

Vulnerability of Privacy-Preserving Visual Localization Methods against Diffusion-based Attacks

Maxime Pietrantoni^{1,2,3}, Torsten Sattler², and Gabriela Csurka³

¹ Faculty of Electrical Engineering, Czech Technical University in Prague,
`{firstname.lastname}@cvut.cz`

² Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical
University in Prague

³ NAVER LABS Europe, `{firstname.lastname}@naverlabs.com`

Abstract. The rapid deployment of visual localization (VL) in augmented/virtual reality (AR/VR) and autonomous systems makes privacy-preserving localization a critical societal necessity. Existing VL systems rely on cloud-based 3D scene representation storage and transmit client side features to a remote server, exposing users to potential reconstruction attacks from intercepted privacy preserving representations. The aim of this paper is therefore to investigate what sensitive information can an adversary actually recover, by comparing multiple privacy-preserving solutions within the literature. We define privacy as the inability to recover personally identifiable information from these representations, acknowledging that general scene details do not inherently represent a privacy breach. To model adversarial behavior and to thoroughly measure a method’s degree of privacy preservation, we introduce a new privacy attack that trains a conditional diffusion model to reconstruct images from privacy-preserving representations. This reconstruction quality serves as a direct proxy for the amount of sensitive information each representation contains. Leveraging this unified attack protocol, we present the first comprehensive comparison across major families of privacy preserving VL methods. Our analysis reveals substantial differences in privacy leakage between representations and highlights limitations in current design assumptions. Code and trained checkpoints may be found here: <https://github.com/MaximePi/benchmark-privacy-inversion>.

1 Introduction

Visual localization (VL), a core component of self-driving cars [14] and autonomous robots [23], refers to the task of estimating the 6DoF camera pose from which an image was captured. VL systems require a representation of the scene in which the pose can be estimated. The representation might be a 3D Structure-from-Motion (SfM) point cloud [18,42], a database of images with known intrinsic



Fig. 1: In this paper, we benchmark the privacy of different image based representations by quantifying the amount of retrievable information they contain with a novel diffusion attack and associated privacy metrics. For example, we show reconstruction output from this attack for the same conditioning starting from different initial diffusion seeds. Here, SIFT [24], SuperPoint (SP) [6], DPT [38], and SAM2 [39] representations are used as conditioning. The ground truth image is displayed on the top left and the corresponding conditioning is displayed below the image. We observe that in contrast to segmentations, SIFT/SuperPoint display low structural variability under different seeds and yield better reconstructions than segmentations, indicating lower privacy.

and extrinsic parameters [43, 57], the weights of neural networks [1, 19], a neural radiance field [4, 27], or a 3D Gaussians-based representation [32].

With the growing adoption of cloud-based localization services in AR/VR, autonomous vehicles, and mobile applications, safeguarding user data has become a critical priority. Therefore, privacy-preserving visual localization is a fundamental societal necessity to enable scalable, secure, and ethical deployment in real-world environments. In practice, visual localization systems store 3D scene representations in clouds, while the client extracts a relevant image representation from the target image to be localized before transmitting them to the server. Protecting both the 3D map and the 2D query data is equally vital, as vulnerabilities in either could compromise user privacy and system integrity. This paper primarily focuses on the latter, framing the privacy breach as an adversary recovering information transmitted between the client and server. Note however that our attack is directly applicable to renderings of 3D models, and hence can be used to potentially recover private details also from 3D representations. Yet, in this work we primarily focus on the 2D query data and defer the analysis of vulnerabilities within 3D representations to a future study.

Following prior work [2, 3, 33, 35, 56], we define privacy as the inability to retrieve personally identifiable information (*e.g.*, faces, license plates, pictures, documents, or texture details) from the image representation. We assume that general scene attributes, including coarse geometric or semantic details, do not in themselves compromise privacy [50, 56]. Since visual localization depends on

a minimal level of discriminative information, the disclosure of broad semantic or geometric cues is expected. Privacy, however, remains subjective and context-dependent, shaped by environmental conditions, legal constraints, and individual expectations. Consequently, the central challenge lies in defining measurable notions of privacy and developing mechanisms to enforce them.

The seminal work [35] demonstrated that the high dimensional descriptors commonly used to establish the feature matches needed for downstream pose estimation or 3D reconstruction contain a substantial amount of privacy sensitive information, which can be recovered by inverting the descriptors and reconstructing the image they were extracted from. A feedforward convolutional network (CNN) is trained to regress RGB values from sparse descriptors as input, allowing faithful image reconstruction from high dimensional descriptors. However, it struggles to handle difficult scenarios where the input descriptors are sparser and noisier or when the input representation contains less information, making the reconstruction uncertain. It is therefore less suited to invert privacy-preserving representations. In contrast, we introduce a new diffusion-based privacy attack that reconstructs images by iteratively denoising directly conditioned on privacy-preserving representations. Conditioning is applied at two levels: locally, via concatenation of pixel-level features, and globally, through structural cues injected by a graph neural network. This multi-scale conditioning yields a more robust and expressive inversion framework, enabling significantly more accurate reconstructions even from sparse inputs.

In the last few years, multiple works have been proposed to tackle privacy-preserving visual localization (PPVL). These approaches can be classified into two main categories, *geometric obfuscation methods* [12, 20, 44, 47, 48], which obfuscate the geometry of the scene while keeping image representations (descriptors) intact for downstream visual localization, and *descriptor obfuscation methods* [8, 29, 31–33, 36, 50], which use image representations that contain little to no privacy sensitive information. These works, if they consider privacy at all in their evaluation, all use different protocols and make independent claims of their level of privacy. This work is the first to put both method categories to the test under a shared inversion attack, supported by a comprehensive benchmark that jointly assesses perceptual quality and the amount of semantic content recovered. In addition to the perceptual metrics used by [35] that fail to adequately capture the extent of leaked privacy-sensitive information, only partially reflecting the true level of privacy risk. We follow [34] and assess privacy also through the descriptive capacity of advanced vision-language models (VLMs). We believe that these metrics complement purely perceptual metrics to quantify privacy, as they are able to capture fine-grained details with comprehensive semantic granularity and enables a more nuanced understanding of privacy risks compared to predefined object detectors used in [33]. In addition to perceptual and VLM-based assessments, we introduce a new privacy evaluation metric that leverages the probabilistic behavior of diffusion models. Specifically, we measure reconstruction variability by initiating the model from multiple random noise

states that serves as a powerful proxy for quantifying the invariant information encoded within the privacy-preserving representation.

