Reports on Geodesy and Geoinformatics, 2026, Vol. 121, pp. 50–58

DOI: 10.2478/rgg-2026-0004
Received: 27 November 2025 / Accepted: 4 April 2026
Published online: 29 April 2026

ORIGINAL ARTICLE

Single-image indoor localization using cross-domain learning from BIM models

Piotr Ryszko [1,2], Dorota Włodarczyk [1] and Małgorzata Jarząbek-Rychard [1*]

[1]Institute of Geodesy and Geoinformatics, Wrocław University of Environmental and Life Sciences, Grunwaldzka 53, 50-357, Wrocław, Poland
[2]Faculty of Information and Communication Technology, Wrocław University of Science and Technology, Wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław, Poland

*malgorzata.jarzabek-rychard@upwr.edu.pl

Abstract

Accurate indoor camera localization is crucial for applications in augmented reality, robotics, and autonomous navigation. While single-image deep learning models for 6-DOF pose regression have shown competitive results on established benchmarks, their development still requires extensive data annotation and hyperparameter tuning. In this work, we investigate the combination of advanced network architectures, transfer learning, and synthetic data to improve single-image indoor pose regression. Our approach employs a ResNet50 backbone pre-trained on the Places365 dataset and further trained and evaluated on established benchmarks. To enhance the training data, synthetic images are generated from 3D BIM models using Unreal Engine, with alignment procedures ensuring accurate correspondence between synthetic and real environments. Real RGB images are preprocessed to resemble synthetic data, enabling effective cross-domain evaluation. Experiments demonstrate that both architectural design and pretraining significantly influence model performance. On the UniMelb dataset (real-to-real scenario), the model achieves 0.21 m and 0.80° errors, surpassing baseline accuracy. We also present cross-validation and synthetic-to-synthetic experiments, providing insights into factors affecting performance and interactions between architecture, pretraining, and dataset characteristics.

Key words: indoor localization, camera pose estimation, deep learning, 3D models, BIM

1 Introduction

Indoor camera localization is essential for applications in augmented reality, robotics, and facility management. Unlike outdoor environments, where GNSS provides reliable positioning, indoor spaces require visual or sensor-based localization techniques. However, challenges such as repetitive structures, variable illumination, reflections, and occlusions reduce the distinctiveness of visual cues, introducing ambiguity into image-based pose estimation.

Traditional feature matching methods rely on detecting and describing keypoints in images, such as SIFT (Lowe, 2004) or SURF (Bay et al., 2006), and establishing correspondences between them to estimate camera poses or reconstruct 3D structure. Structure-from-Motion (SfM) pipelines (Furukawa and Ponce, 2010), which leverage such features, achieve high accuracy but require dense image coverage and tend to degrade in texture-poor or repetitive scenes (Nurutdinova and Fitzgibbon, 2015). In contrast, learning-based approaches, particularly single-image 6-DOF pose regression models like PoseNet (Kendall et al., 2015), estimate camera poses directly from RGB images. These models offer efficiency and real-time inference, but their performance is highly dependent on the quantity and quality of annotated training data.

Recent work has explored advanced architectures, temporal cues, domain adaptation, and the use of synthetic data generated from 3D models. Synthetic imagery is especially useful when real data are scarce, yet its effectiveness depends on model fidelity and

alignment with real scenes. Furthermore, modern feature extractors and cross-domain generalization are underexplored in existing indoor localization benchmarks.

In this study, we introduce ResNetPoseNet, a PoseNet-based architecture with a ResNet50 backbone pretrained on the Places365 dataset. We systematically evaluate how backbone selection, transfer learning, and synthetic data affect pose regression in indoor environments. Our dataset includes synthetic images generated from a BIM model in Unreal Engine, spatially aligned with the real environment, as well as real handheld RGB images used for method validation. Preprocessing ensures appearance consistency between domains by adapting real images to better match the visual characteristics of the synthetic data

The contributions are:

- ResNetPoseNet with ResNet50 backbone and multi-level regression heads, and empirical comparison with PoseNet on standard benchmarks, demonstrating improved feature extraction, stability, and localization accuracy in indoor environments.
- A newly developed custom indoor localization dataset, enabling evaluation and benchmarking of localization methods in a novel environment, and comprising both real-world data and three synthetic variants (cartoonish, photo-realistic, and Sobel).
- Systematic analysis of pretraining strategies, showing that while Places365 initialization works best for real-to-real experiments, ImageNet pretraining can outperform it in synthetic-to-real scenarios, highlighting the importance of matching the pretraining data to the domain.

Through these contributions, this work provides new insights into how architectural choices, pretraining strategies, and dataset characteristics influence indoor pose regression, and highlights the potential of synthetic–real data pipelines for improved performance.

2 Related work

Single-image localization has been widely investigated due to its importance in robotics, augmented reality, autonomous navigation, and digital twin applications. Traditional approaches typically rely on SfM pipelines to obtain large annotated datasets (Agarwal et al., 2009) with accurate camera poses. However, this reliance becomes a major limitation in environments where data acquisition is sparse, constrained, or time-consuming, motivating research on direct camera pose regression and synthetic-to-real learning strategies.

Early Absolute Pose Regression (APR) (Kendall et al., 2015; Kendall and Cipolla, 2016) frameworks demonstrated that deep neural networks can directly regress camera pose from a single image, with Kendall and Cipolla (2016) additionally modeling uncertainty in pose predictions using probabilistic approaches. Although promising, their generalization ability is often surpassed by geometry-based methods. To mitigate this issue, several works have proposed improved feature representations. For example, Bach et al. (2022) introduced FeatLoc, an APR method that directly processes sparse feature descriptors rather than relying on global feature embeddings. By exploiting simplistic view synthesis for data augmentation, their method exhibits significant improvements, up to 40% in translation accuracy on standard indoor datasets, showing the effectiveness of incorporating geometric cues in APR frameworks.

