

Synthetic data generation for architectural typography documentation using diffusion models

Pedro Achancarray¹, Leonhard Wesche², Markus Gerke¹, Sebastian Hoyer², Klaus Thiele²

¹ Institute of Geodesy and Photogrammetry, Technische Universität Braunschweig,
Germany - (p.diaz, m.gerke)@tu-braunschweig.de

² Institute of Steel Structures, Technische Universität Braunschweig,
Germany - (l.wesche, s.hoyer, k.thiele)@tu-braunschweig.de

Keywords: system halls, deep learning, synthetic references, aerial image recognition, diffusion models.

Abstract

The identification and systematic recording of industrial buildings pose significant challenges for modern monument preservation. In particular, system halls have shaped the industrial landscape since the 19th century but often elude complete documentation because of their widespread distribution. These buildings serve as vital witnesses to technical innovations and economic transformation; however, assessing their architectural value requires a comprehensive inventory to determine the rarity or preservation state of specific building types. Deep learning (DL) approaches are commonly used for the automatic recording of these buildings in aerial photographs, where the primary obstacle is the scarcity of curated training datasets. We overcome this by employing generative AI, specifically Stable Diffusion (SD), to produce synthetic data. By fine-tuning the SD model with Low-Rank Adaptation (LoRA), we successfully replicate the appearance and textures of various hall types. To resolve the spatial incoherence and geometric inaccuracies inherent in standard text-to-image generation, we integrated ControlNet. This allows for precise structural grounding using semantic masks, where specific colors represent building types, and polygon shapes define their exact locations. The resulting model generates accurate synthetic samples that maintain both spectral authenticity and an accurate spatial layout. Their usability was assessed by training a building detection model on both the real and synthetic datasets, achieving 71.9 and 66.7 mIoU, respectively. Moreover, introducing a few real samples for validation during training increased the mIoU to 82.7. The detection results demonstrate that the synthetic dataset is a reliable source for training, yielding robust generalization.

1. Introduction

The identification and systematic recording of industrial buildings pose enormous challenges to modern monument preservation. In particular, the group of serial hall structures, the system halls, which have shaped the landscape since industrialization, often eludes complete documentation owing to their widespread distribution. However, these buildings bear witness to technical innovations and economic transformation processes.

Automated recording in aerial photographs (orthophotos) using deep learning (DL) methods (Achancarray et al., 2023b) makes it possible to efficiently identify different hall types to obtain information about their distribution, location, and infrastructure. This not only serves as a pure inventory but also forms the essential basis for assessing the relevance of monuments. Only by gaining an overview of the rarity or degree of preservation of a specific type in comparison to the total inventory can a well-founded monument's value be derived.

Nevertheless, a key obstacle to the application of DL in monument preservation is the lack of sufficiently curated training datasets. Generative models can produce synthetic samples that are statistically similar to real-world samples, providing a viable alternative for training building detection algorithms. This training dataset can be built based on historical sources, such as system hall-type catalogs and advertisements, construction plans, and archived documents. These sources are employed to find instances of real halls by visual inspection of aerial photographs, which can be transformed (augmented) to increase the number of real instances in the training dataset.

Generative models offer powerful and flexible tools for creating synthetic image datasets for training detection models and

overcoming the scarcity of training samples. These models can augment datasets by creating large numbers of synthetic images that resemble real data (Isola et al., 2017) and by incorporating additional information (conditions or restrictions) to control the generated images (Mirza and Osindero, 2014, Zhang et al., 2023). These models typically treat the task as an image-to-image translation problem, where an input (such as a 2D building footprint, label map, or Digital Surface Model) is mapped to a photorealistic aerial representation.

The foundation of modern synthetic image generation, such as orthophotos, lies in Conditional Generative Adversarial Networks (CGANs) (Mirza and Osindero, 2014). Unlike standard GANs (Goodfellow et al., 2020) that generate images from random noise, CGANs condition the generation process using auxiliary information, such as building footprints or semantic maps. An extension of this model was Pix2Pix (Isola et al., 2017), which uses a U-Net-based generator and a "PatchGAN" discriminator to ensure local structural coherence. Pix2Pix includes an $L1$ reconstruction loss alongside adversarial loss to encourage the output to be spatially consistent with the input.

