Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images (Supplementary Material)

Tribhuvanesh Orekondy Mario Fritz Bernt Schiele

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

A. Contents

The appendix contains:

- Detailed descriptions, examples and auxiliary analysis of the 24 privacy attributes discussed in Section 3.2
- Extended discussion on evaluating privacy of redacted/non-redacted images on Amazon Mechanical Turk
- Quantitative results to supplement Section 6.1
- Qualitative results to supplement Figure 6
- Implementation details and qualitative results to supplement Section 4 and Section 6.2

B. Privacy Attributes

In this section, we provide detailed descriptions and examples of the 24 Privacy Attributes used in the proposed dataset. We also present a brief supplementary analysis of the conditional co-occurrence of these attributes in the dataset.

Detailed Descriptions and Instructions In Figures 5-7, we provide detailed descriptions and examples of the 24 privacy attributes grouped by category, which was discussed in Section 3.2. The descriptions briefly summarize the instructions provided to the annotators. The figures displays instance-agnostic ground-truth annotations of respective attributes. Ground-truth annotations are stored in a format similar to MS-COCO [4].

TEXTUAL, `signtr` and `handwrit` attributes are annotated using 4-sided polygons or bounding-boxes. For TEXTUAL attributes, only Latin-based words understandable by English-speakers are annotated. For remaining attributes, the objects are enclosed in a polygon. In case of severe occlusion, the object is enclosed using multiple polygons.

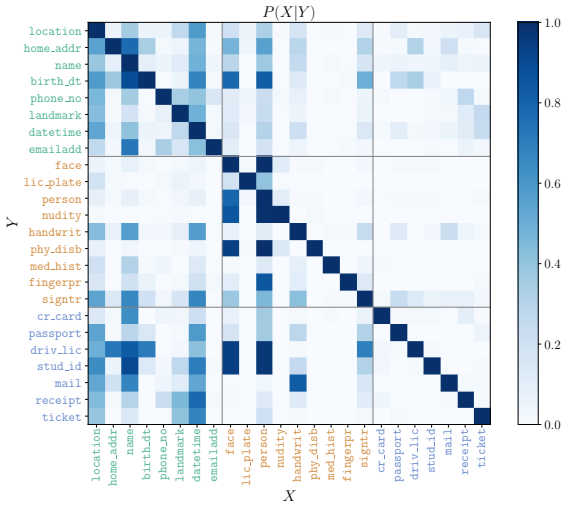


Figure 1: Conditional co-occurrence matrix. Groups of attributes are sorted by categories. Color codes used for attribute categories: TEXTUAL, VISUAL, MULTIMODAL

Auxiliary Privacy Attribute Analysis Figure 1 represents the conditional co-occurrence matrix (*i.e.* probability that attribute X occurs in an image containing attribute Y) of the 24 privacy attributes in images. The privacy attributes along rows and columns are sorted by category. From this plot, we find: (i) Images of MULTIMODAL attributes often appear alongside a variety of TEXTUAL attributes (bottom-left block of matrix). (ii) However, the contrary is not true – TEXTUAL attributes do not frequently occur only in the presence of MULTIMODAL attributes (top-right block of matrix). (iii) `person` and `face` occur frequently alongside other VISUAL attributes as they are central to many common visual scenes (central block of matrix).

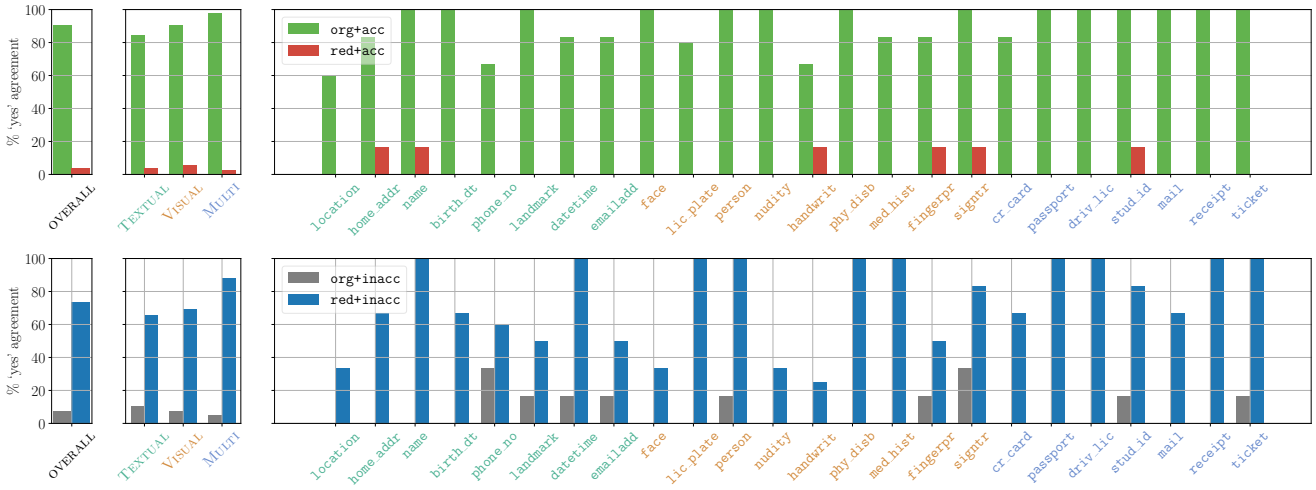


Figure 2: Majority agreement of AMT workers for the privacy question of the form : "Is attribute X present in the image?" over four types of tasks: {org, red} × {acc, inacc}

C. Privacy and Control Questions on AMT

In this section, we present an analysis of responses generated by Amazon Mechanical Turk (AMT) workers for the privacy question discussed in Section 4.2. Previously, we discussed the privacy and utility of images as a result of spatially extending/contracting the GT mask. Now, we extend the discussion on the privacy of images with/without the GT-based redaction and additionally when workers are intentionally asked False Positive questions. We asked FP control questions to test whether workers hallucinate objects or memorize correlations and answer all redacted images as private and non-redacted images as not-private.