In summary our contributions are as follows: 1) We introduce a powerful diffusion-based inversion attack capable of reconstructing images from even sparse privacy-preserving representations. 2) We propose new privacy metrics to quantify how much private information leaks through these representations. 3) We present the first unified, comprehensive privacy evaluation protocol, enabling in-depth comparison of modern privacy-preserving visual localization methods.

2 Related Work

Privacy-preserving visual localization and mapping aim to mitigate the risk of inversion attacks, where attackers reconstruct recognizable images from point clouds, stored descriptors, or representations exchanged between a client and a server. [35] demonstrated that a set of local descriptors with their 2D positions enables recovering the original image content via an inversion attack. The method can be further extended to scene recovery from sparse 3D point clouds, suggesting that storing such point clouds poses a significant privacy risk.

To prevent such attacks and increase the degree of privacy, geometric obfuscation methods have been introduced to conceal the positions of 2D/3D points. In particular, [20, 44, 47, 48] transform 2D/3D points into 2D/3D lines, respectively. By extension, [21] lift pairs of points to lines and [26], relying on a set of spatial anchors, obtain a set of lines with non uniform direction distribution to make the inversion more challenging. Another strategy consists in swapping coordinates between random pairs of points [30] or performing partial pose estimation against distributed partials maps [11]. Such obfuscation methods have also been used in the context of SLAM [44] and SfM [12, 13].

These schemes aim at preventing inversion attacks by hiding the 2D positions, thus preventing the applicability of privacy attacks. Yet, as shown in [2, 3], point locations can often still be inferred by identifying local neighborhoods through co-occurring descriptors, thus enabling inversion attacks despite the obfuscation.

More recent PPVL methods focus on obfuscating the representations rather than the geometry. [8, 29, 36] lift descriptors to affine subspaces or to an obfuscated manifold in order to prevent direct inversion. [56] and [50] perform matching based solely on geometry, eliminating the need for image-based descriptors at the cost of reduced localization accuracy. In an orthogonal direction, [32–34] demonstrate that replacing high-dimensional descriptors with segmentation labels which contain less fine-grain level information effectively mitigates inversion attacks by introducing an ambiguity, i.e., there are many images that can equally explain the segmentation and finding the "right" one becomes hard / impossible. [31] rely on a hybrid approach by extracting descriptors from segmentations rather than original images. In this work, we propose a new shared inversion framework and evaluation protocol to assess the privacy level of most of these methods.

Conditional Diffusion Models. Diffusion models have emerged as a powerful class of generative models, enabling high-fidelity synthesis of complex data distributions by gradually adding noise to data and learning to reverse the diffusion process to generate samples [7, 16]. Subsequent works have focused on improving the sampling strategy [25, 46], developing alternative formulation such as a continuous-time framework [10], or improving the visual fidelity with CLIP-based text encoders and denoising in a latent space [40, 41].

Conditional diffusion models [16, 40] allow to convert conditioning signals into coherent and semantically meaningful data samples. Local conditioning signals such as depth, edges, or human pose aim at controlling the structure of the generated samples, while signals such as text aim at controlling the style of the generated samples. Many structures have been introduced to efficiently incorporate the conditioning [28, 52, 53]. In this work, we finetune a diffusion model [40] jointly conditioned on different representations used by privacy-preserving visual localization methods. We aim to recover the original image content and to measure the level of private data still contained in those representations.

3 A Diffusion-based Privacy Attack

First, we describe our privacy attack, which consists of a diffusion model trained to capture image distributions conditioned on a variety of dense or sparse image representations. Second, we introduce our multi-modal low rank adaptation (LoRA) finetuning scheme that allows us to further enhance training efficiency and to leverage the representational priors of large pretrained models. An overview of our pipeline is provided in Fig. 2.

3.1 Conditioning diffusion for inversion

Our privacy attack involves reconstructing an original image from its privacy-preserving representation using a diffusion model, conditioned on this representation. Diffusion models [7, 16] are generative models characterized by two key processes: a forward process, where noise is iteratively added to the training sample, and a backward process, where a network learns to denoise the noisy sample by iteratively predicting the added noise. The forward process consists in adding Gaussian noise to a sample x_0 drawn from the true distribution q , gradually transforming it into a noise-only distribution. The probability of transitioning is described by $q(x_t|x_{t-1}) = \mathcal{N}(x_t, \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I})$, where $\mathcal{N}$ denotes a Gaussian distribution and β_t is a timestep-dependent noise level. The reverse process aims at predicting x_t from x_{t-1} and it is formulated as:

$$p_\theta(x_{t-1}, x_t) = \mathcal{N}(x_{t-1}, \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad ,$$

where μ_θ and Σ_θ are parametrized by a neural network called denoising network. In practice, the denoising network predicts the noise. The training objective consists in minimizing the KL divergence between the true (unknown) and the

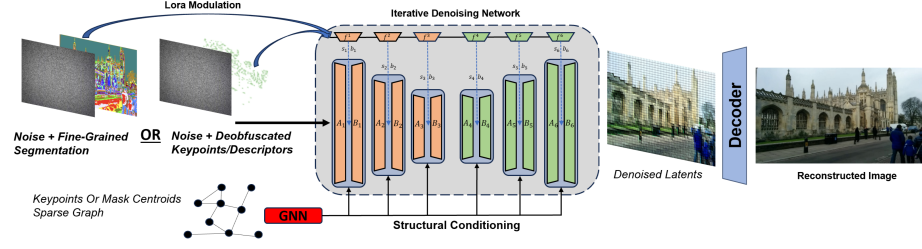


Fig. 2: Overview of our privacy attack pipeline. Given a privacy preserving representation (fine-grained segmentation or a set of keypoints, *e.g.*, obtained from deobfuscation [2], and descriptors), the network progressively denoises the noisy input tensor in a latent space before decoding it to the image associated with the representation. The network is conditioned locally on the privacy-preserving representation through concatenation and globally through cross attention by encoding the set of keypoints or mask centroids with a GNN. LoRA modulation further allows for the inversion of different types of segmentation or descriptors with a single model.

approximated reverse process distributions. This objective can be simplified by minimizing the difference between predicted noise and the noise added during the forward process $L = E_{x_0, \epsilon \sim \mathcal{N}(0, \mathbf{I})} \|\epsilon - \epsilon_\theta(x_t, t)\|$ with $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon$, where $\alpha_t = \prod_{s=1}^t (1 - \beta_s)$ is a closed-form reparameterization used to simplify the training objective.

