Electronic Letters on Computer Vision and Image Analysis 24(2):273-285, 2025

Global Completion of Occupancy Maps Using Generative Models

Selma Güzel* and Erkan Uslu*

* *Department of Computer Engineering, Yildiz Technical University, Davutpasa, Istanbul, Turkiye*

Received 27th of april, 2025; accepted 2nd of december, 2025

Abstract

We address *global occupancy–map completion* from partial, sequential observations by framing it as conditional image generation. Using HouseExpo/SUNCG floorplans, we build a Gazebo+ROS pipeline that collects time–ordered exploration data and release a 128×128 dataset for reproducible benchmarking.

Six model families are compared under a unified setup: Conditional VAE (CVAE), pix2pix–hinge Conditional GAN (CGAN), RePaint–style diffusion (DDPM), Vision Transformer (ViT) encoder–decoder, residual UNet, and a geometric, training–free baseline (GEOM). **UNet** achieves the lowest pixelwise errors (L1/MSE/RMSE) and the best discrete scores (IoU/F1/Accuracy). **GEOM**, despite being training-free, attains the highest SSIM—demonstrating that explicit geometric priors can rival data-driven learning when structural regularity dominates. Diffusion offers strong perceptual quality but at higher computational cost.

Our findings reveal that occupancy-map completion is fundamentally a structure-exploitation task: convolutional architectures inherently favor rectilinear layouts, while attention-based models require more data to learn comparable priors. We release the dataset, ROS pipeline, and all implementations to enable reproducible research in generative mapping.

Keywords— Generative Models, Conditional VAE, Conditional GAN, Diffusion, Vision Transformer, UNet, Map Completion, Gazebo, ROS.

1 Introduction

Map completion—the task of predicting unseen regions in occupancy grids—is essential for autonomous exploration and navigation. We consider both **inpainting** (filling holes) and **outpainting** (extrapolating beyond observed bounds), the latter being particularly challenging due to sparse priors. Because on-board computation is limited, we focus on *lightweight, efficient models* and analyze accuracy–throughput trade-offs using a comprehensive suite of metrics—pixelwise error (L1, MSE, RMSE [19]), perceptual/structural similarity (SSIM, UQI [21, 20]), global distortion and spatial consistency (ERGAS, SCC [22, 23]), and discrete overlap/classification (IoU, F1, Accuracy [17, 18]).

Correspondence to: selma.guzel@std.yildiz.edu.tr

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.2162
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

We benchmark six diverse generative and geometric approaches on single-channel **128×128** maps: a **CVAE** [39, 47], a **CGAN** (pix2pix) [36, 37], a diffusion-based **RePaint** model [41, 42, 31], a compact **ViT** encoder–decoder [40], a residual **UNet** [38], and a geometric, training-free baseline (**GEOM**) [27, 28]. Evaluation is performed on a new sequential Gazebo+ROS dataset derived from HouseExpo [3] (built in Gazebo [13] and ROS [14]), containing **3200/361/426** train/val/test maps with preserved temporal order.

Scope. Our primary goal is a fair, compute-normalized comparison of diverse paradigms—probabilistic (CVAE), adversarial (CGAN), diffusion-based (RePaint), attention-based (ViT), convolutional (UNet), and rule-based (GEOM)—under limited time and hardware constraints. These design choices (short training budgets, compact architectures) also align with potential onboard retraining scenarios during active exploration.

Findings. **UNet** achieves the best aggregate performance (lowest L1/MSE/RMSE; highest IoU/F1/Accuracy), while **GEOM**, despite being training-free, attains the highest **SSIM**—demonstrating that explicit structural priors can rival learned models when geometric regularity dominates. Diffusion (**RePaint**) provides high perceptual realism but slower inference; in contrast, **CVAE**, **CGAN**, **ViT**, and **UNet** exceed 20 img/s, supporting real-time deployment.

Unlike prior works that rely on heavy preprocessing (e.g., GPS filtering [43, 44] or voxel smoothing [46]), our models learn directly from temporally paired occupancy grids, allowing them to handle noise and misalignment without manual intervention.

Problem formulation. An occupancy grid is $M \in \{0, 1, 2\}^{H \times W}$ (free/occupied/unknown). Given a partial map M_{obs}, a generative model $\mathcal{G}_\theta$ predicts missing regions $\hat{M}_{\text{miss}} = \mathcal{G}_\theta(M_{\text{obs}})$, producing a completed map $\hat{M} = M_{\text{obs}} \cup \hat{M}_{\text{miss}}$. Figure 1 summarizes the end-to-end pipeline.

