Procedural Modelling and Generative AI for Augmented Contextual 3D Visualisation of Planned Urban Change

Sven Uythoven [1], Wissam Wahbeh[1], Théo Reibel[1], Stephan Nebiker[1], Alexander Erath[1]

[1] School of Architecture, Civil Engineering and Geomatics
FHNW University of Applied Sciences and Arts Northwestern Switzerland, Muttenz, Switzerland –
(sven.uythoven, wissam.wahbeh, theo.reibel, stephan.nebiker, alexander.erath)@fhnw.ch

Keywords: Generative AI, Procedural Modelling, Visualisation, City Models, Rendering Pipeline.

Abstract

Street-level visualisations play an important role in urban street redesign by helping stakeholders and non-specialist audiences understand proposed spatial transformations. However, conventional image-editing workflows are often time-consuming, while purely image-based generative AI approaches provide limited control over object position, scale, and spatial relationships. This study presents a novel, highly automated workflow combining georeferenced street-level smartphone imagery, procedural 3D modelling, semantic image segmentation, and controlled generative AI to produce realistic visualisations of planned street transformations. Two-dimensional CAD drawings of planned changes are procedurally converted into simplified 3D models. Semantic masks are rendered from georeferenced camera viewpoints in the 3D model and combined with masks extracted from the street-level smartphone images. The resulting semantic and inpainting masks condition a Stable Diffusion XL-based inpainting model through segmentation and Canny-edge ControlNets. Text prompts control the overall visual appearance, while an IP-Adapter enables selected elements, such as vegetation and tree species, to be refined using reference images. The workflow was applied to two street redesign scenarios. The results show that planned elements can be generated at positions and with dimensions closely corresponding to the planning data, while different visual and seasonal variants can be explored without changing the spatial configuration. The method therefore provides a scalable approach for generating street-level visualisations while preserving geometric control.

1. Introduction

Urban street redesign is a central element in addressing current challenges related to climate change, urban liveability, and sustainable mobility. Increasing urban greenery contributes to heat mitigation, improves air quality, enhances biodiversity, and strengthens overall quality of life (Tzoulas et al., 2007). Street redesign projects give the opportunity not only to improve transport or energy infrastructure but also support environmental objectives.

Urban projects can significantly affect the spatial, social, and environmental qualities of a community and therefore depend on public understanding of the changes proposed in the projects and the impact on the public. For this reason, proposed interventions should be communicated transparently, accurately, and in a manner that is accessible to non-specialist audiences. Representation consequently plays a central role in the planning process, as it shapes how spatial transformations are perceived, interpreted, and evaluated by the public.

The need for transparent and accessible project communication becomes particularly relevant when urban interventions involve complex, large-scale transformations of existing public spaces. In such cases, representation is not only a means of illustrating a proposed design but also a tool for supporting dialogue between planning authorities, stakeholders, and local communities. This is especially important for infrastructure projects, which often require extensive modifications to streets and public spaces while simultaneously creating opportunities for broader urban improvement.

A specific example is the expansion of the district heating network in Basel which is required by the ambitious Energy Act (Kanton Basel-Stadt, 2020). The present study is being conducted within this context in collaboration with the local public urban planning authorities. The implementation of this infrastructure for renewable heating will require substantial interventions in existing street spaces over the coming decades. These works provide an opportunity to combine essential technical measures with the redesign of streets according to more sustainable, inclusive, and environmentally responsive principles.

Existing approaches to the visualisation of urban projects typically rely either on manual image-editing workflows or on generative AI-based methods. Manual techniques are generally time-consuming, labour-intensive, and difficult to scale, particularly when multiple viewpoints or design variants are required. Generative AI methods, by contrast, may produce visually compelling outputs more rapidly, but often lack consistency in the representation of the same object across different viewpoints and provide only limited control over metric accuracy and spatial relationships (Guridi et al., 2025). In particular, they do not ensure that planned elements such as trees or street furniture are placed at their intended positions according to the drawings and that parts of the streetscape not affected by the planning measures remain unchanged. These limitations restrict their use in professional planning applications, where visualisations must remain consistent between different views and with the underlying design.

Consequently, many urban projects are still communicated to the public through a limited number of 3D visualisations, while a substantial part of the information continues to be conveyed through conventional 2D drawings. This restricts the public's ability to fully understand the spatial implications of proposed interventions and to assess their potential impact on the existing urban environment.

To address these limitations, this research introduces a novel workflow that combines a procedural transformation of 2D planning data into a 3D model, georeferenced street-level imagery, and generative AI. In contrast to purely image-based methods, the proposed approach ensures that the position and scale of the visualised objects correspond to the CAD drawings.

