Beyond Anonymization: Object Scrubbing for Privacy-Preserving 2D and 3D Vision Tasks

Anonymous authors
Paper under double-blind review

Abstract

We introduce **ROAR (Robust Object Removal and Re-annotation)**, a scalable framework for privacy-preserving dataset obfuscation that removes sensitive objects instead of modifying them. Designed for practical deployment, our method integrates instance segmentation with generative inpainting to eliminate identifiable entities while preserving scene integrity. Extensive evaluations on 2D COCO-based object detection show that ROAR achieves 87.5% of baseline average precision (AP), whereas image dropping achieves only 74.2%, highlighting the advantage of scrubbing in preserving dataset utility. In NeRF-based 3D reconstruction, our method incurs a PSNR loss of at most 1.66 dB while maintaining SSIM and improving LPIPS, demonstrating superior perceptual quality. ROAR follows a structured pipeline of detection, inpainting-based removal, re-annotation, and evaluation. We systematically evaluate the privacy-utility trade-off across both 2D and 3D tasks, showing that object removal offers a more effective balance than traditional methods. Our findings establish ROAR as a practical privacy framework, achieving strong guarantees with minimal performance trade-offs. The results highlight challenges in generative inpainting, occlusion-robust segmentation, and task-specific scrubbing, laying the groundwork for real-world privacy-preserving vision systems.

1 Introduction

As machine learning (ML) continues to rely on large-scale data collection, privacy has emerged as a cornerstone of ethical AI development. Privacy, broadly defined, is an individual's right to control access to their personal and sensitive information. Within this scope, data-level privacy which is a subset of privacy concerns aims to safeguard sensitive attributes within datasets while preserving their utility for downstream applications (van der Schaar et al., 2023; Shoshitaishvili et al., 2015).

Existing privacy-preserving techniques in computer vision span a spectrum. At one extreme, raw images provide full utility but no privacy, while at the other, dataset deletion ensures complete privacy but no usability. Intermediate approaches such as noise injection and pixelation offer weak privacy guarantees while retaining usability (Gross et al., 2006; Neustaedter et al., 2006; Neustaedter & Greenberg, 2003; McPherson et al., 2016). More advanced methods leverage generative adversarial networks (GANs) and diffusion models to anonymize identity-revealing features while preserving scene coherence (Hukkelås & Lindseth, 2023; Sun et al., 2018b;a; Maximov et al., 2020; Hukkelås et al., 2019; Malm et al., 2024; Li & Clifton, 2021; Zwick et al., 2024; Barattin et al., 2023). However, regulatory frameworks like the GDPR mandate erasure rather than modification of personal data. Specifically, *Article 17 ("Right to Erasure")* reinforces this obligation, raising legal concerns about synthetic anonymization in sensitive applications (The European Parliament, 2016).

Our work advances privacy-preserving transformations by introducing **ROAR** (**R**obust **O**bject Removal **a**nd **R**e-annotation) framework, a structured framework for dataset obfuscation that eliminates sensitive objects instead of modifying them. ROAR is designed with application robustness in mind and avoids two key challenges: high computational costs and ethical concerns regarding resemblance to real individuals (Carlini et al., 2023). This trade-off is particularly critical in high-risk applications such as surveillance and medical

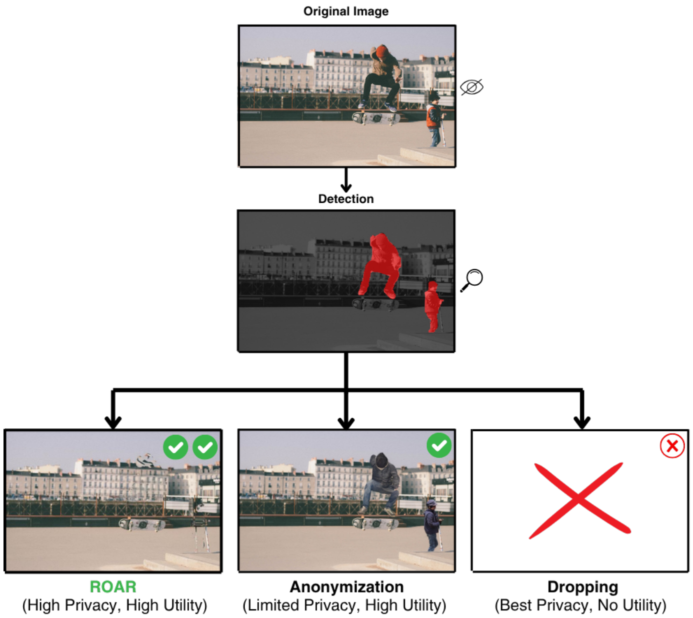

Original Image
Detection
ROAR
(High Privacy, High Utility)
Anonymization
(Limited Privacy, High Utility)
Dropping
(Best Privacy, No Utility)

Figure 1: Privacy-preserving transformations for dataset obfuscation. The input image (top) contains sensitive objects, detected via instance segmentation (middle). Three different obfuscation strategies are applied: (left) **ours** (ROAR), which removes sensitive objects while maintaining scene integrity, (middle) DeepPrivacy2 (Hukkelås & Lindseth, 2023) anonymization, and (right) full data deletion, which ensures maximum privacy at the cost of utility.

imaging, where partial anonymization may still allow for re-identification (Zhu et al., 2024; Malm et al., 2024). By opting for object removal using pre-trained generative models instead of synthetic replacements, ROAR mitigates both concerns: eliminating sensitive entities entirely avoids the computational burden of training dedicated models while also preventing any risk of synthetic identities resembling real individuals. This ensures privacy without introducing new vulnerabilities, preserving the contextual integrity of the scene while maintaining dataset usability.

Recent advances in generative inpainting have also made object removal accessible to the general public. Interactive tools such as Google's Magic Eraser (Google LLC, 2023) or Apple's Clean Up (Apple Inc., 2024) allow users to manually select and erase objects while maintaining perceptual realism, demonstrating how high-quality inpainting has matured into a widely deployed image editing capability. Such systems are designed for single-image use and perceptual quality rather than dataset-scale privacy protection. In contrast, our work operationalizes object removal into an automated, detector-driven framework that ensures consistency, re-annotation, and quantitative privacy–utility evaluation across large-scale datasets. Our key contributions are as follows:

1. We propose **ROAR** (see Fig. 2), a structured privacy-preserving object removal (scrubbing) pipeline that integrates instance segmentation with generative inpainting to eliminate identifiable entities while preserving scene integrity.

2. We systematically evaluate the privacy-utility trade-off by analyzing the impact of object removal on downstream tasks, including object detection, classification, and 3D reconstruction. We compare diffusion-based (Razzhigaev et al., 2023; Rombach et al., 2022) and GAN-based (Zeng et al., 2023; Goodfellow et al., 2020) inpainting methods to assess their effectiveness in maintaining dataset integrity while ensuring privacy.

3. We establish ROAR as a generalizable privacy framework, demonstrating its scalability and robustness as an alternative to conventional anonymization.

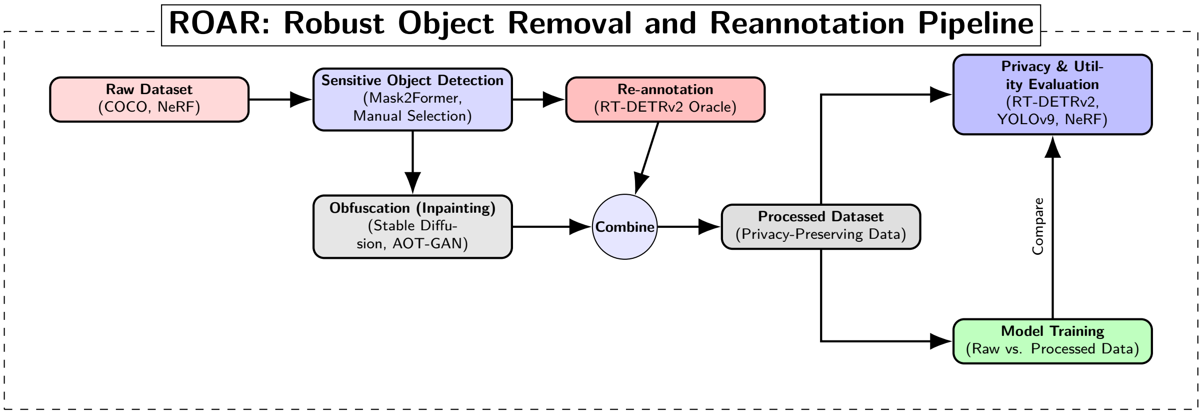

ROAR: Robust Object Removal and Reannotation Pipeline
Raw Dataset (COCO, NeRF)
Sensitive Object Detection (Mask2Former, Manual Selection)
Re-annotation (RT-DETRv2 Oracle)
Privacy & Utility Evaluation (RT-DETRv2, YOLOv9, NeRF)
Obfuscation (Inpainting) (Stable Diffusion, AOT-GAN)
Combine
Processed Dataset (Privacy-Preserving Data)
Compare
Model Training (Raw vs. Processed Data)

Figure 2: Privacy-preserving dataset obfuscation pipeline. **Raw Dataset:** Input data includes COCO (Lin et al., 2014) for 2D detection and NeRF scenes (Mildenhall et al., 2020) for 3D reconstruction. **Sensitive Object Detection:** Instance segmentation (e.g., Mask2Former (Cheng et al., 2022)) identifies sensitive objects in 2D datasets, while NeRF-based datasets require manual selection. **Obfuscation:** Sensitive objects are removed using generative inpainting methods such as diffusion models (e.g., Stable Diffusion (Rombach et al., 2022; Razzhigaev et al., 2023)) and GAN-based models (e.g., AOT-GAN (Zeng et al., 2023)). **Re-annotation:** An oracle model (e.g., RT-DETRv2 (Lv et al., 2024)) updates labels post-obfuscation to maintain dataset integrity. **Processed Dataset:** The resulting dataset ensures privacy while preserving contextual integrity. **Privacy & Utility Evaluation:** Privacy is verified via an oracle, while utility is measured by training object detection (e.g., YOLOv9, RT-DETRv2) and 3D reconstruction (e.g., NeRF) models. **Model Training and Comparison:** Detection models are trained on both raw and obfuscated datasets to assess performance trade-offs.

Notably, our method is the **first** to demonstrate broad applicability across both 2D object detection and 3D Neural Radiance Field (NeRF) (Mildenhall et al., 2020) reconstruction, ensuring its relevance to real-world applications. By eliminating identifiable entities while preserving scene coherence, ROAR provides stronger privacy guarantees than modification-based techniques (see Fig. 3), thereby offering a compelling solution for privacy-preserving AI that complies with regulatory standards while maintaining high data utility. Moreover, as ROAR builds on external segmentation and inpainting models, it naturally benefits from future improvements in this ongoing field of research.

2 Background & Related Work

2.1 Privacy in Computer Vision

With the increasing reliance on large-scale datasets, privacy concerns in computer vision have gained significant attention. Traditional privacy-preserving techniques fall into two broad categories: model-level and data-level approaches. Model-level methods such as federated learning (Kairouz et al., 2021; van der Schaar et al., 2023) and secure multi-party computation (Bonawitz et al., 2019; 2017) prevent direct exposure of data but do not mitigate risks inherent in dataset storage. Data-level approaches, including differential privacy (DP) (Dwork & Roth, 2014; Abadi et al., 2016) and noise-based transformations (Torkzadehmahani et al., 2019; Lee & You, 2024), offer theoretical guarantees but often degrade utility, particularly in high-dimensional vision tasks like object detection and segmentation (Luo et al., 2024; Liu et al., 2021).

A widely adopted strategy for preserving privacy is image anonymization, which involves modifying sensitive attributes to prevent re-identification. Traditional methods such as pixelation and blurring (McPherson et al., 2016) provide weak privacy guarantees, as modern deep learning models can reconstruct obfuscated information (Shokri et al., 2017). More recent approaches rely on generative models to synthesize anonymized images while preserving contextual integrity (Hukkelås & Lindseth, 2023; Sun et al., 2018b;a; Maximov et al., 2020; Hukkelås et al., 2019; Malm et al., 2024; Li & Clifton, 2021; Zwick et al., 2024; Barattin et al., 2023).

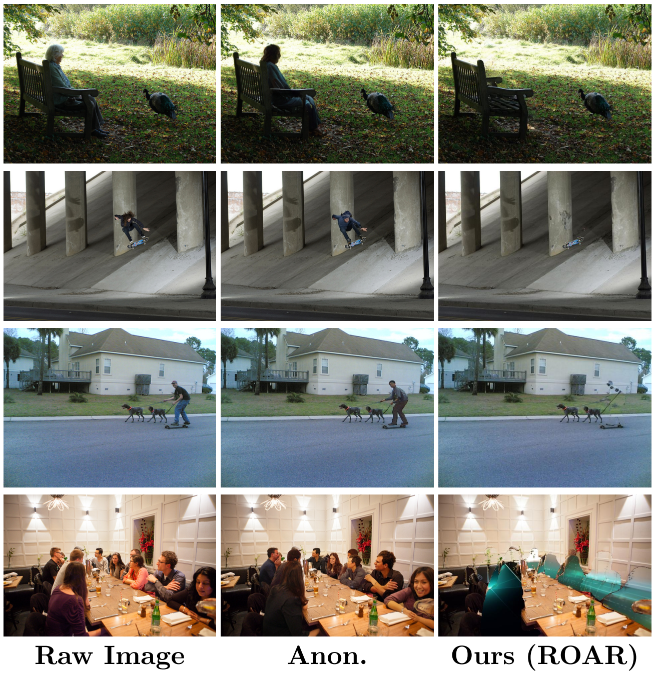


Figure 3: Each row represents a different image processed through raw (left), anonymization (middle)(Hukkelås & Lindseth, 2023), and our approach (right). *First two images are scrubbed with stable diffusion (Rombach et al., 2022), and the last two are scrubbed using Kandinsky (Razzhigaev et al., 2023).*

However, synthetic replacements raise ethical concerns about potential re-identification and resemblance to real individuals (Carlini et al., 2023).

2.2 Generative Models for Privacy-Preserving Obfuscation

Generative models, particularly Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) and diffusion models (Ho et al., 2020; Nichol & Dhariwal, 2021), specifically Latent Diffusion Models (LDMs) (Rombach et al., 2022), have been widely explored for privacy applications. GAN-based methods such as DeepPrivacy2 (Hukkelås et al., 2019; Hukkelås & Lindseth, 2023), and CIAGAN (Maximov et al., 2020) generate anonymized facial images, while approaches like AOT-GAN (Zeng et al., 2023) perform high-resolution inpainting. Despite their effectiveness, GANs suffer from mode collapse and high training costs (Thanh-Tung & Tran, 2020; Durall et al., 2021; Zhang et al., 2018b). Training a dedicated model for synthetic replacements is significantly more expensive than using a pre-trained generic model, making GANs less practical for large-scale obfuscation.

LDMs address some of these limitations by applying iterative denoising in a lower-dimensional latent space, enabling high-quality reconstructions with reduced artifacts (Rombach et al., 2022; Razzhigaev et al., 2023). Stable Diffusion, a prominent LDM-based framework, has been employed for privacy-sensitive applications such as face and full-body anonymization (Malm et al., 2024; He et al., 2024; Zwick et al., 2024).

While LDMs have been explored for anonymization, a related body of work focuses on object removal via generative inpainting. Methods such as LaMa (Suvorov et al., 2022), SmartEraser (Jiang et al., 2025), PowerPaint (Zhuang et al., 2024), and CLIPAway (Ekin et al., 2024) advance inpainting quality through architectural innovations, guidance strategies, and prompt engineering. These approaches are primarily developed for image editing and restoration tasks, and are not evaluated under privacy objectives constraints. Nevertheless, they are complementary to ROAR: their modules can be readily integrated into our framework to improve visual fidelity.

A recent work by Bar et al. (2022) introduced *visual prompting via image inpainting*, demonstrating that modern inpainting models can serve as versatile visual interfaces for downstream tasks such as segmentation or colorization. This line of work, together with large-scale commercial tools such as Google's Magic Eraser (Google LLC, 2023) and Apple's Clean Up (Apple Inc., 2024), shows that high-quality, user-guided object removal is a mature and widely adopted capability. However, these approaches focus on perceptual editing or user interaction rather than automated, privacy-driven dataset obfuscation.

An earlier study by Upenik et al. (2019) explored privacy protection in omnidirectional images using viewport-domain inpainting. While conceptually aligned with our goals, their approach focused on user-selected regions within panoramic imagery and did not include automatic detection, re annotation, or dataset-level evaluation. In contrast, ROAR provides a structured, detector-integrated framework for scalable privacy-preserving dataset obfuscation, explicitly quantifying privacy utility trade-offs across 2D and 3D tasks.

2.3 Privacy-Utility Trade-offs in Object Detection

Privacy-preserving transformations often degrade dataset usability. Object detection and segmentation models like RT-DETRv2 (Zhao et al., 2024; Lv et al., 2024) and YOLOv9 (Wang et al., 2024b; Chang et al., 2023) rely on fine-grained features, making them sensitive to obfuscation. Noise-based anonymization reduces accuracy (Lee & You, 2024), while face-swapping and inpainting may disrupt spatial structures, affecting recognition and generalization (Maximov et al., 2020).

Conversely, these models can aid privacy by detecting sensitive objects for obfuscation. Semantic segmentation allows targeted modifications (Zwick et al., 2024), but models like Mask2Former (Cheng et al., 2022) struggle with occlusions, and may lead to incomplete removals.

2.4 Neural Radiance Fields (NeRF) and Privacy Challenges

Neural Radiance Fields (NeRF) (Mildenhall et al., 2020; Wysocki et al., 2023; Tonderski et al., 2024) enable high-quality 3D scene reconstruction but pose privacy risks by preserving fine-grained details. Recent work has explored adversarial perturbations to disrupt NeRF synthesis (Wu et al., 2023) and privacy-preserving training frameworks that limit information leakage (Zhang et al., 2024), but these methods are computationally expensive and may introduce artifacts.

Benchmarking studies show NeRF is highly sensitive to corruptions like noise and compression (Wang et al., 2024a), suggesting that structured modifications can significantly impact reconstruction. Structured inpainting offers a promising alternative for removing sensitive objects while maintaining scene integrity. To the best of our knowledge, we are the first to apply this approach for privacy-preserving NeRF, ensuring high-quality novel view synthesis from privacy-compliant data. For a detailed discussion on background and related work, please see the appendix C.

3 Methodology

ROAR follows a structured pipeline to achieve privacy-preserving dataset obfuscation using generative models (see Fig. 2) while preserving dataset utility for downstream tasks. The ROAR pipeline consists of four key stages: (1) **Sensitive Object Detection**, where instance segmentation or manual selection identifies sensitive objects; (2) **Object Removal via Generative Inpainting**, which applies diffusion-based or GAN-based models to erase the detected objects; (3) **Oracle-Based Re-annotation**, where an object detection model updates labels post-obfuscation to maintain dataset usability; and (4) **Privacy-Utility Evaluation**, where the effectiveness of obfuscation is assessed through privacy verification and model performance comparisons between raw and processed datasets.

3.1 Stage 1: Sensitive Object Detection

We employ instance segmentation, specifically Mask2Former (Cheng et al., 2022), to detect and localize sensitive objects such as persons. Given an input image $I \in \mathbb{R}^{H \times W \times C}$, where H and W represent the image height and width, and C is the number of color channels, the goal is to extract a set of binary masks M corresponding to detected objects:

$$M = \{m_i \in \{0,1\}^{H \times W} \mid i = 1, \ldots, N\}, \tag{1}$$

where N is the number of sensitive objects, and each m_i denotes a segmentation mask of the same spatial dimensions as the input image. The instance segmentation function S maps an image to a set of masks, class labels, and confidence scores:

$$S(I) = \{(m_i, c_i, s_i) \mid i = 1, \ldots, N\}, \tag{2}$$

where $m_i \in \{0,1\}^{H \times W}$ represents the binary segmentation mask for object i, $c_i \in \mathcal{C}$ is the predicted class label with $\mathcal{C}$ denoting the set of all possible categories, and $s_i \in [0,1]$ is the confidence score assigned to the detection.