In outdoor urban environments, temporal coherence and structural cues have also been utilized to enhance localization robustness. Clark et al. (2017) proposed VidLoc, a video-clip-based 6-DOF relocalization method that leverages temporal information across consecutive frames, demonstrating significant improvements over single-image APR in dynamic and challenging scenarios. Similarly, Li et al. (2021) proposed VNLSTM-PoseNet, a hybrid convolutional network with LSTM modules that captures spatial–temporal dependencies for real-time 6-DOF relocalization in street-level scenes. Earlier work also explored spatial LSTMs for image-based localization (Walch et al., 2016). Although effective, these methods still rely heavily on large volumes of real-world training data and do not explicitly address the domain gap between synthetic and real imagery, which can limit generalization in new environments.

To overcome the limitations associated with scarce annotated data, several studies have explored the use of synthetic imagery generated from 3D models as an alternative training source (Acharya et al., 2019). Early work, such as research by Acharya (2020), demonstrated the feasibility of synthetic-to-real transfer for indoor visual localization using building models. Building on this foundation, Acharya et al. (2022) introduced a domain adaptation technique based on hierarchical edge maps and semantic segmentation, achieving an improvement in pose estimation accuracy compared to unadapted baselines. Further contributions include reverse domain adaptation by Acharya and Khoshelham (2023) and GAN-based style transfer for synthetic-to-real appearance alignment presented by Acharya et al. (2023). Collectively, these studies highlight the importance of feature-level and appearance-level adaptation when training on synthetic data.

Deep neural networks learn increasingly complex features across layers: the lower layers capture generic edges and textures, whereas deeper layers encode high-level structural elements such as doors, windows, or ceiling lights. These semantic and geometric cues function as visual landmarks essential for localization. Prior work by Peng et al. (2015) demonstrated that DCNN features extracted from synthetic renders exhibit strong invariance to color, texture, and background, suggesting that networks trained on synthetic geometry can generalize effectively to real images. Using a pretrained network substantially reduces the need for large, annotated datasets and avoids the computational burden of generating millions of renders. Instead, the model can be fine-tuned on smaller synthetic datasets tailored to the specific indoor environment. Following this idea, Acharya et al. (2019) showed that a DCNN fine-tuned exclusively on synthetic images is capable of accurate 6-DOF pose regression for real query images, thereby demonstrating that the domain gap can be successfully mitigated when high-level geometric features are preserved.

Recent research has further expanded these ideas by integrating domain adaptation directly into deep network architectures. Yao et al. (2024) proposed a multi-scale convolutional attention network, incorporating large kernel attention mechanisms and multi-level adversarial adaptation to address the challenges posed by data sparsity in civil engineering construction environments. Their findings demonstrate that the effectiveness of domain adaptation varies across feature layers, emphasizing the need for adaptive, hierarchical alignment strategies.

Despite significant advances in single-image localization, several key challenges remain. First, many methods rely heavily on large volumes of real-world annotated data, which are costly and time-consuming to collect, particularly in indoor environments with complex geometry and repetitive structures. Second, most studies focus either on real-to-real or synthetic-to-real scenarios without systematically examining how different synthetic rendering styles, feature representations, and pretraining strategies interact to affect cross-domain generalization (Sattler et al., 2019). Finally, current APR architectures often do not fully exploit multi-level feature extraction, nor do they investigate how backbone selection influences robustness in structured indoor spaces (Blanton, 2021).

Addressing these limitations, the present study aims to improve the robustness and accuracy of synthetic-to-real pose regression. To this end, we generate multiple synthetic datasets with varying characteristics and systematically evaluate how these differences affect model performance. In particular, we examine whether enhanced geometric fidelity and increased rendering diversity can strengthen the network's ability to transfer pose-relevant features

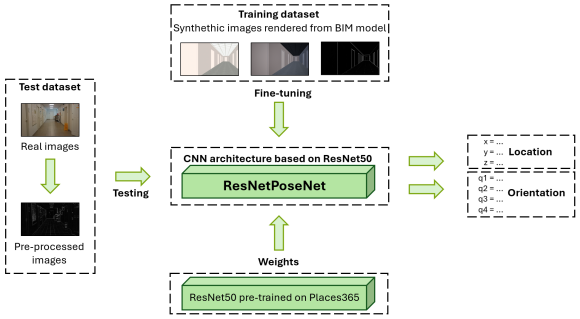


Figure 1. The design of the approach

from synthetic imagery to real indoor scenes. By combining advanced network architectures, transfer learning, and carefully designed synthetic datasets, our approach seeks to close the gaps identified in prior work and advance single-image indoor camera localization.

3 Methodology

The presented approach is based on the *PoseNet* architecture (Kendall et al., 2015), with design choices motivated by the insights from (Acharya et al., 2019), which highlighted the role of synthetic data generation and domain adaptation in enhancing the robustness of pose regression models. Figure 1 illustrates the overall design of our method. While *BIM-PoseNet* adopts the *GoogLeNet* backbone (Szegedy et al., 2015), we extend *PoseNet* by employing a deeper *ResNet* architecture (He et al., 2015), specifically the *ResNet50* variant.

Our modified networks are fine-tuned individually on different sets of synthetic images with known poses, rendered from a 3D indoor model. Prior to fine-tuning, the weights are initialized using a *ResNet50* pretrained on the *Places365* dataset (Zhou et al., 2018), which provides a large and diverse collection of scene-centric images. After fine-tuning, the networks are evaluated on real captured images to assess their 6-DOF indoor pose regression performance.

