electronics

MDPI

Article

Assessing Biases through Visual Contexts

Anna Arias-Duart [1,2,*], Victor Gimenez-Abalos [1,2], Ulises Cortés [1,2] and Dario Garcia-Gasulla [1,2]

[1] Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain; victor.gimenez@bsc.es (V.G.-A.); ia@cs.upc.edu (U.C.); dario.garcia@bsc.es (D.G.-G.)
[2] Department of Computer Science (CS), Universitat Politècnica de Catalunya (UPC)-BarcelonaTECH, 08034 Barcelona, Spain
* Correspondence: anna.ariasduart@bsc.es

Abstract: Bias detection in the computer vision field is a necessary task, to achieve fair models. These biases are usually due to undesirable correlations present in the data and learned by the model. Although explainability can be a way to gain insights into model behavior, reviewing explanations is not straightforward. This work proposes a methodology to analyze the model biases without using explainability. By doing so, we reduce the potential noise arising from explainability methods, and we minimize human noise during the analysis of explanations. The proposed methodology combines images of the original distribution with images of potential *context* biases and analyzes the effect produced in the model's output. For this work, we first presented and released three new datasets generated by diffusion models. Next, we used the proposed methodology to analyze the context impact on the model's prediction. Finally, we verified the reliability of the proposed methodology and the consistency of its results. We hope this tool will help practitioners to detect and mitigate potential biases, allowing them to obtain more reliable models.

Keywords: bias detection; mosaics; diffusion models; biased dataset; shortcuts; context biases

check for updates

Citation: Arias-Duart, A.; Gimenez-Abalos, V.; Cortés, U.; Garcia-Gasulla, D. Assessing Biases through Visual Contexts. *Electronics* **2023**, *12*, 3066. https://doi.org/10.3390/electronics12143066

Academic Editor: Dah-Jye Lee

Received: 15 June 2023
Revised: 6 July 2023
Accepted: 8 July 2023
Published: 13 July 2023

1. Introduction

The presence of *bias* in models and datasets is often inherent to their construction. Those biases can be desirable, useful and harmless, or undesirable, useless and harmful. To build reliable and *fair* models, we must develop tools that facilitate bias detection, so that experts can decide if the found biases belong to the desirable or to the undesirable category and then take the measures that they deem appropriate.

Suresh and Guttag distinguished seven sources of harm in Machine Learning (ML) [1]: historical bias; representation bias; measurement bias; aggregation bias; learning bias; evaluation bias; and deployment bias. In the Computer Vision (CV) field, powered by ML, the most common sources are *representation* and *evaluation* biases: the former is due to the need for large datasets that often lack representativeness [2]; the latter is due to a lack of robust evaluations determining the model's ability to correctly generalize in real-world data.

Current data collection methods lead to non-random selection and make the data unrepresentative of the total population. In this regard, many examples already exist in the Deep Learning (DL) and CV literature. Shankar et al. showed the lack of geo-representation within the ImageNet [3] dataset (e.g., 45.5% of the images were from the United States [4]). Many examples exist of systems which are racially and gender biased, such as the three commercial gender classifiers tested in [5], which performed better for white males. The worst performance was obtained when classifying black women (i.e., belonging to two underrepresented populations: women and black). When we use these biased datasets in sensitive fields such as medicine, the consequences can be deadly: for example, diagnosing skin cancer in later stages in patients with dark skin tones [6].

The previous examples are also a symptom of *evaluation* biases, as a proper assessment would have highlighted the biased behavior of the model and prevented its release. As introduced before, this second bias is common in the ML–CV field. On the one hand, the model performance is usually evaluated in a test partition, distinct from the training and validation sets; however, a random split is typically obtained from the same distribution. Thus, the model is not being evaluated for generalization capability. On the other hand, as also pointed out in [1], the choice of the evaluation metrics could be another source of *evaluation* bias: for example, choosing the method with the best accuracy does not ensure that the method is capable of generalizing better to real-world data or that the method is less biased.

The combination of *representation* biases and *evaluation* biases results in unsafe models with an unknown amount of undesirable biases. Furthermore, failure to detect these biases can lead models to perpetuate and/or exacerbate inequalities. To prevent this, we need methodologies for identifying and illustrating biases, which experts can use to search and select biases in CV models.

Motivated by these needs, the eXplainable Artificial Intelligence (XAI) field has attracted attention in recent years, becoming a tool to provide insights into DL models' behavior. Many explainability techniques already exist in the CV field [7–10], the goal of which is to pinpoint the input image features with the greatest relevance for model predictions. The main objective of these explanations is to better understand the model's behavior, verify its performance and ensure its alignment with expert knowledge. This is why several works rely on these XAI methods, to validate their models and establish trust in the decision making process: for instance, in the medical domain, examples using XAI for model validation can be found across various disciplines and tasks, such as embryo quality grading, COVID-19 classification using X-ray images, chronological age estimation based on orthopantomogram (OPG) images or predicting the breast cancer response to chemotherapy using magnetic resonance imaging (MRI) [11–15]. While the use of these XAI techniques may be useful for understanding which parts of the input image contributed to the prediction, they present two main limitations, which diminish their reliability and their effectiveness in detecting biases. Firstly, the explanations obtained by each XAI method are different, and their reliability requires its own assessment [16–19]. This makes potential biases inconsistent, as they may depend on the choice of the explainability technique. Secondly, biases typically need to be defined, identified and verified by an expert, which is a time-consuming task that can induce subjective criteria and confirmation bias.

One of the tools proposed in the literature to mitigate the latter limitation is the *mosaics* methodology [19,20], which is a composition of images that helps to identify and illustrate model biases in a semi-automatic manner. Mosaics can be used in different ways, depending on the types of biases we are looking for. In this work, we propose a methodology targeting *context* biases: those which influence the prediction of a given item through the visual properties of its context, instead of those from the item itself.

The motivation of this work was to explore whether it is possible to analyze and measure the impact of context bias on model predictions using mosaics, without relying on explainability methods. The underlying idea is that explainability can be misleading, and the bias identification may be subjective (requiring human intervention); therefore, by removing explainability from the process, and solely analyzing the model's predictions when combining images with potential *context* biases, the results regarding the influence of these contexts will be more reliable.

The first contribution of this work is the three new datasets built and publicly released to carry out these experiments, synthetically generated by a stable diffusion model. To facilitate the reproduction of experiments and the creation of education and dissemination material, this work is released together with a notebook which illustrates the use and potential of the approach: https://www.kaggle.com/code/annaariasduart/assessing-biases-through-visual-contexts (accessed on 14 June 2023). The second contribution is the methodology itself: the *contextualized* mosaics. Both contributions are introduced in

Section 2. Then, the methodology is applied, and the findings are presented in Section 3. The document concludes with a discussion of the results and potential future work, in Section 4.

2. Materials and Methods

The methodology proposed to detect biases is based on the idea of using mosaics built by combining images of the original data distribution with images of potential biases. This idea has the advantage of adding in-distribution noise: that is, expanding the input images with content that belongs to the same distribution as the original data, which leads to more realistic and consistent model behavior. This section introduces the different elements needed to perform the experiments (Section 2.1) and the methodology proposed (Section 2.2).

2.1. Experimental Design

First, we present the new synthetic datasets created for this work (Section 2.1.1), then the training configurations used (Section 2.1.2) and, finally, we evaluate the generalization capabilities of all the trained models (Section 2.1.3).

2.1.1. Dataset

An image diffusion model was used to create the datasets employed in this work. With these diffusion models, one can specify what to generate and guide it to produce realistic images. The model used in this work is a text-to-image diffusion model [21]: from a text prompt, the model generates truthful images and, at the same time, is faithful to the text.

We generated three different datasets (see Figure 1), each one composed of four classes: *bench*; *fire hydrant*; *plane*; and *mug*, all of which are publicly available. The following are the generation details:

Figure 1. Sample instances by class and dataset. Each dataset is shown in a different column (from left to right): context (*C*); no context (*NC*); and white background (*WB*) dataset. Class examples are separated by row (from top to bottom): bench; plane; fire hydrant; and mug.