The objective of this work is to develop a highly automated pipeline for generating realistic, geometrically controlled, and context-aware visualisations of street redesign scenarios. Ultimately, the proposed method seeks to provide a reliable tool to support planning processes, improve communication with stakeholders, and increase the acceptance of sustainable street transformations.

2. Related work

This section reviews previous work on generative AI in urban design and the state of the art in controlled image generation. Semantic segmentation is also discussed, as it is a key component of the proposed workflow for identifying existing street elements and defining where exactly image generation should be applied.

2.1 Generative AI in urban design

Several recent studies have explored generative AI in urban design and participatory planning. Valença et al. (2025) discussed generative AI tools for street-design visualisation and argued that these tools may support stakeholder involvement by allowing future street layouts to be visualised more rapidly. However, they also emphasised that the integration of such tools into planning practice remains insufficiently studied. Guridi et al. (2025) investigated diffusion-based tools in the co-design of public spaces, highlighting their potential to translate stakeholder input into realistic visualisations and thereby facilitate communication. However, they reported limited control over the size and hierarchy of generated objects. A user-oriented tool was presented by Kapsalis (2024), who proposed UrbanGenAI, a workflow combining image segmentation, SDXL and ControlNet to enable users to select and modify specific urban elements in an image. Preussner et al. (2025) explored this participatory potential with UrbAI, a mobile system that allows citizens to photograph urban spaces, mark areas for improvement and generate visual representations of their ideas. Their findings show that AI-manipulated images can help citizens express ideas, reflect on possible changes and support discussion in early participation stages. However, they also stress that such images should not be understood as final design proposals, but rather as a basis for discussion, since the outputs may create unrealistic expectations.

Together, these studies show the potential of generative AI for rapid streetscape visualisation. They demonstrate that AI-generated images can help communicate future urban changes, encourage participatory design, and support early discussions about design alternatives. At the same time, the literature also shows that generated images can be unpredictable and may produce visually convincing results that are not necessarily realistic in terms of spatial or technical constraints. For professional planning applications, this inconsistency is a fundamental limitation of purely image-based generative approaches. The proposed workflow therefore addresses this limitation by grounding the generation process in a procedurally constructed 3D model derived from official 2D planning drawings provided in CAD format and high-precision 3D terrain models. This ensures that the scale and spatial arrangement of the design elements in the generated images correspond to the underlying planning data.

2.2 Generative AI for controlled image generation

The generation of photorealistic images from textual or visual conditioning inputs has advanced substantially with the development of latent diffusion models. Rombach et al. (2022) introduced latent diffusion models, the architecture underlying Stable Diffusion (SD). A latent diffusion model performs the denoising process in a compressed latent representation, enabling image synthesis at significantly reduced computational cost compared to pixel-space diffusion models. Podell et al. (2024) extended this framework with Stable Diffusion XL (SDXL), which employs a larger UNet backbone with a second text encoder and multiple novel conditioning schemes. To provide more control over generated images, Zhang et al. (2023) introduced ControlNet. ControlNet is a neural network architecture that enables pretrained diffusion models to be conditioned on additional spatial inputs, such as edge maps or depth maps. The method keeps the weights of the original diffusion model fixed and adds trainable network branches that learn how to integrate the external conditioning information into the denoising process (Zhang et al., 2023). Another method to have more control over the appearance of specific elements in the generated image is by using IP-Adapter (Image Prompt Adapter) proposed by Ye et al. (2023). While ControlNet mainly provides spatial guidance, the IP-Adapter introduces image-based conditioning to guide the visual appearance of the generated content. It uses a reference image as an additional prompt and injects visual features into the diffusion model through a lightweight adapter module, without modifying the weights of the base diffusion model. This makes it possible to combine textual prompts with visual references and to control aspects such as style, texture, shape, and object appearance more directly (Ye et al., 2023).

2.3 Semantic image segmentation

The Cityscapes dataset introduced by Cordts et al. (2016) is a widely used benchmark for semantic segmentation of street scenes and urban scene understanding. It contains images from 50 cities, mainly in Germany, and provides both high-quality pixel-level annotations and coarser annotations. The dataset defines 19 classes and a void class containing rare or ambiguous objects (Cordts et al., 2016).

Deep learning methods have become widely used for semantic segmentation of street scenes. Several architectures have been evaluated on the Cityscapes dataset, with performance measured using the mean Intersection over Union (mIoU) metric (Cordts et al., 2016). Recent approaches based on transformer architectures have shown strong performance in street-level semantic segmentation tasks.