3.2 Stage 2: Object Removal via Generative Inpainting

Once sensitive objects are detected and masked, we apply generative inpainting to reconstruct the masked regions. Depending on the inpainting strategy, we utilize either a LDM such as Stable Diffusion (Rombach et al., 2022) or Kandinsky (Razzhigaev et al., 2023), or a GAN-based model such as AOT-GAN (Zeng et al., 2023). The inpainting function G reconstructs a new image I_{obf} by synthesizing content within the masked regions:

$$I_{\text{obf}} = G(I, M, z), \tag{3}$$

where $I \in \mathbb{R}^{H \times W \times C}$ is the original image, $M \in \{0,1\}^{H \times W}$ is the binary mask indicating sensitive object regions, G is the pre-trained inpainting model (Stable Diffusion, Kandinsky, or AOT-GAN), and z is a latent noise variable (typically Gaussian), used in diffusion-based models.

Latent Diffusion Models LDMs (Rombach et al., 2022) operate by performing image synthesis within a compressed latent space rather than directly in pixel space, significantly improving computational efficiency while maintaining high-quality outputs. Both Stable Diffusion and Kandinsky leverage this approach for inpainting, meaning that instead of directly filling missing pixel values, they infer the missing content in a learned latent space conditioned on surrounding structures and optional text prompts (Razzhigaev et al., 2023; Shakhmatov et al., 2022; Rombach et al., 2022). For simplicity and generality, we use a fixed prompt *"generic background"* throughout our experiments to guide the model to replace any removed object with a plausible background. This prompt was chosen based on empirical observation that it yields realistic fillings for a wide range of scenes.

Since we use pre-trained models, the exact denoising steps are handled internally by the model's learned denoising function, which follows the standard denoising objective of latent diffusion models (Rombach et al., 2022):

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{E(x), \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2 \right], \tag{4}$$

where $E(x)$ is defined as the encoder function that maps an image x into a latent representation, z_t represents the noisy latent variable at time step t, ϵ is Gaussian noise, and ϵ_θ is the neural network predicting the noise component. The final output is obtained by decoding the refined latent representation back into the image domain:

$$I_{\text{obf}} = D(z'_M), \tag{5}$$

where D represents the pre-trained decoder that maps latent representations back to pixel space.

GAN-Based Inpainting (AOT-GAN) AOT-GAN follows an adversarial learning framework, where the generator G_{GAN} synthesizes missing content using contextual priors, while the discriminator D_{GAN} provides adversarial feedback to optimize the generator during training through the adversarial loss L^G_{adv}, enforcing perceptual consistency. The inpainting process follows (Zeng et al., 2023):

$$I_{\text{obf}} = G_{\text{GAN}}(I_M, M), \tag{6}$$

where $I_M = I \odot (1 - M)$ is the masked input image, and $\odot$ represents element-wise multiplication. The generator is optimized using the joint loss function (Zeng et al., 2023):

$$L = \lambda_{\text{adv}} L^G_{\text{adv}} + \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{per}} L_{\text{per}} + \lambda_{\text{sty}} L_{\text{sty}}. \tag{7}$$

Here, L^G_{adv} enforces realism through adversarial learning, L_{rec} ensures pixel-wise reconstruction accuracy, L_{per} preserves perceptual features by comparing deep representations, and L_{sty} maintains texture and style consistency using Gram matrices (Zeng et al., 2023).

Obfuscation Process To construct an obfuscated version of the image while preserving non-masked content, we define an obfuscation operator $\mathcal{O}$, which applies the pre-trained inpainting model G only within the masked regions while keeping the unmasked areas unchanged. This formulation ensures that obfuscation is constrained to sensitive regions while preserving background consistency:

$$I_{\text{obf}} = \mathcal{O}(I, M) = I \odot (1 - M) + G(I, M, z) \odot M. \tag{8}$$

For NeRF, we use a *stitching-based inpainting strategy* to maintain view consistency and prevent artifacts. The inpainted region from a reference view is propagated to others using alpha blending, with histogram matching ensuring smooth transitions and soft transitions using Gaussian blurring. Since 2D datasets do not require cross-view consistency, this technique is specific to NeRF. See the appendix B for details on NeRF.

3.3 Stage 3: Oracle-Based Re-annotation and Final Obfuscation Formulation

After object removal, an oracle detector O, specifically RT-DETRv2-L (Lv et al., 2024; Zhao et al., 2024), verifies whether previously existing objects remain in the obfuscated image I_{obf}. Instead of re-annotating all objects, the oracle updates annotations by preserving only those that were affected by the inpainting process, ensuring efficient and minimal modification to the dataset. The oracle function is defined as:

$$O : \mathbb{R}^{H \times W \times C} \times \mathbb{R}^{H \times W} \to \mathcal{P}(\mathbb{R}^4 \times \mathcal{C} \times [0, 1]), \tag{9}$$

where the input consists of the obfuscated image I_{obf} and the binary mask M, and the output is a set of detected objects with bounding boxes $b \in \mathbb{R}^4$, class labels $c \in \mathcal{C}$, and confidence scores $s \in [0, 1]$. Hence, given the original annotation set A, the oracle detects objects in I_{obf}, producing:

$$A = \{(b_i, c_i) \mid i = 1, \ldots, K\}, \tag{10}$$

$$A_{\text{oracle}} = O(I_{\text{obf}}, M) = \{(b'_j, c'_j, s'_j) \mid j = 1, \ldots, L\}. \tag{11}$$

The verification step is applied only to objects that had significant spatial overlap with the removed sensitive objects, while all other objects are automatically retained in the final annotation set. Specifically, let A_{collided} be the subset of annotations where the bounding boxes overlap with the masked region M:

$$A_{\text{collided}} = \{(b_i, c_i) \mid \text{IoU}(b_i, M) > \zeta\}, \tag{12}$$

where ζ is a predefined threshold for determining collision. The verification process is restricted to this subset, ensuring that only potentially altered objects are checked. The updated annotation set $\hat{A}$ is then obtained as:

$$A_{\text{verified}} = \{(b_i, c_i) \mid (b_i, c_i) \in A_{\text{collided}}, \tag{13}$$

$$\exists (b'_j, c'_j, s'_j) \in A_{\text{oracle}}, \tag{14}$$

$$\text{IoU}(b_i, b'_j) > \tau\}. \tag{15}$$

$$\hat{A} = (A \setminus A_{\text{collided}}) \cup A_{\text{verified}}. \tag{16}$$

Here, the first term $A \setminus A_{\text{collided}}$ retains all objects that were not affected by inpainting, and the second term A_{verified} reinstates only those that the oracle detects as still present after inpainting. Since verification is restricted to A_{collided}, this ensures that objects that never intersected with masked regions are never subjected to verification and are automatically retained.

Thus, the final dataset consists of I_{obf} with updated annotations $\hat{A}$, ensuring that sensitive objects are removed while preserving dataset utility with minimal disruption to unaffected objects.

3.4 Stage 4: Privacy-Utility Evaluation

In this work, we use the term *privacy guarantees* to denote data-level protection achieved through complete removal of identifiable entities from a dataset. Formally, a dataset $\mathcal{D}$ satisfies privacy under ROAR if all instances of sensitive categories (e.g., persons) are eliminated such that the probability of re-identification by any downstream model or human observer is negligible given the obfuscated dataset $\mathcal{D}'$. Unlike differential privacy (Abadi et al., 2016), which bounds information leakage theoretically, our guarantees are operational: they are empirically verified by an oracle detector and quantified through detection-based validation of object removal in the obfuscated outputs, as detailed later in Section 4.2.

To assess the effectiveness of our obfuscation pipeline, we evaluate both privacy and utility aspects of the obfuscated datasets. Specifically, we train object detection models on the obfuscated dataset and compare their performance against models trained on the original dataset.

For the COCO dataset, we benchmark on two state-of-the-art object detection models: RT-DETRv2-M (Lv et al., 2024; Zhao et al., 2024) (RTD) and YOLOv9 (Wang et al., 2024b; Chang et al., 2023) (YOLO). These models are trained from scratch on the obfuscated COCO dataset and benchmarked against their counterparts trained on the original dataset. By analyzing detection performance, we assess the impact of object removal on downstream vision tasks and quantify the trade-off between privacy preservation and dataset utility.

To contextualize ROAR's performance, we compare it against two baselines: *image dropping* and *blackout*. The **image dropping** baseline removes any image containing a sensitive object, simulating a strict privacy-first approach at the cost of substantial data loss. While, the **blackout** baseline replaces detected object bounding boxes with black rectangles. To avoid shape and pose leakage, we rely on object detection instead of segmentation masks. These should not be confused with the original utility baseline (the *None* setting in Table 1), which represents training on the unmodified dataset and serves as the upper bound for utility without any privacy intervention.

NeRF-Based Evaluation In addition to COCO, we evaluate our method in a NeRF-based multi-view 3D reconstruction setting (Mildenhall et al., 2020). Unlike COCO, where sensitive objects are defined by class labels such as persons or vehicles, objects in NeRF experiments are manually selected per scene rather than detected through instance segmentation. As a result, NeRF experiments omit the sensitive object detection and re-annotation stages, focusing solely on inpainting.

To evaluate the impact of object removal on NeRF reconstruction, we scrubbed selected sensitive objects, then trained and tested NeRF models using a 90-10 data split. We assessed structural consistency, visual artifacts, and completeness in 3D scenes to analyze how scrubbing affects view-consistent synthesis. See the appendix B for NeRF pipeline details.

4 Experimental Results

This section presents the empirical findings of ROAR, analyzing their effectiveness in maintaining privacy while preserving dataset utility. We evaluate three generative inpainting techniques: Kandinsky2.2 (KD) (Razzhigaev et al., 2023; Shakhmatov et al., 2022), Stable Diffusion (SD) (Rombach et al., 2022), and AOT-GAN (AOT) (Zeng et al., 2023), assessing their impact on both object detection and 3D reconstruction tasks. Specifically, we measure performance on object detection models, namely YOLOv9 (Wang et al., 2024b) (YOLO) and RT-DETRv2-M (Lv et al., 2024; Zhao et al., 2024) (RTD), as well as on NeRF (Mildenhall et al., 2020) for evaluating scene reconstruction quality using the open-source implementation on PyTorch (Yen-Chen, 2020). The results are contextualized through a discussion of their implications on real-world applications and future privacy-preserving dataset processing.

4.1 Dataset Overview: COCO and NeRF

COCO We use the COCO 2017 training dataset (Lin et al., 2014), a large-scale benchmark for object detection, segmentation, and keypoint detection. It contains 118,287 training images and 2.5 million object

Table 1: **Privacy and Utility Metrics.**

Setting[*]	Method[†]	BD[‡]	RTD$_{AP}$[§] ↑	YOLO$_{AP}$[§] ↑	PE (%)[¶] ↑	IE (%)[¶] ↑	Img. Lost[‖] ↓	Annot. Red.[**] ↓
None	None	N	0.480	0.514	0	0	0 (0%)	0 (0%)
Full	Kandinsky	N	0.420	0.434	79.82	64.77	8799 (7.45%)	107474 (17.99%)
Full	Kandinsky	Y	0.410	0.426	89.22	77.73	12761 (10.78%)	131599 (22.02%)
Full	Stable Diff.	N	**0.441**	**0.460**	59.19	38.16	7482 (6.32%)	98113 (16.42%)
Full	Stable Diff.	Y	0.428	0.448	72.15	59.07	11137 (9.41%)	118217 (19.78%)
Full	AOT-GAN	N	0.424	0.437	63.26	40.82	7239 (6.11%)	99996 (16.73%)
Baseline[††]	Blackout	—	0.399	0.413	94.30	91.88	17608 (14.88%)	150837 (25.24%)
Baseline[††]	Drop	—	0.356	0.367	100.00	100.00	64115 (54.20%)	347201 (58.11%)
Selective	Kandinsky	N	**0.466**	0.481	82.49	—	2875 (2.43%)	41472 (6.94%)
Selective	Stable Diff.	N	0.465	**0.482**	67.45	—	2714 (2.29%)	40634 (6.80%)
Selective	AOT-GAN	N	0.464	0.482	67.91	—	2512 (2.12%)	40374 (6.76%)
Baseline[††]	Blackout	—	0.463	0.482	95.42	—	5339 (4.51%)	51490 (8.62%)
Baseline[††]	Drop	—	0.456	0.469	100.00	—	32057 (27.10%)	173827 (29.09%)

[*] Setting: *Full* (FP) removes all sensitive objects; *Selective* (SP) removes one randomly chosen object per image in 50% of the sensitive subset. *None* denotes the original unmodified dataset. Rows above the mid-rule are FP; rows below are SP.

[†] Method: the inpainting/obfuscation technique used (e.g., Stable Diffusion, Kandinsky, AOT-GAN).

[‡] BD (boundary expansion): whether the segmentation mask was enlarged beyond detected object regions (Y=yes, N=no).

[§] RTD$_{AP}$ and YOLO$_{AP}$: COCO average precision (AP@[0.50:0.95]) from detectors RT-DETRv2 and YOLOv9, respectively.

[¶] PE (%) and IE (%): person-level and image-level removal efficiencies, measured via an oracle detector after scrubbing.

[‖] Img. Lost: number and percentage of images with no retained annotations post-scrubbing.

[**] Annot. Red.: number and percentage of removed non-person object annotations.

[††] Baselines: *Blackout* replaces sensitive-object boxes with black rectangles; *Drop* removes all images containing sensitive objects.

instances across 91 categories, including 262,465 person annotations. Persons appear in 54.2% of images, averaging 4.09 persons per image. This diversity and occlusion make COCO an ideal test bed for privacy-preserving obfuscation. We follow standard protocols using the 5,000-image 2017 validation set.

NeRF To assess multi-view 3D reconstruction, we evaluate on three NeRF-based scenes: Fern, Flower, and Room, comprising 20, 34, and 41 images, respectively (Mildenhall et al., 2020). Each scene involves multi-view images centered around a static scene with a central object. Since NeRF reconstructs scenes from 2D views, we analyze how inpainting-based obfuscation affects its reconstruction quality.

4.2 Privacy and Utility Metrics for Evaluation

To systematically assess the trade-off between privacy preservation and dataset utility, we employ a set of evaluation metrics spanning both privacy effectiveness and object detection performance.

We evaluate privacy effectiveness using Person-level Removal Efficiency (PE, %) and Image-level Removal Efficiency (IE, %) to ensure a precise evaluation of privacy effectiveness across different obfuscation strategies. For Full Privacy (FP) setting, where all sensitive objects are scrubbed, PE is defined as:

$$\mathrm{PE}_{\mathrm{FP}} = \frac{\sum_{i=1}^{N} (P_i^{\mathrm{GT}} - P_i^{\mathrm{Scrubbed}})}{\sum_{i=1}^{N} P_i^{\mathrm{GT}}} \times 100, \tag{17}$$

where P_i^{GT} is the number of persons in image i from the ground truth annotations, and P_i^{Scrubbed} is the number of remaining persons after scrubbing calculated using the oracle model. This metric captures the overall reduction in persons across the dataset.

For Selective Privacy (SP) setting, where only one person per image is targeted for removal in half of the sensitive part of the dataset (1 in 50%), the exact identity of the removed person is unknown. Instead, we compute PE based on whether the total person count in an image decreases:

$$\mathrm{PE}_{\mathrm{SP}} = \frac{\sum_{i=1}^{N} \mathbb{1}(P_i^{\mathrm{GT}} > P_i^{\mathrm{Scrubbed}})}{N} \times 100, \tag{18}$$

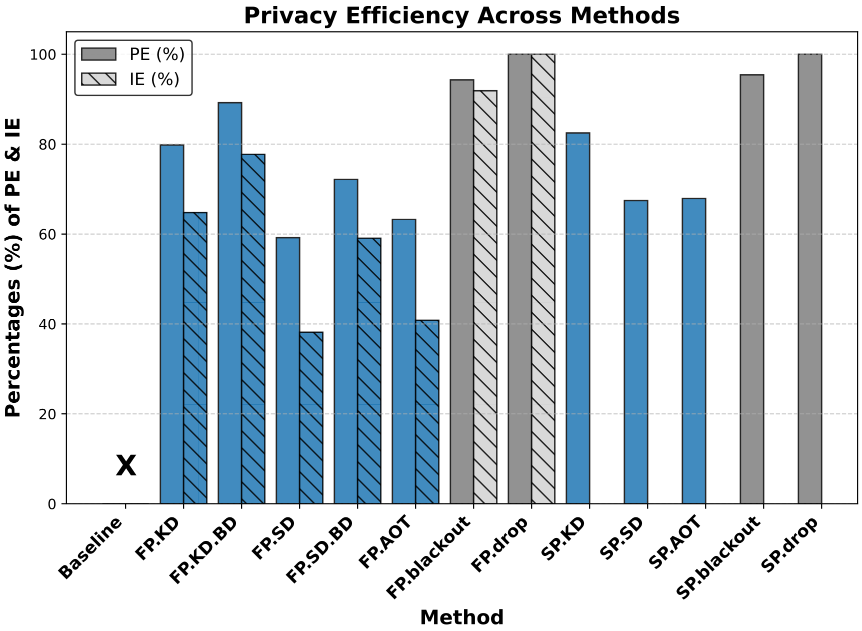


Figure 4: **Privacy Efficiency** The bar plots show Person Removal Efficiency (PE%) and Image Removal Efficiency (IE%). **KD** (Kandinsky), **SD** (Stable Diffusion) and **AOT** (AOT-GAN) denote different inpainting methods. **BD** (Boundary) refers to boundary expansion, while **Drop** removes sensitive images and **Blackout** blackouts the sensitive object. **FP** denotes the *Full* setting and **SP** denotes the *Selective* setting. Blue bars correspond to ROAR-based methods, whereas other methods serve as baselines.

where N is the number of images and $\mathbb{1}(\cdot)$ is an indicator function that equals 1 if at least one person has been successfully removed from the image, and 0 otherwise. This reflects whether scrubbing was effective in each image without requiring knowledge of which specific person was removed.

The Image-level Removal Efficiency (IE, %) is computed as:

$$\text{IE} = \frac{\sum_{i=1}^{N} \mathbb{1}(P_i^{\text{Scrubbed}} = 0)}{N} \times 100, \tag{19}$$

where $\mathbb{1}(P_i^{\text{Scrubbed}} = 0)$ indicates whether an image has been completely cleared of all persons. This metric quantifies the proportion of images where every sensitive object has been successfully scrubbed. Note that, IE is not applicable to the SP approach as we do not want to scrub all persons in an image but one.

For utility assessment, we report Average Precision (AP) as the primary detection performance metric, following the COCO evaluation protocol. AP is computed as the mean of precision-recall scores across Intersection over Union (IoU) thresholds from 0.50 to 0.95 in steps of 0.05, denoted as AP@[IoU=0.50:0.95]. Higher AP values indicate better object detection accuracy. *All AP results are reported on the validation dataset, person label is excluded.* For 3D scene reconstruction, we report standard reconstruction metrics, PSNR, SSIM, and LPIPS (Horé & Ziou, 2010; Zhang et al., 2018a) to evaluate fidelity and perceptual quality (see Appendix B for definitions).

Additionally, we track Image Loss (%), reflecting the percentage of images removed due to privacy constraints, and Annotation Reduction (%), indicating the proportion of object annotations removed, excluding person annotations.

