

Adversarial Exploitation of Data Diversity Improves Visual Localization

Sihang Li* Siqu Tan* Bowen Chang Jing Zhang Chen Feng[✉] Yiming Li[✉]

New York University

{sihangli, tan.k, cfeng, yimingli}@nyu.edu

<https://ai4ce.github.io/RAP>

Abstract

Visual localization, which estimates a camera’s pose within a known scene, is a fundamental capability for autonomous systems. While absolute pose regression (APR) methods have shown promise for efficient inference, they often struggle with generalization. Recent approaches attempt to address this through data augmentation with varied view-points, yet they overlook a critical factor: appearance diversity. In this work, we identify appearance variation as the key to robust localization. Specifically, we first lift real 2D images into 3D Gaussian Splats with varying appearance and deblurring ability, enabling the synthesis of diverse training data that varies not just in poses but also in environmental conditions such as lighting and weather. To fully unleash the potential of the appearance-diverse data, we build a two-branch joint training pipeline with an adversarial discriminator to bridge the syn-to-real gap. Extensive experiments demonstrate that our approach significantly outperforms state-of-the-art methods, reducing translation and rotation errors by 50% and 33% on indoor datasets, and 38% and 44% on outdoor datasets. Most notably, our method shows remarkable robustness in dynamic driving scenarios under varying weather conditions and in day-to-night scenarios, where previous APR methods fail.

1. Introduction

Visual localization, the task of calculating a 6-DoF camera pose—its translation and rotation—based on a query image within a given environment, is essential for various applications, including robotics [1], autonomous vehicles [18], and virtual reality [11]. Besides traditional geometry-based approaches, recent learning-based visual localization methods adopt absolute pose regression (APR) [5, 8, 24, 53], scene coordinate regression (SCR) [3, 4, 43, 61], or post pose refinement (PPR) [9, 19, 33, 43, 59, 64, 66]. SCR methods focus on learning-based 2D-3D correspondences followed by subsequent Perspective-n-Point (PnP) for pose estimation.

*Equal contribution. [✉]Corresponding author.

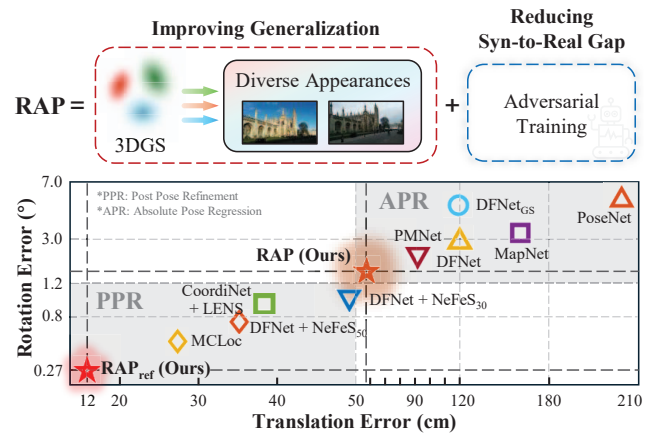


Figure 1. We propose RAP, a novel pipeline to train robust APR models. We lift real-world 2D images into 3D Gaussian Splats [25] to synthesize images with *diverse appearances and poses*, improving model generalizability. We also introduce an adversarial discriminator, mitigating the syn-to-real gap to learn robust features. Together, we achieve state-of-the-art performance.

PPR methods heavily rely on a pose prior, usually obtained from image retrieval, followed by iterative refinement. In contrast, APR methods employ a supervised framework to train a regression neural network on image-pose pairs, enabling direct pose estimation during inference. APR offers faster runtime and lower error in challenging scenes with sparse views, significant lighting variations, or numerous dynamic objects, making it a promising method for ensuring robustness in real-world applications.

Despite promises, there is a performance gap in localization accuracy between APR and other methods. A well-known pivotal work [49] attributes this to APR performing image-based memorization, i.e., retrieving poses seen during training. Driven by this crucial finding, to improve such memorization while avoiding the need for denser real-world training samples, recent methods leverage Neural Radiance Fields (NeRF) [40] to synthesize additional posed images for APR training [8, 30, 42]. LENS [42] tried to employ appearance perturbation using NeRF-W [39], but found the

improvements minimal [42]. As appearance augmentation is common in other learning tasks, why does it fail in APR?

We hypothesize a learning gap: Limitations in previous training pipelines prevented the effective use of diverse data to boost performance. Artifacts always exist in images rendered by common novel view synthesis (NVS) methods, which might disturb the model feature space. Inspired by generative adversarial networks (GAN) [20], where a discriminator is trained to distinguish between real and generated samples, we propose adversarial training for APR, designing a discriminator to align the features of synthetic and real images, thereby reducing the syn-to-real domain gap and mitigating the impact of rendering artifacts. Augmentation quality also matters. To efficiently synthesize diverse high-quality images with *controllable* varying appearance, we extend the vanilla 3D Gaussian Splatting (3DGS) [25] to appearance-varying 3DGS with deblurring ability.

These form our two-branch joint training framework for robust absolute pose regression (RAP). The first branch coarsely trains our Transformer-based pose regressor with both real data and data synthesized at the original real pose, together with an adversarial discriminator to reduce the syn-to-real domain gap. The second branch progressively generates randomly perturbed poses and appearances, providing additional supervision to the same APR Transformer. Through extensive experiments, we demonstrate that exploiting data diversity using adversarial training significantly increases localization accuracy in APR. Meanwhile, our results indicate that APR consistently benefits from more diverse visual data, and we observe clear signs of a more generalizable APR emerging with its localization performance cannot be explained merely by memorization.

Our contributions are summarized as follows:

- We identify the crucial role of appearance diversity for APR, and develop a 3DGS-based appearance-varying data augmentation framework to efficiently generate diverse synthetic data with controllable lighting conditions.
- We propose an adversarial discriminator to reduce the syn-to-real gap. Together with progressive data synthesis, we form a robust two-branch joint training pipeline that fully unleashes the power of data diversity.
- We conduct extensive experiments showing our method outperforms state-of-the-art approaches on challenging datasets with significant appearance change. Ablation studies further analyze key factors affecting performance.

2. Related Works

Visual Localization. Visual localization aims to estimate a camera’s translation and rotation within a 3D scene. Traditional geometry-based methods [6, 15, 32, 35, 44–47, 57] accomplish this by using point clouds and a reference image database, relying on stored descriptors and image retrieval to establish 2D-3D correspondences. In contrast, scene coordinate regression (SCR) methods [2–4, 61] em-

bed map information within neural networks to directly predict 2D-3D correspondences. Both approaches generally require PnP [17] and RANSAC [16] to output camera poses at test time, which adds additional computation cost. Alternatively, absolute pose regression (APR) [5, 7, 22, 24, 41, 52] aims to directly regress the camera pose from a query image using neural networks. Although the performance is suboptimal compared with geometry-based methods, APR remains a promising approach due to its fast inference.

