Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images

Tribhuvanesh Orekondy Mario Fritz Bernt Schiele

Max Planck Institute for Informatics
Saarland Informatics Campus
Saabrücken, Germany
{orekondy,mfritz,schiele}@mpi-inf.mpg.de

Abstract

Images convey a broad spectrum of personal information. If such images are shared on social media platforms, this personal information is leaked which conflicts with the privacy of depicted persons. Therefore, we aim for automated approaches to redact such private information and thereby protect privacy of the individual.

By conducting a user study we find that obfuscating the image regions related to the private information leads to privacy while retaining utility of the images. Moreover, by varying the size of the regions different privacy-utility trade-offs can be achieved. Our findings argue for a "redaction by segmentation" paradigm.

Hence, we propose the first sizable dataset of private images "in the wild" annotated with pixel and instance level labels across a broad range of privacy classes. We present the first model for automatic redaction of diverse private information. It is effective at achieving various privacy-utility trade-offs within 83% of the performance of redactions based on ground-truth annotation.

1. Introduction

More and more visual data is captured and shared on the Internet. Images and video contain a wide range of private information that may be shared unintentionally such as e.g. email-address, picture-id or finger-print (see Figure 1). Consequently, there is a growing interest within the computer vision community [4, 19, 23, 25, 41, 43] to assess the amount of leaked information, understand implications on privacy and ultimately control and enforce privacy again. Yet, we are missing an understanding how image content relates to private information and how automated redaction can be approached.

Therefore, we address two important questions in this

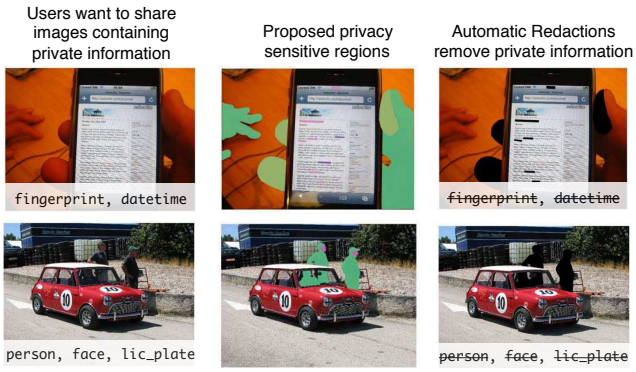

Figure 1: Users often share images containing private information, which poses a privacy risk. For example, in the top row, user might unintentionally leak their fingerprint. We present methods to aid users automatically redact such content by proposing privacy sensitive regions in images.

context. First, how can private information be redacted while maintaining an intelligible image? We investigate this question in a user study with highly encouraging results: we can redact private information in images while preserving its utility. Furthermore, varying the amount of pixels redacted results in different privacy vs. utility trade-offs. We conclude that redaction by segmentation is a valid approach to perform visual redactions.

We ask a second question in this paper: What kind of privacy-utility trade-offs can be achieved by automatic redaction schemes? Based on our first finding, we approach this as a pixel labeling task on multiple privacy classes (which we refer to as *privacy attributes*). Segmenting privacy attributes in images presents a new challenge of reasoning about regions including multiple modalities. For instance, in Figure 1, identifying the `name` and `datetime` requires mapping the relevant pixels to the text domain for understanding, while identifying the `student_id` requires

reasoning over both visual and text domains. Our automated methods address these challenges and localize these privacy attributes for redaction via segmentation. By performing both quantitative and human evaluation, we find these automated methods to be effective in segmentation as well as privacy-utility metrics.

Our model and evaluation for automatic redaction is facilitated by a new dataset that extends the Visual Privacy (VISPR) dataset [41] to include high-quality pixel and instance-level annotations. To this end, we propose a dataset containing 8.5k images annotated with 47.6k instances over 24 privacy attributes. Refer to project website: `https://resources.mpi-inf.mpg.de/d2/orekondy/redactions/`

2. Related Work

Text Sanitation Redaction techniques are primarily studied in the context of confidential text documents, wherein certain sensitive entities need to be removed. Studies focus on identification of such entities [5, 8, 9, 44, 45, 46] and methods to prevent over-sanitation [5, 44]. However, unlike these works which have access to dense structured text data (e.g. documents), we deal with unstructured pixel-level representations of such entities.

Image Perturbations for Privacy Adversarial perturbations [16, 21, 39] are suggested to evade person identification [25, 48]. However, these methods typically assume a white-box CNN-based adversary for the specific task of face recognition. In contrast, we propose redacting content at the expense of some utility to achieve better privacy (measured against humans) across a broad range of privacy classes. Many works [3, 17, 18, 19, 29, 31, 35] propose and analyze redaction strategies (e.g., blurring, cartooning) and study their effects on privacy and utility of the image. However, we focus on automatically localizing *and* redacting private content in images across multiple modalities.

Private Information Recognition Many existing studies focus on either detecting faces [50, 52], license plates [6, 56, 57], relationships [49, 53], age [2] or occupations [47]. Research in determining privacy risk across a broad range of privacy classes are typically treated as a classification problem [41, 51, 54]. However, many studies [1, 11] demonstrate a "privacy paradox" – users share such images in spite of knowing the privacy risks. Hence in this work, we propose a middle ground for reducing privacy leakage, such that users can still share images by redacting private content while preserving its utility.

Visual Privacy Datasets PicAlert [55] and YourAlert [54] propose datasets with user-classified privacy labels. VISPR [41] provides a more exhaustive dataset of 22k images annotated with a broad range of image-level privacy labels. The PEViD video dataset [30] provides person-centric bounding box annotation over 20 video sequences in a constrained setting. In contrast, our dataset based on VISPR images provides pixel level annotation from a diverse set of privacy classes.

Segmentation Identifying pixel-level labels from images is a well-studied problem in computer vision. However, most methods [34, 37] and datasets [10, 13, 36] focus on segmenting common objects in visual scenes. We however focus on identifying private regions in a privacy-utility framework, which introduces many new challenges.

