

Debiased-CAM to mitigate image perturbations with faithful visual explanations of machine learning

Wencan Zhang
National University of Singapore
Singapore
wencanz@u.nus.edu

Mariella Dimiccoli
Institut de Robòtica i Informàtica
Industrial, CSIC-UPC
Barcelona, Spain
maria.dimiccoli@upc.edu

Brian Y. Lim
National University of Singapore
Singapore
brianlim@comp.nus.edu.sg

ABSTRACT

Model explanations such as saliency maps can improve user trust in AI by highlighting important features for a prediction. However, these become distorted and misleading when explaining predictions of images that are subject to systematic error (bias) by perturbations and corruptions. Furthermore, the distortions persist despite model fine-tuning on images biased by different factors (blur, color temperature, day/night). We present Debiased-CAM to recover explanation faithfulness across various bias types and levels by training a multi-input, multi-task model with auxiliary tasks for explanation and bias level predictions. In simulation studies, the approach not only enhanced prediction accuracy, but also generated highly faithful explanations about these predictions as if the images were unbiased. In user studies, debiased explanations improved user task performance, perceived truthfulness and perceived helpfulness. Debiased training can provide a versatile platform for robust performance and explanation faithfulness for a wide range of applications with data biases.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Computer vision**; *Semi-supervised learning settings*; • **Security and privacy** → *Privacy protections*.

KEYWORDS

Explainable AI, Misleading explanations, Class activation map, Robust machine learning, Image perturbations, User studies

ACM Reference Format:

Wencan Zhang, Mariella Dimiccoli, and Brian Y. Lim. 2022. Debiased-CAM to mitigate image perturbations with faithful visual explanations of machine learning. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*, April 29-May 5, 2022, New Orleans, LA, USA. ACM, New York, NY, USA, 32 pages. <https://doi.org/10.1145/3491102.3517522>

1 INTRODUCTION

Machine learning models are increasingly capable to achieve impressive performance in many prediction tasks, such as image recognition [49], medical image diagnosis [29], captioning [86] and dialog

systems [16]. Despite their superior performance, deep learning models are complex and unintelligible; this limits user trust and understanding [60, 63, 87]. This has driven the development of myriad explainable artificial intelligence (XAI) and interpretable machine learning methods [33, 38, 87, 95]. Saliency maps [78, 82, 102] can provide intuitive explanations of Convolutional Neural Networks (CNN) for image prediction tasks by indicating which pixels or neurons were used for model inference. Amongst these, class activation map (CAM) [102], Grad-CAM [78] and extensions [13, 89] are particularly useful by identifying pixels relevant to specific class labels. Users can verify the prediction correctness by checking whether expected pixels are highlighted. Models would be considered more trustworthy if their CAMs matched what users believe as salient.

Despite the fidelity of CAMs on clean images, real-world images are typically subjected to systematic error, such as image blurring, color-distortion or lighting changes, which can affect what CAMs highlight. We call this systematic error *bias*¹ since it is directional based on a contextual factor or confound, and contrast it with *noise* that is based on non-directional random error. Also note that we are not referring to societal bias or discrimination (e.g., racism, sexism) [22]. Blurring can be due to accidental motion [50] or defocus blur [85], or deliberate obfuscation for privacy protection [21]. Unlike [100] which found explanation harms privacy, we find that privacy can harm explanation. Images may also be biased with shifted color temperature [4] due to mis-set white balance, or biased with daylight changes (e.g., day to night, sunrise/sunset). These biases decrease model prediction performance [4, 21, 85] and we further show that they also lead to deviated or distorted CAM explanations that are less faithful to the original scenes. For different bias types (image blur, and color temperature shift, day/night lighting), we found that CAMs deviated more as image bias increased (Fig. 1 and Fig. 3: Biased-CAMs from RegularCNN for $\sigma > 0$). Although Biased-CAM represents what the CNN considers important in a biased image, it is misaligned with people's expectations [68], misleads users to irrelevant targets, and impedes human verification and trust [26] of the model prediction. For example, when explaining the inference of the "Fish" label for an image prediction, Biased-CAMs select pixels of the man instead of the fish (Fig. 1).

To align with user expectations, models should not only have the right predictions but also have the right reasons [73]; however, current approaches face challenges in achieving this goal, particularly for biased data. First, while fine-tuning the model on biased data

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

© 2022 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-9157-3/22/04.

<https://doi.org/10.1145/3491102.3517522>

¹Note that this does not refer to *social bias* that is presently popularly studied in AI fairness and algorithmic bias. We are using the word as defined in engineering and physics regarding measurements.

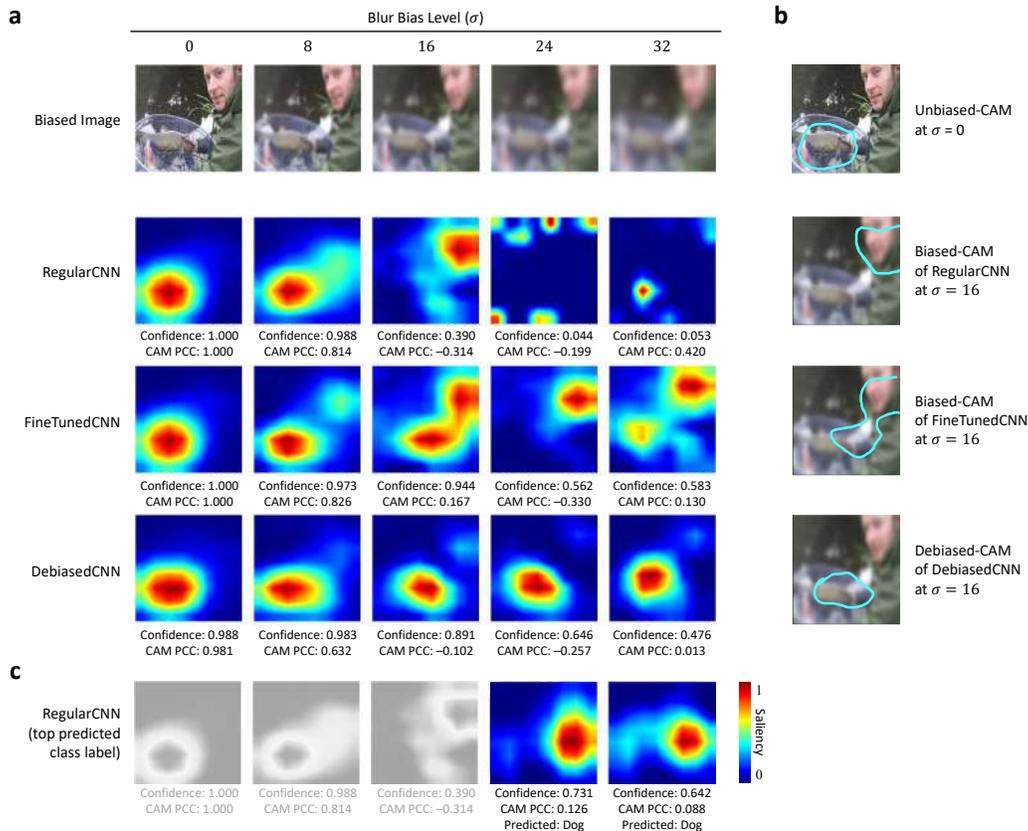


Figure 1: Deviated and debiased CAM explanations for prediction label "Fish". a) Debiased-CAMs (from DebiasedCNN) were most faithful to the Unbiased-CAM (from RegularCNN at $\sigma = 0$) as blur bias increased. In contrast, Biased-CAMs from RegularCNN and FineTunedCNN became very deviated with a much lower CAM Pearson Correlation Coefficient (PCC). The wrong CAMs can mislead users to think the predictions were wrong even if they were correct. b) Debiased-CAM selected similar important pixels of the Fish as Unbiased-CAM, while Biased-CAMs selected irrelevant pixels of the person or background instead. c) CAM of the top predicted class label with only RegularCNN at $\sigma = 32$ predicting the wrong prediction label "Dog".

can improve its performance [21, 85], this does not necessarily produce explanations aligned with human’s understanding. Indeed, we found that explanations remain deviated and unfaithful (Fig. 1, Fine-TunedCNN Biased-CAMs). Conversely, retraining the model with attention transfer [47, 57] only improves explanation faithfulness for clean images, but cannot handle biased images. Finally, evaluating the human interpretability of explanations requires deep inquiry into user perception, understanding and usage [1, 5, 25], but typical evaluations of XAI involve only data simulations [8, 30, 73, 83, 94] or simple surveys [10, 13, 77, 78, 103].

Inspired by how people can “see through the blur” to recognize blurred images due to prior experiences with unblurred but unrelated images, we propose a debiasing approach such that models are trained to faithfully explain the event despite biased sources. Using CNNs with Grad-CAM saliency map explanations [78], we developed DebiasedCNN that interprets biased images as if predicting on the unbiased form of images and produces explanations, Debiased-CAMs, that are more human-relatable and robust. The approach has a modular design: 1) it is self-supervised which does not require additional human annotation for training; 2) it produces

explanations as a secondary prediction task, so that they are retrainable to be debiased; 3) it models the bias level as a tertiary task to support bias-aware predictions. The approach not only enhances prediction performance on biased data, but also produces highly faithful explanations about these predictions as if the data were unbiased (Fig. 1: DebiasedCNN CAMs).

To evaluate the developed model, we conducted simulation and user studies to address the research questions on 1) how bias decreases explanation faithfulness and how well debiasing mitigates this, and 2) how sensitive people are to perceiving explanation deviations and how well debiasing improves perceived explanation truthfulness and helpfulness. For generality, the simulation studies spanned different image prediction tasks (object recognition, activity recognition with egocentric cameras, image captioning, and scene understanding), bias types (blur, color shift, and night vision interpolation) and various datasets. Across all studies, we found that while increasing bias led to poorer prediction performance and worse explanation deviation, Debiased-CAM showed the best improvement in task performance as well as explanation faithfulness. Instead of trading off task performance for explanation faithfulness,

our debiasing training improved both. We further demonstrated the usability and usefulness of Debiased-CAMs in two controlled user studies. Quantitative statistical and qualitative thematic analyses validated that users can perceive the improved truthfulness and helpfulness of Debiased-CAMs on biased images. In summary, this paper made the following **contributions**:

- (1) Assessed the deviations in model explanations due to bias in data across different bias types and levels.
- (2) Proposed a technical approach to accurately predict and faithfully explain inferences under data bias.
- (3) Validated the improvements in perceived truthfulness and helpfulness of debiased explanations.

2 RELATED WORK

We review explainable AI methods for image predictions, how images get biased, how misleading explanations harm user experience and performance, and methods to improve explanation faithfulness.

2.1 Explainable AI for visual CNN models

Many explainable AI (XAI) techniques have been proposed to understand the predictions of CNNs. These include saliency maps [13, 70, 78, 81, 102], feature visualization [10, 39, 66], neuron activations [41] and concept variables [43, 46]. Saliency maps are intuitive to interpret deep CNN models, where important pixels are highlighted to indicate their importance towards the model prediction. Computing the prediction gradient [81, 83, 93] can identify sensitive pixels. Another approach divides prediction outcome across features by Taylor series approximation [8] or Shapley values [61]. Specific to CNNs, coarser saliency maps can be generated by aggregating activation maps as a weighted sum across convolutional kernels [13, 70, 78, 89, 102]. For this work, we evaluated Grad-CAM [78] to test if users can perceive truthful, biased, or debiased explanations, and expect our findings to be generalizable.

2.2 Systematic error and corruptions in images

Although many models are trained on clean curated images, real-world images are subject to systematic errors (biases), perturbations and corruptions. Contextual or incidental biases include blurring, color distortions, or lighting changes. Blurring may be due to accidental motion blur [50], defocus blur [85] or deliberate obfuscation for privacy protection [21]. Image color shift [4] may be due to mis-set white balance. These biases can degrade model performance [4, 21, 85], and limit their usefulness in real-world applications. Images of outdoor scenes regularly change by time of day and seasons due to sunlight or weather changes [51]. Images can also be corrupted due to data processing, such as JPEG compression artifacts, Gaussian noise, brightness or contrast levels [36]. Mitigation strategies to handle such data errors include model fine-tuning with images at known blurred levels [21], or data augmentation with images blurred at multiple levels [85]. However, these approaches only aimed to improve prediction performance and not explanation faithfulness. In this work, we found that explanations remain deviated and we propose methods to debias them.

Such data errors are related to the problem of model robustness, where small changes to data should not cause large changes in model behavior. This is an active area of research [36, 37, 101], but

methods typically focus on improving performance by increasing decision boundary smoothness. In this work, we aim to make explanations more robust. Recent work by Dombrowski et al. [24] improved explanation robustness by similarly increasing decision smoothness relative to explanations, but this assumes clean data, and learns average explanations under bias. Instead, we debias explanations away from deviations due to biased data. Also, other than focusing on explanation robustness or stability towards the impression of global trustworthiness, we focus on faithful explanations that are verifiable per instance.

2.3 Risk of misleading model explanations

User studies of model explanations aim to show that explanations can improve user understanding and trust [45, 59, 65, 90, 98]. These tend to study scenarios of correct model predictions and ideal explanations, but models can make prediction errors or may not be confident in their decisions. Studies have explored how this may lead to distrust, mistrust and over-trust [58, 69, 92]. For such cases, explanations can be avoided when there is a high chance of model error. However, explanations can still be wrong despite the model predicting correctly. For example, explanations may highlight spurious pixels [97], be adversarially manipulated [23, 31], or subject to input error [88]. These cases are harder to detect, pose a serious risk to decrease user trust [18, 54], or mislead users [54]. Unlike works that explore how different explanation formats affect trust [91], we investigate how slight data variations affect user performance and trust. Since data bias and corruption are prevalent in the real-world, it is tantamount to identify the severity of the problem and mitigate it with more robust explanations [35]. In this work, we quantify the extent of explanation deviation due to data bias, and evaluated how sensitive users are to these deviations.

2.4 Attention transfer to correct explanations

While explanation techniques are primarily designed to improve human understanding of model behavior, they can be used to guide model training. One approach is to use transfer learning to regularize attention from a better model to the model under training, such as with student-teacher networks [47]. Another approach indirectly trains attention by ablating salient pixels from input images and maximizing the classification loss between the ablated and original images [57]. However, these approaches only train on clean data and will reinforce biased explanations if trained on obfuscated or biased data. Unlike conventional self-supervised learning with data augmentation and contrastive learning to improve feature learning [14], we use the unbiased explanation as a surrogate "label" to train the debiased model to predict a more faithful explanation.

3 TECHNICAL APPROACH

We first describe baseline RegularCNN and FineTunedCNN approaches to predict on unbiased and biased image data, then our proposed DebiasedCNN architectures to predict on biased image data with debiased explanations.

