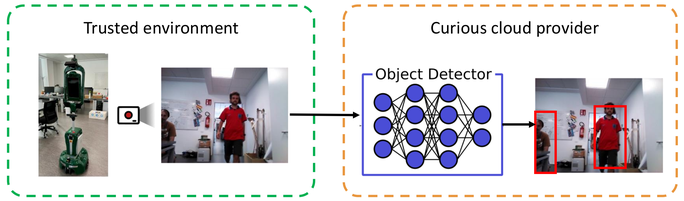
Privacy-Preserving Robotic Perception for Object Detection in Curious Cloud Robotics

Dept. of Computer Science
University of Milan

MICHELE ANTONAZZI micant@kth.se
MATTEO ALBERTI matteo.alberti1@studenti.unimi.it
ALEX BASSOT alex.bassot@unimi.it
MATTEO LUPERTO matteo.luperto@unimi.it
NICOLA BASILICO nicola.basilico@unimi.it

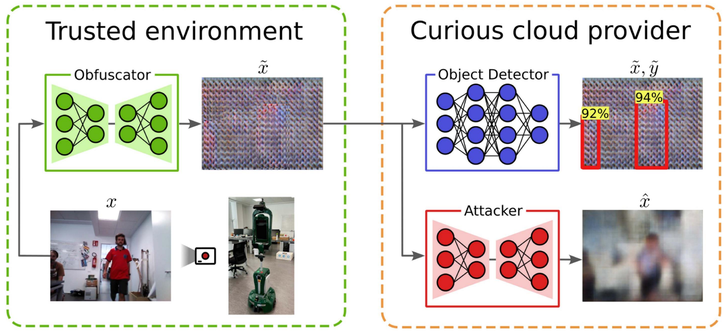
Motivation

- **Low-power service robots** increasingly delocalize computationally heavy deep neural network (DNN) object detection inference to third-party providers
- Service robots operate in human-centric spaces like homes or hospitals, collecting high-resolution images that are **privacy-sensitive**[1]
- Even if images are sent over an encrypted channel, cloud DNN inference needs clear data to operate
- **Can the we trust the provider with our data?**



Curious cloud provider
- We assume the the third-party inference provider is not trusted:
- No guarantees that the user's data will not be used for secondary, non-authorized purposes

Proposed architecture



- Equip the robot with a local and efficient **Encoder-Decoder Obfuscator**
- Transform images into noise-like data such that:
  - ✓ task-specific features are preserved: the cloud-based TaskNet can execute accurate object detection *without being retrained*
  - ✓ prevent adversarial reconstruction: an attacker struggles to reconstruct the original perception from the obfuscated one

Formalization

- **Obfuscation:** mapping image x to a low-dimensional latent space $z = e(x; \theta_e)$ and decoding it to an obfuscated output $\tilde{x} = d(z; \theta_d)$
- **Co-training optimization:** multi-objective loss balancing task accuracy and privacy

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \lambda \mathcal{L}_{task}(Y, \tilde{Y}) + (1 - \lambda) \mathcal{L}^{\gg}_{priv}(x, \tilde{x})$$

- **Attacker:** model Inversion Attack (MIA) network that tries to reconstruct x from $\tilde{x}$, trained with a reconstruction loss on images obfuscated with our method

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \mathcal{L}^{\ll}_{priv}(x, \hat{x})$$

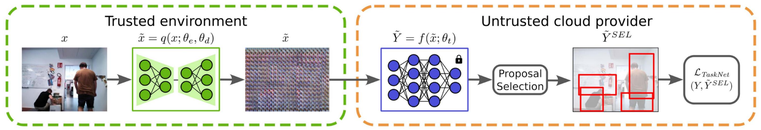
- **TaskNet:** fixed weights trained on a public dataset for object detection

Theoretical analysis

- Prior classification-focused methods[2] achieved privacy by shrinking the bottleneck Z, forcing the network to drop task related features
- **Theorem 1 (Linear Detection-Aware Compression)**
  In proposal-based object detectors, compressing Z below the rank S of the detector's backbone guarantees a non-zero task loss:
  $\Longrightarrow$ **full task performance recovery strictly requires $Z \geq S$**
- **Theorem 2 (Compression vs. Reconstruction)**
  The attacker reconstruction loss decreases as Z grows, shrinking Z provides privacy but destroys detection capabilities:
  $\Longrightarrow$ **strategies solely relying on bottleneck tuning face an unresolvable trade-off**
- **Theorem 3 (Compression for detection and privacy)**
  Zero task loss with a compressed bottleneck ($Z < S$) is possible if and only if the rank of the multi-head stacked backbone matrix is strictly less than S
  $\Longrightarrow$ **restricting gradients to a "weak set" of proposals leads to low-rank feature space**

Proposals selection

- Backpropagation is restricted to a sparse, **dynamically selected subset of proposals**: the obfuscator distills only the minimal features relevant to the task, discarding those that an adversary could exploit for reconstruction



Selection Algorithm: for each ground truth bounding box g select

- **Positive Proposals (P^{SEL}):** top p proposals that achieve the highest Intersection over Union (IoU) with g. These provide gradients to lock onto target localization and class correctness
- **Negative Proposals (N^{SEL}):** filter out positive proposals, enforce an IoU upper bound $\bar{p}$, and pick the top n proposals with the highest TaskNet confidence score. These force the obfuscator to actively suppress false background detections

Algorithm 1: Cotraining with Proposal Selection.