Experimental Setup Similar to the task in Section 4.2., qualified workers are presented with an image and a yes/no privacy question of the form "Is attribute X present in the image?". The image can either be in its original form i.e., containing no redactions (org) or have certain regions redacted (red). The privacy question can be accurate (acc) i.e., attribute X is present in the image and has optionally been redacted. Alternatively, it could be inaccurate (inacc) i.e., attribute X is not present in the image or another attribute $Y \neq X$ has been redacted instead. Using these combinations, we generate four types of tasks:

1. org+acc: In the original image containing attribute X, workers are asked if X is present.
2. red+acc: In the redacted image with attribute X redacted, workers are asked if X is present.
3. org+inacc: In the original image, workers are asked if X is present, although it is not in the ground-truth annotation of the image.
4. red+inacc: In an image with attribute X redacted, workers are asked if attribute $Y \neq X$ is present.

For each of these four types, we generate 144 questions (24 attributes × 6 images), each to be answered by 5 unique workers. We aggregate responses by computing majority agreement.

Discussion Figure 2 displays the results of the experiment. The y-axis indicates the majority agreement of workers responding 'yes' to the privacy question. We observe: (i) For a question of the form "Is attribute X present in the image", one could hypothesize that workers might tend to predominantly answer 'yes' in case of org image and 'no' for red images. However, we observe this is not the case – workers mostly provide reasonable answers and do not develop a knee-jerk reaction to presence/absence of redactions in images w.r.t to privacy. (ii) org+acc: Workers perform slightly worse in detecting smaller attributes. Overall, the agreement is 91%, which shows that they are in most cases able to correctly detect the presence of privacy relevant regions in images. (iii) red+acc: We expect 0% score in the case the attribute is perfectly redacted. However, we observe 6/144 failure cases where the user indicate the attribute is present in the image in spite of redaction. In 5 of these cases, we speculate the workers falsely recognize an incorrect region as the attribute due to context (e.g., text of an hand-written letter recognized as signature). In the last case (handwrit), we observe it occurs due to incomplete recall of pixel annotation in ground-truth. (iv) org+inacc: Ideally, we expect 0% score of turkers hallucinating absent attribute in the image. However, we observe very few failure cases (∼4/11) when this occurs. In such cases, we find the images contain some visual cues corresponding to the attribute in question (e.g., stud_id was possibly confused