3. The Visual Redactions Dataset

In this section we present our pixel-label visual privacy dataset as an extension to the VISPR dataset [41]. We begin with a discussion on how images (Section 3.1) and attributes (Section 3.2) were selected for the task. This is followed by the annotation procedure (Section 3.3) and a brief analysis (Section 3.4) of the dataset.

3.1. Selecting Images for Pixel-level Annotation

The VISPR dataset contains 22k real-world user-uploaded publicly available Flickr images which makes this a great starting point for addressing the visual redaction problem "in the wild". 10k of these images are annotated as safe. From the remaining 12k images we pixel-annotate the subset of 8,473 images that contain at most 5 people. The main reason to focus on this subset was to reduce the annotation cost while maximizing the amount of non-person pixels. We preserve the identical 45-20-35 train-val-test split of these images as in the VISPR dataset.

3.2. Shortlisting Privacy Attributes

The 22k images in the multilabel VISPR dataset are annotated using 68 image-level privacy attributes ($\sim$5.2 attributes per image). These privacy attributes are compiled from multiple privacy-relevant sources – the US Privacy Act of 1974, EU Data Protection Directive 95/46/EC and various social network website rules. Additionally, they cover a diverse range of private information that can be leaked in images (e.g. face, tattoo, physical disability, personal relationships, passport, occupation). Therefore, we use these as a starting point for redactions in images. We select 42 out of 67 privacy attributes (excluding attribute 'safe', which indicates none of the other 67 attributes are present) for three reasons. First, for 11 attributes (e.g. religion, occupation, sports) typically the entire image is linked to the attribute (e.g. scene with church or sport stadium). In such cases, the solution to keeping the information private is to not share such images (as proposed in [41]). We instead focus on attributes which can be localized for redaction, such that the image might still be useful. Second, 8 attributes were extremely tedious to annotate, because of their strong co-

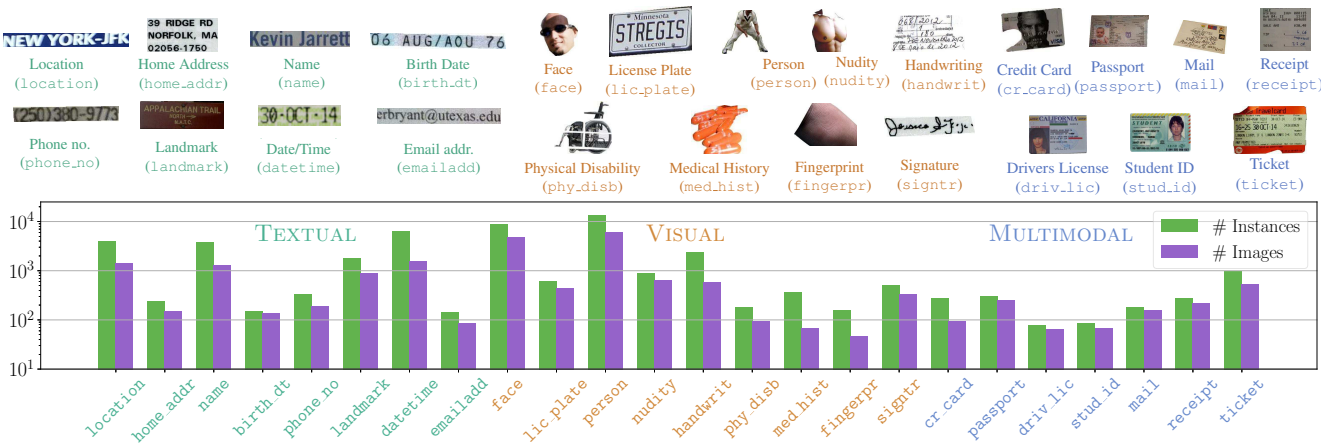

Figure 2: Examples and distribution of privacy attributes in the dataset.

occurrence with crowd-scenes (e.g. political and general opinion, occupation) or the effort required to outline them (e.g. hair color). Third, 6 attributes (e.g. place of birth, email content, national id) contained under 30 examples for training. In spite of filtering such attributes, we still cover a broad spectrum of information to help de-identify people in images (such as by obfuscating faces or names). We further merge few groups among these 42 attributes: (i) when they occur as a complete and partial version (e.g. (complete face, partial face) merged into `face`) (ii) when they localize to the same region (e.g. (race, skin color, gender, relationships) merged into `person`). As a result, we work with 24 localizable privacy attributes in our dataset representative of 42 of the original 67 VISPR privacy attributes (see Figure 2 for the complete list).

3.3. Dataset Annotation

In this section, we discuss the annotation procedure.

Annotation Tool and Instructions We use VGG Image Annotator tool [12] for annotation. Five expert annotators draw polygons around instances based on an instruction manual. A summary of instructions, definitions of attributes and examples are provided in the supplementary material.

Consensus and Agreement Measure Agreement is calculated w.r.t. images annotated by one of the authors. We measure agreement using Mean Intersection Over Union (mIoU): $\sum \frac{tp}{tp+fp+fn}$ averaged over images.

Consensus Experiment and Annotating `person` We observed 93.8% agreement in consensus task of annotating instances of `person` in 272 images. Annotators separately annotated `person` in remaining images. We obtain 13,171 `person` instances annotated over 5,920 images.

Annotating `face` We observed an agreement of 86.2% (lower due to small sizes of instances) in the consensus task for annotating `face` in 100 images. Using the 5,920 images of people as a starting point, annotators annotated 8,996 instances of faces in separate sets of images.

Annotating Remaining Attributes Images for each of the remaining attributes are annotated successively by at most a single annotator. 8 of the text-based attributes (e.g. `name`, `phone_no`) are annotated using 4-sided polygons or bounding boxes. We gather annotation of 26,676 instances.