3.1 Regular and Fine-tuned Models

A regularly trained CNN model (RegularCNN) can generate a truthful CAM \tilde{M} (Unbiased-CAM) of an unbiased image \mathbf{x} , but will

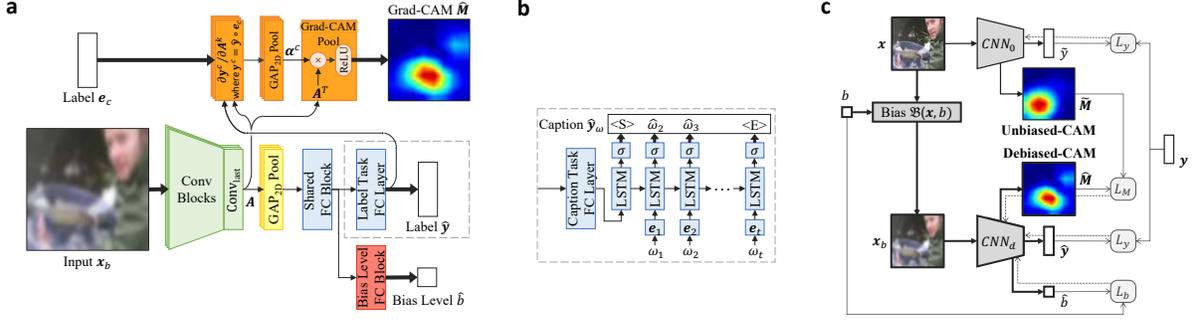


Figure 2: Architecture of DebiasedCNN. a) DebiasedCNN is a multi-input, multi-task convolutional neural network with two inputs image x_b and label e_c for class c , and three tasks for primary prediction task \hat{y} , CAM explanation task \hat{M} , and bias level prediction task \hat{b} . b) DebiasedCNN can be trained for different primary tasks, such as image captioning. c) Meta-architecture with self-supervised learning to minimize the CAM loss L_M between Unbiased-CAM \hat{M} from RegularCNN (CNN_0) predicting on unbiased image x and Debiased-CAM \hat{M} from DebiasedCNN (CNN_d) predicting on biased image x_b at bias level b .

produce a deviated CAM \tilde{M} (Biased-CAM) for the image under bias x_b , i.e., $\tilde{M}(x) \neq \hat{M}(x_b)$, due to the model not training on any biased images and learning spurious correlations with blurred pixels. A fine-tuned model trained on biased images can improve the prediction performance on biased images, but will still generate a deviated CAM \tilde{M} (Fig. 1a and Fig. 3a-c: CAMs of FineTunedCNN), as it was only trained with the classification loss and not explanation loss. While these models can be explained with Grad-CAM, they are not retrainable to improve their CAM faithfulness.

3.2 DebiasedCNN Model with Debiased-CAM Explanations

3.2.1 Trainable CAM as secondary prediction task. We enable CAM retraining by redefining Grad-CAM as a prediction task. Grad-CAM [78] computes a saliency map explanation of an image prediction with regards to class c as the weighted sum of activation maps in the final convolutional layer of a CNN. Each activation map A^k indicates the activation A_{ij}^k for each grid cell (i, j) of the k th convolution filter $k \in K$ (set of all filters). The importance weight α_k^c for the k th activation map is calculated by back-propagating gradients from the output \hat{y} to the convolution filter, i.e.,

$$\alpha_k^c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \frac{\partial \hat{y}^c}{\partial A_{ij}^k} \equiv \text{GAP}_{ij} \left(\frac{\partial \hat{y}^c}{\partial A^k} \right) \quad (1)$$

where H and W are the height and width of activation maps, respectively; \mathbf{y}^c is a one-hot vector indicating only the probability of class c ; $\text{GAP}_{ij}(\cdot)$ is the global average pooling operation. The class activation map (CAM) is the weighted combination of activation maps, followed by a ReLU transform to only show positive activations for class c , i.e.,

$$M^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right) \equiv \hat{M} = \text{ReLU} \left(\alpha^c A^T \right) \quad (2)$$

which we rewrite as a matrix multiplication of all K importance weights $\alpha^c = \{\alpha_k^c\}^K$ and the transpose of activation maps A along the k th axis, i.e., $A^T = \{A_{ij}^k\}^{K \times H \times W}$.

Therefore, the CAM prediction task can be redefined as three non-trainable layers (computational graph) in the neural network (orange in Fig. 2a) to compute $\frac{\partial \hat{y}^c}{\partial A^k}$, α_k^c , and \hat{M} . By reformulating Grad-CAM as a *secondary prediction task*, we can train the model with faithful CAM based on differentiable CAM loss by back-propagating through this task. This task takes e_c as the second input to the CNN architecture to specify the target class label for the CAM. c is set as the ground truth class label at training time, and chosen by the user at run time. We call the aforementioned approach Multi-Task DebiasedCNN, and call the conventional use of Grad-CAM as Single-Task DebiasedCNN. For single-task DebiasedCNN, the loss is added as a simple sum to the primary classification task, rather than predicted with secondary task. This will limit its learning since weights are not updated with gradient descent.

3.2.2 Training CAM debiasing with Self-Supervised Learning. To debias CAMs \tilde{M} of biased images x_b toward truthful Unbiased-CAMs \hat{M} of clean images x , i.e., $\tilde{M}(x_b) \approx \hat{M}(x)$, we train DebiasedCNN with self-supervised learning to transfer knowledge of corresponding unbiased images in RegularCNN into DebiasedCNN. We aim to minimize the difference between Unbiased-CAM \hat{M} and Debiased-CAM \tilde{M} . The training involves the following steps (see Fig. 2c): 1) Given a dataset with clean images $x \in X$ and labels y , apply a bias transformation (e.g., blur) to create biased variants of each image $x_b \in X_b$. 2) Train a RegularCNN to predict label \hat{y} on clean image x . We assume that its Grad-CAM explanations \hat{M} are correct and serve as a good oracle for Unbiased-CAMs. 3) Train a DebiasedCNN to predict label \hat{y} on corresponding biased image x_b , and explain with CAM \tilde{M} . DebiasedCNN is trained with loss function:

$$L = L_y(y, \hat{y}) + \omega_M L_M(\tilde{M}, \hat{M}) \quad (3)$$

where L_y is the classification loss, L_M is the CAM loss, and ω_M is a hyperparameter. The training can be interpreted as attention transfer from an unbiased model to the new model. DebiasedCNN can be generalized to image prediction other tasks (e.g., image captioning: Fig. 3b), other bias types (e.g., color temperature, lighting: Fig. 3c,d), different base CNN models (e.g., VGG16, Inception v3, ResNet50, Xception), and for privacy-preserving machine learning.

3.2.3 Bias-agnostic, Multi-bias predictions with tertiary task. Image biasing can happen sporadically at run time, so the image bias level b may be unknown at training time. Instead of training on specific bias levels [21] or fine-tuning with data augmentation on multiple bias levels [85], we added a *tertiary prediction task* – bias level regression – to DebiasedCNN to leverage supervised learning (Fig. 2a: salmon-colored layers). This enables DebiasedCNN to be bias-aware (can predict bias level) and bias-agnostic (predict under any bias level). With the bias level prediction task, the training loss function for multi-bias, multi-task DebiasedCNN is:

$$L = L_y(y, \hat{y}) + \omega_M L_M(\tilde{\mathbf{M}}, \hat{\mathbf{M}}) + \omega_b L_b(b, \hat{b}) \quad (4)$$

where L_b is the bias prediction loss, and ω_b is a hyperparameter.

3.2.4 Training loss terms. In all, there are three loss terms: primary task loss L_y as cross-entropy loss for standard classification tasks, and as the sum of negative log likelihood for each word [86] in image captioning tasks; bias level loss L_b as the mean squared error (MSE), common for regression tasks; CAM loss L_M as the mean squared error (MSE), since CAM prediction can be considered a 2D regression task, and this is common for visual attention tasks [47].

3.2.5 Summary of DebiasedCNN Model Variants. DebiasedCNN has a modular design: 1) single-task (st) or multi-task (mt) to improve model training; and 2) single-bias (sb) or multi-bias (mb) to support bias-aware and bias-agnostic predictions. We denote the four DebiasedCNN variants as (sb, st), (mb, st), (sb, mt), (mb, mt), and conducted ablation studies to compare between them. Supplementary Fig. 1 and Supplementary Table 2 show details of each variant.

4 SIMULATION STUDIES

To evaluate how much CAMs deviate with biased images and how well DebiasedCNN recovers CAM Faithfulness, we conducted five simulation studies with varying datasets, prediction tasks (classification, captioning), bias types (blur, color temperature, day/night lighting), and bias levels. These studies inform which applications explanation biasing is problematic, and show that our debiased training can successfully mitigate these deviations.

4.1 Evaluation Metrics

We evaluated *prediction performance* and *CAM explanation faithfulness* to compare model variants. For classification, we measured the area under the precision-recall curve (PR AUC) as it is robust against imbalanced data [76], and calculated the class-weighted macro average to aggregate across multiple classes. For image captioning, we calculated the BLEU-4 [67] score that measures how closely 4-grams in the predicted and actual captions matched. For bias level regression, we calculated accuracy with R^2 . We define the correctness of CAM explanations by their similarity or *faithfulness* to the original Unbiased-CAMs from RegularCNN that infers on unbiased data. To better compare CAMs beyond simple residual differences (e.g., MAE, MSE), we calculated CAM Faithfulness as the Pearson’s Correlation Coefficient (PCC) [11, 56] of pixel-wise saliency as it closely matches the human perception to favor compact locations and match the number of salient locations [56], and it fairly weights between false positive and false negatives [11].

4.2 Results

In general, CAMs deviate more from Unbiased-CAMs as bias levels increased, but DebiasedCNN reduces this deviation. Debiased retraining also improved model prediction performance, which suggests that DebiasedCNN indeed “sees through the bias”. Fig. 4 shows our evaluation Task Performance and CAM Faithfulness in ablation studies across increasing bias levels for different prediction tasks and datasets (Supplementary Table 1). Fig. 3 shows some examples of deviated and debiased CAMs. Next, we describe the experiment method and results for each simulation study.

4.2.1 Simulation Study 1 (Blur Bias). We evaluated CAMs for blur biased images of the object recognition dataset ImageNet [40]. We scaled images to a standardized maximum size of 1000×1000 pixels and applied uniform Gaussian blur at various standard deviations σ . We found that Task Performance and CAM Faithfulness decreased with increasing blur level for all CNNs, but DebiasedCNN mitigated these decreases (Fig. 4a). This indicates that model training with additional CAM loss improved model performance rather than trading-off explainability for performance [74]. RegularCNN had the worst Task Performance and the lowest CAM Faithfulness for all blur levels ($\sigma > 8$). In comparison, trained with differentiable CAM loss, DebiasedCNN (sb, mt) showed marked improvements to both metrics, up to 2.33x and 6.03x over FineTunedCNN’s improvements, respectively. Trained with non-differentiable CAM loss, DebiasedCNN (sb, st) improved both metrics to a lesser extent than DebiasedCNN (sb, mt), confirming that separating the CAM task from the classification task enabled better weights update. Trained with an additional bias-level task, multi-bias DebiasedCNN (mb, mt) achieved high Task Performance and CAM Faithfulness for all bias levels that is only marginally lower than single-bias DebiasedCNN (sb, mt), because of the former’s good regression performance for bias level prediction (Supplementary Fig. 4).

4.2.2 Simulation Study 2 (Blur Bias, Egocentric). We evaluated the impact of blur biasing with a more ecologically realistic task – wearable camera activity recognition (NTCIR-12 [34]). This task² represents a real-world use case where egocentric cameras may capture blurred images accidentally due to motion or defocus, or deliberately for privacy protection. We found the same trends as for the ImageNet classification task with some differences due to the increased task difficulty (Fig. 4b). In particular, the differences between RegularCNN and DebiasedCNN in Task Performance and CAM Faithfulness were amplified, indicating that debiasing is more useful for this application. Task Performance and CAM Faithfulness decreased steeply for RegularCNN with increasing blur bias, while DebiasedCNN significantly recovered both metrics, demonstrating marginal decreases with increasing bias. FineTunedCNN marginally increased CAM Faithfulness from RegularCNN (< 44%), while DebiasedCNN achieved a much larger improvement by up to 229%. We verified these trends for different CNN backbones and found that more accurate models produced more faithful CAMs even for stronger blur (Supplementary Figs. 5 and 6). Hence, Debiased-CAM

²Note that we mean that the use case could have blurred images, not that the NTCIR-12 dataset has blurred images. Blurring or biasing ImageNet photos (which may include curated stock photos) is an unrealistic use case, but there are more ecologically legitimate reasons for egocentric photos to be biased.

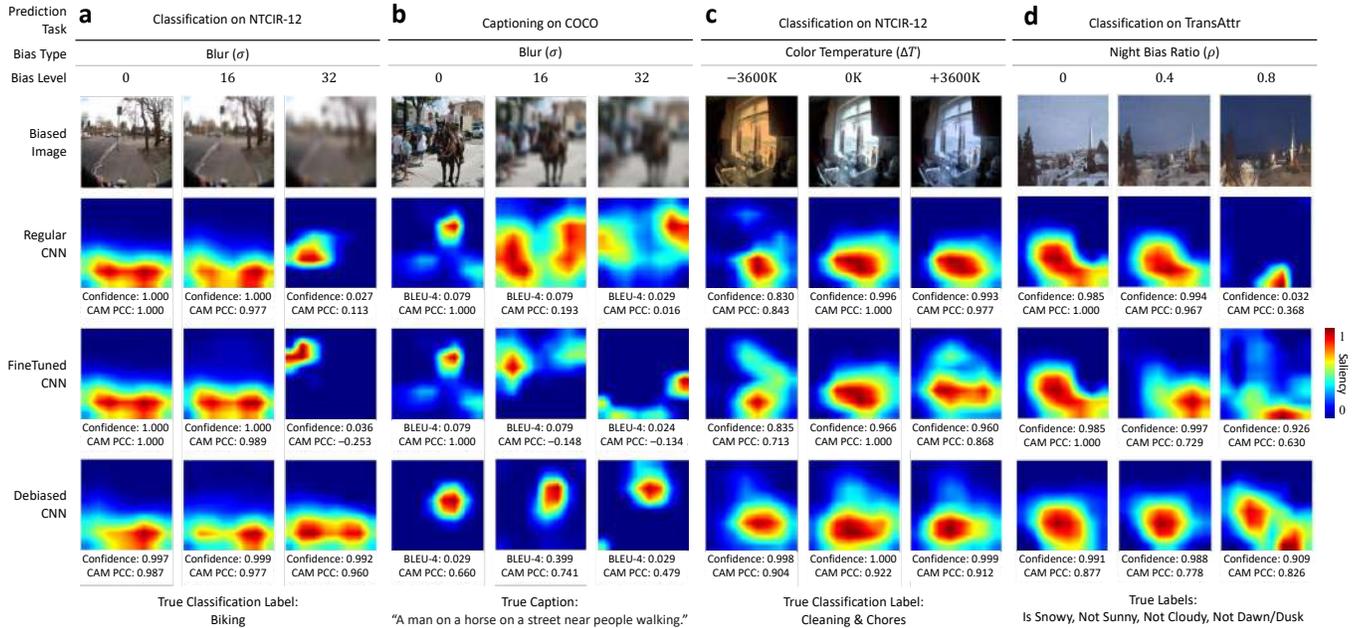


Figure 3: Deviated and debiased CAM explanations from models trained on different prediction tasks (a-d) with varying bias levels. In general, RegularCNN and FineTunedCNN had deviated CAMs that missed selecting important pixels, while DebiasedCNN had CAMs similar to Unbiased-CAMs. At no bias, all CAMs from RegularCNN and FineTunedCNN are unbiased.