1. **Context (*C*)**: This dataset comprised images corresponding to the four objects in a typical context, according to the model's representation. The exact prompt used to generate these images was A GREEN CLASS ON THE FOREGROUND. TYPICAL BACKGROUND. For each class, the word CLASS was replaced by the object: bench; fire hydrant; plane; or mug. Avaialable at https://storage.hpai.bsc.es/object-datasets/context.zip (accessed on 14 June 2023).
2. **No Context (*NC*)**: This dataset contained images of the same four objects, but without a background. To that end, we slightly changed the prompt and asked for a sketch of the object with a uniform background. The exact prompt used to generate these images was NO BACKGROUND. SIMPLE SKETCH OF A GREEN CLASS. Available at https://storage.hpai.bsc.es/object-datasets/no_context.zip (accessed on 14 June 2023).
3. **White Background (*WB*)**: For creating this dataset, we manually removed the background of the *C* dataset images, using a tool designed and provided by Adobe at https://www.adobe.com/express/ (accessed on 14 June 2023); therefore, this dataset was composed of the same images as the *C* dataset, but setting the background to white. Available at https://storage.hpai.bsc.es/object-datasets/white_background.zip (accessed on 14 June 2023).

Note that for each class of each dataset, 150 images were created. In order to prevent the model from learning to differentiate these classes by their recurring colors (e.g., most fire hydrants are red) or by texture (e.g., benches are often made of knotted wood), we set the color of the four objects to green.

2.1.2. Training Setup

We trained six different models, using the three datasets introduced above. Due to the simplicity of the datasets, we used the AlexNet [22] architecture, a shallow architecture that could fit our data. Each dataset was used to train two models: one from scratch, and one pre-trained on ImageNet [3] and then fine-tuned for it. For the pre-trained models, we used the AlexNet model available in the *torchvision.models* subpackage at https://download.pytorch.org/models/alexnet-owt-4df8aa71.pth (accessed on 14 June 2023). We used a total of 100 images per class for training, 25 for validation and 25 for the test. To avoid confusion, we will refer to the models as follows:

1. **model-*C***: model trained from scratch on the *C* dataset;
2. **model-*NC***: model trained from scratch on the *NC* dataset;
3. **model-*WB***: model trained from scratch on the *WB* dataset;
4. **pt-*C***: model pre-trained on ImageNet and fine-tuned on the *C* dataset;
5. **pt-*NC***: model pre-trained on ImageNet and fine-tuned on the *NC* dataset;
6. **pt-*WB***: model pre-trained on ImageNet and fine-tuned on the *WB* dataset.

2.1.3. Cross Evaluation

Potential biases that may have appeared in the previously trained models actually originated in the diffusion model, and were then recreated in the dataset and, finally, learned by the models. If we had evaluated these models in a random split of the same dataset in which it had been trained, we would probably have obtained a high performance, even though these models had not learned the features of the four objects. However, by testing the models in a partition of the other two datasets (both having the same four classes), we could evaluate the model's generalization capabilities.

To do so, we cross-evaluated all six models with all three datasets. The accuracies and cross-accuracies obtained by the different models are illustrated in Figure 2. Each histogram (i.e., group of three bars) corresponds to one of the six models. Each bar corresponds to the accuracy obtained by using a different test set. Each set is shown with a different color: yellow corresponds to the *NC* set; gray to the *C*; and green to the *WB* set.

First of all, the models trained from a random state (the first three columns of Figure 2) performed more poorly on all test settings than the pre-trained models (the last three columns of Figure 2). The best accuracies of the non-pre-trained models (model-*NC*, model-*C* and model-*WB*) were achieved when using the same distribution as for training (i.e., 98%, 100% and 99%, respectively); however, these models struggled to correctly distinguish the classes when using cross-tests.

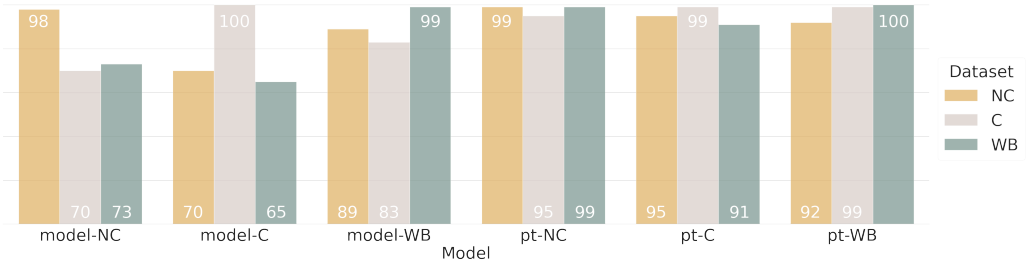

Figure 2. Accuracies obtained by the six models (model-*NC*, model-*C*, model-*WB*, pt-*NC*, pt-*C* and pt-*WB*) on the three test sets (*NC*, *C*, *WB*). Each set is represented by a different color, and the results for each model are grouped in the form of a histogram (a group of three bars).

By contrast, the pre-trained models managed to generalize much better than the models trained from scratch: the differences between the bars of histograms 4, 5 and 6 were less prominent than the differences seen in the first three histograms. As expected, using pre-trained models prevented the model from learning patterns—or, to use a better term, shortcuts—that were present in the small training dataset, but which were not the patterns expected to be learned by the model. In addition, we note that the plane, the mug and the bench were also classes of the ImageNet dataset (check https://cs.stanford.edu/people/karpathy/ilsvrc/ (accessed on 14 June 2023)), which means that the pre-trained models knew the visual features needed to identify the different classes before the fine-tuning process.

An interesting finding in this first analysis was the relevant role played by the context. The only model that performed consistently well on all distribution shifts for non-pre-trained models was the one trained by *WB*, obtaining an accuracy of 89% when using the *NC* and 83% when using the *C*. This lower performance, when using the *C* set—despite being exactly the same objects but with white background—was most likely due to the large distribution shift that the presence of context (i.e., patterns surrounding the objects) supposes for a model not trained with a background. The same was observable for the models trained by *NC*, where the worst performance was obtained by the *C* set (the first and fourth columns of Figure 2).

On the other hand, the model trained from scratch with context images was the one that generalized the worst, obtaining a performance of 70% when tested by *NC* and 65% accuracy when using the *WB*. The *C* model may have learnt contextual biases. Thus, the model did not maintain the performance when those context features were not present (e.g., within the *NC* and *WB* sets).

2.2. Context Biases and Contextualized Mosaics

As previously explained, in this work we focused on *context* biases: if each context is specific to each class and is not found in the the other classes, the model will learn those contexts as shortcuts. As we empirically stated in the previous section, these learned shortcuts lead to biased models that are not able to generalize correctly.

To formalize the model predictive behavior as a desirable causal model [23], we built a Directed Acyclic Graph (DAG) to represent the problem. Each node of this DAG represented the object present in the images (O), the context (C) and the predicted class (Y). The desirable setting representing the relationship between these nodes is shown in Figure 3a. However, we propose that the graph learned by the model was more similar to the one shown in

Figure 3b. To assess the relationship between C and Y—that is, the relationship between the context and the class predicted by the model (e.g., the relationship between the vegetation context for the bench class)—we performed an intervention fixing C=c, with four possible alternatives (i.e., one context per object): $do(C = c_1)$; $do(C = c_2)$; $do(C = c_3)$; and $do(C = c_4)$. The new graph after the intervention was the one shown in Figure 3c. To perform the intervention, we constructed what we termed *contextualized* mosaics.

