FacaDiffy: Inpainting unseen facade parts using diffusion models

Thomas Fröch[*,1], Olaf Wysocki[2], Yan Xia[*,3,5] , Junyu Xie[4], Benedikt Schwab[1], Daniel Cremers[3,5], Thomas H. Kolbe[1]

[1]Chair of Geoinformatics, TUM School of Engineering and Design, Technical University of Munich (TUM), Munich, Germany - (thomas.froech, benedikt.schwab, thomas.kolbe)@tum.de
[2] Photogrammetry and Remote Sensing, TUM School of Engineering and Design, Technical University of Munich (TUM), Munich, Germany - olaf.wysocki@tum.de
[3]Computer Vision Group, TUM School of Computation, Information and Technology, Technical University of Munich (TUM), Munich, Germany - (yan.xia, cremers)@tum.de
[4]Visual Geometry Group, Department of Engineering Science, University of Oxford. Oxford, UK - jyx@robots.ox.ac.uk
[5]Munich Center for Machine Learning, Munich, Germany

Keywords: GSW 2025, 3D-reconstruction, image-inpainting, mobile-laser-scanning, point-clouds, deep-learning, Stable Diffusion, Dreambooth

Abstract

High-detail semantic 3D building models are frequently utilized in robotics, geoinformatics, and computer vision. One key aspect of creating such models is employing 2D conflict maps that detect openings' locations in building facades. Yet, in reality, these maps are often incomplete due to obstacles encountered during laser scanning. To address this challenge, we introduce FacaDiffy, a novel method for inpainting unseen facade parts by completing conflict maps with a personalized Stable Diffusion model. Specifically, we first propose a deterministic ray analysis approach to derive 2D conflict maps from existing 3D building models and corresponding laser scanning point clouds. Furthermore, we facilitate the inpainting of unseen facade objects into these 2D conflict maps by leveraging the potential of personalizing a Stable Diffusion model. To complement the scarcity of real-world training data, we also develop a scalable pipeline to produce synthetic conflict maps using random city model generators and annotated facade images. Extensive experiments demonstrate that FacaDiffy achieves state-of-the-art performance in conflict map completion compared to various inpainting baselines and increases the detection rate by 22% when applying the completed conflict maps for high-definition 3D semantic building reconstruction. The code is be publicly available in the corresponding GitHub repository: https://github.com/ThomasFroech/InpaintingofUnseenFacadeObjects

1. Introduction

Semantic 3D city models hold significant potential to address pressing global issues. Unlike mesh-based models, they are characterized by watertightness and object-oriented modeling, which has proven pivotal in various applications, such as estimating building solar potential and simulating wind flow (Biljecki et al., 2015). Currently, they are ubiquitous, as, for example, approximately 140 million open access building models are available in the United States, Switzerland, and Poland while 55 million are available in Germany (Wysocki et al., 2024). While numerous cities and entire countries provide semantic 3D building models at Level of Detail (LoD)2, characterized by complex roof shapes and planar facades, LoD3 datasets with semantically detailed facades remain scarce.

However, such highly-detailed LoD3 datasets are required for numerous application areas, such as assessing flood risk (Amirebrahimi et al., 2016), analyzing building potential for vertical farming (Palliwal et al., 2021), and estimating energy demand (Nouvel et al., 2013).

Although a great deal of research has been devoted to the automatic LoD3 reconstruction (Szeliski, 2010), the current practice indicates that tedious, manual LoD3 modeling prevails (Chaidas et al., 2021). Yet, recent developments have shown that the so-called *conflict maps* prove to be valuable for the automatic LoD3 reconstruction (Wysocki et al., 2023c). As we illustrate in Figure 1, such conflict maps are generated based on the analysis of the 3D building model and sensor rays: The surface is deemed confirmed (green) when the ray point hits the surface, conflicted (red) when the ray traverses the surface, and unknown (black) when the surface is unmeasured. These maps are considered a core intermediate geometric reconstruction cue, as they are frequently coupled with semantic-rich images or point clouds (Wysocki et al., 2023c); and the principle can also be directly used, for example, for building change detection (Tuttas et al., 2015). However, as we see in Figure 1, the conflict maps are prone to occlusions (black), which render them incomplete and impact their at-scale applicability.

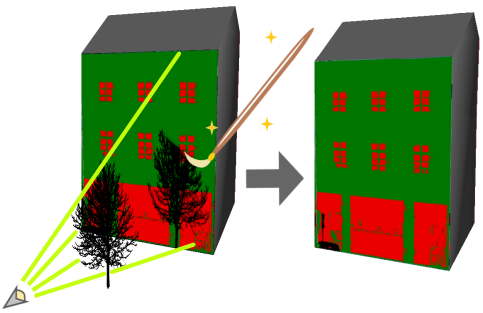

Figure 1. The conflict maps are prone to occlusions (black) owing to the reliance on the ray-to-model (yellow) analysis. We complement conflicts (red) and confirmations (green) by employing the proposed FacaDiffy.

* Corresponding Author

Recent machine learning advancements have provided potent models for image inpainting, such as Stable Diffusion (Rombach et al., 2022) or the LaMa GAN (Suvorov et al., 2022), presenting an opportunity to address the challenge of incomplete conflict maps. We introduce the application of deep- learning methods to inpaint previously unseen facade objects into 2D conflict maps computed from existing LoD2 models and corresponding point clouds. Completed conflict maps can, for example, be incorporated into existing pipelines for LoD3 reconstruction and contribute to increasing their reconstruction accuracy or be utilized for change detection (Tuttas et al., 2015). We facilitate the deployment of a Stable Diffusion inpainting model (Rombach et al., 2022) by personalizing with Dreambooth (Ruiz et al., 2023), utilizing synthetic conflict maps derived from randomly generated semantic city models and those obtained from annotated images of the CMP Facade Database (Radim and Radim, 2013) as training data.