The need for finer control over building geometry led to the adoption of ControlNet (Zhang et al., 2023). While standard diffusion models, such as Stable Diffusion (Rombach et al., 2022), excel at creating realistic textures, they often struggle with the strict geographic and geometric constraints. ControlNet addresses this by adding a trainable adapter to a pre-trained diffusion backbone, allowing the model to be conditioned on spatial maps, such as Canny edges, Hough lines, or building segmentation masks. Recent research, such as the EcoMapper framework (Goktepe et al., 2025), has successfully used ControlNet to generate climate-aware satellite imagery, ensuring that

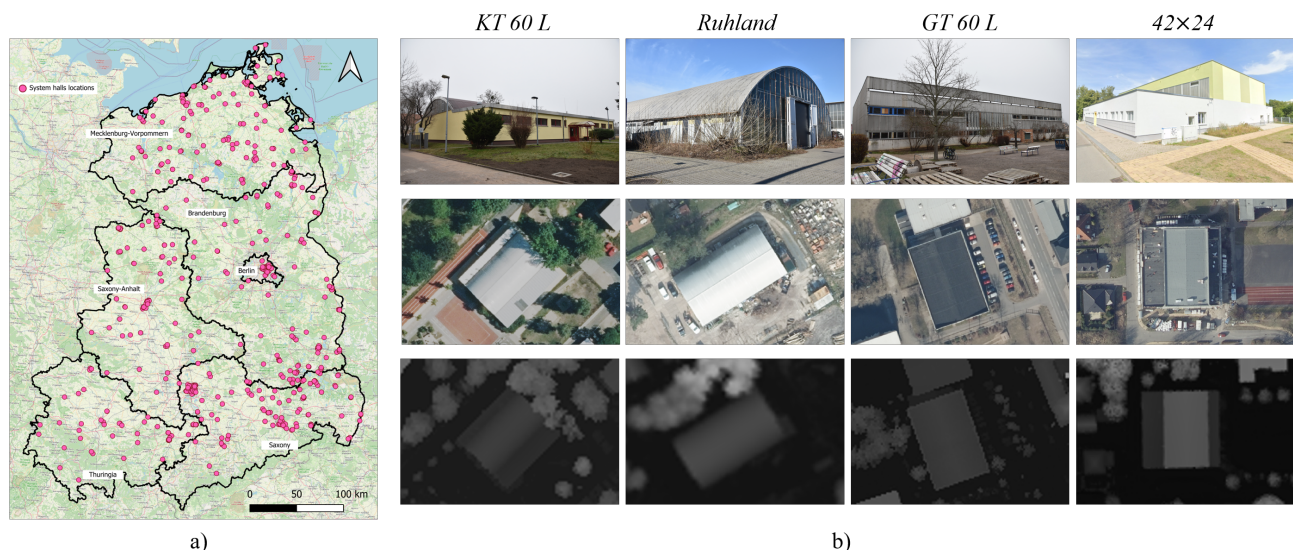


Figure 1. System halls dataset. a) Locations of instances of different system hall types extracted from (Heinrich et al., 2024). b) Samples of the four system hall types available in the dataset, from left to right: *KT 60 L*, *Ruhland*, *GT 60 L*, and *42x24*. From top to bottom: side view, Digital Orthophotos (DOP), and Normalized Digital Surface Models (nDSM).

the synthesized buildings maintain their exact footprint while adapting their visual appearance to specific environmental conditions.

The most recent work in orthophoto synthesis has moved beyond 2D translation toward implicit neural representations. Models such as Ortho-NeRF (Chen et al., 2025) leverage Neural Radiance Fields (NeRF) to generate "true" digital orthophoto maps (TDOMs) by modeling the entire 3D scene geometry from sparse UAV imagery. This approach effectively eliminates common artifacts, such as relief displacement and "ghosting" that plague traditional 2D generative methods.

In this context, this study proposes a methodology for generating synthetic data based on Stable Diffusion and ControlNet. The proposed methodology generates synthetic samples of different system hall types based on given conditions, such as edges (polygons with the location to generate the buildings) and geometric restrictions (geometric dimensions of each system hall type). Thus, the research questions guiding this work are the following:

- Can generative models produce synthetic samples of different system hall types?
- How can a priori knowledge about system hall types, such as geometric dimensions or roof types, be introduced into generative models to enhance the production of synthetic samples?
- Are synthetic samples of different system hall types a reliable source to train building detection models for system hall classification?