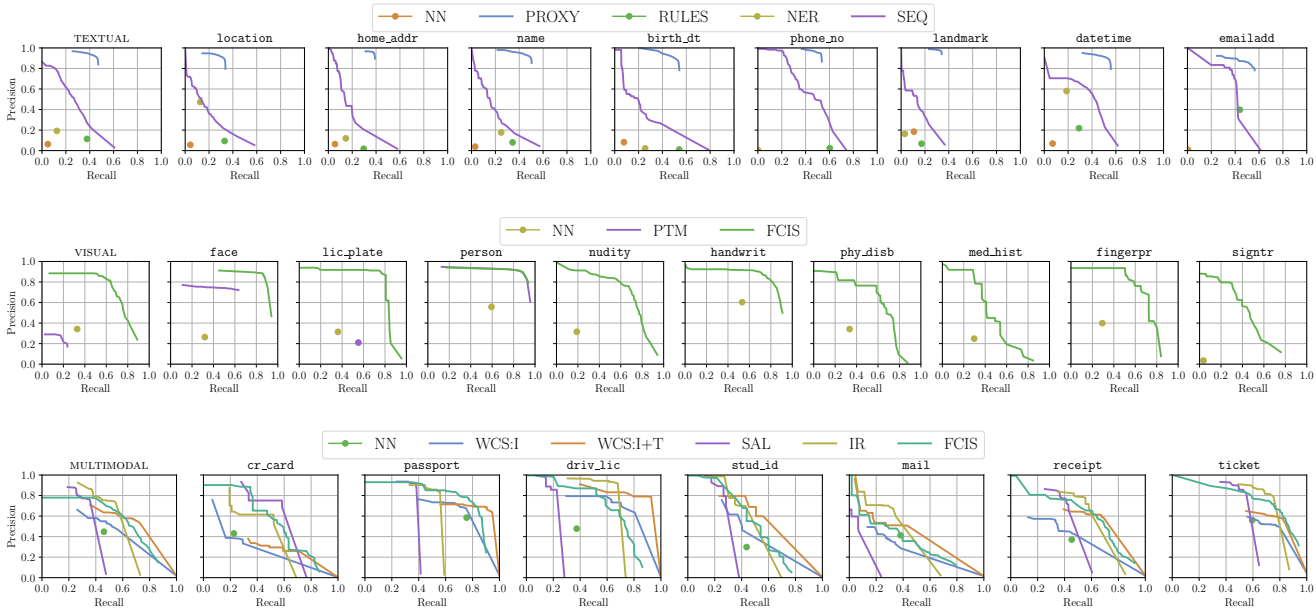


Figure 3: Precision-Recall curves for methods in Table 1

for a `driv_id`). We observe other failure cases often occur due to imperfect recall of attributes in the VISPR dataset. (v) `red+inacc`: While typically we expect 100% agreement since another attribute Y has been redacted instead of X in question, there is often significant overlap between the regions of the attributes. As a result, often redacting attribute Y also redacts X and hence we observe a score below 100%.

D. Extended Quantitative Discussion

The Precision-Recall curves of methods proposed in Table 1 are presented in Figure 3. The first column represents averaged category performance. We plot these curves by thresholding our methods at 50 uniform intervals in the range $[0, 1]$. Similar to Pascal VOC [2], we correct the curves to have monotonically decreasing precision by setting precision at r to be the highest precision at $r' \geq r$. Moreover, precision at $r = 0$ is extrapolated as highest precision at $r' \geq 0$. We calculate Average Precision as area under this curve using trapezoidal rule.

Auxiliary Discussion From PR curves in Figure 3, we observe: (i) The under-performance NN indicates diversity and difficulty of the dataset. (ii) TEXTUAL: We find the best performance using SEQ. PROXY denotes a rough upper bound. We find SEQ obtain slightly higher recall as it predicts overlapping masks. (iii) VISUAL: We find FCIS achieve the best performance. For `person`, we find a similar curve with PTM since both have the same architecture and images from the same domain (Flickr) used for training. (iv) MULTIMODAL: FCIS achieves slightly higher category performance compared to others. WCS:I+T generally achieves better recall across all attributes. IR/SAL improves precision of WCS:I+T by trading off recall.

Effect of Size Figure 4 displays the intersection over union (IoU) score over the ground-truth images across the relative size of privacy attributes in the image. These scores were computed over the model ENSEMBLE using the following thresholds: 0.39 for FCIS, 0.14 for SEQ and 0.08 for WSL:I+T. We observe: (i) TEXTUAL: On an average, textual attributes occupy just 2.5% of the image and this additionally presents challenges for detection and segmentation. (ii) VISUAL: We find that performance of our method is influenced by the size of the attribute in the image, with an IoU of 0.65 for regions smaller than the average visual region (19% of the image) and 0.8 for regions larger than it. (iii) MULTIMODAL: When predicting all pixels of the image as some attribute, the IoU value is equal to the relative size of the ground-truth region in the image and hence we find the IoU of our method to be concentrated along the diagonal. Since multimodal attributes occupy large regions (70% of the image on average), we find that this simple strategy is suitable for segmentation.

E. Qualitative Results for Segmentation

We present qualitative results in Figure 11 to supplement results in Figure 6 and discussion in Section 6.1. We present the qualitative results per attribute, sorted by their Intersection Over Union (IoU) Scores. Hence, figures on top repre-

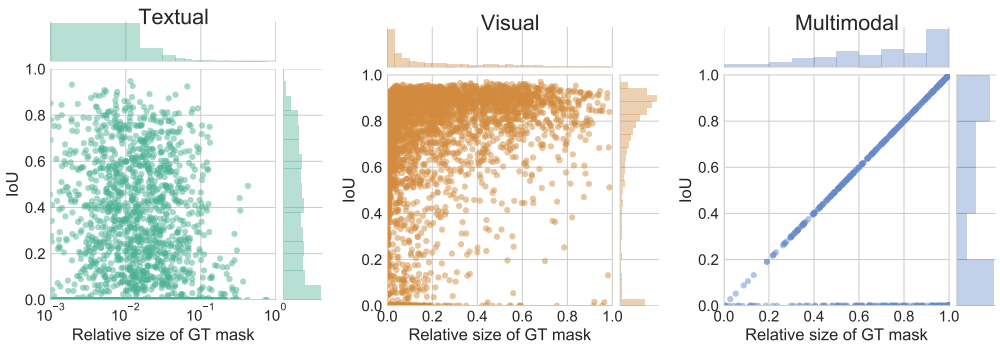


Figure 4: Analyzing IoUs vs. relative size of ground-truth region for the model ENSEMBLE

sent common success modes and figures at the bottom represent common failure modes. These results were obtained using ENSEMBLE by choosing the operating point with the highest IoU score per mode.

F. Privacy vs. Utility Trade-off

In this section, we provide implementation details on the redaction scaling strategy used for ground-truth redactions (Section 4.1) and predicted redactions (Section 6.1). In both cases, we perform a black-out of relevant pixels. For `phy_disb`, we black-out w.r.t. a bounding-box region since we observed the silhouette is a strong visual indicator of the attribute. In addition, we provide qualitative results for these strategies in Figures 12 and 13 to supplement Figure 3.