3.1 Network architecture and fine-tuning

The standard ResNet layers (conv1 through layer4) are used to extract hierarchical feature maps from the input image.

For pose regression, we introduce separate regression heads at multiple depths of the network (layer2, layer3, and layer4). Each regression head consists of:

- global average pooling (AdaptiveAvgPool2d),
- a fully connected layer mapping the pooled features to a 1024-dimensional vector,
- two parallel fully connected layers producing:
  - a 4-dimensional quaternion vector $\hat{\mathbf{q}}_i$ representing camera orientation,
  - a 3-dimensional vector $\hat{\mathbf{t}}_i$ representing camera translation.

During the forward pass, feature maps from layer2, layer3, and layer4 are passed through their respective regression heads, yielding six outputs:

$$\hat{\mathbf{q}}_2,\ \hat{\mathbf{t}}_2,\ \hat{\mathbf{q}}_3,\ \hat{\mathbf{t}}_3,\ \hat{\mathbf{q}}_4,\ \hat{\mathbf{t}}_4.$$

This multi-layer design allows the network to leverage features at different semantic levels for improved pose estimation.

During fine-tuning, the network weights are initialized from the pre-trained ResNet50, and the regression heads are trained using synthetic images with known poses. Predicted quaternions are normalized to unit length before evaluating the loss. The training objective, shown in (1), is defined as the weighted sum of mean squared errors (MSE) for translation and orientation, following the geometric loss formulation proposed in (Kendall and Cipolla, 2017):

$$\mathcal{L}(I) = \sum_{i=2}^{4} w_i \,\mathrm{MSE}(\mathbf{t}_i, \hat{\mathbf{t}}_i) \;+\; \beta \sum_{i=2}^{4} w_i \,\mathrm{MSE}\left(\mathbf{q}_i, \frac{\hat{\mathbf{q}}_i}{\|\hat{\mathbf{q}}_i\|_2}\right), \quad (1)$$

where $\mathbf{t}_i$ and $\mathbf{q}_i$ denote ground-truth translation and orientation, w_i are the regression-head weights ($w_2 = 0.15$, $w_3 = 0.15$, $w_4 = 0.7$), and $\beta = 600$ balances translation and rotation errors. The selection of the β parameter and regression head weights follows the empirical findings of Kendall et al. (2015), who suggested a functional range between 120 and 750. Specifically, we adopted a value of 600, consistent with the BIMPoseNet architecture (Acharya et al., 2019), as values beyond this threshold have been shown to correlate with increased localization errors due to an imbalance in the loss gradient. The use of MSE for both translation and quaternion regression also follows standard practice in PoseNet-like architectures.

3.2 Training procedure

Training was performed using the NAdam optimizer (Dozat, 2016) with a learning rate of 1×10^{-4} and a batch size of 40. The networks were trained for up to 160 epochs with early stopping if validation loss did not improve for 20 consecutive epochs. Optionally, learning rate scheduling was performed using ReduceLROnPlateau with a factor of 0.5 and patience of 10 epochs.

Data augmentation applied during training included:

- Color jittering with brightness, contrast, and saturation variations.
- Center cropping and resizing to 224×224.
- Conversion to RGB and normalization using the mean and standard deviation from the dataset used in pretraining, ensuring consistency.

3.3 Evaluation metrics

Model performance is evaluated using two metrics at each epoch:

- **Median Euclidean distance** between predicted and ground-truth translations (position accuracy).
  The position error (2) for each head is defined as:

$$e_{\mathrm{pos},i} = \left\|\mathbf{t} - \hat{\mathbf{t}}_i\right\|_2 \quad (2)$$

  where $\mathbf{t}$ denotes the ground-truth translation, and $\hat{\mathbf{t}}_i$ is the prediction from the i-th regression head.
  The model outputs predictions from three consecutive regression heads with weights $w_2 = 0.15$, $w_3 = 0.15$, and $w_4 = 0.7$. The weighted position error, shown in (3), is defined as:

$$E_{\mathrm{pos}} = \sum_{i=2}^{4} w_i \, e_{\mathrm{pos},i}. \quad (3)$$

- **Median angular error** between predicted and ground-truth quaternions (orientation accuracy), expressed in degrees.
  The orientation error (4) for each head is defined as:

$$e_{\mathrm{ori},i} = 2 \arccos\left(\left|\mathbf{q}^\top \hat{\mathbf{q}}_i\right|\right) \cdot \frac{180}{\pi} \quad (4)$$

  where $\mathbf{q}$ denotes the ground-truth unit quaternion and $\hat{\mathbf{q}}_i$ is the predicted quaternion.

The weighted orientation error, shown in (5), is computed as:

$$E_{\text{ori}} = \sum_{i=2}^{4} w_i \, e_{\text{ori},i}. \tag{5}$$

Finally, the reported position and orientation accuracies correspond to the median weighted errors over all samples in the training or validation epoch, as presented in (6):

$$\text{Acc}_{\text{pos}} = \text{median}(E_{\text{pos}}), \qquad \text{Acc}_{\text{ori}} = \text{median}(E_{\text{ori}}). \tag{6}$$

The decision to report median localization errors was specifically made to ensure consistency and a fair comparison with the foundational works in visual relocalization, namely Kendall et al. (2015) and Acharya et al. (2019). Since these studies establish the benchmark for the field using the median metric, adopting the same approach allows for a direct assessment of the performance of our model.

