and counterfactual explanations and finally, what useful recommendations could be provided. For training the Siamese networks, 80% of each dataset's images were used for training and the other 20% for testing while preserving the variance of each class in each set. In addition, 10% of training images were used as a validation set for optimizing the network's performance. The implementation code along with the datasets can be found in https://github.com/ioannislivieris/Grad_CAM_Siamese.git (access in 14 August 2023).

Based on the images of each training dataset, we created the training pairs as follows: For each image, two images were randomly selected; one image from the same class and another image from a different class. The first pair containing the images from the same class was assigned with label zero (0), while the second pair containing the images from different classes was assigned with label one (1). Along this line, the similarity between two random input images is defined by $1 - d$, where d is the Siamese model's output. Notice that this methodology was initially proposed by Melekhov et al. [8].

At this point, it is worth mentioning that the model's prediction can be exploited to obtain information if two images belong to the same class or not. More specifically, if the prediction of the Siamese network for a pair of images is less than a pre-defined *threshold*, then the images are considered similar (belonging to the same class). Otherwise, they are considered dissimilar (belonging to the different classes). Notice that in our experiments *threshold* was set to 0.5.

Regarding the Siamese network architecture, ResNet50 [32] was used as a backbone network, followed by an average pooling layer of size of $(1, 1)$ and a dense layer of 256 neurons with ReLU activations for calculating each input image embedding. Next, the L_2-distance between the embeddings is calculated, followed by an output layer with one neuron with a Sigmoid activation function. The utilized architecture and hyperparameter selection provide us with a very good and reliable performance regarding all three benchmarks. It is worth highlighting the scope of this research was not to address a specific class of benchmarks, i.e., few-shot learning benchmarks, one-shot learning benchmarks, etc., neither to provide a new advanced model architecture but to provide human-meaningful explanations on similarity tasks through the proposed framework. Finally, the Siamese model was trained using the ADAM algorithm [33] while a contrastive loss function [34] was used for training the network, which is defined by

$$\mathcal{L} = \frac{1}{2}\Big[(1-y)(D_w)^2 + y\{max(0, m - D_w)\}^2\Big],$$

where D_w is the model's output and m is the margin value, which was set to 2.

4.1. Flowers Dataset

Next, we present an example from the application of the proposed framework on two random images (Figure 2a,d) from the Flowers dataset, which belongs to the same class ("rose"). The Siamese model's prediction was 0.24, which implies that the model predicts that the similarity between the input images is 76%. In addition, since the similarity score is greater than the pre-defined $threshold = 0.5$, the model suggests that input images belong to the same class. Figure 2b,e presents the factual explanations provided by Grad-CAM in order to identify the features, which impact the model's decisions. In more detail, the model's decision was based on the flower's blossoms in which it found common characteristics. As regards, the counterfactual explanations, which are presented in Figure 2c,f, they highlight that the model would have been based on the stems of both flowers for predicting that the images are not similar.

By taking into consideration that similar conclusions can be drawn by randomly selecting any pair of images in the Flowers dataset, a possible recommendation for improving the model's performance could be that the model is based on identifying the blossoms in

the input images for making its prediction; thus, a removal of other characteristics such as stems, background, etc., may improve the model's performance.

Figure 2. Application of the proposed framework on flowers dataset. (**a**) Original input $image_1$, (**b**) factual explanations on $image_1$, (**c**) counterfactual explanations on $image_1$, (**d**) original input $image_2$, (**e**) factual explanations on $image_2$, (**f**) counterfactual explanations on $image_2$.

4.2. Skin Cancer Dataset

Figure 3 presents the results from the application of the proposed framework on two random images from the skin cancer dataset, which belong to the same differences classes, i.e., the first image belongs to the "Benign" class, while the second one belongs to the "Malignant" class. The Siamese model's prediction was 0.781, which implies that the model predicts the similarity score 21.9% and that the input images belong to the different classes. Figure 3b,e presents the factual explanations provided by Grad-CAM in order to identify the features, which impact the model's decisions. The interpretation of Figure 3b suggests that the model focused on a small region on the skin while the interpretation of Figure 3e reveals that the model was focused on the tumor area. This implies that the model was focused on regions with dissimilar visual characteristics for predicting that the similarity score between the input images is considerably low. Furthermore, Figure 3c,f presents the counterfactual explanations that demonstrate the region of each image in which the model would have been based for predicting that the images are similar. Clearly, the highlighted areas in both images possess no similar visual characteristics.