Auxiliary Detections We augment all images in the dataset with text detections obtained using the Google Cloud Vision API to aid localization of text-based attributes. This is provided as OCR and bounding box annotation in structured hierarchy of text elements in the order: characters, words, paragraphs, blocks and pages. In addition, we also gather face and landmark bounding box detections using the same API. These detections are solely used as auxiliary input to methods discussed in Section 5 and not for evaluation.

Summary With an annotation effort of ∼800 hours concentrated over four months with five annotators (excluding the authors), we propose the first sizable pixel-labeled privacy dataset of 8,473 images annotated with ∼47.6k instances using 24 privacy attributes.

3.4. Dataset Analysis and Challenges

We now present a brief analysis of the dataset and the new challenges it presents for segmentation tasks. Examples of the proposed attributes and their distribution among the 8k images in the dataset are presented in Figure 2.

Popular datasets [10, 13, 36] provide pixel-level annotation of various common visual objects. These objects are common in visual scenes, such as vehicles (car, bicycle), animals (dog, sheep) or household items (chair, table). Common to all these objects are their distinctive visual cues. Looking at the examples of attributes in Figure 2, one can notice similar cues among the VISUAL attributes, but it is not evident in the others. Recognizing TEXTUAL attributes (such as names or phone numbers) in images instead require

Figure 3: Dilation/Erosion of attribute `fingerprint`

detecting and parsing text information and additionally associating it with prior knowledge. While some of the MULTIMODAL attributes can be associated with visual cues, often the text content greatly helps disambiguate instances (a card-like object could be a `student_id` or `driv_lic`). We also observe a strong correlation between modalities and sizes of instances. We find TEXTUAL instances to occupy on average less than 1% of pixels in images, while MULTIMODAL attributes predominantly occur as close-ups occupying 45% of the image area on average. Consequently, the privacy attributes pose challenges from multiple modalities and require specialized methods to individually address them. Moreover, they provide different insights due to the variance in sizes. Hence, going forward, we treat the modes TEXTUAL, VISUAL and MULTIMODAL as categories to aid analysis and addressing challenges presented by them.

Applicability to other problems We believe the proposed dataset could be beneficial to many other problems apart from visual redactions. In visual privacy, it complements datasets to perform tasks such as person de-identification [4, 19]. Outside of the privacy domain, we also provide a sizable face segmentation dataset with 9k `face` instances, compared to 2.9k in Labeled Faces in the Wild [26] and 200 in FASSEG [27].

4. Understanding Privacy and Utility w.r.t. Redacted Pixels

In this section, we study how redacting ground-truth pixels of attributes influences privacy and utility of the image by conducting a user study on Amazon Mechanical Turk (AMT). The results from this section motivates our approach in Section 5. We will also use the results from this study as a reference point for evaluating our proposed automated methods in Section 6.2.

4.1. Generating Redactions

Given an image I_a containing attribute a, we generate a ground-truth redacted version of the image $I_{\bar{a}}$ by simply blacking-out pixels corresponding to a in the ground-truth.

Spatially extending a We now want to redact fewer or more pixels in image $I_{\bar{a}}$ to understand how this influences the image's privacy and utility. We generate multiple versions of the ground-truth redacted image $\{I^s_{\bar{a}} : s \in S\}$ at different scales of redaction, such that $I^{ns}_{\bar{a}}$ contains n times as many blacked-out pixels of $I^s_{\bar{a}}$. We achieve different scales of redactions by dilating/eroding the ground-truth binary mask of a, as shown in Figure 3. We use seven scales $S = \{0.0, 0.25, 0.5, 1.0, 2.0, 4.0, \inf\}$, where $I^0_{\bar{a}}$ is the unredacted image, $I^1_{\bar{a}} (= I_{\bar{a}})$ is the GT redacted image and $I^{\inf}_{\bar{a}}$ is a completely blacked-out image.

4.2. User Study

We create an AMT project of 1,008 tasks (24 attributes $\times$ 6 images $\times$ 7 scales), each to be responded by 5 unique workers from a pool of 29 qualified workers. Each task contains 2 yes/no questions based on an image $I^s_{\bar{a}}$, one each for Privacy and Utility. We consider *privacy* and *utility* w.r.t. (i) two versions of the same image: $(I_a, I^s_{\bar{a}})$, and (ii) users (AMT workers in our case).

Defining Privacy To understand if attribute a has been successfully redacted in $I^s_{\bar{a}}$, we pose the privacy question in the form: "Is a visible in the image?". We also provide a brief description of the attribute a along with examples. We consider $I^s_{\bar{a}}$ to be *private*, if a majority of the users respond *no*.

Defining Utility To understand utility of an image, we pose the question: "Is the image intelligible, so that it can be shared on social networking websites? *i.e.* does this image convey the main content of the original image (i.e., the image without the black patch)". As a result, we define the utility of an image independent to its aesthetic value and instead associate it with the semantic information. We consider $I^s_{\bar{a}}$ to have *utility*, if a majority of the users respond *yes*.

Measuring Privacy and Utility We label each of the 1,008 images with varying redacted scales their privacy and utility as discussed above. For any given redaction scale s, we aggregate privacy/utility scores simply as the percentage of images considered private/useful. Consequently, an ideal visual redaction has both high privacy and utility.

4.3. Analysis

We now discuss results based on the privacy-utility scores obtained over modes and various sizes (*i.e.* relative size of a in I_a) based on Figure 4.

Privacy is a Step Function We observe in Figure 4 across all plots, that a minimum number of pixels of attribute a need to be removed to effectively redact it from the image. This minimum number corresponds to exactly the ground-truth redaction ($s = 1$) – redacting fewer pixels than this makes the image non-private and redacting more pixels achieves marginal privacy gains. More specifically, we achieve 94% privacy with ground-truth redactions. The imperfect privacy score is predominantly (5/9 failure cases) due to turkers overlooking important details in the question. Apart from this, other cases involve contextual cues revealing the attribute (e.g. wheelchair shadow) and regions that were not annotated (e.g. outline of a `person` at a distance).