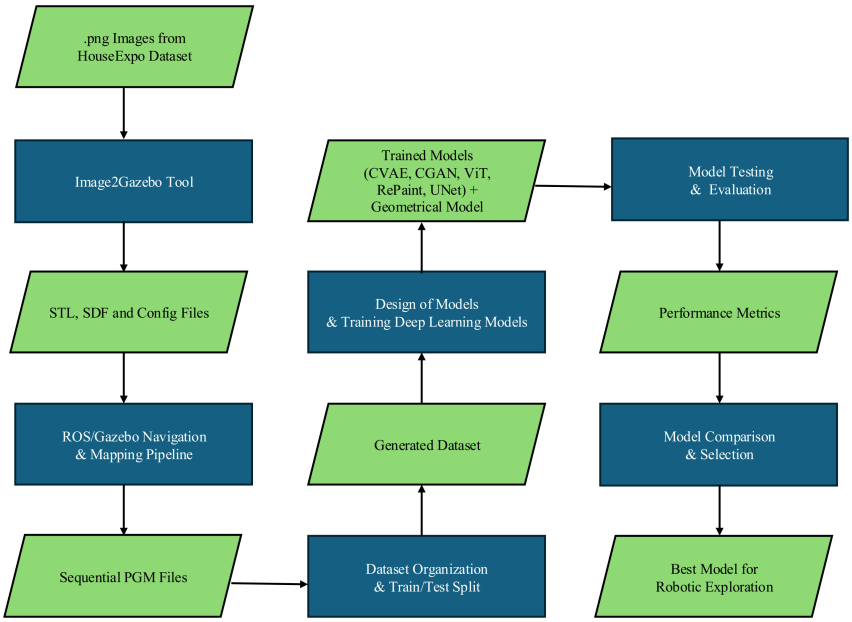


Figure 1: End-to-end pipeline: from HouseExpo floorplans through Gazebo/ROS data collection to model training and evaluation.

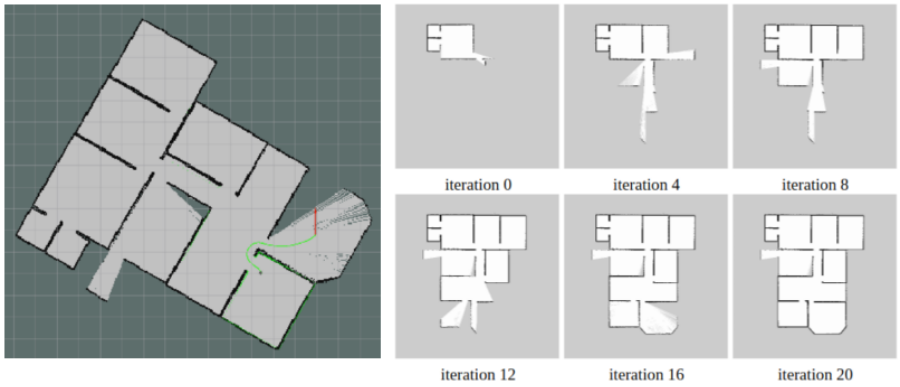


Figure 2: Left: RVIZ during navigation to a Voronoi point. Right: Sequential data captured between Voronoi points.

2 Background

Map completion. Occupancy grids are a standard, efficient representation for indoor mapping [2]. Early learning-based studies, such as CNN and pix2pix-style CGANs trained on HouseExpo [3], demonstrated that partial-to-complete synthesis is feasible, though often with motor-based simulation lacking realistic sensor noise and weak conditioning on observed structure. Later GAN variants [4, 25] improved local realism but still lacked explicit *structure-aware conditioning* for global layout.

Global vs. local prediction. Global map prediction forecasts unseen regions for exploration, while local completion fills nearby cells. Recent works extend high-capacity inpainting models (LaMa, RePaint, Stable Diffusion [29, 31, 32]) to mapping; MapEx [7] and Spinos et al. [30] adopt these for global extrapolation, whereas SenseExpo [33] targets local prediction using lightweight Transformer–Fourier designs [34, 35]. CGAN and transformer-based image completion [10, 11, 9] provide architectural precedents.

Analysis and critique. Across prior studies, three gaps persist: (i) weak conditioning on observed structure, (ii) non-sequential or limited-scale datasets that hinder generalization, and (iii) inconsistent metrics that conflate perceptual realism with geometric accuracy. Moreover, heavy models improve appearance but often exceed onboard compute limits.

This study. We address these gaps through a unified, compute-normalized benchmark of six lightweight families—**CVAE**, **CGAN**, **RePaint** diffusion, **ViT**, **UNet**, and a rule-based **GEOM** baseline [27, 28]. Unlike prior work focused on local completion [8], we target global map completion and compare six diverse families on a new sequential Gazebo+ROS dataset (derived from HouseExpo [3]) with coherent temporal splits and comprehensive accuracy–throughput evaluation spanning pixelwise, perceptual, and topological metrics. This provides systematic insights into how architectural choices and computational budgets jointly determine completion fidelity and runtime feasibility—both critical for autonomous navigation.

3 Dataset

Source and collection. We build on **HouseExpo** [3] (from **SUNCG** [12]) and use **Gazebo** [13]+**ROS** [14] to generate sequential exploration data for **128**×**128** global map completion. HouseExpo PNGs are converted to Gazebo worlds at `0.01 m` resolution; a TurtleBot3 navigates between Voronoi points via `MoveBaseClient`, saving maps with `map_saver` at each step. This process is repeated for all scenes to produce time-ordered trajectories reflecting realistic exploration—enabling evaluation closer to actual deployment. Figure 2 shows the pipeline.

Preprocessing & loader. A fixed-content cropper detects the target's content box (Canny+contours), ap-

plies the same crop to the input, then pads/resizes both to **128×128** (nearest-neighbor). Pixels are quantized to $\{0, 205, 254\}$ and normalized to $\{0.0, 0.5, 1.0\}$ (free/unknown/occupied). No augmentation (`augment=false`); preprocessing is deterministic (`margin=10`, `wall_thickness=3`). Dataloader: `size`=128, `batch_size`=32, `num_workers`=4, reading paired `input/`–`output/` files by basename.