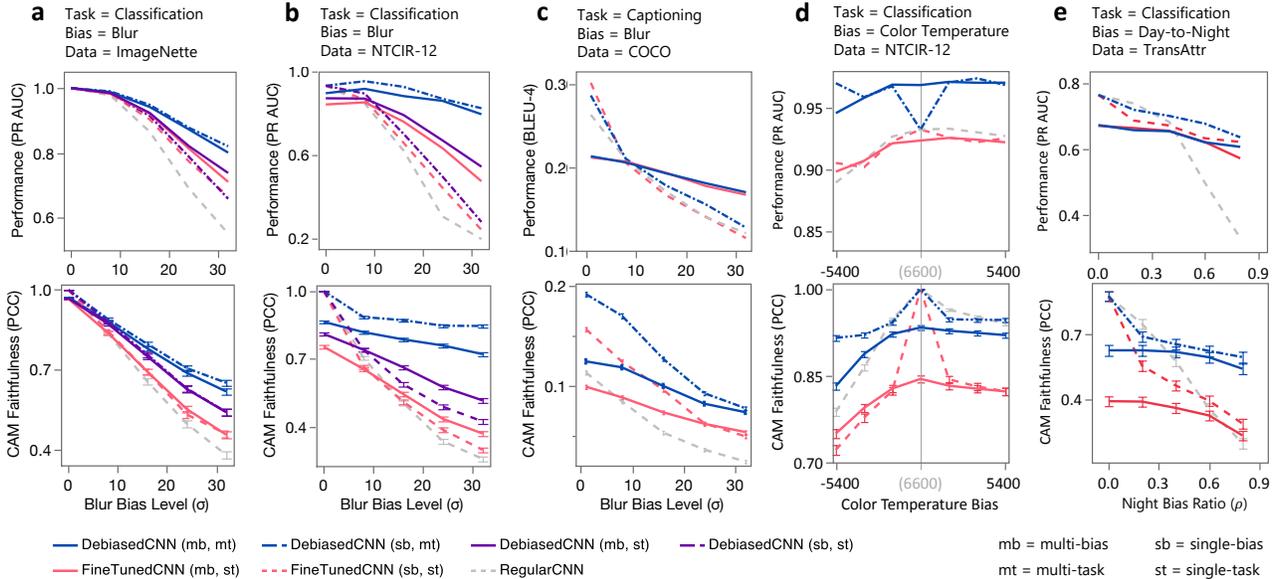


Figure 4: Task Performance and CAM Faithfulness for different prediction tasks with increasing bias levels. a)-e) All models' Task Performance and CAM Faithfulness decreased with increasing blur, while DebiasedCNN decreased the least. For DebiasedCNN variants, multi-task had the highest CAM Faithfulness and Task Performance that is higher than single-task.

enables privacy-preserving wearable camera activity recognition with improved performance and faithful explanations.

4.2.3 Simulation Study 3 (Blur Bias Captioning). We evaluated the influence of blur on a different prediction task – image captioning (COCO [15]). We found similar trends in Task Performance and CAM Faithfulness as before, though all models performed poorly at

all blur levels (Fig. 4c). Furthermore, CAM Faithfulness was low for all models, even for RegularCNN at a small blur bias ($\sigma = 1$). This could be because captioning is much harder than classification, and CAM retraining is weakened by vanishing gradients due to the long LSTM recurrence. Yet, DebiasedCNN improved CAM Faithfulness for all blur levels by up to 224% from RegularCNN.

4.2.4 Simulation Study 4 (Color Temperature Biased). We evaluated color temperature bias on wearable camera images in NTCIR-12. This represents another realistic problem for the wearable camera use case, where the white balance may be miscalibrated. We set the neutral color temperature t to 6600K (cloudy/overcast) and perturbed the color temperature bias by applying Charity’s color mapping function to map a temperature to RGB values [12]. Color temperature can be bidirectionally biased towards warmer (more orange, lower values) or cooler (more blue, higher values) temperatures from neutral 6600K. Furthermore, image pixel values deviate asymmetrically with larger deviations for orange than for blue biases. Consequently, we found that orange bias led to a larger decrease in Task Performance and CAM Faithfulness than blue bias (Fig. 4d). Notably, CAM deviation was smaller across all color temperature biases than for blur biases, as indicated by the smaller decrease in CAM Faithfulness (compare Fig. 4b, d); hence, Task Performance also did not decrease as much as blur bias. FineTuned-CNN had similar Task Performance but lower CAM Faithfulness than RegularCNN; this suggests that color-biased images were too similar to improve model training with classification fine-tuning, and yet this significantly degraded explanation quality. In contrast, DebiasedCNN improved Task Performance and CAM Faithfulness compared to RegularCNN. Furthermore, due to bidirectional bias, multi-bias training enabled DebiasedCNN (mb, mt) to have significantly higher Task Performance even for unbiased images ($\Delta t_b = 0$).

4.2.5 Simulation Study 5 (Lighting Bias). We evaluated lighting bias for outdoor scenes for a multi-label scene attribute recognition task (transient attribute database, TransAttr [51]). Lighting in outdoor scenes regularly change across hours or seasons due to transient attributes, such as sunlight or weather changes. Hence, models trained on images captured in one lighting condition may predict and explain differently under other conditions. Specifically, for the multi-label prediction task of classifying whether a scene is Snowy, Sunny, Foggy, or Dawn/Dusk, we biased whether the scene was daytime or nighttime. We performed a pixel-wise interpolation with ratio ρ to simulate interstitial periods between day and night (details in Appendix B.1.2). We found similar trends in Task Performance and CAM Faithfulness as with previous blur-biased classification tasks. The image prediction training was biased towards day-time photos, and as photos became darker to represent dusk or night time, all models generated more deviated, but least so for DebiasedCNN. Given the regularity and frequency of outdoor scenes changes, this study demonstrates the prevalence of biasing in model predictions and explanations, and emphasizes the need for Debiased-CAMs.

5 USER STUDIES

Having found that DebiasedCNN improves CAM faithfulness, we next evaluated how well Debiased-CAM improves human interpretability over Biased-CAM. We conducted user studies to evaluate their perceived truthfulness (User Study 1) and helpfulness (User Study 2) in an AI verification task for a hypothetical smart camera with privacy blur filters, label predictions and CAM explanations, i.e., the Simulation Study 1 prediction task. Both studies had a 3×3 factorial design with two independent variables – Blur Bias level (None $\sigma = 0$, Weak $\sigma = 16$, Strong $\sigma = 32$) and CAM type (Unbiased, Debiased, and Biased). Unbiased-CAM is the CAM from

RegularCNN predicting on the unbiased image regardless of blur bias level; Debiased-CAM is the CAM from DebiasedCNN (mb, mt) and Biased-CAM is the CAM from RegularCNN predicting on the biased image at corresponding Blur Bias levels. At the None blur level, Biased-CAM is identical to Unbiased-CAM. The user studies were approved by our university Institutional Review Board.

5.1 User Study 1 (CAM Truthfulness)

The first study evaluated the perceived truthfulness of Unbiased, Debiased, and Biased CAMs.

5.1.1 Experiment Procedure. Participants: 1) read the introduction and gave consent; 2) studied a tutorial about automatic image labeling, privacy blurring, heatmap explanations, and how to interpret the survey questions; 3) answered four screening questions to test their labeling of an unblurred and a weakly blurred image and their selection of important locations in an image and a CAM; 4) if screening was passed (all correct answers), answered background questions on technology savviness and image comprehension, performed the main study with 10 trials; and ended with demographic questions. See Supplementary Figs. 11-13 for questionnaire details.

In the main study (Fig. 5a), each participant viewed 10 repeated image trials, where each trial was randomly assigned to one of the three Blur Bias levels (within-subjects). All participants viewed the same 10 images (selection criteria described in Appendix C.1) in random order. For each trial, the participant: viewed a labeled unblurred image, indicated the most important locations on the image regarding the label with a “grid selection” UI (q1); and in the next page, viewed the blurred image, viewed CAMs of all 3 types generated from that and arranged randomly side-by-side, rated how well each CAM represented the image label on a 10-star scale (q2), and wrote her rating rationale (q3).

5.1.2 Experiment Apparatus and Measures. We used a “grid selection” user interface (UI) to measure objective truthfulness (Fig. 5b) to mitigate poor estimation of perceptions [6, 7, 32]. It overlays a clickable grid on the image for selecting important cells regarding the label. For usability, we limited the grid to 5×5 cells that can be selected or unselected (binary values). In the surveys, we referred to CAMs as “heatmaps”, which is a more familiar term. To compare the participant’s grid selection (User-CAM) with the heatmap shown (CAM), we aggregated CAM by averaging the pixel saliency in each cell and calculated **CAM Truthfulness Selection Similarity** as the Pearson’s Correlation Coefficient (PCC) between User-CAM and CAM. We also measured the **CAM Truthfulness Rating** as a subjective, self-reported rating on a uni-polar 10-point star scale (1 to 10). We collected the rationale of ratings as open-ended text. We measured the task time (per trial) as **Task Time Level** as low (<33 percentile), high (>66), medium, to account for response thoughtfulness. We tracked the **Image Label** of each image, since some types are easier to recognize even if blurred.

5.1.3 Participants. We recruited 36 participants from Amazon Mechanical Turk (AMT) with high qualification (≥ 5000 completed HITs with >97% approval rate). 32 participants passed screening, and completed the survey in a median time of 15.9 minutes and were compensated US\$2.00. They were 41.7% female and 23-69 years old (Median = 35).

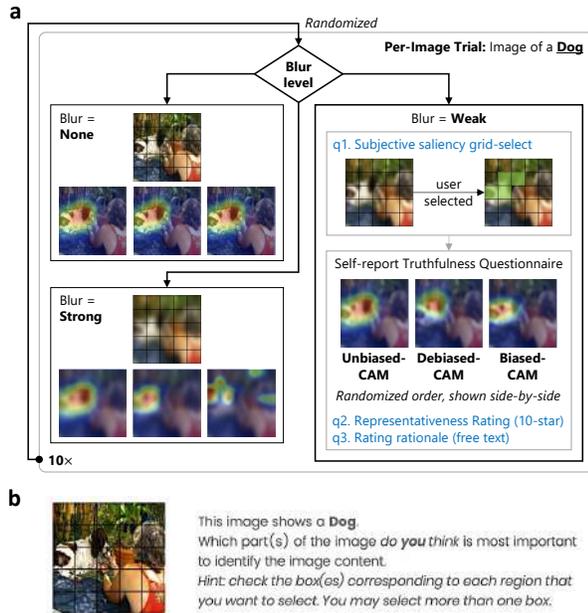


Figure 5: User Study 1 main study procedure (a) and grid selection UI to measure CAM Truthfulness Selection (b).

5.1.4 Statistical Analysis and Quantitative Results. For all dependent variables, we fit a multivariate linear mixed effects model with Blur Bias Level, CAM Types, Image Label and Task Time Level as fixed effects, Blur Bias Level \times CAM Type, Image Label \times Blur Bias Level, Image Label \times CAM Type, Task Time Level \times Blur Bias Level and Task Time Level \times CAM Type as fixed interaction effects, and Participant as a random effect.

Supplementary Table 3 reports the model fit (R^2) and significance of ANOVA tests for each fixed effect. Due to the large number of comparisons in our analysis, we consider differences with $p < .001$ as significant. This is sufficiently strict for a Bonferroni correction for 50 comparisons ($\alpha = .05/50$). Furthermore, all results reported were significant at $p < .0001$, unless otherwise stated. We performed post-hoc contrast tests for specific differences described. All statistical analyses were performed using JMP (v14.1.0).

Fig. 6 summarizes our results. Unbiased-CAM had the highest CAM Truthfulness Selection Similarity, while Biased-CAM the lowest Similarity that was only 21.3-43.7% of the truthfulness of Unbiased-CAM. Debiased-CAM had significantly higher CAM Truthfulness Selection Similarity than Biased-CAM at 69.4-79.0% of the truthfulness of Unbiased-CAM. Similarly, for blurred images, participants rated Unbiased-CAM as the most truthful ($M = 7.83$ out of 10, standard error = 0.12), followed by Debiased-CAM ($M = 6.00 \pm 0.21$ to 7.21 ± 0.18), and Biased-CAM as the least truthful ($M = 3.05 \pm 0.21$ to 4.98 ± 0.26). In summary, Debiased-CAM improved CAM truthfulness, despite stronger blur that reduced CAM truthfulness by highlighting wrong or unexpected regions, sizes, and shapes.

5.1.5 Thematic Analysis and Qualitative Findings. We analyzed the rationale of participant ratings to better understand how participants interpreted different CAMs as truthful or untruthful, and what visual features they perceived in images and CAMs. We performed a thematic analysis with open coding [64]. Two authors

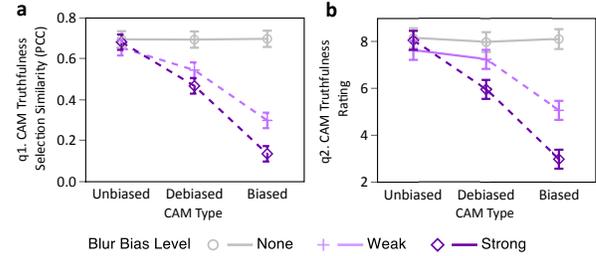


Figure 6: User Study 1 results. CAM Truthfulness decreased with blur, but was improved with Debiased-CAM. Dotted lines indicate very significant $p < .0001$ comparisons; solid lines indicate no significance at $p > .01$. Error bars indicate 90% confidence interval.

independently coded the rationales and discussed the coding until themes converged. Next, we first describe rationales for different blur levels, then describe themes spanning all blur levels. Note that all CAM types were shown anonymously (labeled A, B, and C) with randomized orders; we quote them specifically by type for clarity.

For None blur, as expected, most participants perceived CAMs as identical, e.g., “all 3 images are the same and mostly representative” (Participant P23, “Fish” image); though some participants could perceive the slight decrease in the CAM truthfulness of Debiased-CAM, e.g., for the “Church” image, P1 wrote that Unbiased-CAM and Biased-CAM “had the most focus on *all* the crosses on the roof of the church and therefore I thought they were the most representative. [Debiased-CAM] gives less importance to the leftmost cross on the roof and therefore was rated lower.” For Weak blur, participants felt Unbiased-CAM was very truthful, Debiased-CAM was slightly less truthful, and Biased-CAM was untruthful; e.g., P29 felt that Biased-CAM “doesn’t show anything but blackness, [other CAMs] are much better in the way the heatmap shows details.” For Strong blur, participants perceived Debiased-CAM as moderately truthful, but Biased-CAM as very untruthful, e.g., P18 felt that “[Biased-CAM] is totally off, nothing there is a garbage truck. [Unbiased-CAM] shows the best and biggest area, and [Debiased-CAM] is good too but I’m thinking not good enough as [Unbiased-CAM].”

Across blur conditions, we found that participants interpreted whether a CAM was truthful based on several criteria – primary object, object parts, irrelevant object, coverage span, and shape. Participants checked whether the primary object in the label was highlighted (e.g., “That heatmap that focuses on the chainsaw itself is the most representative.” P20, Chain Saw), and also checked whether specific parts of the primary object were included in the highlights (e.g., “[Unbiased-CAM and Debiased-CAM] correctly identify the fish though [Unbiased-CAM] also gives importance to the fish’s rear fin.” P1, Fish, Weak blur). P15 noted differences between the CAMs for the “French Horn” image: “[Unbiased-CAM] places the emphasis over the unique body of the French horn, and it places more well-defined, yellow and green emphasis on the mouthpiece and the opening of the horn itself. [Biased-CAM] is too vertical to completely capture the whole horn, and [Debiased-CAM]’s red area is too small to capture the body of the horn, and does not capture the opening of the horn or the mouthpiece.” Participants rated a CAM as less truthful if it highlighted irrelevant objects, e.g., “[Debiased-CAM] is quite close to capturing the entire church. (But) [Unbiased-CAM] captures more

of the tree.” (P26, Church). Much discussion also focused on the coverage of salient pixels. Less truthful CAMs had coverages that were either too wide (e.g., “[Debiased and Biased CAMs] are inaccurate. They are too wide.” P22, Garbage Truck), covering the background or other objects to get “less representative when it misleads you into the background or surroundings of the focus. It needs to only emphasize the critical area.” (P23, Church); or too narrow, not covering enough of the key object such that it “is very small and does not highlight the important part of the image. It is too narrow.” (P30, Fish). Finally, participants appreciated CAMs that highlighted the correct shape of the primary object, e.g., “[Debiased-CAM] perfectly captures the shape of the ball and all of its quadrants. [Unbiased-CAM] is a little more oblong than the golf ball itself, so it’s not as perfect. [Biased-CAM] is almost a vertical red spot and does not really capture the shape of the golf ball at all.” (P15, Golf Ball).