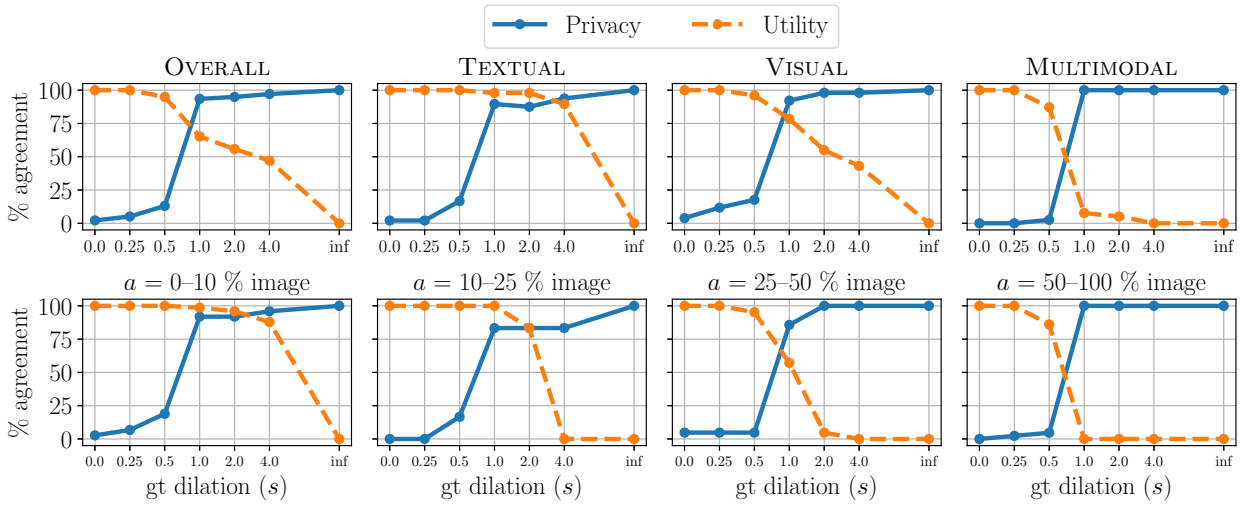

Figure 4: Privacy and Utility using various scales of ground-truth redaction over (Top row) modes (Bottom row) sizes

Gradual Loss in Utility From Figure 4 OVERALL, we find utility to decrease gradually as the size of redacted region increases. Another interesting observation is that utility strongly depends on the size of a in the image. In the bottom row of Figure 4, we see that for smaller GT regions ($a = 0 - 10\%$), we still obtain high utility at larger dilations. However, as the area of the GT regions increases beyond 50% of the image, redaction entails blacking-out the majority of the image pixels and hence zero utility.

Privacy and Utility What can we take away from this while proposing automated methods to preserve privacy while retaining utility? Due to the correlation between modes and sizes, we can predict more pixels for smaller attributes with minimal loss to utility. For instance, for TEXTUAL attributes, we can predict 4x as many ground-truth pixels for redaction. However, for larger ground-truth regions (>50% of image) both privacy and utility are step functions and hence making redaction a choice between privacy and utility.

GT Segmentations are a Good Proxy In general, for images over all attributes and sizes (Figure 4 OVERALL), we see that we can already achieve high privacy *while* retaining considerable utility of the image. Moreover, we obtain near-perfect privacy with the highest utility in all cases at $s = 1$, the ground-truth redactions. This justifies to address privacy attribute redaction as a segmentation task.

5. Pixel-Labeling of Private Regions

In Section 3 we discussed the challenges of attributes occurring across multiple modalities (TEXTUAL, VISUAL, MULTIMODAL). In Section 4, we motivated how ground-truth segmentations in our dataset make a good proxy for visual redactions. In this section we propose automated methods to perform pixel-level labeling (semantic segmentation) of privacy attributes in images, with an emphasis on methods tackling each modality.

We begin with a simple baseline **Nearest Neighbor (NN)**: A 2048-dim feature is extracted using ResNet-50 for each image. At test time, we predict the segmentation mask of the closest training image in terms of L_2 distance.

5.1. Methods for TEXTUAL-centric attributes

To facilitate segmenting textual attributes, for each image we first obtain an ordered sequence of bounding box detections of words and their OCR using the Google Cloud Vision API (as discussed in Section 3.3).

Proxy GT We represent n words in an image as a sequence $[(w_i, b_i, y_i)]_{i=1}^{n}$, where w_i is the word text, b_i is the bounding box and y_i is the label. We use 9 labels (8 TEXTUAL attributes + safe). We assign each y_i in the sequence the ground-truth attribute that maximally overlaps with b_i, or a *safe* label in case of zero overlap. At test-time, we segment pixels in region b_i if a non-safe label is predicted for word w_i. For the test set, we refer to predictions from this proxy dataset as **PROXY** to obtain an upper-bound for our methods on these text detections.