To summarize, our contributions are as follows:

- A deterministic method to generate 2D conflict maps from existing LoD2 building models and corresponding laser scanning point clouds.
- Conflict-oriented, personalized stable diffusion inpainting improving the conflict map completeness.
- An approach to generate synthetic conflict maps using stochastic city model generators.

2. Related Work

Semantic 3D city models. In addition to offering geometric and visual insights into topographic features, semantic 3D city models provide comprehensive information about structures, taxonomies, and aggregations at the scale of cities, regions, and even complete countries. For the representation and management of city models, the standard CityGML is used internationally, which has been issued by the Open Geospatial Consortium (OGC) (Kolbe, 2009, Gröger and Plümer, 2012, Kolbe et al., 2021). CityGML enables the modeling of urban objects with their 3D geometry, appearance, topology, and semantics at four different LoDs. The data model of CityGML 3.0 is based on the ISO 191xx series of geographic information standards, and CityGML datasets can be encoded using the Geography Markup Language (GML) (Kutzner et al., 2020).

Synthetic generation of semantic city models. With Random3Dcity, Biljecki *et al.* (Biljecki et al., 2016) introduce a method for the procedural generation of randomized semantic city models at various LoD levels. Their approach utilizes a set of pre-defined architectural modeling rules guiding its stochastic nature. Such rules govern aspects like the permissible positioning of facade elements such as doors or windows.

Reconstruction of semantic 3D building models. The considerable potential for applying detailed LoD3 models across diverse domains, coupled with their scarcity, has motivated a significant number of studies to explore the reconstruction of such models. Investigations involve leveraging various data sources, such as optical images, oblique Airborne Laser Scanning (ALS) point clouds, and Mobile Laser Scanning (MLS) measurements, as well as employing diverse approaches, including formal grammar approaches and Bayesian networks (Ripperda, 2010, Huang et al., 2020, Wysocki et al., 2023c). Within their pipeline for reconstructing underpasses in semantic LoD2 city models from co-registered MLS point clouds, Wysocki *et al.* introduce the concept of 2D conflict maps. Their probabilistic approach relies on an occupancy grid implemented as an octree structure, with voxel sizes reflecting the combined uncertainty of the MLS measurements and the semantic city model. This concept has been further developed by (Wysocki et al., 2023a, Wysocki et al., 2023c, Hoegner and Gleixner, 2022), thereby substantiating its effectiveness. The pivotal advantage of the previously mentioned methods over mesh-based approaches is the usage of 3D semantic building models as priors, which has proven to maintain the model watertightness as well as higher 3D reconstruction accuracy, reaching up to around 50% when compared to the mesh-based Poisson reconstruction (Wysocki et al., 2023c). Nevertheless, despite generally yielding commendable results, the challenge of incompleteness remains unsolved.

Deep-learning-based image inpainting. Besides traditional image inpainting methods, which typically rely on solving Partial Differential Equations (PDE)s (Telea, 2004, Bertalmio et al., 2001), and yield unsatisfactory results when applied to facade images (Fritzsche et al., 2022). Recent advancements in deep-learning-based image inpainting suggest potential methodologies for addressing the challenge of incomplete 2D conflict maps. Large mask inpainting (LaMa) (Suvorov et al., 2022), configured as a Generative Adversarial Network (GAN), employs Fast Fourier Convolution operators (Chi et al., 2020) to overcome the limitation of restricted receptive fields, thereby enabling expansive coverage across the entire image.

Diffusion Probabilistic Model (DM)s have emerged as powerful tools for various generative applications in the last years. Notably, Stable Diffusion (SD) (Rombach et al., 2022) demonstrates remarkable flexibility with its capability to handle open-ended text conditioning. When supplemented with additional image and mask inputs, the SD framework can be further extended to solve image inpainting tasks, guided by relevant text prompts. To our knowledge, no method deploying diffusion models for inpainting facade conflict maps has been published yet.

To tailor pre-trained SD models for domain-specific applications, a variety of personalization (*i.e.,* customization) techniques (Gal et al., 2023, Kumari et al., 2023, Wei et al., 2023) are developed. These methods generally utilize small-scale datasets to fine-tune the pre-trained SD pipeline for specialized domains. In particular, Dreambooth (Ruiz et al., 2023) stands out as a robust personalization approach, incorporating a prior-preservation loss and LoRA (Hu et al., 2022)-based fine-tuning.

3. Method

As illustrated in Figure 2, given the LoD2 model and point cloud data, FacaDiffy aims at completing facade conflict maps by leveraging deep-learning-based image inpainting techniques. Specifically, Sect. 3.1 introduces a ray-casting approach to derive 2D real-world conflict maps and corresponding binary masks indicating occluded regions. These estimated occluded areas are then adopted as inpainting masks. To complement insufficient real-world conflict map data for training, Sect. 3.2 details a scalable pipeline that generates synthetic conflict maps as additional training data for the inpainting model. Sect. 3.3 elaborates on a personalized SD-based inpainting method designed to recover complete conflict maps.