4.3 Privacy-Utility Analysis

Kandinsky scrubbing yields the best privacy-utility trade-off. Across all privacy configurations and models, *Kandinsky with boundary expansion (FP.KD.BD)* consistently delivers the strongest privacy guarantee (*PE = 89.22%*) while retaining *85.4%* of the original model utility (RTD_{AP} *= 0.410 vs. 0.480 baseline*), as shown in Table 1 and Fig. 4. This substantially outperforms the image dropping, which guarantees privacy (*PE = 100%*) but severely degrades detection performance to RTD_{AP} *= 0.356*, corresponding to only *74.2%* of the baseline.

Our method outperforms deletion and masking baselines. Image-level scrubbing with generative inpainting retains more contextual information, yielding better downstream performance than both image dropping, which discards more than half of the data, and blackout, which introduces high-contrast artifacts that disrupt model learning (RTD_{AP} *= 0.399*). In contrast, our structured scrubbing achieves RTD_{AP} = *0.420* (FP.KD), preserving semantics and visual continuity.

Our experiments yield additional insights:
Selective Privacy supports real-world deployment. In SP.KD, where only one person is removed from 50% of sensitive images, detection performance remains high (RTD_{AP} *= 0.466*) while still removing sensitive content in a controlled manner (*PE = 82.49%*). This setting reflects scenarios such as opt-in removals or regulatory requirements targeting partial scrubbing.

Boundary expansion improves privacy at the cost of utility. Enlarging masks by 10 pixels enhances privacy (*PE = 89.22%* vs. *79.82%* without expansion) but reduces accuracy (RTD_{AP} *= 0.410* vs. *0.420*). This trade-off helps mitigate contextual leakage beyond object edges.

Image loss quantifies oracle-driven reannotation limits. Scrubbing may lead to complete loss of annotations in crowded scenes, reflected by *Images Lost = 10.78%* for FP.KD.BD. This is a conservative privacy and utility safeguard: if no remaining annotation can be verified, the image is discarded.

Structured inpainting is especially critical in sensitive datasets. In COCO, where over 50% of images contain people, deletion leads to losing more than half of the dataset. Scrubbing avoids such drastic pruning while still achieving over 60% PE, thus ensuring sufficient data coverage for tasks like federated learning, where data scarcity may otherwise arise.

Failure analysis reveals key bottlenecks. Failures primarily arise from two sources: (i) *Segmentation failure*: Mask2Former (Cheng et al., 2022) may miss small, occluded, or ambiguously shaped objects, particularly in crowded scenes; and (ii) *Inpainting failure*: diffusion models can hallucinate human-like content within masked regions, especially when masks are imprecise. In both cases, blackout would technically avoid hallucination but at the cost of disrupting context and degrading performance. See Appendix E for a comprehensive discussion.

Table 2: **Comparison of PSNR, SSIM, and LPIPS across different scrubbing methods.**

Scene	Method	PSNR ↑*	SSIM ↑*	LPIPS ↓*
Fern	Baseline	25.17	0.79	0.28
	Ours/AOT	25.59	0.79	0.24
	Ours/SD	25.67	0.79	0.24
	Ours/KD	**26.49**	**0.84**	**0.15**
Flower	Baseline	27.40	0.83	0.22
	Ours/AOT	26.27	0.81	0.18
	Ours/SD	26.17	0.80	0.18
	Ours/KD	**27.64**	**0.87**	**0.10**
	Dropped	17.80	0.58	0.33
Room	Baseline	**32.70**	0.95	0.18
	Ours/AOT	29.81	0.94	0.13
	Ours/SD	29.66	0.93	0.13
	Ours/KD	31.04	**0.96**	**0.07**

*↑ indicates higher is better; ↓ indicates lower is better.

4.4 Effects on 3D Scene Reconstruction

We evaluate the impact of privacy-preserving transformations on NeRF-based 3D reconstruction, analyzing how different inpainting methods affect scene fidelity. Despite challenges introduced by object removal, our results confirm that structured generative transformations maintain high-quality synthesis, as illustrated in the supplementary material.

As shown in Table 2, diffusion-based inpainting methods outperform GAN-based approaches in preserving scene integrity post-removal. Compared to the *baseline* where NeRF is trained on unmodified input images, latent diffusion models such as Kandinsky achieve the highest PSNR and SSIM, while also minimizing perceptual discrepancies (lower LPIPS). These improvements suggest that LDMs reconstruct missing regions with high geometric and textural fidelity, ensuring minimal degradation in view synthesis.

In contrast, GAN-based inpainting introduces artifacts and structural inconsistencies, leading to lower PSNR and SSIM scores. While SD yields strong results, further refinements in latent-space conditioning, such as Kandinsky (Razzhigaev et al., 2023), enhance reconstruction quality. These findings reinforce the effectiveness of structured generative transformations for privacy-preserving NeRF, ensuring compliance with privacy constraints while preserving downstream utility.

In the Flower scene, an insect appeared in 29 out of 34 images, making image dropping a potential, albeit extreme, privacy-preserving option. Removing these images and training NeRF with only five remaining views severely degraded reconstruction (see Table 2). This confirms that while dropping ensures privacy, it significantly harms scene fidelity, reinforcing scrubbing as the preferred approach.

4.5 Discussion and Future Directions

Our findings show that ROAR enables effective privacy-preserving dataset obfuscation, removing sensitive objects with minimal degradation in 2D detection and 3D reconstruction. Yet, several challenges remain:

Segmentation and Inpainting Limitations. Segmentation errors, especially in crowded or occluded scenes, can leave residual traces or remove essential context. Techniques like confidence-based mask refinement or ensemble models may improve robustness. Inpainting models also struggle with large or irregular masks, sometimes hallucinating human-like features (Aithal et al., 2024; Rombach et al., 2022; Borji, 2023). Refining these models for privacy-driven object removal, using negative prompts (e.g., "no human") or rejection sampling, is a promising direction.

Task Generalization. Tasks like segmentation or activity recognition may be more sensitive to object removal. Supporting them requires adaptive scrubbing that preserves non-sensitive context. Optimizing scrubbing for NeRF-based reconstruction and standardizing evaluation benchmarks will also be key to broader adoption.

5 Conclusion

We presented **ROAR (Robust Object Removal and Re-annotation)**, a privacy-preserving dataset obfuscation framework that removes sensitive objects while preserving scene integrity. Our results demonstrate that for 2D COCO-based object detection, ROAR achieves 87.5% of the baseline average precision (AP), whereas image dropping achieves 74.5% of the baseline, highlighting the advantage of scrubbing in maintaining dataset utility. In NeRF-based 3D reconstruction, scrubbing incurs a PSNR loss of at most 1.66 dB while maintaining SSIM and improving LPIPS.

These findings establish ROAR as a scalable and practical solution for privacy-preserving vision systems, offering strong privacy guarantees with minimal utility loss. Its modular design ensures compatibility with evolving segmentation and inpainting models, making it well-suited for real-world applications.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

Sumukh K. Aithal, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/f29369d192b13184b65c6d2515474d78-Abstract-Conference.html.

Apple Inc. Apple intelligence is available today on iphone, ipad, and mac, 2024. URL https://www.apple.com/newsroom/2024/10/apple-intelligence-is-available-today-on-iphone-ipad-and-mac/. Accessed: October 2025.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. Visual prompting via image inpainting. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9f09f316a3eaf59d9ced5ffaefe97e0f-Abstract-Conference.html.

Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 8001–8010. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00773. URL https://doi.org/10.1109/CVPR52729.2023.00773.

Kallista A. Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In Bhavani Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu (eds.), *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*, pp. 1175–1191. ACM, 2017. doi: 10.1145/3133956.3133982. URL https://doi.org/10.1145/3133956.3133982.

Kallista A. Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. In Ameet Talwalkar, Virginia Smith, and Matei Zaharia (eds.), *Proceedings of the Second Conference on Machine Learning and Systems, SysML 2019, Stanford, CA, USA, March 31 - April 2, 2019*. mlsys.org, 2019. URL https://proceedings.mlsys.org/paper_files/paper/2019/hash/7b770da633baf74895be22a8807f1a8f-Abstract.html.

Ali Borji. Qualitative failures of image generation models and their application in detecting deepfakes. *Image Vis. Comput.*, 137:104771, 2023. doi: 10.1016/J.IMAVIS.2023.104771. URL https://doi.org/10.1016/j.imavis.2023.104771.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In Joseph A. Calandrino and Carmela Troncoso (eds.), *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pp. 5253–5270. USENIX Association, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Hung-Shuo Chang, Chien-Yao Wang, Richard Robert Wang, Gene Chou, and Hong-Yuan Mark Liao. Yolor-based multi-task learning. *CoRR*, abs/2309.16921, 2023. doi: 10.48550/ARXIV.2309.16921. URL https://doi.org/10.48550/arXiv.2309.16921.

Dingfan Chen, Tribhuvanesh Orekondy, and Mario Fritz. GS-WGAN: A gradient-sanitized approach for learning differentially private generators. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/9547ad6b651e2087bac67651aa92cd0d-Abstract.html.

Dingfan Chen, Raouf Kerkouche, and Mario Fritz. Private set generation with discriminative information. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022a. URL http://papers.nips.cc/paper_files/paper/2022/hash/5e1a87dbb7e954b8d9d6c91f6db771eb-Abstract-Conference.html.

Jia-Wei Chen, Chia-Mu Yu, Ching-Chia Kao, Tzai-Wei Pang, and Chun-Shien Lu. DPGEN: differentially private generative energy-guided network for natural image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8377–8386. IEEE, 2022b. doi: 10.1109/CVPR52688.2022.00820. URL https://doi.org/10.1109/CVPR52688.2022.00820.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 1280–1289. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00135. URL https://doi.org/10.1109/CVPR52688.2022.00135.

Ricard Durall, Avraam Chatzimichailidis, Peter Labus, and Janis Keuper. Combating mode collapse in GAN training: An empirical analysis using hessian eigenvalues. In Giovanni Maria Farinella, Petia Radeva, José Braz, and Kadi Bouatouch (eds.), *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2021, Volume 4: VISAPP, Online Streaming, February 8-10, 2021*, pp. 211–218. SCITEPRESS, 2021. doi: 10.5220/0010167902110218. URL https://doi.org/10.5220/0010167902110218.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Yigit Ekin, Ahmet Burak Yildirim, Erdem Eren Caglar, Aykut Erdem, Erkut Erdem, and Aysegul Dundar. Clipaway: Harmonizing focused embeddings for removing objects via diffusion models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/1f6f0b6eec8a4ff0f6baa707ff91a442-Abstract-Conference.html.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020. doi: 10.1145/3422622. URL https://doi.org/10.1145/3422622.

Google LLC. Google photos – editing tools, 2023. URL https://www.google.com/photos/editing/. Accessed: October 2025.

Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. Model-based face de-identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2006, New York, NY, USA, 17-22 June, 2006*, pp. 161. IEEE Computer Society, 2006. doi: 10.1109/CVPRW.2006.125. URL https://doi.org/10.1109/CVPRW.2006.125.

Xiao He, Mingrui Zhu, Dongxin Chen, Nannan Wang, and Xinbo Gao. Diff-privacy: Diffusion-based face privacy protection. *IEEE Trans. Circuits Syst. Video Technol.*, 34(12):13164–13176, 2024. doi: 10.1109/TCSVT.2024.3449290. URL https://doi.org/10.1109/TCSVT.2024.3449290.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*, pp. 2366–2369. IEEE Computer Society, 2010. doi: 10.1109/ICPR.2010.579. URL https://doi.org/10.1109/ICPR.2010.579.

Håkon Hukkelås and Frank Lindseth. Deepprivacy2: Towards realistic full-body anonymization. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pp. 1329–1338. IEEE, 2023. doi: 10.1109/WACV56688.2023.00138. URL https://doi.org/10.1109/WACV56688.2023.00138.

Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang, and Panpan Xu (eds.), *Advances in Visual Computing - 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7-9, 2019, Proceedings, Part I*, volume 11844 of *Lecture Notes in Computer Science*, pp. 565–578. Springer, 2019. doi: 10.1007/978-3-030-33720-9_44. URL `https://doi.org/10.1007/978-3-030-33720-9_44`.

Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24452–24462. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02277. URL `https://openaccess.thecvf.com/content/CVPR2025/html/Jiang_SmartEraser_Remove_Anything_from_Images_using_Masked-Region_Guidance_CVPR_2025_paper.html`.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021. doi: 10.1561/2200000083. URL `https://doi.org/10.1561/2200000083`.

Jun Ha Lee and Su Jeong You. Balancing privacy and accuracy: Exploring the impact of data anonymization on deep learning models in computer vision. *IEEE Access*, 12:8346–8358, 2024. doi: 10.1109/ACCESS.2024.3352146. URL `https://doi.org/10.1109/ACCESS.2024.3352146`.

Qiushi Li, Yan Zhang, Ju Ren, Qi Li, and Yaoxue Zhang. You can use but cannot recognize: Preserving visual privacy in deep neural networks. In *31st Annual Network and Distributed System Security Symposium, NDSS 2024, San Diego, California, USA, February 26 - March 1, 2024*. The Internet Society, 2024. URL `https://www.ndss-symposium.org/ndss-paper/you-can-use-but-cannot-recognize-preserving-visual-privacy-in-deep-neural-networks/`.

Tao Li and Chris Clifton. Differentially private imaging via latent space manipulation. *CoRR*, abs/2103.05472, 2021. URL `https://arxiv.org/abs/2103.05472`.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL `https://doi.org/10.1007/978-3-319-10602-1_48`.

Bo Liu, Ming Ding, Hanyu Xue, Tianqing Zhu, Dayong Ye, Li Song, and Wanlei Zhou. Dp-image: Differential privacy for image data in feature space. *CoRR*, abs/2103.07073, 2021. URL `https://arxiv.org/abs/2103.07073`.

Zelun Luo, Yuliang Zou, Yijin Yang, Zane Durante, De-An Huang, Zhiding Yu, Chaowei Xiao, Li Fei-Fei, and Animashree Anandkumar. Differentially private video activity recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pp. 6643–6653. IEEE, 2024. doi: 10.1109/WACV57701.2024.00652. URL `https://doi.org/10.1109/WACV57701.2024.00652`.

Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. *CoRR*, abs/2407.17140, 2024. doi: 10.48550/ARXIV.2407.17140. URL `https://doi.org/10.48550/arXiv.2407.17140`.

Simon Malm, Viktor Rönnbäck, Amanda Håkansson, Minh-Ha Le, Karol Wojtulewicz, and Niklas Carlsson. RAD: realistic anonymization of images using stable diffusion. In Erman Ayday and Jaideep Vaidya (eds.), *Proceedings of the 23rd Workshop on Privacy in the Electronic Society, WPES 2024, Salt Lake City, UT, USA, October 14-18, 2024*, pp. 193–211. ACM, 2024. doi: 10.1145/3689943.3695048. URL `https://doi.org/10.1145/3689943.3695048`.

Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. CIAGAN: conditional identity anonymization generative adversarial networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 5446–5455. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00549. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Maximov_CIAGAN_Conditional_Identity_Anonymization_Generative_Adversarial_Networks_CVPR_2020_paper.html.

Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning. *CoRR*, abs/1609.00408, 2016. URL http://arxiv.org/abs/1609.00408.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pp. 405–421. Springer, 2020. doi: 10.1007/978-3-030-58452-8_24. URL https://doi.org/10.1007/978-3-030-58452-8_24.

Carman Neustaedter and Saul Greenberg. Balancing privacy and awareness in home media spaces. In *Workshop on Ubicomp Communities: Privacy as Boundary Negotiation, held as part of the 5th International Conference on Ubiquitous Computing (UbiComp 2003)*, Seattle, WA, USA, October 12 2003. URL https://grouplab.cpsc.ucalgary.ca/grouplab/uploads/Publications/Publications/2003-BalancingPrivacy.UbicomWorkshop.pdf.

Carman Neustaedter, Saul Greenberg, and Michael Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Trans. Comput.-Hum. Interact.*, 13(1):1–36, March 2006. ISSN 1073-0516. doi: 10.1145/1143518.1143519. URL https://doi-org.vu-nl.idm.oclc.org/10.1145/1143518.1143519.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL http://proceedings.mlr.press/v139/nichol21a.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8024–8035, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *CoRR*, abs/2109.08203, 2021. URL https://arxiv.org/abs/2109.08203.

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pp. 286–295. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-DEMO.25. URL https://doi.org/10.18653/v1/2023.emnlp-demo.25.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

David Schneider, Sina Sajadmanesh, Vikash Sehwag, Saquib Sarfraz, Rainer Stiefelhagen, Lingjuan Lyu, and Vivek Sharma. Activity recognition on avatar-anonymized datasets with masked differential privacy. *CoRR*, 2024. URL https://arxiv.org/abs/2410.17098.

Arseniy Shakhmatov, Anton Razzhigaev, Aleksandr Nikolich, Vladimir Arkhipkin, Igor Pavlov, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 2, 2022. URL https://github.com/ai-forever/Kandinsky-2/tree/main. Accessed: Feb 2025.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 3–18. IEEE Computer Society, 2017. doi: 10.1109/SP.2017.41. URL `https://doi.org/10.1109/SP.2017.41`.

Yan Shoshitaishvili, Christopher Kruegel, and Giovanni Vigna. Portrait of a privacy invasion. *Proc. Priv. Enhancing Technol.*, 2015(1):41–60, 2015. doi: 10.1515/POPETS-2015-0004. URL `https://doi.org/10.1515/popets-2015-0004`.

Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5050–5059. Computer Vision Foundation / IEEE Computer Society, 2018a. doi: 10.1109/CVPR.2018.00530. URL `http://openaccess.thecvf.com/content_cvpr_2018/html/Sun_Natural_and_Effective_CVPR_2018_paper.html`.

Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A hybrid model for identity obfuscation by face replacement. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, volume 11205 of *Lecture Notes in Computer Science*, pp. 570–586. Springer, 2018b. doi: 10.1007/978-3-030-01246-5_34. URL `https://doi.org/10.1007/978-3-030-01246-5_34`.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pp. 3172–3182. IEEE, 2022. doi: 10.1109/WACV51458.2022.00323. URL `https://doi.org/10.1109/WACV51458.2022.00323`.

Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pp. 1–10. IEEE, 2020. doi: 10.1109/IJCNN48605.2020.9207181. URL `https://doi.org/10.1109/IJCNN48605.2020.9207181`.

The European Parliament. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), May 2016. URL `https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng`.

Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14895–14904. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01411. URL `https://doi.org/10.1109/CVPR52733.2024.01411`.

Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. DP-CGAN: differentially private synthetic data and label generation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 98–104. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPRW.2019.00018. URL `http://openaccess.thecvf.com/content_CVPRW_2019/html/CV-COPS/Torkzadehmahani_DP-CGAN_Differentially_Private_Synthetic_Data_and_Label_Generation_CVPRW_2019_paper.html`.

Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=YTWGvpFOQD-`.

Evgeniy Upenik, Pinar Akyazi, Mehmet Tuzmen, and Touradj Ebrahimi. Inpainting in omnidirectional images for privacy protection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pp. 2487–2491. IEEE, 2019. doi: 10.1109/ICASSP.2019.8683346. URL `https://doi.org/10.1109/ICASSP.2019.8683346`.

Mihaela van der Schaar, Isabelle Guyon, Nabeel Seedat, Jennifer Wortman Vaughan, Kyunghyun Cho, Razvan Pascanu, and Jim Weatherall. Data-centric AI for reliable and responsible AI: From theory to practice. In *Neural Information Processing Systems (NeurIPS) Tutorial*, 2023. URL `https://neurips.cc/virtual/2023/tutorial/73947`.

Chen Wang, Angtian Wang, Junbo Li, Alan L. Yuille, and Cihang Xie. Benchmarking robustness in neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pp. 2926–2936. IEEE, 2024a. doi: 10.1109/CVPRW63382.2024.00298. URL `https://doi.org/10.1109/CVPRW63382.2024.00298`.

Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXI*, volume 15089 of *Lecture Notes in Computer Science*, pp. 1–21. Springer, 2024b. doi: 10.1007/978-3-031-72751-1_1. URL `https://doi.org/10.1007/978-3-031-72751-1_1`.

Chengkun Wei, Minghu Zhao, Zhikun Zhang, Min Chen, Wenlong Meng, Bo Liu, Yuan Fan, and Wenzhi Chen. Dpmlbench: Holistic evaluation of differentially private machine learning. In Weizhi Meng, Christian Damsgaard Jensen, Cas Cremers, and Engin Kirda (eds.), *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, pp. 2621–2635. ACM, 2023. doi: 10.1145/3576915.3616593. URL `https://doi.org/10.1145/3576915.3616593`.

Yihan Wu, Brandon Y. Feng, and Heng Huang. Shielding the unseen: Privacy protection through poisoning nerf with spatial deformation. *CoRR*, abs/2310.03125, 2023. doi: 10.48550/ARXIV.2310.03125. URL `https://doi.org/10.48550/arXiv.2310.03125`.

Magdalena Wysocki, Mohammad Farid Azampour, Christine Eilers, Benjamin Busam, Mehrdad Salehi, and Nassir Navab. Ultra-nerf: Neural radiance fields for ultrasound imaging. In Ipek Oguz, Jack H. Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heimann, Despina Kontos, Bennett A. Landman, and Benoit M. Dawant (eds.), *Medical Imaging with Deep Learning, MIDL 2023, 10-12 July 2023, Nashville, TN, USA*, volume 227 of *Proceedings of Machine Learning Research*, pp. 382–401. PMLR, 2023. URL `https://proceedings.mlr.press/v227/wysocki24a.html`.

Lin Yen-Chen. Nerf-pytorch, 2020. URL `https://github.com/yenchenlin/nerf-pytorch/`.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in pytorch. *CoRR*, abs/2109.12298, 2021. URL `https://arxiv.org/abs/2109.12298`.

Guangsheng Yu, Xu Wang, Ping Yu, Caijun Sun, Wei Ni, and Ren Ping Liu. Dataset obfuscation: Its applications to and impacts on edge machine learning. *CoRR*, abs/2208.03909, 2022. doi: 10.48550/ARXIV.2208.03909. URL `https://doi.org/10.48550/arXiv.2208.03909`.

Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Trans. Vis. Comput. Graph.*, 29(7):3266–3280, 2023. doi: 10.1109/TVCG.2022.3156949. URL `https://doi.org/10.1109/TVCG.2022.3156949`.

Bokang Zhang, Yanglin Zhang, Zhikun Zhang, Jinglan Yang, Lingying Huang, and Junfeng Wu. S^{2}nerf: Privacy-preserving training framework for nerf. In Bo Luo, Xiaojing Liao, Jun Xu, Engin Kirda, and David Lie (eds.), *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS 2024, Salt Lake City, UT, USA, October 14-18, 2024*, pp. 258–272. ACM, 2024. doi: 10.1145/3658644.3690185. URL `https://doi.org/10.1145/3658644.3690185`.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 586–595. Computer Vision Foundation / IEEE Computer Society, 2018a. doi: 10.1109/CVPR.2018.00068. URL `http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html`.

Zhaoyu Zhang, Mengyan Li, and Jun Yu. On the convergence and mode collapse of GAN. In Nafees Bin Zafar and Kun Zhou (eds.), *SIGGRAPH Asia 2018 Technical Briefs, Tokyo, Japan, December 04-07, 2018*, pp. 21:1–21:4. ACM, 2018b. doi: 10.1145/3283254.3283282. URL `https://doi.org/10.1145/3283254.3283282`.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 16965–16974. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01605. URL `https://doi.org/10.1109/CVPR52733.2024.01605`.

Yanming Zhu, Xuefei Yin, Alan Wee-Chung Liew, and Hui Tian. Privacy-preserving in medical image analysis: A review of methods and applications. In Yupeng Li, Yong Zhang, and Jianliang Xu (eds.), *Parallel and Distributed Computing, Applications and Technologies - 25th International Conference, PDCAT 2024, Hong Kong, China, December 13-15, 2024, Proceedings*, volume 15502 of *Lecture Notes in Computer Science*, pp. 166–178. Springer, 2024. doi: 10.1007/978-981-96-4207-6_15. URL https://doi.org/10.1007/978-981-96-4207-6_15.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVIII*, volume 15116 of *Lecture Notes in Computer Science*, pp. 195–211. Springer, 2024. doi: 10.1007/978-3-031-73636-0_12. URL https://doi.org/10.1007/978-3-031-73636-0_12.

Pascal Zwick, Kevin Rösch, Marvin Klemp, and Oliver Bringmann. Context-aware full body anonymization using text-to-image diffusion models. *CoRR*, abs/2410.08551, 2024. doi: 10.48550/ARXIV.2410.08551. URL https://doi.org/10.48550/arXiv.2410.08551.

Appendix Organization

This appendix provides supplementary details and additional insights into various aspects of our work. It is structured as follows:

- **A: Detection Outcomes and Privacy Implications**: This section discusses the relationship between object detection performance and privacy preservation, including an analysis of false positives, false negatives, and their impact on dataset obfuscation.
- **B: NeRF**: This section covers NeRF preliminaries, detailing its mathematical formulation and rendering process. We describe dataset preprocessing for object removal and provide a detailed explanation of the metrics used for NeRF evaluation, including PSNR, SSIM, and LPIPS.
- **C: Background and Related Works**: A deeper exploration of foundational concepts relevant to our study, including prior work on dataset obfuscation, generative inpainting, and privacy-preserving machine learning.
- **D: Implementation and Reproducibility**: Details on our implementation, including parameter settings, training configurations, and dataset preprocessing steps, ensuring reproducibility.
- **E: Supplementary ROAR Results**: We include supplementary qualitative comparisons demonstrating the effectiveness of our ROAR framework in different object removal scenarios in the COCO dataset and NeRF scenes. Finally, we conclude with qualitative analysis of failure cases, presenting limitations of the ROAR pipeline.

Each section aims to provide extended analyses, insights, and experimental findings that complement the main paper. We encourage readers to refer to relevant sections based on their specific interests.

A Detection Outcomes and Privacy Implications

A.1 Detection Outcomes and Privacy Implications

To understand the impact of detection quality on privacy-preserving object scrubbing, we categorize images into three subsets based on detection outcomes:

- **True Positives (TP):** Images where there is actually a sensitive object.
- **False Negatives (FN):** Images containing sensitive objects that remain undetected. These samples represent a privacy risk as they remain in the dataset unaltered.
- **True Negatives (TN):** Images that do not contain any sensitive objects.

Here, the term object detection is used loosely to include segmentation models, where the primary objective is to detect objects.

Since our scrubbing pipeline is dependent on the accuracy of the sensitive object detection model, the false negative rate ρ directly affects the privacy guarantees of the pipeline. We assume that the detection of sensitive objects has a false negative rate ρ, while the false positive rate is orders of magnitude smaller, i.e., approximately zero. This means that the detection model is very unlikely to incorrectly identify non-sensitive objects as sensitive (i.e., false positives), but some sensitive objects may still go undetected (i.e., false negatives).

If a dataset consists of M images, with N images containing at least one sensitive object, the detector fails to identify sensitive content in ρN samples. Consequently, these ρN samples may still contain sensitive objects, thus affecting the privacy and utility of the dataset. As a result, our privacy guarantee is that the remaining $M - \rho N$ samples, which have been successfully scrubbed, do not contain any sensitive objects.

Defining N for Full Privacy (FP) and Selective Privacy (SP) In our privacy-preserving object scrubbing framework, the dataset is processed under two primary settings: Full Privacy (FP) and Selective Privacy (SP). The definition of N, the number of images considered in each setting, is crucial for understanding the evaluation metrics.

Full Privacy (FP). For the FP setting, where all instances of a sensitive object (e.g., persons) are removed from the dataset, we define N as the total number of images that contain at least one instance of the sensitive object. Specifically, for the COCO dataset, this corresponds to all images that contain at least one person.

Selective Privacy (SP). In the SP setting, we selectively scrub one randomly chosen person per image but only in half of the images that contain persons. As a result, the number of images considered in SP is given by:

$$N_{\text{SP}} = 0.5 \times N_{\text{FP}}. \tag{20}$$

This means that in the SP setting, only a subset of images undergo scrubbing, allowing for a controlled evaluation of privacy-utility trade-offs while retaining partial information about sensitive entities in the dataset.

A.2 Privacy-Utility Trade-offs and Detection Improvements

Our scrubbing pipeline builds upon existing detection methods, leveraging state-of-the-art object detection and segmentation models. These models, while highly effective, are not perfect. A higher false negative rate ρ means that more sensitive objects may slip through the detection process and remain in the dataset, compromising the privacy of the dataset. However, with an extremely low false positive rate, we are confident that most of the images without sensitive objects will be correctly classified as non-sensitive (i.e., true negatives), preserving the dataset's utility.

Since our pipeline relies on the outputs of SOTA object detectors, any improvement in detection accuracy, particularly in reducing false negatives, directly enhances the privacy guarantees. A decrease in ρ leads to a more robust scrubbing pipeline, as fewer sensitive images escape the transformation process. Furthermore, improving object detection to reduce ρ is always beneficial, as it ensures that fewer sensitive images remain unprotected in the dataset.

A.3 Impact of Dataset Composition

The effectiveness of scrubbing is closely tied to the true positive ratio, $\gamma = \frac{TP}{\text{All}}$. When γ is low, dropping sensitive images may be viable with minimal utility loss. However, in real-world datasets where sensitive objects are prevalent, dropping leads to severe data reduction, making scrubbing essential for balancing privacy and usability.

Empirical results show that datasets with a low true positive rate experience less accuracy degradation from dropping, as the loss of sensitive images has a smaller overall impact. However, this is uncommon in privacy-sensitive domains, where structured scrubbing is necessary to prevent excessive data loss that could hinder model training.

Maintaining both privacy and utility requires minimizing false negatives (ρ) while ensuring robust scrubbing. Improving detection models to reduce ρ enhances privacy by correctly identifying and removing sensitive objects while preserving non-sensitive data. Since false positives are assumed negligible, efforts should focus on refining detection accuracy for sensitive objects without disrupting non-sensitive content.

A.4 Impact of Scrubbing and Dropping on Object Detection Performance and Privacy Implications

As shown in Fig. 5, we analyze the impact of privacy-preserving transformations by evaluating the relationship between the proportion of sensitive images (denoted as $\gamma = \frac{TP}{\text{All}}$) in the dataset and the accuracy of object detection models. We compare two strategies: image scrubbing (FP) and image dropping (FP.drop) on Kandinsky2.2 (Razzhigaev et al., 2023). In our original dataset, $\gamma = 54$, meaning 54% of the images contain sensitive objects, specifically persons. The dataset distribution is provided in Table 4.

We trained models with different values of γ by systematically dropping images from either the sensitive (TP) or non-sensitive (TN) portion of the dataset, creating a range of experimental conditions: $\gamma = 30$, $\gamma = 40$, $\gamma = 54$ (original), $\gamma = 70$, and $\gamma = 80$. This allowed us to obtain a more detailed analysis of the performance trends across varying degrees of privacy constraints. Note that the results are based on the object classes listed in Table 3, which include a selection of small, large, and randomly chosen objects, totaling 11 objects along with the overall accuracy.

To further investigate performance variations, we clustered the data into four distinct groups based on the distribution of object-wise accuracy scores. These clusters naturally emerged around AP values of approximately 0.0, 0.2, 0.4, and 0.7.

The composition of each cluster provides insight into the underlying structure of the performance degradation:

Cluster 1 (Near 0.0 AP): Consists mostly of small objects such as *Backpack* and *Handbag*. These objects are frequently attached to persons and are often lost entirely when the person is scrubbed. Their recognition accuracy is severely impacted, with performance reductions exceeding 70%.

Cluster 2 (Around 0.2 AP): Includes objects like *Remote* and *Toothbrush*. These are typically small but independent objects, which may suffer from occlusion-related issues when sensitive entities are removed. Accuracy degradation in this cluster is significant but not as extreme as in Cluster 1.

Cluster 3 (Around 0.4 AP): Contains mid to large sized objects such as *Teddy Bear* and *Motorcycle*. These objects experience moderate performance degradation due to contextual dependencies on surrounding entities.

Cluster 4 (Around 0.7 AP): Comprises large and well-defined objects like *Bus* and *Airplane*. These objects are largely resilient to obfuscation, maintaining over 90% of their baseline performance.

A detailed breakdown for each object can be found in Table 3.

These findings align with a general trend in object detection: larger objects tend to be detected with higher accuracy. This tendency can be attributed to several factors. First, larger objects inherently occupy more pixels in an image, making them easier to distinguish from background noise. Second, deep learning models trained for object detection typically have better feature extraction capabilities for larger, more prominent objects, as they provide more spatial information. Third, occlusion effects impact small objects disproportionately; for example, a handbag or remote may become indistinguishable if a person is removed, whereas a bus or an airplane is less likely to be fully occluded by another object in the scene.

Furthermore, we observe that moving from smaller to larger objects in our clusters, the detection model's performance improves significantly for smaller objects, whereas for larger objects where the detection model already performs well the gap in accuracy becomes much smaller. This trend is evident when comparing the relative drops in AP between the clusters. Small objects in Cluster 1 experience a dramatic decrease in AP, whereas the degradation becomes progressively less severe as we move to larger objects in Cluster 4.

This effect can be further explained by the inherent difficulty that object detection models face when detecting small objects. Since small objects already pose a challenge for the model due to their limited spatial information and higher susceptibility to occlusion, the removal of sensitive images that might contain them intensifies this issue. Dropping sensitive images disproportionately harms the detection of these smaller objects because they often co-occur with the sensitive entities being removed. In contrast, structured scrubbing, which removes only the sensitive object while preserving the rest of the scene, retains valuable context that helps the model better recognize small objects.

For instance, in Cluster 1, scrubbing achieves a 2.44× improvement in AP compared to image dropping, highlighting the substantial benefit of preserving non-sensitive parts of the image. This improvement is particularly crucial for small objects that are frequently occluded or embedded in complex backgrounds. As we move to larger objects in Cluster 4, where the detection model already performs well, the advantage of scrubbing over dropping diminishes. Larger objects are less dependent on contextual features and are more likely to be detected independently of the presence of sensitive entities. Consequently, the gap between scrubbing and dropping decreases, reaffirming that structured scrubbing is most beneficial for objects that are already challenging to detect.

These observations further reinforce the necessity of adaptive privacy-preserving strategies. While dropping sensitive images might seem like a straightforward privacy measure, it disproportionately harms the detection of small objects, leading to greater accuracy degradation. Scrubbing, on the other hand, maintains a higher level of dataset utility by selectively removing sensitive content while preserving crucial visual information. As a result, structured scrubbing proves to be a more effective approach, particularly in datasets where small objects play an important role in downstream tasks.

In conclusion, structured scrubbing provides a clear advantage over image dropping, particularly for datasets containing high proportions of sensitive content. The cluster-wise analysis highlights the varying levels of resilience among different object categories and underscores the importance of developing more adaptive scrubbing methods to preserve dataset utility while enforcing strong privacy guarantees.

A.5 Selection of τ and ζ

The verification step in Stage 3 is applied only to objects that had significant spatial overlap with the removed sensitive objects, while all other objects are automatically retained in the final annotation set.

The ζ is a predefined threshold for determining collision. In our setup, we set $\zeta = 0$ to ensure that verification is performed aggressively, meaning that any object that has any degree of overlap with the removed region is considered for verification. This decision prioritizes caution, ensuring that all potentially affected objects are evaluated. However, more relaxed values of ζ could be explored in future work to balance computational efficiency and verification robustness.

For the intersection-over-union (IoU) threshold τ, which determines whether the oracle-detected object corresponds to the original ground truth object, we select $\tau = 0.3$ based on empirical observations. This threshold was determined by analyzing various IoU settings and their impact on verification performance. While a higher τ might provide stricter verification, we observed that for small objects (especially cases where persons are heavily occluded) using an IoU threshold greater than 0.3 did not significantly change the verification outcome. This is likely due to the fact that distinct objects tend to have relatively lower IoU values, making stricter thresholds redundant in these cases.

Nonetheless, further study is required to identify an optimal τ, as different datasets and application scenarios might exhibit different tendencies. Exploring adaptive or learned thresholding mechanisms based on object size and category could further refine the verification process, ensuring better alignment with dataset characteristics and downstream task requirements.

Table 3: **Comparison of AP$_{50-95}$ across categories for Baseline, Scrubbing (FP.KD), and Dropping (FP.drop) using RT-DETRv2.**

Class	Baseline	FP.KD (Scrubbing)	FP.drop (Dropping)	Remarks[*]
Airplane	0.752 (100%)	0.735 (97.74%)	0.681 (90.56%)	Stable across methods, minimal drop.
Bus	0.722 (100%)	0.707 (97.92%)	0.658 (91.14%)	Minor degradation, well-preserved.
All	0.480 (100%)	0.420 (87.50%)	0.356 (74.16%)	Significant drop in overall performance for dropping.
Car	0.483 (100%)	0.471 (97.52%)	0.410 (84.89%)	Minimal impact, robust across all methods.
Motorcycle	0.531 (100%)	0.421 (79.28%)	0.285 (53.66%)	Dropping severely impacts this category.
Teddy Bear	0.481 (100%)	0.444 (92.31%)	0.351 (72.97%)	Degradation in accuracy due to context loss.
Bicycle	0.349 (100%)	0.303 (86.82%)	0.215 (61.60%)	Dropping affects objects that are tied with persons significantly.
Chair	0.345 (100%)	0.285 (82.61%)	0.213 (61.74%)	Moderate drop, highlights contextual loss.
Remote	0.397 (100%)	0.214 (53.90%)	0.162 (40.81%)	Highly affected by both methods, especially by dropping.
Toothbrush	0.327 (100%)	0.233 (71.25%)	0.184 (56.27%)	Small objects are significantly impacted.
Handbag	0.198 (100%)	0.058 (29.29%)	0.018 (9.09%)	Most degraded class, difficult to detect.
Backpack	0.196 (100%)	0.063 (32.14%)	0.021 (10.71%)	Significant accuracy loss for accessories.

[*] Higher values indicate better performance. Clusters are separated by horizontal lines.