Scaling Ground-truth Redactions We scale ground-truth redactions using super-pixels to roughly adhere to edges and object boundaries. The downscaled image is first represented using 3000-5000 superpixels generated using SLIC0 [1]. We represent the ground-truth binary mask per attribute using a 0-1 labeling over the graph of super-pixels, where 1 represents the node (superpixel) belongs in the redaction. To *dilate*, we iteratively add 0-nodes with most number of adjacent 1-nodes. To *erode*, we perform the same operation with an inverted ground-truth binary-mask. We parameterize the scaling using $s \in S$ (where $S = \{0.0, 0.25, 0.5, 1.0, 2.0, 4.0, \inf\}$), representing the dilation/erosion factor of the ground-truth mask.

Scaling Predicted Redactions From the ENSEMBLE method, we obtain softmax probability score masks $\mathbb{R}^{w \times h \times k}$ for k attributes per image. We compute multiple thresholds per attribute to binarize the score masks, such that at threshold $t \in T$, t times the number of ground-truth attribute pixels are redacted over the entire test-set of images. We use $T = \{0.25, 0.5, 1.0, 2.0, 4.0, 8.0\}$. For TEXTUAL attributes, we use an additional threshold such that all detected text is redacted.

Qualitative Results Auxiliary Discussion Figure 12 and 13 displays examples of common success and failure modes w.r.t. to the attribute mentioned. All images in these figures are from the test set. P and U indicate privacy and utility score, which is simply the percentage of $\sim$5 AMT workers who agree to the privacy and utility questions. High P indicates the image is private w.r.t. to attribute a and high U indicates the image is intelligible. In these figures, we find: (i) For small private regions, we can redact more pixels without affecting utility (Figure 12 `location` and `face`) (ii) MULTIMODAL attributes often display a hard choice between privacy and utility (Figure 12 `mail`) (iii) Text detections or OCR is a common failure mode with handwritten text for automatic redactions (Figure 13 `home_addr`) (iv) Some difficult MULTIMODAL attributes (Figure 13 `stud_id`) can be detected only at high thresholds, entailing complete redactions of many FP images too (v) Figure 13 `fingerpr` represents one of the failure cases for ground-truth redaction discussed in Section 4.3, where AMT turkers overlook details in the question. In this particular case, the workers were asked to only consider fingerprints from fingertips. However, even at $s = 1$ where the finger-tips are redacted, many workers incorrectly answer fingerprints as being visible.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 2012. 4

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

[3] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCVW*, 2011. 6

[4] 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. 1

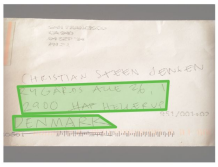
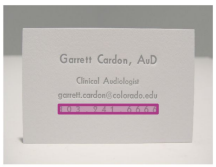
Attribute	Example	Description
Location (`location`)		Region of the image depicting where the photographer might have visited. Includes the following cases: Street signs, addresses, GPS co-ordinates, flags.
Home Address (`home_addr`)		Someone's home address based on the context, such as on an identity card or mail.
Name (`name`)		Someone's name such as on a name-tag or identity card. Any recognizable name in Latin-based text is included, including that of popular figures.
Birth Date (`birth_dt`)		Someone's date of birth (day, month and/or year) determined based on context, such as on identity cards or passports.
Phone no. (`phone_no`)		A syntactically-correct phone number (either personal or business), determined either based on context or pattern.
Landmark (`landmark`)		Name of a store, restaurant or a business such as on a store front or a receipt.
Date/Time (`datetime`)		A date or time, such as revealing a time-frame when the photograph might have been captured.
Email address (`emailadd`)		A syntactically-correct email address

Figure 5: Descriptions and examples of TEXTUAL attributes. privacy attributes. For readability, we display images where attributes are salient.

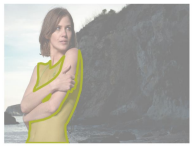
Attribute	Example	Description
Face (`face`)		Region indicating a person's face, containing all visible facial landmarks discussed in [3]. Regions occluded by hair or masks are excluded.
License Plate (`lic_plate`)		Region containing a license plate or vehicle registration or identification number in any language/country. We consider any motorized vehicle (e.g. cars, motorbike, train).
Person (`person`)		Region indicating any part of a person or their reflections. Includes person's body along with wearables (e.g. hats, goggles, backpacks). Excludes objects the person is holding (e.g. shopping bag, guitar).
Nudity (`nudity`)		Torso and thigh region of a person, if skin is completely/partially visible in this region.
Handwriting (`handwrit`)		Someone's handwritten text in any language.
Physical Disability (`phy_disb`)		Region indicating either a) special equipment used by a physically disabled person (e.g. wheelchair) or b) region around limbs, if limbs are absent.
Medical History (`med_hist`)		Any pharmaceutical consumable such as pills, capsules or syrups (including their containers and packaging).
Fingerprint (`fingerpr`)		Someone's finger-tips if ridges are clearly visible upon zooming-in or fingerprint impressions on any surface.
Signature (`signtr`)		Region indicating someone's signature

Figure 6: Descriptions and examples of VISUAL attributes. For readability, we display images where attributes are salient.