SETR, introduced by Zheng et al. (2021) is one of the first methods in which the CNN backbone is replaced with a transformer-based encoder (Li et al., 2024). SETR achieves 82.2% mIoU on the Cityscapes dataset (Zheng et al., 2021). In comparison, SegFormer is a simple yet effective semantic segmentation approach, with model variants ranging from lightweight real-time models (SegFormer-B0) to high-performance architectures (SegFormer-B5). The largest variant achieves 84.0% mIoU on the same benchmark. SegFormer surpasses SETR in performance while using fewer parameters (84.7M vs 318.3M) and offering improved computational efficiency (Xie et al., 2021).

3. Method

Our method establishes a highly automated pipeline for generating realistic visualisations of planned street redesigns and has been developed in collaboration with the planning department of the Canton of Basel according to their practical needs. First, a series of smartphone images are captured along the street corridor and accurately georeferenced using a bundle adjustment (section 3.1). The 2D planning drawings are automatically transformed into 3D models and enriched with urban and vegetation assets using a procedural approach. From these models, colour-coded semantic masks containing all added and modified elements are rendered for selected image poses (section 3.2). The street-level images are then semantically segmented using an AI-based segmentation model to identify the existing urban elements. The resulting existing-state semantic masks are combined with the project-state semantic masks rendered from the 3D model. This process produces two complementary control images: a fused, colour-coded semantic mask and a binary inpainting mask defining the regions affected by the planned intervention (section 3.3). Finally, these masks are used as conditioning inputs for a generative AI model, which produces realistic visualisations of the future project situation (section 3.4). The automation across the different processing steps enables the efficient generation of multiple street-level views with limited manual intervention.

3.1 Image acquisition and georeferencing

Street-level images are acquired with a mobile phone (iPhone 14 Pro) while walking along the street. The image acquisition is automated using the Pix4Dcatch application, with one image recorded every 10 to 15 cm. This procedure ensures a high overlap between adjacent images and an approximate relative orientation, which is required for reliable photogrammetric reconstruction. The images were subsequently processed using Agisoft Metashape through a Structure-from-Motion (SfM) workflow. Depending on the street, between seven and ten ground control points (GCPs) were distributed along the study area and measured using real-time kinematic GNSS. Their coordinates were recorded in the Swiss national coordinate reference system LV95 (EPSG:2056). The GCPs were included in the bundle adjustment to georeference the reconstruction and improve the accuracy of the estimated camera poses. Figure 1 shows an example of the image acquisition pattern used in this study, including the image capturing trajectory and the location of the GCPs. Depending on the street, the resulting total 3D root mean square error at the GCPs ranged from 4 to 7 cm, and a mean reprojection error between 1.2 and 2.5 pixels was achieved. In addition to the camera position and orientation estimation, a 3D mesh of the existing street environment was reconstructed. The mesh was used in the subsequent procedural modelling workflow to take into account the inclination and elevation changes of the street surface. Finally, the reconstructed camera poses were transferred to the Rhinoceros 3D environment and used to define the corresponding camera views for semantic rendering.

3.2 Procedural 3D modelling and semantic rendering

A procedural modelling algorithm was developed within the Rhinoceros 3D environment, using Grasshopper as visual programming and a script-based computational platform. The algorithm is designed to automate the conversion of 2D CAD drawings into 3D geometric objects by extruding them according to manually defined input parameters, thus preserving a high degree of parametric flexibility. This approach enables the generated model to be efficiently adjusted, refined, or adapted to different spatial configurations without requiring extensive manual remodelling.

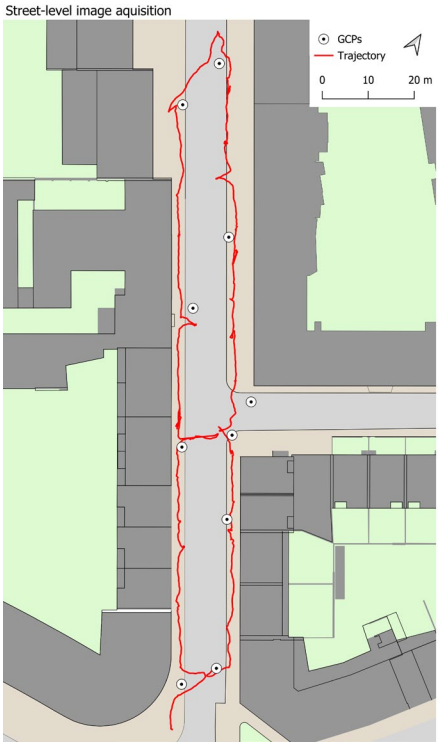


Figure 1. Street-level image acquisition pattern used for the photogrammetric reconstruction, showing the image capturing trajectory and the spatial distribution of the GCPs.