Data Augmentation for Pose Regression. End-to-end pose regression methods rely heavily on the amount and diversity of training data. Previous work [49] shows that APR implicitly learns image *retrieval* in the given environment. Therefore, the following works LENS [42], DFNet [8] and PMNet [30] enhance APR performance by spatially enriching training views with NeRF. However, these approaches fail to address the generalizability of APR models and exhibit several limitations: (1) The efficiency of training and novel view synthesis (NVS) in NeRF is severely restricted, hindering scalability. (2) They limit NVS to geometric (pose) transformations while neglecting photometric (appearance) variations, thereby decreasing APR robustness to changes in visual appearance. (3) The augmented data is underutilized in their learning frameworks, leaving its potential for improving APR largely untapped. Differently, our framework switches to 3DGS [25] as the scene representation to efficiently generate novel posed images with *controllable* appearances and introduce adversarial training to unleash the power of such diverse data.

Handling and Synthesizing Challenging Scenarios. Visual localization often encounters unstructured photo collections [56], where visual appearance varies due to moving objects, lighting changes, and inconsistent camera exposure settings. To tackle these in-the-wild challenges, NeRF-W [38] uses per-image transient and appearance embeddings. In 3DGS [25], VastGaussian [31] applies a CNN to 3DGS outputs but still struggles with significant appearance variations. SWAG [12] mitigates this issue by storing appearance information in an external hash-grid-based implicit field, while GS-W [63] enhances flexibility by separating intrinsic and dynamic appearance features for each Gaussian point. 3DGM [29] leverages consensus across multiple sequences as the self-supervision signal to remove transient and moving objects without human annotations. Deblur-GS [62] addresses motion blur—another challenge in localization datasets—by modeling camera motion to yield sharper edges in rendered scenes. Our method incorporates appearance modeling and edge refinement to handle and synthesize diverse indoor, outdoor, and driving scenes.

3. Method

3.1. Pre-Processing with 3DGS

A robust pose regressor should focus on intrinsic scene attributes, not appearance variations. Therefore, we first

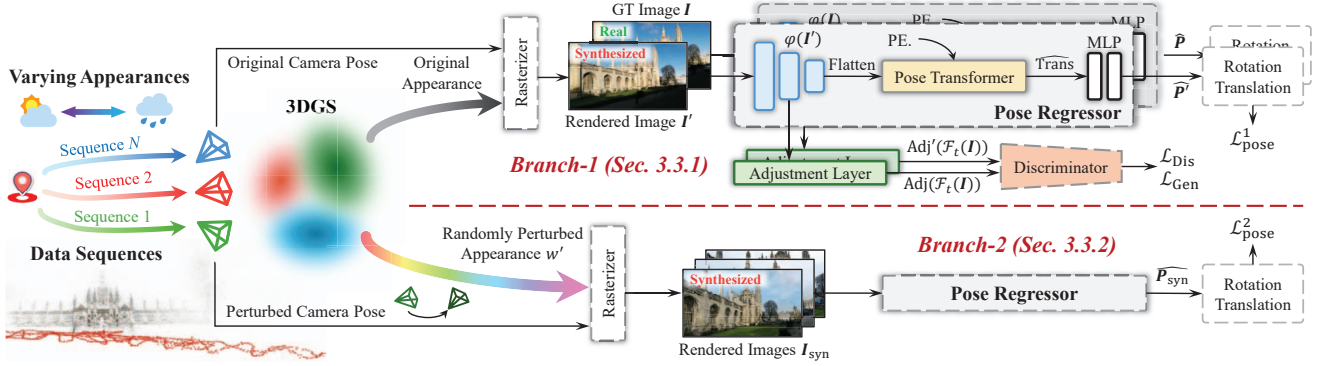


Figure 2. **Pipeline of RAP.** We lift multiple RGB video sequences into 3D Gaussian Splats, which serve as our data engine. The **branch-1** (see Sec. 3.3.1) inputs paired real and synthetic images to regress poses, with a discriminator to bridge the syn-to-real gap. The **branch-2** (see Sec. 3.3.2) generates views with novel poses and appearances, which are fed into the same pose regressor as additional supervision.

synthesize diverse visual data for training. We leverage 3DGS [25], representing scenes with explicit ellipsoids, to model diverse appearances. Following GS-W [63], we assume the scene contains K Gaussians and represent the independent intrinsic material attributes using positions $\mu \in \mathbb{R}^{K \times 3}$, spherical harmonics $\mathcal{Y} \in \mathbb{R}^{K \times 16 \times 3}$, and other parameters Θ including rotation $\mathbf{q} \in \mathbb{R}^{K \times 4}$, scaling $\mathbf{s} \in \mathbb{R}^{K \times 3}$, and opacity $\alpha \in \mathbb{R}^K$. To capture the dynamic appearance influenced by environmental factors, we extract features from the input image and assign each Gaussian its own feature using a learnable sampler $\mathcal{S} \in \mathbb{R}^{K \times 2}$, forming features $\mathcal{E} \in \mathbb{R}^{K \times 16 \times 3}$. We also incorporate the camera’s view direction θ to account for viewpoint-dependent effects. The final color of Gaussians $\mathcal{C} \in \mathbb{R}^{K \times 3}$ is:

$$\mathcal{C} = \text{MLP}(\mu, \mathcal{Y}, \omega \mathcal{E}, \theta), \quad (1)$$

where ω is the blending weight that controls the dynamic appearance of the rendered image.

Another significant challenge in visual localization is motion blur, often caused by slow shutter speeds during video capture, leading to pose ambiguity and degraded rendering quality, further decreasing localization accuracy. Inspired by Deblur-GS [62], we model camera motion blur as the inverse of scene motion, i.e., the transformation in Gaussian position denoted by $\mathcal{T} \in \text{SE}(3)$. For each training image, we sample a certain time step along a linear trajectory with a sampling weight $\phi \in \mathbb{R}^n$ and blend them to compute loss \mathcal{L} with the original blurry image $\mathcal{I}_b \in \mathbb{R}^{H \times W \times 3}$, optimizing \mathcal{T} , ϕ , \mathcal{C} and other 3DGS parameters Θ :

$$\underset{\phi, \mathcal{T}, \mu, \mathcal{Y}, \mathcal{E}, \Theta}{\text{argmin}} \mathcal{L} \left(\mathcal{I}_b, \sum_{i=1}^n \phi_i \text{Render}(\mathcal{T}_i(\mu), \mathcal{C}, \Theta) \right), \quad (2)$$

where the details of \mathcal{L} are in supplemental materials. After training, our 3DGS can efficiently render posed images given θ and ω .

3.2. Architecture of Pose Regressor

Given a set of images and their associated camera poses $\{(\mathcal{I}_i, \mathbf{P}_i)\}_{i=1}^n$, our goal is to train a neural network to directly output a homogeneous camera pose $\mathbf{P} \in \mathbb{R}^{3 \times 4}$ for a query image $\mathcal{I} \in \mathbb{R}^{H \times W \times C}$. Our network architecture is shown as the pose regressor in Fig. 2.

Feature Extractor. Pose regression networks typically extract features using a common backbone φ , such as VGG [55] or EfficientNet [58], leveraging multiple deeper layers for translation and rotation regression:

$$\varphi(\mathcal{I}) = \{\mathcal{F}_0(\mathcal{I}), \dots, \mathcal{F}_{N-1}(\mathcal{I}), \mathcal{F}_N(\mathcal{I})\}, \quad (3)$$

$\mathcal{F}_*(\cdot)$ denotes features extracted from the $*$ -th layer of a backbone with N layers. $\mathcal{F}_t(\mathcal{I})$ and $\mathcal{F}_r(\mathcal{I})$ denote features for translation and rotation regression, respectively.