4 Data acquisition

For the purpose of this study, we developed a dedicated indoor localization dataset acquired in the building of the Institute of Geodesy and Geoinformatics at the Wrocław University of Environmental and Life Sciences. A detailed 3D BIM model of the facility was manually reconstructed in Revit from a terrestrial laser scanning (TLS) point cloud, enabling an accurate geometric representation of interior spaces. The presented experiment focuses on a building corridor, characterized by repetitive structural elements and moderate illumination changes. The following subsections describe the acquisition of real images as well as the generation of the synthetic datasets used in our experiments.

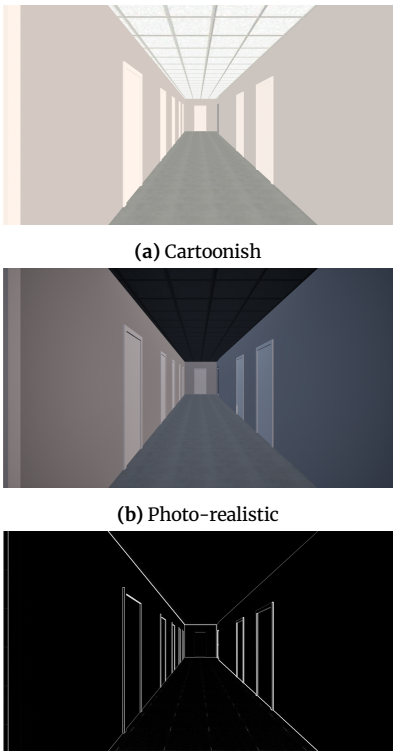
4.1 Synthetic data

Synthetic images were generated from a BIM indoor model visualized in Unreal Engine. Three visual variants of the dataset were produced (illustrated in Figure 2): (a) cartoon-style rendering, with pronounced illumination effects, (b) photo-realistic, with physically based lighting and shadow effects, (c) edge-enhanced Sobel-filtered, preserving prominent geometric contours. These variants allow us to assess how different appearance domains influence pose regression performance.

A virtual camera was animated along the reconstructed trajectory to render all three datasets. The resolution of the synthetic images was 1920×1080 pixels. Since exhaustively sampling all possible camera positions and orientations would be prohibitively expensive in terms of rendering time and computational cost, the images were rendered at 0.05 m spatial intervals along the trajectory, with additional perturbations of $\pm 10^{\circ}$ independently applied to each of the three rotation axes. Examples of these rotational perturbations are shown in Figure 3. This procedure yielded seven images per trajectory step and a total of 1920 images for each synthetic dataset.

4.2 Real data

To construct the real-world dataset, we first recorded a handheld video using a mobile phone at a frame rate of 30 fps in a horizontal orientation (16:9 aspect ratio). Subsequently, representative keyframes were extracted at a 10-frame interval to ensure sufficient spatial displacement between samples. The images were acquired using a sensor with a focal length of 4.266 mm and a pixel size of 0.00122 mm. These frames were processed and aligned in Agisoft Metashape software using a standard Structure-from-Motion pipeline, which enabled the estimation of accurate camera poses. A high-accuracy reference point cloud of the environment, acquired with a terrestrial laser scanner (Leica BLK360), served as a geometric baseline for the reconstruction.



(a) Cartoonish

(b) Photo-realistic

(c) Sobel

Figure 2. Different sets of synthetic images rendered from the 3D indoor model

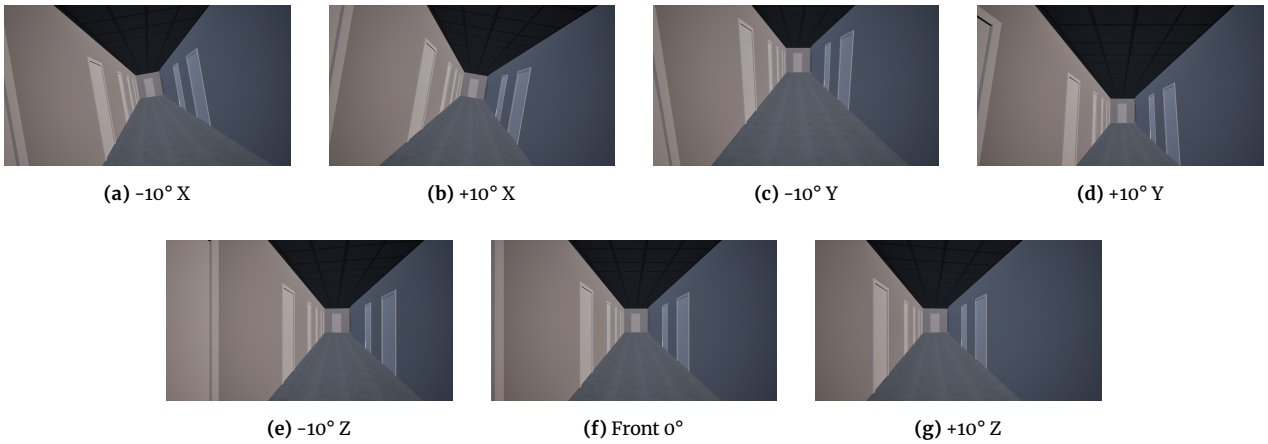
To ensure spatial consistency between the synthetic and real domains, the Unreal Engine model was aligned to the same reference point cloud. This was achieved by manually selecting corresponding characteristic points and computing the transformation matrix that registers the synthetic model with the real environment. As a result, the synthetic images used for training are geometrically consistent with the real-world scenes used for evaluation.

Finally, to enhance cross-domain comparability, we applied the same Sobel-based edge extraction to the real images as used in the synthetic pipeline (see Figure 4). This step acts as a lightweight pre-processing–level domain adaptation, ensuring that both domains share similar low-level structural representations during evaluation.

5 Experiments and results

Three experiments were conducted to evaluate the pose regression ability of a network.