Rule-based Classification (RULES) We use the following rules to label words in the sequence: (i) `name`: if it exists in a set of 241k names obtained from the US Census Bureau website (ii) `location, landmark, home_address`: if it exists in a set of 2.8M locations consisting of countries, states, cities and villages from the GeoNames geographical database [15] (iii) `datetime, phone_no, birth_dt`: if the word contains a digit (iv) `emailadd`: if the word contains the symbol @, we predict this word and adjacent words assuming a format □@□.□

Named Entity Recognition (NER) We use the popular Stanford NER CRFClassifier [14] to label each word of the

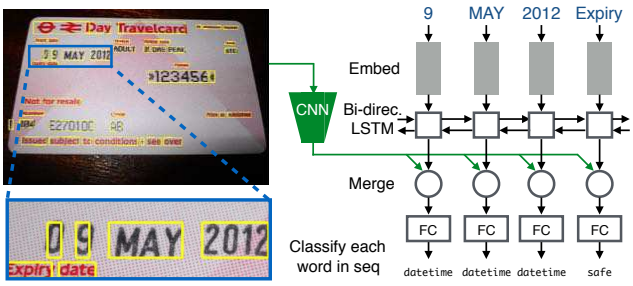

Figure 5: Architecture to perform Sequence Labeling

sequence as from a set of recognized entity classes (e.g. person, organiziation, *etc*.). We use the model which is trained on case-invariant text to predict one of seven entity classes.

Sequence Labeling (SEQ) We train a sequence labeler similar to [22, 33, 38] as shown in Figure 5. We preprocess by replacing all digits with 0s and stem each word to reduce the size of the vocabulary. We tokenize the words in the training sequences using a vocabulary of size 4,149 (number of words with at least 4 occurrences). We embed the words using 100-d GloVe embeddings [42]. To capture the temporal nature, we use two-level Bidirectional LSTMs. At each time-step, we obtain a joint embedding by element-wise multiplication of: the *text* embedding (256-d output of the LSTM) and the *image* embedding (2048-d ResNet-50 [20] feature reduced to 256-d using an FC layer). We classify this joint embedding into 9 labels using an FC layer followed by softmax activation.

5.2. Methods for VISUAL-centric attributes

Recent deep-learning segmentation methods have proven to be effective in localizing objects based on their visual cues. We propose using a state of the art method in addition to few pretrained methods for VISUAL attributes.

Pretrained Models (PTM) We use pretrained methods to classify three classes typically encountered in popular visual scene datasets. (i) `face`: We use bounding box face detections obtained using the Google Cloud Vision API. (ii) `person`: We use the state-of-the-art segmentation method FCIS [34] to predict pixels of COCO class "person" (iii) `lic_plate`: We use OpenALPR [40] to detect license plates in images.

FCIS We retrain all layers of the FCIS model [34] for our task and dataset. We train it for 30 epochs with learning rate 0.0005 over trainval examples and their horizontally mirrored versions. We fine tune it from the model provided by the authors trained for segmentation on MS-COCO [36]. We obtained best results using default hyper-parameters.

5.3. Methods for MULTIMODAL-centric attributes

Recognizing Multimodal attributes (e.g. `driv_lic`, `receipt`) require reasoning over both visual and textual domains. We treat this as a classification problem due to: (i) limited training examples ($\sim$125 per multimodal attribute) (ii) large region of these attributes ($\sim$45% image area), which provides only $\sim$10% utility even after GT-based redaction (Section 4.2).

Weakly Supervised Labeling (WSL) We propose learning a multilabel classifier based on visual-only (**WSL:I**) and visual+text content (**WSL:I+T**). If the class probability of an attribute is beyond a certain threshold, we predict all pixels in the image for the attribute. WSL:I is the same approach used in [41] – a multilabel ResNet-50 [20] classifier. In the case of WSL:I+T, we obtain a multimodal embedding by concatenating visual and text representations. We obtain visual representation (identical to WSL:I) with a ResNet-50 architecture. We obtain text representation by encoding all words in the image. We tried three such variants: (i) *Bag-of-Words (BOW) encoding*: Words in the image are represented as a one-hot vector with vocabulary of size 1,751. (ii) *LSTM encoding*: Identical to SEQ, we encode the word sequence using an LSTM with 128-hidden units. We use output from the last cell as the text representation. (iii) *Conv1D encoding*: We use 1D convolutions to encode the word sequence (typically used for sentence classification tasks [28]) followed by max pooling to obtain a fixed-size text representation In all three cases, we reduce the text-representation to 512-d using an FC+ReLU layer. We report BOW encoding results for **WSL:I+T** in the rest of the paper since this provided the best results.

Salient Object Prediction (SAL) Using WSL:I+T as the base classifier, we use the salient object as an approximation of the attribute's location. We obtain class-agnostic saliency obtained using DeepLab-v2 ResNet [7, 24].

Weakly Supervised Iterative Refinement (IR) For document-like objects, the text regions tend to be densely clustered in images. Hence, after classification using WSL:I+T, we refine the convex hull of the text regions using DenseCRF [32] to "spill into" the document region.

6. Experiments and Discussion

In this section, we discuss segmentation performance (Section 6.1) and privacy-vs-utility performance (Section 6.2) of our proposed methods.

6.1. Evaluating Segmentation Performance

We now evaluate methods proposed in Section 5 in terms of its segmentation performance using Mean Average Precision, suggested in Pascal VOC [13]. This is calculated by averaging area under precision-recall curves over the privacy attributes. We use 50 thresholds uniformly spaced between 0 and 1 to obtain this curve. At each threshold t, we: (i) binarize the prediction score masks per image by thresholding pixel-level scores at t (ii) aggregate pixel-level TP,

