Explaining models relating objects and privacy

Alessio Xompero[1], Myriam Bontonou[2], Jean-Michel Arbona[2], Emmanouil Benetos[1], Andrea Cavallaro[1,3,4]
[1]Queen Mary University of London, United Kingdom, [2]ENS de Lyon and CNRS, France,
[3]Idiap Research Institute, [4]École polytechnique fédérale de Lausanne, Switzerland
{a.xompero,emmanouil.benetos}@qmul.ac.uk, {myriam.bontonou,jeanmichel.arbona}@ens-lyon.fr,
a.cavallaro@idiap.ch

Abstract

Accurately predicting whether an image is private before sharing it online is difficult due to the vast variety of content and the subjective nature of privacy itself. In this paper, we evaluate privacy models that use objects extracted from an image to determine why the image is predicted as private. To explain the decision of these models, we use feature-attribution to identify and quantify which objects (and which of their features) are more relevant to privacy classification with respect to a reference input (i.e., no objects localised in an image) predicted as public. We show that the presence of the person category and its cardinality is the main factor for the privacy decision. Therefore, these models mostly fail to identify private images depicting documents with sensitive data, vehicle ownership, and internet activity, or public images with people (e.g., an outdoor concert or people walking in a public space next to a famous landmark). As baselines for future benchmarks, we also devise two strategies that are based on the person presence and cardinality and achieve comparable classification performance of the privacy models.

1. Introduction

People take photos in a large variety of situations (e.g., at a party, of themselves, of a landmark, or of friends, family, animals, or food) and share them on social media platforms, often lacking awareness of privacy risks associated with their sharing [2, 6]. Images may contain a set of objects that reveal private information about a person or be associated with a specific location or event that the person is attending. Therefore, an automatic warning prior to sharing could help users protect their privacy [9, 15, 23].

Privacy classification methods are trained on datasets annotated by one or multiple annotators with a binary label (public or private) [21, 23, 24]. As the notion of privacy varies among people and also depends on the context, the annotation in these datasets is potentially ambiguous. Most of the existing works design methods that aimed at improving the classification performance on these datasets. We categorise existing methods for image privacy as single-stage and two-stage. *Single-stage* methods directly train or fine-tune a deep neural network (DNN) from the images [9]. *Two-stage* methods uses DNNs (e.g., convolutional neural networks or CNNs) to extract concepts (i.e., objects, scenes) from the images followed by a privacy classifier in the second-stage, such as a Multi-Layer Perceptron (MLP) or a graph neural network (GNN) [3, 15, 17, 18, 21]. Two-stage methods can be further split into end-to-end training or hybrid. *End-to-end training* based methods fine-tune the DNNs to initialise the concept features for the privacy classifier [15, 21]. *Hybrid* methods extract concepts from the images with a pre-trained detector or multi-label image classifier [3, 17, 18].

In this paper, we explain the decisions made by a range of privacy classifiers that use as input the cardinality and confidence features of objects identified in an image (see Fig. 1). Among many existing explainability methods [10, 12, 13, 16, 22], we select integrated gradients [16] that is computationally efficient and attributes the decision of the privacy models to the identified objects and their features with respect to a reference input. This reference input consists of features with zero values to represent the case of no objects localised in an image and hence classified as public. Based on the findings from the explainability analysis, we define two simple strategies using people presence as main driving factor to determine whether an image is private. These explainable-by-design strategies achieve comparable performance to the more complex privacy-decision models[1]. As baselines in future comparisons, these strategies will also enable the design of explainable and more

Alessio Xompero and Myriam Bontonou equally contributed. Myriam Bontonou is also affiliated with Inserm, France, and Jean-Michel Arbona with Univ Lyon and LBMC Lyon, France.