- Experiment 1: This experiment aims to evaluate the baseline accuracy of our proposed *ResNetPoseNet* architecture in comparison to the original PoseNet, as presented in Kendall et al. (2015) and Acharya et al. (2019). We utilize the King's College (Kendall et al., 2015) and UniMelb (Acharya et al., 2019) real-world datasets for this comparison, focusing on both translation and rotation errors. The goal is to quantify the improvement offered by the ResNet50 backbone and multi-level regression heads in standard indoor environments, establishing a reference point for subsequent experiments involving synthetic and cross-domain data.
- Experiment 2: In this experiment, we investigate the effect of the pretraining dataset on the performance of *ResNetPoseNet*

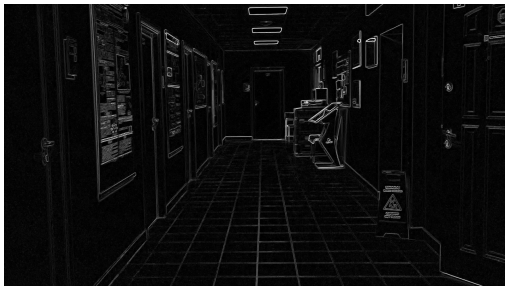


(a) -10° X **(b)** +10° X **(c)** -10° Y **(d)** +10° Y

(e) -10° Z **(f)** Front 0° **(g)** +10° Z

Figure 3. Synthetic images generated from the 3D BIM model in Unreal Engine. Top row: ±10° rotations around X and Y axes. Bottom row: ±10° rotations around Z with the central front view (0°). All images correspond to step-wise rendered frames for the Photo-Realistic dataset.

(a) Original real-world image



(b) Processed real-world image (Sobel)

Figure 4. Comparison between an original real-world frame and its processed version using Sobel filtering

across all evaluation scenarios, including real-to-real and various synthetic-to-real settings. Using the UniMelb datasets, we fine-tuned the network with two different initial weights: pretrained on *Places365* or on *ImageNet* (Deng et al., 2009). By keeping all other factors constant, this experiment isolates the impact of the pretraining dataset on both translation and rotation accuracy across six experimental setups, providing insights into the benefits of scene-specific versus general-purpose feature representations for indoor camera pose regression.

- Experiment 3: The purpose of this experiment is to assess the performance of our proposed *ResNetPoseNet* architecture on the custom indoor localization dataset introduced in Sec. 4. We conducted a comprehensive set of evaluations including real-to-real (real images for both training and testing), synthetic-to-synthetic (synthetic images for both training and testing), and cross-domain experiments (training on synthetic images and testing on real images). This setup allows us to quantify the generalization ability of our network across domains, and to evaluate how effectively the proposed architecture leverages synthetic data for real-world pose regression.

5.1 Baseline accuracy evaluation

The quantitative results in Table 1 provide a clear comparison between the baseline PoseNet and the proposed *ResNetPoseNet* architecture. On the King's College dataset, both methods achieve similar accuracy, with PoseNet reaching 1.92 m, 2.70°, and ResNetPoseNet obtaining 2.04 m, 3.67°, indicating comparable performance levels. However, a substantial improvement is observed on the UniMelb indoor dataset: PoseNet reaches 0.33 m, 1.85°, whereas ResNetPoseNet achieves 0.21 m, 0.80°. This corresponds to a 36% reduction in translation error and a 57% reduction in rotation error, demonstrating the advantage of using a ResNet50 backbone and multi-level regression heads in structured indoor settings.

These findings suggest that while PoseNet and ResNetPoseNet exhibit comparable performance in expansive outdoor environments, the proposed architecture is particularly effective at leveraging the dense structural features of indoor datasets. This indicates that the optimal choice of architecture is closely linked to the spatial characteristics of the target environment.

A qualitative assessment further supports these findings. Figure 5 visualizes the predicted camera trajectory produced by our best-performing *ResNetPoseNet* model. The red curve denotes the ground-truth path, whereas the colored points represent the model predictions, with the color scale indicating the position error. The predicted trajectory remains closely aligned with the ground truth across the entire sequence, which directly reflects the low localiza-

Table 1. Baseline localization accuracy on Kings College and UniMelb datasets

Scene	Frames [Train/Test]	PoseNet	ResNetPoseNet
Kings College	1200 / 343	1.92 m, 2.70 °	2.04 m, 3.67°
UniMelb	300 / 300	0.33 m, 1.85 °	**0.21 m, 0.80°**

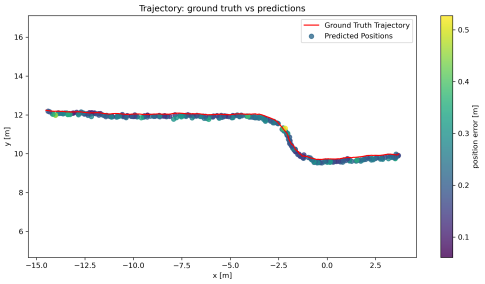


Figure 5. Trajectory comparison on the UniMelb indoor test sequence. The red curve denotes the ground-truth path, whereas the colored points represent the model predictions. The color scale indicates the position error.

tion error reported in Table 1.

These results demonstrate that *ResNetPoseNet* is highly suitable for robust indoor camera localization and provides a solid foundation for subsequent experiments involving synthetic data and cross-domain evaluation.

5.2 Influence of pretraining dataset across multiple scenarios

The results presented in Table 2 demonstrate a clear dependence of pose regression accuracy on the choice of pretraining dataset. To ensure a robust evaluation, these experiments were conducted on the UniMelb dataset benchmark, adopting the data categories and nomenclature established in Acharya et al. (2019). This includes synthetic models (Syn-car), photo-realistic renderings (Syn-photo-real), and gradient-based features (Syn-edge), allowing for a comprehensive assessment of the model's performance across diverse visual domains.