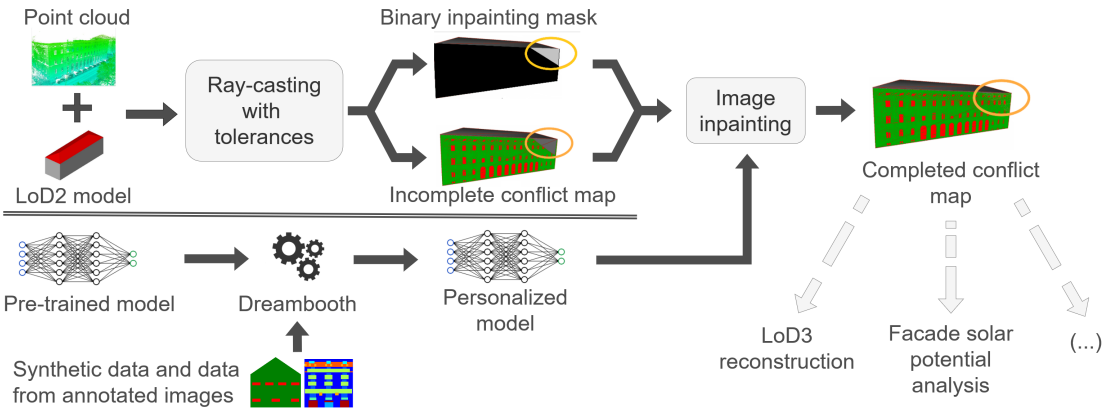

Figure 2. Schematic overview of FacaDiffy. By combining an existing LoD2 building model and corresponding laser scanning point clouds, we formulate a deterministic method based on ray-casting analysis to obtain incomplete conflict maps and corresponding binary inpainting masks (top branch). We generate synthetic conflict maps to personalize the Stable Diffusion model (bottom branch), which is employed for inpainting given the partial evidence of deterministic conflict maps. These can be utilized for various downstream applications such as accurate LoD3 reconstructions, facade solar potential analysis, etc

3.1 Conflict Map Computation

As Figure 2 illustrates, we obtain incomplete conflict maps and corresponding binary masks, highlighting missing (*i.e.,* occluded) areas through a deterministic ray-casting approach with tolerances that combine semantic LoD2 building models and corresponding laser scanning point cloud data.

We first define the unit rays, denoted as $\mathbf{r_p}$, as originating from a viewpoint $\mathbf{v}$ and oriented towards a corresponding point $\mathbf{p}$.

$$\mathbf{r_p} = \mathbf{v} + \frac{\mathbf{p} - \mathbf{v}}{|\mathbf{p} - \mathbf{v}|} \tag{1}$$

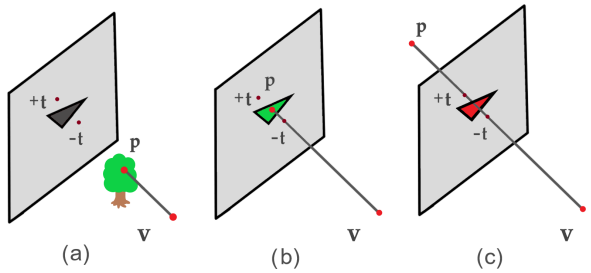

Figure 3. Schematic overview of the conflict determination in the ray casting approach with the viewpoint $\mathbf{v}$, the point $\mathbf{p}$, and the tolerance $\pm t$. Three distinct scenarios are illustrated: (a) unknown; (b) confirming; (c) conflicting.

Then, we identify conflicts between a LoD2 model and a corresponding point cloud by evaluating the distance to the intersection of the LoD2 surface and the rays. Three mutually exclusive cases have to be distinguished:

- Unknown: As illustrated in Figure 3 (a), occurrences of occlusions caused by objects, such as vegetation, which are unrelated to a facade, are identifiable by an intersection distance shorter than the anticipated value. This information is utilized to derive the binary masks indicating the occluded areas in the conflict maps.
- Confirming: The majority of instances corresponds to Figure 3 (b), where the intersection distance remains within an expected tolerance range $[-t, t]$, leading to incomplete conflict map predictions.
- Conflicting: As depicted in Figure 3 (c), the assessed intersection distance surpasses the expected value for openings in the facade, such as windows, due to the voyeur effect (Tuttas and Stilla, 2013), or underpasses and arches that are not considered in LoD2 models (Kutzner et al., 2020).

To achieve a high geometric resolution, we employ a multi-iteration midpoint triangle subdivision approach (Chen and Prautzsch, 2014) on the triangles comprising the wall surface. As the number of triangle subdivision iterations, denoted as n_{div}, increases, striking a balance between geometric resolution and computational efficiency becomes essential.

3.2 Synthetic Conflict Map Generation

We aim to utilize realistic synthetic conflict maps to complement insufficient real-world training data. These maps are derived from randomly generated semantic building models and classified facade image benchmarks, by leveraging their structured knowledge along with insights from previous works.

The extensive semantic information contained in the randomly generated CityGML LoD3 models makes it possible to classify structures as conflicting (red) or confirming (green), according to prior knowledge about the behavior of certain types of building parts. Windows or doors are considered to be conflicting due to the voyeur effect, while underpasses and extruded facade objects such as balconies or decorative molding deviate from the wall surface geometrically (Tuttas and Stilla, 2013, Tuttas et al., 2015, Hoegner and Gleixner, 2022, Wysocki et al., 2023a). The same objects, identifiable by their annotation, are considered conflicting in the annotated facade images.

We subsequently project the facades to 2D and plot the projected triangles they comprise in the corresponding color to obtain synthetic 2D conflict maps. We apply these for personalizing with Dreambooth and to evaluate the inpainting performance in scenarios that closely match real-world applications.