[1]Code: https://github.com/graphnex/ig-privacy

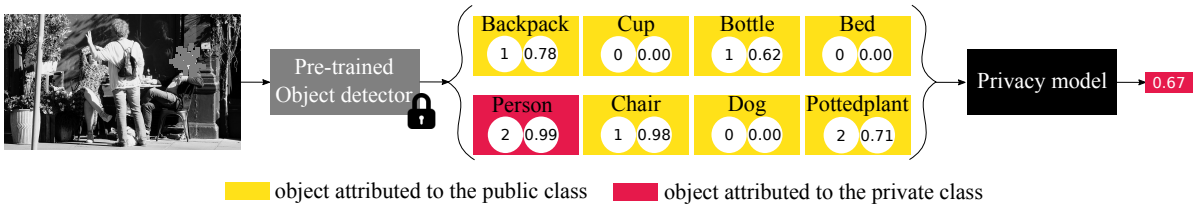

Figure 1. Two-stage privacy method: a pre-trained object detector identifies concepts (e.g., objects, scene type) within an image and a privacy model is trained to classify an image as private or public, considering the cardinality and confidence level of the extracted objects (numbers below each object). The input image is from the PrivacyAlert dataset [24], with obfuscation added on the face of the person.

accurate privacy models that capture and use relationships between concepts beyond the only presence and cardinality of people in images.

2. Problem formulation

Let I be an image and $f_\theta(\cdot)$ a privacy model trained on a dataset $\mathcal{D} = \{(I, y)_n\}_{n=1}^{N}$ to predict a class $y \in \{0, 1\}$, where 0 denotes public and 1 private, θ contains the model parameters, and N is the number of images in the training dataset. We consider the privacy model to map the outputs of other models to the predicted class y: $y = f_\theta(d_\eta(I))$, where η contains pre-trained parameters. For example, $d_\eta(\cdot)$ can be a pre-trained object detector that localises a set of objects with their confidence in the image I. We refer to the pre-defined categories outputted by these pre-trained models as concepts. Therefore, let $\mathcal{X} = \{\mathbf{x}^c | c = 0, \ldots, C-1\}$ be the set of C concepts with their F-dimensional feature vectors $\mathbf{x}^c = \left[x_0^c, \ldots, x_{F-1}^c\right]$ that are provided as input to the privacy model.

Our objective is to explain why the trained model $f_\theta(\cdot)$ predicts the label $y = 1$ for a given image I (observable explanation [1]). Specifically, we want to determine which concepts contribute to the prediction of the private label for the input image. To this end, we use post-hoc explainability to assign a score to each feature of each concept, $\phi(x_j^c) \in \{-1, 1\}$.

Following previous works [15, 17, 21], we consider objects as concepts and we use a pre-trained object detector to localise a pre-defined set of objects [11] (i.e., $C = 80$ for the COCO dataset [8]). We define two features ($F = 2$) for each object: cardinality, $x_0^c \in \mathbb{N}$ and confidence, $x_1^c \in [0, 1]$. For cardinality, we count the number of instances localised in an image and belonging to each object. If no instances are localised for an image, then the cardinality is set to 0. For confidence, we retain the value of the most confident object instance if multiple instances of the same object are localised in an image. The privacy model could be an MLP or a GNN [3, 15, 17, 18, 21]. For MLP, the input is the concatenation of all object features, resulting in a vector of dimensionality CF: $\mathbf{x} = [\ldots, \mathbf{x}^c, \ldots], \forall \mathbf{x}^c \in \mathcal{X}$. For GNN, the input is a $C \times F$ matrix of the object features, where each row corresponds to a node of a graph. For simplicity, we use the set $\mathcal{X}$ as input of the privacy model, independently of the representation: $f_\theta(\mathcal{X})$.

3. Explaining image privacy predictions

In this section, we describe the dataset and models used for image privacy (see the Supplementary Material document for additional details), and discuss their classification performance. We explain the models decision and analyse the explainability results.

Dataset. We use PrivacyAlert [24] as a recent image privacy dataset $\mathcal{D}$ for our evaluations and analyses. PrivacyAlert has 6,800 images[2] split into a training set of 3,136 images (788 private images and 2,348 public images), a validation set of 1,864 images (466 private images and 1,398 public images), and a testing set of 1,800 images (450 private images and 1,350 public images), as originally described by the authors [24]. The dataset has a high class imbalance towards the public images (ratio of about 3:1).