Pose Transformer. Unlike CNN-based regression models [8, 30], where fine-grained local features can introduce noise and harm performance, we propose *Pose Transformer* to leverage the strong ability of Vision Transformer (ViT) [14] for modeling long-range dependencies. Each Transformer generates a global token (Trans for translation and Rot for rotation) to provide a comprehensive context for pose regression, inspired by the CLS token in ViT. Given $\mathcal{F}_r(\mathcal{I})$ and $\mathcal{F}_t(\mathcal{I})$, the translation token is then concatenated with the flattened input features*:

$$\widetilde{\mathcal{F}}_t(\mathcal{I}) = \text{Cat}(\text{Flatten}(\mathcal{F}_t(\mathcal{I})), \text{Trans}) \in \mathbb{R}^{(H_t W_t + 1) \times C_t}. \quad (4)$$

The positional encodings are then added to the flattened feature ($\text{PE} + \widetilde{\mathcal{F}}_t(\mathcal{I})$) $\in \mathbb{R}^{(H_t W_t + 1) \times C_t}$. Multi-head Self-Attention (MSA) is then conducted through a stack of multiple layers with the post-processing as follows:

$$\begin{aligned} \widehat{\mathcal{F}}'_t(\mathcal{I}) &= \text{MSA}(\text{PE} + \widetilde{\mathcal{F}}_t(\mathcal{I})) + \text{PE} + \widetilde{\mathcal{F}}_t(\mathcal{I}), \\ \widehat{\mathcal{F}}_t(\mathcal{I}) &= \text{LN}(\text{FFN}(\text{LN}(\widehat{\mathcal{F}}'_t(\mathcal{I}))) + \widehat{\mathcal{F}}'_t(\mathcal{I})), \end{aligned} \quad (5)$$

*We only present the translation regression for simplicity.

where LN indicates layer normalization and FFN denotes the fully connected feed-forward network, consisting of two linear layers with a ReLU. The final output is flattened back to $(H_t W_t + 1) \times c_t$. See supplementary for more details.

Regression Head. Only the processed translation token, $\widehat{\text{Trans}}$, capturing global features for regression, is fed into the regression head. This regression head consists of two MLPs, each with a hidden layer and GeLU activation:

$$\hat{\mathbf{t}} = \text{Linear}(\text{GeLU}(\text{Linear}(\widehat{\text{Trans}}))). \quad (6)$$

The $\hat{\mathbf{t}}$ represents the final prediction for translation. Similarly, we obtain the rotation prediction denoted by $\hat{\mathbf{r}}$.

3.3. Two-Branch Joint Training Paradigm

3.3.1. Branch-1: Aligning Features via Discriminator

Synthetic images from 3DGS provide novel viewpoints and appearances but often contain artifacts, leading to a syn-to-real domain gap. To align features from rendered and real images of the same pose, we introduce an adversarial training mechanism besides the basic pose regression training.

Pose Regression Loss. For basic training, we render the synthetic image I' with the same pose label \mathbf{P} as the real image I , both used as supervision for the pose regressor. The training objective consists of translation loss \mathcal{L}_t and the rotation loss \mathcal{L}_r , which are measured by the Euclidean distance between the ground truth pose $\mathbf{P} = \{\mathbf{t}, \mathbf{r}\}$ and the estimated pose $\widehat{\mathbf{P}} = \{\hat{\mathbf{t}}, \hat{\mathbf{r}}\}$:

$$\mathcal{L}_t = \|\mathbf{t} - \hat{\mathbf{t}}\|_2, \quad (7)$$

$$\mathcal{L}_r = \left\| \mathbf{r} - \frac{\hat{\mathbf{r}}}{\|\hat{\mathbf{r}}\|} \right\|_2, \quad (8)$$

$$\mathcal{L}_{\text{pose}}^1 = \mathcal{L}_t \exp(-s_t) + s_t + \mathcal{L}_r \exp(-s_r) + s_r, \quad (9)$$

where s_t and s_r are learned parameters for balancing the optimization between rotation and translation [23].

Adversarial Loss. The adversarial training mechanism optimizes the discriminator to distinguish real from rendered image features, while training the feature extractor to fool the discriminator, effectively bridging the domain gaps. To prevent vanishing gradients, we propose a novel adversarial objective for pose regression, inspired by LSGAN [37]:

$$\begin{aligned} \arg\min_D \mathcal{L}_{\text{Dis}}(D) &= \frac{1}{2} \mathbb{E}_{I \sim p_{\text{data}}(I)} [(D(\text{Adj}(\mathcal{F}_t(I))) - 1)^2] \\ &+ \frac{1}{2} \mathbb{E}_{I'} [D(\text{Adj}'(\mathcal{F}_t(I')))^2], \end{aligned} \quad (10)$$

$$\arg\min_G \mathcal{L}_{\text{Gen}}(G) = \frac{1}{2} \mathbb{E}_{I'} [(D(\text{Adj}'(\mathcal{F}_t(I'))) - 1)^2]. \quad (11)$$

Here, Adj and Adj' are the adjustment layers, consisting of Conv-ReLU-BN layers. The feature extractor φ acts as the generator G , while D is the discriminator, composed of several convolutional layers with ReLU activations. More details are in supplemental materials.

3.3.2. Branch-2: Training while Synthesizing Data

With the proposed appearance-varying 3DGS, more posed images are generated to enrich the training data for better generalizability. Specifically, our data synthesis is categorized into two dimensions: *pose augmentation* and *appearance augmentation*, as illustrated in Fig. 1. For pose augmentation, given a training pose \mathbf{P} , a perturbed pose \mathbf{P}_{syn} is generated around \mathbf{P} by the translation noise of δt and rotation noise of δr . For appearance augmentation, we randomly adjust the appearance of rendered images using random blending weights ω , and then render the synthetic image I_{syn} using the Gaussian Splats trained in Sec. 3.1. The novel image-pose pair $(I_{\text{syn}}, \mathbf{P}_{\text{syn}})$, online generated every 20 epochs during training until the validation MSE loss and median errors cease to decrease, serves as additional supervision for the training. Given the estimated pose of the synthesized image denoted by $\widehat{\mathbf{P}}_{\text{syn}}$, the loss function $\mathcal{L}_{\text{pose}}^2(\widehat{\mathbf{P}}_{\text{syn}}, \mathbf{P}_{\text{syn}})$ is same as $\mathcal{L}_{\text{pose}}^1$.

3.3.3. Overall Objective

The total loss for the pose regressor is:

$$\mathcal{L}_{\text{total}} = \beta_1 \mathcal{L}_{\text{pose}}^1 + \beta_2 \mathcal{L}_{\text{pose}}^2 + \beta_3 (\mathcal{L}_{\text{Gen}} + \mathcal{L}_{\text{Dis}}), \quad (12)$$

where $\beta_1, \beta_2, \beta_3$ are loss weights. The total loss will optimize the pose regressor, adjustment layers, and discriminator. Only the pose regressor will be deployed in the inference phase, while the other two components are discarded.