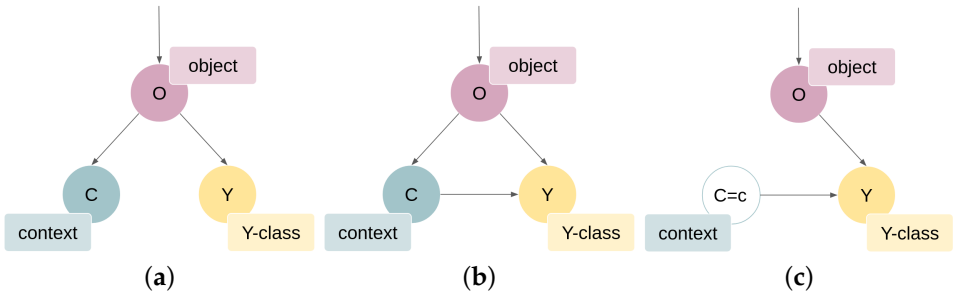

Figure 3. (**a**) desired causal model representing the relationships between the object within the images (O), the context (C) and the predicted class (Y); (**b**) actual graph learned by the model; (**c**) the modified graph after the intervention (fixing C).

To rationalize the *contextualized* mosaics, let us first investigate the mosaic concept [19]. In short, given an image dataset $\mathbb{I} = \{img_1, img_2, img_3 \dots img_N\}$, each image having an assigned label from the set of classes $\mathbb{L} = \{l_1, l_2, l_3 \dots l_M\}$ and being $N > M$, each mosaic will be composed of J images arranged within a grid: for example, a mosaic of size $J = 2$ will be composed of two images, $m = \{img_1, img_2\}$, belonging to two different classes $l(img_1) \neq l(img_2)$. The way the mosaics are built will depend on the objective of the experiment: for example, if we want to analyze the bias between two classes (e.g., planes and benches), we can create mosaics of size 1×2, combining images from both classes (see Figure 4a). We can also increase the number of images per class, to increase the probability that the bias will arise (see Figure 4b: two mugs versus two fire hydrant images). We might be interested in evaluating if the model has learned some patterns of one class (e.g., plane) better than others (e.g., mug), for which purpose, we could build a 2×2 mosaic with three mugs versus one plane (see Figure 4c). Mosaics can be useful for detecting biases, due to their inherent properties. On the one hand, mosaics maintain the visual features from the original distribution, which reduces the noise induced and facilitates the identification of biases. On the other hand, mosaics introduce a source of confusion in a controllable and scalable manner (mosaics are custom and easy to generate), being able to challenge the model.

Figure 4. Examples of mosaics with different configurations: (**a**) mosaic of size 1×2; (**b**) mosaic of size 2×2, combining two mugs and two fire hydrants; and (**c**) mosaic of size 2×2, combining three mugs versus one plane.

For the *contextualized* mosaics, apart from the set of images $\mathbb{I}$ and the set of classes $\mathbb{L}$, we also had a set of contexts $\mathbb{C} = \{ctx_1, ctx_2, ctx_3 \dots ctx_M\}$, each context composed of a set of context images. Note that there was the same number of contexts as classes: that is, one potential bias context per class. For the mosaic construction, we set $J = 2$ (mosaics of size 1×2), where each mosaic was made up of an image and a context $m = \{img, ctx\}$, where $l(img) \neq l(ctx)$. In other words, the class assigned to the context was different from the class of the image (e.g., a mug image was combined with a sky image, the sky image being assigned the plane class, because it was a typical context of the plane and not of the mug).

Let us investigate how we built these *contextualized* mosaics. The mosaic design was based on the assumption that the target biases were known beforehand: this was a realistic assumption, as the domain expert should have prior knowledge of the possible biases that could exist. In our experiment, we could replace expert knowledge by analyzing the text prompt used during the *C* dataset generation. The text prompt included the sentence: TYPICAL BACKGROUND. From there, we observed the typical contexts where the four objects (i.e., bench, plane, fire hydrant and mug) were usually found, according to the generative model used. We used the same diffusion model to obtain the context images with which we generated the training datasets. The selected context per object, and the prompts used to generate those images, were as follows:

- A park for the *bench* class: A PARK WITH VEGETATION;
- Sky for the *plane* class: A CLEAR BLUE SKY;
- A road for the *fire hydrant* class: A REALISTIC TARRED ROAD IN A CITY;
- A piece of wood for the *mug* class: A PIECE OF WOOD.

Once these context images had been generated (see Figure 5), we built the mosaics, by combining the original images of the different objects from the test set within the different contexts within a 1×2 grid. For each of the 25 object images, we combined them with five different samples obtained for each of the three contexts not belonging to that class: that is, for an image of a plane, we combined it with 5 park images, 5 road images and 5 wood images. This resulted in a total of 1500 mosaics. Examples of these mosaics are shown in Figure 6.

(**a**) (**b**) (**c**) (**d**)

Figure 5. Examples of context samples generated for each class: (**a**) a park for the *bench* class (A PARK WITH VEGETATION); (**b**) a sky image for the *plane* class (A CLEAR BLUE SKY); (**c**) a road for the *fire hydrant* class (A REALISTIC TARRED ROAD IN A CITY); (**d**) a piece of wood for the *mug* class (A PIECE OF WOOD).

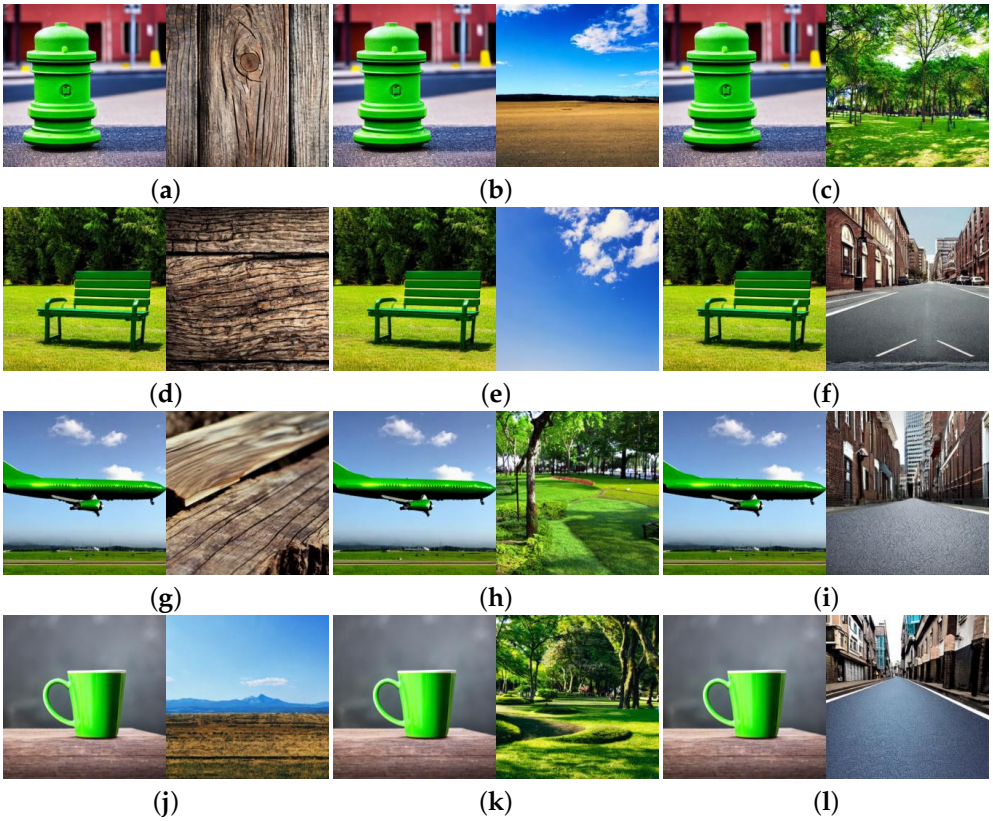

Figure 6. *Contextualized* mosaic examples, of size 1×2, composed of object images with context images. Each row shows mosaics built from the same object image within different contexts. The first row corresponds to a fire hydrant image within (**a**) a wood context, (**b**) a sky context and (**c**) a park context. The second row corresponds to mosaics built with a bench image within (**d**) a wood context, (**e**) a sky context and (**f**) a road context. The third row is made up of a plane image within (**g**) a wood context, (**h**) a park context and (**i**) a road context. In the last row, mosaics of a mug image within (**j**) a sky context, (**k**) a park context and (**l**) a road context are shown.