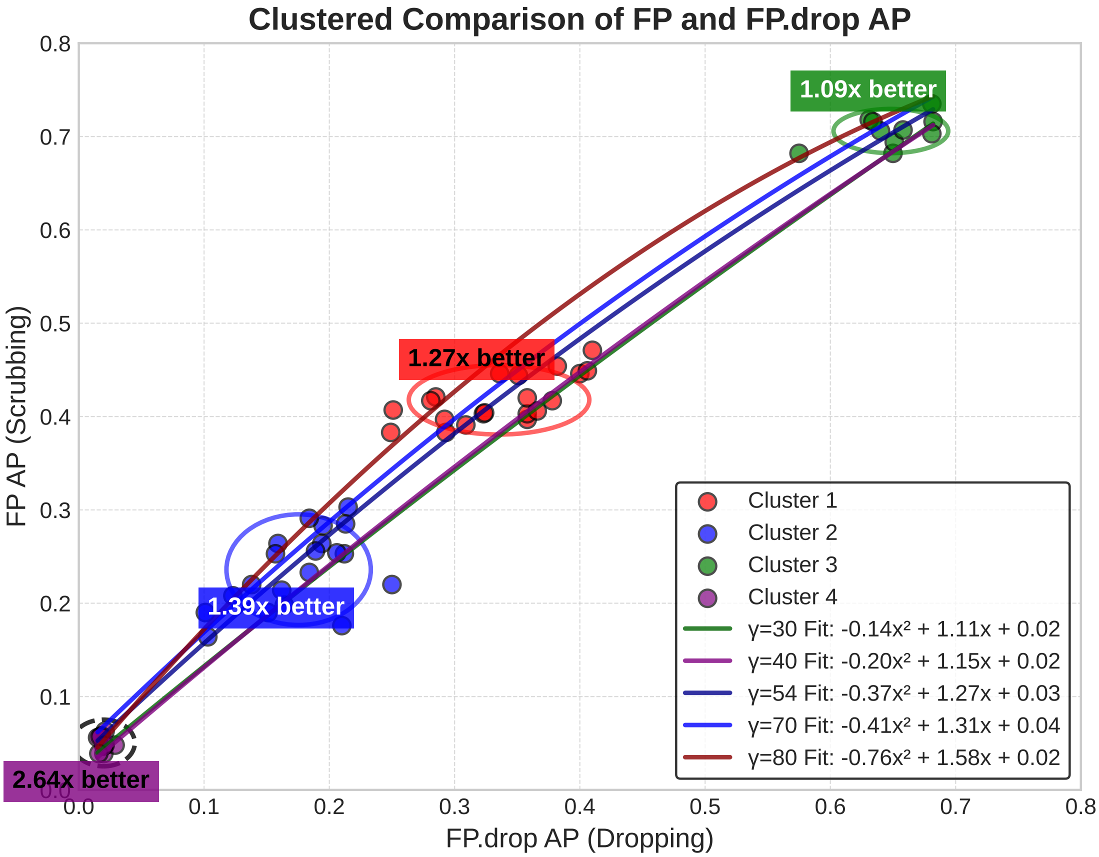


Figure 5: Comparison of FP (scrubbing sensitive objects) and FP.drop (dropping images with sensitive objects) accuracies across different clusters. The improvement factor indicates how much better scrubbing performs compared to dropping on average, particularly highlighting the significant advantages for small and occlusion-prone objects.

B NeRF

B.1 NeRF Preliminaries

In NeRF (Mildenhall et al., 2020), 2D images with their respective 3D spatial coordinates of the camera from where the image was taken are used to render a 3D volumetric model to represent the object being reconstructed. In this each 2D pixel of the image is parameterized to a camera ray $\mathbf{r}_t = \mathbf{r}_o + t\mathbf{r}_d$, where $\mathbf{r}_o \in \mathbb{R}^3$ represents the rays origin, $\mathbf{r}_d \in \mathbb{R}^3$ represents the rays direction, and t represents the depth along the ray. To render a particular pixel from a different view, NeRF model f samples points along the corresponding ray. For each sampled point, the model returns its color c and density σ, expressed as $(\sigma, c) = f(\mathbf{r}_t, \mathbf{d})$. The final pixel color $\hat{\mathbf{C}}(\mathbf{r})$ is then obtained by integrating the colors of these sampled points along the ray using:

$$\hat{\mathbf{C}}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}_t)c(\mathbf{r}_t, \mathbf{d})dt, \tag{21}$$

where $T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}_s)ds\right)$ is the transmittance accumulated along the ray between the depths t_n and t.

Table 4: Dataset Statistics Overview. This table presents the distribution of persons and objects, including general statistics, person-specific metrics, and the most frequent object categories.

General Statistics	
Total Images	118,287
Total Annotations	860,001
Person Annotations	262,465
Other Object Annotations	597,536
Person Distribution	
Images with Persons	64,115 (54.20%)
Average Persons/Image	4.09
Median Persons/Image	2.00
Images with >N Persons	
N > 1	39,283
N > 2	28,553
N > 5	16,049
N > 10	8,407

Using a numerical quadrature rule (Mildenhall et al., 2020), the integral is approximated as follows:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) c_i, \tag{22}$$

where $T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)$ and $\delta_i = t_{i+1} - t_i$ denoting the distance between two samples.

Finally, the NeRF model f is trained under the MSE loss between the rendered color pixels and the ground truth $\mathbf{C}(\mathbf{r})$ as follows for a coarse-sampled NeRF model and fine-sampled NeRF model:

$$\mathcal{L} = \sum_{r \in \mathcal{R}} [\left\|\hat{\mathbf{C}}_{\mathbf{c}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\right\|_2^2 + \left\|\hat{\mathbf{C}}_{\mathbf{f}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\right\|_2^2] \tag{23}$$

where $\mathcal{R}$ is the accumulation of sampled camera rays.

Evaluation Metrics

To assess the quality of the rendered images, we utilize three widely used evaluation metrics, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). These metrics provide pixel-wise accuracy and perceptual similarity between the synthesized and ground-truth images.

Peak Signal-to-Noise Ratio (PSNR)

PSNR is a common metric for measuring the fidelity of reconstructed images by comparing them to reference images. It is computed as:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}_{bit}^2}{\text{MSE}}\right), \tag{24}$$

where MAX_{bit} is the maximal possible value of the pixel in the image. For an 8-bit channel, the maximum value is 255 (Horé & Ziou, 2010). A higher PSNR value signifies improved reconstruction quality.

Structural Similarity Index Measure (SSIM)

SSIM evaluates the structural similarity between two images by considering luminance, contrast, and structural components. It is defined as:

$$\mathrm{SSIM}(A, B) = \frac{(2\mu_A\mu_B + C_1)(2\sigma_{AB} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_A^2 + \sigma_B^2 + C_2)} \tag{25}$$

where μ_A, μ_B are the average intensities of the images A and B, respectively. σ_A,σ_B, and σ_{AB} correspond to the standard deviations of the images and their cross-variance. C_1 and C_2 are small positive constants used for numerical stability (Horé & Ziou, 2010). A higher SSIM value signifies improved reconstruction quality.

Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS is a perceptual metric that compares high-level features extracted from deep neural networks rather than relying on pixel-wise differences. It computes the similarity between two images by measuring the difference in their feature representations from a pre-trained network, typically a deep convolutional neural network (CNN) (Zhang et al., 2018a). For our experiments, we used VGG 16. A lower LPIPS value signifies improved reconstruction quality.

B.2 NeRF Experiments

For each NeRF dataset, a specific object was selected for removal based on its presence in multiple views. In the Room scene, the TV was removed. In the Flower scene, a living bug appearing in 29 out of 34 images was selected for obfuscation. In the Fern scene, the ceiling light was the target for removal. Since these objects were explicitly chosen, NeRF-based obfuscation does not rely on instance segmentation or re-annotation and directly applies generative inpainting to remove them from all viewpoints. We split the obfuscated dataset into training (90%) and test (10%) subsets and trained NeRF models on the processed data (Yen-Chen, 2020). All images were uniformly resized to 512 × 384 × 3, to ensure consistency in resolution and color depth across evaluations.

To maintain view consistency across NeRF reconstructions, we implemented a *stitching-based inpainting technique*. This method ensures that inpainted regions remain coherent across multiple perspectives. Since object masks were manually created, slight inconsistencies were present across different views, which could lead to artifacts during NeRF reconstruction. To mitigate this, we first identify the image containing the largest instance of the object to be removed. This image serves as a template for inpainting using a pre-trained generative model. The masked region from this template image is then extracted and resized dynamically to match the bounding box of the corresponding object in all other images.

Once resized, the inpainted region is seamlessly blended into each target image using standard alpha blending. To further refine appearance consistency, we apply histogram matching at the boundary regions, ensuring that color and texture transitions between the inpainted patch and the surrounding scene remain smooth. Additionally, a Gaussian-blurred edge blending technique is used to suppress harsh transitions that may arise due to illumination variations across viewpoints.

By leveraging this stitching strategy, our approach minimizes the risk of inconsistent inpainting artifacts across multi-view NeRF datasets, leading to improved coherence in the final reconstructed 3D scene.

C Background & Related Works

C.1 Data-Centric Machine Learning and Privacy

Machine learning pipelines involve multiple stages: data acquisition, training, testing, and deployment (van der Schaar et al., 2023). The data stage is particularly crucial when addressing privacy concerns, as it establishes the foundation for subsequent processing. Unlike model-level privacy methods, which secure models or outputs, data-level privacy focuses on safeguarding datasets while preserving their utility for downstream tasks.

Privacy-preserving machine learning can be categorized into data-level and model-level approaches (Wei et al., 2023). Model-level techniques such as federated learning (Kairouz et al., 2021; Bonawitz et al., 2019) and secure multi-party computation prevent direct data exposure but do not mitigate risks inherent in raw datasets. Traditional data-level privacy techniques, such as differential privacy (DP) and noise-based obfuscation, often degrade data utility, particularly in high-dimensional tasks like object detection (Chen et al., 2020; Torkzadehmahani et al., 2019).

To address these limitations, our work introduces a novel paradigm in data privacy: object scrubbing. Unlike anonymization techniques that modify or replace sensitive elements while preserving contextual integrity, scrubbing directly removes sensitive objects from images. This approach minimizes the risk of re-identification while ensuring the dataset remains usable for downstream tasks. Moreover, unlike DP training, which depletes a privacy budget and inherently limits the number of learning tasks for which a dataset can be used, object scrubbing allows the dataset to remain usable for any number of learning tasks without privacy degradation.

C.2 GANs and Diffusion Models

GANs (Goodfellow et al., 2020) have been widely used for image synthesis and inpainting. Advanced models such as AOT-GAN (Zeng et al., 2023) improve high-resolution inpainting through aggregated contextual transformations, enabling better texture synthesis and object removal. However, GANs suffer from mode collapse and training instabilities, making them less reliable for privacy-preserving dataset obfuscation (Thanh-Tung & Tran, 2020; Durall et al., 2021; Zhang et al., 2018b).

Latent Diffusion Models (LDMs) (Rombach et al., 2022) address these limitations by applying a progressive denoising process (Ho et al., 2020) in a lower-dimensional latent space rather than directly in pixel space. This approach significantly reduces computational costs while maintaining high-fidelity image generation, making LDMs well-suited for privacy-preserving tasks such as targeted object removal. Stable Diffusion is a specific instance of LDMs that leverages text-to-image conditioning via CLIP embeddings, allowing for controllable and efficient image synthesis (Rombach et al., 2022). Other LDM-based models, such as Kandinsky2.2 (Razzhigaev et al., 2023), further refine this process by introducing structured image priors, enhancing semantic control in transformations.

Compared to GANs, LDMs provide greater robustness, maintain structural coherence, and might offer stronger privacy guarantees by operating in a latent space. These advantages make LDMs particularly effective for dataset obfuscation, where sensitive entities must be removed while preserving the integrity of surrounding image content.

C.3 Object Detection and Semantic Segmentation Models

Object detection and semantic segmentation are two fundamental tasks in computer vision, both aimed at understanding scene composition but differing in their approach. Object detection focuses on localizing and classifying objects within an image using bounding boxes, while semantic segmentation assigns a class label to each pixel, providing a more granular representation of object regions.

In recent years, transformer-based architectures have significantly advanced object detection by leveraging self-attention mechanisms to model long-range dependencies. A notable example is RT-DETRv2 (Lv et al., 2024), which enhances the original RT-DETR framework with a hybrid encoder-decoder architecture for optimized multi-scale feature extraction. By decoupling intra-scale interactions from cross-scale fusion and using scale-adaptive sampling in the deformable attention module, RT-DETRv2 improves both flexibility and efficiency.

Another state-of-the-art detector, YOLOv9 (Wang et al., 2024b), integrates Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN) to optimize gradient flow and parameter utilization. Unlike prior YOLO variants, YOLOv9 enhances feature representation through gradient path planning, improving both convergence stability and detection accuracy. By retaining full input information across layers, it remains competitive with transformer-based models while maintaining the efficiency of CNN-based architectures.

For instance segmentation, Mask2Former (Cheng et al., 2022) unifies semantic, instance, and panoptic segmentation through a transformer-based masked attention mechanism. Unlike fully convolutional networks (FCNs), it dynamically extracts region-specific features, restricting cross-attention to localized areas. This improves segmentation accuracy, accelerates convergence, and enhances small-object segmentation through multi-scale feature aggregation. Mask2Former achieves state-of-the-art results on COCO and ADE20K benchmarks, outperforming both specialized and prior universal segmentation models.

Our methodology combines transformer-based and CNN-based architectures for robust privacy-preserving dataset obfuscation. RT-DETRv2 enables real-time detection with optimized multi-scale attention, while Mask2Former ensures precise segmentation for anonymization.

C.4 Anonymization

Recent research in privacy-preserving computer vision has explored various methodologies for dataset anonymization, focusing on both differential privacy mechanisms and generative adversarial methods. Barattin et al. (Barattin et al., 2023) propose an attribute-preserving face dataset anonymization approach leveraging latent code optimization within a pre-trained GAN space. Their method optimizes identity obfuscation while preserving crucial facial attributes using a feature-matching loss in FaRL's deep feature space. In contrast, He et al. (He et al., 2024) introduce Diff-Privacy, a diffusion-based framework integrating multi-scale image inversion to enhance both anonymization and visual identity protection.

Hukkelas et al. (Hukkelås & Lindseth, 2023) extend previous face anonymization frameworks to full-body anonymization using DeepPrivacy2, a GAN-based model that ensures realistic occlusion of individuals. Their work addresses limitations in detecting and anonymizing full-body figures by incorporating pose estimation and conditional synthesis techniques.

Lee and You (Lee & You, 2024) analyze the trade-off between privacy and accuracy in deep learning models trained on anonymized data, demonstrating that aggressive anonymization techniques can significantly degrade model performance. This aligns with findings from Li and Clifton (Li & Clifton, 2021) on differentially private imaging, where latent space manipulation is used to inject noise selectively, balancing privacy guarantees and data utility. In contrast, Malm et al. (Malm et al., 2024) introduce the RAD framework, which integrates Stable Diffusion with ControlNet for high-utility anonymization while preserving downstream model performance.

Maximov et al. (Maximov et al., 2020) propose CIAGAN, a conditional identity anonymization GAN that allows controlled identity swapping for dataset anonymization, ensuring anonymization while maintaining downstream utility.

Sun et al. (Sun et al., 2018a) present a head-inpainting approach for naturalistic identity obfuscation, demonstrating that generative models can replace identity-revealing regions while retaining contextual integrity. Similarly, Zwick et al. (Zwick et al., 2024) explore text-to-image diffusion models to synthesize anonymized figures that integrate seamlessly into complex scenes.

Our work differentiates from traditional anonymization techniques by adopting a data transformation strategy centered on complete object removal. Instead of modifying or replacing sensitive elements within an image, we systematically eliminate them while preserving the surrounding scene structure. By leveraging advanced generative models, including latent diffusion and stable diffusion techniques, we evaluate their effectiveness in reconstructing realistic, high-fidelity backgrounds post-removal. Unlike prior approaches that focus on masking or synthetic identity replacement, our method ensures that no identifiable traces remain, offering robust privacy guarantees without compromising the usability of the remaining dataset.

C.5 Differential Privacy in Computer Vision

Differential Privacy (DP) (Dwork & Roth, 2014) provides a formal framework for quantifying and limiting privacy loss in machine learning systems.

While DP offers robust privacy guarantees, it imposes significant trade-offs between privacy and utility. For high-dimensional data, such as images, achieving meaningful privacy requires injecting substantial noise, which often degrades task-specific performance. Frameworks like TensorFlow Privacy and PyTorch Opacus (Yousefpour et al., 2021; Paszke et al., 2019) have been developed to integrate DP into machine learning workflows, but they require extensive parameter tuning and adaptation to support computer vision tasks such as object detection and segmentation (Wei et al., 2023). These challenges make DP less practical for real-world applications where preserving fine-grained features is essential.

Several adaptations of DP for computer vision have emerged to address these limitations. For example, Masked Differential Privacy (MaskDP) selectively obfuscates sensitive regions rather than applying noise uniformly across an entire dataset (Schneider et al., 2024). This targeted approach improves utility by preserving non-sensitive components of the data. Similarly, VisualMixer (Li et al., 2024) disrupts sensitive visual features through pixel shuffling while maintaining overall data fidelity. Although these methods offer improvements over traditional DP approaches, they often lack the precision needed for high-dimensional tasks and can lead to degradation in downstream utility for object detection or segmentation.

Our work departs from these traditional DP approaches by leveraging latent diffusion models to obfuscate sensitive features within the latent space. Unlike noise-based methods, latent diffusion models enable semantically meaningful transformations that preserve structural and contextual integrity. By focusing on privacy at the data level, our approach avoids the severe trade-offs associated with DP in high-dimensional scenarios, ensuring robust privacy guarantees while maintaining task-specific utility. This paradigm shift makes generative models particularly well-suited for computer vision tasks, where both precision and scalability are critical.

C.6 GANs and Private Dataset Generation Techniques

Generative models, including GANs and diffusion models, have recently gained attention for privacy-preserving tasks. Methods like DP-CGAN and GS-WGAN adapt GANs to generate synthetic data satisfying differential privacy guarantees (Chen et al., 2020; Torkzadehmahani et al., 2019). However, these methods often introduce artifacts and fail to preserve the original distribution of the dataset, limiting their applicability to tasks requiring high fidelity. PrivSet (Chen et al., 2022a) uses dataset condensation to optimize a small set of synthetic samples for downstream tasks. DPGEN (Chen et al., 2022b) incorporates DP into energy-based models to generate private synthetic data, while Hand-DP (Tramèr & Boneh, 2021) leverages scattering networks to extract wavelet-based features before fine-tuning models using DP-SGD (Abadi et al., 2016).

In contrast to synthetic data generation, Yu et al. propose direct dataset obfuscation using random noise to preserve privacy during training outsourcing or edge applications. Their approach evaluates privacy, utility, and distinguishability trade-offs through the PUD-triangle framework. While effective for simple scenarios like MNIST or CIFAR-10, this method's reliance on uniform noise obfuscation may degrade utility and task-specific performance in complex datasets (Yu et al., 2022).

Our work fundamentally differs from the aforementioned methods as it directly operates on the original images without relying on synthetic data generation. By leveraging diffusion models, we effectively scrub sensitive objects while preserving the structural and contextual integrity of the dataset. Unlike synthetic data generation approaches that often introduce artifacts or distort the original data distribution, our method maintains the dataset's fidelity, ensuring that the modified data remains as close to its original distribution as possible. This direct interaction with the raw data allows us to overcome the utility degradation commonly observed in noise-based obfuscation methods, supporting downstream tasks such as object detection and recognition with minimal performance loss.

Additionally, our approach streamlines the data preparation process by eliminating the need for training computationally expensive generative models for specific tasks. This not only reduces overhead but also ensures a more interpretable mapping between the original and obfuscated datasets, a critical requirement for sensitive applications where transparency and traceability are essential.

Latent diffusion models provide a powerful framework for privacy-preserving transformations by iteratively applying controlled noise and denoising processes within a compressed latent space (Rombach et al., 2022; Ho et al., 2020). These models achieve precise modifications while preserving both the structural coherence and the semantic meaning of the data (Razzhigaev et al., 2023). Building on this foundation, our approach applies diffusion models directly to the original data to target sensitive objects, avoiding the artifacts commonly associated with synthetic data generation. This ensures that privacy is preserved without compromising utility, making our method particularly effective for high-dimensional tasks like object detection and recognition in complex, real-world datasets.

D Implementation and Reproducibility

This section provides details on the experimental setup, including hardware specifications, model configurations, and datasets used to ensure the reproducibility of our results.

D.1 Hardware and Computational Setup

Our experiments were conducted on various GPU configurations, including NVIDIA RTX A5000, A6000, and A4000. The computational resources were allocated based on the complexity of the models and the dataset sizes.

D.2 Object Detection and Utility Benchmarking

We utilized the official implementations from the **RT-DETRv2** (Zhao et al., 2024; Lv et al., 2024) and **YOLOv9** (Wang et al., 2024b; Chang et al., 2023) repositories for our object detection experiments.