Notice that although the input images look similar for a non-expert human, the tumor's characteristics such as texture, color and size are considered vital for separating benign from malignant cases. Therefore, a possible recommendation from this use case could be to use data augmentation based on transformation techniques (rotation, flip, crop, zoom, change the brightness, contrast, saturation, etc.) in order to improve the model's performance.

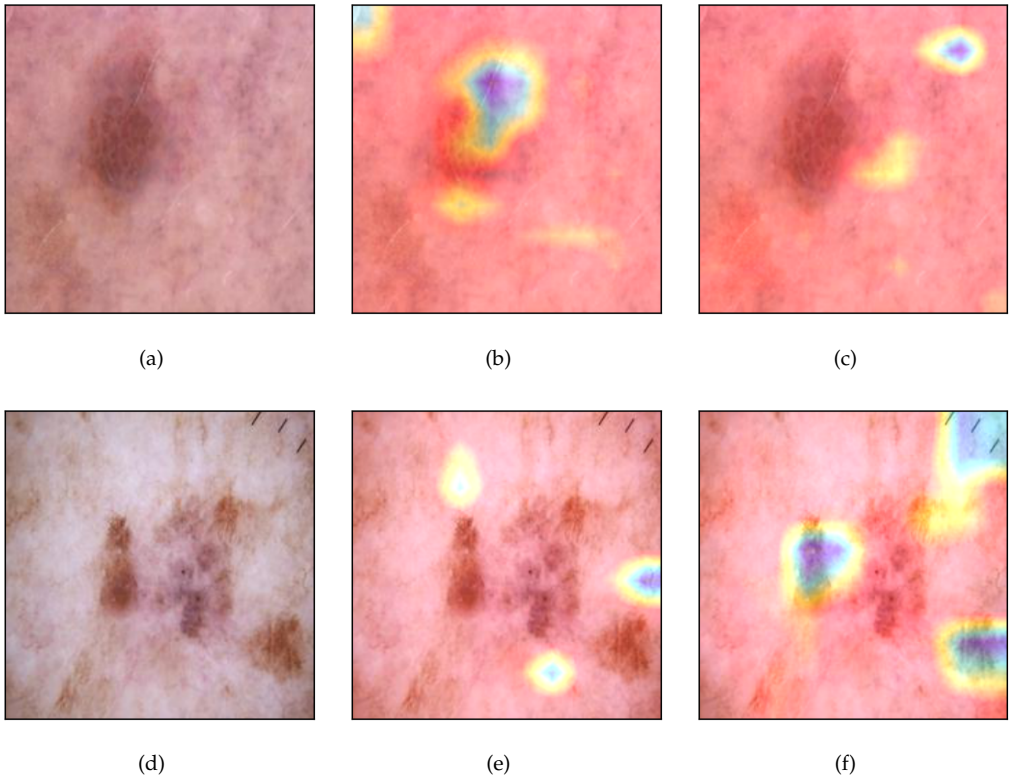

Figure 3. Application of the proposed framework on skin cancer dataset. (**a**) Original input $image_1$, (**b**) factual explanations on $image_1$, (**c**) counterfactual explanations on $image_1$, (**d**) original input $image_2$, (**e**) factual explanations on $image_2$, (**f**) counterfactual explanations on $image_2$.

4.3. Airbnb Dataset

Figure 4 presents the results from the application of the proposed framework on two random images from the AirBnB dataset, which belong to different classes, i.e., the first image belongs to the"bedroom" class, while the second one belongs to the "living-room" class. The Siamese model's prediction was 0.516, namely that the similarity score is 48.4%, which suggests that the model predicts that the input images marginally belong to different classes. Figure 4b,e presents the factual explanations provided by Grad-CAM, which suggest that the model was focused on the chairs presented in the first image and on several items in the second image (such as lamps, fire-place and clock) to predict that the images are marginally dissimilar.