The workflow also supports the integration of generic three-dimensional assets, including trees, vehicles, street furniture, and other contextual elements, which are primarily used for masking and semantic classification purposes. The spatial placement of these assets is derived from their 2D positions in the planning CAD drawings (Figure 2). The corresponding 2D reference points are projected onto predefined 3D surfaces representing the street or the extruded pavement geometry. The resulting projected locations, for example, the centre point of a tree, are then used to position instances of generic 3D models representing the respective asset category.

Because the primary objective in this step is only to capture the visible silhouettes, spatial extent, and occlusion relationships of these objects, their geometry, materials, and level of detail are intentionally simplified. This approach reduces both modelling effort and computational demand while preserving the geometric and semantic information required for the reliable generation of class-based masks.
Once the 3D scene has been created, the scene is viewed from one or more selected camera poses corresponding to the captured street-level images. The scene is then rendered using a class-based material assignment system to generate colour-coded semantic segmentation masks. The colour attribution follows the standardised ADE20K colour scheme, in which each semantic category, such as buildings, vegetation, roads, pavement, vehicles, and urban furniture, is represented by a predefined and visually distinct colour. The resulting rendered images therefore provide geometrically consistent semantic reference masks that can be used in the next steps by combining the project state masks with the street-level image masks.

https://doi.org/10.5194/isprs-archives-L-4-W2-2026-221-2026

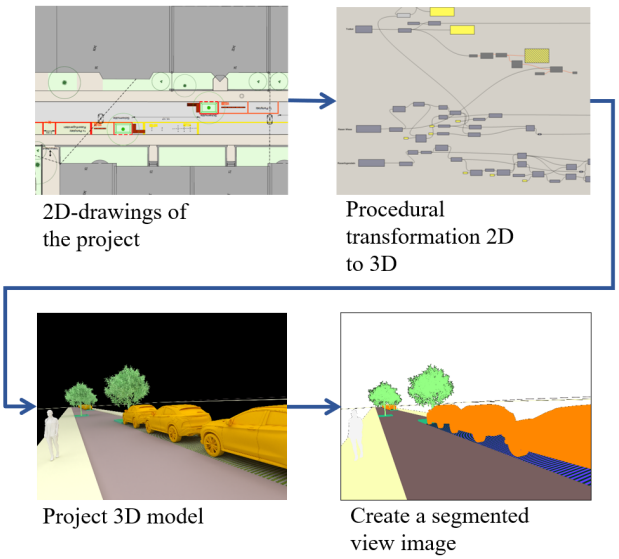


Figure 2. Pipeline from the 2D CAD planning drawings to the project-state segmentation mask.

3.3 Image segmentation and masking

First, the street-level image is semantically segmented. For this step, the high-performance SegFormer-B5 architecture (Xie et al., 2021) was selected, as it provides fine-tuned weights that enable straightforward inference on new images. Each input image is loaded and automatically processed using the corresponding image processor, which adapts the image to the required input format of the model. The original images (4032 × 3024 pixels) are resized during preprocessing to 1024 × 1024 pixels to match the model's expected input size. The pre-processed image is then passed through the network to compute class-wise logits, which are converted into probabilities using a softmax function. The final segmentation is obtained by selecting the class with the highest probability for each pixel. Since the output is produced at a lower spatial resolution, the probability maps are upsampled back to the original image dimensions, before being converted into the final segmentation map by assigning each pixel to the class with the highest predicted probability. Examples of the resulting segmentation outputs are shown in Figure 3.

Figure 3. Examples of the resulting semantic segmentation outputs for street-level images visualised using the Cityscapes colour palette, where each colour represents a different class.

The semantic mask extracted from the street-level image is mapped to the ADE20K colour palette. It is then fused with the project-state semantic masks generated from the procedural 3D model. This process results in two control images: a colour-coded semantic class mask and a binary black-and-white inpainting mask (Figure 4). While the semantic mask encodes the class membership of relevant scene elements, the inpainting mask defines the image regions influenced by the planned street intervention. The inpainting mask is additionally dilated with a fade by 50 pixels to enable a better transition between the generated elements and the street-level image.

A Canny-edge control image is also produced from a shaded rendering of the procedural 3D model by applying a Canny edge-detection filter. The shaded rendering captures the geometry and visible boundaries of the planned elements, while the resulting edge image provides structural guidance for their contours and spatial arrangement. The semantic mask, inpainting mask, and Canny-edge image are used as conditioning inputs for the generative AI model, enabling the creation of realistic visualisations of the future project state that remain aligned with the current street-level image and the geometry of the planned intervention.