4. Experiments

4.1. Evaluation Setup

Datasets. We follow previous works [8, 30] to mainly use four scenes in the Cambridge Landmarks dataset [24] with spatial extents around 875 m². Moreover, we evaluate our method on MARS [28], a self-driving dataset featuring challenges like moving objects, lighting changes, and motion blur. To investigate the robustness of our model under extreme lighting changes, such as the transition from day to night, we also prepared a subset of the Aachen Day-Night dataset [48]. The training data includes images captured using various camera models with differing resolutions, which renders direct evaluation with APR methods infeasible. Thus, we standardized the camera models through center cropping and built a COLMAP [50] model as pose annotations, including 13 nighttime images for evaluation and 246 daytime images for training 3DGS and RAP. We also employ the 7-Scenes dataset [54], which provides seven indoor scenes with volumes spanning 1 m³–18 m³, and follow the original training and testing splits with more accurate SfM pose annotations [4, 9]. Although it is an indoor dataset, it is still non-trivial as it includes texture-less surfaces, object occlusions, and motion blur.

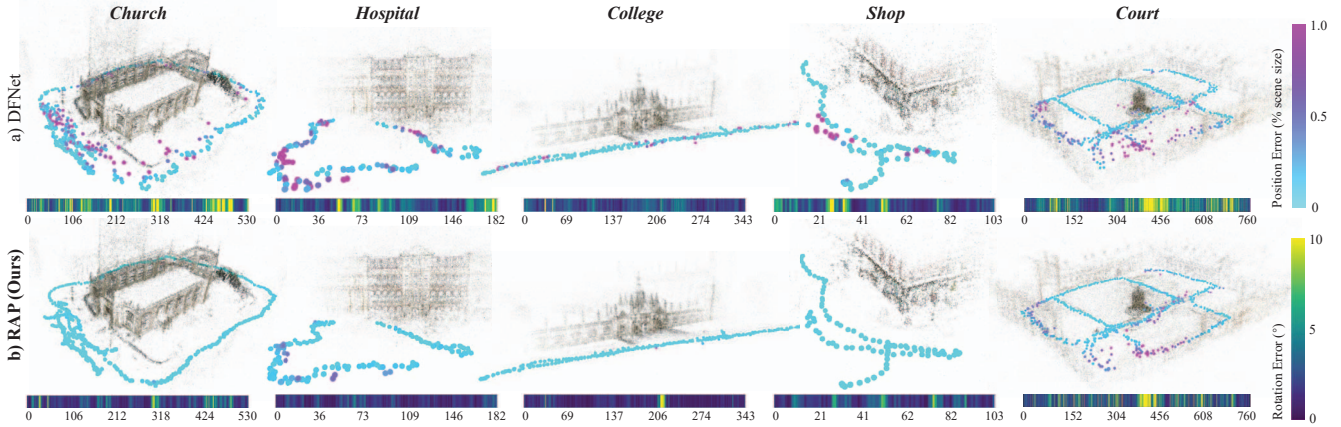


Figure 3. **Qualitative comparison of camera pose estimation errors between a) DFNet [8] and b) our RAP framework** across five scenes on the Cambridge Landmarks dataset [24]. Our RAP framework estimates trajectories that more closely follow the ground truth, with significantly reduced rotation and position errors compared to DFNet [8].

Table 1. **Median translation (cm) and rotation ($^{\circ}$) errors on the Cambridge Landmarks dataset [24].** The best results[†] in pure APR and PPR are highlighted in **bold**.

	Methods	College	Hospital	Shop	Church	Average [‡]	Court
Pure APR	PN [24]	166/4.86	262/4.90	141/7.18	245/7.95	204/6.23	683/3.50
	MapNet [5]	107/1.89	194/3.91	149/4.22	200/4.53	163/3.64	N/A
	MS-Trans. [53]	83/1.47	181/2.39	86/3.07	162/3.99	128/2.73	N/A
	PAE [51]	90/1.49	207/2.58	99/3.88	164/4.16	140/3.03	N/A
	LENS [†] [42]	33/0.50	44/0.90	27/1.60	53/1.60	39/1.20	N/A
	DFNet [8]	73/2.37	200/2.98	67/2.21	137/4.03	119/2.90	217/4.11
	PMNet [30]	68/1.97	103/1.31	58/2.10	133/3.73	90/2.27	N/A
	RAP (Ours)	52/0.90	87/1.21	33/1.48	53/1.52	56/1.28	115/1.68
SCR	DSAC* [2]	18/0.3	21/0.4	5/0.3	15/0.6	15/0.4	34/0.2
	ACE [4]	28/0.4	31/0.6	5/0.3	18/0.6	21/0.5	43/0.2
	GLACE [61]	19/0.3	17/0.4	4/0.2	9/0.3	12/0.3	19/0.1
PPR	FQN-MN [19]	28/0.38	54/0.82	13/0.63	58/2.00	38/0.96	4253/39.16
	LENS [42]	34/0.54	45/0.96	28/1.66	54/1.66	40/1.21	N/A
	CrossFire [43]	47/0.7	43/0.7	20/1.2	39/1.4	37/1.0	N/A
	NeFeS ₅₀ [9]	37/0.54	52/0.88	15/0.53	37/1.14	35/0.77	N/A
	HR-APR [33]	36/0.58	53/0.89	13/0.51	38/1.16	35/0.78	N/A
	MCLoc [59]	31/0.42	39/0.73	12/0.45	26/0.88	27/0.62	N/A
	DFNetGS-CPR [34]	23/0.32	42/0.74	10/0.36	27/0.62	26/0.51	N/A
	ACE _{GS-CPR} [34]	20/0.29	21/0.40	5/0.24	13/0.40	15/0.33	N/A
	DFNet _{ref} (Ours)	16/0.24	21/0.41	8/0.42	10/0.26	14/0.33	25/0.13
	RAP_{ref} (Ours)	15/0.25	18/0.38	5/0.23	9/0.23	12/0.27	23/0.15

[‡]Since most methods did not report results on *Court*, it is excluded from the average error calculation. [†]As CoordiNet + LENS [42] does not provide open-source code, it is unclear whether any post-processing is used.

Baselines. We first compare our proposed RAP against common APR-only approaches on the four datasets, where PMNet [30] and DFNet [8] are the most related and advanced methods based on data augmentation. We split the remaining methods into two categories based on whether they rely on extra novel view synthesis in test time, including SCR [3, 4, 61] and PPR (Post Pose Refinement) [9, 19, 33, 43, 59, 64, 66], which involves rendering images, querying features in novel views by the initial pose, iterative refinement or sequential refinement [41].

Implementation Details. First, we optimize our 3DGS for each scene without masking moving objects. We

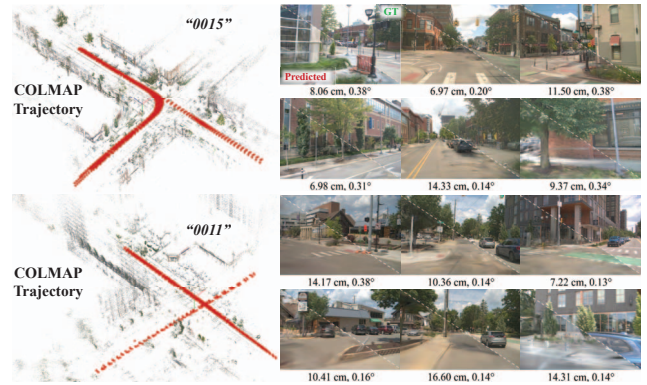


Figure 4. **Visualization of RAP_{ref} on MARS [28].** In each sub-figure, a diagonal line separates the “Predicted” (rendered from the refined pose) and “GT” (ground truth) sections. Smooth alignment along this boundary shows RAP_{ref}’s improved pose accuracy.