Input:
- $f(\cdot\,; \theta_t)$: a fixed and pretrained object detector
- $q(\cdot\,; \theta_e^0, \theta_d^0)$: an encoder-decoder with random weights
- $\mathcal{D} = \{(x_i, Y_i)\}_{i=1}^{|D|}$: training dataset
- N_{tot}: the number of training rounds
- p, n: the number of positive and negative proposals
- $\bar{p}$: the IoU threshold

Output: the trained parameters θ_e^N, θ_d^N

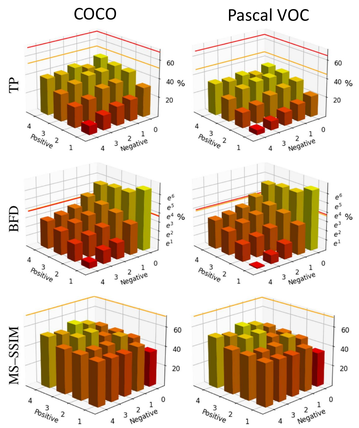
$\theta_e^0, \theta_d^0 \leftarrow RandInit()$
for $\tau \leftarrow 0$ to N_{iter} **do**
  $(x, Y) \sim \mathcal{D}$
  $\tilde{Y} = f(q(x; \theta_e^\tau, \theta_d^\tau); \theta_t)$
  $\tilde{Y}^{\text{SEL}} = \text{SELECTPROPOSALS}(Y, \tilde{Y}, p, n)$
  $\theta_e^{\tau+1}, \theta_d^{\tau+1} \leftarrow \text{BACKPROP}(\mathcal{L}_{\text{TaskNet}}(Y, \tilde{Y}^{\text{SEL}}), \theta_e^\tau, \theta_d^\tau, \theta_t)$
end for
procedure SELECTPROPOSALS(Y_{GT}, Y_t, p, n)
  $S \leftarrow \emptyset$
  for all $g \in Y_{GT}$ **do**
    $P \leftarrow \{t \in Y_t \mid \arg\max_{y \in Y_{GT}} \rho(t, y) = g\}$
    $P^{\text{SEL}} \leftarrow \{t_1 \ldots t_p \mid t_i \in P, \rho(t_i, g) \geq \rho(t_j, g), \forall j \geq i\}$
    $N \leftarrow \{t \in Y_t \setminus P \mid \arg\max_{y \in Y_{GT}} \rho(t, y) = g, \rho(t, g) < \bar{p}\}$
    $N^{\text{SEL}} \leftarrow \{t_1, \ldots, t_n \mid t_i \in N, \sigma(t_i) \geq \sigma(t_j), \forall j \geq i\}$
    $S \leftarrow S \cup P^{\text{SEL}} \cup N^{\text{SEL}}$
  end for
  return S
end procedure

Evaluation

Experimental setup

- People and vehicle detection (highly privacy-critical) using an off-the-shelf Faster R-CNN with a ResNet-50 backbone
- Trained and validated using filtered indoor contexts of the COCO dataset and out-of-distribution validation sets via Pascal VOC 2012
- Detection accuracy measured via AP, AP_{50} true positives (TP) and background false detections (BFD)
- Privacy measured via Multi-Scale Structural Similarity (MS-SSIM) of reconstructions from an optimized MIA attacker using Mean Absolute Error (MAE) and fine-grained Edge-Centric (EC) Sobel loss functions[3]
- Baselines
  - training with all proposals (ALL), as done in [2]
  - TaskNet with plain data (TN)

Datasets validation



TaskNet (TN)

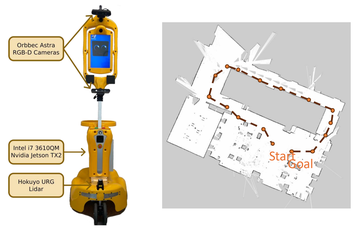
Baseline (ALL)

- ✓ Increasing Positive Proposals improves correct detections
- ✓ More Negative Proposals suppresses background detections
- ✓ Negative and positive proposals should be equal to have a good balance between correct detections and error suppression

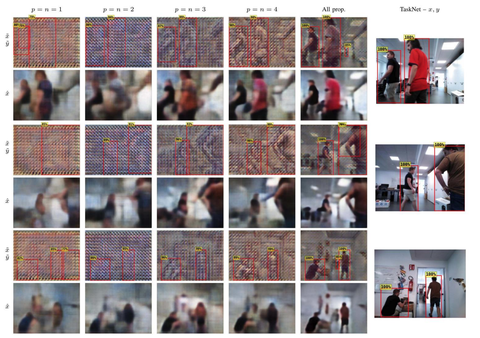
Attacker ablation

- ✓ A more powerful attacker (enhanced reconstruction loss) fails to restore privacy-sensitive features

Real-world experiment



p, n	AP ↑	AP_{50} ↑	TP ↑	BFD ↓	MS ↓
1, 1	37	68	58	8	53
2, 2	56	83	67	5	54
3, 3	57	84	77	10	61
4, 4	59	85	80	11	61
ALL	66	87	86	17	81
TN	78	96	95	19	100



References
[1] Taras et al., "Inherently privacy-preserving vision for trustworthy autonomous systems: Needs and solutions," Journal of Responsible Technology, 2024
[2] Nakanoya et al., "Co-design of communication and machine inference for cloud robotics," Autonomous Robots, 2023
[3] Azizian et al., "Privacy-preserving autoencoder for collaborative object detection," IEEE Transactions on Image Processing, 2024