3.3 Deep-Learning-Based Inpainting

Given the incomplete conflict maps, we leverage the SD-inpainting method (Rombach et al., 2022) to recover the missing areas. Specifically, we treat the estimated occluded regions

as inpainting masks (*i.e.,* areas where the inpainting is conducted). We convert color-coded conflict maps into binary images, indicating the presence or absence of conflicts, before initiating the inpainting process to avoid undesirable structures related to color properties during inpainting. However, directly applying the pre-trained SD-inpainting may result in undesired artifacts, largely due to the domain discrepancy between the binary conflict maps and the real-world image priors embedded in the pre-trained SD model. To mitigate such disparity, we adopt a personalization approach and carefully design the input text prompts.

Personalization of the inpainting model. To adapt the pre-trained SD-inpainting model for conflict map inpainting, we utilize the Dreambooth (Ruiz et al., 2023) technique. During the personalization fine-tuning, we adopt the synthetically generated conflict maps (detailed in Sect. 3.2) as training images, and apply random inpainting masks generated by the method proposed in (von Platen et al., 2022).

The choice of text prompts. During the inpainting process, the choice of text prompts exhibits a great influence on resultant quality. To heuristically identify a suitable text-prompt that is consistently applied for conflict map inpainting, we investigate a variety of text-prompts, involving high-level (*e.g., "Window"*) and low-level (*e.g., "Rectangle"*) descriptions and assessing their corresponding effects on the inpainting outcomes.

4. Experiments

We provide a detailed discussion with additional examples, implementation details, and systematic tests in the supplementary material in the corresponding GitHub Repository[1].

4.1 Datasets

Real-world data for conflict maps. As sources for computing real-world conflict maps, we leveraged two primary datasets: (i) a proprietary MLS point cloud (Wysocki et al., 2023b), acquired by the company 3D Mapping Solutions (Haigermoser et al., 2015, 3D Mapping Solutions, 2023) with its geo-referencing supported by the German SAPOS RTK system; (ii) the official LoD2 building models supplied by the Bavarian State Office for Digitizing, Broadband and Survey(Bayerische Vermessungsverwaltung, 2023). The datasets we employed encompass sections of the TUM city campus and of the Pfisterstraße in Munich, Germany.

Annotated facade images. To derive conflict maps from annotated images, we utilized the CMP database of annotated images provided by the Center for Machine Perception in Prague (Radim and Radim, 2013). We considered windows, doors, cornices, sills, balconies, blinds, decorations, molding, pillars, and background as conflicting.

Ground-truth information from LoD3 models. We obtained ground-truth conflict maps from existing LoD3 building models maintained in the Tum2Twin GitHub repository (Schwab et al., 2023) [2]. Note that, the limited availability of LoD3 building models with corresponding point clouds only supports a small-scale evaluation.

[1] https://github.com/ThomasFroech/InpaintingofUnseenFacadeObjects
[2] https://tum2t.win/

Inpainting masks. We consider two types of inpainting masks: the vegetation (*i.e.,* tree-shaped) mask and randomly generated masks. To produce realistic, tree-shaped binary masks, we utilized a point cloud stemming the TreeML-Data collection (Yazdi et al., 2024), projecting it to a 2D raster orthogonally. To produce medium-sized, randomly generated masks, we applied the method introduced by (Suvorov et al., 2022).

4.2 Evaluation Metrics

We assessed the similarity between the inpainting results and ground-truth information from two major perspectives: (i) The structural and shape similarity is measured by the Structural Similarity Index (SSIM) (sliding window size of 71 pixels) (Wang et al., 2004) and the Intersection over Union (IoU) (Rezatofighi et al., 2019); (ii) Given the limitations of IoU in accurately assessing semantic similarity (Table 1, Table 2), we also employ the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) for perceptual assessment.

4.3 Implementation Parameters

Personalization with DreamBooth. In our implementation of the deterministic ray analysis approach to generate conflict maps we set $t = 0.7\mathrm{m}$ and $n_{\mathrm{div}} = 8$. We personalized the SD-inpainting model (sd-v1.5-inpaint) with DreamBooth leveraging 192 synthetically generated conflict maps that are derived from randomly generated semantic building models obtained from the Random3Dcity application (Biljecki et al., 2016). Following the implementation of (Ferreira Barbosa Junior and Patil, 2024), inpainting masks were randomly generated following (von Platen et al., 2022). We specified the text prompt as *"Black background with white patches that are consistent and symmetric to the rest of the image"*. In choosing hyperparameters, we followed the settings specified in (Patil et al., 2022) and consistently deployed them throughout all experiments.

LaMa GAN. As an alternative deep-learning-based inpainting strategy, we trained the LaMa GAN (Suvorov et al., 2022) for 25 epochs. We leveraged the CMP-base dataset (Radim and Radim, 2013) and the Random3Dcity application (Biljecki et al., 2016) to obtain a training dataset consisting of approx. 20.000 conflict maps.

Traditional inpainting strategies. For comparison with traditional inpainting strategies based on solving Partial Differential Equations (PDE)s, we applied the inpainting method by Telea (Telea, 2004) and Navier-Stokes-based inpainting (Bertalmio et al., 2001). We utilized the implementation available in (Bradski, 2000). Evaluation results of both traditional methods are summarized in Table 1.

4.4 Comparisons with State-of-the-art

We evaluated our inpainting model performance on two major datasets with (i) 228 conflict maps derived from the annotated facade images in the CMP-extended database (Radim and Radim, 2013) and (ii) a small-scale dataset with real conflict maps concerning ground-truth data from corresponding LoD3 building models, as a demonstration of our real-world applicability. We treated the unmasked conflict maps as the ground truth and measured their similarity with the completed conflict maps as an indication of the inpainting quality.