Table 2. **Median translation (cm) and rotation ($^{\circ}$) errors on the MARS dataset [28].**

Methods	“0011”	“0015”	“0037”	“0041”	Average
PoseNet [24]	149/1.80	136/2.34	123/1.60	75/0.92	121/1.67
RAP (Ours)	32/0.61	37/1.08	15/0.35	28/0.35	28/0.60
RAP_{ref} (Ours)	8.5/0.13	8.2/0.20	8.7/0.09	7.6/0.11	8.3/0.13

Table 3. **Median translation (COLMAP [65] unit) and rotation ($^{\circ}$) errors on the Aachen Day-Night Dataset [48].**

APR-Based					SCR-Based	
PoseNet [24]	DFNet [8]	RAP w/o App.	RAP	RAP _{ref}	ACE [4]	GLACE [61]
217/74.30	174/85.80	134/75.99	130/13.70	67/4.64	914/90.50	482/36.4

then train our RAP network, which uses an Efficient-B0 backbone [36] pre-trained on ImageNet [13], optimized with Adam [26] at a learning rate of 10^{-4} . Only the features from the third (`reduction_3`) and fourth (`reduction_4`) layers are used respectively for translation and rotation regression, and both layers are utilized for

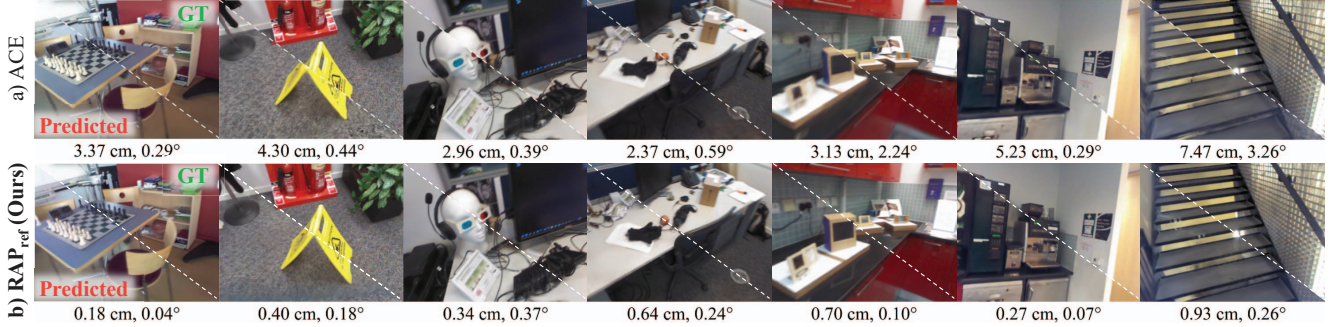


Figure 5. Visualization of the localization errors of RAP_{ref} on the 7-Scenes dataset [54].

Table 4. Quantitative results on the 7-Scenes dataset [54]. The best results in pure APR and PPR are highlighted in **bold**. DSLAM GT and SfM GT refer to different sets of ground truth. More visualizations and details are in supplemental materials.

Category	Methods	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Average
Pure APR	PoseNet (PN) [24]	32/8.12	47/14.4	29/1.20	48/7.68	47/8.42	59/8.64	47/13.80	44/10.4
	MapNet [5]	8/3.25	27/11.7	18/13.3	17/51.5	22/4.02	23/4.93	30/12.1	21/7.77
	AtLoc+ [60]	10/3.18	26/10.8	14/11.4	17/5.16	20/3.94	16/4.90	29/10.2	19/7.08
	MS-Transformer [53]	11/4.66	24/9.60	14/12.2	17/5.66	18/4.44	17/5.94	17/5.94	18/7.28
	PAE [51]	12/4.95	24/9.31	14/12.5	19/5.79	18/4.89	18/6.19	25/8.74	19/7.48
	CoordiNet + LENS [42]	4/1.38	11/3.77	8/5.86	8/1.98	9/2.27	10/2.27	15/3.67	9/3.07
	DFNet [8]	5/1.88	17/6.45	6/3.63	8/2.48	10/2.78	22/5.45	16/3.29	12/3.71
	PMNet [30]	4/1.70	10/4.51	7/4.23	7/1.96	14/3.33	14/3.36	16/3.62	10/3.24
	RAP (Ours, DSLAM GT)	3/1.41	7/3.46	6/6.02	5/1.97	6/1.96	7/2.18	10/2.14	6/2.73
	RAP (Ours, SfM GT)	2/0.85	6/2.84	4/4.52	5/1.94	4/1.70	7/2.12	10/1.30	5/2.18
SCR	DSAC [2]	0.5/0.17	0.8/0.28	0.5/0.34	1.2/0.34	1.2/0.28	0.7/0.21	2.7/0.78	1.1/0.34
	ACE [4]	0.5/0.18	0.8/0.33	0.5/0.33	1.0/0.29	1.0/0.22	0.8/0.20	2.9/0.81	1.1/0.34
	GLACE [61]	0.6/0.18	0.9/0.34	0.6/0.34	1.1/0.29	0.9/0.23	0.8/0.20	3.2/0.93	1.2/0.36
	marepo [‡] [10]	2.6/1.35	2.5/1.42	2.3/2.21	3.6/1.44	4.2/1.55	5.1/1.99	6.7/1.83	3.9/1.68
PPR	FQN-MN [19]	4.1/1.31	10.5/2.97	9.2/2.45	3.6/2.36	4.6/1.76	16.1/4.42	139.5/34.67	28/7.3
	CrossFire [43]	1/0.4	5/1.9	3/2.3	5/1.6	3/0.8	2/0.8	12/1.9	4.4/1.38
	DFNet + NeFeS ₅₀ [9]	2/0.57	2/0.74	2/1.28	2/0.56	2/0.55	2/0.57	5/1.28	2.4/0.79
	HR-APR [33]	2/0.55	2/0.75	2/1.45	2/0.64	2/0.62	2/0.67	5/1.30	2.4/0.85
	MCLoc [59]	2/0.8	3/1.4	3/1.3	4/1.3	5/1.6	6/1.6	6/2.0	4.1/1.43
	DFNet + GS-CPR (SfM GT) [34]	0.7/0.20	0.9/0.32	0.6/0.36	1.2/0.32	1.3/0.31	0.9/0.25	2.2/0.61	1.1/0.34
	ACE + GS-CPR (SfM GT) [34]	0.5/0.15	0.6/0.25	0.4/0.28	0.9/0.26	1.0/0.23	0.7/0.17	1.4/0.42	0.8/0.25
	RAP_{ref} (Ours, DSLAM GT)	2.83/1.43	2.07/1.23	1.54/1.87	2.55/1.20	4.59/1.54	4.23/1.83	3.48/1.23	3.04/1.48
	RAP_{ref} (Ours, SfM GT)	0.33/0.11	0.51/0.21	0.40/0.29	0.59/0.17	0.83/0.20	0.50/0.12	1.11/0.32	0.61/0.20

[‡]As marepo [10] combines SCR and APR, we classify it as SCR.

narrowing the domain gap via the discriminator, which is also optimized with Adam [26], using a learning rate of 10^{-4} and betas set to (0.5, 0.999). More details about training are in supplemental materials. For generating random views, we apply random normalized perturbations to each training pose: $\delta t = 20$ cm and $\delta r = 10^\circ$ for indoor scenes, and $\delta t = 150$ cm and $\delta r = 4^\circ$ for outdoor scenes.