Attribute	Example	Description
Credit Card (`cr_card`)		Either front, rear or any details of a credit card or similar monetary instrument
Passport (`passport`)		Any page (including cover) of a Passport
Drivers License (`driv_lic`)		Front, rear or written details of a Drivers License or driving permit
Student ID (`stud_id`)		Front or rear of a student identity card
Mail (`mail`)		Mail including hand-written letters, post-cards or packages
Receipt (`receipt`)		A document indicating a financial transaction, such as receipts or checks
Ticket (`ticket`)		A ticket, such as for travel, concert or sports match

Figure 7: Descriptions and examples of MULTIMODAL attributes. For readability, we display images where attributes are salient.

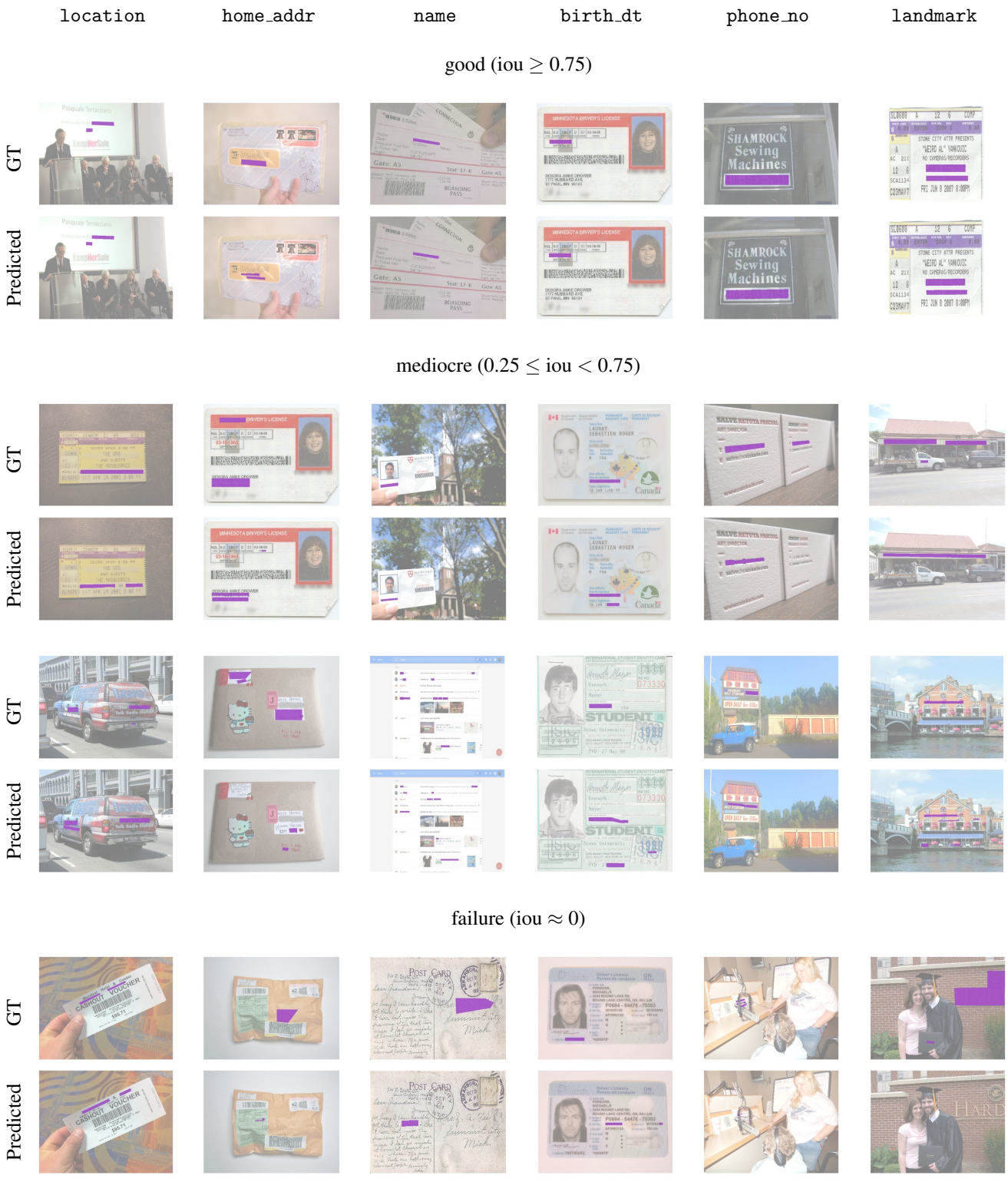


Figure 8: Qualitative results per attribute. In each pair of images, top is ground-truth segmentation and bottom is prediction. Pairs of images in each column are sorted by IoU scores (high to low).