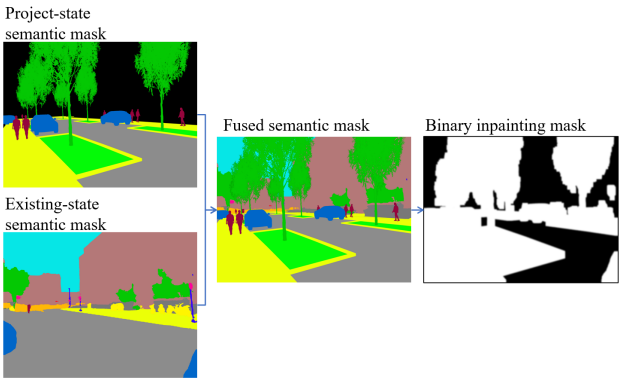


Figure 4. Fusion of the semantic mask extracted from the street-level image with the rendered mask generated from the procedural 3D model. From the fused segmentation, the binary inpainting mask is derived.

3.4 Generative AI

For the image generation, the pretrained latent diffusion model RealVisXL V5.0 (SG161222, 2024) was selected. RealVisXL is based on SDXL and has been specifically optimised for photorealistic image synthesis. An SDXL-based model was chosen because it offers a high degree of controllability through extensions such as ControlNet (Zhang et al., 2023), while remaining computationally efficient enough to run on a standard workstation without requiring high-end GPU hardware. The image generation was performed using an inpainting pipeline, which allowed the synthesis process to be restricted to predefined masked regions of the input image. In this way, selected areas of the original street-level image could be modified while preserving the surrounding image content, such as buildings and other unchanged scene elements. The overall style and appearance of the output can be guided by a prompt and a negative prompt.

To improve the spatial controllability of the generation process, two ControlNet modules are applied. The first ControlNet is conditioned on the fused semantic mask and is used to guide the spatial arrangement of generated objects. For this purpose, the fused semantic mask is provided as conditioning input. This enables control over the position and size of newly generated elements, such as trees, grass or cars, during image synthesis. The second ControlNet is based on the Canny-edge image. It is used to preserve and reinforce important structural edges in the scene, in particular the boundaries of the pavement and adjacent street elements. Figure 5 shows an example of the input images used for the image generation.

https://doi.org/10.5194/isprs-archives-L-4-W2-2026-221-2026

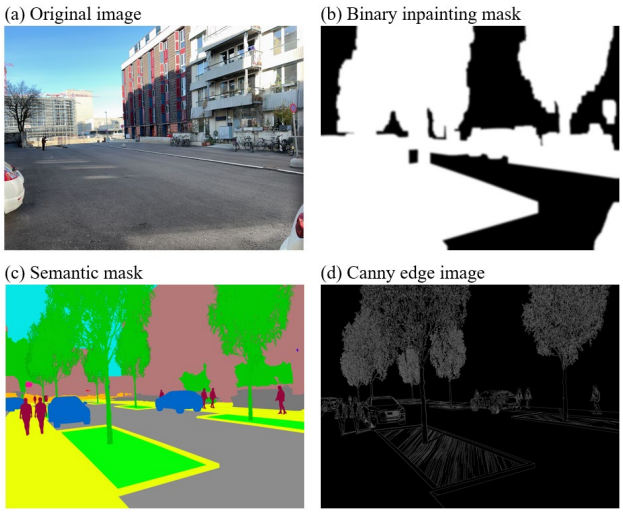


Figure 5. Example of the input images used as guidance for the image generation.

When using different controlling elements, the image generation process requires careful adjustment of the weights assigned to the different components, including the text prompt, the original image, the inpainting strength, and the two ControlNet conditions. These weights strongly influence the balance between realism, adherence to the input constraints, and generative flexibility. In this study, the inpainting strength was set to its maximum value because the masked regions represent a complete redesign and therefore differ substantially from the existing street scene. For the remaining conditioning components, no universal weight setting could be identified. Instead, these parameters must be adjusted for each new scene. The tuning is necessary because excessively high ControlNet weights over-constrain the base diffusion model, reducing its ability to generate photorealistic and coherent image content. On the other hand, if the ControlNet weights are set too low, the generated image no longer follows the semantic mask and therefore does not respect the intended position and scale of the generated elements. Consequently, the final image quality depends on finding a suitable compromise between strong structural guidance and sufficient generative freedom.

To further control the visual appearance of the generated elements, an IP-Adapter can be applied in a subsequent refinement step. The IP-Adapter uses a reference image as an additional visual condition and transfers its appearance characteristics to the generated output. In this study, it was used to guide the specific tree species and to change the appearance of the planting / water retention area.