Splits & curation. Final split: **3200** train, **361** val, **426** test. Entire scene sequences stay in a single split to avoid leakage and preserve temporal coherence. Near-duplicate Voronoi frames were pruned. Since automatic thresholds (SSIM/UQI/L_1) proved unreliable, representative frames were selected manually before training. Manifests are released (see Data Availability).

4 Map Completion Models

Common setup. We evaluate six families on single–channel **128×128** occupancy maps: CVAE, CGAN (pix2pix), RePaint, ViT encoder–decoder, residual UNet, and rule–based GEOM [39, 36, 31, 40, 38, 27]. Inputs are partial maps x, targets full maps y. Models use Adam (`batch_size=32`, `ReduceLROnPlateau`); CVAE uses pyramid conditioning, others concatenate or directly ingest x. Outputs are logits ℓ with

$$\sigma(\ell) = \tfrac{1}{1+e^{-\ell}}, \quad \mathrm{softmax}(\ell)_c = \tfrac{e^{\ell_c}}{\sum_k e^{\ell_k}}, \quad \mathrm{snap}(g) = \arg\min_{v\in\{0,0.5,1\}} |g - v|.$$

This unified setup ensures fair, consistent evaluation across architectures.

4.1 Conditional VAE (CVAE)

A 3-down/3-up CVAE [39, 47] is implemented, where the encoder processes y to form $q_\phi(z \mid y)$, and the decoder conditions on a three-level average–pooled pyramid of the partial map x (no encoder–decoder skip connections). Outputs are logits $[B, 1, 128, 128]$, and sampling uses either a prior draw or the posterior mean.

Objective. Reconstruction uses BCE-with-logits (default) or optionally MSE on $\sigma(\ell)$, combined with a β-weighted KL term under linear annealing. Masking is disabled in our main runs:

$$\mathcal{L}_{\text{recon}} = \begin{cases} \mathrm{BCE_with_logits}(\text{logits}, y) & \text{(default)} \\ \mathrm{MSE}(\sigma(\text{logits}), y) & \text{(optional)} \end{cases}$$

$$\mathcal{L}_{\text{KL}} = -\tfrac{1}{2}\,\mathbb{E}\big[\,1 + \log\sigma^2 - \mu^2 - \sigma^2\,\big], \qquad \mathcal{L} = \mathcal{L}_{\text{recon}} + \beta_{\text{eff}}\,\mathcal{L}_{\text{KL}}.$$

Here $\beta_{\text{eff}} = \beta\cdot\mathrm{anneal}(e; s{=}0,\ e{=}200)$ with $\beta{=}0.2$. This model highlights the trade-off between reconstruction fidelity and latent regularization, which is critical for fast and stable inference in robotic mapping.

Relation to prior work. Compared to [8], which uses concatenation-based conditioning and a composite $\mathrm{MSE}{+}(1{-}\mathrm{SSIM}){+}\beta\,\mathrm{KLD}$ loss, we adopt *pyramid conditioning*, a lightweight 3 + 3 structure, and a simpler recon+KL objective to maintain efficiency and ensure fair cross-model comparison.

4.2 Conditional GAN (pix2pix)

We implement a UNet generator with bilinear upsampling and a spectral–normalized PatchGAN discriminator [36, 37]. The discriminator scores $[x,\, y]$ versus $[x,\, G(x)]$. We employ a hinge–GAN objective with strong L_1 regularization ($\lambda{=}100$) to balance sharpness with stability:

$$\mathcal{L}_D = \mathbb{E}[\max(0, 1 - D(x, y))] + \mathbb{E}[\max(0, 1 + D(x, G(x)))]\,,$$

$$\mathcal{L}_G = -\mathbb{E}[D(x, G(x))] + \lambda\|G(x) - y\|_1, \quad \lambda{=}100.$$

The CGAN captures texture-level completeness and provides realistic structural inpainting, complementing the CVAE's smoothness with sharper boundaries.

4.3 RePaint Diffusion (Conditioned DDPM)

A conditioned DDPM [41] is employed, where the UNet predicts $\epsilon_\theta(x_t, t, x)$ from the concatenation $[x_t, x]$ with learned timestep embeddings. We use a linear β schedule (T=1000). At test time, DDIM sampling (50 steps, η=0) [42] is applied while re-imposing known pixels of x at each step; RePaint-style jump–resampling [31] is also available. Training minimizes noise prediction plus an SSIM auxiliary term:

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t, x)}{\sqrt{\bar{\alpha}_t}}, \qquad \mathcal{L} = \mathbb{E}\|\epsilon - \epsilon_\theta\|_2^2 + w_{\text{SSIM}}\big(1 - \text{SSIM}(\hat{x}_0, y)\big).$$

Although diffusion offers high perceptual quality, it trades speed for fidelity, which is relevant for deployment trade-offs in real-time exploration.

4.4 Vision Transformer (ViT Enc–Dec)

A compact ViT [40] with non-overlapping 16×16 patches (64 tokens), D=256, H=4, and L_{enc}=L_{dec}=3 is used. A transposed convolution restores the image to $[B, 1, 128, 128]$. **Objective:** PyTorch *Smooth-L_1* on predictions (default, β=0.1), or *BCE-with-logits* as a drop-in alternative. This configuration reflects a balance between transformer-based global reasoning and lightweight design suitable for embedded robotic applications.