In this work, we elect to condition the reverse process on a privacy-preserving representation c_{pp} . The reverse process therefore becomes:

$$p_\theta(x_{t-1}|x_t, c_{pp}) = \mathcal{N}(x_{t-1}, \mu_\theta(x_t, t, c_{pp}), \Sigma_\theta(x_t, t, c_{pp})) ,$$

and the resulting training objective for the inversion network is:

$$L = E_{x_0, \epsilon \sim \mathcal{N}(0, \mathbf{I})} \|\epsilon - \epsilon_\theta(x_t, c, t)\| . \quad (1)$$

For each time step t of the schedule, the conditioning is performed locally on the pixel-level by concatenating, on the channel dimension, the privacy-preserving representation to the noisy sample x_t . The resulting concatenated tensor is fed to the denoising network, which extracts the information contained within the privacy-preserving representation to progressively denoise the sample.

To strengthen the conditioning signal and ensure maximal fidelity in the reverse process, we furthermore add a global conditioning through cross-attention layers in the denoising network. Given a set of N keypoint locations (for sparse keypoint-based representations) or N centroid center locations (for segmentations), we divide the image space into S sub-windows. Within each sub-window s , we construct a sparse graph using the keypoints/centers, and apply a graph attention network (GNN) to compress structural information, yielding an embedding g_s that captures the compositionality of the privacy-preserving representation. The S embeddings are concatenated and passed to the cross-attention layers of the denoising network (see Fig. 2).

3.2 Multi-modal low rank adaptation

To leverage the semantic and structural priors contained within a pretrained text-to-image diffusion foundation model [40], and to compensate for the scarcity of training data associated with privacy-preserving representations, we use Low-Rank Adaptation (LoRA) [17]. LoRA is a parameter-efficient fine-tuning technique based on the hypothesis that weight updates during fine-tuning have low intrinsic dimensionality, enabling the adaptation of models with a large number of parameters. Formally, weights matrices of the original model are frozen and a new set of low-rank matrices are instead introduced. A weight matrix $\mathbf{W} \in \mathbb{R}^{dk}$ belonging to the original model is adapted as follows: $\mathbf{W}' = \mathbf{W} + \frac{\alpha}{r} \mathbf{B} \mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{dr}$ and $\mathbf{B} \in \mathbb{R}^{rk}$ are the low rank matrices, with $r \ll \min(d, k)$ being the target rank, and α controlling the adaptation strength. The adaptation happens on a per layer basis; two low rank matrices are learned and stored per layer.

To further improve both the robustness and the efficiency of our inversion model, instead of training a set of LoRA weights for each representation, we train a single set of LoRA weights per group of similar representations – sparse features or dense segmentations – and modulate the low-rank adaptation accordingly. We hypothesize that representations within the same group are sufficiently similar to enable the geometric and semantic information required for inversion to be captured by shared low-rank matrices. This shared representation leverages similarities between representations, potentially yielding synergies, while the signal is simply modulated to account for the differences between individual representations. Inspired by [49], we perform modulation within the LoRA low-rank subspace to enable regularization and prioritize critical details essential for inversion.

Concretely, given a group of C representations $\{c_i\}_{i=1}^C$, a learnable embedding h_c^l is assigned to each representation c for every layer l , which serves as input to two small MLPs predicting a scale $s = f_s^l(h_c^l)$ and a bias $b = f_b^l(h_c^l)$. Given an input x_c conditioned on c and the weight matrix $\mathbf{W}$ associated to a layer l , the modulation for the layer is then effectively described by:

$$\mathbf{W}' x_c = \mathbf{W} x_c + \frac{\alpha}{r} (\mathbf{B} + (f_s^l(h_c) * \mathbf{A} x_c + f_b^l(h_c))) ,$$

yielding an efficient model that can invert multiple representations with limited overhead and efficient training.

4 Privacy Evaluation Protocol

To assess the privacy-level of a given obfuscated representation, when relevant, we first apply the deobfuscation method from [2] to recover approximated key-point positions either from randomized coordinates or lines. Then, we apply our conditional diffusion-based model to reconstruct the associated image. Note that the inversion model extracts information contained in the input representation by mapping the latent representation into an interpretable RGB space, enabling both visual and quantitative comparison with the original image content.

We argue that better reconstructions correspond to greater privacy leakage, while worse reconstructions imply stronger privacy preservation. Intuitively, when fine details are accurately reproduced, more of the original information, especially more personally identifiable ones, are exposed. For this reason, we evaluate the privacy level by analyzing the image reconstruction quality using multiple perceptual and semantic measures. Concretely, we use the following existing and new metrics (see more details in the Supplementary):

- **Perceptual similarity** is measured via the LPIPS [54] and FID [15] metrics often used to evaluate perceived visual fidelity. LPIPS (Learned Perceptual Image Patch Similarity) relies on a deep network that was trained to align closely with human visual judgment. FID (Fréchet Inception Distance) quantifies the realism and diversity of the reconstructed images.
- **Structural similarity** is measured via SSIM (Structural Similarity Index Measure) [51], which measures image degradation as a perceived change in structural information, while also incorporating important perceptual phenomena, including both luminance masking and contrast masking terms.
- **Reconstructed Diversity (RD)** is defined as the pairwise SSIM or LPIPS, – denoted by RD(S) and RD(L) – between a set of images reconstructed with a diffusion model that uses the same conditioning but different noise seeds (see also Fig. 1). Low diversity between the reconstructed images, *i.e.*, high RD(S) or low RD(L), indicates that the conditioning signal strongly constrains the reconstruction, yielding similar image reconstruction independently of the seed. In contrast, high diversity, *i.e.*, low RD(S) or high RD(L), suggests that the model must infer or hallucinate varying content/details due to weak guidance, meaning the representation is more privacy-preserving. RD complements existing metrics by capturing privacy in a probabilistic sense, reflecting uncertainty in the reconstruction rather than relying solely on deterministic measures.