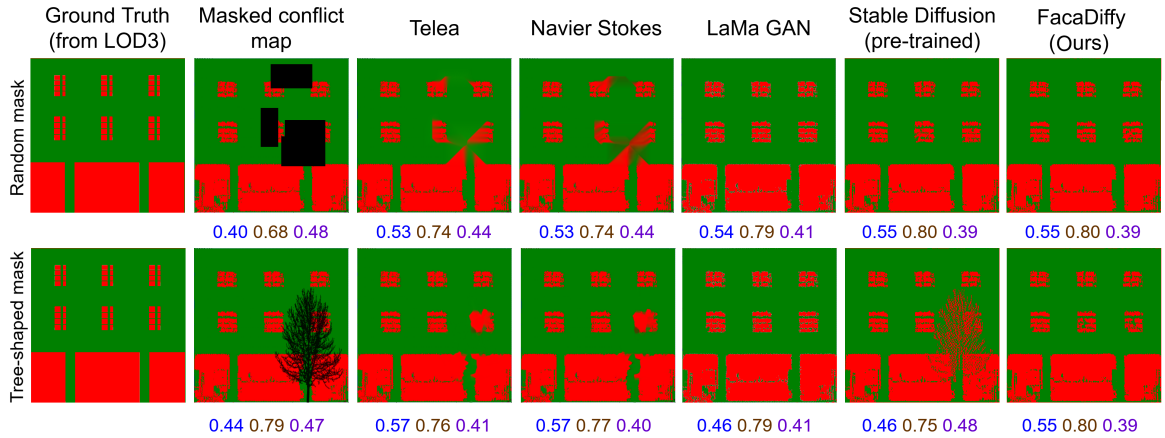

Figure 4. Exemplary inpainting results on a real conflict map. The similarity between ground-truth (LoD3) and inpainted results is measured in terms of SSIM (blue), IoU (brown), and LPIPS (purple). The conflict maps are color-coded with the conflicting (red), confirming (green), and unknown/masked (black) areas.

Methods	Randomly generated masks			Tree-shaped masks		
	SSIM ↑	IoU ↑	LPIPS ↓	SSIM ↑	IoU ↑	LPIPS ↓
Masked conflict map	0.72	0.85	0.24	0.83	0.92	0.21
Telea (Telea, 2004)	0.82	0.89	0.18	0.94	0.96	0.10
Navier-Stokes (Bertalmio et al., 2001)	0.85	0.92	0.14	0.94	0.97	0.09
LaMa GAN (Suvorov et al., 2022)	0.85	0.92	0.14	0.95	0.94	0.06
SD-inpainting (pre-trained) (Rombach et al., 2022)	0.89	0.72	0.09	0.85	0.70	0.21
FacaDiffy (Ours)	0.91	0.72	0.08	0.90	0.66	0.11

Table 1. Quantitative comparison with baseline methods. The evaluation is conducted on the CMP-extended image database (comprising 228 annotated images). The similarity is measured between the inpainting results and unmasked ground-truth conflict maps. The IoU metric yields counterintuitive results when applied to traditional inpainting strategies, despite evident semantic inconsistencies, as exemplified in Figure 4

Comparison with LaMa GAN. Results shown in Table 1 and Figure 4 suggest that LaMa GAN excels at recovering fine structures such as tree-shaped masks, while our method also yields competitive results. In terms of randomly generated masks, FacaDiffy exhibits superior performance compared to LaMa GAN. These findings suggest the better applicability of FacaDiffy considering the scenario with large occlusions.

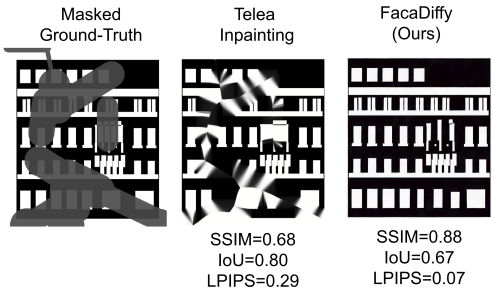

Figure 5. Counterintuitive IoU evaluation result for a randomly masked conflict map derived from the CMP-database of annotated images. While our method performs better qualitatively, the IoU yields counterintuitive results.

Comparison with traditional inpainting strategies. Concerning randomly generated masks, the diminished performance of the traditional methods compared to FacaDiffy becomes evident in the LPIPS and SSIM measurements summarized in Table 1. According to Table 1, the traditional approaches perform better in the case of the tree-shaped mask. However, Figure 4 illustrates notable semantic inconsistencies in the inpainting outcomes achieved through these methods, while FacaDiffy demonstrates qualitatively superior semantic consistency.

Counterintuitive IoU evaluation results. When deploying the baseline methods, greater overlap due to larger continuous areas being inpainted into the images may positively impact the IoU, even though the underlying true semantic similarity would not be positively affected. Such counterintuitive results, evident in Table 1 and Table 2, are exemplarily illustrated in Figure 5. The IoU contradicts the qualitative similarity assessment of the inpainting results to the ground truth, the SSIM, and the LPIPS measurements. While these metrics indicate greater similarity, the IoU suggests the opposite. In contrast to evaluating the IoU instance-wise with respect to the individual facade features, we evaluate the entire conflict maps, which might also affect the score. Additionally, as (Zhang et al., 2018) mention, no universal metric for quantifying image completion is available, which motivates us to consider multiple metrics for an unbiased comparison.

