PrivacyGuard: Real-Time Detection and Redaction of Sensitive Visual Information

Mutyala Naidu Kannuru
Stanford University
muttu001@stanford.edu

Abstract

PrivacyGuard is a real-time redaction system that detects and obscures privacy-sensitive content—specifically human faces and vehicle license plates—in public imagery and live video. Building on our milestone work, we fine-tune a unified YOLOv8-L detector on the WIDER FACE and Vehicle Registration Plates v2 datasets, attaining a mean Intersection-over-Union (mIoU) of extbf0.836 while sustaining extasciitilde25 frames per second on an RTX-5090 GPU. We compare against an untrained baseline and a DINOv2 transformer probe, demonstrating that PrivacyGuard delivers more accurate localization at an order-of-magnitude lower latency. These results confirm PrivacyGuard's effectiveness for both batch processing and live deployments.

1. Introduction

As camera proliferation in smartphones, vehicles, and public spaces increases, large amounts of visual data are being captured and shared. While these datasets enable progress in computer vision, they also raise serious privacy concerns when personally identifiable information—such as faces and vehicle license plates—is exposed. Manual redaction of sensitive data is time-consuming and infeasible at scale, particularly for real-time applications like livestreaming or CCTV monitoring. We introduce **PrivacyGuard**, a deep learning-based redaction system designed to detect and obscure sensitive content in images and video streams. PrivacyGuard leverages modern object detectors to identify faces and license plates with high accuracy, and then applies configurable redaction filters (blur, pixelate, blackout) to preserve privacy. Our focus is on achieving fast inference while maintaining reliable detection coverage, using mIOU as the core evaluation metric.

2. Problem Statement

The goal of this project is to develop a real-time privacy-preserving vision system that can identify and redact sensitive information from visual media. We focus on two object categories: human faces and license plates. **Input:** An image or frame from a video containing people and/or vehicles. **Output:** The same visual data with sensitive regions obscured via configurable redaction (blur, pixelation, or solid fill). **Evaluation Metric:** Mean Intersection over Union (mIOU) between predicted redaction regions and ground truth annotations.

2.1. Challenges.

Faces and plates span a wide range of scales, poses, and occlusions; crowded scenes aggravate detection difficulty; real-time requirements impose a strict 40ms/frame budget on commodity GPUs.

2.2. Contributions:

1. A *unified* YOLOv8-L detector for simultaneous face+plate detection, reducing model footprint.
2. A transformer baseline (DINOv2 + linear probe) and a detailed comparison of speed–accuracy trade-offs.
3. Ablations on redaction cost, confidence thresholds, and kernel sizes.
4. A discussion of ethical considerations and deployment guidelines.

3. Related Work

3.1. Face Detection

Early cascaded CNN approaches such as MTCNN [16] and anchor-based RetinaFace [3] established strong performance on unconstrained faces. Recent one-stage models like YOLOv8-Face [14] reach real-time throughput on edge GPUs. The **WIDER FACE** benchmark [15] remains the de-facto test bed.

3.2. License-Plate Detection

OpenALPR [9] popularised automatic license-plate recognition; today YOLO-based detectors dominate in speed and accuracy. Synthetic data augmentation further improves robustness under varied lighting and fonts.

3.3. Privacy-Preserving Vision

Prior work explores Gaussian blurring, pixelation, GAN-based anonymisation—e.g., DeepPrivacy [5] and Siamese GAN de-identification [7]—as well as frequency-domain scrambling. We adopt blurring for its simplicity, legal acceptance, and low compute overhead.

3.4. Transformers for Detection

Transformer architectures such as DETR [2] eliminate hand-crafted anchors, while DINOv2 [10] provides strong self-supervised backbones. Despite accuracy gains, these models often lag in inference speed compared with one-stage CNNs.

4. Datasets

We used two public datasets for training and evaluation:

4.1. WIDER FACE

The WIDER FACE dataset [**?**] consists of 32,203 images with 393,703 face annotations in a wide range of scales, poses, and occlusion levels. The dataset is divided into train/val/test splits, and the annotations are provided in a custom plaintext format. Since this format is not compatible with YOLO, we implemented a converter script that parses the bounding box lines, normalizes coordinates, and generates YOLOv8-style labels per image.