Perceptual metrics provide a global assessment of reconstruction quality. However, users may be more concerned with how accurately particular objects or areas are identifiable. To capture such privacy-relevant aspects, we introduce three new metrics that focus specifically on sensitive details in the image.

- **Scene Recognition (SR)**: captures to extent the reconstructed image is correctly identified as depicting the same location as the ground-truth image. It evaluates whether the reconstruction preserves enough discriminative visual cues to distinguish the original location from visually similar scenes. To compute such scores we use a strong VLM to generate detailed captions for both images, prompted to focus on salient scene elements. We then measure to which extent the two descriptions appear to depict the same place by measuring the similarity between descriptions.
- **Object Recovery (OR)**: evaluates object-level fidelity by computing the recall of distinctive objects preserved between the ground-truth image and its reconstruction. A VLM first extracts a list of salient, non-trivial objects (ignoring generic elements like walls, floors, sky, or buildings) from each

image. OR then measures the fraction of objects from the original list that are successfully recovered in the reconstructed one.

- **Texture Recovery (TR)**: focuses on whether subtle visual details—like text, specific textures, or characteristic patterns—are preserved during reconstruction. By comparing detailed descriptions generated by a VLM for both images, TR reports the proportion of such fine-grained elements that remain recoverable.

More details, VLM descriptions, and qualitative examples are in the Supp. Mat.

5 Experimental Results

We begin by providing training details, followed by a preliminary analysis of image recovery using our diffusion-based inversion from segmentations and sparse descriptors. We then compare the privacy-preservation level of multiple PPVL methods and conclude by demonstrating that our diffusion-based attack surpasses the FFConv baseline of [35].

Training and evaluation details. Our models are trained on a combination of two large-scale datasets, MegaDepth [22] and ScanNet-v2 [5], accounting over two million training images and covering a wide variety of outdoor and indoor scenes. Four inversion models are trained: one for sparse descriptors (using SIFT [24], SP (SuperPoint) [6], and XFeat [37] features), one for segmentations obtained from SegLoc [33], GSFF [32], DPT [38], SAM2 [39] and PpNeSF [34], one for the Adversarial Affine Subspace Embedding (AASE) obfuscated descriptors [9] (applied to SIFT and SP), and, finally one for the NinjaDesc obfuscated descriptors [29] (applied to SIFT). All models use a LoRA of rank $r = 64$ with $\alpha = 24$ (for more details see the Supplementary).

After training, we apply our inversion models on the two most commonly used real-world datasets in visual localization literature, *i.e.*, 7Scenes [45] and Cambridge Landmarks [19] and evaluated the information contained in the aforementioned image representations. Importantly, none of these evaluation datasets were seen by the inversion model during training.

5.1 Local features *vs.* segmentations

We start this experimental section with a more general study aimed at evaluating the amount of information encoded in general in segmentations *vs.* local descriptors by learning to reconstruct images from these representations with our model. Then, we evaluate, on the test set, the faithfulness of images reconstructed from semantic segmentations generated by a DPT [38], trained on ADE20k [55], from segmentation masks obtained with SAM2 [39], and from a set of different sparse descriptors extracted from images such as SIFT [24], SP [6] and XFeat [37]. Results for both the 7Scenes dataset and Cambridge Landmarks are reported in Tab. 1 and visual examples are shown in Fig. 1. Metrics are averaged over the images within a scene, then averaged across scenes. Full per-scene results are provided in the Supplementary.

Table 1: Inversion results on Cambridge Landmarks and 7-Scenes datasets comparing conditioning on sparse local features (SDesc.) *vs.* segmentations (Seg.). Higher FID, LPIPS, RD(LPIPS), and lower SSIM, RD(SSIM), SR, OR, TR indicate lower quality and more uncertain reconstructions, implying a higher degree of privacy. **Green bold** indicates the highest level of privacy, **red bold** the lowest level of privacy.

Method		LPIPS $\uparrow$	SSIM $\downarrow$	FID $\uparrow$	RD(L) $\uparrow$	RD(S) $\downarrow$	SR $\downarrow$	OR $\downarrow$	TR $\downarrow$
7-Scenes	Seg. SDesc. SIFT	0.33	0.63	125.9	0.22	0.49	0.47	0.63	0.67
	Seg. SDesc. SP	0.20	0.72	96.8	0.11	0.71	0.55	0.73	0.68
	Seg. SDesc. XFeat	0.25	0.73	104.7	0.11	0.83	0.57	0.65	0.72
	Seg. SDesc. DPT	0.59	0.44	222.5	0.52	0.26	0.11	0.34	0.27
	Seg. SDesc. SAM2	0.50	0.46	177.9	0.42	0.33	0.26	0.48	0.47
	Seg. Seg. SIFT	0.27	0.48	71.8	0.18	0.58	0.63	0.68	0.68
Cambridge	Seg. SDesc. SP	0.20	0.55	62.1	0.14	0.68	0.70	0.80	0.72
	Seg. SDesc. XFeat	0.28	0.57	79.6	0.15	0.78	0.66	0.78	0.73
	Seg. SDesc. DPT	0.58	0.29	148	0.50	0.25	0.43	0.38	0.37
	Seg. Seg. SAM2	0.48	0.29	94.8	0.41	0.27	0.46	0.52	0.43
	Seg. Seg. SAM2	0.48	0.29	94.8	0.41	0.27	0.46	0.52	0.43

Sparse local descriptors encode sufficient information to enable high-fidelity image reconstruction. This is particular true for learned descriptors such as XFeat and SP, as shown by all metrics in Tab. 1. This suggests that even a sparse set of local descriptors (between 500 and 1,000 keypoints per image) retains sufficient information to enable high-fidelity image reconstruction. Low perceptual differences (LPIPS, FID) and high structural similarity (SSIM) indicate that the reconstructions remain perceptually and statistically close to the original images. High SR/OR/TR scores demonstrate that the VLM can still identify meaningful semantic cues. The RD scores further support the qualitative observations in Fig. 1, *i.e.*, unlike segmentations, reconstructions from local features exhibit low structural variability across different random seeds.