4.5 Ablation Studies

Additionally, we conducted personalizations involving 5 synthetically generated conflict maps, and 228 conflict maps derived from annotated facade images. We ensured comparability by applying the same text prompt consistently. An improvement of -0.07 in the LPIPS measurements concerning tree-shaped masks in Table 2 demonstrates that increasing the number of synthetic training samples from 5 (Exp. A) to 192 (Exp. C) generally leads to an enhanced inpainting performance. The personalized model, fine-tuned on synthetic data (Exp.,C), competes well with its counterpart trained on real-

Exp.	No. of conflict maps	Type of conflict maps	Randomly generated masks			Tree-shaped masks		
			SSIM ↑	IoU ↑	LPIPS ↓	SSIM ↑	IoU ↑	LPIPS ↓
A	5	Syn.	0.90	0.73	0.08	0.86	0.66	0.18
B	228	Real	0.91	0.70	0.08	0.84	0.65	0.22
C (default)	192	Syn.	0.91	0.72	0.08	0.90	0.66	0.11

Table 2. Quantitative comparison with baseline methods. The evaluation is conducted on the CMP-extended image database (comprising 228 annotated images). The similarity is measured between the inpainting results and unmasked ground-truth conflict maps. The IoU metric yields counterintuitive results when applied to traditional inpainting strategies, despite evident semantic inconsistencies, as exemplified in Figure 4

world conflict maps (Exp.,B), validating our synthetic map generation pipeline (Table 2).

5. Discussion

FacaDiffy demonstrates enhancements of $+0.19$ in the SSIM and of -0.16 in the LPIPS compared to the masked conflict map in Table 1. As illustrated in Figure 4, the pre-trained SD-inpainting model struggles to accurately reconstruct the facade structures from the tree-shaped inpainting mask. This poses a challenge for deploying it in realistic scenarios, as trees represent a common type of occlusion. Conversely, our personalized model, FacaDiffy, effectively reduces the impact of tree-shaped masks and successfully completes the masked regions with semantically meaningful content. This improvement is further substantiated in Table 1 and Table 2, where FacaDiffy contributes to up to 10% boosts in the LPIPS scores compared to the pre-trained model.

5.1 Impact on the LoD3 Semantic Reconstruction

To evaluate FacaDiffy's impact on the potential downstream tasks, we selected one of the most prominent: LoD3 reconstruction. We selected building facades belonging to the so-called Building 23 of the TUM City Campus, characterized by three inpainting scenarios: A, very good facade visibility (visibility $>$90%) with our randomly generated mask, B partial visibility due to scaffolding obstructing the field-of-view ($\sim$ 50%), C low visibility due to trees and city furniture blocking the field-of-view (70%). We used the provided implementation for the part of the 3D reconstruction of (Wysocki et al., 2023c). Yet, unlike the Scan2LoD3 probabilistic maps (Wysocki et al., 2023c), in the presented method we design our maps as hard evidence of conflict, confirmation, and unknown; additionally, we assumed that each conflict represented a window opening for homogeneous comparison. We evaluated our method's impact for the typical three pillars of 3D semantic reconstruction in the following sub-sections.

Detection rate. Our experiments corroborate that FacaDiffy can achieve a higher detection rate by 22% compared to the plain, deterministic approach (Table 3). The prominent example of the method's impact is illustrated in Figure 6, where the 3D reconstruction of the randomly-masked facade A with (Figure 6c) and without inpainting (Figure 6d) is shown, indicating the 22% higher detection rate. Yet, the facade completeness remains challenging for small-size windows, as indicated by missing top-row windows in Figure 6. It is worth noting that FacaDiffy does not violate the main trait of the conflict maps: low false alarm rate. It remained negligible and similar throughout the experiments ($\sim$ 1%).

	Before inpainting				After inpainting				Gain ↑	
	A	B	C	Tot	A	B	C	Tot		
AO	66	17	20	**103**	66	17	20	**103**		
D	33	6	10	**49**	52	13	12	**77**	↑	28
TP	32	6	10	**48**	46	13	12	**71**	↑	23
FP	1	0	0	**1**	1	0	0	**1**	∼	0
FN	35	11	10	**56**	20	4	8	**32**	↑	24
DA	48%	35%	50%	**47%**	70%	76%	60%	**69%**	↑	22%
FA	3%	0%	0%	**2%**	2%	0%	0%	**1%**	↑	1%

Table 3. Detection rate for all openings (DA) and the respective false alarm rate (FA) for façades A, B, and C (AO = all openings, D = detections, TP = true positives, FP = false positives, FN = false negatives, ↑ indicates positive change).

Segmentation. As we show in Table 4, the impact on the segmentation per-instance area was even across the testing sample. We observe a decreasing performance rate correlated with the number of the initially detected windows (Table 3). For example, the frontal inpainted facade A scored an improved Intersection over Union (IoU) over the non-inpainted one. Whereas a decrease was noticed for facade B owing to the small initially detected evidence sample (heavily obstructed by scaffolding); and a smaller decrease for facade C with a smaller obstruction rate. Note, that here we did not include the samples with misses (FN).

	mean IoU ↑			
Façade	A	B	C	Total
Openings	66	17	20	103
Before inpainting	0.46	0.41	0.25	0.41
After inpainting	0.48	0.28	0.22	0.41

Table 4. Comparison of the median intersection over union (IoU) scores for the opening segmentation (openings where $IoU > 0\%$)

3D reconstruction. Subsequently, we tested the impact on the final LoD3 reconstruction for the complete building 23. The Hausdorff distance was employed to calculate the difference between the ground truth LoD3 model before and after the inpainting, which essentially evaluated the geometrical gain of the proposed method (Cignoni et al., 1998). Here, the improvement oscillated around $0.03\mathrm{m}$ in terms of the RMSE score and $0.01\mathrm{m}$ for the mean (μ), as shown in Table 5. As expected, the geometrical gain was relatively small, as the tested opening elements were not facade-protruded.