By combining mask-based inpainting, segmentation-based spatial control, Canny-edge guidance, and image-prompt-based appearance control, the proposed workflow enables targeted and photorealistic modifications of existing street-scene images while maintaining consistency with the original urban context.

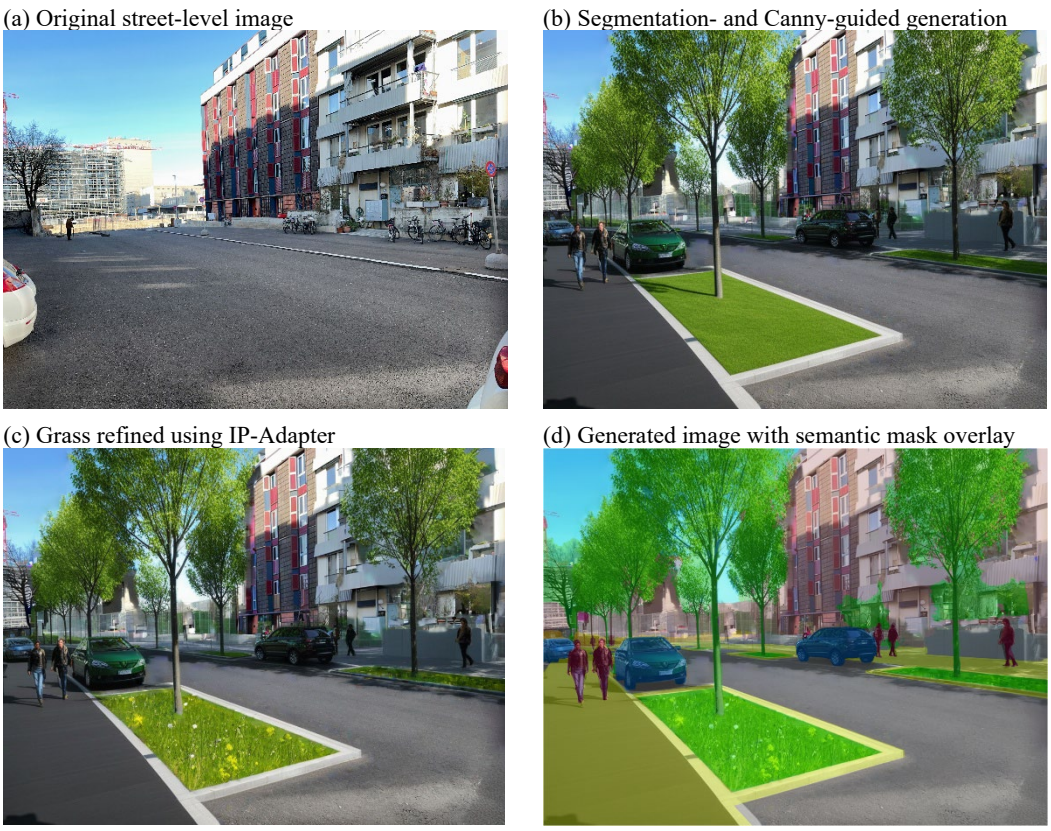


Figure 6. Results for the first street scene: (a) original street-level image; (b) generated redesign visualisation using segmentation and Canny conditioning; (c) refinement of the ground vegetation using IP-Adapter; (d) superposition of image (c) with the semantic mask to assess the spatial correspondence.

https://doi.org/10.5194/isprs-archives-L-4-W2-2026-221-2026

4. Results

The proposed workflow was applied to two different streets in the city of Basel and enables the generation of realistic and geometrically consistent visualisations of street redesign scenarios from georeferenced street-level imagery. Figures 6 and 7 illustrate a representative example of the results of the two different streets. In both cases, the workflow integrates planned elements, including new trees, planting / water retention areas, cars, pedestrians and modified street layouts, into the existing scenes. The positions and dimensions of these elements are defined by the planning data and the procedurally generated 3D model, while the generative model produces their final visual appearance.