- **RT-DETRv2**: We trained and evaluated models using the `rtdetrv2_r101vd_6x_coco.yml` configuration file. The RT-DETRv2 models were trained for 120 epochs as per the predefined settings in the repository.
- **YOLOv9**: For YOLOv9, we used the **YOLOv9-M** model and trained it for 100 epochs.

For benchmarking, the **RT-DETRv2-X** model was used as the oracle detector, while **RT-DETRv2-M** was employed for utility evaluation due to computational constraints.

D.3 Random Seeds and Dataset Preprocessing

To ensure reproducibility:

- A random seed of **3407** (Picard, 2021) was used for all main experiments.
- A seed of **42** was used for other randomized processes such as selecting images to drop.

E Supplementary ROAR Results

The following figures illustrate different privacy-preserving transformations applied to datasets. To avoid redundancy, individual captions are omitted. Instead, we provide a general description of the transformations below.

In Figure 6, Full Privacy (FP) ensures the removal of all sensitive objects in the dataset, leaving no identifiable entities. In contrast, Selective Scrubbing (SP) targets only one randomly chosen individual per image in half of the sensitive dataset, balancing privacy with data retention.

All other figures (Figs. 7, 8, 9, 10, 11, and 12) depict the Full Privacy (FP) setting, where all sensitive objects in the dataset are removed, ensuring no identifiable entities remain. The raw image (left) has no privacy protection, while at the opposite extreme (omitted in figures) is image dropping, which ensures full privacy at the cost of data loss. The intermediate method is the anonymization (modifying sensitive elements while preserving context) with DeepPrivacy2 (Hukkelås & Lindseth, 2023; Hukkelås et al., 2019), and our object scrubbing approach, which removes sensitive objects while maintaining scene integrity using Stable Diffusion and Kandinsky (Rombach et al., 2022; Razzhigaev et al., 2023). Figure 12 shows the blackout baseline, which replaces sensitive content with black rectangles.

In Figures 16, 17, and 18, we present a comparison of the original images and their scrubbed counterparts generated using different obfuscation techniques. Each row represents a view from the scene, where the first column corresponds to the original image, while the subsequent columns depict images processed using AOT-GAN-based scrubbing (Zeng et al., 2023), Stable Diffusion 2.2 (SD) scrubbing (Rombach et al., 2022), and Kandinsky 2.2 (KD) (Razzhigaev et al., 2023) scrubbing, respectively. The other figures, (Figs. 19, 20, and 21), depict the original image, the best-performing method (Kandinsky scrubbed images) and their NeRF reconstruction.

E.1 Qualitative Failure Case Analysis

We have showcased representative failure cases in (Figs. 13, 14, and 15) to illustrate typical limitations of the ROAR pipeline under both high-density occluded scenes and sparse single-object scenarios. The captions of these figures highlight key failure patterns, such as background hallucination, contextual bias from nearby sensitive objects, and misdetections by the post-scrubbing oracle detector. In particular, Fig. 15 demonstrates that even when scrubbing is visually successful, images may still be discarded due to oracle failure in detecting nearby non-sensitive objects. These insights are derived from qualitative visual analysis of specific examples. However, we emphasize that the failure causes may not be generalizable to every failure; verifying them would require extensive analysis and manual inspection. Future work may pursue a more rigorous categorization of failure modes, enabling a deeper understanding of their origins and implications. Therefore, these examples are intended to provide intuition for common failure modes rather than a comprehensive or statistically representative diagnosis.

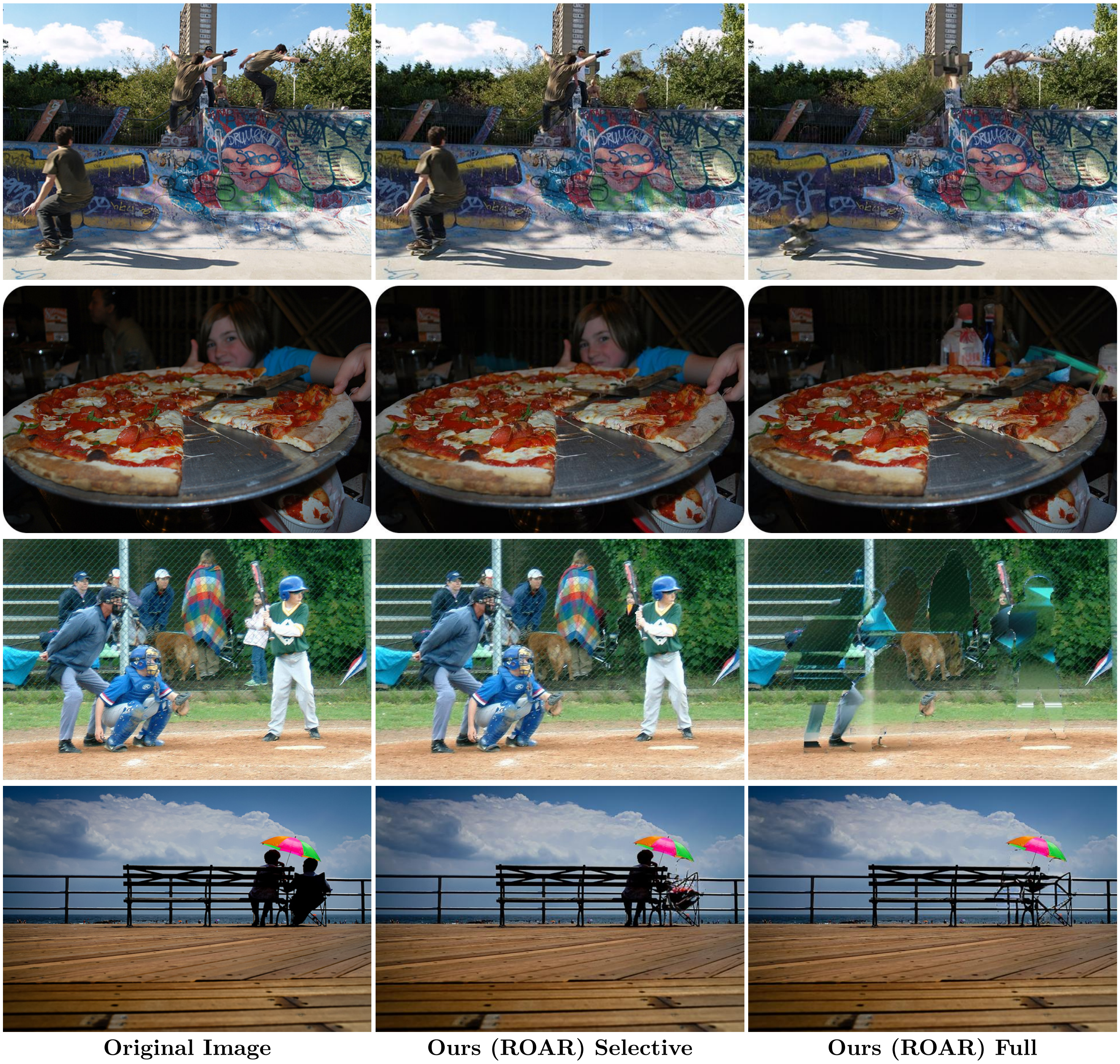
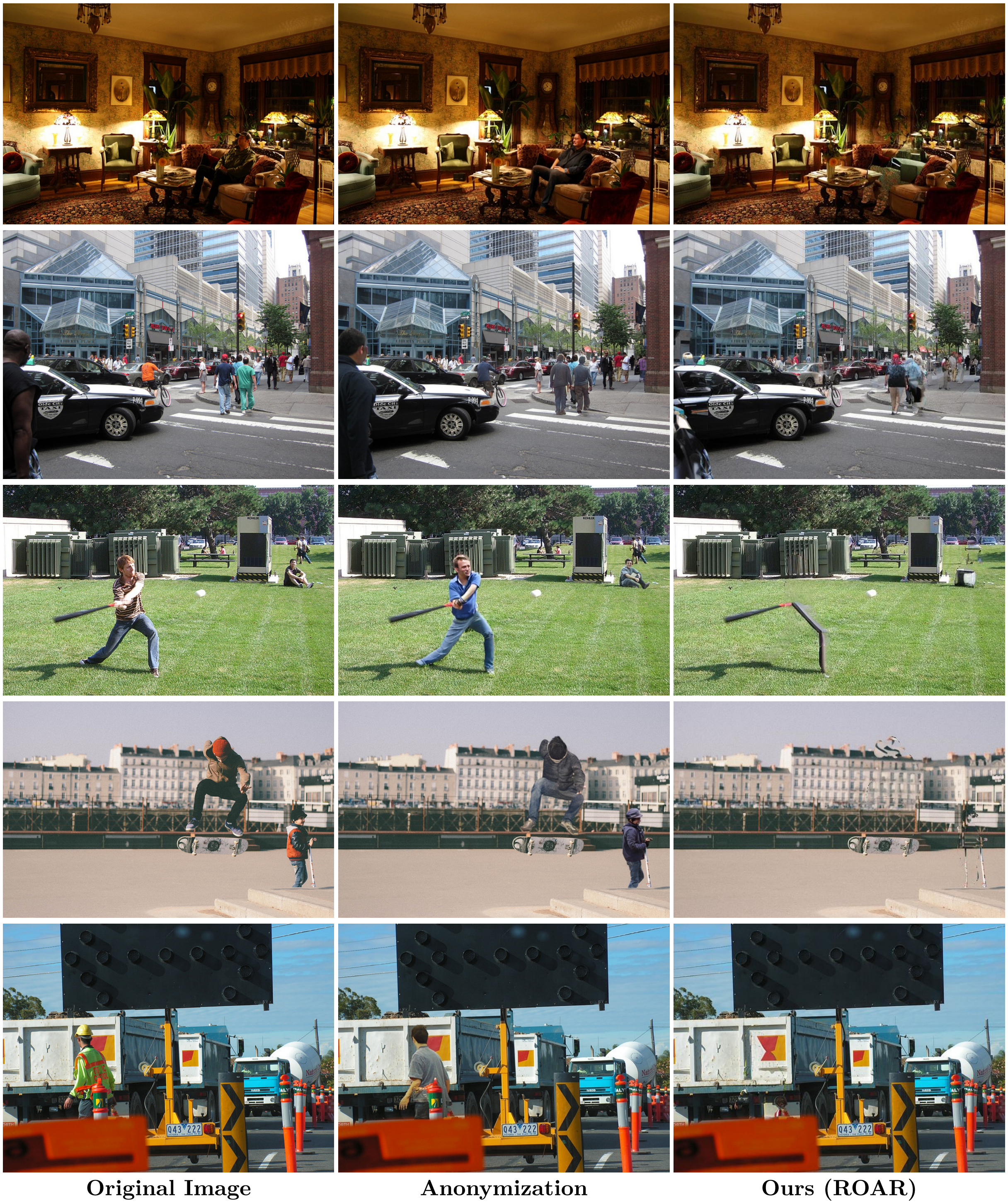
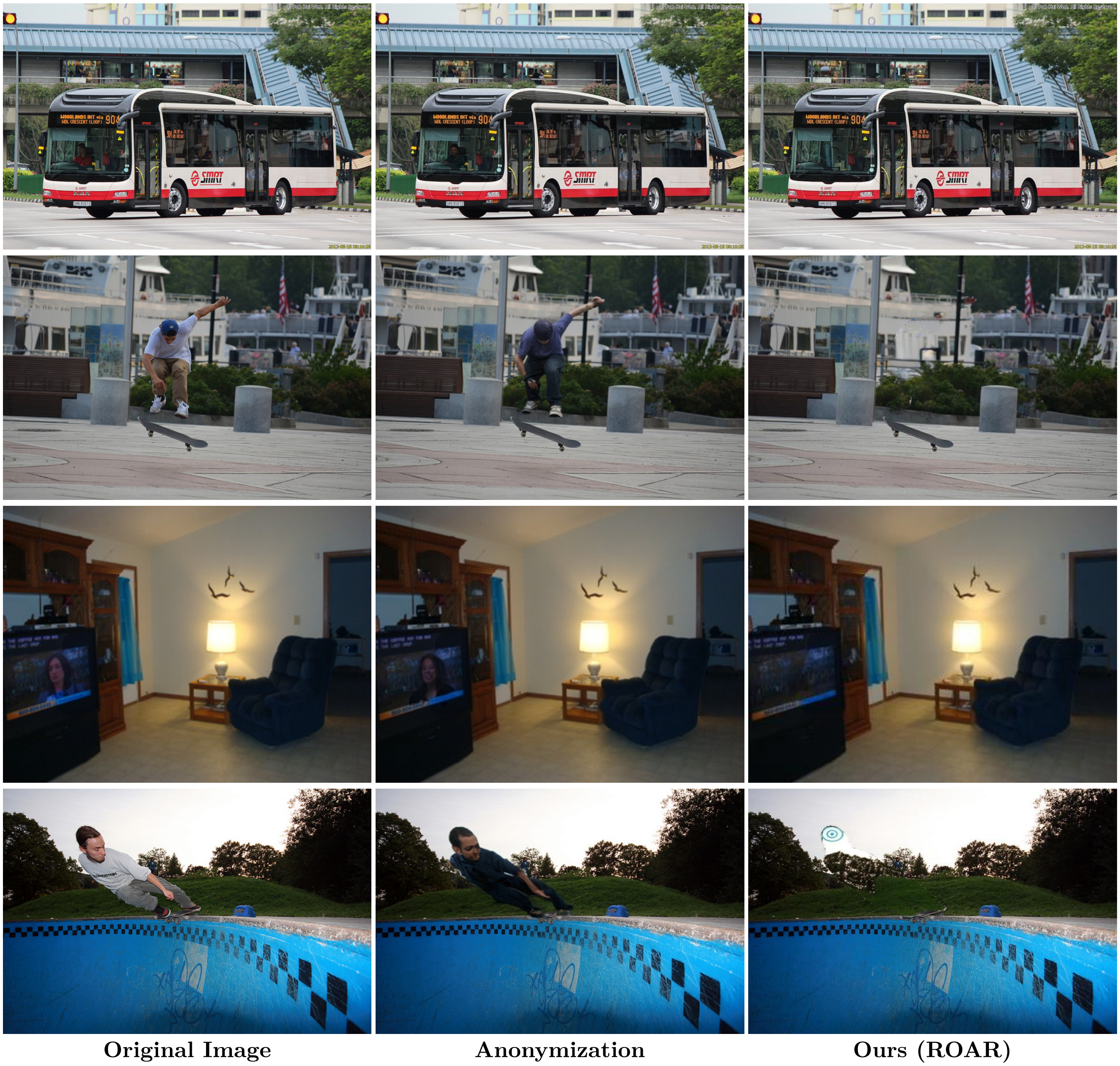


Figure 6: Comparison of different approaches: Full Privacy (FP) ensures the removal of all sensitive objects in the dataset, leaving no identifiable entities. In contrast, Selective Scrubbing (SP) targets only one randomly chosen individual per image in half of the sensitive dataset, balancing privacy with data retention.



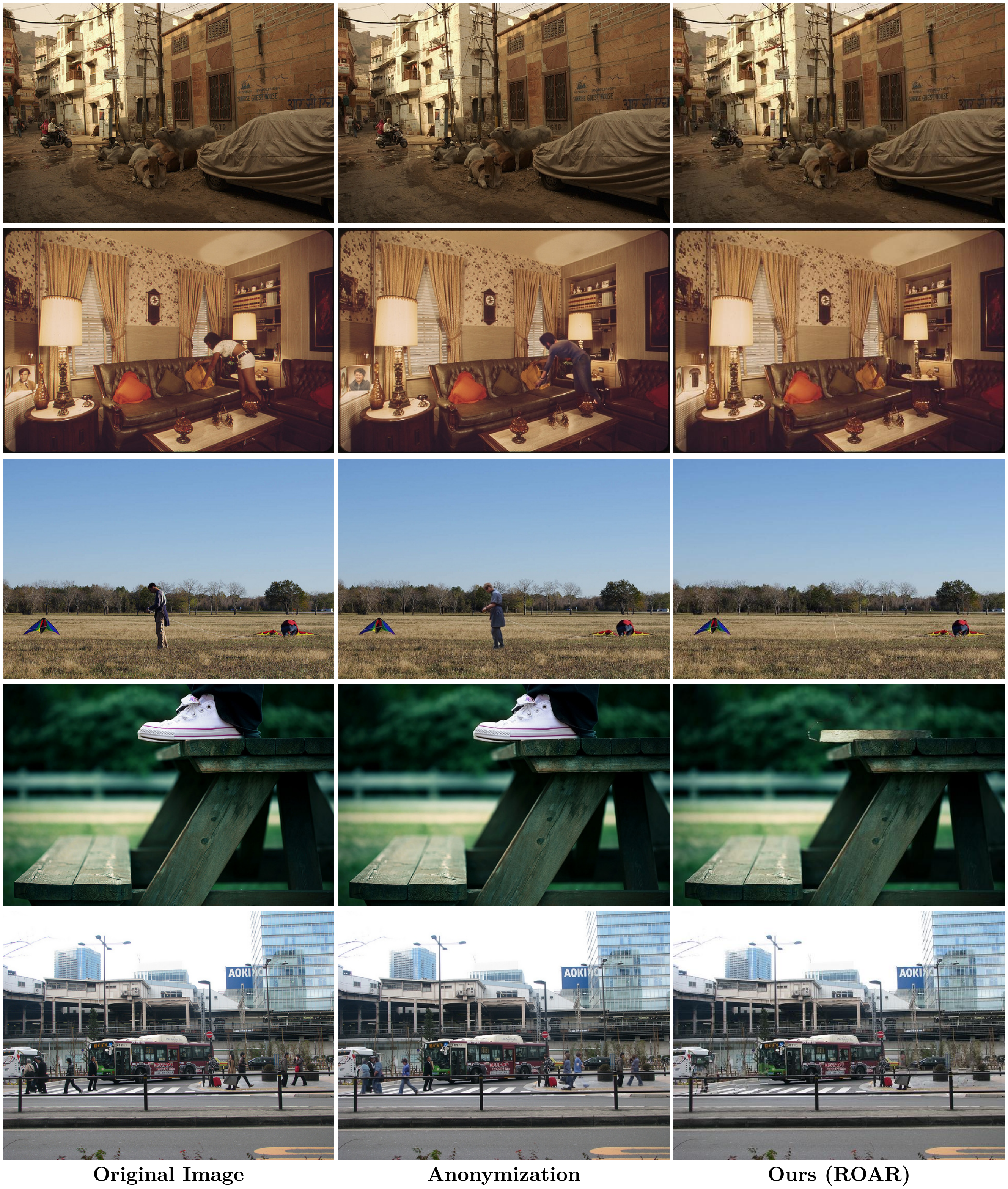
Original Image **Anonymization** **Ours (ROAR)**

Figure 7: Comparison of different approaches.