Methods. We consider MLP [3, 17], two graph-based models, GIP [21] and GPA [15], and a graph-agnostic model (GA-MLP) [5]. The MLP aims at reproducing Tonge et al.'s method [17] that uses Support Vector Machine as a privacy classifier and, as input, a binary feature vector of the top-k most confident classes recognised by a pre-trained CNN for multi-label object recognition. In our case, we replace the multi-label classifier with the object detector and the object presence with the cardinality and confidence features of the identified objects. The MLP consists of 3 hidden layers, each with a 16-dimensionality hidden status and followed by batch normalisation. GIP and GPA modelled graphs to relate the objects with two privacy classes or the objects with each other, respectively, using a Graph Reasoning Model [19] as GNN. These two models belong to the two-stage end-to-end training category. They fine-tune the CNNs in the first stage to initialise the node features and the CNN thus contribute to the privacy decision of the models. For a fair comparison, we adapt GIP and GPA to the two-stage hybrid approach by decoupling the GNN from the CNNs. We use only the GNN with the graph

[2]PrivacyAlert provides links to images on Flickr whose license was falling under Public Domain [24]. Note that 7 images are no longer available and we re-train and evaluate models excluding these images.

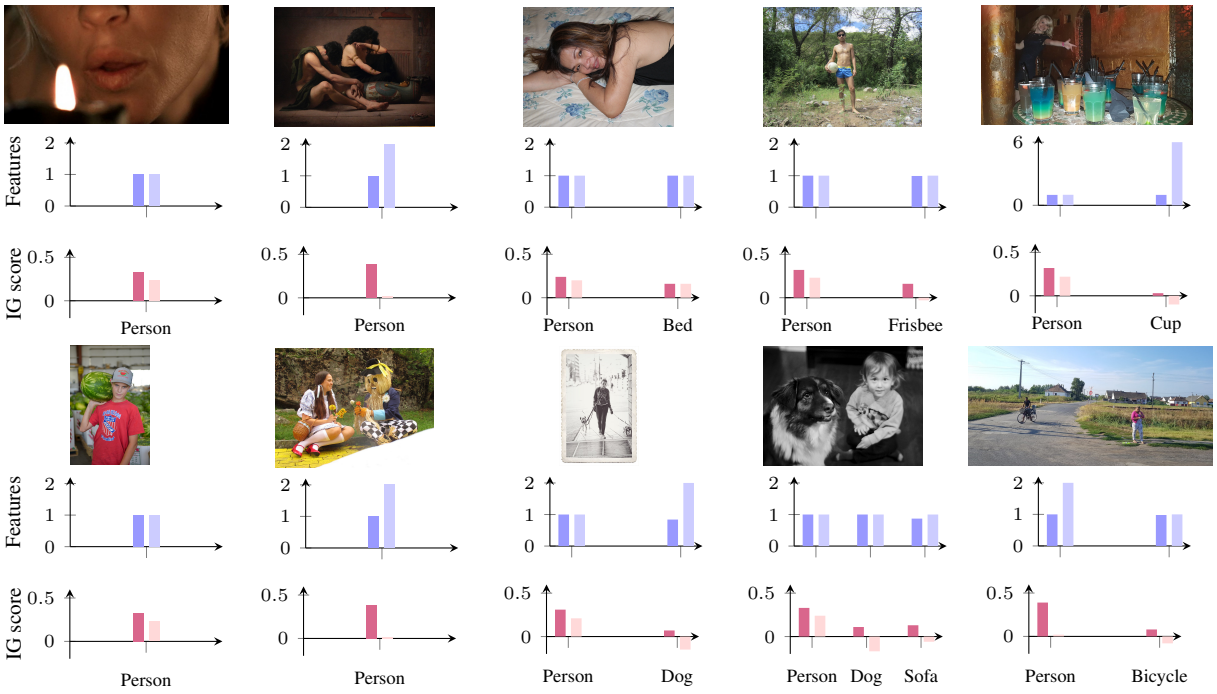

Figure 2. Sample of training images from PrivacyAlert [24] correctly predicted as private (first row) and incorrectly predicted as private (fourth row) by the graph-agnostic baseline [5], with their extracted object and features (blue bar plots) and the explanation scores (red bar plots) of Integrated Gradients (IG) [16]. Darker colours (left bar) are associated with the confidence feature and lighter colours (right bar) with cardinality. Positive IG scores support privacy, whereas negative IG scores support the public decision. Note the different maximum limit for the y-axis in the top-right blue bar plot (fifth column, second row).

modelled by each method and with the cardinality and confidence as input node features. This adaptation allows us to assess the impact of the GNN as privacy model. GA-MLP aims to replicate the steps of a GNN but without the graph structure (graph-agnostic) [5]. To enable the training of the model, we independently project each node feature to a higher dimensionality vector with a fully connected layer shared among the nodes, and we then concatenate the projected features. Similarly to the multiple layers of a GNN, we refine the projected node features using three blocks, each consisting of a fully connected layer, a batch normalization layer [7], a ReLU activation function, and a dropout layer [14]. We aggregate the refined features using global sum pooling and we provide the resulting global feature vector as input to an MLP-based classifier.