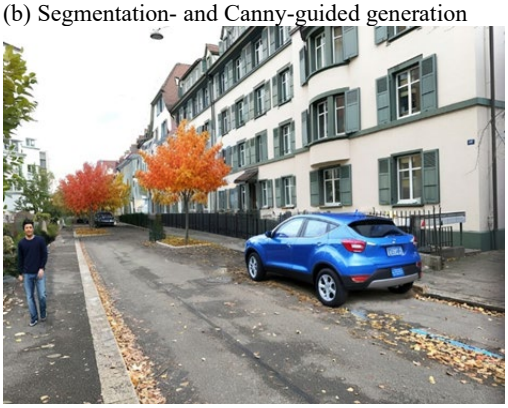
Figure 6 presents the results for the first street. The generated image in Figure 6b introduces a redesigned pavement, planting / water retention area, and rows of trees while maintaining the buildings and background of the original street-level image. The edges of the redesigned pavement and planting area remain sharp and clearly delineated. This shows the contribution of the Canny-edge ControlNet, which reinforces important structural edges during image generation. Figure 6c demonstrates the use of IP-Adapter to refine the appearance of the ground vegetation beneath the trees. In the initial generated image, the planting / water retention areas are represented by a relatively uniform grass texture. After applying IP-Adapter with a reference image of a wildflower meadow, a more heterogeneous combination of grass and small flowers is produced. The visual texture of the planting area is therefore modified without changing its intended location, extent, or surrounding curb geometry. To assess whether the generated elements respect the spatial constraints defined by the planning data, Figure 6d superposes the resulting image with the semantic mask. The comparison shows that the generated street layout, trees, cars and pedestrians correspond very closely to the positions and dimensions represented in the semantic mask. This indicates that the semantic conditioning input effectively constrains the spatial arrangement of the generated elements, also after using the IP-Adapter.

Figure 7 illustrates the results for the second street. Figure 7b and Figure 7c were generated using different text prompts while retaining the same spatial conditioning inputs. By modifying the text prompt, visual attributes such as vehicle colour, tree appearance, and the appearance and clothing of pedestrians can be adjusted. Figure 7b shows trees with autumnal foliage and a walking man wearing warmer clothing, suggesting an autumn setting. Figure 7c, in contrast, shows green trees and a walking woman wearing lighter clothing, suggesting a summer setting. This demonstrates that the text prompt can be used to explore alternative visual representations of the same planned spatial configuration. In a further step, IP-Adapter was applied to the result shown in Figure 7c to refine the appearance of the trees. As shown in Figure 7d, the tree species and crown characteristics are adapted towards those of a reference maple tree image. In contrast to the first street, where IP-Adapter is used to modify the texture of the ground vegetation, this example demonstrates its application to the appearance of larger individual objects.

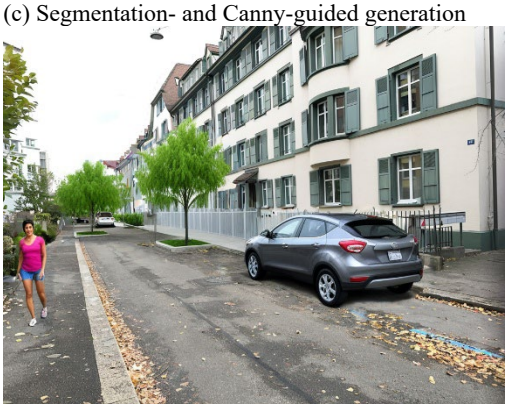
Both examples also reveal limitations in visual realism. The generative model has difficulty producing consistent shadows that correspond to the lighting conditions of the original street-level image. Furthermore, elements located farther from the camera are generated with less detail and visual coherence than larger foreground objects.

(a) Original street-level image

(b) Segmentation- and Canny-guided generation



(c) Segmentation- and Canny-guided generation



(d) Trees refined using IP-Adapter

Figure 7. Results for the second street: (a) original street-level image; (b-c) segmentation- and Canny-guided redesign visualisations using different prompts; (d) refinement of the tree appearance in image (c) using IP-Adapter.

https://doi.org/10.5194/isprs-archives-L-4-W2-2026-221-2026

Overall, the two examples demonstrate the controllability of the proposed workflow. The semantic conditioning inputs determine the spatial arrangement of the planned objects, the Canny-edge ControlNet supports the generation of clear edges and boundaries, the text prompt enables variations in the overall appearance of the scene, and IP-Adapter allows selected elements to be refined using visual references. The superposition presented in Figure 6d further provides a confirmation that the generated images respect the positions and dimensions defined by the planning data.

5. Conclusions and outlook

The proposed procedural workflow provides several advantages over conventional image-editing approaches, including Photoshop-based workflows and image-generation methods that do not provide metric control over proportions and perspective. While image-based techniques can produce convincing results, subsequent modifications often require extensive manual adjustments or repeated image-generation processes. In contrast, the parametric structure of the developed 3D workflow allows individual elements, such as trees, vehicles, street furniture, or architectural components, to be replaced, repositioned, or modified rapidly without reconstructing the entire scene.

A further advantage is the possibility of generating semantic masks from multiple viewpoints with only limited additional effort, since all street-level image poses are estimated as part of the SfM / bundle adjustment process. Once the 3D model and its semantic class assignments have been established, different camera positions can be selected and rendered automatically. For each additional viewpoint, only one to two working hours were required to identify suitable generation parameters and obtain a satisfactory visualisation. This makes the method particularly suitable for projects requiring several street-level perspectives or repeated evaluations of alternative design configurations.