4.5 UNet (3 + 3 Residual)

A residual UNet [38] with three downs/ups, InstanceNorm, bilinear upsampling, and `out_ch`=3 logits for classes $\{0, 0.5, 1\}$ is trained using class–weighted CE with mild label smoothing. Both discrete and grayscale metrics are reported. The grayscale expectation used for L_1/MSE is:

$$p_c = \text{softmax}(\ell)_c, \quad v = \{0, 0.5, 1\}, \qquad \hat{g} = \sum_{c=1}^{3} p_c\, v_c \in [0, 1].$$

Discrete metrics employ $\arg\max_c p_c$. This model yields strong reconstruction consistency, showing why convolutional inductive biases remain powerful in structured spatial prediction.

4.6 GEOM (Rule–Based)

A training–free geometric baseline [27] detects near–axis guide rails (Canny → Hough), locates doorway gaps under width/support constraints, and fills bounded rectangles while respecting a 1-px stop mask (walls ∪ guides). Outputs are quantized to $\{0, 0.5, 1\}$; behavior is governed by door–width/support and guide–thickness thresholds. Although deterministic, GEOM provides insight into the limits of purely geometric reasoning and serves as a valuable reference for learned models.

5 Experimental Setup

Dataset and simulation. We employ the sequential HouseExpo–Gazebo–ROS pipeline (Sec. 3): Gazebo 11, ROS Noetic, Voronoi navigation via `MoveBaseClient`, maps saved with `map_saver` at `0.01 m` and resized to **128×128**. Splits: **3200/361/426** (train/val/test); sequences remain unshuffled to preserve temporal coherence.

Frameworks and hardware. PyTorch 1.13.1 (CUDA), Python 3.8, OpenCV 4.5 [15]; Gazebo/ROS in C++14. Hardware: Intel Core i9, **64 GB** RAM, NVIDIA GeForce RTX 4070 Laptop GPU (8 GB VRAM).*

*To streamline implementation, **ChatGPT** [16] was employed as a coding assistant under author supervision.

Table 1: Test-set results (**FULL** scope). Left: pixel/perceptual; right: discrete after snapping to $\{0, 0.5, 1\}$. Best per column in **bold**.

Model	L1 ↓	MSE ↓	RMSE ↓	SSIM ↑	UQI ↑	ERGAS ↓	IoU ↑	F1 ↑	Acc ↑
CVAE	0.0672	0.0324	0.1795	0.7486	0.7856	25.00	0.702	0.815	0.882
CGAN	0.0649	0.0367	0.1911	0.7519	0.7738	26.64	0.716	0.828	0.882
RePaint	0.0718	0.0312	0.1760	0.8179	0.8120	24.50	0.772	0.869	0.901
ViT (16)	0.0812	0.0344	0.1850	0.7168	0.7672	25.76	0.657	0.775	0.873
UNet (3+3)	**0.0453**	**0.0262**	**0.1610**	0.8279	**0.8366**	**22.43**	**0.796**	**0.884**	**0.916**
GEOM (CPU)	0.0529	0.0287	0.1686	**0.8484**	0.8153	23.50	0.779	0.874	0.898

Common training and validation. All models use Adam optimizer with batch size **32**, gradient clipping at 1.0, `ReduceLROnPlateau` (patience 6, $\eta_{\min} = 10^{-6}$), and seed **42**. Inputs are single-channel *partial* maps; targets are *complete* maps. Metrics: continuous (**L1**, **MSE**, **RMSE**) and discrete (**IoU/mIoU**, **F1**, **Accuracy**) after snapping to $\{0, 0.5, 1\}$, computed over **FULL** image for consistent comparison.

Model-specific configurations. *CVAE* uses BCE-with-logits $+\beta$KL (β=0.2) with KL annealing, multi-scale pyramid conditioning (Sec. 4.1). *CGAN* (Sec. 4.2): hinge-GAN, L_1 regularization (λ_{L_1}=100), spectral-normalized PatchGAN. *RePaint* (Sec. 4.3): timestep-conditioned U-Net, diffusion + SSIM loss, DDIM evaluation (50 steps, η=0). *ViT* (Sec. 4.4): compact encoder-decoder, 16×16 patches, Smooth-L_1 (β=0.1). *UNet* (Sec. 4.5): residual 3+3, class-weighted CE (smoothing 0.05), 3-class logits; grayscale via softmax expectation. *GEOM* (Sec. 4.6): training-free (CPU), geometric door/guide detection with constrained filling.

Hyperparameter tuning. All models received equal wall-clock tuning budget on validation data. Core parameters (learning rate, batch size, regularization) were adjusted; architecture-specific settings (CGAN λ=100, diffusion steps= 50) followed validated defaults. Best validation-IoU configurations were selected for balanced comparison.

Epoch budgets and wall-clock time. To ensure comparable effort, epoch budgets yield similar wall-clock times: CVAE ≈ 3,500; CGAN, UNet, RePaint = 1,000; ViT ≈ 10,250; GEOM: none.

Training horizon and design rationale. We omit early stopping to equalize runtime and simulate short-horizon, resumable learning for resource-constrained platforms. This enables fair accuracy-throughput assessment.

Secondary consideration: dynamic updates. Short budgets and lightweight architectures allow potential on-board fine-tuning during exploration, motivating our emphasis on accuracy and time-to-train for real-time adaptation.