Classification. Table 1 compares the classification performance of the privacy models on PrivacyAlert. Given the class imbalance of the dataset, we discuss the results in terms of recall on the private class and balanced accuracy (average recall of the two classes), reported as percentages. Both MLP and GA-MLP achieve a balanced accuracy of 71.60% and 74.30%, respectively, and an overall precision of 69.90% and 70.20%. GA-MLP correctly identifies more private images than MLP (higher recall in the private class). GPA and GIP models degenerate to predict (almost) all images as public, showing that the decoupled graph component based only on object features is not useful for privacy. This suggests that the models trained in the corresponding papers [15, 21] were driven by the fine-tuning of CNNs.

Explainability analysis. To explain the privacy models, we analyse the importance of the extracted concepts and their features to the decision of the models by using Integrated Gradients (IG) [16]. IG is a post-hoc explainability method that is widely used to attribute the prediction of a model to the input features, resulting in a IG score per object feature $\phi(x_j^c)$. As IG is model-agnostic, the method can be applied to all gradient-based models. Specifically, IG compares the privacy prediction of a model f_θ for the input set of objects features $\mathcal{X}$ with the privacy prediction of the same model for a reference input $\mathcal{R} = \{\mathbf{r}^c | c = 0, \ldots, C-1\}$. As private is the target class, we select a null-vector as our reference for each concept: $\mathbf{r}^c = \mathbf{0}, \forall c$ so that $f_\theta(\mathcal{R}) = 0$. This is equivalent to no objects detected in an image. IG also satisfies the *completeness* axiom [16],

$$f_\theta(\mathcal{X}) - f_\theta(\mathcal{R}) = \sum_{c=0}^{C-1} \sum_{j=0}^{F-1} \phi(x_j^c)\,, \tag{1}$$

that quantifies the contribution of the features of all objects towards the decision of the model. Fig. 2 shows the ex-

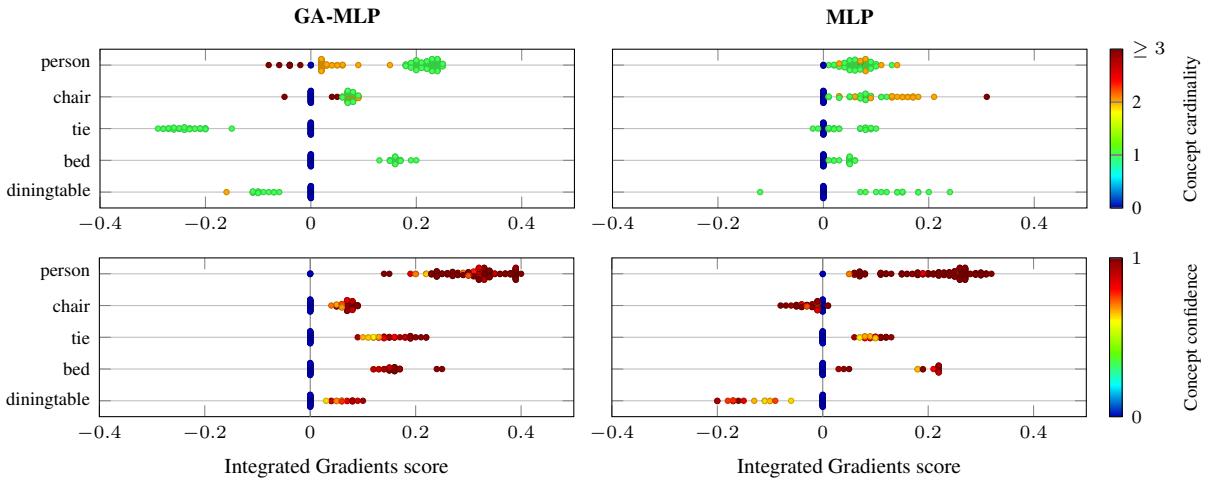

Figure 3. Comparison of the explainability scores across training images correctly classified as private by the graph-agnostic (GA-MLP) and MLP models on the training set of PrivacyAlert [24]. We show only the top 5 objects based on the largest mean absolute explainability scores. Note that colours of the data points represent the value of the object feature. Also note the different limits of the colour bars.