In summary, we found that Debiased-CAM and Unbiased-CAM were perceived as truthful, because they: 1) highlighted semantically relevant targets while avoiding irrelevant ones, so concept or object-aware CNN models are important [10, 43]; 2) had salient regions that were neither too wide nor narrow for the image domain; and 3) had accurate shape and edge boundaries for salient regions, which can be obtained from gradient explanations [74].

5.2 User Study 2 (CAM Helpfulness)

The second study evaluated the perceived helpfulness of each CAM type to verify predictions of blur biased images.

5.2.1 Experiment Procedure. The procedure is the same as User Study 1, except for the main study section. User Study 1 focused on CAM Truthfulness to obtain the participant’s saliency annotation of the unblurred image before revealing CAMs. In User Study 2, showing the unblurred image first will invalidate the use case of verifying predictions on blurred images, since the participant would have foreknowledge of the image. Hence, participants needed to see the blurred image and model prediction first, answer perception questions, then see the image unblurred.

In the main study (Fig. 7a), each participant viewed 7 repeated image trials, each randomly assigned to one of 9 conditions (3 Blur Bias levels × 3 CAM types) in a within-subjects experiment design. Participants viewed 7 randomly chosen images from the same 10 images of User Study 1, instead of all 10, so that they could not easily conclude the class label for the remaining images by eliminating previous classes. For each trial, the participant performed the common explainable AI task to verify the label prediction of the model. On the first page, the participant viewed a labeled image at the assigned Blur Bias level with corresponding CAM for the assigned CAM type, indicated her likelihood choice(s) for the image label with the “balls and bins” question [32] to elicit user labeling (Fig. 7b) (q1); rated how well each CAM represented the image label (q2); rated how helpful the CAM was for verifying the label (q3), and wrote the rationale for her rating (q4). On the next page, participants saw the image unblurred and answered questions q2-4 again as questions q5-7. See Supplementary Fig. 14 for questionnaire details.

5.2.2 Experiment Apparatus and Measures. For q1, we asked the participant to indicate likelihoods of 10 possible image labels with

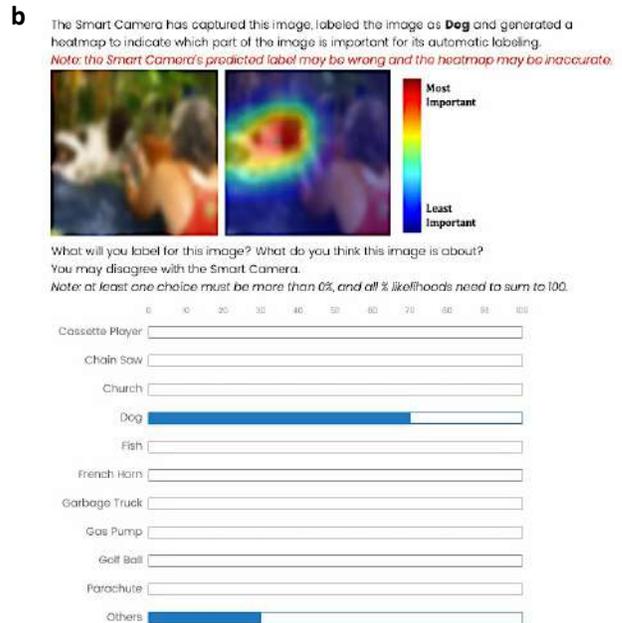
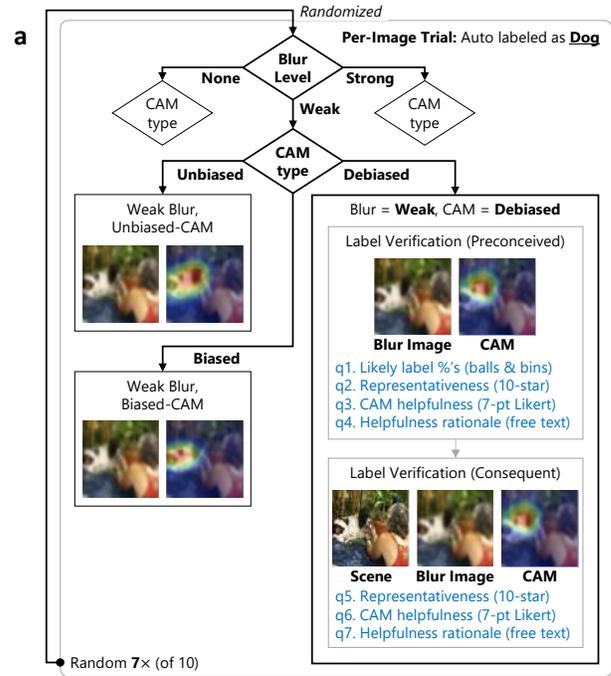


Figure 7: User Study 2 main study procedure (a) and “balls and bins” UI to elicit user labeling (b).

the “balls and bins” question [19, 32, 80] to elicit her probability distribution $p = \{p_c\}^T$ over label classes $c \in C$. This question is reliable in eliciting probabilities from lay users [32, 80] and avoids priming participants with the actual label c_0 , since it asks about all labels. We calculated the participant’s selected label \hat{c} as the class with the highest probability, i.e., $\hat{c} = \operatorname{argmax}_c(p_c)$, **Labeling Confidence** as the indicated likelihood for the actual label p_{c_0} , and **Label Correctness** as $[\hat{c} = c_0]$, where $[\cdot]$ is the Iverson bracket notation.

We measured the perceived **CAM Helpfulness** and **CAM Truthfulness Ratings** on a bipolar 7-point Likert scale ($-3 = \text{Strongly Disagree}$, $+3 = \text{Strongly Agree}$). We collected rating rationale as open-ended text. We used different formats for CAM Truthfulness and CAM Helpfulness to mitigate repetitive or copied responses and to allow for a more precise measurement of CAM Truthfulness. We also measured **Task Time Level** and **Image Label** per trial.

5.2.3 Participants. We recruited 191 new participants from AMT with the same qualification criteria as User Study 1. 162 participants passed screening, completed the survey in a median time of 18.4 minutes and were compensated US\$2.00. They were 46.0% female and between 21 and 74 years old (Median = 37). We excluded 7 participants who gave wrong labels for $>60\%$ of encountered unblurred images, which indicated the participant's poor recognition ability.

5.2.4 Statistical Analysis and Quantitative Results. For each dependent variable, we fit a multivariate linear mixed effects model with the same fixed, interaction, and random effects as in User Study 1. We further analyzed CAM Truthfulness and Helpfulness ratings with fixed main and interaction effects regarding whether users rated before or after seeing the unblurred version of the image, i.e., Unblurred Disclosure: preconceived or consequent. Supplementary Table 4 reports the model fit (R^2) and ANOVA tests for each fixed effect, and report significant results similarly to User Study 1.

Fig. 8 summarizes our results from 1,085 trials of 155 included participants. Differences in decision quality (Labeling Correctness and Labeling Confidence) across CAM types depended on blur bias level. For None blur, the decision quality was high for all CAM types (confidence $M = 95.6\% \pm 0.6\%$, correctness $M = 99.0\% \pm 0.5\%$) due to the ease of the tasks, while for Strong blur, the decision quality was low for all CAM types (confidence $M = 68.5\% \pm 1.8\%$, correctness $M = 79.9\% \pm 2.2\%$), suggesting that blurring was too strong even for truthful CAMs to be useful. However, for Weak blur, Debiased-CAM reduced labeling error by 1.92x ($1 - \text{Correctness}$: from $18.2\% \pm 3.5\%$ to $9.5\% \pm 2.9\%$) and improved confidence from $75.4\% \pm 2.9\%$ to $82.8\% \pm 2.7\%$ compared to Biased-CAM. We found stronger differences in preconceived ratings of CAM types. For Weak blur, participants rated Debiased-CAM as more truthful ($M = 7.7 \pm 0.2$ vs. 5.6 ± 0.3 out of 10) and more helpful ($M = 1.56 \pm 0.13$ vs. 0.15 ± 0.19 on a 7-point Likert scale from -3 to 3) than Biased-CAM. Moreover, for Strong blur, although their decision quality did not improve, participants perceived Debiased-CAM as more truthful ($M = 6.4 \pm 0.2$ vs. 4.4 ± 0.3) and helpful ($M = 0.60 \pm 0.16$ vs. -0.49 ± 0.19) than Biased-CAM. These effects were similar and slightly amplified for consequent ratings (Fig. 8), indicating that users more strongly appreciated Debiased-CAM and disliked Biased-CAM if they had foreknowledge of the unblurred scenes. In summary, Debiased-CAM recovered the usefulness of CAMs for moderately blurred images, and were perceived as helpful even for strong blur.

5.2.5 Thematic Analysis and Qualitative Findings. To understand why participants rated CAMs as helpful or unhelpful, we performed a thematic analysis on rationales, similarly to User Study 1. Rationale depended much on image Blur Bias level, and we identified how truthful and helpful debiased CAMs were even for blurred images.

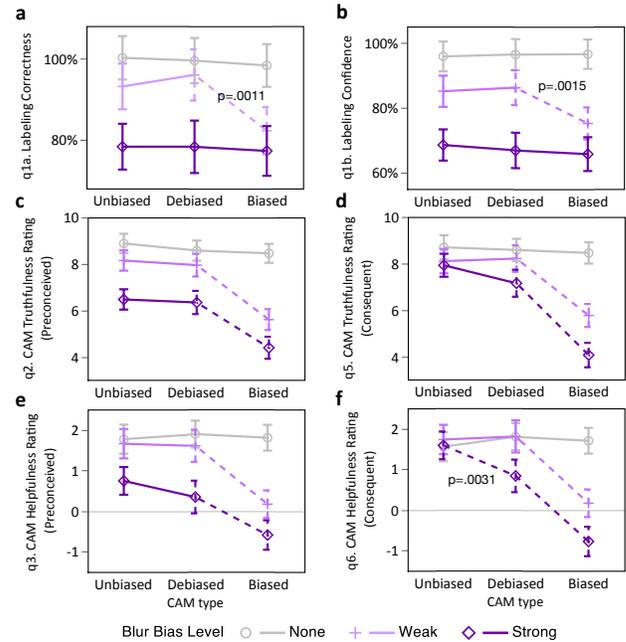


Figure 8: Quantitative results for User Study 2. Decision correctness (a,b) and perceived ratings (c-f) decreased with stronger blur, but Debiased-CAM improved them to be similar as with Unbiased-CAM.

For unblurred images (None blur level), participants mostly felt that CAMs were helpful, because CAMs helped to: 1) focus their attention “on the most important part of the image, which helps me to quickly identify and label the image.” (Participant P106, Garbage Truck); 2) ignore irrelevant targets to “let me know I can disregard the person in the foreground” (P89, Dog), “It helps hone in on what the content is, and helps to ignore the extra things in the frame.” (P14, Chain Saw); 3) matched their expectations since CAMs “did a solid job of identifying the garbage truck.” (P36) and was “highly correlated to where the fish is in this image.” (P38). Conversely, as expected, many participants considered CAMs unhelpful since “I could easily identify the object in the image without the heatmap” (P32, Church).

For images with Weak blur, a truthful CAM: 4) “helps focus my attention to that area on the blurry picture” (P105, Debiased-CAM), “clearly give hint on what was needed to notice in the photo.” (P140, Unbiased-CAM); and 5) helped to confirm image labels, e.g., P3 felt that “the heatmap gives me the idea that the object might be a fish, I could not tell otherwise” and wrote after seeing the unblurred image that “I wouldn’t have known what the object was without the heatmap.” P118 described how Unbiased-CAM “pointed to the steeple and it helped me realize that it was indeed a picture of a church. I had trouble recognizing it on my own.” Debiased-CAMs helped to locate suspected objects in unexpected images, e.g., P96 felt that “based on what the heatmap is marking, that’s the exact spot where someone would hold a french horn”, and P67 noted “that is not an area where I would expect to find a fish, so it’s helpful to have this guide.”

For images with Strong blur, many participants felt that the CAMs were very unhelpful, because 6) the task was too difficult such that they had “NO idea what image is and heatmap doesn’t

help.” (P68, Biased-CAM), felt the task “*was very hard, i could not figure it out*” (P71, Debiased-CAM), did not have much initial trust as “*I feel that the heatmap could be wrong because of the clarity of the image.*” (P62, Unbiased-CAM). Some participants would 7) blindly trust the CAM due to a lack of other information such that “*without the heatmap and the suggestion, I would have no guess for what this is. I am flying a bit blind. So, I concur with the recommendation (french horn) until I see more.*” (P92, Unbiased-CAM) and due to the trustful expectation that CAM “*enables me to know the most useful part in the camera.*” (P138, Church, Unbiased-CAM). Finally, we found that 8) confirmation bias may cause the CAM correctness to be misjudged. For example, P76 first thought a misleading Biased-CAM “*helps make a blurry picture more clear*”, but later realized “*it’s in the wrong spot.*” (“Garbage Truck”); in contrast, P24 wrongly accused that an Unbiased-CAM “*was focused on the wrong thing*”, but changed his opinion after seeing the unblurred image, admitting “*Now that I see it’s a dog, it is more clear.*”

In summary, these findings explain why truthful Debiased-CAM and Unbiased-CAM helped participants to verify classifications of unblurred or weakly blurred images. For unblurred images, these CAMs: 1) focused user attention to relevant objects to speed up verification, 2) averted attention from irrelevant targets to simplify decision making, and 3) matched user expectations [73] of the target object shapes. For weakly blurred images, these CAMs: 4) provided hints on which parts to study in blurred images, and 5) supported hypothesis formation and confirmation [80, 87] of suspected objects. For strongly blurred images, participants generally rated all CAMs as unhelpful because: 6) verifying the images was too difficult, 7) they felt misguided to blindly trust CAMs, and 8) they misjudged CAMs based on preconceived notions, i.e., confirmation bias [87].

6 DISCUSSION AND DESIGN IMPLICATIONS

Our results highlighted issues in explanation faithfulness when CNN models explain their predictions on images subjected to systematic error bias, which we addressed by with Debiased-CAM to improve explanation truthfulness and helpfulness, and consequently the prediction performance. We discuss implications for XAI and HCI researchers and practitioners to: 1) be wary of how contexts and corruptions can make explanations misleading, 2) support scalable human-centric explanations, 3) extend debiasing to social contexts, and 4) carefully design unconfounded user studies to evaluate XAI. We also discuss 5) generalizations of our debiasing method to other XAI techniques and data types.

6.1 Physical contextual bias in explanations

Most AI explanations have been developed to support model debugging, help end-users identify incorrect model reasoning, or trust correct explanations. However, we have shown that moderate systematic error (biases) in data, which may seem innocuous, can lead to severe deviations in explanations, despite model fine-tuning. For baseline models, these truly reflect the model reasoning, but they can be incongruent with user expectations, and can harm user trust.