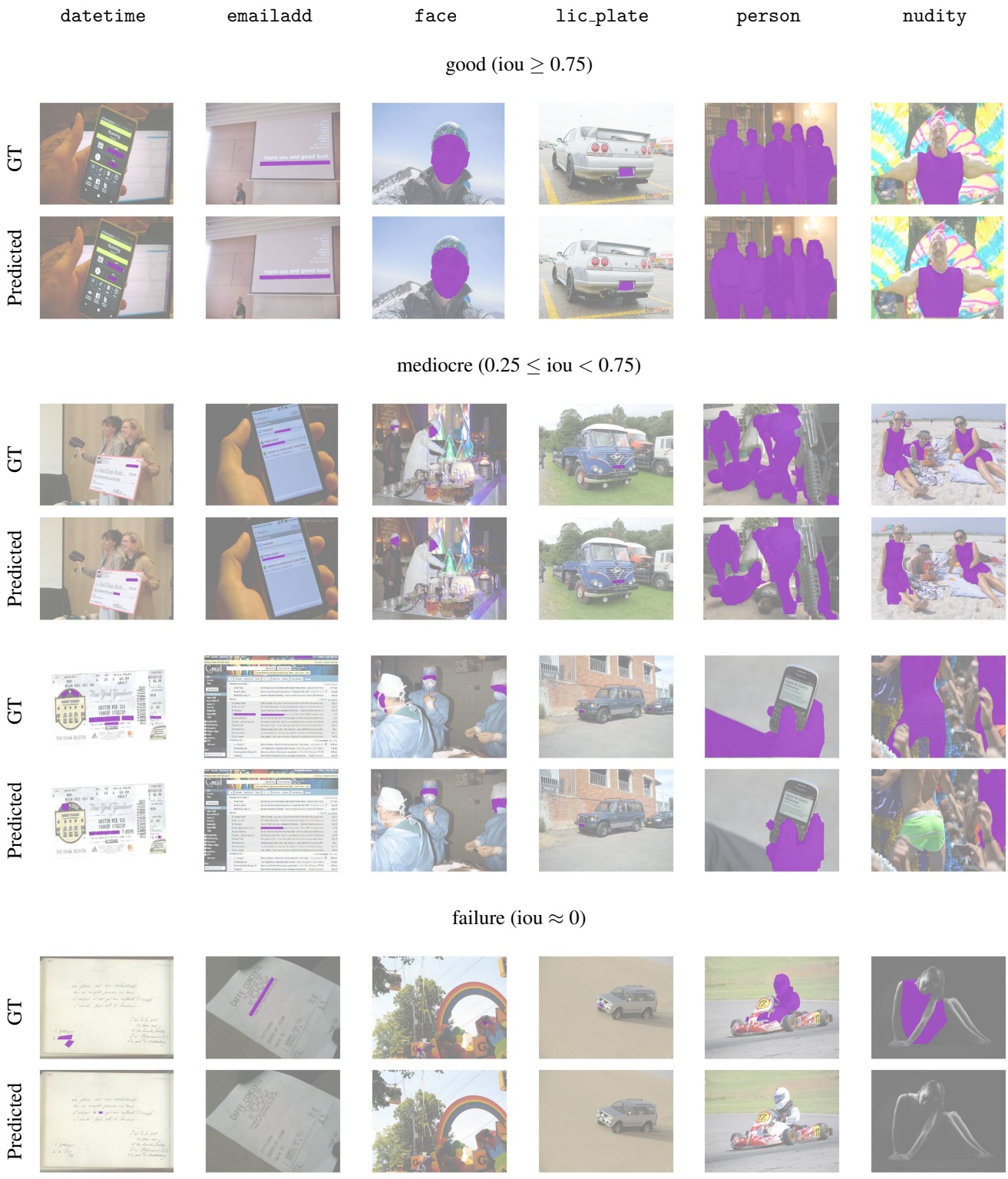


Figure 9: Qualitative results per attribute. In each pair of images, top is ground-truth segmentation and bottom is prediction. Pairs of images in each column are sorted by IoU scores (high to low).