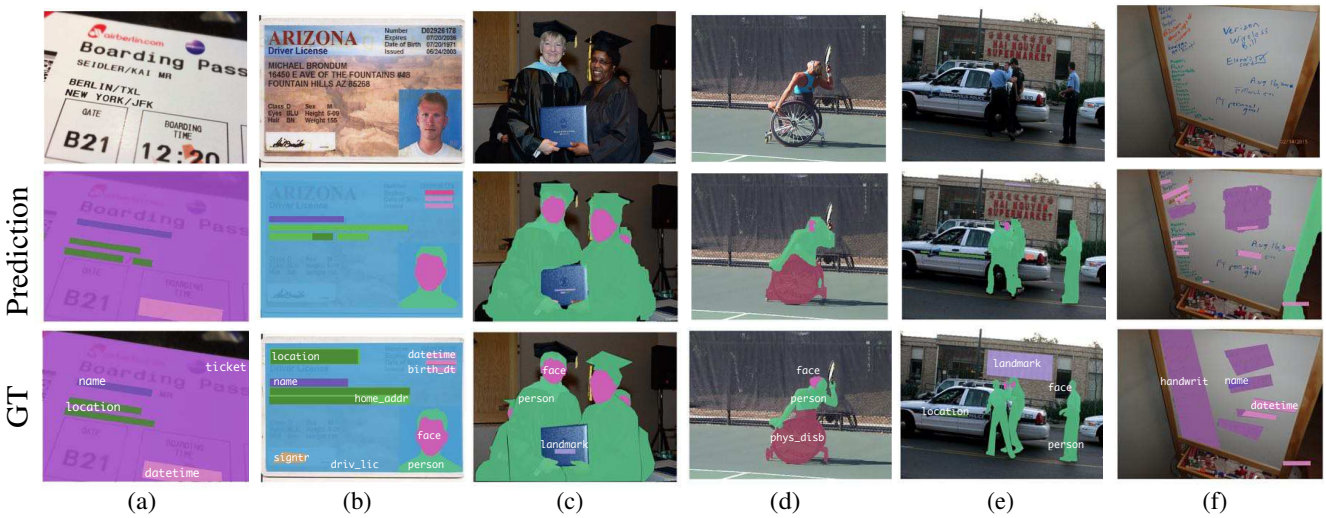

Figure 6: Qualitative examples from our method

FP, FN counts (normalized by image size) per attribute over all images to obtain attribute-level precision and recall. We ignore GT masks containing under 25^2 pixels during evaluation (<1% GT masks).

Table 1 presents the quantitative results of the proposed methods on the test set. Qualitative results in Figure 6 are based on an **ENSEMBLE**, using predictions of SEQ for TEXTUAL, FCIS for VISUAL, WCS:I+T for MULTIMODAL attributes. Auxiliary results and analysis are available in the supplementary material. We generally observe that NN underperforms simple baselines across all modalities, highlighting the difficulty and diversity presented by the dataset.

TEXTUAL We observe: (i) *Patterns, frequency and context*: SEQ achieves the best overall score, justifying the need for special methods to tackle text attributes. It is reasonably effective in detecting `datetime` (Fig. 6a), `emailadd` and `phone_no` due to patterns they often display. We additionally find SEQ detect attributes which often require prior knowledge (e.g. `name`, `location`). The common success modes in such cases are when the words are popular entities (e.g. "Berlin" in Fig. 6a) or have discriminative visual/textual context (e.g. detecting `home_addr` in Fig. 6b). (ii) *Challenges imposed by text detections*: PROXY represents an upper bound to our textual methods. The low scores highlights the difficulty of text detection and this is especially severe for scene and handwritten text detection, a frequent case in our dataset (e.g. Fig. 6e,f). Moreover, our text detections do not perfectly overlap with ground-truth annotations. Since text regions are small, we additionally pay a high performance penalty even for correct detections (e.g. IoU=0.42 for `home_addr` in Fig. 6b). Moreover, even in the case of correct text detections, we observe failures in OCR which affects the quality of input for dependent methods. This can be observed by the under-performance of NER, which is typically very effective on clean sanitized text.

VISUAL We observe: (i) *The unreasonable effectiveness of FCIS*: We obtain the highest score in the VISUAL category using FCIS. We find FCIS to be highly effective localizing visual objects commonly encountered in other datasets (e.g. `person`, `face`). Moreover, we find it achieves reasonable performance even when there is a lack of training data (e.g. only <60 examples of `fingerpr`, `phys_disb`, see Fig. 6d). The common failure modes are difficult examples (e.g. `face` in Fig. 6e) and uncommon visual objects (e.g. `signtr` in Fig. 6b). (ii) *Comparison with Baselines*: PTM achieves comparable results for `person`, due to Flickr images used to train both models. However, it underperforms for `face` (detections are not precise enough) and `lic_plate` (poor performance in the wild).

MULTIMODAL We observe: (i) *WSL:I is a good simple baseline*: WSL:I achieves reasonable performance (45.4) for multimodal attributes, compared to other modes (1.5 in text and 20.8 in visual) although the prediction spans the entire image. This is attributed to large size of MULTIMODAL instances found in images. (ii) *Multimodal reasoning helps*: We find WSL:I+T improves performance over WCS:I by 20%, justifying the need for methods to perform multimodal reasoning to detect these attributes. This is particularly necessary to disambiguate similar looking visual objects (e.g. card-like objects `driv_lic` and `stud_id`, Fig. 6b). (iii) *Precision-Recall trade-off*: We find precision for WSL:I+T for this method can be improved for some attributes (e.g. `cr_card`, `ticket`) by IR, which instead of the entire image, predicts only the smoothened hull of text regions. We observe FCIS achieve the best overall score due to higher precision.

TEXTUAL

Method	mAP	location	home addr	name	birth dt	phone no	land mark	date time	email add
PROXY	45.0	31.7	37.8	48.7	52.5	52.6	33.6	52.4	50.8
NN	0.9	0.3	1.9	0.4	*0.7*	0.0	3.1	0.6	0.0
NER	3.0	*6.0*	1.7	4.4	0.5	0.0	0.5	10.9	0.0
RULES	4.2	3.1	0.5	2.8	0.6	1.4	1.2	6.4	*17.5*
FCIS	*7.2*	4.3	0.2	*9.8*	0.1	*2.5*	**27.6**	*12.9*	0.0
SEQ	**26.8**	**18.4**	**19.4**	**19.1**	**25.1**	**45.8**	*13.9*	**33.4**	**38.9**