Segmentations are harder to invert. Segmentations are inherently ambiguous to invert due to the wide range of potential visual content that a single mask can encode, *i.e.*, multiple visually dissimilar images can be obtained from the same segmentation (*e.g.*, see Fig. 1). This ambiguity is reflected in the RD scores on both datasets, where the inversion model conditioned on segmentations produces wildly different images for the same conditioning signal and different seeds yielding high pairwise LPIPS and low pairwise SSIM. In contrast, sparse descriptors produce stable (up to slight variations in color) and accurate outputs.

Fine segmentation granularity helps inversion. SAM2, despite having per-image masks not necessarily generalizing across images, leads to better image reconstructions than DPT, which segments coarser semantic concepts (*c.f.* Fig. 1). This confirms the observation in [32, 33] that the segmentation granularity can be set as a trade-off between localization accuracy and privacy preservation.

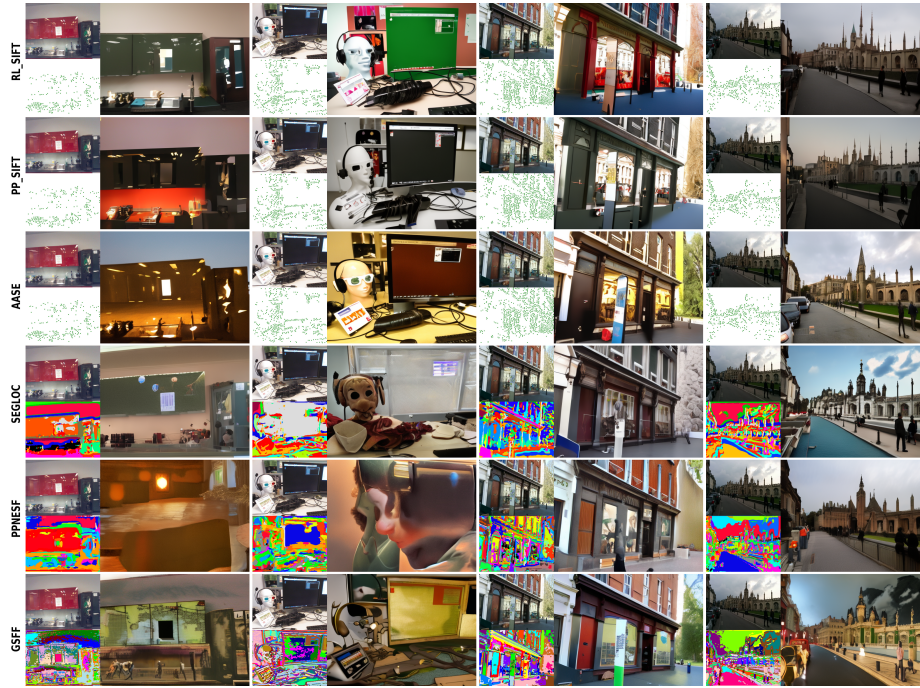


Fig. 3: Reconstructions from various PPVL representations from top to bottom: RandLine (SIFT) [48], CoordPerm (SIFT) [30], AAES [9], PpNeSF [34] and GSFF [32]. The ground truth image is displayed on the top left and the conditioning is displayed on the bottom left. Feature-based methods can capture both coarse structure and fine details (see reflections or details on the screen) while segmentations only recover detailed scene structure when it is visually obvious in the segmentations (such as building facades or repetitive patterns). In many cases, however, particularly for indoor scenes, the global layout is entirely hallucinated (GSFF) or appears highly surreal (PpNeSF).

5.2 Inversion results from PPVL representations

In this section, our aim is to compare the vulnerability of the representations used by different PPVL methods. In particular, we consider two sparse feature-based obfuscation methods, Random Lines (RandLine) [47] and Coordinate Permutation (CoordPerm) [30], three segmentation-based methods, Segloc [33], GSFF [32], and PpNeSF [34], the AASE-based descriptor obfuscation from [8], and NinjaDesc [29]. For the geometric obfuscation methods, we first deobfuscate the keypoints using [2], then we use the descriptors of the recovered keypoints to condition the inversion model. For AASE [8], obfuscated descriptors are embedded in an image space sparse tensor and directly fed into the inversion model. Similarly for [29], we embed obfuscated descriptors from the NinjaDesc obfuscation network in an image based space sparse tensor which is subsequently fed in the inversion model. For the segmentation methods, we use the dense seg-

Table 2: Inversion results on the Cambridge Landmarks and 7-Scenes datasets comparing inversion results conditioned on PPVL representations - geometric obfuscation (GObf.), descriptor obfuscation (DObf.), and segmentation-based obfuscation (Seg.). Higher FID, LPIPS, RD(L), and lower SSIM, RD(S), SR, OR, TR indicate lower quality and more uncertain reconstruction, implying a higher degree of privacy. **Green bold** indicates the highest level of privacy, **red bold** the lowest level of privacy.