The remaining parts of this work are organized as follows: Section 2 describes the dataset used to train the generative model, Section 3 explains the proposed methodology for the synthetic generation of different system hall types, Section 4 gives details about the experimental protocol and setup for our experiments and presents the results obtained for different training approaches, and Section 5 summarizes our findings and future research areas that can be explored.

2. Dataset

In the present study, a specific dataset was compiled to serve as a reference system for testing serial hall constructions (see Figure 1). We focus on four specific building types that serve as ideal test objects because of their extensive documentation: *KT 60 L*, *Ruhland*, *GT 60 L*, and *42x24*. Consequently, the resulting dataset is characterized by a high density of meta-information and precise spatial resolution as follows:

- System halls types information: technical specifications, typological features, and construction details of the four hall systems (based on catalogues and design plans) extracted from (Wesche et al., 2023)¹.
- Object information: a list of specifically constructed examples that serve as real reference objects/instances for training and validation. These instances were obtained from (Heinrich et al., 2024)².

The locations were primarily acquired through systematic manual recording of aerial photographs. In addition, the first locations were identified through pilot automated AI-supported preliminary recordings and added to the inventory after review (Achanccaray et al., 2023b). This hybrid approach allows both reliable "ground truth" data and real application scenarios of automated searches to be mapped. Figure 1a shows a map of the locations of the 188 instances available in the dataset with the following distribution per type: *KT 60 L*(81), *Ruhland*(74), *GT 60 L*(10), and *42x24*(23).

High-resolution orthoimages were available for the identified locations, where a total of 110 image tiles were collected from the following states in Germany: Berlin (7), Brandenburg (16), Mecklenburg-Vorpommern (29), Sachsen (20), Sachsen-Anhalt (15), and Thüringen (23). The orthoimages were provided at 5K x5K or 10K x10K, depending on the federal state, with 20 cm spatial resolution. From each image tile, patches of 512 x 512