VISUAL

Method	mAP	face	licp late	person	nudity	hand writ	phy disb	med hist	fing erpr	sig ntr
NN	16.6	9.0	16.0	33.6	6.2	37.5	11.4	18.9	16.9	0.1
WSL:I	*20.8*	5.0	4.3	30.3	*16.4*	*49.9*	*13.7*	*37.7*	*28.8*	*1.3*
PTM	20.0	*47.6*	*44.5*	**88.3**	0.0	0.0	0.0	0.0	0.0	0.0
FCIS	**68.3**	**83.8**	**77.9**	*87.0*	**69.7**	**80.7**	**59.0**	**45.8**	**68.1**	**42.6**

MULTIMODAL

Method	mAP	cr card	pass port	driv lic	stud id	mail	rece ipt	tic ket
NN	24.1	10.5	49.5	19.9	14.5	20.6	17.1	36.7
WSL:I+T	*55.6*	27.7	*68.8*	**83.3**	**56.1**	**41.4**	54.2	58.0
SAL	36.2	**55.9**	37.2	23.8	30.4	8.1	42.5	55.1
IR	53.6	41.7	51.2	67.8	48.1	36.9	*57.2*	*72.5*
FCIS	**59.2**	*53.2*	**76.3**	66.5	*50.3*	33.1	**59.4**	**75.4**

Table 1: Quantitative results of our methods for segmenting privacy regions. **Bold** numbers denote highest and *italicized* numbers second highest scores in the columns.

6.2. Privacy vs. Utility Trade-off by Automatic Redaction

In the previous section, we evaluated our approaches w.r.t. segmentation quality. Now, we ask how effective are redactions based on our proposed methods in terms of privacy and utility?

To answer this, we once again run the user study in Section 4.2 on AMT, but now by redacting proposed pixels of our automated method over those exact images. To vary the number of predicted pixels, we vary the threshold to binarize the predicted score masks over attributes. As a result, we obtain 6-8 redacted versions for each of the 144 images (24 attributes $\times$ 6 images). Each image is labeled by 5 unique qualified AMT workers.

Results We obtain privacy-utility scores for each threshold and plot it as a curve in Figure 7. We also plot the scores obtained for different dilations of redacted ground-truth annotated region. It should be noted that perfect redactions are unavailable to us and we use these ground-truth based redactions (or manual redactions) only to serve as a reference. We evaluate performance by calculating area under the curve (AUC). We observe: (i) Overall, we find our method obtain a privacy-utility score of 65% – a relative performance of 83% compared to redactions using ground-truth annotation from the dataset. (ii) MULTIMODAL attributes present a hard choice between privacy and utility, as these regions are often large. We find the slightly lower AUC(gt) to be an artifact of sampling. (iii) Although we obtain a low mAP for TEXTUAL attributes, we observe an 81% privacy-utility score. This occurs as we can now over-predict regions, exhibiting low precision and high recall w.r.t. segmentation, but yet retaining high utility due to their small size. Consequently, we can predict more text pixels "for free".

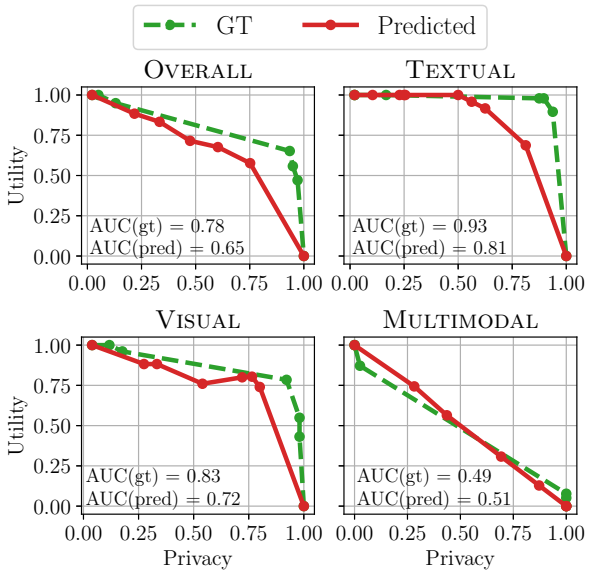

Figure 7: Comparing redactions using predicted and ground-truth segmentations

Based on these observations, we find the automatic redactions of our models trained on the proposed dataset show highly promising results – they closely mimic performance achieved by redacting ground-truth regions across a broad range of private information.

7. Conclusion

We proposed a redaction by segmentation approach to aid users selectively sanitize images of private content. To learn automated approaches for this task, we proposed the first sizable visual redactions dataset containing images with pixel-level annotations of 24 privacy attributes. By conducting a user study, we showed that redacting ground-truth regions in this dataset provides near-perfect privacy while preserving the image's utility. We then presented automated approaches to segment privacy attributes in images and observed that we can already reasonably segment these attributes. By performing a privacy-vs-utility evaluation of our automated approach, we achieved a highly encouraging 83% performance w.r.t. GT-based redactions.

Acknowledgement This research was partially supported by the German Research Foundation (DFG CRC 1223). We thank Anna Khoreva and Alina Dima for their feedback.