Since the model's prediction is not very confident, it is wise to study the counterfactual explanations to explore why the model was near to being confused. Figure 4c,f presents the counterfactual explanations of both images, which suggest that the model, focused on the bed and sofa located in the first and second images, respectively, as well as the tables presented in both images. This implies that the model was nearly confused since both images possess a common item (table) as well as two items which are visually similar (bed and sofa).

A possible recommendation for improving the model's performance could be to use advanced image processing techniques for item identification in order to assist the model of correlating the items and/or furniture, which belong to each room.

Figure 4. Application of the proposed framework on AirBnB dataset. (**a**) Original input image$_1$, (**b**) factual explanations on image$_1$, (**c**) counterfactual explanations on image$_1$, (**d**) original input image$_2$, (**e**) factual explanations on image$_2$, (**f**) counterfactual explanations on image$_2$.

4.4. Improving the Siamese Model's Performance

In the rest of this section, we present an example of improving the performance of the Siamese model through the conclusions and recommendations, which could be provided from the application of the proposed framework.

Firstly, we recall that in the use case scenario performed on the Flowers dataset, we observed that by randomly selecting any pair of images which belong to the same class, the Siamese model focused on the blossoms for making its decision (Figure 2). Hence, a possible recommendation for improving the model's performance could be that the model was based on identifying the blossoms in the input images for making its prediction; thus, a removal of characteristics such as stems, background, etc., may improve the model's performance.

To examine the effectiveness of this approach, we create a new dataset in which each figure is replaced with a bounding box containing the flower's blossom. For calculating the bounding boxes for each image in the training data (anchor image), another image from the same class was randomly selected for calculating their similarity. In this case, their predicted similarity by the model was >80%, then we calculated the anchor's image Grad-CAM heatmap. Based on the calculated heatmap, we utilized the methodology and implementation of Cui et al. [35] for obtaining a bounding box, which contains the area

which was mostly focused on by the Siamese model for making its decision (i.e., flower's blossom). Along this line, in the newly created dataset, each image was replaced with the calculated bounding box. Figure 5 presents an example of the presented technique, i.e., the original image, the bounding boxes of Grad-CAM and the cropped image.

(a) (b) (c)

Figure 5. (**a**) Original image, (**b**) bounding box, (**c**) cropped image.

Table 1 presents the performance of the Siamese model of identifying similar and dissimilar pairs of instances (a) trained with the original Flowers dataset (b) trained with the "cropped" Flowers dataset in which each image has been replaced with the identified bounding box using the technique of Cui et al. [35]. The evaluation was performed on 432 pairs of similar and 432 pairs of dissimilar unseen images using accuracy, area under curve (AUC), precision and recall as performance metrics [36,37]. Clearly, we are able to conclude that the performance of the Siamese model was considerably increased relative to all performance metrics. In addition, the Siamese model achieved its top performance during the training process requiring less epochs in case it was trained with the "cropped" dataset.

Table 1. Siamese model's performance trained with the original and the "cropped" dataset.

Dataset	Accuracy	AUC	Precision	Recall
Original	87.15%	0.872	0.890	0.872
"Cropped"	88.31%	0.883	0.900	0.880

Figure 6 presents two pairs of (similar) images from the same class (Daisy). The first pair contains images from the original Flowers dataset while the second pair contains the corresponding "cropped" images. For the first pair, the Siamese model predicted a similar score equal to 18%, while for the second pair the model predicted a similar score equal to 11%.

(a)

Figure 6. *Cont.*

(b)

Figure 6. (**a**) Pair of images for the same class obtained from the original dataset, (**b**) corresponding cropped images.

Summarizing the previous discussion, we are able to conclude that the recommendation of removing characteristics such as stems, background and focusing on the blossoms considerably improved the quality of the dataset.