3. Results

Let us analyze the results obtained using *contextualized* mosaics, to assess the relevance of the context in the predictions of model-*C*, the model that may have learned context biases. When building the mosaics, combining the original images with the context images, we induced a source of confusion for the model. We assessed the impact of this noise (i.e., context noise) by comparing the model's output of the class object image to the model's output produced by the same image when composed in a mosaic within a context image. The results of this analysis are shown in Figure 7, and the code to reproduce these results is publicly available at https://github.com/HPAI-BSC/Assessing-Biases (accessed on 14 June 2023). For each possible combination of <class, context>, as long as $l(img) \neq l(ctx)$, we show a 1D and a 2D histogram. The 2D histogram color intensity represents frequency. The color codes used were green for bench images or mosaics within a park context, orange for the mug images or mosaics within a wood context, blue for the plane images or mosaics within a sky context and gray for the fire hydrant images or mosaics within road context. We observed the change in model probabilities induced by attaching a given context to an image of a given class. The larger this change was, the stronger the bias. To obtain a better understanding of Figure 7, we represent, in Figure 8, examples of images and mosaics utilized to obtain the first row results in Figure 7 (i.e., using the sky context). The rectangle colors in Figure 8 were also aligned to the colors used in Figure 7. Let us now analyze the results in Figure 7 by context (i.e., by row).

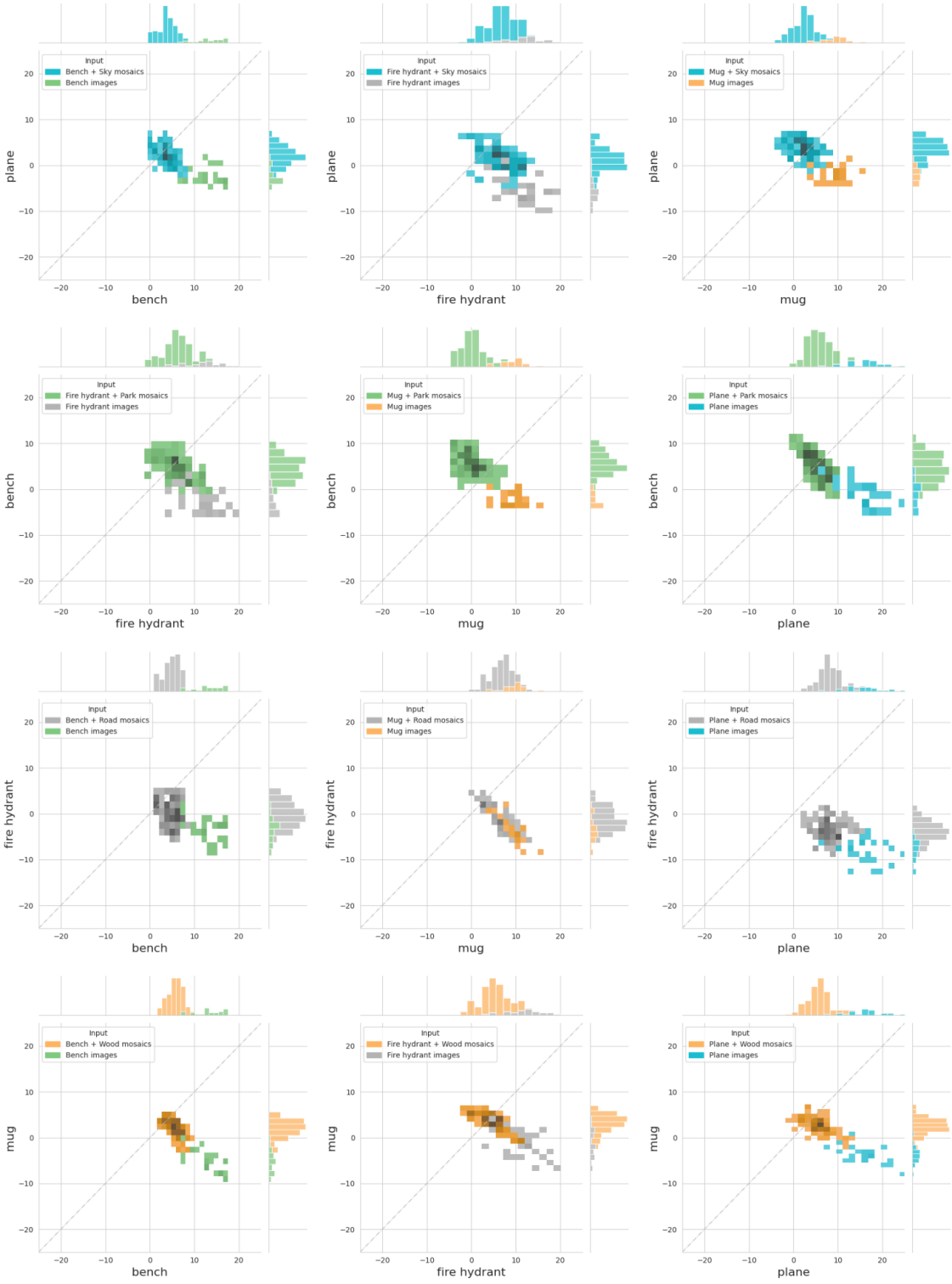

Figure 7. The logits obtained with the original images, with respect to those obtained with the images combined with the different contexts, are shown as histograms. In each row, the results are displayed for the mosaics composed of sky images (in the first row), park images (in the second row), road images (in the third row) and wood images (in the fourth row). The color code used is as follows: blue for mosaics with sky images and for plane images; green for mosaics with park images and for bench images; gray for wood mosaics and for fire hydrant images; and orange for mosaics with wood context and mug images.

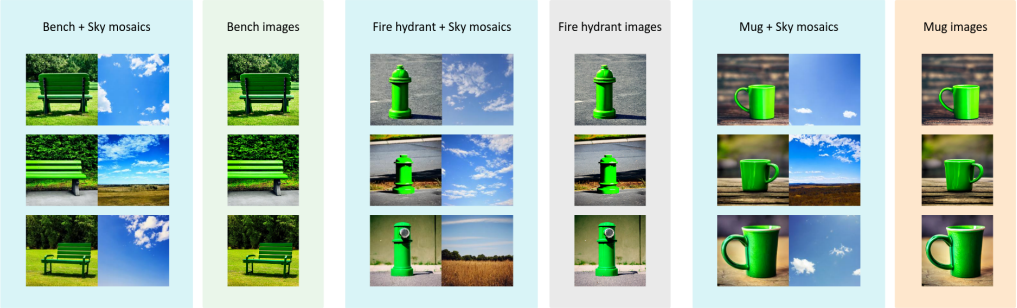

Figure 8. Examples of images and mosaics used to obtain the results of the first row of Figure 7: that is, using the sky context. The rectangle colors are the same as those used in Figure 7. As these mosaics were built with the sky context images (typical of the *plane* class), they are represented in blue. The single images of benches are in green, those of fire hydrants are in gray and those of mugs are in orange.

Sky. According to these results, the sky seemed to be a relevant feature learned by the model. When the bench images were combined with the sky context, the sky confused the model, obtaining higher logits for the *plane* class than for the *bench* class in 44 mosaics out of 125 mosaics (i.e., 35.2%). In the case of the *fire hydrant*, the sky context also increased the logits of the *plane* class; however, the *fire hydrant* image features seemed to be more relevant for the model, maintaining in 107 mosaics the highest logit values for the *fire hydrant* class. When combining the mug with the sky, the model was completely confused, the logits for the *plane* class being higher than the *mug* logits in more than half of the mosaics (64%).