Moreover, the use of a geometrically defined 3D environment provides integrated control over scale, dimensions, object placement, and spatial relationships. This represents an important advantage over purely image-based workflows, in which geometric accuracy and consistency between different views can be difficult to verify. The procedural approach therefore supports not only efficient mask generation but also reproducibility, spatial consistency, and rapid adaptation to changing project requirements.

Overall, the developed method offers a flexible and scalable alternative to conventional image-editing workflows, reducing the effort required for larger projects by almost an order of magnitude compared with current practice. Its main strengths lie in the efficient modification of scene elements, the generation of multiple viewpoints without substantial additional modelling effort, and the continuous control of scale and geometry throughout the process. These capabilities could support planning authorities in communicating different design alternatives and perspectives to non-specialist stakeholders, thereby improving their understanding of the spatial implications of proposed street transformations.

However, despite these advantages, the proposed workflow also presents several limitations, primarily related to the visual realism of the generated images. The AI models currently used within the process offer limited capabilities and do not yet achieve a level of realism comparable to carefully crafted manual digital editing or to advanced commercial generative AI platforms. As a result, some visual details, textures, and transitions may appear simplified or less convincing, particularly in complex urban scenes.

To address the limitations identified in this study, future work should focus on both the quantitative evaluation and visual refinement of the generated images. Spatial correspondence could be assessed using measures such as mask overlap, boundary deviation, and object-position errors instead of relying exclusively on visual overlays.

Furthermore, the integration of an additional post-processing stage using complementary tools for the fine-tuning and enhancement of the generated images is considered a viable extension of the proposed methodology. Such a step could improve visual coherence and realism while preserving the geometric consistency and procedural flexibility of the workflow. The investigation of these complementary refinement strategies therefore constitutes one of the authors' next research steps.

Acknowledgements

We would like to thank the Department of Construction and Transport of the Canton of Basel-Stadt for its valuable collaboration, technical input, and provision of the planning data used in this study.

References

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The Cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 3213–3223. doi.org/10.1109/CVPR.2016.350.

Guridi, J.A., Cheyre, C., Goula, M., Santo, D., Humphreys, L., Souras, A., Shankar, A., 2025. Image Generative AI to Design Public Spaces: A Reflection of How AI Could Improve Co-Design of Public Parks. *Digital Government: Research and Practice,* 6(1), 1–14. doi.org/10.1145/3656588.

Kanton Basel-Stadt, 2020. Verordnung zum Energiegesetz (Energieverordnung, EnV).

Kapsalis, T., 2024. UrbanGenAI: Reconstructing Urban Landscapes using Panoptic Segmentation and Diffusion Models. doi.org/10.48550/arXiv.2401.14379.

Li, X., Ding, H., Yuan, H., Zhang, W., Pang, J., Cheng, G., Chen, K., Liu, Z., Loy, C. C., 2024. Transformer-based visual segmentation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 46(12), 10138–10163. doi.org/10.1109/TPAMI.2024.3434373.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R., 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. *International Conference on Learning Representations (ICLR)*.

Preussner, A., Crescenza, A., Schöning, J., Mathis, F., 2025. UrbAI: Exploring the Possibilities of Generative AI Image Processing to Promote Citizen Participation. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, 1–21*. doi.org/10.1145/3706598.3713803.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 10674–10685. doi.org/10.1109/CVPR52688.2022.01042.

SG161222, 2024. RealVisXL V5.0. Hugging Face model repository. huggingface.co/SG161222/RealVisXL_V5.0 (20 June 2026).

Tzoulas, K., Korpela, K., Venn, S., Yli-Pelkonen, V., Kaźmierczak, A., Niemela, J., James, P., 2007. Promoting ecosystem and human health in urban areas using Green Infrastructure: A literature review. *Landscape and Urban Planning,* 81(3), 167–178. doi.org/10.1016/j.landurbplan.2007.02.001.

Valença, G., Azevedo, C., Moura, F., Morais De Sá, A., 2025. Creating visualizations using generative AI to guide decision-making in street designs: A viewpoint. *Journal of Urban Mobility,* 7, 100104. doi.org/10.1016/j.urbmob.2025.100104.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. doi.org/10.48550/arXiv.2105.15203.

Ye, H., Zhang, J., Liu, S., Han, X., Yang, W., 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. doi.org/10.48550/arXiv.2308.06721.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* 3836–3847. doi.org/10.1109/ICCV51070.2023.00355.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S., Zhang, L., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 6877–6886. doi.org/10.1109/CVPR46437.2021.00681.