5. Discussion and Conclusions

The motivation for this research was to introduce the concept of explainable image similarity for providing useful, interpretable and transparent insights into the underlying factors driving image relationships and comparisons.

In this modern deep learning area, models are becoming more and more complex. They can exhibit high accuracy but their lack of transparency and explainability becomes crucial for building trust and understanding. The context of explainable image similarity aims to bridge the gap between the black-box nature of sophisticated similarity models and human interpretability. The main goal is not only the development of models which are able to provide accurate similarity scores between pairs of images but also to offer insights into the specific features, patterns or attributes that contribute to the computed similarity. By offering interpretable explanations, explainable image similarity not only enhances the usability of similarity-based applications, such as image retrieval and recommendation systems, but also empowers users to comprehend the reasoning behind the models' decisions, ultimately fostering informed and confident decision making.

For achieving this goal, we proposed a new framework by integrating Siamese networks together with the Grad-CAM technique. The former are used for calculating the similarity between input images while the latter is used for visualizing and interpreting the decisions made by convolutional-based Siamese neural networks. An attractive advantage of the proposed framework is that it is able to provide an image similarity score along with visual intuitive explanations for its decisions. In addition, the proposed framework is able to evaluate bias on the model's decisions as well as provide counterfactual explanations, highlighting the ability of the "what if/model's decisions". The presented use cases scenarios included the application of the proposed framework on three similarity tasks from different application domains (two classification datasets and a few-shot learning dataset). Notice that the scope of this research was not to address a specific class of benchmarks, i.e., few-shot learning benchmarks, one-shot learning benchmarks, etc., but to provide human-meaningful explanations on similarity tasks. Clearly, the proposed framework can easily be applied to any image similarity tasks as well as few-shot/one-shot image classification tasks providing similarity scores along with visual explanations about its decisions. The use cases scenarios along with the provided comprehensive discussion highlighted the need for explainable image similarity and the useful conclusions and recommendations, which can be provided by its application. Furthermore, we presented an example of improving the performance of the Siamese model for the Flowers use case scenario through the conclusions and recommendations provided from the application

of the proposed framework. In more detail, the provided recommendations resulted in increasing the model's accuracy by 1.2% and its prediction ability to identify similar images. For the Skin cancer and AirBnB use case scenarios, the recommendations for improving the models' performance were to use data augmentation based on transformation techniques (rotation, flip, crop, etc.) and image processing techniques for item identification in order to correlate the items and/or furniture, which belong to each room, respectively. Nevertheless, the former resulted in a minor improvement of the model's performance while the latter needs expert image processing and object identification techniques; hence, we decided to omit them.

It is worth mentioning that the proposed framework is based on the original Grad-CAM for providing visual explanations. Clearly, other state of the art techniques such as Grad-CAM++ [38], XGrad-CAM [39] and Score-Grad [40] can be easily adopted and incorporated. This can be considered a limitation of this work. Nevertheless, we should take into consideration that this was not the scope of this work. Another limitation can be considered the fact that the proposed framework uses a Siamese network with two input images. A possible extension could include the utilization of recent state-of-the-art models [41] with more advanced and complex architectures as well as the use of heatmap different saliency algorithms for heatmap calculation. Some interesting works presented by RichardWebster et al. [42] and Hu et al. [43] used and proposed several algorithms for calculating saliency algorithms. An adoption of the proposed approach to their frameworks could provide useful conclusions from the factual and counterfactual explanations.

Our future work is concentrated on the application of the proposed framework on real-world image similarity benchmarks and its usage in conjunction with non post hoc explainable techniques [11,16]. Since the presented conclusions from the presented use case scenarios are quite encouraging, we intent to proceed with studying the accuracy performance impact on similarity tasks through the adoption of the proposed framework and the utilization of advanced image processing techniques. Finally, another interesting idea could be the usage of advanced large language models for providing automated recommendations from the factual and/or counterfactual explanations [44,45]. Our expectation is that this research could be used as a reference for explainability frameworks, assisting decision making by providing useful visual insights and offering customized assistance and recommendations on image similarity-related tasks.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study are freely available in Kaggle. Flowers dataset: https://www.kaggle.com/datasets/alxmamaev/flowers-recognition; Skin cancer dataset: https://www.kaggle.com/datasets/emmanuelpintelas/segmentation-dataset-for-skin-cancer; Airbnb dataset: https://www.kaggle.com/datasets/barelydedicated/airbnb-duplicate-image-detection.