We have investigated three prevalent sources of bias — blur, color, lighting — that can plausibly occur in image applications, and were feasibly manipulable to test in evaluations. We showed

that CAM deviations are significant in such cases, yet Debiased-CAM can improve them. In pilot studies, we have also investigated other bias types listed in [36]. We found that environmental biases (e.g., snow, fog, frost) cause moderate CAM deviations that will require debiasing. Some biases had very weak CAM deviations (e.g., brightness, contrast), since the image pixel changes were monotonic which does not affect the additive activations in neural networks much. Debiasing may not be needed for such biases. Finally, we found that biases due to image processing and compression artifacts (e.g., Gaussian noise, JPEG) had small CAM deviations. We expect other blur biases (e.g., defocus, frosted glass, motion, zoom) to have slightly stronger, but similar effects as Gaussian blur that we evaluated, since the images remain similar to the original.

However, images subjected to adversarial noise [23] would be particularly concerning, since an attacker can intelligently and maliciously inject noise to deliberately harm the performance or explanation, such that CAM deviations will be worse. Training Debiased-CAMs under such attacks may be more difficult.

6.2 Scalable human-centric explanations

The needs for model explainability are diverse. Langer et al. cataloged many desiderata of explanations [55], including several societal objectives such as agreeability, auditability, fairness and privacy. To support human interpretability at the cognitive level, explanations need to conform to human prior knowledge [28, 53, 73, 79], human reasoning processes [63, 87], and human perceptual processes [96]. In this work, we focus on improving the *agreeability* of explanations towards human prior knowledge. This typically requires manual inspection and annotation by people [52, 73], which is labor-intensive. Instead, given a clean dataset with agreeable explanations, we train our model to produce debiased explanations. This uses self-supervision, so it is also scalable and not labor-intensive. Users may only need to select images rather than annotate details in each image. For our studies, we had assumed that unbiased images have reliable explanations, but this should be verified by human labelers. An interface to support quick ratings of explanation acceptability would help to accelerate this data curation. Another scalable strategy involves defining axioms (e.g., attribution priors [28], psychological preferences [88], or visual cognitive chunks [2]) and constraining explanations towards them. From our qualitative analysis, we identified desiderata for truthful saliency maps (e.g., trace the shape of relevant objects, control the spread or tightness of hot spots) that can be used as general axioms for faithful saliency maps. This further increases scalability by reducing the dependency on selecting reliable explanation references.

6.3 Debiasing explanations against social bias

Although we have focused on bias due to physical contexts, bias in social situations also needs debiasing. People are subjected to *egocentric bias* [48] and *societal discrimination (unfairness)* [22]. With egocentric bias, different stakeholders would prioritize their own objectives [27] and may be ignorant of other viewpoints. Debiased explanations could encode different interpretation preferences (e.g., [53, 88]) to show how slightly different two stakeholders interpret a decision (e.g., patient and doctor for medical diagnosis). With social bias, models may predict or reason undesirably for some protected

groups of people based on sensitive attributes (race, gender, etc.). For example, saliency maps can detect bias in a model by highlighting a female face for Nurse, but highlighting a stethoscope held by a woman for Doctor [78]. Instead of the current approach to debias models with data balancing, our debias approach can retrain models to de-emphasize focusing on sensitive concepts (e.g., faces). However, we caution about the dark pattern of debiasing explanations to make an unfair model appear fair by retraining its explanation to appear fair (e.g., [23, 24]).

6.4 Sensitive measures for faithful explanation

Saliency map explanations have mostly been evaluated with simulation metrics and rarely with human subjects [5, 44]. User studies are important to verify the severity of problems (*perceptually noticeable enough?*) and the efficacy of solutions (*problem no longer perceivable or perceptually forgivable?*). However, designing successful experiments with strong effects and sensitive measures is difficult and many studies fail to find effects [9, 44, 69]. To improve the sensitivity, experiments need more sensitive measures and carefully designed participant tasks.

Current user studies use simple true/false or multiple choice responses and confidence ratings, but these measures are prone to lucky guesses, do not capture secondary choices, or suffer from social desirability bias. The insensitivity of such methods could have led to null results [5, 44]. Instead, we employed more sensitive and objective measures of labeling likelihood ("balls and bins" question). We also measured explanation agreement objectively, since users tend to over-trust wrong explanations [42], affecting the validity of subjective ratings. Hence, we employed the grid-selection UI, which is similar to segment selection in [99]. Another method is to ask participants to write important and ignored features in the free text [5], but this is difficult to automatically evaluate.

Explanation understanding is typically evaluated with human simulatability tasks [59, 60], where users try to predict what a model would predict. However, participant answers may be confounded by leaking information that participants are tested on. Zhang et al. [99] evaluated saliency using a reverse-ablation method to incrementally reveal important segments and ask participants to label the image; this avoids the hindsight bias effect [72]. In this work, we controlled when to pose questions. To measure perceived truthfulness, we first measured objective ground truth before showing CAMs to avoid participants copying or being primed. To measure perceived helpfulness in a privacy-preserving application, we posed questions twice, first with blurred images, then unblurred images. This mitigated the hindsight bias effect. Thus, we add sensitive experiment apparatuses to the literature on evaluating XAI.

6.5 Generalization to other XAI and data types

Our self-supervised debiasing can apply to other gradient-based explanations [8, 81, 83] by formulating the activation, gradient or propagated terms as a secondary prediction task. However, some saliency explanations, such as Layer-wise Relevance Propagation (LRP) [8] and Integrated Gradients [83], which produce fine-grained "edge detector" heatmaps [3] are likely to be more severely degraded with biasing, such as strong blurring. Beyond gradient-based explanations, model-agnostic explanations such as LIME [71] and Kernel

SHAP [61] can be debiased by regularizing on a saliency loss metric. Notably, CNN explanation techniques such as feature visualizations [10, 66] and neuron attention [57] have higher dimensionality that requires more sensitivity to debias. Dimensionality reduction with autoencoders or generative adversarial networks (GANs) could provide latent features that are feasible to debias. Finally, concept-based explanations such as TCAV [43] and RexNet [96] can be debiased to align the generated concept with user expectations.

Debiased-CAM can be generalized to other types of data subjected to bias, particularly those that can be modeled with CNNs, such as audio and time series data. Other than biases in images, debiasing is also necessary for explaining model predictions of other data types and behaviors, such as audio signals with noise or obfuscation [62], and human activity recognition with inertial measurement units (IMU) or other wearable sensors [75]. With the prevalence of data bias in the real-world and privacy obfuscation, Debiased-CAM provides a generalizable framework to train robust performance and faithful explanations for responsible AI.

7 CONCLUSION

We highlight issues in explanation faithfulness when CNN models explain their predictions on images that are biased with systematic error, and address this by developing Debiased-CAM to improve the truthfulness of explanations. We achieved these improvements by ensuring that model parameters were learned based on more important attention as identified by unbiased explanations and on more diverse inputs due to data augmentation across multiple bias levels. We also implemented more precise training with multiple prediction tasks and differentiable explanation loss. Our results showed that even when image data were degraded or distorted due to bias, 1) they retained sufficient useful information that DebiasedCNN could learn to recover salient locations of unbiased explanations, and 2) these salient locations were highly relevant to the primary task such that prediction performance could be improved.

ACKNOWLEDGMENTS

This work was supported in part by the Ministry of Education, Singapore under the grant T2EP20121-0040 and the NUS Centre for Research in Privacy Technologies (N-CRIPT).

REFERENCES

- [1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.
- [2] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*. 9505–9515.
- [4] Mahmoud Afifi and Michael S Brown. 2019. What else can fool deep learning? Addressing color constancy errors on deep neural network performance. In *Proceedings of the IEEE International Conference on Computer Vision*. 243–252.
- [5] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.
- [6] Ronald Angel and William Gronfein. 1988. The use of subjective information in statistical models. *American Sociological Review* (1988), 464–473.

- [7] Daniel Avrahami, James Fogarty, and Scott E Hudson. 2007. Biases in human estimation of interruptibility: effects and implications for practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 50–60.
- [8] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one* 10, 7 (2015), e0130140.
- [9] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [10] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6541–6549.
- [11] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2018. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence* 41, 3 (2018), 740–757.
- [12] Mitchell Charity. [n.d.]. What color is a blackbody? - some pixel rgb values. <http://www.vendian.org/mncharity/dir3/blackbody/>.
- [13] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 839–847.
- [14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).
- [16] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 326–335.
- [17] Robert G Davis and Dolores N Githner. 1990. Correlated color temperature, illuminance level, and the Kruithof curve. *Journal of the Illuminating Engineering Society* 19, 1 (1990), 27–38.
- [18] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [19] Adeline Delavande and Susann Rohwedder. 2008. Eliciting subjective probabilities in Internet surveys. *Public Opinion Quarterly* 72, 5 (2008), 866–891.
- [20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [21] Mariella Dimiccoli, Juan Marin, and Edison Thomaz. 2018. Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–18.
- [22] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.
- [23] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems* 32 (2019), 13589–13600.
- [24] Ann-Kathrin Dombrowski, Christopher J. Anders, Klaus-Robert Müller, and Pan Kessel. 2022. Towards Robust Explanations for Deep Neural Networks. *Pattern Recognit.* 121 (2022), 108194.
- [25] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- [26] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.
- [27] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [28] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. 2021. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence* (2021), 1–12.
- [29] Andre Esteva, Brett Kuppel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.
- [30] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3429–3437.
- [31] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3681–3688.
- [32] Daniel G Goldstein and David Rothschild. 2014. Lay understanding of probability distributions. *Judgment & Decision Making* 9, 1 (2014).
- [33] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.
- [34] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, and Rami Albalal. 2016. Ntcir lifelog: The first test collection for lifelog research. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 705–708.
- [35] Leif Hancox-Li. 2020. Robustness in machine learning explanations: does it matter?. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 640–647.
- [36] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (2019).
- [37] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2020. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)* (2020).
- [38] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics* 25, 8 (2018), 2674–2693.
- [39] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. 2019. S ummit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1096–1106.
- [40] Jeremy Howard. [n.d.]. The imagenette dataset. <https://github.com/fastai/imagenette>. Github.
- [41] Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Chau. 2017. ActiVis: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 88–97.
- [42] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna M. Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists’ Use of Interpretability Tools for Machine Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).
- [43] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. PMLR, 2668–2677.
- [44] Jacob Kittley-Davies, Ahmed Alqaraawi, Rayoung Yang, Enrico Costanza, Alex Rogers, and Sebastian Stein. 2019. Evaluating the effect of feedback from different computer vision processing stages: a comparative lab study. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [45] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2390–2395.
- [46] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International Conference on Machine Learning*. PMLR, 5338–5348.
- [47] Nikos Komodakis and Sergey Zagoruyko. 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*.
- [48] James Konow. 2005. Blind spots: The effects of information and stakes on fairness bias and dispersion. *Social Justice Research* 18, 4 (2005), 349–390.
- [49] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.
- [50] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. 2018. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8183–8192.
- [51] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. 2014. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on graphics (TOG)* 33, 4 (2014), 1–11.
- [52] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006* (2019).
- [53] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2018. Human-in-the-loop interpretability prior. *Advances in neural information processing systems* 31 (2018).
- [54] Himabindu Lakkaraju and Osbert Bastani. 2020. “How do I fool you?” Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.

- [55] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.
- [56] Jia Li, Changqun Xia, Yafei Song, Shu Fang, and Xiaowu Chen. 2015. A data-driven metric for comprehensive evaluation of saliency models. In *Proceedings of the IEEE international conference on computer vision*. 190–198.
- [57] Kunpeng Li, Ziyang Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. 2018. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9215–9223.
- [58] Brian Y Lim and Anind K Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing*. 415–424.
- [59] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.
- [60] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
- [61] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774.
- [62] Ian McLoughlin, Haomin Zhang, Zhipeng Xie, Yan Song, and Wei Xiao. 2015. Robust sound event classification using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 3 (2015), 540–552.
- [63] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
- [64] Michael J Muller and Sandra Kogan. 2010. Grounded theory method in HCI and CSCW. *Cambridge: IBM Center for Social Software* 28, 2 (2010), 1–46.
- [65] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105.
- [66] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature visualization. *Distill* 2, 11 (2017), e7.
- [67] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
- [68] Michael I Posner, Charles R Snyder, and R Solso. 2004. Attention and cognitive control. *Cognitive psychology: Key readings* 205 (2004).
- [69] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.
- [70] Harish Guruprasad Ramaswamy et al. 2020. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. In *The IEEE Winter Conference on Applications of Computer Vision*. 983–991.
- [71] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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*. 1135–1144.
- [72] Neal J. Roese and Kathleen D. Vohs. 2012. Hindsight Bias. *Perspectives on Psychological Science* 7 (2012), 411–426.
- [73] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2662–2670.
- [74] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
- [75] Michael Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. 2017. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31.
- [76] Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS one* 10, 3 (2015), e0118432.
- [77] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence* 2, 8 (2020), 476–486.
- [78] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
- [79] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2591–2600.
- [80] William F Sharpe, Daniel G Goldstein, and Phil W Blythe. 2000. The distribution builder: A tool for inferring investor preferences. *preprint* (2000).
- [81] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. (2014).
- [82] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.
- [83] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365* (2017).
- [84] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.
- [85] Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. 2016. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint arXiv:1611.05760* (2016).
- [86] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.
- [87] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
- [88] Danding Wang, Wencan Zhang, and Brian Y Lim. 2021. Show or suppress? Managing input uncertainty in machine learning model explanations. *Artificial Intelligence* 294 (2021), 103456.
- [89] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 2020. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 24–25.
- [90] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.
- [91] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.
- [92] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
- [93] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.
- [94] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8827–8836.
- [95] Quan-shi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 27–39.
- [96] Wencan Zhang and Brian Y Lim. 2022. Towards Relatable Explainable AI with the Perceptual Process. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).
- [97] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. 2020. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*.
- [98] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.
- [99] Zijian Zhang, Jaspreet Singh, Ujjwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
- [100] Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Lim. 2021. Exploiting Explanations for Model Inversion Attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 682–692.
- [101] Stephan Zheng, Yang Song, Thomas Leung, and Ian J. Goodfellow. 2016. Improving the Robustness of Deep Neural Networks via Stability Training. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 4480–4488.
- [102] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.
- [103] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 119–134.