Park. Vegetation also appeared to be an important feature for the *bench* class, to an even greater extent than the sky context for the *plane* class. Although the lowest impact was for the mosaics with *fire hydrant* images, the park context still had a huge influence, the *bench* logits being higher than the *fire hydrant* logits in 50 mosaics. The park context also impacted the mug results, managing to drastically shift the predictions towards the *mug* class in 115 mosaics (i.e., 92%). In the plane case, 58 mosaics out of 125 obtained higher logits for the *bench* class (i.e., 46.4%).

Road. This context did not seem to be as decisive as the two previous contexts (i.e., the sky and park contexts). Combining the road context with the three classes (i.e., bench, mug or plane) had a lesser impact on the prediction: less than 12% of the mosaics were predicted as *fire hydrants*. For the *bench* class, the road presence slightly reduced the confidence towards the *bench* class, obtaining higher logits for the *fire hydrant* in 14 out of 125 mosaics. In the mug case, only 9 mosaics obtained *fire hydrant* logits greater than *mug* logits. Finally, the road context combined with plane images did not change the prediction of any mosaic. This was likely due to the presence of the sky (which was a very weighty feature for the model) in the plane images, thus favoring the *plane* class.

Wood. Although not as influential as the sky and the park contexts, the wooden context did seem to have a greater impact than the road context. When combined with bench images, the wooden context increased the logits value towards the *mug* class, obtaining 22 out of 125 mosaics higher logits for the *mug* class than for the *bench*. The impact was greater when combined with the *fire hydrant* class, obtaining higher logits for the mug class in 38 mosaics (i.e., 30.4%). And 28 mosaics combined with *plane* images obtained higher logits in the *mug* class.

In short, the main findings of these results are as follows. The sky seemed to be a relevant feature for the model when predicting the *plane* class. The park context (i.e., the vegetation) was clearly a relevant characteristic for the prediction of the *bench* class: this could have been a shortcut learned by the model. The road context was a characteristic that favored the *fire hydrant* class, but did not seem to be so determinant for its prediction. Finally, the wood context was influential for the prediction of the *mug* class, although not as relevant as the sky for the *plane* class or the park for the *bench* class. We also observed that depending on the object class with which the contexts were combined, they had a

greater or lesser effect: for example, we observed that when contexts were combined with fire hydrant images, the bias effect was lower; and we observed that the sky present in the plane images within the mosaics continued to favor the *plane* class.

Another way to attain these findings is by analyzing the results of the *contextualized* mosaics, using Table 1. This table shows the Euclidean distance of the means of the logit distributions of Figure 7 to the diagonal (see Figure 9, for a better comprehension of the steps followed). The diagonal was the point where the logits towards the two classes coincided. To better understand the results, we showed the distances corresponding to the means in the object class part as positive, and as negative when the mean was in the context class part: that is, the more negative the value, the more predictions there would be towards the context class. An illustration of these results is also shown in Figure 10.

Table 1. Euclidean distance of the distributions means to the diagonal for each of the combinations shown in Figure 7. The diagonal corresponded to the point where the logit values for the two classes coincided. For better interpretation, when the mean was located on the right of the diagonal, we present the results as positive. Conversely, if the mean of the logits distribution was located on the left of the diagonal, we show the results as negative. The mosaics with benches are highlighted in green, the mosaics with fire hydrants in gray, the mosaics with planes in blue and the mosaics with mugs in orange.

Bench + Sky mosaics	0.7165	Fire hydrant + Sky mosaics	2.5436	Mug + Sky mosaics	−0.7512
Bench images	7.9297	Fire hydrant images	8.7015	Mug images	5.6778
Fire hydrant + Park mosaics	0.7191	Mug + Park mosaics	−2.8234	Plane + Park mosaics	0.3372
Fire hydrant images	7.3755	Mug images	5.8023	Plane images	8.6055
Bench + Road mosaics	2.3406	Mug + Road mosaics	3.7252	Plane + Road mosaics	5.6887
Bench images	8.4977	Mug images	6.4769	Plane images	12.0303
Bench + Wood mosaics	1.7755	Fire hydrant + Wood mosaics	0.8831	Plane + Wood mosaics	1.4188
Bench images	9.2526	Fire hydrant images	6.6443	Plane images	9.8776

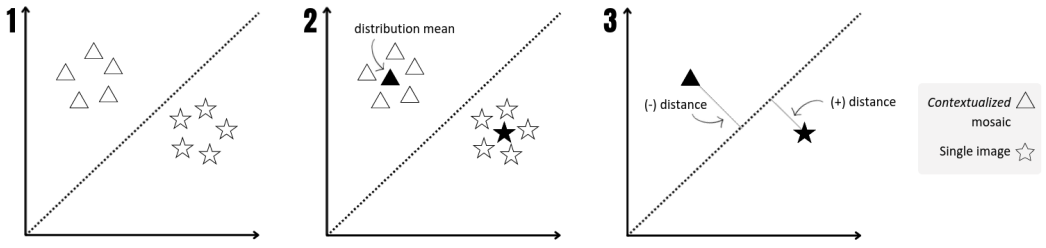

Figure 9. Steps to compute the distances shown in Table 1 and Figure 10. In the first step, the distributions of the single images and the mosaic distributions were calculated (i.e., the same as those shown in Figure 7). In the second step, the distribution mean was obtained (filled triangle and star). Finally, the Euclidean distance from the mean to the diagonal was computed.

As we had already anticipated, if we look at the fire hydrant class (highlighted in gray in Table 1), we can see that this is the class where the contexts had the least influence: note that the distance was always positive and greatest (among the other objects) in the case of the sky and park contexts (i.e., 2.5436 and 0.7191). This may have been because the model had learned some pattern of the object itself (or perhaps a bias from the fire hydrant context that we had not detected), and, therefore, in most cases, the model continued to predict the mosaics as *fire hydrants*. We can see that both the sky and the park context, when combined with mug images (highlighted in orange), managed to move the mean of the distribution towards the prediction of the class of the context (i.e., −0.7512 and −2.8234): that is, mosaics of mugs combined with sky images were predicted mostly as *planes*, and mosaics of mugs with parks were predicted mostly as *benches* (see filled orange points on the left side of Figure 10). On the other hand, the road context combined with any object

obtained a mean distribution still far from zero (third row): in other words, the road context did not manage to move the distribution towards the prediction of the *fire hydrant* class.

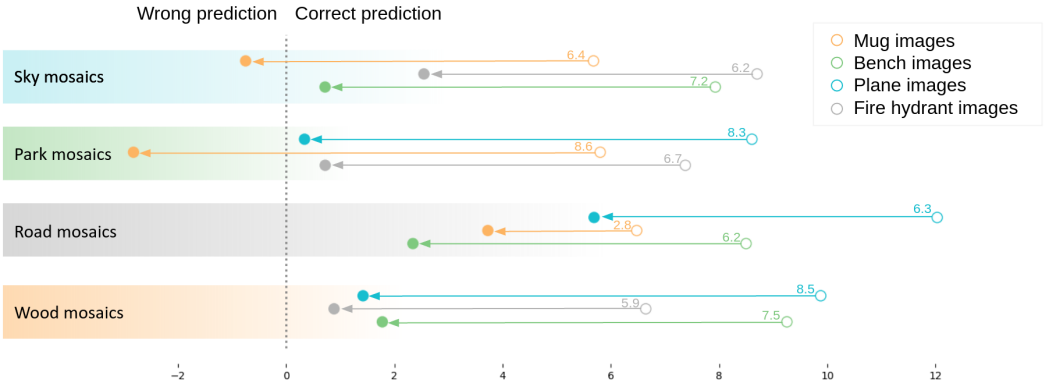

Figure 10. Visualization of the results of Table 1. The filled dots represent the mosaics. The empty dots correspond to the results of the single images. The diagonal is depicted by a dashed vertical line. The horizontal lines represent the difference between the two distances (i.e., that of the single images to the diagonal, and that of the mosaics to the diagonal). Greater values indicate the stronger influence of the context.