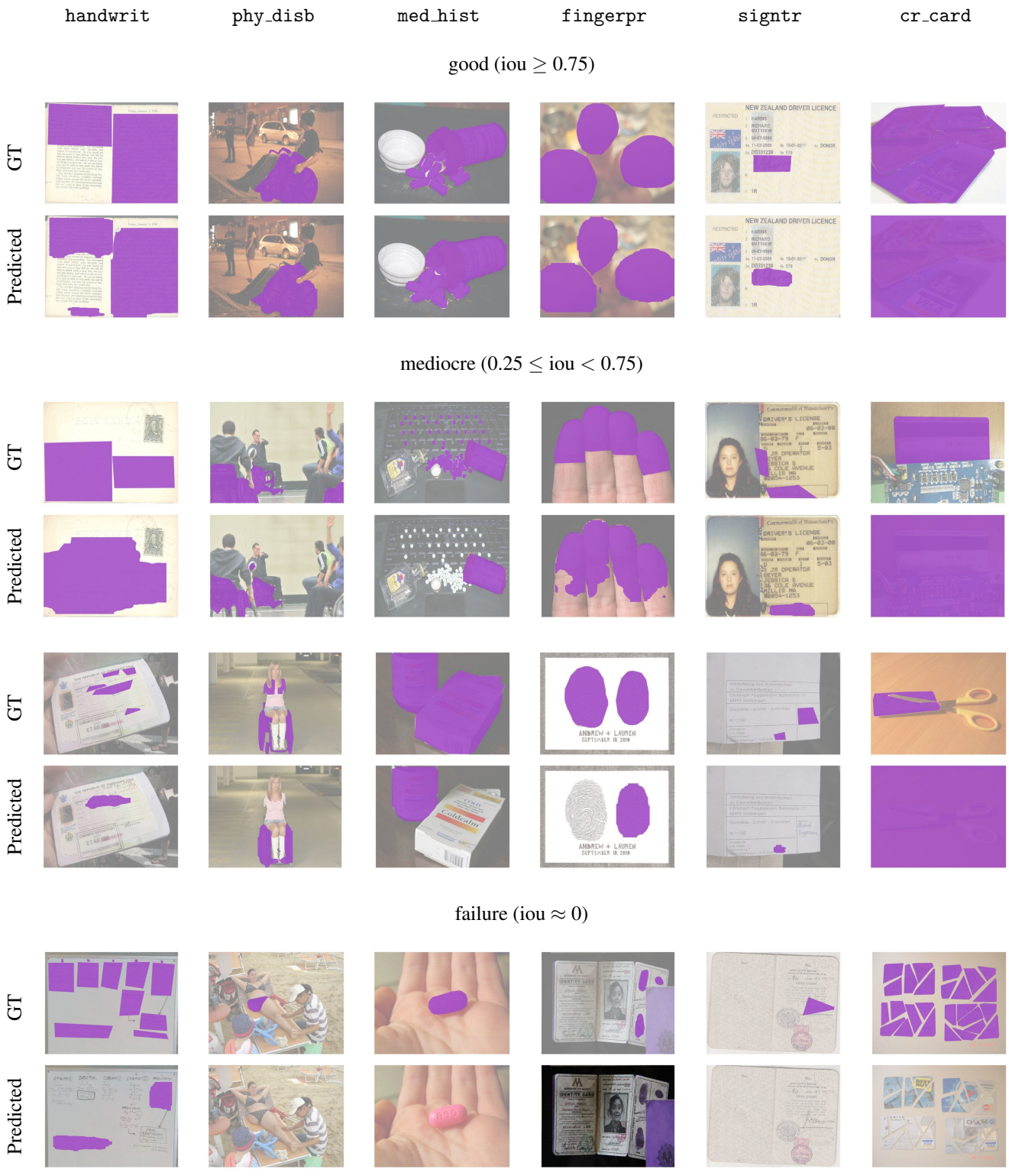


Figure 10: Qualitative results per attribute. In each pair of images, top is ground-truth segmentation and bottom is prediction. Pairs of images in each column are sorted by IoU scores (high to low).

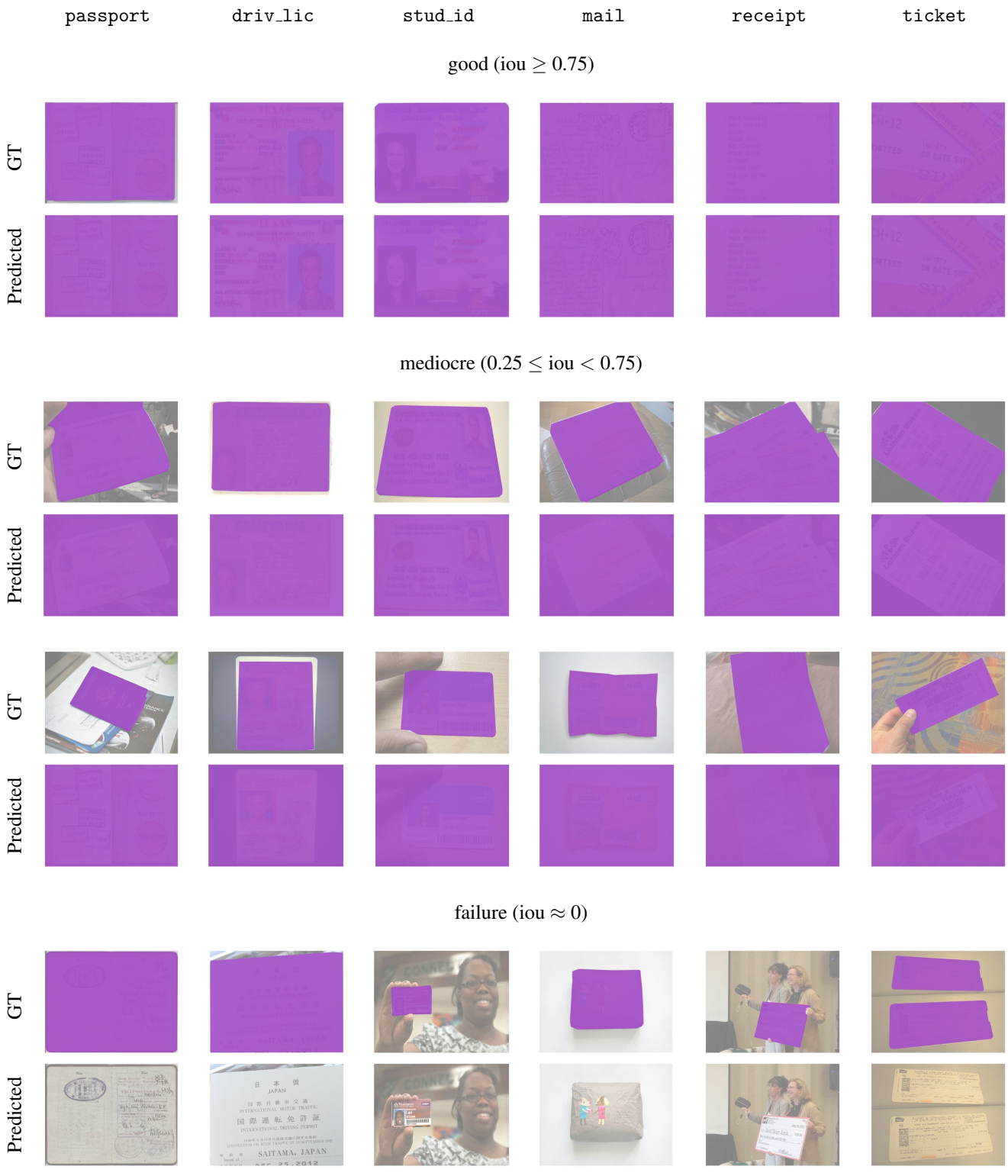


Figure 11: Qualitative results per attribute. In each pair of images, top is ground-truth segmentation and bottom is prediction. Pairs of images in each column are sorted by IoU scores (high to low).