References

[1] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *PET*, 2006. 2

[2] C. Bauckhage, A. Jahanbekam, and C. Thurau. Age recognition in the wild. In *ICPR*, 2010. 2

[3] M. Boyle, C. Edwards, and S. Greenberg. The effects of filtered video on awareness and privacy. In *CSCW*, 2000. 2

[4] K. Brkic, I. Sikiric, T. Hrkac, and Z. Kalafatic. I know that person: Generative full body and face de-identification of people in images. In *CVPRW*, 2017. 1, 4

[5] V. T. Chakaravarthy, H. Gupta, P. Roy, and M. K. Mohania. Efficient techniques for document sanitization. In *CIKM*, 2008. 2

[6] S.-L. Chang, L.-S. Chen, Y.-C. Chung, and S.-W. Chen. Automatic license plate recognition. *IEEE Trans. Intelligent Transportation Systems*, 2004. 2

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 6

[8] R. Chow, P. Golle, and J. Staddon. Detecting privacy leaks using corpus-based association rules. In *KDD*, 2008. 2

[9] R. Chow, I. Oberst, and J. Staddon. Sanitization's slippery slope: the design and study of a text revision assistant. In *SOUPS*, 2009. 2

[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 3

[11] B. Debatin, J. P. Lovejoy, A.-K. Horn, and B. N. Hughes. Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of Computer-Mediated Communication*, 2009. 2

[12] A. Dutta, A. Gupta, and A. Zissermann. Vgg image annotator (via), 2016. `http://www.robots.ox.ac.uk/~vgg/software/via/` Accessed: 2017-11-08. 3

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 2, 3, 6

[14] J. R. Finkel, T. Grenager, and C. D. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005. 5

[15] Geonames geographical database. `http://www.geonames.org/` Accessed: 2017-11-08. 5

[16] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 2

[17] R. Gross, E. Airoldi, B. Malin, and L. Sweeney. Integrating utility into face de-identification. In *PET*, 2005. 2

[18] R. Hasan, E. Hassan, Y. Li, K. Caine, D. J. Crandall, R. Hoyle, and A. Kapadia. Viewer experience of obscuring scene elements in photos to enhance privacy. In *CHI*, 2018. 2

[19] E. T. Hassan, R. Hasan, P. Shaffer, D. Crandall, and A. Kapadia. Cartooning for enhanced privacy in lifelogging and streaming videos. In *CVPRW*, 2017. 1, 2, 4

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[21] J. Hendrik Metzen, M. Chaithanya Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. In *ICCV*, 2017. 2

[22] Z. Huang, W. Xu, and K. Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015. 6

[23] S. Joon Oh, R. Benenson, M. Fritz, and B. Schiele. Faceless person recognition; privacy implications in social media. In *ECCV*, 2016. 1

[24] S. Joon Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele. Exploiting saliency for object segmentation from image level labels. In *CVPR*, 2017. 6

[25] S. Joon Oh, M. Fritz, and B. Schiele. Adversarial image perturbation for privacy protection – a game theory perspective. In *ICCV*, 2017. 1, 2

[26] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller. Augmenting CRFs with Boltzmann machine shape priors for image labeling. In *CVPR*, 2013. 4

[27] K. Khan, M. Mauro, and R. Leonardi. Multi-class semantic segmentation of faces. In *ICIP*, 2015. 4

[28] Y. Kim. Convolutional neural networks for sentence classification. In *EMNLP*, 2014. 6

[29] P. Korshunov, C. Araimo, F. D. Simone, C. Velardo, J.-L. Dugelay, and T. Ebrahimi. Subjective study of privacy filters in video surveillance. *MMSP*, 2012. 2

[30] P. Korshunov and T. Ebrahimi. Pevid: privacy evaluation video dataset. In *SPIE*, 2013. 2

[31] P. Korshunov and T. Ebrahimi. Using warping for privacy protection in video surveillance. *DSP*, 2013. 2

[32] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 6

[33] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In *NAACL*, 2016. 6

[34] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, 2017. 2, 6

[35] Y. Li, N. Vishwamitra, B. P. Knijnenburg, H. Hu, and K. Caine. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *CVPRW*, 2017. 2

[36] 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 *ECCV*, 2014. 2, 3, 6

[37] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[38] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *ACL*, 2016. 6

[39] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *CVPR*, 2017. 2

[40] Openalpr. `https://github.com/openalpr/openalpr` Accessed: 2017-11-08. 6

[41] T. Orekondy, B. Schiele, and M. Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *ICCV*, 2017. 1, 2, 6

[42] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 6

[43] N. Raval, A. Machanavajjhala, and L. P. Cox. Protecting visual secrets using adversarial nets. In *CVPRW*, 2017. 1

[44] D. Sánchez and M. Batet. Toward sensitive document release with privacy guarantees. *Engineering Applications of AI*, 2017. 2

[45] D. Sánchez, M. Batet, and A. Viejo. Detecting sensitive information from textual documents: An information-theoretic approach. In *MDAI*, 2012. 2

[46] D. Sánchez, M. Batet, and A. Viejo. Automatic general-purpose sanitization of textual documents. *IEEE Transactions on Information Forensics and Security*, 2013. 2

[47] M. Shao, L. Li, and Y. Fu. What do you do? occupation recognition in a photo via social context. In *ICCV*, 2013. 2

[48] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *ACM CCS*, 2016. 2

[49] Q. Sun, B. Schiele, and M. Fritz. A domain based approach to social relation recognition. In *CVPR*, 2017. 2

[50] X. Sun, P. Wu, and S. C. H. Hoi. Face detection using deep learning: An improved faster rcnn approach. *CoRR*, 2017. 2

[51] A. Tonge and C. Caragea. Privacy prediction of images shared on social media sites using deep features. *arXiv preprint arXiv:1510.08583*, 2015. 2

[52] P. A. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 2001. 2

[53] G. Wang, A. C. Gallagher, J. Luo, and D. A. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *ECCV*, 2010. 2

[54] E. S. Xioufis, S. Papadopoulos, A. Popescu, and Y. Kompatsiaris. Personalized privacy-aware image classification. In *ICMR*, 2016. 2

[55] S. Zerr, S. Siersdorfer, J. Hare, and E. Demidova. I know what you did last summer!: Privacy-aware image classification and search. In *ACM SIGIR*, 2012. 2

[56] H. Zhang, W. Jia, X. He, and Q. Wu. Learning-based license plate detection using global and local features. In *International Conference on Pattern Recognition (ICPR)*, 2006. 2

[57] W. Zhou, H. Li, Y. Lu, and Q. Tian. Principal visual word discovery for automatic license plate detection. *IEEE Transactions on Image Processing*, 2012. 2