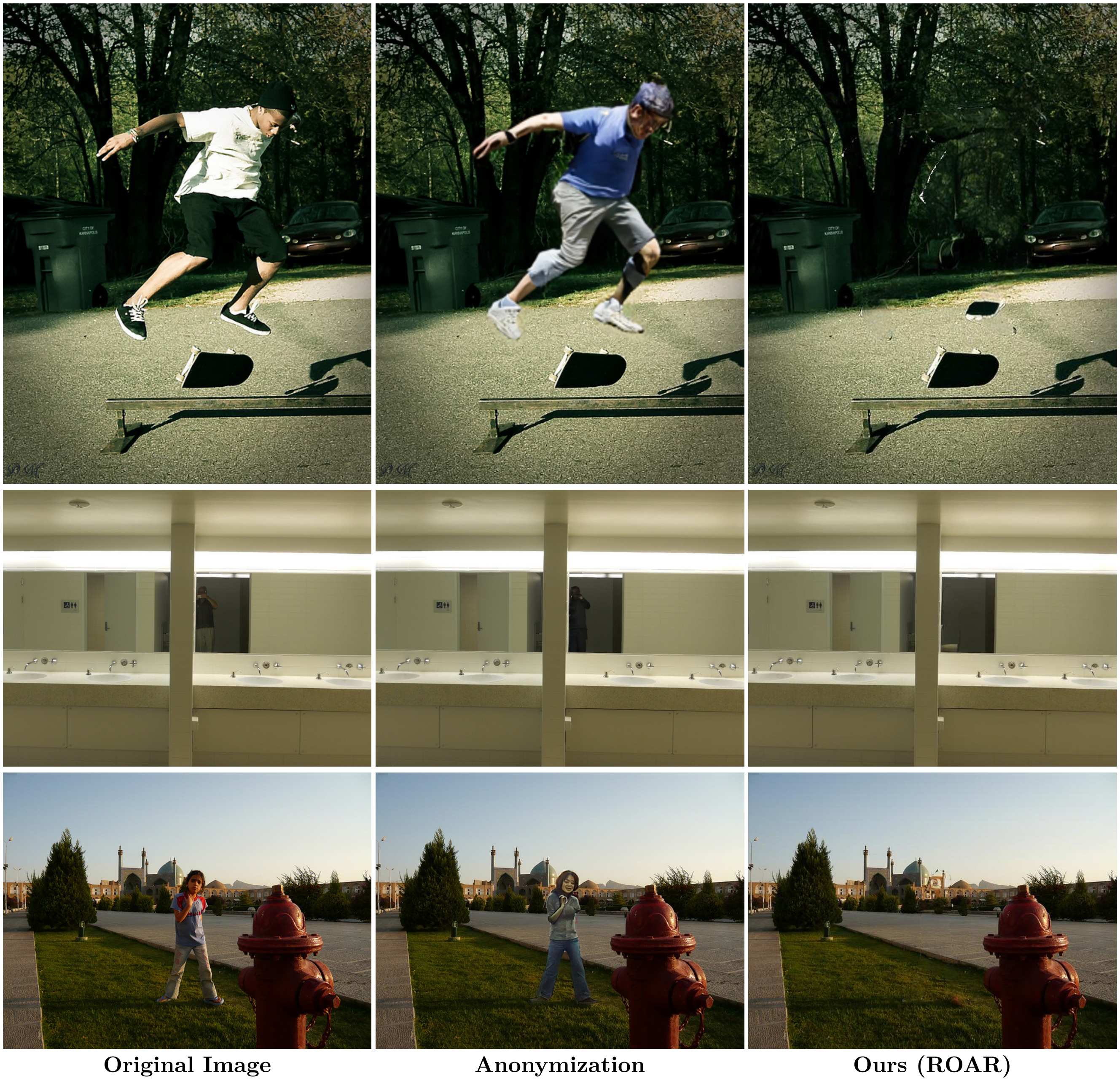
Original Image **Anonymization** **Ours (ROAR)**

Figure 8: Comparison of different approaches.



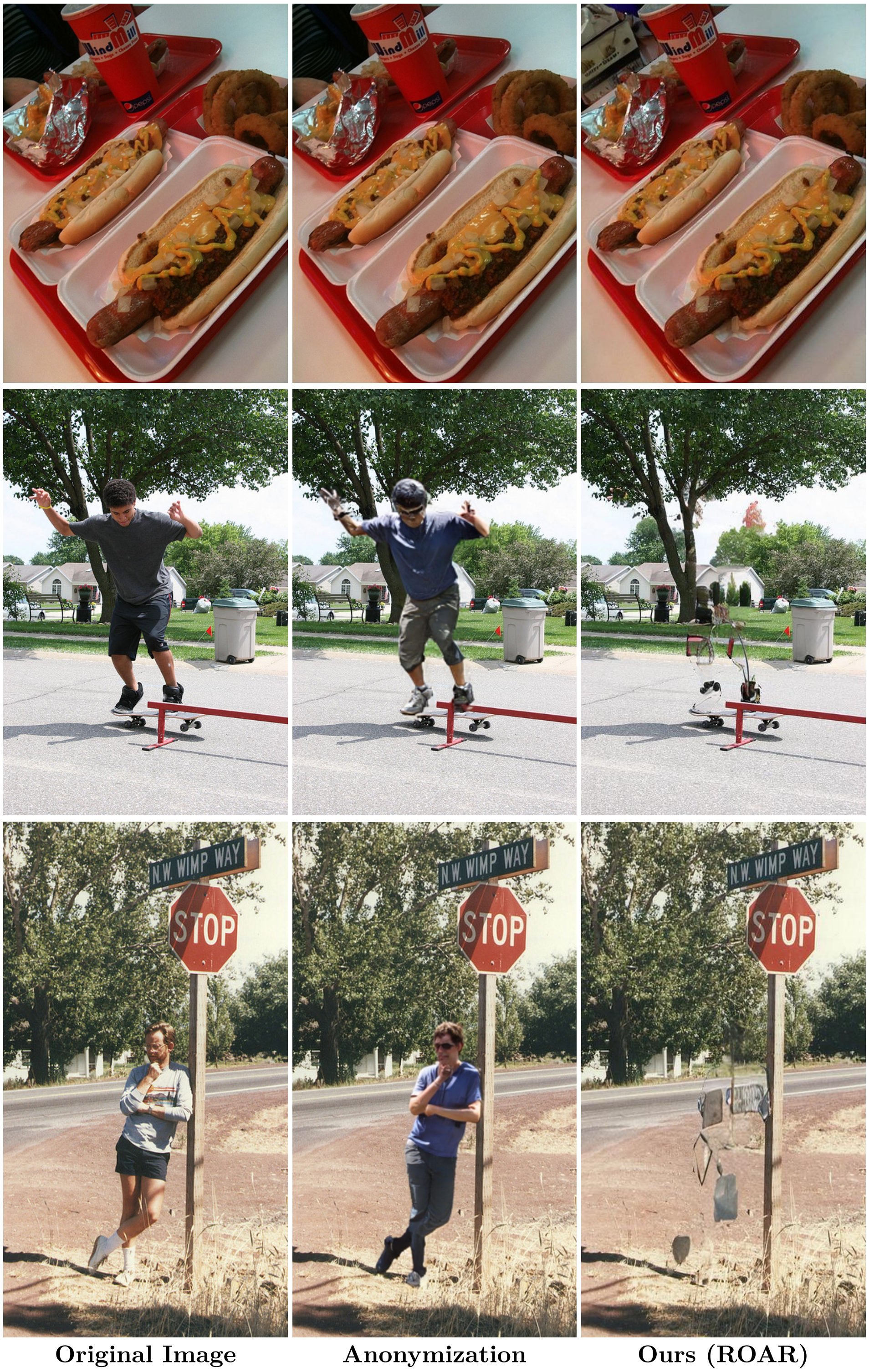
Original Image **Anonymization** **Ours (ROAR)**

Figure 9: Comparison of different approaches.



Original Image **Anonymization** **Ours (ROAR)**

Figure 10: Comparison of different approaches.



Original Image **Anonymization** **Ours (ROAR)**

Figure 11: Comparison of different approaches.

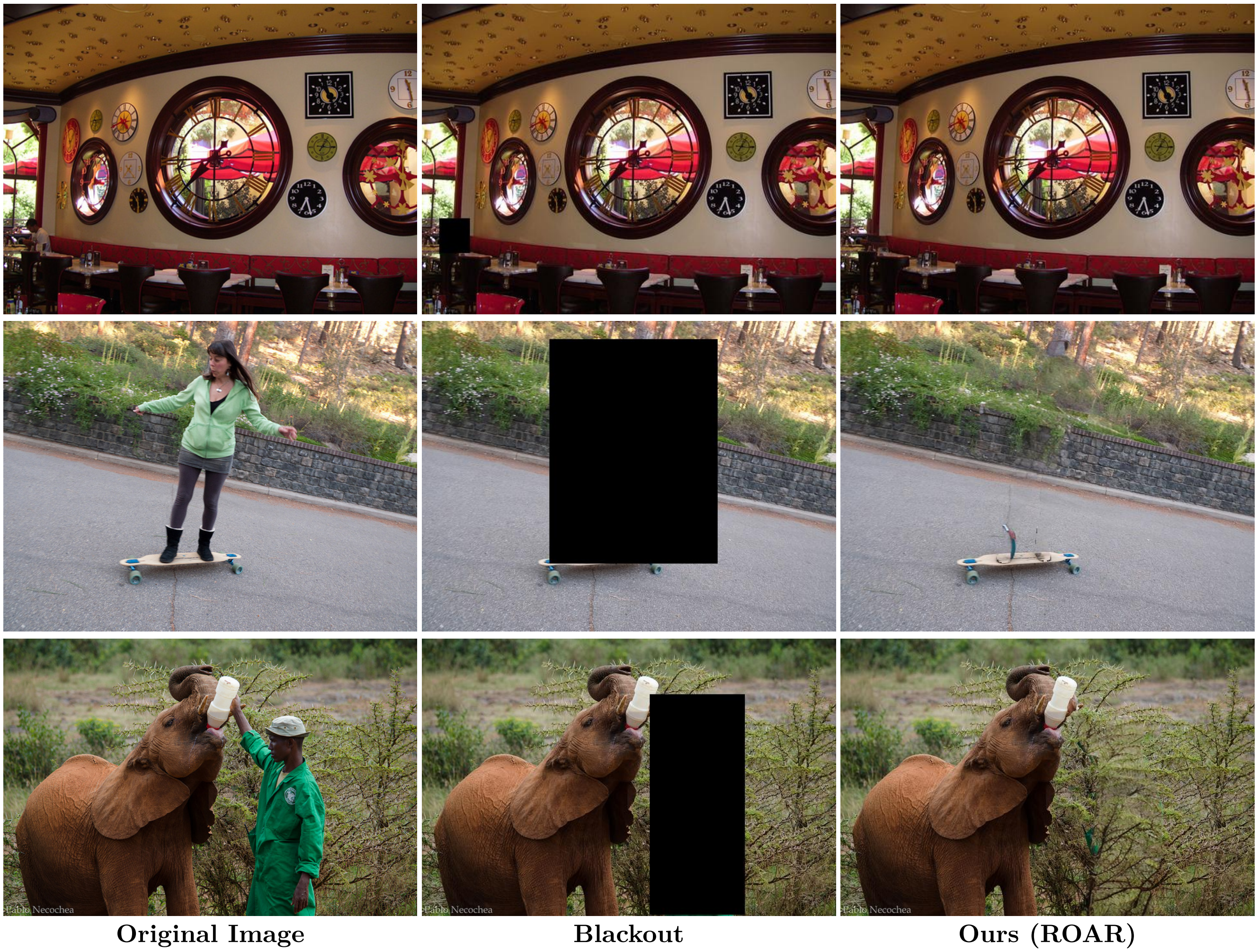
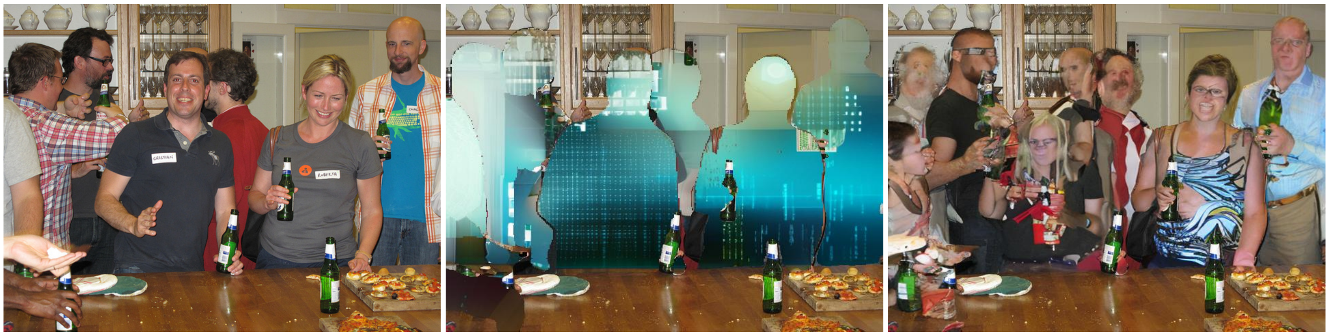
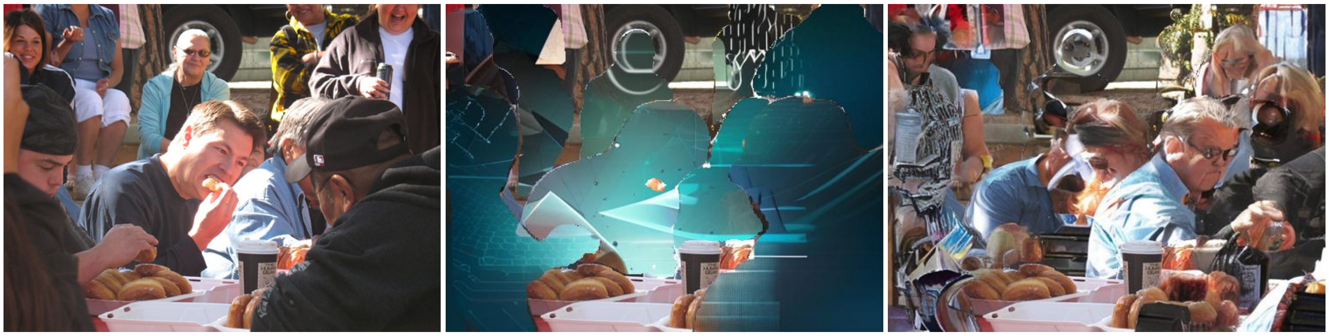


Figure 12: Comparison of blackout baseline and ROAR (ours).



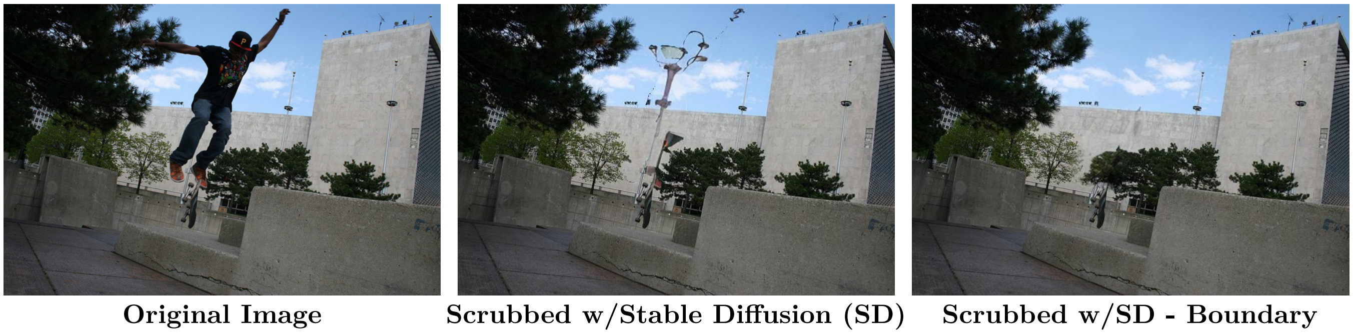
Original Image **Scrubbed w/Kandinsky** **Scrubbed w/Stable Diffusion**

Figure 13: **Failure case analysis.** The original image (left) contains multiple large sensitive objects (persons), many of which are occluded by handheld objects (e.g., bottles). The high density of persons in the scene appears to influence the generative models: Kandinsky replaces persons with vague blueish silhouettes, which still leak identity cues and fail to fully erase sensitive content. Stable Diffusion, on the other hand, hallucinates synthetic humans in place of the removed ones, likely due to strong contextual priors from the surrounding crowd. Although technically replaced, the presence of human-like figures undermines privacy. This failure illustrates that hallucinated humans, although artificial, still violate the intended privacy guarantees. Furthermore, the presence of occluded secondary objects (e.g., bottles) leads to collateral removal. Pre-ROAR annotations included 8 bottles; post-ROAR, only 3 and 4 are detected in the Kandinsky and Stable Diffusion outputs, respectively. This reduction may be attributed to two factors: (i) partial inpainting due to occlusion within the masked region, and (ii) degraded post-scrubbing detection performance. Overall, this example highlights key failure modes: background hallucination and object co-dependence that challenge the robustness of ROAR.



Original Image **Scrubbed w/Kandinsky** **Scrubbed w/Stable Diffusion**

Figure 14: **Failure case analysis.** The original image (left) contains multiple large sensitive objects (persons), many of whom are mutually occluded. Similar to the pattern observed in 13, the inpainting models fail to synthesize a coherent background, likely due to the overwhelming presence of nearby persons dominating the contextual priors. Beyond the generative failure, this example also highlights a limitation of the post-scrubbing oracle detector. In the original image, there are 12 donuts, 1 cup, and a car. Post-ROAR, the oracle successfully re-identifies the donuts and cup, but fails to detect the car. This failure stems from the fact that only a small, partially occluded region of the car (e.g., wheels) remains visible after inpainting, and also that region is partially corrupted. While the object is still visible to the human eye, the oracle does not recognize it and thus discards the annotation. This case exemplifies two key challenges: (i) inpainting under high-density person occlusion can yield incoherent or insufficient context reconstruction, and (ii) post-hoc re-annotation is sensitive to small, partially scrubbed objects, thus demonstrating a clear gap between human perception and model verification.



Original Image **Scrubbed w/Stable Diffusion (SD)** **Scrubbed w/SD - Boundary**

Figure 15: **Failure case analysis.** In this scenario, the scrubbing appears visually successful (middle), the person has been effectively removed. However, the image is discarded at the final stage of the ROAR pipeline because the oracle detector fails to verify the presence of any remaining non-sensitive objects. The original image (left) contains only two annotations: a person and a skateboard. Post-scrubbing, although the skateboard appears mostly unaltered and only minimally occluded, the oracle fails to detect it. This highlights a case where the limitations of the oracle model directly impact ROAR's utility. The stronger the oracle detector, the better the utility that ROAR can retain. It is also worth noting that while ROAR would treat this as a successful scrubbing case, minor visual cues remain that suggest a person was present. These residual traces can be mitigated using our boundary expansion variant, which enlarges the mask by 10 pixels. As shown in the right example, this eliminates cues more effectively and even leads to a better background synthesis. However, in such sparse scenes, boundary expansion poses a risk: since the skateboard is the only non-sensitive object and lies close to the person, the expanded mask may engulf it entirely leading to its removal and guaranteeing failure of oracle verification, regardless of oracle quality.