In this section, we have shown different ways to analyze and visualize the results of the *contextualized* mosaics, with the aim of evaluating the influence of context biases. After the first analysis, measures could be applied to mitigate the effect of context biases in this model-*C*, if the domain expert considered such biases to exist.

Results Checks

Taking advantage of the availability of the *WB* dataset, for this section we performed another experiment, to further analyze the importance of the context and to verify whether the results were consistent with those obtained by the *contextualized* mosaic methodology. To do so, for each context (i.e., park, sky, road and wood), we pasted an object of the remaining three classes, creating a total of 25 images per context–object pair. We then calculated the performance of the model for those new images. Note that this was another way of performing the intervention explained in Section 2.2: that is, fixing the four contexts intending to analyze the relationship between context C and the predicted class Y.

The results for each class are shown in the form of bar plots. For example, Figure 11 shows the number of images predicted as *plane*: (**a**) using 25 context images only; (**b**) using 25 mug instances superimposed on the sky context; (**c**) using 25 fire hydrant images within the sky context; and (**d**) using 25 bench images superimposed on the sky context. The color code of the following figures was green for the bench images and the park context, orange for the mug images and the wood context, gray for the fire hydrants and the road context and blue for the plane images and the sky context. Let us now analyze the importance of context for each class.

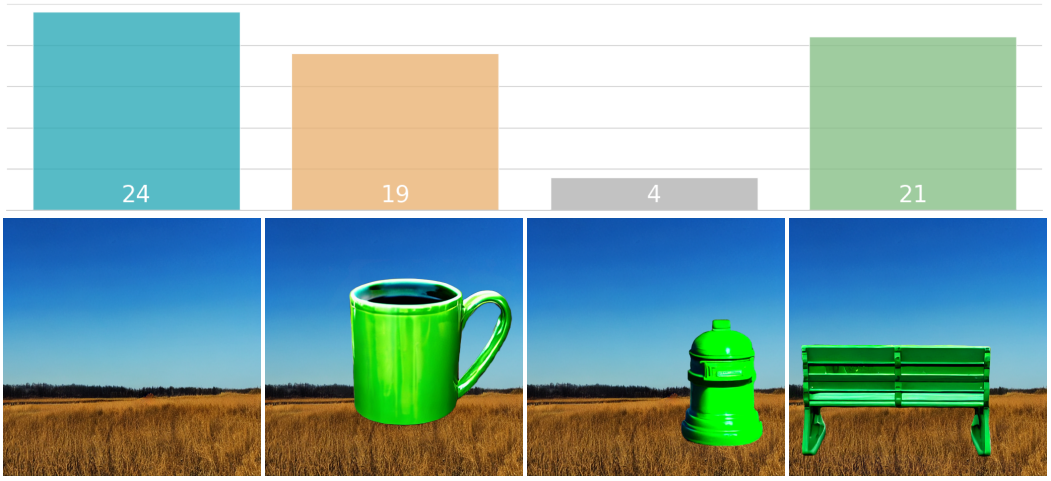

Figure 11. Number of images classified as *plane* within the sky context. The blue bar corresponds to results with only sky images, the orange to the mug within sky context images, the gray to a fire hydrant and the green to the bench within sky context images. The instance examples used for each bar plot are shown at the bottom of the image: (**a**) sky context image; (**b**) image of a mug superimposed on the sky context; (**c**) a fire hydrant; and (**d**) a bench also superimposed on the sky image.

Sky. These results confirm the findings obtained with the mosaics. Figure 11 shows that 24 sky images out of 25 were classified as planes: this means that the blue sky was an important feature for the *plane* class. Nevertheless, on the basis of that experiment alone, we could not affirm that this was an unwanted shortcut. However, 19 (out of 25) sky images with mugs were classified as *planes*, and 21 with benches were also classified as *planes*: this, on the contrary, does confirm that this correlation learned by the model was not the intended behavior. Note that in the case of the fire hydrant, only four sky images with fire hydrant objects were classified as *planes*. This was consistent with previous results obtained by the mosaic analysis: the fire hydrant features were more relevant to the model than the presence of the sky.

Park. Also consistent with the mosaic results, vegetation was shown to be important to the *bench* class (see Figure 12): 25 images of parks only were predicted as *benches*. In addition, 25 mug objects superimposed on parks, 22 fire hydrants and 24 out of 25 planes over parks were predicted as *benches*: that is to say, regardless of the object present in the park images, almost all of them were predicted as *benches*.

Road. As already anticipated with the *contextualized* mosaics, the road context was not a relevant context for the *fire hydrant* class—or, at least, the model did not rely solely on the context. This can be clearly seen in the gray bar plot in Figure 13: only 15 road images, only 5 bench images, only 9 mug images and none of the plane images were predicted to be *fire hydrants*.

Wood. This context favored the *mug* class, as all the wooden images were all classified as *mugs* (see Figure 14). However, when the different objects were pasted, the model was confused. The number of images classified as mugs decreased drastically: 9 bench images, 6 fire hydrant images and 8 plane images out of 25 were classified as *mugs*.

These findings were consistent with the *contextualized* mosaic outcomes, demonstrating their reliability and usefulness. The difficulty of having segmented objects and of building cross-context images rendered this second experiment nonviable in a real use case scenario, whereas the mosaic creation was straightforward. Therefore, as long as we have identified possible sources of unwanted biases, we can build mosaics and use the proposed methodology to analyze their impact.

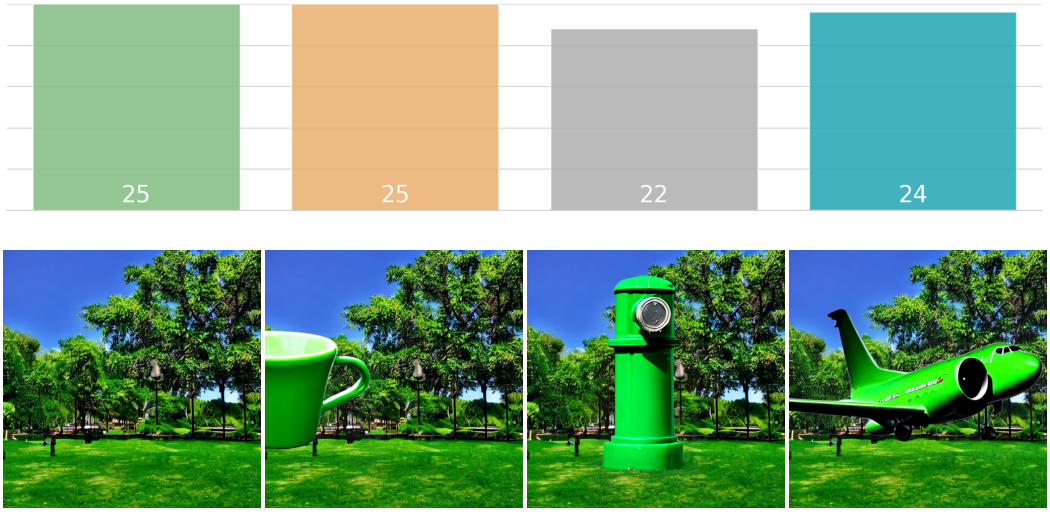

Figure 12. Number of images classified as *bench* within the park context. The green bar corresponds to results with only context images, the orange bar to the mug within park context images, the gray bar to a fire hydrant within park context images and the blue bar to the plane within park context images. As before, instance examples used for each bar plot are shown at the bottom of the image: (**a**) park context image, (**b**) a mug in a park context, (**c**) a fire hydrant and (**d**) a plane superimposed on a park context.