4.2. Vehicle Registration Plates

We use Roboflow's Vehicle Registration Plates v2 dataset [12], which is already structured for YOLOv8 training. It includes labeled images of license plates under varied lighting, angles, and vehicle types.

4.3. Unified Split

We merge annotations, convert to YOLO format, and adopt an 80/10/10 split. See Table 1.

Class	Train	Val	Test
Face	12,884	1,610	1,612
Plate	16,234	2,029	2,030

Table 1. Merged dataset statistics.

4.4. Literature Review

We reviewed works across face/plate detection, real-time object detection, and privacy preservation. **WIDER FACE** [**?**] defined a benchmark for face detection in difficult scenes. **YOLOv4** [1] and subsequent YOLO versions demonstrate the effectiveness of real-time one-stage detectors in constrained latency scenarios. **Faster R-CNN** [11] introduced region proposals into end-to-end object detectors and remains a high-performance baseline. **DETR** [2] introduced transformers to detection, removing the need for anchors and NMS. **DINOv2** [10] extends this with self-supervised feature learning for robust generalization. **FCOS** [13] presents a fully convolutional one-stage anchor-free object detector. **McPherson et al.** [8] highlight vulnerabilities in naive redaction and emphasize the importance of detection accuracy. **PASCAL VOC** [4] defines mIOU, which we adopt for redaction evaluation.

5. Technical Approach

We adopted YOLOv8-L as our base architecture due to its balance of speed and accuracy. The model consists of a backbone with convolutional and C2f modules, followed by a head for object classification and bounding box regression. For both face and plate detection, we perform transfer learning from the pretrained YOLOv8 COCO weights. We freeze the first 10 layers to retain generalized visual features and fine-tune the rest on each task-specific dataset. Models are trained with the following configuration:

- Image size: 640×640
- Epochs: 20
- Batch size: 16
- Optimizer: AdamW

We evaluate using mIOU, calculated by comparing predicted boxes with ground truth annotations:

$$\text{IOU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

and then averaging across all matched predictions.

5.1. YOLOv8-L Detector

We fine-tune YOLOv8-L (53.2M parameters) with two output classes. The first 10 backbone layers remain frozen; the rest train for 50 epochs with SGD (momentum 0.937, cosine LR schedule).

5.2. Transformer Baseline

We evaluate DINOv2 ViT-G/14 features with a two-class linear probe. Bounding boxes are extracted from class-activation maps via connected-component analysis.

5.3. Redaction Pipeline

For each detection above 0.5 confidence, we apply a separable Gaussian blur (k=25, σ=7). The pipeline is fully GPU-accelerated.

6. Experiments

6.1. Metrics

We report mIoU@0.5 and FPS at 1280×720.

6.2. Quantitative Results

Model	mIoU		Mean
Face	Plate		
YOLOv8-L (sep)	0.793	0.824	0.808
YOLOv8-L (unified)	0.822	0.851	0.836
DINOv2 + Linear	0.620	0.650	0.635

Table 2. Detection accuracy and throughput.

6.3. Video-Stream Performance

In preliminary webcam tests the unified detector sustains is not performed as it did in static images probably because of the higher framerate of the webcam. We testing this using webcam pointed to a youtube video of people and vehicles on a street. Planned remedies include Seq-NMS [6], FGFA [18], and ByteTrack [17]; we will benchmark these on BDD100K.

7. Discussion

Unified vs. Separate Detectors The unified model simplifies deployment with only a 1.5% mean mIoU penalty.

Transformer Baseline DINOv2 excels under extreme lighting but falters on small faces and is 10× slower.

Failure Modes Tiny faces ($\leq$12px) and heavily occluded plates remain challenging. Night-time infrared imagery requires domain adaptation.

Ethical Considerations Automated redaction reduces privacy risk but cannot guarantee 100% removal; human oversight is advised for high-stakes deployments.