5.2 Limitations

Succeeding in effectively obtaining conflict maps, our deterministic ray analysis approach exhibits limitations in handling

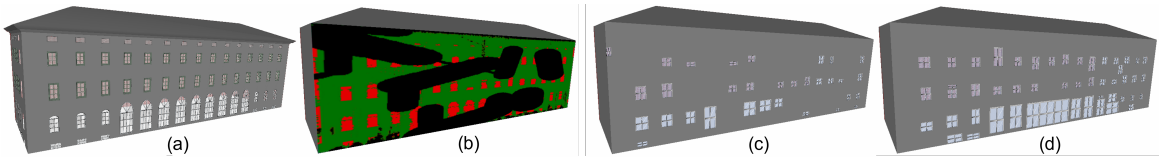

Figure 6. Impact of our method on the 3D reconstruction: (a) ground-truth LoD3 model; (b) incomplete conflict map; (c) 3D reconstruction based on the incomplete map; (d) 3D reconstruction based on the inpainted conflict map.

Method	vs. GT LoD3 ↓	
	μ	RMS
Before inpainting	0.27	0.30
After inpainting	0.26	0.27

Table 5. The impact of our method on the final LoD3 reconstruction using the ground-truth LoD3 model and the Hausdorff distance.

scenarios involving intricate transparent facade structures and buildings with extruded facade elements that encase corners. The lack of probabilistic information potentially limits its effectiveness in scenarios encompassing additional confidence information. The heuristically designed text prompt proves effective for the highly-challenging scenarios of random and tree occlusions. Yet, our method can be sensitive to the chosen text prompt. The degree to which the synthetically generated conflict maps resemble real conflict maps requires further investigation and is planned as future research.

6. Conclusion

We propose FacaDiffy, a method for completing conflict maps derived from semantic LoD2 building models and corresponding laser scanning point clouds. FacaDiffy incorporates DreamBooth-based personalization to address the limitations of the pre-trained Stable Diffusion model in conflict map inpainting. This results in an approximate 10% improvement in the Learned Perceptual Image Patch Similarity (LPIPS) score when dealing with tree-shaped inpainting masks. Interestingly, as illustrated in Figure 5, the IoU exhibits counterintuitive evaluation results. Extensive experiments further demonstrate the superior performance of our method over other image inpainting baselines. During the LoD3 semantic reconstruction, the introduction of FacaDiffy contributes to a 22% increase in the detection rate, proving its efficacy in enhancing the 3D reconstruction accuracy of existing pipelines. Future work will explore the completed conflict maps in the various application pipelines, such as solar potential estimation of facades and wind flow simulations.

Acknowledgements

This work was conducted within the framework of the Leonhard Obermeyer Center at the Technical University of Munich (TUM). We thank the City of Munich for the cooperation in the Connected Urban Twins (CUT) project funded by the Federal Ministry for Housing, Urban Development and Building of Germany (BMWSB). We gratefully acknowledge the staff members of the TUM Professorship of Photogrammetry and Remote Sensing for their valuable insights and support.

References

3D Mapping Solutions, 2023. MoSES mobile mapping platform - technical details. https://www.3d-mapping.de/ueber-uns/unternehmensbereiche/data-acquisition/unser-vermessungssystem/. Accessed: 2023-01-30.

Amirebrahimi, S., Rajabifard, A., Mendis, P., Ngo, T., 2016. A BIM-GIS integration method in support of the assessment and 3D visualisation of flood damage to a building. *Journal of spatial science*, 61(2), 317–350.

Bayerische Vermessungsverwaltung, 2023. 3D-Gebäudemodelle (LoD2). Online. https://geodaten.bayern.de/opengeodata/ Accessed: 2023-01-09.

Bertalmio, M., Bertozzi, A. L., Sapiro, G., 2001. Navier-stokes, fluid dynamics, and image and video inpainting. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, IEEE Comput. Soc.

Biljecki, F., Ledoux, H., Stoter, J., 2016. Generation of multi-LOD 3D city models in CityGML with the procedual modelling engine Random3Dcity. *ISPRS - Int. Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, IV-4/W1, 51–59.

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., Çöltekin, A., 2015. Applications of 3D city models: State of the art review. *ISPRS ISPRS Int. J. Geo-Inf.*, 4(4), 2842–2889.

Bradski, G., 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*. https://opencv.org/. Accessed: 2025-10-02.

Chaidas, K., Tataris, G., Soulakellis, N., 2021. Seismic Damage Semantics on Post-Earthquake LOD3 Building Models Generated by UAS. *ISPRS Int. J. Geo-Inf.*, 10(5).

Chen, Q., Prautzsch, H., 2014. General triangular midpoint subdivision. *Computer Aided Geometric Design*, 31(7-8).

Chi, L., Jiang, B., Mu, Y., 2020. Fast fourier convolution. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (eds), *Advances in Neural Information Processing Systems*, 33, Curran Associates, Inc., 4479–4488.

Cignoni, P., Rocchini, C., Scopigno, R., 1998. Metro: measuring error on simplified surfaces. *Computer Graphics Forum*, 17(2), Blackwell Publishers, 167–174.

Ferreira Barbosa Junior, A., Patil, S., 2024. Dreambooth for the inpainting model. https://github.com/huggingface/diffusers/tree/main/examples/research_projects/dreambooth_inpaint. Acessed: 23.10.2024.