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

References

1. Gordo, A.; Almazan, J.; Revaud, J.; Larlus, D. End-to-end learning of deep visual representations for image retrieval. *Int. J. Comput. Vis.* **2017**, *124*, 237–254. [CrossRef]
2. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2874–2883.
3. Gygli, M.; Grabner, H.; Riemenschneider, H.; Van Gool, L. Creating summaries from user videos. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part VII 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 505–520.
4. Shin, H.C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med Imaging* **2016**, *35*, 1285–1298. [CrossRef] [PubMed]
5. Radenović, F.; Tolias, G.; Chum, O. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1655–1668. [CrossRef] [PubMed]
6. Chicco, D. Siamese neural networks: An overview. In *Artificial Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 73–94.
7. Appalaraju, S.; Chaoji, V. Image similarity using deep CNN and curriculum learning. *arXiv* **2017**, arXiv:1709.08761.
8. Melekhov, I.; Kannala, J.; Rahtu, E. Siamese network features for image matching. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 378–383.
9. Rossi, A.; Hosseinzadeh, M.; Bianchini, M.; Scarselli, F.; Huisman, H. Multi-modal siamese network for diagnostically similar lesion retrieval in prostate MRI. *IEEE Trans. Med Imaging* **2020**, *40*, 986–995. [CrossRef]
10. Selbst, A.D.; Barocas, S. The intuitive appeal of explainable machines. *Fordham L. Rev.* **2018**, *87*, 1085. [CrossRef]
11. Pintelas, E.; Liaskos, M.; Livieris, I.E.; Kotsiantis, S.; Pintelas, P. Explainable machine learning framework for image classification problems: Case study on glioma cancer prediction. *J. Imaging* **2020**, *6*, 37. [CrossRef]
12. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
13. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
14. Wachter, S.; Mittelstadt, B.; Floridi, L. Transparent, explainable, and accountable AI for robotics. *Sci. Robot.* **2017**, *2*, eaan6080. [CrossRef]
15. Livieris, I.E.; Karacapilidis, N.; Domalis, G.; Tsakalidis, D. An advanced explainable and interpretable ML-based framework for educational data mining. In Proceedings of the 13th International Conference on Methodologies and Intelligent Systems for Technology Enhanced Learning, Guimaraes, Portugal, 12–14 July 2023.
16. Pintelas, E.; Liaskos, M.; Livieris, I.E.; Kotsiantis, S.; Pintelas, P. A novel explainable image classification framework: Case study on skin cancer and plant disease prediction. *Neural Comput. Appl.* **2021**, *33*, 15171–15189. [CrossRef]
17. Samek, W.; Wiegand, T.; Müller, K.R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv* **2017**, arXiv:1708.08296.
18. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
19. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
20. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
21. Ma, J.; Jiang, X.; Fan, A.; Jiang, J.; Yan, J. Image matching from handcrafted to deep features: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 23–79.
22. Pintelas, E.; Livieris, I.E.; Pintelas, P.E. A convolutional autoencoder topology for classification in high-dimensional noisy image datasets. *Sensors* **2021**, *21*, 7731. [CrossRef] [PubMed]
23. Pintelas, E.; Livieris, I.E.; Kotsiantis, S.; Pintelas, P. A multi-view-CNN framework for deep representation learning in image classification. *Comput. Vis. Image Underst.* **2023**, *232*, 103687. [CrossRef]
24. Hsiao, C.T.; Lin, C.Y.; Wang, P.S.; Wu, Y.T. Application of convolutional neural network for fingerprint-based prediction of gender, finger position, and height. *Entropy* **2022**, *24*, 475. [CrossRef]
25. Kim, S.H.; Park, J.S.; Lee, H.S.; Yoo, S.H.; Oh, K.J. Combining CNN and Grad-CAM for profitability and explainability of investment strategy: Application to the KOSPI 200 futures. *Expert Syst. Appl.* **2023**, *225*, 120086. [CrossRef]
26. Singh, A.; Sengupta, S.; Lakshminarayanan, V. Explainable deep learning models in medical image analysis. *J. Imaging* **2020**, *6*, 52. [CrossRef]