plainabilty scores for a sample of images predicted as private by GA-MLP. Images are selected from the training set of PrivacyAlert based on the objects and their cardinality identified in the images, contrasting correct and incorrect predictions. When images are correctly predicted as private (top row), high confidence in detecting a person significantly influences the decision of the model. On the contrary, the localisation of multiple individuals in an image tends to favour the public class. Public images are often misclassified due to the detection of *person* (second row of images). Fig. 3 compares the predictions and explainability of GA-MLP and MLP across all images correctly identified as private in the training set[3]. As for the previous analysis, *person* is the most relevant concept for private predictions. Unlike MLP, GA-MLP favours the public class when three or more people are detected.

4. Person-centric classification

Based on the outcomes of the previous analysis, we devise two person-centric decision strategies that act directly on the objects extracted by the vision models and the corresponding features. The first, simple strategy classifies an image as private if at least one person is detected, $x_0^c \geq 1$, where x_0^c is the cardinality feature and the object c corresponds to *person*. The second, simple strategy includes an additional constraint that limits the number of people localised in an image, i.e., $x_0^c > 0 \wedge x_0^c \leq 2$, where $\wedge$ is the logical AND operator.

We report the performance of these two strategies on the testing set of PrivacyAlert in Table 1. The second strategy achieves performance comparable to GA-MLP and outperforms MLP, especially in terms of recall on the private class and balanced accuracy. The first strategy has lower balanced accuracy than the second strategy but achieves a recall of 90.89% in the private class denoting that most of the private images contain people. Nonetheless, this first strategy has many false positives (a precision of 40.43% in the private class), indicating that images with people are not necessarily private. The more restrictive condition of the second strategy better balances the issues of the first strategy, but the recall for private images is limited to 74.67%.

[3]The confidence to predict the reference input as private is 0.2 for MLP and 0.1 for GA-MLP.

Table 1. Classification performance on the testing set of PrivacyAlert [24]. All the models are using the same object detector to extract object features from the images. Note the failure of GPA [15] and GIP [21] adapted to the hybrid approach and using only GNN (recall of 100% for public class, and precision and recall of 0% for private class). Their original performance was driven by the dependence on CNNs [15, 21].

Method	Public		Private		Overall	
	P	R	P	R	P	BA
All private	0.00	0.00	25.06	100.00	12.53	50.00
All public	74.94	100	0.00	0.00	37.47	50.00
MLP	86.29	82.32	53.52	60.89	69.90	71.60
GPA*	75.30	97.62	37.25	4.22	56.28	50.92
GPA$^\diamond$	74.94	100	0.00	0.00	37.47	50.00
GIP$^\triangle$	74.94	100	0.00	0.00	37.47	50.00
GA-MLP	88.87	77.71	51.53	70.89	70.20	74.30
Strategy-1	94.76	55.05	40.34	90.89	67.55	72.97
Strategy-2	89.67	73.55	48.55	74.67	69.11	74.11

KEY – P: precision; R: recall, BA: Balanced accuracy; MLP: multi-layer perceptron; GA: graph-agnostic baseline; GPA: Graph Privacy Advisor [15]; $\star$: GPA adapted to the hybrid approach; $\diamond$: adapted GPA with corrected implementation of adjacency matrix; $^\triangle$: GIP [21] adapted to the hybrid approach, using cardinality and confidence as object features and privacy nodes with zero-initialised features.

5. Conclusion

In this paper, we used post-hoc explainability to identify and quantify objects contributing to the decision of image privacy classification models, which are trained on concepts extracted from an image by a pre-trained detector. The explainability analysis showed that privacy models, such as MLP and GA-MLP, are biased towards the presence of the object *person*. Based on this finding, we devised two simple person-centric strategies that achieve comparable overall classification performance to that of the state-of-the-art models consodered in the comparison.

Future work will extend the explainability analysis to other publicly available datasets, such as VISPR [9], IPD [21] and DIPA [20], and other models with different concepts and features [3, 15, 18, 21]. We will also include and compare the results of other explainability methods [4, 10, 12, 13].