Method		LPIPS $\uparrow$	SSIM $\downarrow$	FID $\uparrow$	RD(L) $\uparrow$	RD(S) $\downarrow$	SR $\downarrow$	OR $\downarrow$	TR $\downarrow$	
7-Scenes	GObf.	RandLine (SIFT)	0.45	0.57	154.7	0.32	0.32	0.32	0.46	0.43
		RandLine (SP)	0.39	0.61	169.0	0.24	0.46	0.30	0.39	0.40
		CoordPerm. (SIFT)	0.47	0.56	170.3	0.35	0.31	0.26	0.39	0.43
		CoordPerm. (SP)	0.42	0.60	173.6	0.25	0.44	0.26	0.43	0.44
	DObf.	AASE (SIFT)	0.39	0.58	151.3	0.30	0.39	0.36	0.42	0.39
		AASE (SP)	0.27	0.65	131.2	0.18	0.59	0.40	0.44	0.50
		NinjaDesc(SIFT)	0.37	0.61	131.3	0.27	0.44	0.42	0.54	0.48
		Seg.	PpNeSF	0.56	0.54	270.1	0.46	0.23	0.09	0.30
	GSFF		0.48	0.41	190.5	0.40	0.12	0.19	0.35	0.34
	SegLoc	0.43	0.56	174.9	0.35	0.35	0.22	0.41	0.46	
Cambridge	GObf.	RandLine (SIFT)	0.33	0.42	72.47	0.21	0.51	0.60	0.50	0.36
		RandLine (SP)	0.29	0.44	69.10	0.19	0.53	0.59	0.52	0.43
		CoordPerm (SIFT)	0.34	0.42	75.64	0.22	0.50	0.56	0.45	0.46
		CoordPerm (SP)	0.32	0.43	73.43	0.20	0.53	0.56	0.47	0.48
	DObf.	AASE (SIFT)	0.35	0.40	75.54	0.26	0.48	0.66	0.48	0.46
		AASE (SP)	0.25	0.48	65.03	0.24	0.52	0.66	0.49	0.48
		NinjaDesc(SIFT)	0.28	0.46	73.4	0.21	0.52	0.60	0.51	0.49
		Seg.	PpNeSF	0.44	0.37	92.64	0.34	0.32	0.48	0.38
	GSFF		0.43	0.35	91.31	0.37	0.21	0.51	0.35	0.37
	SegLoc	0.39	0.40	91.10	0.31	0.47	0.53	0.48	0.47	

mentation of the query image directly as conditioning signal for the inversion model.

Results for 7Scenes and Cambridge Landmarks are reported in Tab. 2. The evaluation protocol and metrics are the same as in Sec. 5.1. Fig. 3 provides qualitative comparison of a few images reconstructed from different PPVL representations. Again, note the high content fidelity obtained from sparse descriptors.

Segmentation-based PPVL are more robust against the privacy attack than descriptors obfuscation methods. The vulnerability of local features observed in Sec. 5.1 also holds when we evaluate the geometric obfuscation PPVL representations. For both RandLine and CoordPerm, the perceptual and semantic recovery metrics in Tab. 2 indicate that the inversion model is able to recover more image content from the deobfuscated keypoints than from the segmentation based SegLoc, GSFF, and PpNeSF. The AASE representations as well as NinjaDesc yield higher reconstruction quality, hence lower privacy preserving scores. The variability analysis (RD scores) also implies that segmentations result in

Table 3: Localization accuracy *vs* privacy vulnerability on the Cambridge Landmarks dataset. We report privacy vulnerability metrics from Tab. 2. For localization, we provide median translation error (MTE) in meter and median rotation error (MRE) in angle (°) averaged over all scenes in the dataset (reported from the literature).

PPVL methods	from	MTE ↓	MRE ↓	LPIPS ↑	RD(S) ↓	OR ↓	TR ↓
CoordPerm. (SP)	[30]	0.12	0.18	0.32	0.53	0.47	0.48
RandLine (SIFT)	[26]	0.07	0.17	0.33	0.51	0.50	0.36
SegLoc	[33]	0.22	0.39	0.39	0.47	0.49	0.47
GSFF	[32]	0.17	0.41	0.43	0.21	0.35	0.37
PpNeSF	[34]	0.14	0.36	0.44	0.32	0.38	0.36

more uncertain reconstructions. Finally, as VLM metrics in Tab. 2 shows that the highest privacy-preserving scores were obtained with GSFF and PpNeSF although some limited content may be recovered depending on the situation.

Geometric Obfuscation (GObf.) *vs* Descriptor Obfuscation (DObf.).

The inversion network is able to directly learn a deobfuscation mapping to reconstruct images from obfuscated descriptors [8, 29]. On the contrary, geometric obfuscation requires an initial approximate geometry recovery step before learning an inversion mapping. As Tab. 2 shows, descriptor obfuscation tends to leak more privacy preserving information than geometric obfuscation. Indeed, in most cases, the geometric recovery step for geometric obfuscation methods is more prone to failure than learning a deobfuscation mapping through our inversion network.

Robustness of the inversion model to noisy keypoint locations. Despite introducing a median error of 5 to 10 pixels for the recovered keypoint positions due to the permutations or random line recovery process [2], this level of noise is still sufficient for the inversion model to reconstruct images with a satisfactory degree of fidelity emphasizing the robustness of our inversion model.

Privacy *vs.* localization accuracy. Tab. 3 compares localization accuracy and privacy preservation on the Cambridge Landmark dataset for several PPVL methods. As expected, there is a trade-off between privacy and localization accuracy. Feature-based obfuscation yields more discriminative representations and thus better localization accuracy, but increases vulnerability to inversion attacks. In contrast, segmentation-based methods offer stronger privacy level, as discussed above, at the cost of a modest reduction in accuracy. These number are just illustrative, as this tradeoff heavily depends on the localization setup and the amount of information encoded in the input representation. For example, one can increase accuracy of segmentation methods by increasing the granularity of segmentations at the cost of privacy or inversely.

Table 4: Feed-forward inversion models compared to our Diffusion Inversion based model on Cambridge Landmark dataset. **Best** and **worse** inversion results.