In earlier work, Kendall et al. (2015) initialized networks with weights pretrained from the *Places365* dataset, which are used as a starting point for camera pose regression networks. This strategy was later adopted in several subsequent studies, including Acharya et al. (2019), which also relied on *Places365* to initialize their models.

Our experiments confirm that for real-to-real training, *Places365* indeed provides a strong prior: the network pretrained on this dataset significantly outperforms the *ImageNet*-initialized counterpart. However, a different trend emerges in synthetic and cross-domain scenarios. For several synthetic-to-real configurations, particularly those involving edge-based or gradient-magnitude synthetic inputs, the *ImageNet*-pretrained model yields lower translation and rotation errors. This may be attributed to the fact that *ImageNet* features, being more object- and texture-centric, generalize better to the lower-level structural cues commonly found in synthetic renderings, whereas *Places365* is tuned for high-level scene semantics that synthetic renderings often fail to replicate accurately.

These observations indicate that pretraining selection should not be viewed as a universal solution. Instead, the optimal initialization should be matched to the specific characteristics of the training domain.

5.3 Evaluation on the custom dataset

The evaluation on our custom indoor dataset examines the behavior of the proposed *ResNetPoseNet* architecture across three complementary scenarios: within-domain synthetic testing, real-to-real fine-tuning, and cross-domain synthetic-to-real generalization.

Table 3 presents validation errors for networks fine-tuned and tested on the same synthetic domain. Each synthetic variant achieves sub-meter and low-degree errors, with the photo-realistic dataset showing the smallest error. This indicates that high geometric fidelity and realistic lighting conditions closely match the training distribution expected by the network.

Baseline performance obtained by fine-tuning and validating on real images is shown in Table 4. The real-to-real model reaches 1.34 m and 5.52°, which is substantially higher than validation errors observed in the synthetic domain. The larger error is consistent with the limited number of real training samples and more complex illumination changes and visual noise present in the real domain.

Cross-domain evaluation results are summarized in Table 5, where each synthetic model is tested directly on real images without real-image fine-tuning (except for Sobel, where minimal preprocessing consistency is applied). Among the three synthetic domains, the cartoonish fine-tuned model exhibits the lowest translation and rotation errors, while the Sobel and photo-realistic models show higher errors.

Figure 6 provides trajectory visualizations of all three synthetic-trained networks. The red line represents the ground-truth trajectory, and blue points correspond to predicted camera poses. The cartoonish model (Figure 6a) maintains the closest alignment with the reference trajectory, showing stable predictions with limited drift. The Sobel-based model (Figure 6b) produces a tightly clustered set of predictions, indicating reliance on local gradient structures that provide insufficient global context. The photo-realistic model (Figure 6c) follows the overall corridor geometry but exhibits higher dispersion and local instability, likely due to appearance mismatches between synthetic and real images.

Overall, these results suggest that synthetic data can be effective for indoor camera pose regression, but the degree of generalization depends on the visual style. Strong geometric cues, as in the cartoonish domain, appear more beneficial than purely edge-based representations or strictly photorealistic images. These findings highlight that designing synthetic renderings with consistent structural features may improve cross-domain performance, providing a foundation for future work on domain adaptation and synthetic data generation strategies.

6 Discussion

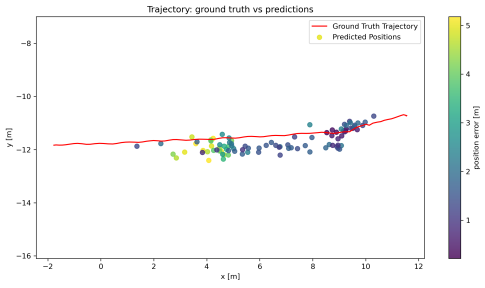
The experimental results presented in this work reveal several important insights regarding indoor camera pose regression and the role of architectural design, pretraining, and synthetic data in cross-domain generalization.

First, the baseline comparison confirms that the proposed *ResNetPoseNet* architecture is particularly well-suited for indoor environments. Although its performance on the outdoor King's College scene remains comparable to the original PoseNet, the significant improvement on the UniMelb dataset demonstrates that increasing model capacity through a ResNet50 backbone and introducing multi-level regression heads effectively enhances the extraction of stable geometric cues present in indoor spaces. Since our research focuses primarily on indoor localization, these results validate the underlying design choices.

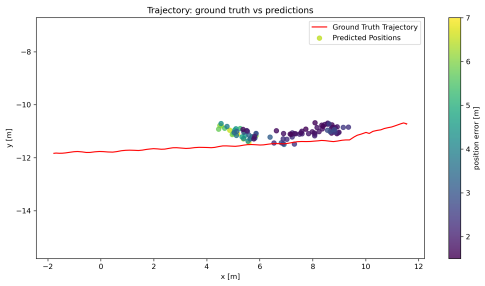
Second, the analysis of pretraining strategies shows that the commonly adopted approach – initializing networks with Places365 weights – remains optimal when both training and validation occur on real images. However, our results indicate that this assumption does not hold universally. In several synthetic-to-

Table 2. Impact of pretraining dataset (ImageNet vs. Places365) across real, synthetic, and cross-domain evaluation scenarios. For each pair, the better result is highlighted in bold.