A TECHNICAL APPROACH APPENDIX

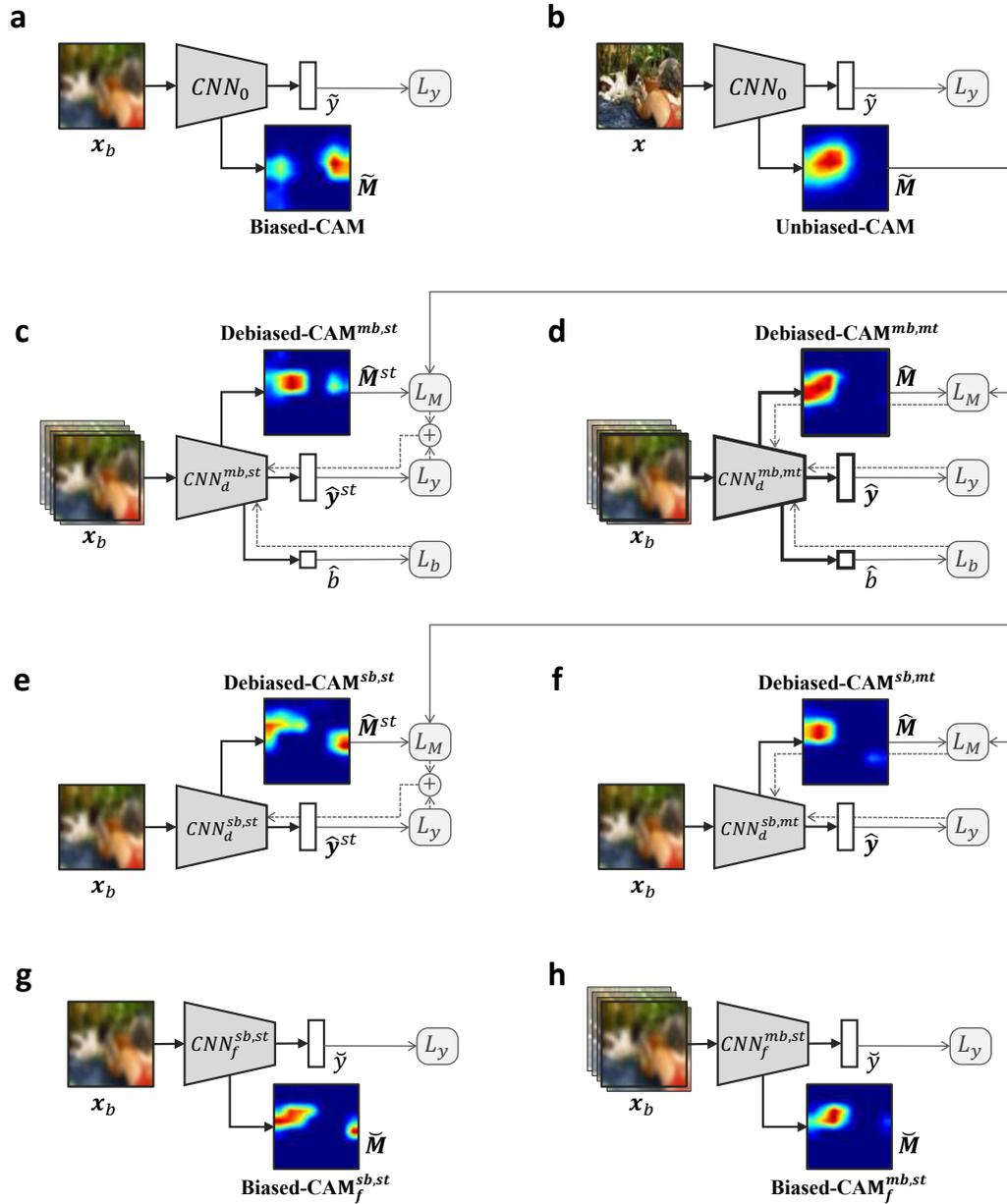
A.1 Datasets and model implementation details

In simulation studies, we evaluated the models on three datasets for two image tasks (summarized in 1). For Simulation Study 1 (Blur Bias), we used Inception v3 [84] pretrained on ImageNet ILSVRC-2012 [20] and fine-tuned on blur biased images of ImageNette [40], which is a subset of ILSVRC-2012. We only retrained layers from the last two Inception blocks of the Inception v3 model. For Simulation Studies 2 and 4 (Blur and Color Temperature Bias on egocentric activity images), we also used Inception v3 pretrained on ILSVRC-2012, and fine-tuned it on the NTCIR-12 [34]. For Simulation Study 3 (Blur Bias Captioning), we used the Neural Image Captioner (NIC) [86] with Inceptionv3-LSTM model and fine-tuned on blur biased images from COCO [15]. We retrained the last two inception blocks of Inception v3 as well as LSTM blocks. For Simulation Study 5 (Lighting Bias), we fine-tuned the Inception v3 (pretrained on ILSVRC-2012) on the Transient Attribute database (TransAttr) [51] for multi-label classification. We limited our evaluations to four labels: Snowy, Sunny, Cloudy, Dawn/Dusk. All model hyperparameters were tuned using the Adam optimizer with batch size 64 and learning rate 10^{-5} .

Exp	Task	Model	Re-trained Dataset	Dataset Size	Train-Test Ratio
1	Classify	Inception v3 CNN	ImageNette	13,395 images	70.0% / 30.0%
2	Classify	Inception v3 CNN	NTCIR-12	44,902 images	80.0% / 20.0%
3	Caption	Neural Image Captioner (Inception v3 + LSTM)	COCO	123,287 images, 616,435 captions	66.7% / 33.3%
4	Classify	Inception v3 CNN	NTCIR-12	44,902 images	80.0% / 20.0%
5	Classify	Inception v3 CNN	TransAttr	4,584 images	80.0% / 20.0%

Supplementary Table 1. **Baseline CNN models trained on training datasets for Simulation Studies. All models were pre-trained on ImageNet ILSVRC-2012 and retrained to fine-tune on respective datasets. Train-test ratios were determined from the original literature as referenced.**

A.2 Model Variants

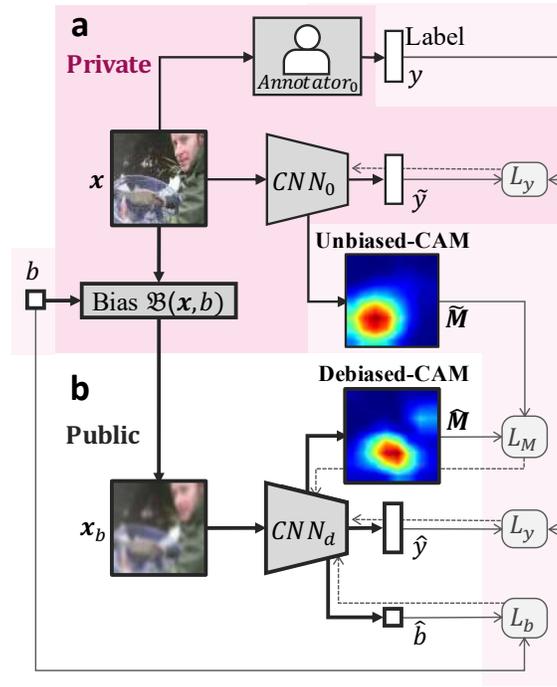


Supplementary Fig. 1. Architectures of self-supervised DebiasedCNN variants and of baseline CNN models and their CAM explanations from a biased “Dog” image blurred at $\sigma = 24$. a) RegularCNN on biased image. b) RegularCNN on unbiased image. c) DebiasedCNN (mb, st) with single-task loss as a sum of classification and CAM losses for the classification task, trained on multi-bias images with auxiliary bias level prediction task. d) DebiasedCNN (sb, mt) with multi-task for CAM prediction trained with differentiable CAM loss, and trained on multi-bias images with auxiliary bias level prediction task. e) DebiasedCNN (sb, st) with single-task loss as a sum of classification and CAM losses for the classification task. f) DebiasedCNN (sb, mt) with multi-task for the CAM prediction and differentiable CAM loss. g) FineTunedCNN (sb, st) retrained on images biased at a single-bias level. h) FineTunedCNN (mb, st) retrained on images biased variously at multi-bias levels.

Model Variant	Training Loss Function	Training Set Bias Levels
--- RegularCNN	$L(\mathbf{w}) = L_y(y, \hat{y}(\mathbf{w}))$	$b = 0$
--- FineTunedCNN (sb, st)	$L(\mathbf{w}) = L_y(y, \hat{y}(\mathbf{w}))$	$b \in (0, b_{max}]$
--- FineTunedCNN (mb, st)	$L(\mathbf{w}) = L_y(y, \hat{y}(\mathbf{w}))$	$b \in B_{rand} \sim U([0, b_{max}])$
--- DebiasedCNN (sb, st)	$L(\mathbf{w}) = L_y(y, \hat{y}(\mathbf{w})) + \omega_M L_M(\tilde{\mathbf{M}}, \hat{\mathbf{M}})$	$b \in (0, b_{max}]$
--- DebiasedCNN (mb, st)	$L(\mathbf{w}) = \begin{pmatrix} L_y(y, \hat{y}(\mathbf{w})) + \omega_M L_M(\tilde{\mathbf{M}}, \hat{\mathbf{M}}) \\ \omega_b L_b(b, \hat{b}(\mathbf{w})) \end{pmatrix}$	$b \in B_{rand}$
--- DebiasedCNN (sb, mt)	$L(\mathbf{w}) = \begin{pmatrix} L_y(y, \hat{y}(\mathbf{w})) \\ \omega_M L_M(\tilde{\mathbf{M}}, \hat{\mathbf{M}}(\mathbf{w})) \end{pmatrix}$	$b \in (0, b_{max}]$
--- DebiasedCNN (mb, mt)	$L(\mathbf{w}) = \begin{pmatrix} L_y(y, \hat{y}(\mathbf{w})) \\ \omega_M L_M(\tilde{\mathbf{M}}, \hat{\mathbf{M}}(\mathbf{w})) \\ \omega_b L_b(b, \hat{b}(\mathbf{w})) \end{pmatrix}$	$b \in B_{rand}$

Supplementary Table 2. CNN model variants with single-task (st) or multi-task (mt) architectures trained on a specific (sb) or multiple (mb) bias levels. Each training set image $x \in X$ is preprocessed by a bias operator \mathfrak{B} at a selected level b , i.e., $x_b = \mathfrak{B}(x, |b| > 0)$, $\forall x \in X$. \mathfrak{B} depends on the bias type (e.g., blur, color temperature, day-night lighting). For DebiasedCNN, mt refers to including a CAM task with differentiable CAM loss separate from the primary prediction task, while st refers to the primary prediction task with non-differentiable CAM loss. Models trained for single-bias (sb) used training set images biased at a single level $b > 0$, while models trained for multi-bias levels (mb) used training datasets with data augmentation where each image is biased to a level that is randomly selected from a uniform probability distribution $B_{rand} \sim U([0, b_{max}])$. Multi-bias DebiasedCNN also adds a task for bias level prediction. Loss functions in vector form specify one loss function per task in a multi-task architecture.

A.3 Debiasing spurious explanations of privacy-preserving AI



Supplementary Fig. 2. Architecture of multi-task DebiasedCNN model with self-supervised learning from private training data for privacy-preserving machine learning. a) RegularCNN (CNN_0) was trained on a private dataset with unblurred image x to generate Unbiased-CAM \tilde{M} . b) DebiasedCNN (CNN_d) was trained on the corresponding public (privacy-protected) biased form of the private dataset with blurred image x_b and self-supervised with Unbiased-CAM \tilde{M} to generate Debiased-CAM \hat{M} . During model training, CNN_d has access to the bias level b of each image x_b , Unbiased-CAM \tilde{M} , and actual label y , but has no access to them during model inference. CNN_d never has access to any unblurred image x . At inference time, DebiasedCNN can generate relevant and faithful Debiased-CAMs from privacy-protected blurred images.

B SIMULATION STUDIES APPENDIX

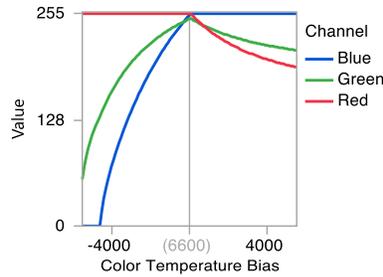
B.1 Supplemental Method: Calculating Bias Levels

We provide details to calculate different bias for color temperature and lighting biases.

B.1.1 Color Temperature Bias. Color temperature refers to the temperature of an ideal blackbody radiator as if illuminating the scene. We biased color temperature as follows. Each pixel in an unbiased image has color $(r, g, b)^T$, where R, G, B represent the red, green, and blue color values within range 0-255, respectively. Each pixel is biased from neutral temperature t_0 by Δt_b at bias level b by multiplying a diagonal correction matrix with its color, i.e.,

$$(r_b, g_b, b_b)^T = \text{diag}(255/R_b, 255/G_b, 255/B_b)(r, g, b)^T, \quad (5)$$

where $(R_b, G_b, B_b)^T = f_{CT}(T) = f_{CT}(t_0 + \Delta t_b)$ are scaling factors obtained from Charity’s color mapping function f_{CT} to map a blackbody temperature to RGB values [12] (Supplementary Fig. 3). We set the neutral color temperature t_0 to 6600K, which represents cloudy/overcast daylight. Color temperature biasing is asymmetric about zero bias, because people are more sensitive to perceiving changes in orange than blue colors (Kruithof Curve [17]); and due to the non-linear monotonic relationship between blackbody temperature and modal color frequency (Wien’s Displacement Law). This asymmetry explains why orange biasing led to stronger CAM deviation than blue biasing.



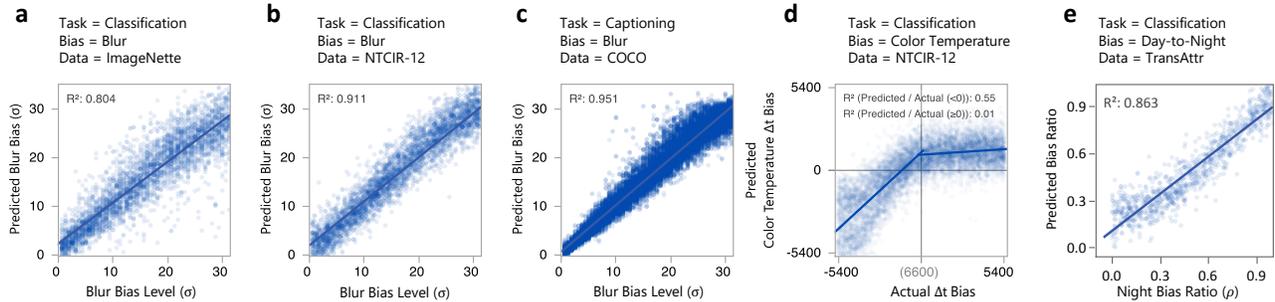
Supplementary Fig. 3. **Color mapping function to bias color temperature of images in Simulation Study 4. Changes in Red, Green, Blue values are larger for orange biases (lower color temperature) than blue biases (higher temperature). Neutral color temperature is set to represent shaded/overcast skylight at 6600K.**

B.1.2 Lighting Bias. Lighting bias occurs when the same scene is lit brightly or dimly. In nature, this occurs as sunlight changes hour-to-hour, or season-to-season. The Transient Attributes database [51] contains photos of scenes from the same camera position taken across different times of the day and year. Attribute changes include whether the scene is daytime or nighttime, snowy, foggy, dusk/dawn or not. We sought to generate images with different degrees of darkness, but the dataset only contained photos that were very bright or very dark. Therefore, we interpolated photos to generate scenes with intermediate darkness. For each scene, with a daytime image $I_{day}(x, y)$ and nighttime image $I_{night}(x, y)$, we performed the pixel-wise interpolation as,

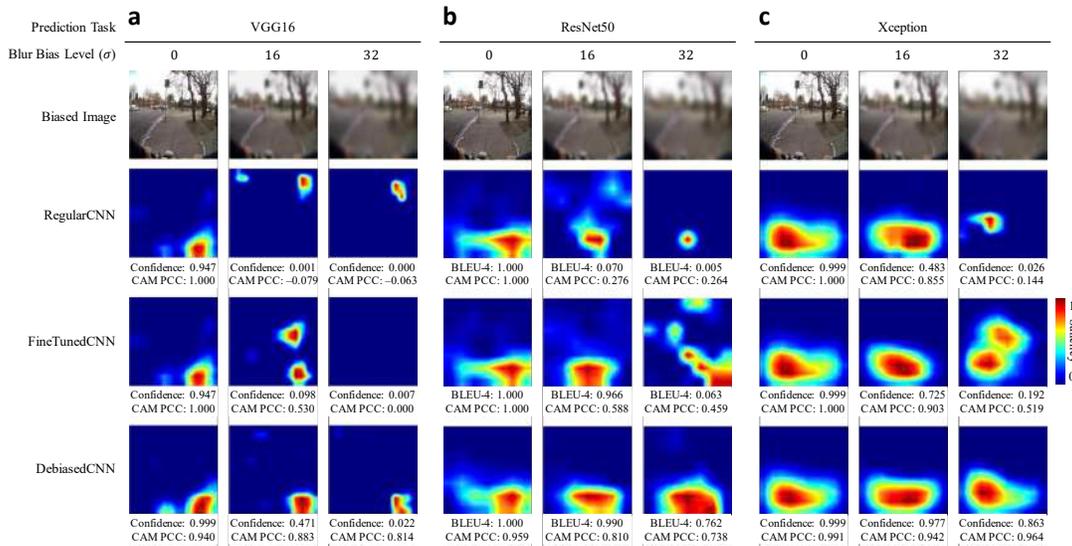
$$I_{biased}(x, y) = (1 - \rho) \times I_{day}(x, y) + \rho \times I_{night}(x, y), \quad (6)$$

where ρ is the night/day ratio. An unbiased image has $\rho = 0$ indicating daytime, and the most biased image has $\rho = 1$ indicating nighttime.