Conditioning	FFConv	FFTrans	FFUnetSA	Ours
	SSIM($\uparrow$) / LPIPS($\downarrow$) / FID($\downarrow$)			
SegLoc	0.37/ 0.51 /230	0.35/0.51/245	0.37/0.47/189	0.40/0.39/91
RandLine (SIFT)	0.36/0.45/158	0.35/0.49/210	0.39/0.42/144	0.42/0.33/73
CoordPerm (SIFT)	0.33 /0.47/160	0.34/ 0.49/220	0.37/0.43/147	0.42/0.34/76
	SR($\uparrow$) / OR($\uparrow$) / TR($\uparrow$)			
SegLoc	0.48 /0.35/0.21	0.48/0.28/0.10	0.50/0.38/0.21	0.53/0.48/0.47
RandLine (SIFT)	0.56/0.35/0.16	0.46/0.15/0.09	0.54/0.36/0.18	0.60/0.50/0.36
CoordPerm (SIFT)	0.50/0.33/ 0.10	0.48/0.18/0.11	0.50/0.33/0.14	0.56/0.45/0.46

**Fig. 4:** Groundtruth (*col. 1/4*), reconstructions from NinjaDesc SIFT by FFUnetSA (*col. 2/5*) and by our method (*col. 3/6*).

5.3 Comparison to feed forward inversion methods

In this section, we compare our diffusion-based inversion model to an implementation of a convolutional inversion model **FFConv** [35] that takes as input the same local conditioning as our diffusion-based model and is trained (on the same training dataset) to reconstruct images in a discriminative feed-forward manner. In addition we consider two additional feed-forward baselines trained in similar fashion as our model: a stack of transformer layers with a convolutional decoder (**FFTrans**) and a UNet model with self-attention layers (**FFUnetSA**).

In Tab. 4 we show inversion results on Cambridge Landmarks in terms of reconstruction quality of the inversion models. Therefore higher SSIM and lower LPIPS/FID indicate that the inversion model is more effective (arrows are inverted compared to former experiments evaluating privacy). They show that FF approaches, which can only modify the input conditioning signal, struggle to recover more than global structural cues. While our diffusion model is able to model complex conditional distributions by directly learning a mapping from noise to image space, recovering finer details (see Fig. 4 and Fig. 5).

In addition in Tab. 5 we evaluate both FFConv [35] and our model’s robustness to the descriptor sparsity level. We can see that increasing sparsity results, in both cases, in a reconstruction quality drop. However our diffusion based reconstruction method consistently and significantly outperforms FFConv on all three metrics. This is also confirmed by the qualitative comparisons in Fig. 5 where feed-forward reconstruction from sparse inputs are much blurrier than

Table 5: Comparison of FFConv [35] with our diffusion-based approach on Cambridge Landmarks in terms of reconstruction metrics (SSIM $\uparrow$ / LPIPS $\downarrow$ / FID $\downarrow$) with varying level of sparsity for SIFT keypoints.

	FFConv	Ours
High Sparsity (avg. 500 kpts/img)	0.45 / 0.44 / 164	0.49 / 0.28 / 74
Medium Sparsity (avg. 1K kpts/img)	0.43 / 0.50 / 198	0.46 / 0.36 / 97
Low Sparsity (avg. 2K kpts/img)	0.38 / 0.59 / 248	0.41 / 0.49 / 130



Fig. 5: Qualitative comparison between our inversion model (left) *versus* FFConv (right) across varying keypoint sparsity. Top: dense $\rightarrow$ sparse. Bottom: using noisy de-obfuscated keypoints.

our model. Furthermore, the results in Tab. 4 clearly show that our diffusion model is also better suited than FFConv to invert noisy keypoints obtained by deobfuscating keypoints from RandLine and CoordPerm obfuscation schemes.

Acknowledgements. Maxime and Torsten received funding from NAVER LABS Europe.

6 Conclusion

To address rising privacy concerns in visual localization, we introduce a unified evaluation framework for privacy-preserving visual localization, including a new diffusion-based attack that reconstructs images conditioned locally and globally on privacy-preserving representations. We further extend traditional image-quality privacy metrics with new complementary measures that score semantic recovery, and reconstruction diversity, to better quantify how much information each representation leaks hence to valuate its privacy level. Our benchmark provides the first comprehensive comparison across major families of privacy-preserving methods and shows that segmentations offer stronger privacy than geometric obfuscation with only slight drop in localization accuracy. The latter depends on sparse descriptors that still encode fine-grained sensitive details, whereas segmentations are more ambiguous and inherently less informative, though at the cost of slightly reduced localization accuracy. We hope that this work will encourage ongoing research on the importance of privacy in the development and learning of representations for visual localization.

References

1. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: DSAC - Differentiable RANSAC for Camera Localization. In: CVPR (2017)
2. Chelani, K., Benbihi, A., Kahl, F., Sattler, T., Kukeleva, Z.: Obfuscation Based Privacy Preserving Representations are Recoverable Using Neighborhood Information. In: 3DV (2025)
3. Chelani, K., Kahl, F., Sattler, T.: How Privacy-Preserving Are Line Clouds? Recovering Scene Details From 3D Lines. In: CVPR (2021)
4. Chen, S., Bhargat, Y., Li, X., Bian, J., Li, K., Wang, Z., Prisacariu, V.A.: Neural Refinement for Absolute Pose Regression with Feature Synthesis. In: CVPR (2024)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In: CVPR (2017)
6. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-Supervised Interest Point Detection and Description. In: CVPR Workshops (2018)
7. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. In: NeurIPS (2021)
8. Dusmanu, M., Schonberger, J.L., Sinha, S.N., Pollefeys, M.: Privacy-preserving image features via adversarial affine subspace embeddings. In: CVPR (2021)
9. Dusmanu, M., Schönberger, J., Sinha, S., Pollefeys, M.: Privacy-Preserving Image Features via Adversarial Affine Subspace Embeddings. In: CVPR (2021)
10. Feng, B.T., Smith, J., Rubinstein, M., Chang, H., Bouman, K.L., Freeman, W.T.: Score-based diffusion models as principled priors for inverse imaging. In: ICCV (2023)
11. Geppert, M., Larsson, V., Schönberger, J.L., Pollefeys, M.: Privacy Preserving Partial Localization. In: CVPR (2022)
12. Geppert, M., Larsson, V., Speciale, P., Schönberger, J., Pollefeys, M.: Privacy Preserving Structure-from-Motion. In: ECCV (2020)
13. Geppert, M., Larsson, V., Speciale, P., Schonberger, J.L., Pollefeys, M.: Privacy Preserving Localization and Mapping from Uncalibrated Cameras. In: CVPR (2021)
14. Heng, L., Choi, B., Cui, Z., Geppert, M., Hu, S., Kuan, B., Liu, P., Nguyen, R., Yeo, Y., Geiger, A., Lee, G., Pollefeys, M., Sattler, T.: Project AutoVision: Localization and 3D Scene Perception for an Autonomous Vehicle with a Multi-Camera System. In: ICRA (2019)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: NeurIPS (2017)
16. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: NeurIPS (2020)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LORA: Low-rank adaptation of large language models. ICLR 1(2), 3 (2022)
18. Humenberger, M., Cabon, Y., Pion, N., Weinzaepfel, P., Lee, D., Guérin, N., Sattler, T., Csurka, G.: Investigating the Role of Image Retrieval for Visual Localization. IJCV **130**(7), 1811–1836 (2022)
19. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: a Convolutional Network for Real-Time 6-DOF Camera Relocalization. In: ICCV (2015)