Fritzsche, W., Goebbels, S., Hensel, S., Russinski, M., Schuch, N., 2022. Inpainting applied to facade images: A comparison of algorithms. *International Conference on Pattern Recognition and Artificial Intelligence*.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., Cohen-Or, D., 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. *International Conference on Learning Representations (ICLR)*.

Gröger, G., Plümer, L., 2012. CityGML–interoperable semantic 3D city models. *ISPRS J. of Photogramm. and Remote Sens.*, 71.

Haigermoser, A., Luber, B., Rauh, J., Gräfe, G., 2015. Road and track irregularities: measurement, assessment and simulation. *Vehicle System Dynamics*, 53(7).

Hoegner, L., Gleixner, G., 2022. Automatic extraction of facades and windows from MLS point clouds using voxelspace and visibility analysis. *ISPRS - Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*, XLIII-B2-2022, 387–394.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*.

Huang, H., Michelini, M., Schmitz, M., Roth, L., Mayer, H., 2020. LoD3 building reconstruction from multi-source images. *ISPRS - Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*, XLIII-B2-2020, 427–434.

Kolbe, T. H., 2009. *Representing and exchanging 3D city models with CityGML*. Springer, Berlin, Heidelberg, chapter 3D Geo-Information Sciences. Lecture Notes in Geoinformation and Cartography.

Kolbe, T. H., Kutzner, T., Smyth, C. S., Nagel, C., Roensdorf, C., Heazel, C., 2021. Open Geospatial Consortium. OGC City Geography Markup Language (CityGML) Part 1: Conceptual Model Standard v3.0. OGC Document Number: 20-010.

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.-Y., 2023. Multi-concept customization of text-to-image diffusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kutzner, T., Chaturvedi, K., Kolbe, T. H., 2020. CityGML 3.0: New Functions Open Up New Applications. *PFG – J. of Photogramm., Remote Sensing and Geoinf. Sci.*, 88(1), 43–61.

Nouvel, R., Schulte, C., Eicker, U., Pietruschka, D., Coors, V., 2013. CityGML-based 3D city model for energy diagnostics and urban energy policy support. *IBPSA World*, 2013, 1–7.

Palliwal, A., Song, S., Tan, H. T. W., Biljecki, F., 2021. 3D city models for urban farming site identification in buildings. *Computers, Environment and Urban Systems*, 86, 101584.

Patil, S., Cuenca, P., Kozin, V., 2022. Training stable diffusion with Dreambooth using diffusers. https://huggingface.co/blog/dreambooth. Accessed: 2024-01-11.

Radim, T., Radim, Š., 2013. Spatial pattern templates for recognition of objects with regular structure. J. Weickert, M. Hein, B. Schiele (eds), *Pattern Recognition*, 8142, Springer, Berlin, Heidelberg, Berlin, Heidelberg, 364–374.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 658–666.

Ripperda, N., 2010. Rekonstruktion von Fassadenstrukturen mittels formaler Grammatiken und Reversible Jump Markov Chain Monte Carlo Sampling. PhD thesis, Leibnitz Universität Hannover.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Ruiz, N., Li, Y., Jampan, V., Pritch, Y., Rubinstein, M., Aberman, K., 2023. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. L. O'Conner (ed.), *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22500–22510.

Schwab, B., Wysocki, O., Biswanath, M., Barbosa, J., 2023. tum2twin. https://tum2t.win/. Accessed: 2024-01-11.

Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V., 2022. Resolution-robust large mask inpainting with fourier convolutions. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, arXiv, 2149–2159.

Szeliski, R., 2010. *Computer vision: algorithms and applications*. Springer Science & Business Media.

Telea, A., 2004. An image inpainting technique based on the fast marching method. *J. of Geogr. Tools*, 6(1).

Tuttas, S., Stilla, U., 2013. Reconstruction of façades in point clouds from multi aspect oblique ALS. *ISPRS - Int. Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, II-3/W3, 91–96.

Tuttas, S., Stilla, U., Braun, A., Borrmann, A., 2015. Validation of BIM components by photogrammetric point clouds for construction site monitoring. *ISPRS - Int. Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, II-3/W4, 231–237.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T., 2022. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`. Accessed: 2024-01-11.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612.

Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W., 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wysocki, O., Grilli, E., Hoegner, L., Stilla, U., 2023a. Combining visibility analysis and deep learning for refinement of semantic 3D building models by conflict classification. *ISPRS - Int. Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, X-4/W2-2022, 289–296.

Wysocki, O., Hoegner, L., Stilla, U., 2023b. MLS2LoD3: refining low LoDs building models with MLS point clouds to reconstruct semantic LoD3 building models. T. H. Kolbe, A. Donaubauer, C. Beil (eds), *Recent Advances in 3D Geoinformation Science: Proceedings of the 18th 3D GeoInfo Conference*.

Wysocki, O., Schwab, B., Beil, C., Holst, C., Kolbe, T. H., 2024. Reviewing Open Data Semantic 3D City Models to Develop Novel 3D Reconstruction Methods. *ISPRS - Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*, XLVIII-4-2024, 493–500.

Wysocki, O., Xia, Y., Wysocki, M., Grilli, E., Hoegner, L., Cremers, D., Stilla, U., 2023c. Scan2LoD3: Reconstructing semantic 3D building models at LoD3 using ray casting and Bayesian networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 6547–6557.

Yazdi, H., Shu, Q., Rötzer, T., Petzold, F., Ludwig, F., 2024. A multilayered urban tree dataset of point clouds, quantitative structure and graph models. *Scientific Data*, 11(1).

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.