6 Experimental Results and Evaluations

We evaluate models on **128×128** occupancy maps with a batch size of 32. Continuous predictions are snapped to $\{0, 0.5, 1\}$ for discrete scores (IoU/F1/Acc), while pixel- and perceptual-level metrics use raw outputs.[†] Table 1 summarizes the quantitative results.

Summary. UNet (3+3) delivers the best overall accuracy (lowest L1/MSE/RMSE, highest IoU/F1/Acc) and strong generalization. Training-free **GEOM** achieves the top **SSIM** (0.8484), showing that geometric priors can rival learned models under structured layouts. **RePaint** ranks next in perceptual quality but is slower due to iterative sampling. Overall, learned models capture spatial continuity and noise, while rule-based methods excel under strong geometric regularity.

[†]*Protocol.* Metrics are computed in **FULL** (entire image). We report pixelwise error (**L1**, **MSE**, **RMSE**), perceptual/structural similarity (**SSIM**, **UQI**), global distortion (**ERGAS**), spatial consistency (**SCC**), and discrete overlap/classification quality (**IoU**, **F1**, **Accuracy**) [21, 17, 20, 19, 22, 23, 4, 18].

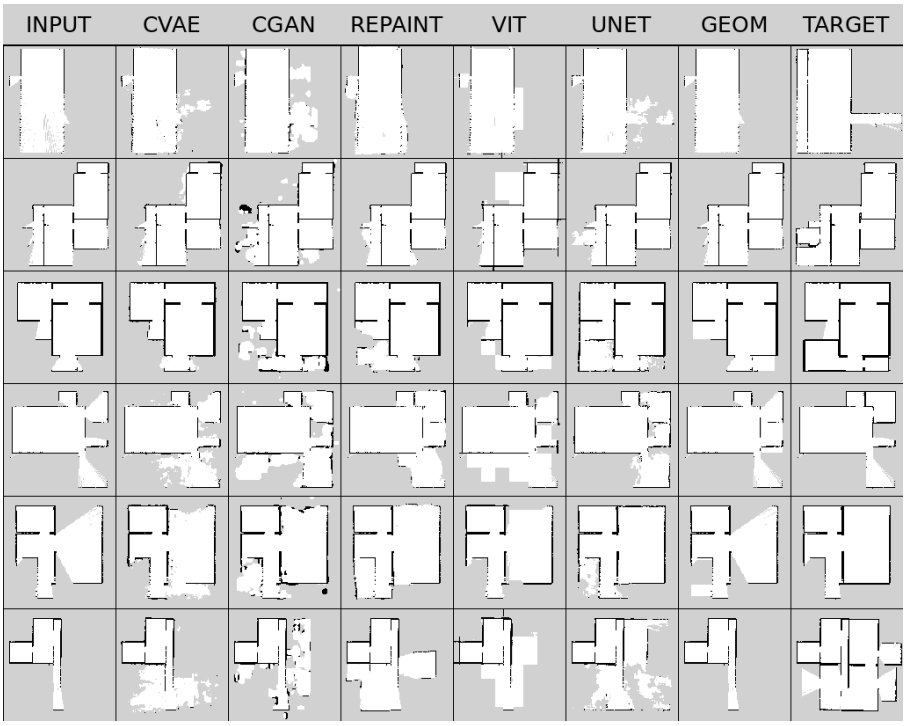


Figure 3: Qualitative comparison showing inputs, model predictions, and ground-truth targets across representative test scenes.

6.1 Qualitative Results

Figure 3 illustrates representative completions across models. **UNet (3+3)** yields crisp wall boundaries and preserves corridor topology via skip connections retaining fine-grained details. **RePaint** demonstrates strong large-hole completion through iterative denoising but sometimes over-completes or under-completes uncertain regions, blending weak or ambiguous boundaries. Both surpass **GEOM** in overall completeness, highlighting the benefit of learned priors when geometric cues are missing. Typical errors include hallucinated walls, incomplete doorways, and sensitivity to threshold parameters in **GEOM**. These observations suggest that hybrid geometric–generative approaches, coupling structural priors with learned adaptability, may offer improved reliability for real-world deployment.

6.2 Comparison with [2]

[2] (MCN-SLAM, U-Net, **64**×**64**) reports **F1**=20.4963 and **Acc**=20.5852. At a higher resolution of **128**×**128**, our **UNet (3+3)** achieves **F1**=20.8837 and **Acc**=20.9165, representing a substantial improvement. Full-frame metrics are: L1=20.0453, MSE=20.0262, RMSE=20.1610, IoU=20.7962, SSIM=20.8279, UQI=20.8366, ERGAS=222.43, and SCC=20.8376. Binary results are: Acc=20.9235, Prec=20.9261, Rec=20.9280, F1=20.9269, and IoU=20.8645. Since the datasets and resolutions differ, this comparison remains qualitative rather than direct.

6.3 Runtime and Throughput

On the same hardware (Sec. 5), feed-forward models (CVAE/CGAN/ViT/UNet) run at $>$20 img/s with $<$50 ms/map latency. Diffusion (**RePaint**) is slower ($\sim$220 ms/map, 50 DDIM steps) but offers strong perceptual quality. Training-free **GEOM** runs on CPU at $\sim$52 ms/map. These profiles reveal a speed-fidelity trade-off critical for resource-constrained robotic platforms.