Acknowledgements

This work was supported by the CHIST-ERA programme through the project GraphNEx, under UK EPSRC grant EP/V062107/1 and France ANR grant ANR-21-CHR4-0009.

References

[1] G. AlRegib and M. Prabhushankar. Explanatory paradigms in neural networks: Towards relevant and contextual explanations. *IEEE Signal Process. Magazine*, 39(4), 2022. 2

[2] P. Arias-Cabarcos, S. Khalili, and T. Strufe. 'Surprised, Shocked, Worried': User Reactions to Facebook Data Collection from Third Parties. *Proc. Privacy Enhancing Technologies*, 2023. 1

[3] D. Baranouskaya and A. Cavallaro. Human-interpretable and deep features for image privacy classification. In *IEEE Int. Conf. Image Process.*, 2023. 1, 2, 5

[4] B. Bilodeau, N. Jaques, P. W. Koh, and B. Kim. Impossibility theorems for feature attribution. *Proc. National Academy of Sciences*, 121(2), Jan. 2024. 5

[5] V. P.. Dwivedi, C. K. Joshi, A. T. Luu, T. Laurent, Y. Bengio, and X. Bresson. Benchmarking Graph Neural Networks. *J. Mach. Learning Res.*, 24(43), 2023. 2, 3

[6] L. Ferrarello, A. Cavallaro, R. Fiadeiro, and R. Mazzon. Reframing the narrative of privacy through system-thinking design. In *Design Research Society Biennial Conf.*, 2022. 1

[7] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. Int. Conf. Machine Learning*, 2015. 3

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2018. 2

[9] T. Orekondy, B. Schiele, and M. Fritz. Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images. In *Int. Conf. Comput. Vis.*, 2017. 1, 5

[10] V. Petsiuk, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proc. Brit. Mach. Vis. Conf.*, 2018. 1, 5

[11] J. Redmon and A. Farhadi. YOLOv3: An Incremental Improvement. *CoRR*, abs/1804.02767, 2018. 2

[12] M. T. Ribeiro, S. Singh, and C. Guestrin. Model-agnostic interpretability of machine learning. In *Proc. ICML Workshop on Human Interpretability in Machine Learning*, 2016. 1, 5

[13] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016. 1, 5

[14] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learning Res.*, 15, 2014. 3

[15] D. Stoidis and A. Cavallaro. Content-based Graph Privacy Advisor. In *Proc. IEEE Int. Conf. Multimedia Big Data*, 2022. 1, 2, 3, 4, 5

[16] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proc. Int. Conf. Machine Learning*, 2017. 1, 3

[17] A. Tonge and C. Caragea. Image privacy prediction using deep features. In *Proc. AAAI Conf. Artificial Intell.*, 2016. 1, 2

[18] A. Tonge, C. Caragea, and A. Squicciarini. Uncovering scene context for predicting privacy of online shared images. In *Proc. AAAI Conf. Artificial Intell.*, 2018. 1, 2, 5

[19] Z. Wang, T. Chen, J. Ren, W. Yu, H. Cheng, and L. Lin. Deep reasoning with knowledge graph for social relationship understanding. In *Proc. Int. Joint Conf. Artificial Intelligence*, 2018. 2

[20] A. Xu, Z. Zhou, K. Miyazaki, R. Yoshikawa, S. Hosio, and K. Yatani. DIPA: An Image Dataset with Cross-cultural Privacy Concern Annotations. In *Proc. 28th ACM Int. Conf. Intelligent User Interfaces*, 2023. 5

[21] G. Yang, J. Cao, Z. Chen, J. Guo, and J. Li. Graph-based Neural Networks for Explainable Image Privacy Inference. *Pattern Recognit.*, 105, Sept. 2020. 1, 2, 3, 4, 5

[22] H. Yuan, H. Yu, S. Gui, and S. Ji. Explainability in Graph Neural Networks: A Taxonomic Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):5782–5799, 2023. 1

[23] S. Zerr, S. Siersdorfer, J. Hare, and E. Demidova. Privacy-aware image classification and search. In *Proc. ACM SIGIR Int. Conf. Research and Development in Information Retrieval*, 2012. 1

[24] C. Zhao, J. Mangat, S. Koujalgi, A. Squicciarini, and C. Caragea. PrivacyAlert: A Dataset for Image Privacy Prediction. In *Int. AAAI Conf. Web and Social Media*, 2022. 1, 2, 3, 4