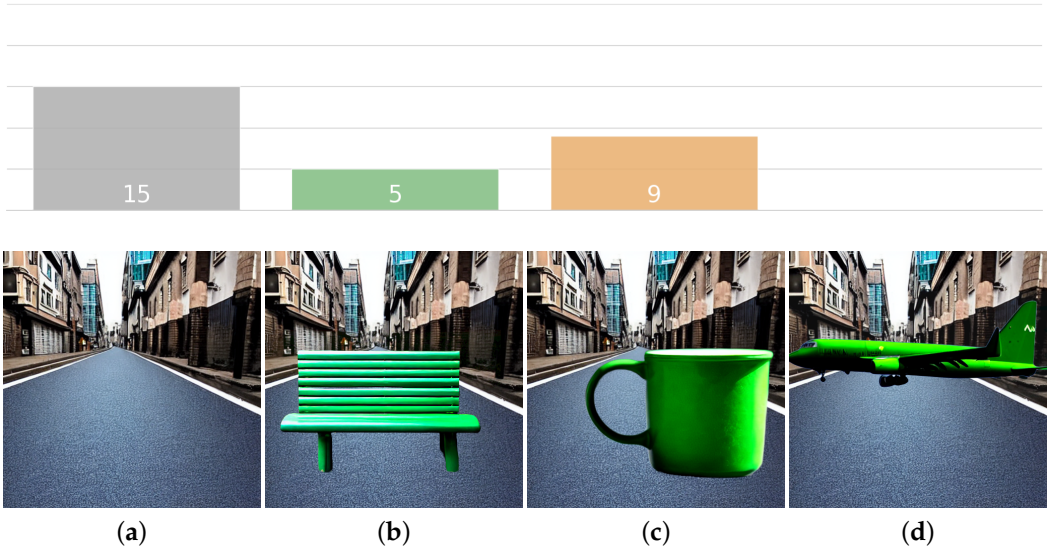

Figure 13. Number of images classified as *fire hydrant* within the road context. The gray bar corresponds to results with only road images (**a**), the green bar corresponds to bench images superimposed on a road context (**b**) and the orange bar to mug images in the road context (**c**). The bar corresponding to the plane within the road context (**d**) obtained an accuracy of zero.

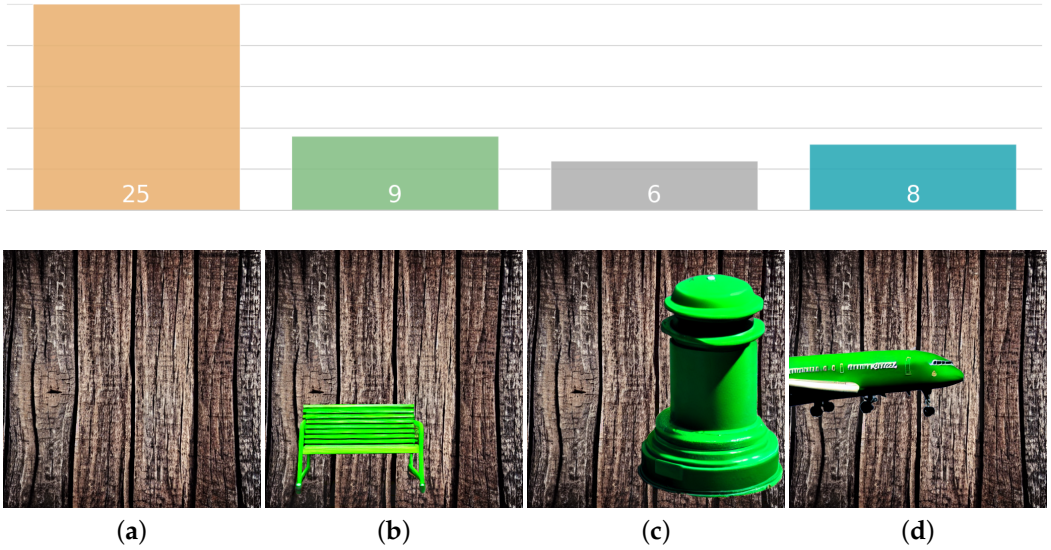

Figure 14. Number of images classified as *mug* within the wood context. The orange bar corresponds to results with only wood images (**a**), the green bar to bench images within the wood context (**b**), the gray to fire hydrant images (**c**) and the blue bar to plane images (**d**).

4. Discussion

The *contextualized* mosaics methodology allowed us to analyze the context influence on the model-*C* decisions. The different steps followed are detailed in Figure 15. We started by identifying four potential *context* biases: the sky for the *plane* class; the park for the *bench* class; the road for the *fire hydrant* class; and the wood for the *mug* class. Then, we built the *contextualized* mosaics, by combining those contexts with the original images. Finally, after analyzing the impact produced on the output, the main findings were as follows. The park context was identified as a potential shortcut learned by the model to predict the *bench* class. To mitigate this shortcut, one could try to add more benches without a vegetation context, thus forcing the model to learn the bench characteristics and to rely less on the vegetation when predicting the *bench* class. The sky also turned out to be an element favoring the *plane* class.

To prevent any object from being identified as a plane when having a sky in the background, one could add more images of the other classes with blue sky (i.e., benches or mugs with a blue sky behind them). The road does favor the fire hydrant prediction, because all the fire hydrants are on the street; nevertheless, this bias could be considered not dangerous. The assessment of whether or not a bias could be harmful must be decided ultimately by the domain expert. Finally, the wood context is not a determinant shortcut in predicting the *mug* class. These context biases having been mitigated, the model will learn the characteristics of each object. Therefore, the model will generalize better, achieving a high performance when tested with images from outside the distribution (e.g., when tested with the *NC* or *WB* sets).

In this work, we only used the model's output—unlike previous works, which used XAI methods to asses biases. To better rationalize this decision, we examined whether we could identify the previously introduced biases by solely relying on the explanations. To do so, we applied the GradCAM [10] method to the model-*C*, using the images from the *C* dataset. We chose this XAI method for two reasons: firstly, it is a trustworthy method, according to the *Focus* score [19]; secondly, GradCAM is a useful method for assisting humans in the task of bias detection, according to the *Utility-K* score [24]. Some of the explanations obtained are shown in Figure 16.

Although we had previously checked that the park context was relevant for the prediction of the *bench* class, identifying this bias by exclusively relying on the explanation would have been impossible. Note in Figure 16a that the explanation primarily focused on the bench. When it came to the *fire hydrant* and the *mug* class, quantifying the relevance of context solely based on the explanations was also challenging. The only case in which we might have had an

intuition that the context was important was in the *plane* class, because the most highlighted part in the explanations corresponded to the area above the plane (see Figure 16d).

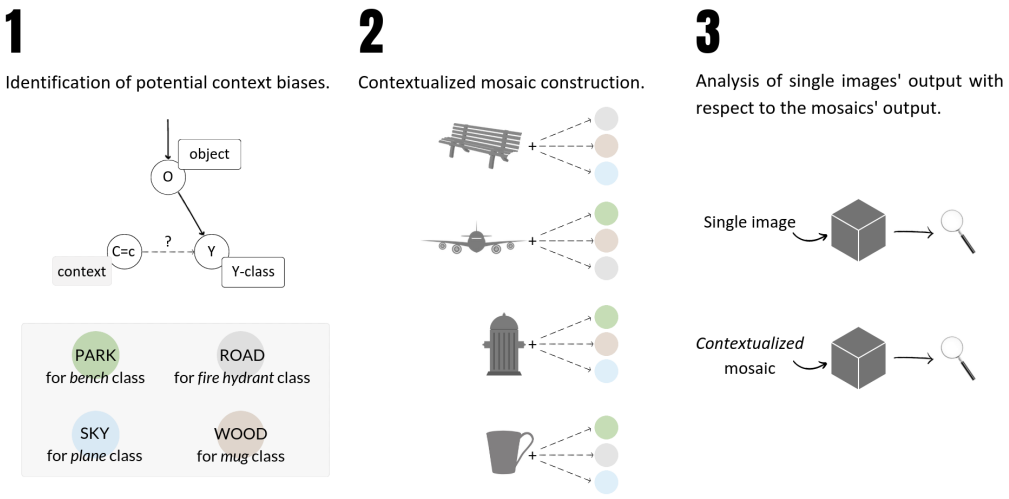

Figure 15. Steps of the proposed methodology. Firstly, four potential biases are identified: park for the *bench* class; road for the *fire hydrant* class; sky for the *plane* class; and wood for the *mug* class. Next, we construct the *contextualized* mosaics by combining each object with the three remaining contexts. Finally, the influence of the context is analyzed, by comparing the output of the single images with that of the mosaics.