7 Discussion

Overall trends. The residual **UNet** achieves the best overall trade-off, combining high accuracy (lowest L1/MSE/RMSE, best IoU/F1/Acc) with fast inference. Diffusion-based **RePaint** offers strong perceptual plausibility and flexible large-region completion, though with variable boundary fidelity and higher latency. The training-free **GEOM** remains competitive, showing that simple geometric priors can rival or exceed learned models in structural consistency.

Runtime–quality trade-offs. Feed-forward models (CVAE, CGAN, ViT, UNet) exceed 20 img/s at 128×128 with < 50 ms/map latency, enabling real-time use. Diffusion (**RePaint**) is slower ($\sim$220 ms/map) but yields strong perceptual fidelity, while lightweight **GEOM** runs on CPU ($\sim$52 ms/map) without training. Overall, they form a clear speed–quality spectrum: feed-forward for speed, diffusion for detail, and GEOM for low-power mapping.

Failure modes. Errors include both over- and under-completion near frontiers. Diffusion models may fill unseen rooms prematurely or leave them incomplete under weak context. Some apparent over-fills arise from the evaluation protocol—targets capture intermediate exploration states, so early correct completions are penalized as false positives. Blurring of 1–2 px walls, **GEOM**'s threshold sensitivity, and difficulty with thin or oblique walls persist, underscoring the need for geometric constraints and uncertainty-aware decoding.

Practical guidelines. Ternary snapping $\{0, 0.5, 1\}$ should be applied for discrete metrics, while continuous scores use raw predictions. Maintaining full exploration sequences prevents temporal leakage, and modest label smoothing or class weighting stabilizes the underrepresented "unknown" class—improving reproducibility and realism.

Regression vs. classification. While CVAE, CGAN, ViT, and RePaint regress continuous grayscale values, **UNet** performs explicit 3-class classification. Because discrete metrics (IoU/F1/Acc) are computed on snapped maps, this naturally favors classification-based heads. Hybrid objectives combining regression and classification could narrow this gap and better align optimization with evaluation.

Architectural biases. Rectilinear occupancy grids favor models with local receptive fields; CNNs naturally capture axis-aligned walls. Skip connections in **UNet** preserve spatial detail, explaining its strong L1 and IoU. The training-free **GEOM** achieves the highest SSIM, showing structural priors can substitute for deep features, while **RePaint** sometimes merges nearby structures by blending weak boundaries.

Sources of underperformance. Transformers require larger datasets to learn spatial priors that CNNs obtain inductively. Coarse 16×16 patching blurs thin structures, contributing to a 14% IoU gap. CVAE and diffusion models may hallucinate plausible but nonexistent regions; CGANs, though sharper, destabilize under limited data. **UNet**'s discrete formulation aligns better with IoU/F1 objectives, whereas continuous regressors—requiring post-hoc snapping—are slightly disadvantaged.

Robotic implications. In navigation, even 1–2 px wall deviations can yield false frontiers or sealed passages, degrading coverage and safety. The **UNet**'s precise wall localization and **GEOM**'s structural reliability mitigate such risks. Adversarial models may introduce spurious obstacles, and diffusion models can over- or under-fill uncertain regions, potentially misleading exploration. Balancing generative flexibility with geometric reliability is key to safe deployment.

Limitations. Main issues include **GEOM**'s threshold sensitivity, and thin or angled walls. Hardware limits (8 GB VRAM, real-time constraints) restricted tests to 128×128 maps and shallow networks; larger models (256–512) were infeasible. The simulated dataset may cause domain shift to real SLAM maps, and fixed epoch budgets, while reproducible, may induce mild under- or over-fitting.

Outlook. (i) *Hybrid designs:* integrate GEOM priors into UNet or RePaint; (ii) *Uncertainty-aware decoding:* reduce over-confident hallucinations; (iii) *Scaling:* extend to 256–512 resolutions for SLAM pipelines; (iv) *Objective–metric alignment:* multi-task losses correlating with IoU/F1/SSIM.

8 Conclusion

We benchmarked six families—**CVAE**, **CGAN**, **RePaint**, **ViT**, **UNet**, and **GEOM**—on 128×128 occupancy grids using a Gazebo/ROS HouseExpo pipeline.

Key findings. UNet achieves the best trade-off (lowest L1/MSE/RMSE; highest IoU/F1/Acc). **RePaint** delivers high perceptual quality and strong large-hole completion but is slower; its SSIM remains below **GEOM**, which is competitive without training and attains the top SSIM. **CGAN** yields sharp boundaries yet overcompletes, while **CVAE/ViT** offer less precise outputs under limited data and sometimes show mild over- or under-completion due to their latent and patch-based structures. Feed-forward models exceed 20 img/s; diffusion is slower; **GEOM** is CPU-efficient.

Scientific insight. Occupancy-map completion is generative yet highly structured. Models with latent representations (CVAE, CGAN), iterative sampling (diffusion), or coarse spatial encoding (ViT) underperform direct prediction (**UNet**) and geometric reasoning (**GEOM**). Latent bottlenecks compress spatial information, adversarial training introduces instability, diffusion incurs high computational cost, and ViT's patching blurs thin walls. Deterministic spatial reasoning proves superior for structured prediction.

Future work. (i) Integrate GEOM constraints into learned architectures; (ii) Develop uncertainty-aware decoding; (iii) Scale to higher resolutions (256–512) with real-world online learning.

Competing Interests

The authors declare no competing interests. All co-authors have seen and agree with the contents of the manuscript, and there is no financial interest to report. We certify that the submission is original work and is not under review at any other venue.