To allow for comparison with SCR methods and leverage 3DGS’s efficient rendering for PPR, we extend the APR pipeline with match-based refinement similar to GS-CPR [34], denoted as RAP_{ref} . At test time, RAP ’s initial pose is used to render an RGB-D image via 3DGS. Together with MAST3R [27], we can obtain 2D-3D correspondences to perform RANSAC-PnP [16, 17], resulting in a refined pose. More details are in supplementary materials.

4.2. Benchmark Results

Cambridge Landmarks [24]. In the challenging outdoor Cambridge Landmarks dataset (Table 1), our RAP reduces both translation and rotation errors across all scenes by over

30% compared to other APR-only methods. The visualization in Fig. 3 shows that our method produces fewer outliers than DFNet [8]. In the three larger-scale scenes with significant appearance diversity (*College*, *Church*, and *Court*), rotation error is even halved compared to DFNet. Table 1 also shows the effectiveness of our RAP_{ref} in further reducing pose errors through refinement. RAP_{ref} outperforms CoordiNet + LENS [42], which assumes a continuous trajectory when an Extended Kalman Filter [21] is required for refinement [41]. RAP_{ref} even surpasses ACE [4] and its post-refinement variant, ACE + GS-CPR [34], despite GS-CPR manually masking dynamic objects when building 3DGS. This demonstrates the strong representation capability of our appearance-varying 3DGS with deblurring.

MARS [28]. Autonomous driving scenarios present unique challenges, including moving objects and frequent changes in lighting conditions, as illustrated in Fig. 4. Our RAP demonstrates effective and robust performance across four challenging scenes, as shown in Table 2, achieving an average of 28 cm / 0.60° localization error. This significantly

Table 5. Ablation study.

Setups	on <i>Shop</i>	Trans. (cm) ↓	Rot. (°) ↓
I (Baseline): $\varphi = \text{VGG16}$		174	5.45
II: $\varphi = \text{Efficient-B0}$		103	4.64
III: II + Pose Aug.		75	3.52
IV: III + Appearance Aug.		60	3.14
V: IV + Decoder (ConvNet)		52	2.51
VI: V + Decoder (Transformer)		40	1.98
VII (Ours): VI + Discriminator		33	1.48

outperforms the baseline[†] PoseNet [24]. With one-shot refinement, our RAP_{ref} further reduces outdoor localization errors to below 10 cm.

Aachen Day-Night [48]. Benefiting from appearance augmentation, our RAP significantly reduces the localization error from 134 unit / 75.99° to 130 unit / 13.70° , outperforming other APR [8, 24] and SCR [4, 61] baselines, as shown in Table 3. This demonstrates the effectiveness of appearance diversity in handling extreme lighting changes.

7-Scenes [54]. As shown in Table 4, our RAP reduces translation error by 50% (10 cm \rightarrow 5 cm) and rotation error by 32.72% ($3.24^\circ \rightarrow 2.18^\circ$) on average compared to previous state-of-the-art single-frame APR methods. The only exception is *Heads*, where the rotation error is suboptimal. This scene consists of just two sequences—one for training and one for testing—potentially limiting the effectiveness of our augmentation in capturing scene variability. Meanwhile, RAP_{ref} further reduces localization error below 1 cm with one-shot refinement using our 3DGS. It also surpasses ACE [4] and its post-refinement variant, ACE + GS-CPR [34]. Qualitative examples are shown in Fig. 5.

4.3. Ablation Study

We conduct ablation studies on the validation set of *Shop* in the Cambridge Landmarks dataset to investigate the impact of all the components in our RAP. Setup I, our baseline, consists of the same components as in PoseNet [24] and has been retrained for our experiments. In Setup II, we replace the feature extraction from VGG16 [55] to Efficient-B0 [36], which enhances performance due to its superior feature representation, while they both exhibit poor performance due to the lack of data synthesis. In Setup III and IV, we explore the effectiveness of the designed pose augmentation and appearance augmentation, which bring notable improvements: translation error reduces from 103 cm to 75 cm, and rotation error from 3.52° to 3.14° . In Setup V and VI, we add regular convolutional layers and Pose Transformer between feature extraction and pose regression. Both improve performance due to the increasing parameters, but the Transformer achieves superior results by effectively handling long-term dependencies through atten-

[†]DFNet [8] results are omitted as we were unable to successfully train its NeRF component, likely due to the need for manual scene scaling within $[-\pi, \pi]$, which is tedious for diverse outdoor scenes.

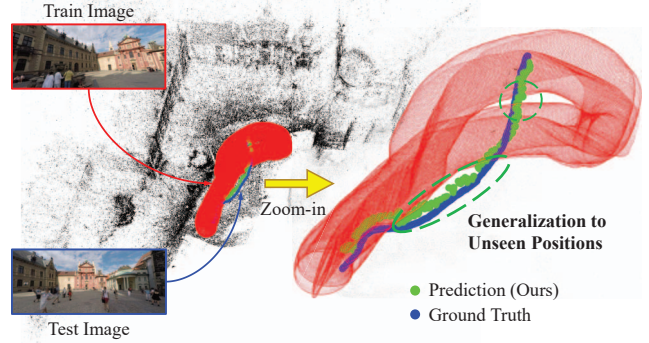


Figure 6. Visualization of the training set distribution and results on *St. George’s Basilica Building* [49]. The red hollow spheres, centered on the real images in the training set, indicate the potential locations of all synthetic images during training.

tion mechanisms. Finally, in Setup VII, our adversarial discriminator effectively reduces the syn-to-real domain gap, allowing the model to learn better pose regression features from synthetic data and further reduce localization error.

4.4. Discussion on Data Synthesis

Emerging Generalizability. Previously, APR has been understood to implicitly learn image retrieval [49], lacking the ability to successfully interpolate between training samples and generalize beyond them. To investigate how APR training is affected by increasing synthetic data, we trained RAP on the *St. George’s Basilica Building* [49] and visualized the results in Fig. 6. Here, the translation perturbation was set to $\delta t = 350$ cm and the rotation perturbation to $\delta r = 60^\circ$. Notably, the test set contained two regions entirely uncovered by the training set. Despite this, the model still closely predicts the test camera poses, demonstrating generalization ability beyond the original training positions.

We also learn from our experiments that reducing the rotation perturbation, such that the overlap between test and training views remains minimal, leads to high localization error. This is because the translation and rotation parameter space is inherently a $\text{SE}(3)$ manifold. Even if the translation remains fixed, significant rotation changes result in entirely different visual content in the images, naturally preventing the model from estimating poses of such unseen views, which correspond to a large distance on the $\text{SE}(3)$ manifold. Therefore, enabling generalization across a broader range of space is an important direction for future work.