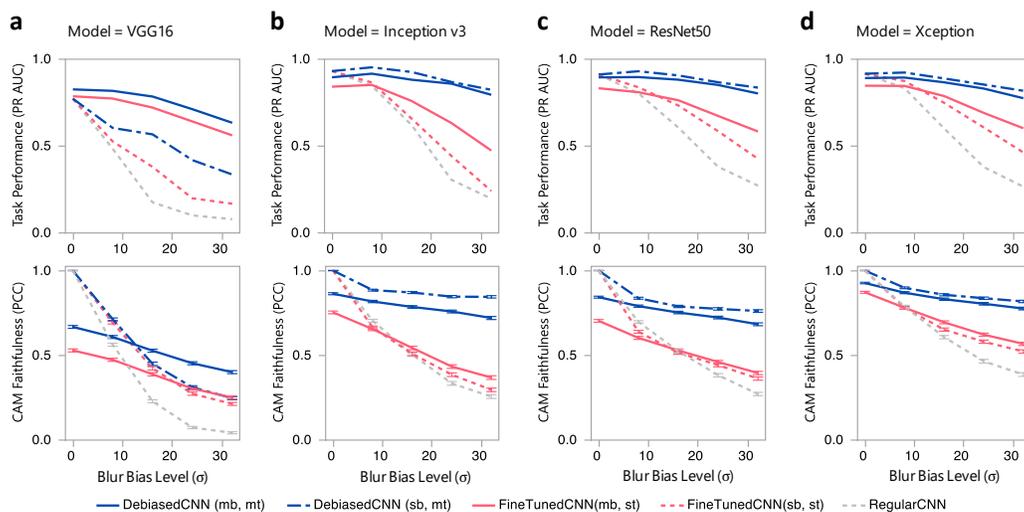
B.2 Supplemental Results



Supplementary Fig. 4. Regression performance for DebiasedCNN (mb, mt) measured as R^2 for the bias level prediction task for five simulation studies. Very high R^2 values indicate that models trained for Simulation Studies 1-3 and 5 could predict the respective bias levels well. Color temperature bias level prediction depended on whether bias was towards lower (more orange) or higher (more blue) temperatures. Since blue-biased images were less distinguishable, the model was less well-trained to predict the blue color temperature bias level; it was more able to predict orange bias at a reasonable accuracy.



Supplementary Fig. 5. Deviated and debiased CAM explanations from various CNN models at varying bias levels of blur biased image from NTCIR-12 labeled as “Biking”. a) VGG16, b) ResNet50, c) Xception. a)-c), Models arranged in increasing CAM Faithfulness (see Supplementary Fig. 6, second row). CAMs from more performant models were more representative of the image label with higher CAM Faithfulness (PCC).



Supplementary Fig. 6. Comparison of model Task Performance and CAM Faithfulness for image classification on NTCIR-12 trained with different CNN models. a) VGG16, b) Inception v3, c) ResNet50, d) Xception. a)-d) Results agreed with Fig. 4 that higher bias led to lower Task Performance and CAM Faithfulness, but debiasing improved both. CNN models are arranged in increasing CAM Faithfulness from left to right. All models were pretrained on ImageNet and fine-tune on NTCIR-12. We set the last two layers of VGG16, and last block of ResNet50 and Xception as retrainable. b)-d) Newer base CNN models than VGG16 significantly outperformed it for both Task Performance and CAM Faithfulness. These newer models had similar Task Performance across bias levels, though their CAM Faithfulness differed more notably.

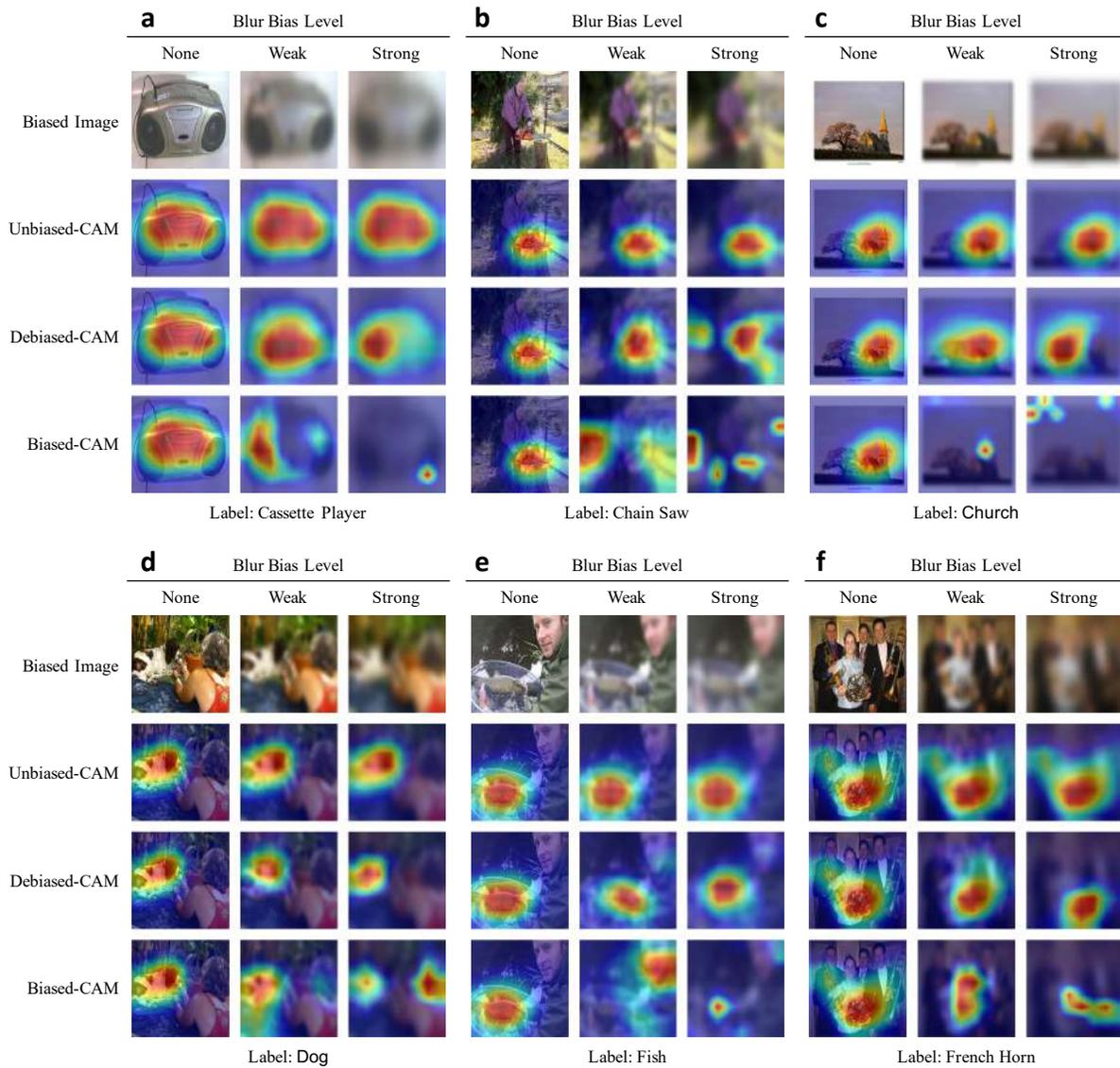
C USER STUDIES APPENDIX

C.1 User Studies Image Selection and CAMs

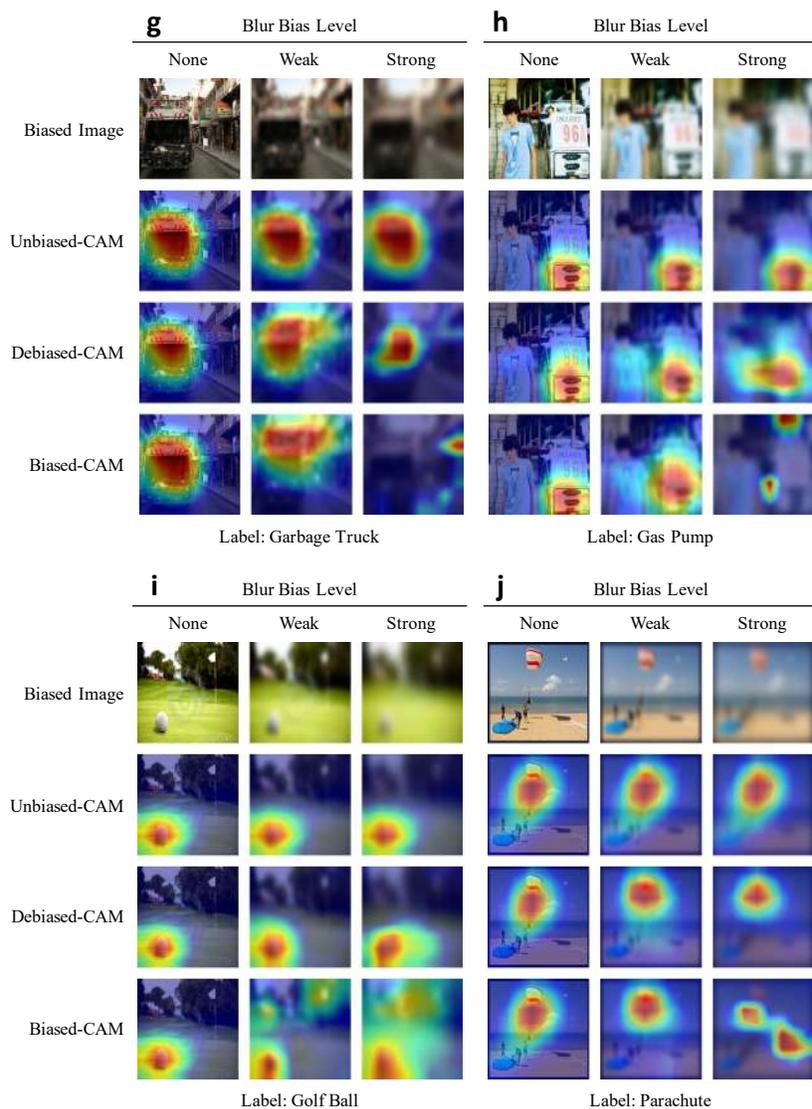
For both user studies, we chose 10 images to select one instance per class label for 10 classes of ImageNet. This balanced between selecting a variety of images for better external validity, and too much workload for participants due to too many trials. CAMs were generated from specific CNN models in Simulation Study 1. At each blur level, Unbiased-CAM and Biased-CAM were generated from RegularCNN, while Debiased-CAM was generated from DebiasedCNN (mb, mt). A key objective of the user studies was to validate the results of the simulation studies regarding CAM types and image blur bias levels, hence, we selected canonical images that:

- (1) Had RegularCNN and DebiasedCNN predict correct labels for unblurred images, since we were not investigating the use of CAMs to debug model errors. CNN predictions on blurred images may be wrong, but we showed the CAM of the correct label.
- (2) Were easy to recognize when unblurred, so that users can perceive whether a CAM is representative of a recognizable image. This was validated in our pilot study.
- (3) Were somewhat difficult but not impossible to recognize with Weak blur, so that participants can feasibly verify image labels with some help from CAMs.
- (4) Were very difficult to recognize with Strong blur, such that about half of pilot participants were unable to recognize the scene, to investigate the upper limits of CAM helpfulness.
- (5) Had Unbiased-CAMs that were representative of their labels, to evaluate perceptions with respect to truthful CAMs. Conversely, debiasing towards untruthful CAMs is futile.
- (6) Had Biased-CAMs for Strong blur that were perceptibly deviated and localized irrelevant objects or pixels; otherwise, no difference between Unbiased-CAM and Biased-CAM will lead to no perceived difference between Unbiased-CAM and Debiased-CAM too.
- (7) Had Debiased-CAMs that were an approximate interpolation between the Unbiased-CAM and Biased-CAM of each image, to represent the intermediate CAM Faithfulness of Debiased-CAM found in the simulation studies.

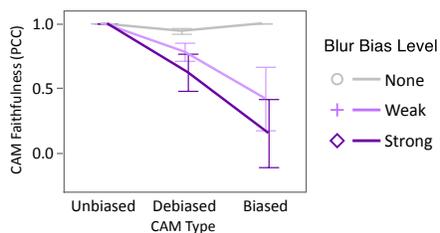
These criteria were verified with participants in a pilot study and the selected images had CAM Faithfulness representative of Simulation Study 1 for Debiased-CAM, but with slightly lower CAM Faithfulness for Biased-CAM to represent worse case scenarios. CAMs were different based on CAM type and Blur Bias level. Unbiased-CAMs were the same for all Blur Bias levels, and Unbiased-CAM and Biased-CAM were the same for None blur level. For other conditions, CAMs were deviated and debiased based on CAM type and Blur Bias level. We chose not to test participants with images in NTCIR-12 due to quality and recognizability issues. Since images were automatically captured at regular time intervals, many images were transitional (e.g., pointing at ceiling while “Watching TV”), which made them unrepresentative of the label. Furthermore, in pilot testing, participants had great difficulty recognizing some scenes (e.g., “Cleaning and Chores”) in images with Strong blur, such that the tasks became too confusing to test. Nevertheless, our results can generalize to wearable camera images with Weak blur, for users who are familiar with or can remember their personal recent or likely activities.



Supplementary Fig. 7. Images and CAMs at various Blur Bias levels and CAM types that participants viewed in both User Studies.



Supplementary Fig. 8. (Continued) Images and CAMs at various Blur Bias levels and CAM types that participants viewed in both User Studies.



Supplementary Fig. 9. CAM Faithfulness of selected 10 image instances used in user studies. Faithfulness decreased as Blur Bias increased, was the highest for Unbiased-CAM, the lowest for Biased-CAM, and improved by Debiased-CAM. Error bars indicate 90% confidence interval.

C.2 User Study 1 and 2 Questionnaires

We illustrate key sections in the questionnaire for the CAM Truthfulness User Study 1 and CAM Helpfulness User Study 2. Both questionnaires were identical except for the main study section.

Training

Consider a Smart Camera that captures images and automatically labels them. It also has smart filters for various features.

Here is an example scene that the Smart Camera can photograph.