Figure 12: *Common Success Modes of Automatic Redactions*. GT-based are ground-truth regions scaled and redacted as discussed previously. Predicted are automatic redactions generated by method ENSEMBLE. P indicates privacy score and U indicates utility score. In both cases, higher is better. Scores are indicated in green in case of majority agreement and red otherwise.

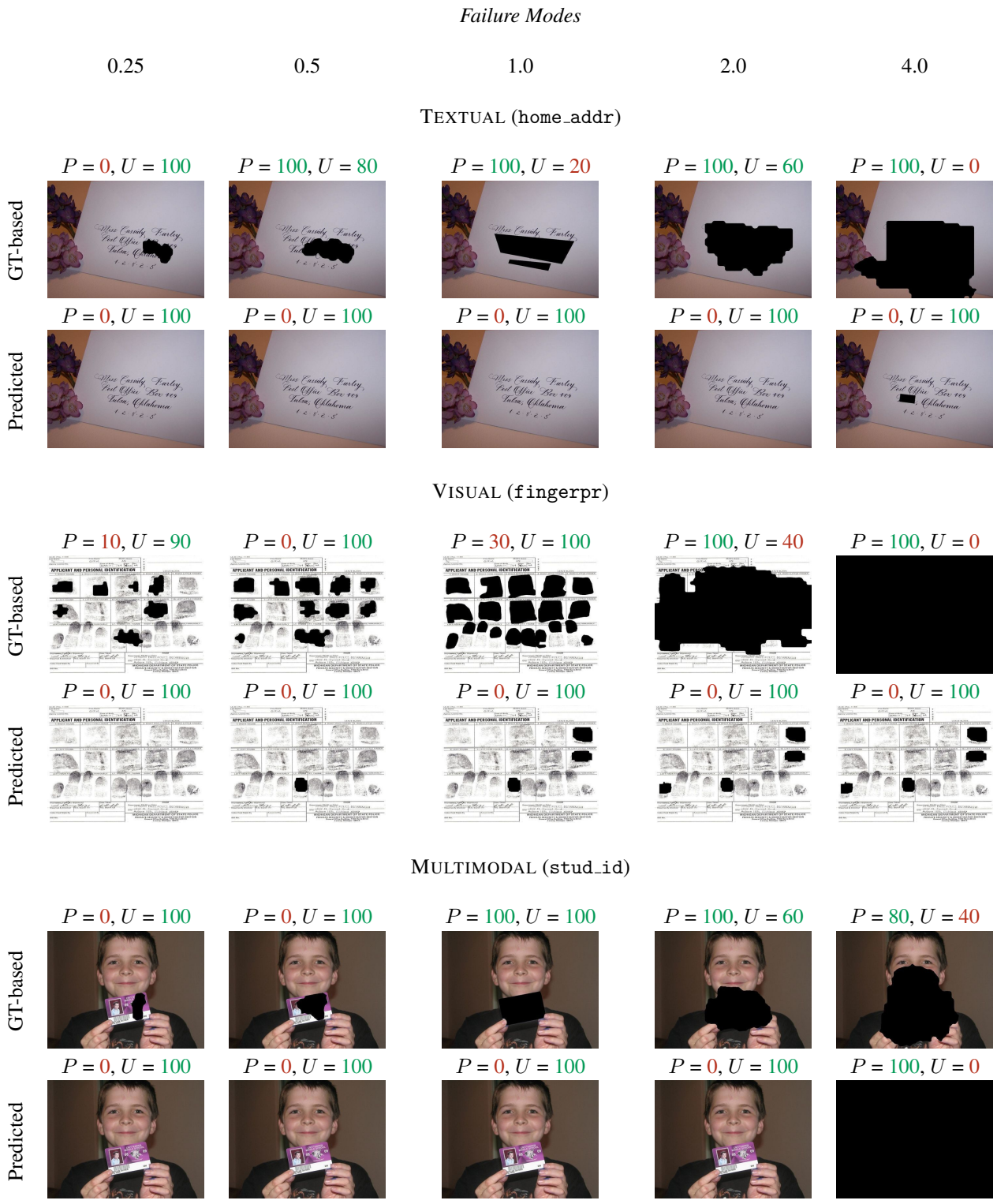


Figure 13: *Common Failure Modes of Automatic Redactions*. GT-based are ground-truth regions scaled and redacted as discussed previously. Predicted are automatic redactions generated by method ENSEMBLE. P indicates privacy score and U indicates utility score. In both cases, higher is better. Scores are indicated in green in case of majority agreement and red otherwise.