Data Availability

The customized HouseExpo dataset (processed splits and manifests) will be released publicly upon acceptance.

Code Availability

All code, configs, and Gazebo/ROS assets will be released publicly upon acceptance.

References

[1] S. Güzel, S. Yavuz, "Survey on Learning-Based Scene Extrapolation in Robotics", *International Journal of Intelligent Robotics and Applications*, vol. 8, no. 1, pp. 251–268, 2024. Springer. https://doi.org/10.1007/s41315-023-00303-0.

[2] Y. Katsumata, A. Kanechika, A. Taniguchi, L. El Hafi, Y. Hagiwara, T. Taniguchi, "Map Completion from Partial Observation Using the Global Structure of Multiple Environmental Maps", *Advanced Robotics*, vol. 36, no. 5–6, pp. 279–290, 2022. Taylor & Francis. https://doi.org/10.1080/01691864.2022.2029762.

[3] T. Li, D. Ho, C. Li, D. Zhu, C. Wang, M.Q.-H. Meng, "HouseExpo: A Large-Scale 2D Indoor Layout Dataset for Learning-Based Algorithms on Mobile Robots", *Proc. of 2020 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 5839–5846, 2020. https://doi.org/10.1109/IROS45743.2020.9341284.

[4] D. De Martini, A. Paiement, L. Tao, A. Koschorreck, R. Vollgraf, "Occupancy Grid Artefact Removal and Error Correction using GANs", *Proc. of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2301–2307, 2020. https://doi.org/10.1109/IROS45743.2020.9340894.

[5] R. Shrestha, F.-P. Tian, W. Feng, P. Tan, R. Vaughan, "Learned Map Prediction for Enhanced Mobile Robot Exploration", *Proc. of 2019 Intl. Conf. on Robotics and Automation (ICRA)*, pp. 1197–1204, 2019. https://doi.org/10.1109/ICRA.2019.8793769.

[6] A. Aydemir, P. Jensfelt, J. Folkesson, "What Can We Learn from 38,000 Rooms? Reasoning About Unexplored Space in Indoor Environments", *Proc. of 2012 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 4675–4682, 2012. https://doi.org/10.1109/IROS.2012.6386110.

[7] C. Ho, S. Kim, B. Moon, A. Parandekar, N. Harutyunyan, C. Wang, K. Sycara, G. Best, S. Scherer, "MapEx: Indoor Structure Exploration with Probabilistic Information Gain from Global Map Predictions", *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2025. https://doi.org/10.48550/arXiv.2409.15590.

[8] S. Güzel and E. Uslu, "Map Completion Using Generative Models on a Custom Dataset," *Signal, Image and Video Processing*, vol. 19, p. 878, 2025. https://doi.org/10.1007/s11760-025-04431-x.

[9] H. Zhao, J. Song, Y. Yang, C.C. Loy, L. Ma, "High-Fidelity Pluralistic Image Completion with Transformers", *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11169–11178, 2021. https://doi.org/10.1109/CVPR46437.2021.01099.

[10] Z. Zhang, "Conditional Deep Convolutional GANs for Image Completion", *Proc. of the Conference on Artificial Intelligence and Applications*, pp. 11–19, 2019.

[11] W. Chen, M. Zhang, H. Lin, "Large Mask Image Completion with Conditional GAN", *Symmetry*, vol. 14, no. 10, p. 2148, 2022. MDPI. https://doi.org/10.3390/sym14102148.

[12] S. Song, F. Yu, A. Zeng, A.X. Chang, M. Savva, T. Funkhouser, "Semantic Scene Completion from a Single Depth Image", *Proc. of 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 190–198, 2017. https://doi.org/10.1109/CVPR.2017.28.

[13] N. Koenig, A. Howard, "Design and Use Paradigms for Gazebo, an Open-Source Multi-Robot Simulator", *Proc. of 2004 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, vol. 3, pp. 2149–2154, 2004. https://doi.org/10.1109/IROS.2004.1389727.

[14] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A.Y. Ng, et al., "ROS: An Open-Source Robot Operating System", *ICRA Workshop on Open Source Software*, vol. 3, no. 3.2, p. 5, 2009. Kobe, Japan.

[15] G. Bradski, "The OpenCV Library", *Dr. Dobb's Journal of Software Tools*, 2000.

[16] OpenAI, "ChatGPT: A Conversational AI Assistant", *https://openai.com/chatgpt*, Accessed: March 4, 2025.

[17] R. Padilla, S.L. Netto, E.A.B. da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms", *Proc. of 2020 Intl. Conf. on Systems, Signals and Image Processing (IWSSIP)*, pp. 237–242, 2020. https://doi.org/10.1109/IWSSIP48289.2020.9145130.

[18] D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, 5th ed., John Wiley & Sons, Hoboken, NJ, 2010.

[19] T. Chai, R.R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature", *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014. https://doi.org/10.5194/gmd-7-1247-2014.

[20] Z. Wang, A.C. Bovik, "A Universal Image Quality Index", *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002. https://doi.org/10.1109/97.995823.

[21] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. https://doi.org/10.1109/TIP.2003.819861.

[22] L. Wald, T. Ranchin, M. Mangolini, "Quality of high resolution synthesised images: Is there a simple criterion?", *Intl. Archives of Photogrammetry and Remote Sensing*, vol. 33, no. B2, pp. 326–332, 2000.