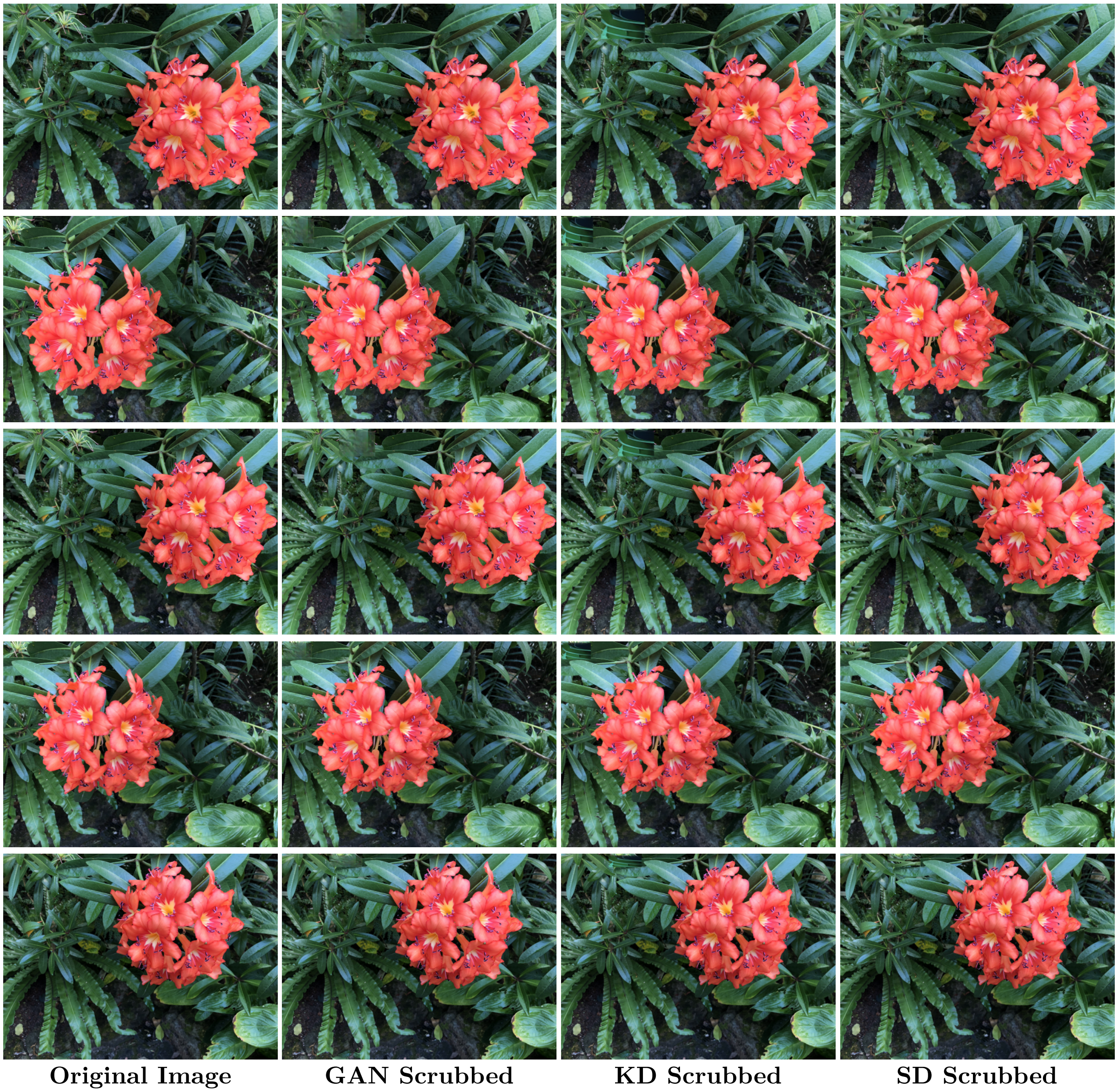


Figure 16: Flower Scene. Comparison of original and scrubbed images using GAN, KD, and SD inpainting methods. The removed object is a green bug located in the upper-left region of the image.

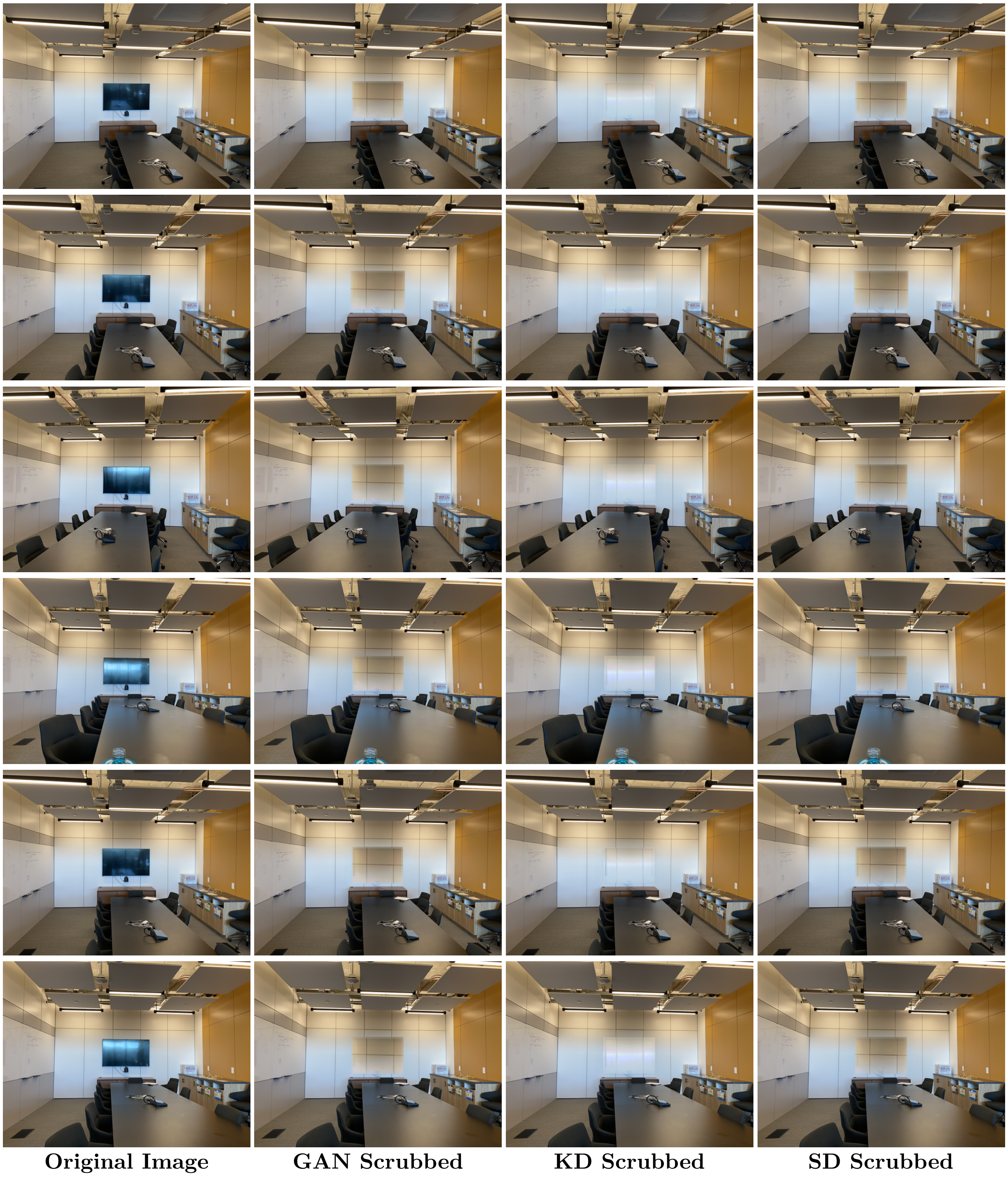
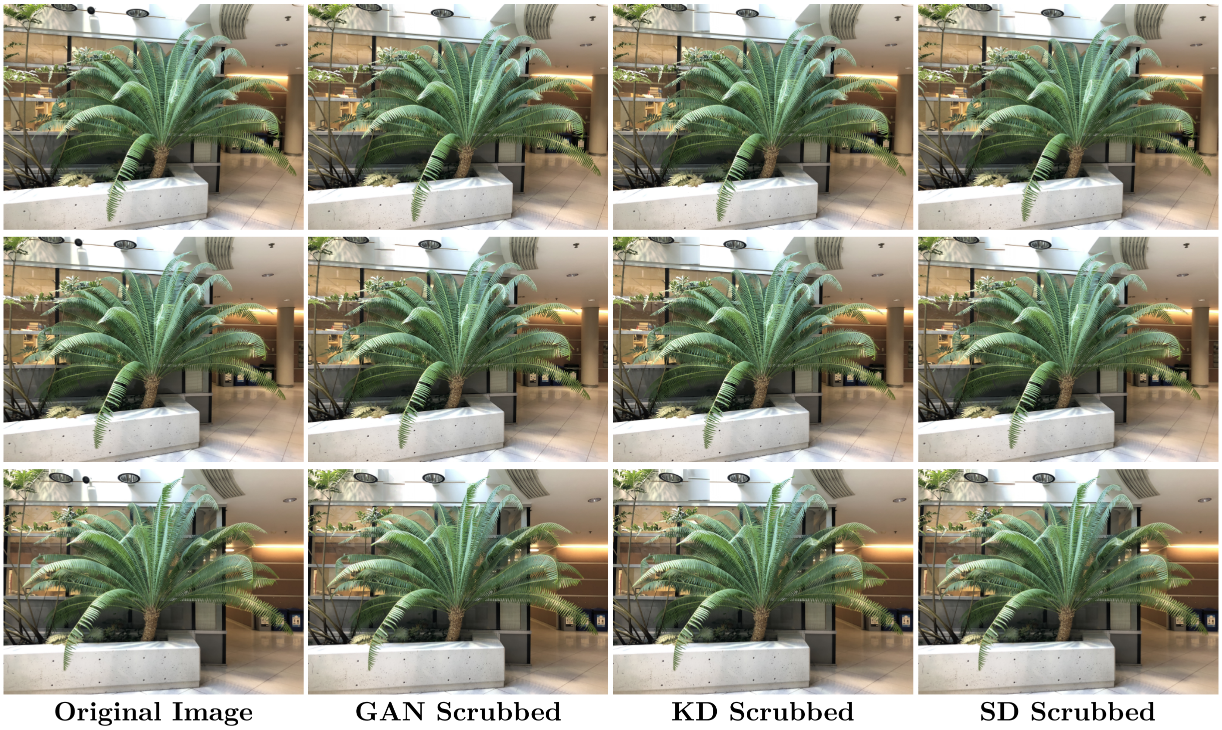


Figure 17: Room Scene. Comparison of original and scrubbed images using GAN, KD, and SD inpainting methods. The removed object is a television positioned at the center of the image.



Original Image **GAN Scrubbed** **KD Scrubbed** **SD Scrubbed**

Figure 18: Fern Scene. Comparison of original and scrubbed images using GAN, KD, and SD inpainting methods. The removed object is a white lamp located near the upper center of the image.

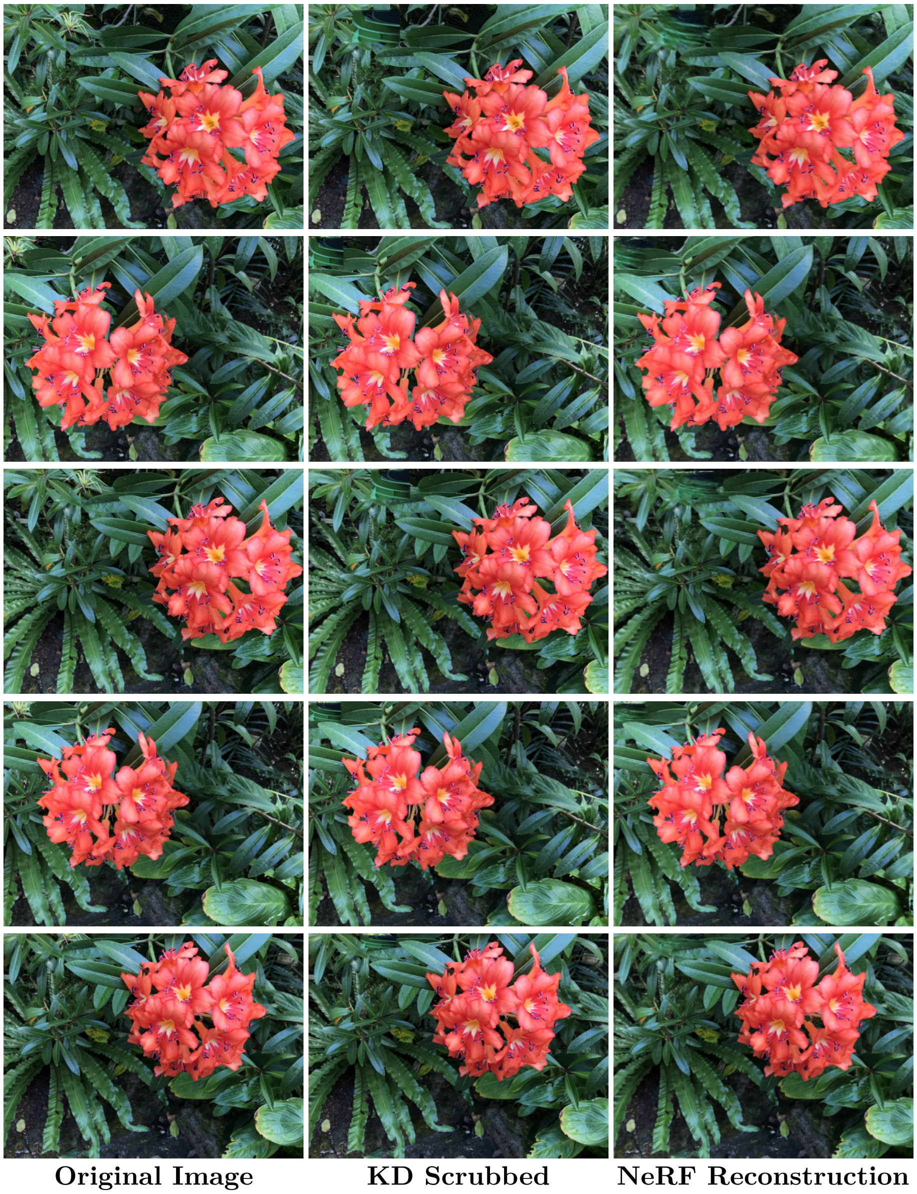


Figure 19: Flower Scene: original image, scrubbed using Kandinsky, and the NeRF reconstruction. The removed object is a green bug located in the upper-left region of the image.

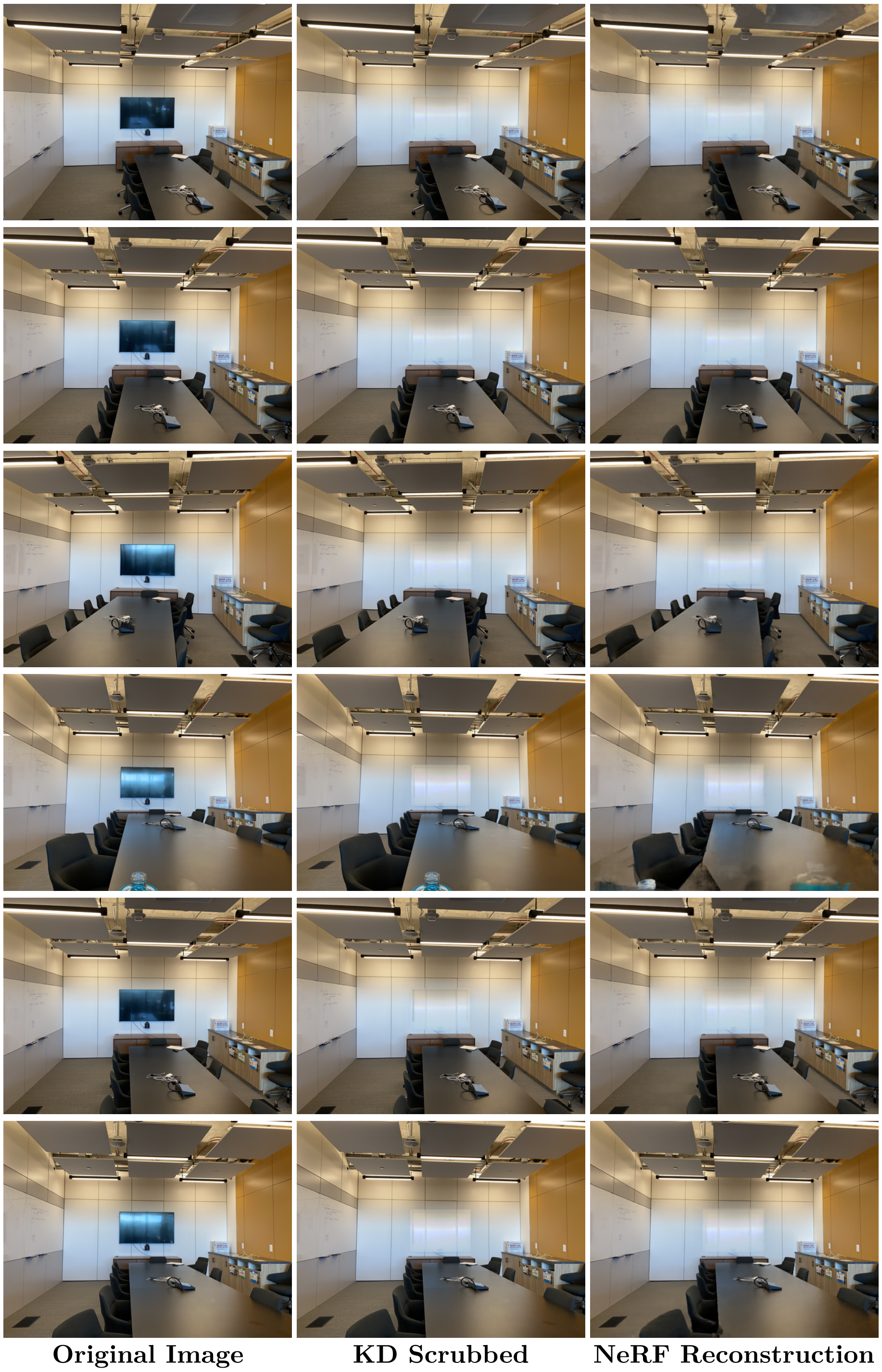


Figure 20: Room Scene: original image, scrubbed using Kandinsky, and the NeRF reconstruction. The removed object is a television positioned at the center of the image.

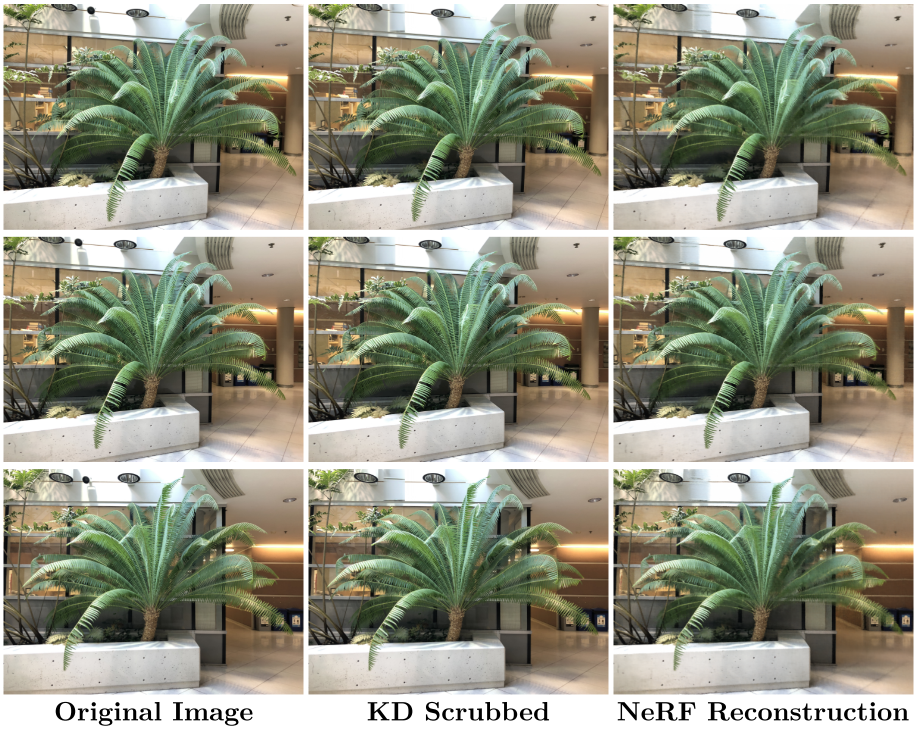


Figure 21: Fern Scene: original image, scrubbed using Kandinsky, and the NeRF reconstruction. The removed object is a white lamp located near the upper center of the image.