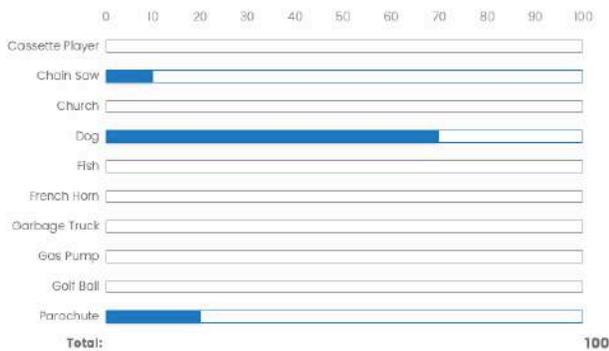
Blur Filter

The Smart Camera applies a **Blur Filter** on the captured image to protect sensitive information. Here you can see the original unblurred image on the left and blurred one on the right. Note that this privacy filter may make it harder to understand what the camera is photographing.



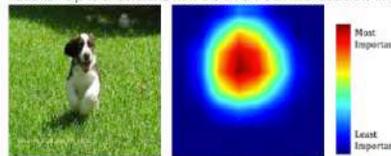
We will ask you to identify what the image is about from a list of possible answers. We use a set of sliders to indicate what the image is likely to be. Suppose you think that it is 70% likely a Dog, 20% likely a Parachute, and 10% likely a Chain Saw, you will then indicate the sliders as shown below.

Note that the % likelihoods need to sum to 100.



Heatmap Filter

The Smart Camera applies a Heatmap Filter to help viewers understand how it automatically labels the image. It indicates which part of the image is important to understand the activity. Red areas are more important and blue areas are less important. In the following example, the heatmap highlights the dog's head to help to understand how the Smart Camera labels the image as Dog.



Note that sometimes the highlighted regions of heatmap may be inaccurate or misleading.



Supplementary Fig. 10. Tutorial to introduce the scenario background of a smart camera with privacy blur and heatmap (CAM) explanation. It taught the participant to i) interpret the “balls and bins” question ([32]), ii) understand why images were blurred, and iii) interpret the CAM.

Warm-up: Screening questions

You need to answer the following questions correctly to qualify for the Main Survey.

What is this image about?



- Church
- Dog
- Fish
- Garbage Truck
- Parachute

What is this image about?



- Church
- Dog
- Fish
- Garbage Truck
- Parachute

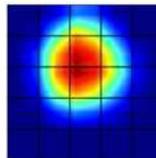
Which of the following grid cells contain a **french horn**?

Hint: click on the box region. You need to select as least one box but maybe more than one.



Here is a different image. According to *this heatmap filter*, which grid cells are most important?

Hint: click on the box region. You need to select as least one box but maybe more than one.



Supplementary Fig. 11. Screening quiz with four questions to test labeling correctness and saliency selection. Questions tested for correct labeling on an unblurred (1) and a weakly blurred (2) photograph image, and correct grid selection of relevant locations in a photograph image (3) and a heatmap (4). The participant is excluded from the study if he answered more than one question wrongly.

Congratulations! On to main survey

Congratulations! You have correctly completed the tutorial questions.

Please tell us a little about yourself.

Do you agree or disagree with the following statements?

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I consider myself as a technology-savvy person.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have no problem understanding photographs.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Next, you will see a series 10 of images and answer several questions for each image.

Important: Please answer questions carefully. We may have to reject your work if you fail any attention test, or give inconsistent, repeatedly identical or seemingly random answers.

[→](#)

Supplementary Fig. 12. **Background questions on participant self-reported technology savviness and photograph comprehension. These questions were posed after passing the screening quiz, and before the main study section to measure the participant’s pre-conceived self-assessment which may be biased after repeatedly viewing variously blurred images and variously biased heatmaps.**

a

Image 4

This image shows a **Dog**.
Which part(s) of the image do **you** think is most important to identify the image content.
Hint: check the box(es) corresponding to each region that you want to select. You may select more than one box.



b

Image 4

Given this scene ...

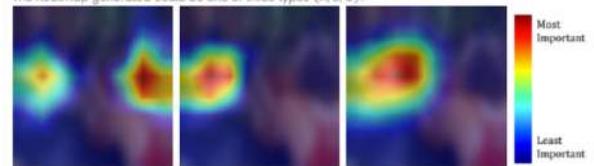


The Smart Camera has captured this image ...



Labeled it as **Dog** and

Generated a heatmap to indicate which part of the image was important for this automatic labeling.
The heatmap generated could be one of three types (A, B, C):



Please rate how representative each heatmap is. 1=not (least) representative, 10=most representative.

Heatmap A ☆☆☆☆☆☆☆☆☆☆

Heatmap B ☆☆☆☆☆☆☆☆☆☆

Heatmap C ☆☆☆☆☆☆☆☆☆☆

For this image, explain what you think makes a heatmap more representative or less representative.



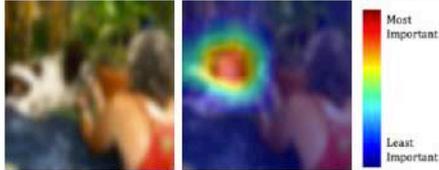
Supplementary Fig. 13. Example main study per-Image Trial for CAM Truthfulness User Study 1. a) The first page asked the participant to q1) select on a grid which locations in an unblurred image are important to identify the image as labeled. b) The second page showed how the smart camera has captured the image (at a randomly selected Blur Bias level), and asked the participant to q2) rate the Truthfulness of all three CAM types (randomly arranged) along a 10-point scale and to q3) explain her rating rationale.

a

Image 4

The Smart Camera has captured this image, labeled the image as **Dog** and generated a heatmap to indicate which part of the image is important for its automatic labeling.

Note: the Smart Camera's predicted label may be wrong and the heatmap may be inaccurate.



What will you label for this image? What do you think this image is about?
You may disagree with the Smart Camera.

Note: at least one choice must be more than 0%, and all % likelihoods need to sum to 100.

0 10 20 30 40 50 60 70 80 90 100

Cassette Player

Chain Saw

Church

Dog

Fish

French Horn

Garbage Truck

Gas Pump

Golf Ball

Parachute

Others

Total: 0

Please rate how **representative** the heatmap is for labeling the image. 1=not (least) representative, 10=most representative.

Your Rating: ★★★★★★★★

Regardless of whether you thought the heatmap is representative of the label, do you agree or disagree that the **heatmap** was helpful for you to label the image?

Strongly agree

Agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Disagree

Strongly disagree

Explain why you think the heatmap is helpful or not helpful to identify the image content.

b

Image 4

Here is the original scene that the Smart Camera saw.



Recall that the Smart Camera has captured this image, labeled the image as **Dog** and generated a heatmap to indicate which part of the image is important for its automatic labeling.

Note: the Smart Camera's predicted label may be wrong and the heatmap may be inaccurate.



Please rate how **representative** the heatmap is for labeling the image. 1=not (least) representative, 10=most representative.

Your Rating: ★★★★★★★★

Regardless of whether you thought the heatmap is representative of the label, do you agree or disagree that the heatmap was **helpful** for you to label the image?

Strongly agree

Agree

Somewhat agree

Neither agree nor disagree

Somewhat disagree

Disagree

Strongly disagree

Explain why you think the heatmap is helpful or not helpful to identify the image content.

Supplementary Fig. 14. Example main study per-Image Trial for CAM Helpfulness User Study 2. a) The first page showed the smart camera's captured blur biased image, generated heatmap (CAM) explanation, and predicted label; and asked the participant to q1) indicate the label likelihood with a "balls and bins" question; q2) rate the CAM Truthfulness, q3) rate the CAM Helpfulness and q4) explain his rating rationale. b) The second page showed the image unblurred, redisplayed the blurred image and CAM and repeated the questions for q5) CAM Truthfulness rating, q6) CAM Helpfulness rating and q7) rating rationale; the repeated questions allow the comparison of ratings before (preconceived) and after (consequent) the participant knew about the ground truth image.

C.3 Statistical Analyses

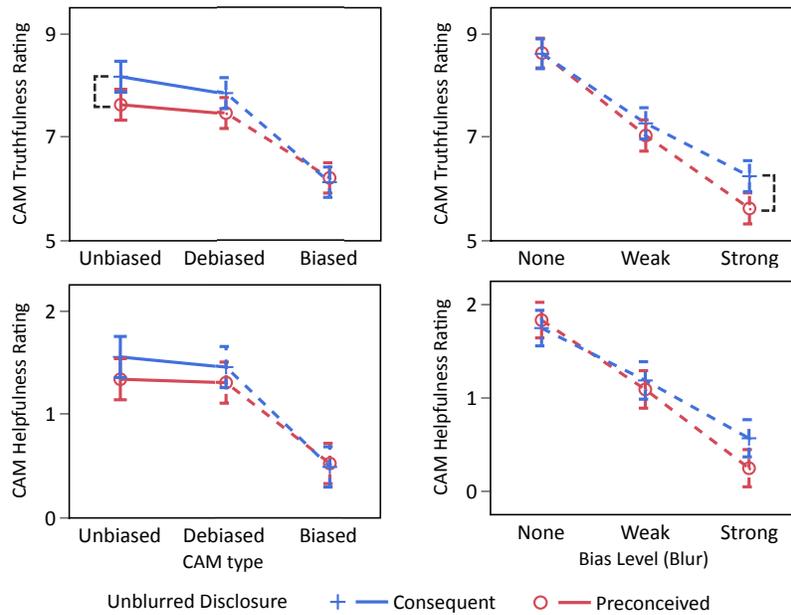
#	Response	Linear Effects Model (Participant as random effect)	F	p>F	R ²
q1	CAM Truthfulness Selection (PCC)	Blur Bias Level +	169.3	<.0001	.671
		CAM Type +	255.7	<.0001	
		Blur Bias Level × CAM Type +	69.4	<.0001	
		Image Label +	20.6	<.0001	
		Image Label × Blur Bias Level +	13.9	<.0001	
		Image Label × CAM Type +	8.9	<.0001	
		Task Time Level +	1.7	<i>n.s.</i>	
		Task Time Level × Blur Bias Level +	1.7	<i>n.s.</i>	
		Task Time Level × CAM Type	0.1	<i>n.s.</i>	
q2	CAM Truthfulness Rating	Blur Bias Level +	134.9	<.0001	.612
		CAM Type +	195.6	<.0001	
		Blur Bias Level × CAM Type +	61.5	<.0001	
		Image Label +	7.3	<.0001	
		Image Label × Blur Bias Level +	4.0	<.0001	
		Image Label × CAM Type +	5.8	<.0001	
		Task Time Level +	1.1	<i>n.s.</i>	
		Task Time Level × Blur Bias Level +	1.7	<i>n.s.</i>	
		Task Time Level × CAM Type	6.1	<.0001	

Supplementary Table 3. Statistical analysis of responses due to effects as linear mixed effects models for CAM Truthfulness User Study 1. All models had various fixed main and interaction effects (shown as one effect per row) and Participant as a random effect. Rows with grey text indicate non-significant effects. Numbers (blue) correspond to survey questions in each image trial in Fig. 5a.

#	Response	Linear Effects Model (Participant as random effect)	F	p>F	R ²
q1a	Labeling Correctness	Blur Bias Level +	42.4	<.0001	.346
		CAM Type +	3.2	.0428	
		Bias Level × CAM Type +	1.7	<i>n.s.</i>	
		Image Label +	4.6	<.0001	
		Image Label × Blur Bias Level +	2.3	.0013	
		Image Label × CAM Type +	0.9	<i>n.s.</i>	
		Task Time Level +	0.2	<i>n.s.</i>	
		Task Time Level × Blur Bias Level +	1.2	<i>n.s.</i>	
		Task Time Level × CAM Type	1.6	<i>n.s.</i>	
q1b	Labeling Confidence	Blur Bias Level +	118.6	<.0001	.522
		CAM Type +	3.0	.0484	
		Bias Level × CAM Type +	2.1	<i>n.s.</i>	
		Image Label +	9.9	<.0001	
		Image Label × Blur Bias Level +	3.9	<.0001	
		Image Label × CAM Type +	0.8	<i>n.s.</i>	
		Task Time Level +	3.3	.0387	
		Task Time Level × Blur Bias Level +	0.9	<i>n.s.</i>	
		Task Time Level × CAM Type	1.1	<i>n.s.</i>	
q2,3	CAM Truthfulness Rating	Blur Bias Level +	146.4	<.0001	.575
		CAM Type +	95.0	<.0001	
		Blur Bias Level × CAM Type +	20.9	<.0001	
		Image Label +	6.9	<.0001	
		Image Label × Blur Bias Level +	7.9	<.0001	
		Image Label × CAM Type +	2.6	.0003	
		Task Time Level +	0.6	<i>n.s.</i>	
		Task Time Level × Blur Bias Level +	3.6	.0067	
		Task Time Level × CAM Type +	0.5	<i>n.s.</i>	
		Unblurred Disclosure +	10.0	.0016	
		Unblurred Disclosure × Blur Bias Level +	3.7	.0249	
Unblurred Disclosure × CAM Type +	3.9	.0201			
		Unblurred Disclosure × Blur Bias Level × CAM Type	2.9	.0210	
q5,6	CAM Helpfulness Rating	Blur Bias Level +	91.6	<.0001	.517
		CAM Type +	71.8	<.0001	
		Blur Bias Level × CAM Type +	20.9	<.0001	
		Image Label +	5.3	<.0001	
		Image Label × Blur Bias Level +	5.9	<.0001	
		Image Label × CAM Type +	2.7	.0001	
		Task Time Level +	1.0	<i>n.s.</i>	
		Task Time Level × Blur Bias Level +	3.7	.0052	
		Task Time Level × CAM Type +	0.1	<i>n.s.</i>	
		Unblurred Disclosure +	2.2	<i>n.s.</i>	
		Unblurred Disclosure × Blur Bias Level +	4.1	.0169	
Unblurred Disclosure × CAM Type +	2.0	<i>n.s.</i>			
		Unblurred Disclosure × Blur Bias Level × CAM Type	2.0	<i>n.s.</i>	

Supplementary Table 4. Statistical analysis of responses due to effects as linear mixed effects models for CAM Helpfulness User Study 2. All models had various fixed main and interaction effects (shown as one effect per row) and Participant as a random effect. Rows with grey text indicate non-significant effects. Numbers (blue) correspond to survey questions in each image trial in Fig. 7a.

C.4 Supplementary Results



Supplementary Fig. 15. Comparisons of perceived CAM Truthfulness and CAM Helpfulness before (preconceived) and after (consequent) disclosing the unblurred image. There was a significant difference across Unblurred Disclosure for CAM Truthfulness Rating ($p = .0013$) but not for CAM Helpfulness Rating. Comparing preconceptual to consequent ratings, Unbiased-CAMs were rated as less truthful ($M = 7.7$ vs. 8.3 , $p = .0004$), Debiased-CAMs were rated marginally less truthful ($p = .0212$), Biased-CAMs were rated similarly untruthful, and overall, CAMs of Strongly blurred images were rated as less truthful ($M = 5.6$ vs. 6.3 , $p < .0001$). These results suggest that even with the least biased CAM (Unbiased-CAM), the unfamiliarity of unblurred scenes can hurt trust (truthfulness) in the CAM, though there was no change in perceived helpfulness before or after disclosing the unblurred image. CAM Truthfulness Ratings were measured along a 1-10 scale, and CAM Helpfulness Ratings along a 7-point Likert scale ($-3 =$ Strongly Disagree, $0 =$ Neither, $+3 =$ Strongly Agree). Error bars indicate 90% confidence interval. Dotted lines indicate extremely significant $p < .0001$ comparisons, and solid lines indicate no significance at $p > .01$.