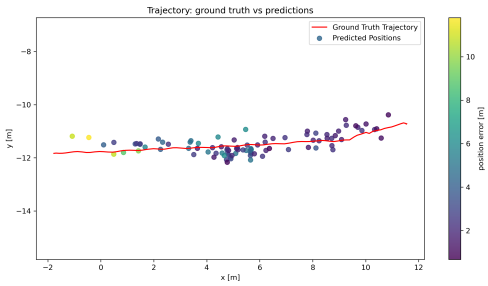
Fine-tuned on	Validated with	ImageNet	Places365
Real images (300)	Real images (300)	0.46 m, 2.14°	**0.21 m, 0.80°**
Syn-car (2505)	Real images (600)	14.83 m, 13.12°	**4.15 m, 21.64°**
Syn-pho-real (2505)	Real images (600)	11.19 m, 18.43°	**5.34 m, 13.19°**
Syn-pho-real-tex (2505)	Real images (600)	4.11 m, 9.82°	**3.05 m, 11.21°**
Gradmag-Syn-car (2505)	Gradmag of real images (600)	**2.50 m, 9.31°**	2.77 m, 9.22°
Syn-edge (2505)	Gradmag of real images (600)	**3.03 m, 7.44°**	2.96 m, 12.81°



(a) Cartoonish fine-tuned on real images



(b) Sobel fine-tuned on sobel-real images



(c) Photo-realistic fine-tuned on real images

Figure 6. Visualizations of the networks evaluated in cross-domain tests (see Table 5)

Table 3. Validation errors for each network architecture, evaluated within the same synthetic domain it was fine-tuned on

Fine-tuned on	Validated with	Validation error (m, deg)
cartoonish	cartoonish	0.25m, 1.01°
sobel	sobel	0.22m, 1.23°
photo-realistic	photo-realistic	**0.21m, 0.93°**

Table 4. Baseline performance: fine-tuning and validating on real images, used as a reference for cross-domain comparison

Fine-tuned on	Validated with	Validation error (m, deg)
real	real	**1.34m, 5.52°**

real scenarios, ImageNet pretraining led to lower pose errors than Places365. A likely explanation is that ImageNet features emphasize object- and texture-level representations, which align more closely with the structural simplicity and high-frequency cues found in synthetic datasets. By contrast, Places365, optimized for semantic understanding of real-world scenes, may struggle to transfer effectively to synthetic environments that lack realistic scene semantics. This finding challenges the default choice of pretraining dataset and highlights the importance of aligning feature initialization with the characteristics of the training domain.

Third, the evaluation on our custom synthetic datasets further emphasizes that visual realism alone is not a sufficient indicator of cross-domain performance. The *cartoonish* model, despite its low-fidelity appearance, generalizes to real images substantially better than the more visually realistic photo-realistic model. Conversely, the *Sobel*-based approach, which performs well in-domain, fails to generalize and collapses to a tightly clustered trajectory when evaluated on real data. This behavior suggests that relying solely on local gradient information restricts the model from learning global spatial patterns required for accurate 6-DOF localization.

Finally, the trajectory visualizations (Figure 6) provide further qualitative confirmation of these trends. The cartoonish model produces a trajectory reasonably aligned with the ground truth, the photo-realistic model exhibits moderate but consistent drift, while the Sobel-based model concentrates predictions in a narrow region, indicating a severe domain mismatch. These qualitative differences highlight that cross-domain success depends more on the stability and informativeness of geometric structures present in the synthetic data than on photorealism.

Overall, the findings suggest that indoor pose regression benefits from a combination of architectural robustness, appropriate

Table 5. Cross-domain evaluation: test errors on real images for networks fine-tuned on synthetic data. The reported results correspond to the trajectory visualizations shown in Figure 6.

Fine-tuned on	Validated with	Test error (m, deg)
cartoonish	real	**1.72m, 8.52°**
sobel	sobel-real	2.35m, 13.61°
photo-realistic	real	2.06m, 13.53°

pretraining strategies, and synthetic datasets that preserve global geometric consistency. The inconsistencies observed across domains underscore the need for further research into synthetic data design, domain adaptation, and representation learning for localization.

7 Conclusions

This work investigated the problem of indoor camera pose regression with a particular focus on architectural design, pretraining strategies, and the role of synthetic data in improving cross-domain generalization. The proposed *ResNetPoseNet* architecture, based on a ResNet50 backbone coupled with multi-level regression heads, was shown to provide an important performance improvement over the original PoseNet on indoor data. The gains observed on the UniMelb dataset confirm that higher-capacity feature extractors and multi-scale geometric representations offer clear advantages in structured indoor environments, where precise localization requires sensitivity to subtle spatial cues.

A second contribution is the construction and evaluation of a custom synthetic dataset comprising three distinct visual modalities: cartoonish, Sobel, and photo-realistic renderings. These datasets enabled an in-depth exploration of how different types of visual abstraction influence the network's generalization ability. Notably, the cartoonish dataset – despite its low visual fidelity – achieved the best cross-domain results, whereas the photo-realistic dataset did not transfer as successfully. The Sobel-based model, while highly consistent when validated within its own synthetic domain, achieved the weakest performance in the cross-domain real-image evaluation, where its predictions concentrated in a narrow region of the scene. These results suggest that global geometric consistency, rather than visual realism or high-frequency edge structure, is the key factor governing successful domain transfer for pose regression.

The third contribution of this work lies in the systematic analysis of pretraining strategies. While the field has predominantly adopted Places365 initialization, our experiments demonstrate that this practice does not generalize equally well across all domains. Places365 pretraining yielded the best results for real-to-real experiments, yet ImageNet initialization outperformed it in several synthetic-to-real scenarios. This finding challenges the standard assumption that scene-centric pretraining is universally optimal and highlights the need to match the pretraining dataset with the structural properties of the training domain.