20. Kim, J., Moon, H., Hong, J.H.: Revisiting geometric obfuscation with dual convergent lines for privacy-preserving image queries in visual localization. In: CVPR (2026)
21. Lee, C., Kim, J., Yun, C., Hong, J.H.: Paired-Point Lifting for Enhanced Privacy-Preserving Visual Localization. In: CVPR (2023)
22. Li, Z., Snavely, N.: MegaDepth: Learning Single-View Depth Prediction from Internet Photo. In: CVPR (2018)
23. Lim, H., Sinha, S., Cohen, M., Uyttendaele, M., Kim, H.: Real-time Monocular Image-based 6-DoF Localization. *IJRR* **34**(4–5), 476–492 (2015)
24. Lowe, D.G.: Object Recognition from Local Scale-invariant Features. In: ICCV (1999)
25. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In: NeurIPS (2022)
26. Moon, H., Lee, C., Hong, J.H.: Efficient Privacy-Preserving Visual Localization Using 3D Ray Clouds. In: CVPR (2024)
27. Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., e La Fortelle, A.: LENS: Localization Enhanced by NeRF Synthesis. In: CoRL (2022)
28. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. In: AAAI (2024)
29. Ng, T., Kim, H.J., Lee, V.T., DeTone, D., Yang, T.Y., Shen, T., Ilg, E., Balntas, V., Mikolajczyk, K., Sweeney, C.: NinjaDesc: content-concealing visual descriptors via adversarial learning. In: CVPR (2022)
30. Pan, L., Schönberger, J.L., Larsson, V., Pollefeys, M.: Privacy Preserving Localization via Coordinate Permutations. In: ICCV (2023)
31. Panek, V., Beliansky, P., Kukeleva, Z., Sattler, T.: Privacy-preserving structureless visual localization via image obfuscation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 117–128 (June 2026)
32. Pietrantoni, M., Csurka, G., Sattler, T.: Gaussian Splatting Feature Fields for (Privacy-Preserving) Visual Localization. In: CVPR (2025)
33. Pietrantoni, M., Humenberger, M., Sattler, T., Csurka, G.: SegLoc: Learning Segmentation-Based Representations for Privacy-Preserving Visual Localization. In: CVPR (2023)
34. Pietrantoni, M., Humenberger, M., Sattler, T., Csurka, G.: Can we make NeRF-based visual localization privacy-preserving? In: CVPR Workshops (2025)
35. Pittaluga, F., Koppal, S.J., Kang, S.B., Sinha, S.N.: Revealing Scenes by Inverting Structure from Motion Reconstructions. In: CVPR (2019)
36. Pittaluga, F., Zhuang, B.: LDP-Feat: Image Features with Local Differential Privacy. In: ICCV (2023)
37. Potje, G., Cadar, F., Araujo, A., Martins, R., Nascimento, E.R.: XFeat: Accelerated Features for Lightweight Image Matching. In: CVPR (2024)
38. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision Transformers for Dense Prediction. In: ICCV (2021)
39. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. *arxiv:2408.00714* (2024)
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: CVPR (2022)

41. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In: *NeurIPS* (2022)
42. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In: *CVPR* (2019)
43. Sattler, T., Zhou, Q., Pollefeys, M., Leal-Taixé, L.: Understanding the Limitations of CNN-based Absolute Camera Pose Regression. In: *CVPR* (2019)
44. Shibuya, M., Sumikura, S., Sakurada, K.: Privacy Preserving Visual SLAM. In: *ECCV* (2020)
45. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In: *CVPR* (2013)
46. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. *arxiv:2010.02502* (2020)
47. Speciale, P., Schönberger, J.L., Kang, S.B., Sinha, S.N., Pollefeys, M.: Privacy Preserving Image-Based Localization. In: *CVPR* (2019)
48. Speciale, P., Schönberger, J., Sinha, S., Pollefeys, M.: Privacy Preserving Image Queries for Camera Localization. In: *ICCV* (2019)
49. Stracke, N., Baumann, S.A., Susskind, J., Bautista, M.A., Ommer, B.: CTRLorAL-ter: Conditional LoRAadapter for Efficient 0-Shot Control & Altering of T2I Models. In: *ECCV* (2024)
50. Wang, S., Kannala, J., Barath, D.: DGC-GNN: Descriptor-free Geometric-Color Graph Neural Network for 2D-3D Matching. In: *CVPR* (2024)
51. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. *TIP* **13**(4), 600–612 (2004)
52. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arxiv:2308.06721* (2023)
53. Zhang, L., Rao, A., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. In: *ICCV* (2023)
54. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: *CVPR* (2018)
55. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic Understanding of Scenes through the ADE20K Dataset. *IJCV* **127**(3), 302–321 (2019)
56. Zhou, Q., Agostinho, S., Osep, A., Leal-Taixé, L.: Is Geometry Enough for Matching in Visual Localization? In: *ECCV* (2022)
57. Zhou, Q., Sattler, T., Pollefeys, M., Leal-Taixé, L.: To Learn or not to Learn: Visual Localization from Essential Matrices. In: *ICRA* (2020)