Analyzing Generalization Boundaries. To evaluate the model’s generalizability, we designed an experiment introducing a “void zone” centered on the test camera, where all real and synthetic data within this zone were excluded. The void zone was progressively expanded to determine the critical threshold at which the localization performance declines most significantly. Specifically, for *Shop*, we used 100% of the training set to ensure complete scene cov-

Table 6. Exploring the generalization boundaries of the model with synthetic data. Green, blue, and red percentages indicate the relative change in localization error (Med Err) compared to the scenario without a void zone.

	w/o Void Zone	w/ Void Zone (cm ^o)				
		10/0.5	20/1	30/1.5	50/2	80/2.5
		Med Err ↓	33/1.26	30/1.34	32/1.32	40/1.84
(rel. change)	0%/0%	-9%/6.3%	-3%/4.7%	21%/46.0%	18%/64.3%	48%/74.6%
Avg Err ↓	41/1.51	38/1.75	41/1.63	51/2.40	48/2.40	61/2.84
Max Err ↓	155/4.52	147/6.56	219/6.01	192/8.38	246/9.80	242/13.12
Min Err ↓	3/0.17	8/0.20	4/0.22	7/0.30	4/0.16	8/0.60

Table 7. Ablation on different pose augmentation policies.

Methods	College	Hospital	Shop	Church	Average
RAP (LENS [42])	73/1.15	126/1.87	71/3.37	128/3.50	100/22.47
RAP (Ours)	52/0.90	87/1.21	33/1.48	53/1.52	56/1.28

erage, with void zone ranges set as [10/0.5, 20/1, 30/1.5, 50/2, 80/2.5, 100/3] (cm^o). The results in Table 6 demonstrate a stepwise decline in performance. Initially, expanding the void zone has minimal impact on localization accuracy. However, at 30 cm / 1.5°, a sharp decrease in performance marks the model’s generalization boundary.

Pose Augmentation Policy. We conduct experiments using a modified version of RAP, with the pose augmentation approach identical to that of LENS [42], as shown in the Table 7, where our method obtains superior performance. This may be because, although LENS’s pose augmentation policy covers a broader spatial area than the training set, its synthetic data may have lower NVS image quality in many unseen regions, which could negatively affect APR training.

Density of Training Data. As shown in Table 8, our method with the proposed augmentation significantly reduces errors as the density of real training data increases from 20% to 80%. However, the localization accuracy remains almost unchanged from 80% to 100%, as the scene is already sufficiently covered. Notably, using 100% of the training data without augmentation can result in a significantly higher maximum error in translation, nearly double that with only 20% of the training data with augmentation, despite its limited spatial coverage. This suggests that our augmentation method successfully prevents overfitting to the training data, improving generalization to the test set.

Quality of Training Data. We evaluate the impact of synthetic image quality on model performance in Table 9, using 20% and 50% of the real data for pose regression. For *Shop*, it is evident that fewer training samples in 3DGS result in lower-quality rendered views (as indicated by lower PSNR), leading to suboptimal localization performance, particularly for rotation. Surprisingly, localization performance using only 20% of the data for training suboptimal 3DGS and pose regression surpasses the results obtained with 100% of the data without augmentation, as shown in Table 8. This experiment confirms the need for a robust NVS model and the proposed augmentation method in APR training.

Table 8. Impact of the density of real training data. Our augmentation improves the model’s ability to generalize across the entire scene, although this effect has an upper limit.

Training Pose %	w/ Appearance & Pose Aug. (cm ^o)					w/o Aug. (cm ^o)	
	100%	80%	60%	40%	20%	100%	50%
Med Err ↓	33/1.26	32/1.27	37/1.90	57/2.23	87/3.65	98/3.75	104/4.17
Avg Err ↓	41/1.51	40/1.50	48/2.17	62/2.81	91/4.65	128/4.49	139/5.33
Max Err ↓	155/4.52	158/4.09	193/9.39	230/11.06	231/15.45	490/20.73	500/21.02
Min Err ↓	3/0.17	7/0.20	7/0.18	6/0.38	12/0.46	13/0.63	9/0.48

Table 9. Impact of synthetic image quality. Training with higher-quality synthetic images from advanced NVS models enhances localization performance.

3DGS Performance			Localization Performance (cm ^o)				
% Images (Train)	PSNR ↑ (Train)	PSNR ↑ (Test)	% Images (Train)	Med Err ↓	Avg Err ↓	Max Err ↓	Min Err ↓
20%	29.08	15.98	20%	58/3.59	68/4.31	211/12.19	14/0.51
20%	29.08	15.98	50%	43/2.47	55/3.26	196/21.11	7/0.40
50%	26.88	17.55	50%	37/1.88	48/2.37	184/10.17	9/0.52
100%	24.60	18.30	50%	35/1.64	41/2.12	130/11.07	4/0.38

Table 10. Ablation on different 3D representations.

Methods	College	Hospital	Shop	Church	Average
DFNet [8]	73/2.37	200/2.98	67/2.21	137/4.03	119/2.90
DFNetGS	102/2.31	137/8.08	77/3.92	123/4.68	110/4.75
RAP (Ours)	52/0.90	87/1.21	33/1.48	53/1.52	56/1.28

Different 3D Representations. We evaluate the impact of different 3D representations on the performance in Table 10. Trivially replacing NeRF with 3DGS in existing frameworks degrades performance due to 3DGS’s inferior 3D consistency. This shows the effectiveness of our proposed joint training paradigm in RAP, which better utilizes diverse synthetic data to learn appearance-invariant features, rather than naively transferring from NeRF to 3DGS.

5. Conclusion

Summary. We address absolute pose regression with a robust two-branch joint training framework based on Transformer, coupled with an efficient data synthesis pipeline leveraging 3D Gaussian Splats (3DGS) to synthesize numerous posed images with diverse appearances as additional supervision. Our RAP achieves state-of-the-art localization performance, even under challenging appearance variations. Moreover, we thoroughly investigate the impact of synthesizing diverse data and present a novel perspective on APR: generalizability can emerge if the learning gap in APR is effectively addressed together with diverse data. We believe our RAP could be a promising starting point, and the experiments presented in the paper can provide useful insights for future research in this field.

Acknowledgment. This work was supported in part through NSF grants 2238968, 2121391, and 2024882, and the NYU IT High Performance Computing resources, services, and staff expertise. Yiming Li is supported by NVIDIA Graduate Fellowship (2024-2025).