[23] C. Persello, L. Bruzzone, "Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 225–238, 2013. https://doi.org/10.1109/TGRS.2012.2198245.

[24] S.K. Ramakrishnan, Z. Al-Halah, K. Grauman, "Occupancy Anticipation for Efficient Exploration and Navigation", *Proc. of ECCV 2020*, Part V, pp. 400–418, 2020. Springer. https://doi.org/10.1007/978-3-030-58548-8_24.

[25] K. Katyal, K. Popek, C. Paxton, J. Moore, K. Wolfe, P. Burlina, G.D. Hager, "Occupancy Map Prediction Using Generative and Fully Convolutional Networks for Vehicle Navigation", *arXiv preprint arXiv:1803.02007*, 2018. https://doi.org/10.48550/arXiv.1803.02007.

[26] K. Katyal, K. Popek, C. Paxton, P. Burlina, G.D. Hager, "Uncertainty-Aware Occupancy Map Prediction Using Generative Networks for Robot Navigation", *Proc. of 2019 Intl. Conf. on Robotics and Automation (ICRA)*, pp. 5453–5459, 2019. https://doi.org/10.1109/ICRA.2019.8793500.

[27] S. Güzel, E. Uslu, and S. Yavuz, "Occupancy Grid Map Completion Based on Geometric Features and Door Discovery," in *Proc. 12th Int. Symp. on Digital Forensics and Security (ISDFS)*, 2024, pp. 1–6. https://doi.org/10.1109/ISDFS60797.2024.10527317.

[28] M. Luperto, F. Amadelli, M. Di Berardino, and F. Amigoni, "Mapping Beyond What You Can See: Predicting the Layout of Rooms Behind Closed Doors," *Robotics and Autonomous Systems*, vol. 159, p. 104282, 2023. https://doi.org/10.1016/j.robot.2022.104282.

[29] R. Suvorov *et al.*, "Resolution-Robust Large Mask Inpainting with Fourier Convolutions," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, 2022, pp. 2149–2159. https://doi.org/10.1109/WACV51458.2022.00323.

[30] A. Spinos *et al.*, "Map Prediction and Generative Entropy for Multi-Agent Exploration," *arXiv preprint arXiv:2501.13189*, 2025. https://doi.org/10.48550/arXiv.2501.13189.

[31] A. Lugmayr *et al.*, "RePaint: Inpainting Using Denoising Diffusion Probabilistic Models," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11451–11461. https://doi.org/10.1109/CVPR52688.2022.01117.

[32] R. Rombach *et al.*, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01042.

[33] H. Gao *et al.*, "SenseExpo: Efficient Autonomous Exploration with Prediction Information from Lightweight Neural Networks," *arXiv preprint arXiv:2503.16000*, 2025. https://doi.org/10.48550/arXiv.2503.16000.

[34] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017. https://doi.org/10.48550/arXiv.1706.03762.

[35] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier Convolution," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 4479–4489, 2020. https://doi.org/10.48550/arXiv.2006.01463.

[36] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://doi.org/10.1109/CVPR.2017.632.

[37] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2018. [Online]. Available: `https://openreview.net/forum?id=B1QRgziT-`.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, 2015, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

[39] K. Sohn, H. Lee, and X. Yan, "Learning Structured Output Representation Using Deep Conditional Generative Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. [Online]. Available: `https://proceedings.neurips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html`.

[40] A. Dosovitskiy *et al.*, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021. [Online]. Available: `https://openreview.net/forum?id=YicbFdNTTy`.

[41] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Infor-*

mation Processing Systems (NeurIPS), 2020. https://doi.org/10.48550/arXiv.2006.11239.

[42] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021. https://doi.org/10.48550/arXiv.2010.02502.

[43] A. Kumar, T. Oishi, S. Ono, K. Ikeuchi, "Improving GPS Position Accuracy by Identification of Reflected GPS Signals Using Range Data", *Journal of the Japan Society of Photogrammetry and Remote Sensing (Seisan Kenkyu)*, vol. 66, no. 2, pp. 101–106, 2014. https://doi.org/10.11188/seisankenkyu.66.101.

[44] A. Kumar, T. Oishi, S. Ono, K. Ikeuchi, "Global Coordinate Adjustment of 3D Survey Models under Unstable GPS Condition", *Journal of the Japan Society of Photogrammetry and Remote Sensing (Seisan Kenkyu)*, vol. 65, no. 2, pp. 91–96, 2013. https://doi.org/10.11188/seisankenkyu.65.91.

[45] A.K. Aggarwal, A.P.S. Chauhan, "Robust Feature Extraction from Omnidirectional Outdoor Images for Computer Vision Applications", *International Journal of Intelligent Systems and Applications in Engineering*, vol. 13, no. 1, pp. 13–18, 2025. [Online]. Available: `https://www.iaras.org/iaras/filedownloads/ijim/2025/023-0002(2025).pdf`.

[46] L. Vinckenbosch, T. Deconinck, W. Desmet, "Light Propagation through Biological Tissue: Comparison of Monte Carlo Simulation with Deterministic Models", *Proceedings of the 12th International Conference on Numerical Simulation of Optoelectronic Devices (NUSOD)*, pp. 135–136, 2015. https://doi.org/10.1109/NUSOD.2015.7292909.

[47] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013. https://doi.org/10.48550/arXiv.1312.6114.