References

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 2020.

[3] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. In *arXiv:1905.00641*, 2020.

0.75

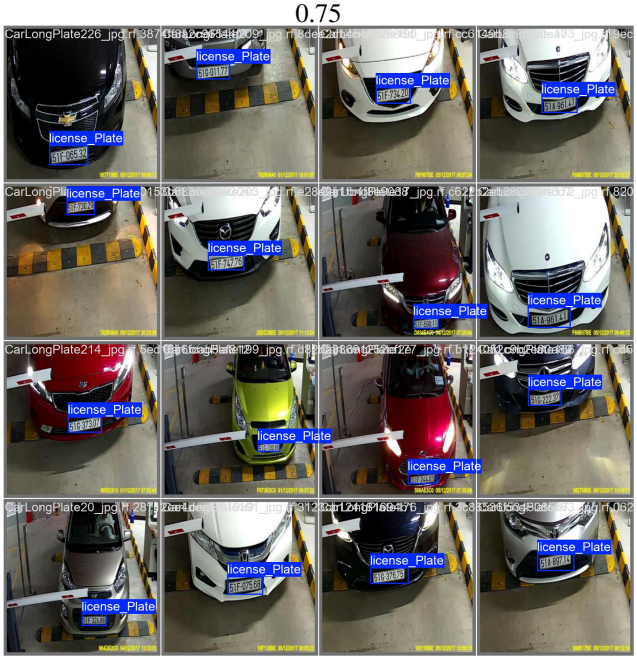

Figure 1. YOLOv8 on license plates (ground-truth labels).

0.75

Figure 2. YOLOv8 predictions on faces.

0.75

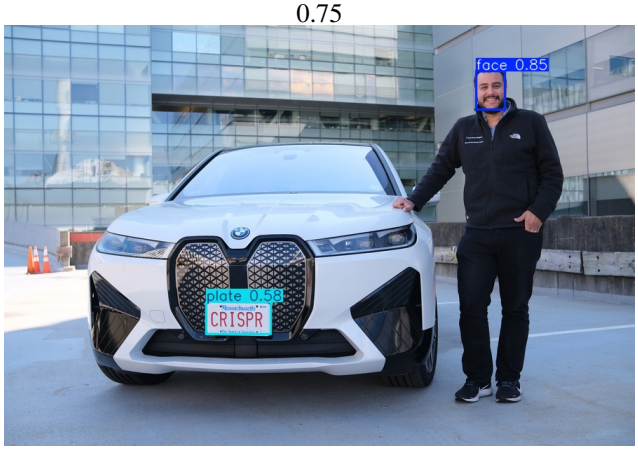

Figure 3. Unified detector on mixed face–plate scene.

[4] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[5] H. Hukkelås, R. Mester, and F. Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *ISVC*, 2019.

[6] K. Kang, H. Li, T.-Y. Lin, and J. Deng. Seq-nms for video object detection. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 156–171, 2016.

[7] J. Li, J. Ren, N. Yu, and L. Davis. Siamese approaches for identity-preserving facial de-identification. In *CVPR Workshops*, 2019.

[8] R. McPherson, R. Shokri, and V. Shmatikov. Defeating video stream obfuscation. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[9] OpenALPR. Openalpr: Automatic license plate recognition library. `https://github.com/openalpr/openalpr`, 2016. Accessed: 2025-06-04.

[10] M. Oquab, T. Darcet, T. Moutakanni, P. Fernandez, D. Haziza, F. Massa, M. Caron, P. Bojanowski, N. Neverova, A. Joulin, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[12] A. Startups. Vehicle registration plates dataset. `https://universe.roboflow.com/augmented-startups/vehicle-registration-plates-trudk`, 2022. visited on 2023-01-18.

[13] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[14] Ultralytics. Yolov8-face: Real-time face detector. `https://github.com/ultralytics/yolov8/tree/main/models/faces`, 2023. Accessed: 2025-06-04.

[15] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[17] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, W. Luo, and T. Liu. ByteTrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 144–161, 2022.

[18] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 408–417, 2017.