Overall, the experiments demonstrate that indoor pose regression benefits from carefully designed architectures, thoughtful selection of pretraining datasets, and synthetic data that emphasize global structural cues. The findings also highlight the complex interplay between representation learning and domain characteristics, suggesting that effective localization cannot rely solely on model capacity or photorealistic training data. Instead, the field may benefit from more principled approaches to synthetic dataset construction and domain adaptation, which remain promising avenues for future research.

We identify two natural directions for extending this work. First, we plan to transition from indoor to city-level localization, potentially leveraging City Information Models (CIMs) combined with GPS-based approximate positioning and image-based pose refinement to enable both global and local camera alignment. Second, we aim to explore alternative network architectures, in particular visual transformers (Vaswani et al., 2017; Dosovitskiy et al., 2020), trained on large-scale synthetic datasets, together with *GAN*-based synthetic-to-real domain adaptation (Goodfellow et al., 2014; Bousmalis et al., 2016; Zhu et al., 2020), to improve cross-domain generalization and pose regression accuracy. These directions represent promising avenues for scalinge and further enhancinge our current indoor localization approach.

Acknowledgements

The research was supported by the Wrocław University of Environmental and Life Sciences (Poland) as part of the research project no. N060/0002/24.

References

Acharya, D. (2020). *Visual indoor localisation using a 3D building model*. PhD thesis, The University of Melbourne.

Acharya, D. and Khoshelham, K. (2023). Reverse domain adaptation for indoor camera pose regression. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-1/W1-2023:453–460, doi:10.5194/isprs-annals-X-1-W1-2023-453-2023.

Acharya, D., Khoshelham, K., and Winter, S. (2019). BIM-PoseNet: Indoor camera localisation using a 3D indoor model and deep learning from synthetic images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150:245–258, doi:10.1016/j.isprsjprs.2019.02.020.

Acharya, D., Tatli, C. J., and Khoshelham, K. (2023). Synthetic-real image domain adaptation for indoor camera pose regression using a 3D model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:405–421, doi:10.1016/j.isprsjprs.2023.06.013.

Acharya, D., Tennakoon, R., Muthu, S., Khoshelham, K., Hoseinnezhad, R., and Bab-Hadiashar, A. (2022). Single-image localisation using 3D models: Combining hierarchical edge maps and semantic segmentation for domain adaptation. *Automation in Construction*, 136:104152, doi:10.1016/j.autcon.2022.104152.

Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., and Szeliski, R. (2009). Building Rome in a day. In *2009 IEEE 12th International Conference on Computer Vision*, pages 72–79. doi:10.1109/ICCV.2009.5459148.

Bach, T. B., Dinh, T. T., and Lee, J.-H. (2022). FeatLoc: Absolute pose regressor for indoor 2D sparse features with simplistic view synthesizing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 189:50–62, doi:10.1016/j.isprsjprs.2022.04.021.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). SURF: Speeded Up Robust Features. In Leonardis, A., Bischof, H., and Pinz, A., editors, *Computer Vision – ECCV 2006*, pages 404–417, Berlin, Heidelberg. Springer Berlin Heidelberg.

Blanton, H. (2021). *Revisiting Absolute Pose Regression*. PhD thesis, University of Kentucky.

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. (2016). Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 95–104.

Clark, R., Wang, S., Markham, A., Trigoni, N., and Wen, H. (2017). VidLoc: 6-DoF Video-Clip Relocalization. *CoRR*, abs/1702.06521.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. doi:10.1109/CVPR.2009.5206848.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Dozat, T. (2016). Incorporating Nesterov Momentum into Adam. In *Proceedings of the 4th International Conference on Learning Representations (ICLR) Workshop*.

Furukawa, Y. and Ponce, J. (2010). Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, doi:10.1109/TPAMI.2009.161.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition.

Kendall, A. and Cipolla, R. (2016). Modelling uncertainty in deep learning for camera relocalization. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 4762–4769. doi:10.1109/ICRA.2016.7487679.

Kendall, A. and Cipolla, R. (2017). Geometric loss functions for camera pose regression with deep learning. *CoRR*, abs/1704.00390.

Kendall, A., Grimes, M., and Cipolla, R. (2015). Convolutional networks for real-time 6-DOF camera relocalization. *CoRR*, abs/1505.07427.

Li, M., Qin, J., Li, D., Chen, R., Liao, X., and Guo, B. (2021). VNLSTM-PoseNet: A novel deep ConvNet for real-time 6-DOF camera relocalization in urban streets. *Geo-spatial Information Science*, 24(3):422–437, doi:10.1080/10095020.2021.1960779.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60:91–110, doi:10.1023/B:VISI.0000029664.99615.94.

Nurutdinova, I. and Fitzgibbon, A. (2015). Towards Pointless Structure from Motion: 3D Reconstruction and Camera Parameters from General 3D Curves. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2363–2371. doi:10.1109/ICCV.2015.272.

Peng, X., Sun, B., Ali, K., and Saenko, K. (2015). Learning Deep Object Detectors from 3D Models.

Sattler, T., Zhou, Q., Pollefeys, M., and Leal-Taixé, L. (2019). Understanding the Limitations of CNN-based Absolute Camera Pose Regression. *CoRR*, abs/1903.07504.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. doi:10.1109/CVPR.2015.7298594.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need.

Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., Hilsenbeck, S., and Cremers, D. (2016). Image-based Localization with Spatial LSTMs. *CoRR*, abs/1611.07890.

Yao, D., Zhu, H., Ren, B., and Zhuang, X. (2024). Improving single image localization through domain adaptation and large kernel attention with synthetic data. *Engineering Applications of Artificial Intelligence*, 137:108951, doi:10.1016/j.engappai.2024.108951.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, doi:10.1109/TPAMI.2017.2723009.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2020). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.