References

- [1] Joydeep Biswas and Manuela Veloso. Depth camera based indoor mobile robot localization and navigation. In *2012 IEEE International Conference on Robotics and Automation*, pages 1697–1702, 2012. 1
- [2] Eric Brachmann and Carsten Rother. Visual camera re-localization from rgb and rgb-d images using dsac. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5847–5865, 2021. 2, 5, 6
- [3] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6684–6692, 2017. 1, 5
- [4] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to re-localize in minutes using rgb and poses. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5044–5053, 2023. 1, 2, 4, 5, 6, 7
- [5] Samarth Brahmhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2018. 1, 2, 5, 6
- [6] David M. Chen, Georges Baatz, Kevin Koser, Sam S. Tsai, Ramakrishna Vedantham, Timo Pylvanainen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, Bernd Girod, and Radek Grzeszczuk. City-scale landmark identification on mobile devices. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2011. 2
- [7] Shuai Chen, Zirui Wang, and Victor Prisacariu. Direct-posenet: Absolute pose regression with photometric consistency. In *2021 International Conference on 3D Vision (3DV)*, 2021. 2
- [8] Shuai Chen, Xinghui Li, Zirui Wang, and Victor A Prisacariu. Dfnet: Enhance absolute pose regression with direct feature matching. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022. 1, 2, 3, 4, 5, 6, 7, 8
- [9] Shuai Chen, Yash Bhalgat, Xinghui Li, Jia-Wang Bian, Kejie Li, Zirui Wang, and Victor Adrian Prisacariu. Neural refinement for absolute pose regression with feature synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20987–20996, 2024. 1, 4, 5, 6
- [10] Shuai Chen, Tommaso Cavallari, Victor Adrian Prisacariu, and Eric Brachmann. Map-relative pose regression for visual re-localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20665–20674, 2024. 6
- [11] Manuela Chessa, Chiara Bassano, and Fabio Solari. Detection and localization of changes in immersive virtual reality. In *International Conference on Image Analysis and Processing*, pages 121–132, 2023. 1
- [12] Hiba Dahmani, Moussab Bennehar, Nathan Piasco, Luis Roldao, and Dzmitry Tsishkou. Swag: Splatting in the wild images with appearance-conditioned gaussians. In *European Conference on Computer Vision*, pages 325–340, 2025. 2
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3
- [15] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8092–8101, 2019. 2
- [16] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2, 6
- [17] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng. Complete solution classification for the perspective-three-point problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):930–943, 2003. 2, 6
- [18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1
- [19] Hugo Germain, Daniel DeTone, Geoffrey Pascoe, Tanner Schmidt, David Novotny, Richard Newcombe, Chris Sweeney, Richard Szeliski, and Vasileios Balntas. Feature query networks: Neural surface description for camera pose refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5071–5081, 2022. 1, 5, 6
- [20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2
- [21] Simon J. Julier and Jeffrey K. Uhlmann. New extension of the Kalman filter to nonlinear systems. In *Signal Processing, Sensor Fusion, and Target Recognition VI*, pages 182 – 193, 1997. 6
- [22] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2
- [23] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5974–5983, 2017. 4
- [24] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *IEEE international conference on computer vision*, pages 2938–2946, 2015. 1, 2, 4, 5, 6, 7
- [25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time

- radiance field rendering. *ACM Transaction on Graphics*, 42(4), 2023. 1, 2, 3
- [26] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 5, 6
- [27] Vincent Leroy, Yann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 6
- [28] Yiming Li, Zhiheng Li, Nuo Chen, Moonjun Gong, Zonglin Lyu, Zehong Wang, Peili Jiang, and Chen Feng. Multiagent multitaversal multimodal self-driving: Open mars dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22041–22051, 2024. 4, 5, 6
- [29] Yiming Li, Zehong Wang, Yue Wang, Zhiding Yu, Zan Gojcic, Marco Pavone, Chen Feng, and Jose M. Alvarez. Memorize what matters: Emergent scene decomposition from multitraverse. In *Advances in Neural Information Processing Systems*, 2024. 2
- [30] Jingyu Lin, Jiaqi Gu, Bojian Wu, Lubin Fan, Renjie Chen, Ligang Liu, and Jieping Ye. Learning neural volumetric pose features for camera localization. In *European Conference on Computer Vision*, 2024. 1, 2, 3, 4, 5, 6
- [31] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, and Wenming Yang. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5166–5175, 2024. 2
- [32] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 2
- [33] Changkun Liu, Shuai Chen, Yukun Zhao, Huajian Huang, Victor Prisacariu, and Tristan Braud. Hr-apr: Apr-agnostic framework with uncertainty estimation and hierarchical refinement for camera relocalisation. In *2024 IEEE International Conference on Robotics and Automation*, pages 8544–8550, 2024. 1, 5, 6
- [34] Changkun Liu, Shuai Chen, Yash Sanjay Bhalgat, Siyan HU, Ming Cheng, Zirui Wang, Victor Adrian Prisacariu, and Tristan Braud. GS-CPR: Efficient camera pose refinement via 3d gaussian splatting. In *The Thirteenth International Conference on Learning Representations*, 2025. 5, 6, 7
- [35] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2
- [36] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *IEEE International Conference on Computer Vision*, pages 2372–2381, 2017. 5, 7
- [37] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, pages 2794–2802, 2017. 4
- [38] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2
- [39] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1
- [40] 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 *European Conference on Computer Vision*, pages 405–421, 2020. 1
- [41] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanculescu, and Arnaud de La Fortelle. Cooridnet: uncertainty-aware pose regressor for reliable vehicle localization. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2229–2238, 2022. 2, 5, 6
- [42] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanculescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*, pages 1347–1356, 2022. 1, 2, 5, 6, 8
- [43] Arthur Moreau, Nathan Piasco, Moussab Bennehar, Dzmitry Tsishkou, Bogdan Stanculescu, and Arnaud de La Fortelle. Crossfire: Camera relocalization on self-supervised features from an implicit representation. In *IEEE/CVF International Conference on Computer Vision*, pages 252–262, 2023. 1, 5, 6
- [44] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *IEEE international conference on computer vision*, pages 3456–3465, 2017. 2
- [45] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019.
- [46] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020.
- [47] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1744–1756, 2016. 2
- [48] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018. 4, 5, 7
- [49] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3302–3312, 2019. 1, 2, 7

- [50] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 4
- [51] Yoli Shavit and Yosi Keller. Camera pose auto-encoders for improving pose regression. In *European Conference on Computer Vision*, pages 140–157, 2022. 5, 6
- [52] Yoli Shavit, Ron Ferens, and Yosi Keller. Paying attention to activation maps in camera pose regression. *arXiv preprint arXiv:2103.11477*, 2021. 2
- [53] Yoli Shavit, Ron Ferens, and Yosi Keller. Learning multi-scene absolute pose regression with transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 2733–2742, 2021. 1, 5, 6
- [54] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2930–2937, 2013. 4, 6, 7
- [55] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 7
- [56] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM Transactions on Graphics*, pages 835–846, 2006. 2
- [57] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 2
- [58] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019. 3
- [59] Gabriele Trivigno, Carlo Masone, Barbara Caputo, and Torsten Sattler. The unreasonable effectiveness of pre-trained features for camera pose refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12798, 2024. 1, 5, 6
- [60] Bing Wang, Changhao Chen, Chris Xiaoxuan Lu, Peijun Zhao, Niki Trigoni, and Andrew Markham. Atloc: Attention guided camera localization. In *AAAI Conference on Artificial Intelligence*, pages 10393–10401, 2020. 6
- [61] Fangjinhua Wang, Xudong Jiang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Glace: Global local accelerated coordinate encoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21562–21571, 2024. 1, 2, 5, 6, 7
- [62] Chen Wenbo and Liu Ligang. Deblur-gs: 3d gaussian splatting from camera motion blurred images. In *ACM in Computer Graphics and Interactive Techniques*, 2024. 2, 3
- [63] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In *European Conference on Computer Vision*, 2024. 2, 3
- [64] Boming Zhao, Luwei Yang, Mao Mao, Hujun Bao, and Zhaopeng Cui. Pnerfloc: Visual localization with point-based neural radiance fields. In *AAAI Conference on Artificial Intelligence*, pages 7450–7459, 2024. 1, 5
- [65] Enliang Zheng and Changchang Wu. Structure from motion using structure-less resection. In *IEEE International Conference on Computer Vision*, pages 2075–2083, 2015. 5
- [66] Qunjie Zhou, Maxim Maximov, Or Litany, and Laura Leal-Taixé. The perfect match: Exploring nerf features for visual localization. *arXiv preprint arXiv:2403.09577*, 2024. 1, 5