27. Saeki, M.; Ogata, J.; Murakawa, M.; Ogawa, T. Visual explanation of neural network based rotation machinery anomaly detection system. In Proceedings of the 2019 IEEE International Conference on Prognostics and Health Management (ICPHM), San Francisco, CA, USA, 17–20 June 2019; pp. 1–4.
28. Gupta, R.T.; Mehta, K.A.; Turkbey, B.; Verma, S. PI-RADS: Past, present, and future. *J. Magn. Reson. Imaging* **2020**, *52*, 33–53. [CrossRef]
29. Neculoiu, P.; Versteegh, M.; Rotaru, M. Learning text similarity with siamese recurrent networks. In Proceedings of the 1st Workshop on Representation Learning for NLP, Berlin, Germany, 11 August 2016; pp. 148–157.
30. Guo, Z.; Arandjelović, O.; Reid, D.; Lei, Y. A Siamese Transformer Network for Zero-Shot Ancient Coin Classification. *J. Imaging* **2023**, *9*, 107. [CrossRef]
31. Mazzeo, P.L.; Libetta, C.; Spagnolo, P.; Distante, C. A siamese neural network for non-invasive baggage re-identification. *J. Imaging* **2020**, *6*, 126. [CrossRef]
32. 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, Nevada, USA, 27–30 June 2016; pp. 770–778.
33. Reddi, S.J.; Kale, S.; Kumar, S. On the convergence of adam and beyond. *arXiv* **2019**, arXiv:1904.09237.
34. Wang, F.; Liu, H. Understanding the behaviour of contrastive loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2495–2504.
35. Cui, X.; Wang, D.; Wang, Z.J. CHIP: Channel-wise disentangled interpretation of deep convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 4143–4156. [CrossRef] [PubMed]
36. Hossin, M.; Sulaiman, M.N. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 1.
37. Livieris, I.E.; Kotsilieris, T.; Tampakas, V.; Pintelas, P. Improving the evaluation process of students' performance utilizing a decision support software. *Neural Comput. Appl.* **2019**, *31*, 1683–1694. [CrossRef]
38. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
39. Fu, R.; Hu, Q.; Dong, X.; Guo, Y.; Gao, Y.; Li, B. Axiom-based Gram-CAM: Towards accurate visualization and explanation of CNNs. *arXiv* **2020**, arXiv:2008.02312.
40. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 24–25.
41. Hu, B.; Vasu, B.; Hoogs, A. X-MIR: EXplainable Medical Image Retrieval. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE Computer Society, Waikoloa, HI, USA, 3–8 January 2022; pp. 1544–1554.
42. RichardWebster, B.; Hu, B.; Fieldhouse, K.; Hoogs, A. Doppelganger Saliency: Towards More Ethical Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 2847–2857.
43. Hu, B.; Tunison, P.; RichardWebster, B.; Hoogs, A. Xaitk-Saliency: An Open Source Explainable AI Toolkit for Saliency. In Proceedings of the AAAI Conference on Artificial Intelligence, Arlington, VA, USA, 25–27 October 2023; Volume 37, pp. 15760–15766.
44. Peng, B.; Li, C.; He, P.; Galley, M.; Gao, J. Instruction tuning with GPT-4. *arXiv* **2023**, arXiv:2304.03277.
45. Topsakal, O.; Akinci, T.C. Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. In Proceedings of the International Conference on Applied Engineering and Natural Sciences, Konya, Turkey, 10–12 July 2023; Volume 1, pp. 1050–1056.