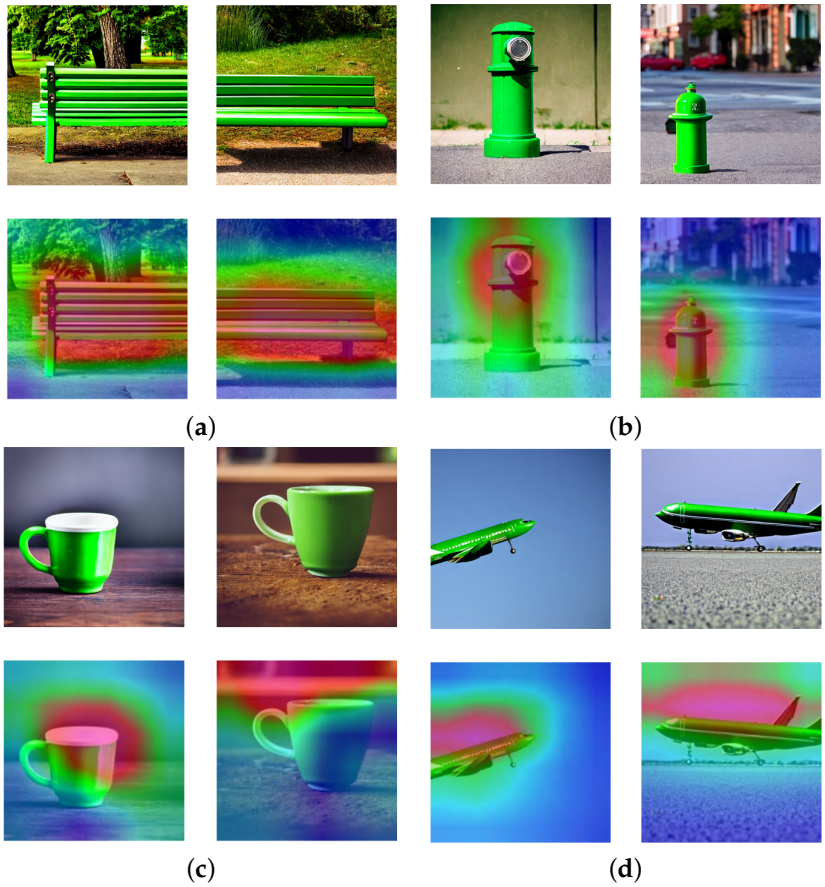

Figure 16. Examples of images from the *C* dataset and their corresponding feature attribution maps, obtained by the GradCAM method on top of the model-*C*. The examples are grouped by class: (**a**) *bench*; (**b**) *fire hydrant*; (**c**) *mug*; and (**d**) *plane*.

To conclude, while this work focused on analyzing the impact of *context* biases using mosaics, this methodology could be extended to examine other types of biases: for example, if textures were a source of bias, mosaics could be created, by combining the objects with the textures identified as potential biases.

Author Contributions: Conceptualization, A.A.-D. and D.G.-G.; methodology, A.A.-D.; software, A.A.-D.; formal analysis, A.A.-D. and D.G.-G.; data curation, A.A.-D.; writing—original draft preparation, A.A.-D.; writing—review and editing, A.A.-D., V.G.-A., U.C. and D.G.-G.; visualization, A.A.-D.; supervision, U.C. and D.G.-G.; funding acquisition, U.C. and D.G.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This work received funding from the European Union's H2020-INFRAIA-2019-1 program under the Grant Agreement n.871042 (SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics) and from the HORIZON-INFRA-2021-DEV-02 program under the Grant Agreement n.101079043 (SoBigData RI Preparatory Phase Project). Additionally, this work was supported by the Departament de Recerca i Universitats of the Generalitat de Catalunya, under the Industrial Doctorate Grant DI 2018-100.

Data Availability Statement: The three datasets used in this work are publicly available. Context dataset: https://storage.hpai.bsc.es/object-datasets/context.zip (accessed on 14 June 2023). No Context dataset: https://storage.hpai.bsc.es/object-datasets/no_context.zip (accessed on 14 June 2023). White Background dataset: https://storage.hpai.bsc.es/object-datasets/white_background.zip (accessed on 14 June 2023).

Acknowledgments: We thank the High Performance Artificial Intelligence (HPAI) Group and the Barcelona Supercomputing Center (BSC) for the resources provided to generate the datasets and train the models used in this work.

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

References

1. Suresh, H.; Guttag, J. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1–9.
2. Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1356. [CrossRef]
3. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
4. Shankar, S.; Halpern, Y.; Breck, E.; Atwood, J.; Wilson, J.; Sculley, D. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. *arXiv* **2017**, arXiv:1711.08536.
5. Buolamwini, J.; Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 77–91.
6. Daneshjou, R.; Vodrahalli, K.; Liang, W.; Novoa, R.A.; Jenkins, M.; Rotemberg, V.; Ko, J.; Swetter, S.M.; Bailey, E.E.; Gevaert, O.; et al. Disparities in Dermatology AI: Assessments Using Diverse Clinical Images. *arXiv* **2021**, arXiv:2111.08006.
7. 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.
8. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3319–3328.
9. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef] [PubMed]
10. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2019**, *128*, 336–359. [CrossRef]
11. Wang, S.; Zhou, C.; Zhang, D.; Chen, L.; Sun, H. A deep learning framework design for automatic blastocyst evaluation with multifocal images. *IEEE Access* **2021**, *9*, 18927–18934. [CrossRef]
12. Payá, E.; Bori, L.; Colomer, A.; Meseguer, M.; Naranjo, V. Automatic characterization of human embryos at day 4 post-insemination from time-lapse imaging using supervised contrastive learning and inductive transfer learning techniques. *Comput. Methods Programs Biomed.* **2022**, *221*, 106895. [CrossRef] [PubMed]
13. Van der Velden, B.H.; Kuijf, H.J.; Gilhuijs, K.G.; Viergever, M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* **2022**, *79*, 102470. [CrossRef] [PubMed]

14. Vila-Blanco, N.; Carreira, M.J.; Varas-Quintana, P.; Balsa-Castro, C.; Tomas, I. Deep neural networks for chronological age estimation from OPG images. *IEEE Trans. Med. Imaging* **2020**, *39*, 2374–2384. [CrossRef] [PubMed]
15. El Adoui, M.; Drisis, S.; Benjelloun, M. Multi-input deep learning architecture for predicting breast tumor response to chemotherapy using quantitative MR images. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 1491–1500. [CrossRef] [PubMed]
16. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity checks for saliency maps. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 9525–9536.
17. Sixt, L.; Granz, M.; Landgraf, T. When explanations lie: Why many modified bp attributions fail. In Proceedings of the International Conference on Machine Learning, Virtual Event, 13–18 July 2020; pp. 9046–9057.
18. Rong, Y.; Leemann, T.; Borisov, V.; Kasneci, G.; Kasneci, E. A Consistent and Efficient Evaluation Strategy for Attribution Methods. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 18770–18795.
19. Arias-Duart, A.; Parés, F.; Garcia-Gasulla, D.; Giménez-Ábalos, V. Focus! Rating XAI Methods and Finding Biases. In Proceedings of the 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Padua, Italy, 18–23 July 2022; pp. 1–8.
20. Arias-Duart, A.; Parés, F.; Giménez-Ábalos, V.; Garcia-Gasulla, D. Focus and Bias: Will It Blend? In *Artificial Intelligence Research and Development*; IOS Press: Amsterdam, The Netherlands, 2022; pp. 325–334.
21. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis With Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
23. Glymour, M.; Pearl, J.; Jewell, N.P. *Causal Inference in STATISTICS: A Primer*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
24. Colin, J.; Fel, T.; Cadène, R.; Serre, T. What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 2832–2845.