¹ Available at 10.24355/dbbs.084-202305261242-0

² Available at 10.24355/dbbs.084-202403130624-0

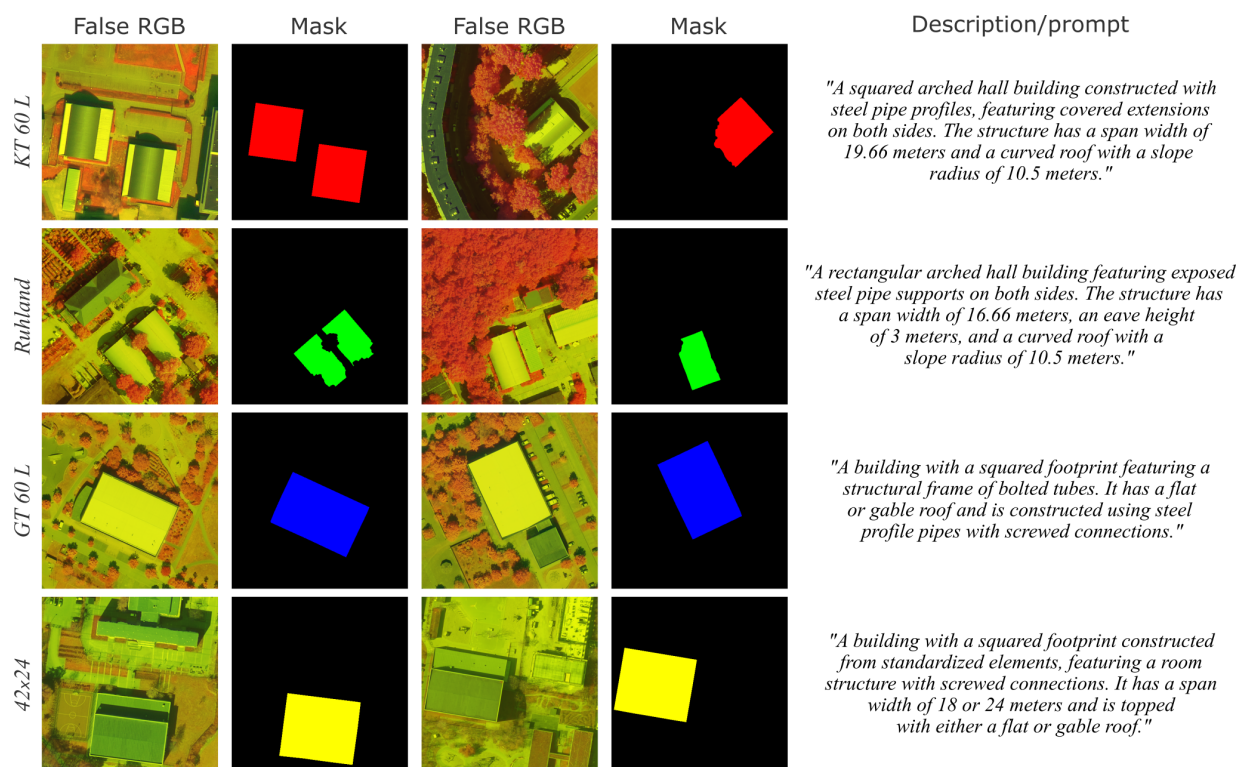


Figure 2. Samples of the created dataset to train the generative model. On the left, false RGB color compositions and their corresponding masks for each system hall type. On the right, a description of each system hall type based on the meta-information acquired from catalogues and technical specifications.

were extracted, changing the relative position of each building in the patch to the center, south, west, north, and east, where patches with more than 50% overlap were removed to skip redundant information. Digital Orthophotos (DOP) and Normalized Digital Surface Models (nDSM) were acquired from each state's geoportals, as described in (Achancaray et al., 2023a). The DOP, containing four bands (Red, Green, Blue, and Infrared), provides information about the surface details, materials, and roof structures, while the nDSM provides height information to distinguish hall types and delimit them from their surroundings (e.g., vegetation or flatter outbuildings). Figure 1b presents a sample of each system hall type, consisting of a side view, DOP, and corresponding nDSM.

As our main purpose was to create synthetic data that could be used to train building detection methods, we focused on the generation of false RGB color compositions, similar to (Achancaray et al., 2023b), where these compositions were used to train instance segmentation methods. The false RGB color compositions were created as follows: infrared (R), gray (G), and nDSM (B), where gray is the arithmetic mean of the Red, Green, and Blue channels from the DOP. Figure 2 shows samples of the false RGB color compositions for each system hall type and their corresponding mask, which were manually delineated, and a description of each system hall type extracted from the meta-information acquired from catalogues and technical specifications. In this description, specific geometric details are included and used as a prompt to train the generative model.

3. Methodology

To create synthetic images from different system hall types, we propose a methodology based on Stable Diffusion (SD) and

ControlNet, consisting of two main parts: generative model training and generation of synthetic data (see Figure 3).

For generative models, a dataset with real images was employed to fine-tune the SD with Low-Rank Adaptation (LoRA) to adapt the base SD model to the building domain to improve structural realism, texture distribution, and domain-specific appearance. The LoRA weights were then merged into the SD base model. Finally, ControlNet was trained on the merged SD model with the following inputs: a prompt (description of each system hall type) and a condition (mask with the desired position of the buildings). During ControlNet training, the Variational AutoEncoder (VAE), UNet, and text encoder were frozen, while the branch with the condition was fine-tuned.

To generate synthetic data, the previously trained ControlNet

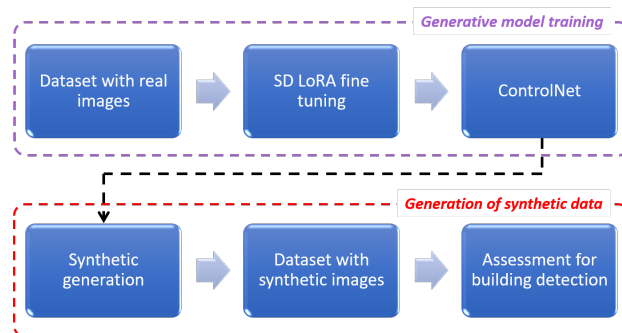


Figure 3. Proposed methodology for the synthetic generation of images from different system hall types, consisting of two main parts: generative model training (using the dataset with real images) and generation of synthetic data (using a synthetic dataset to train a building detection method).

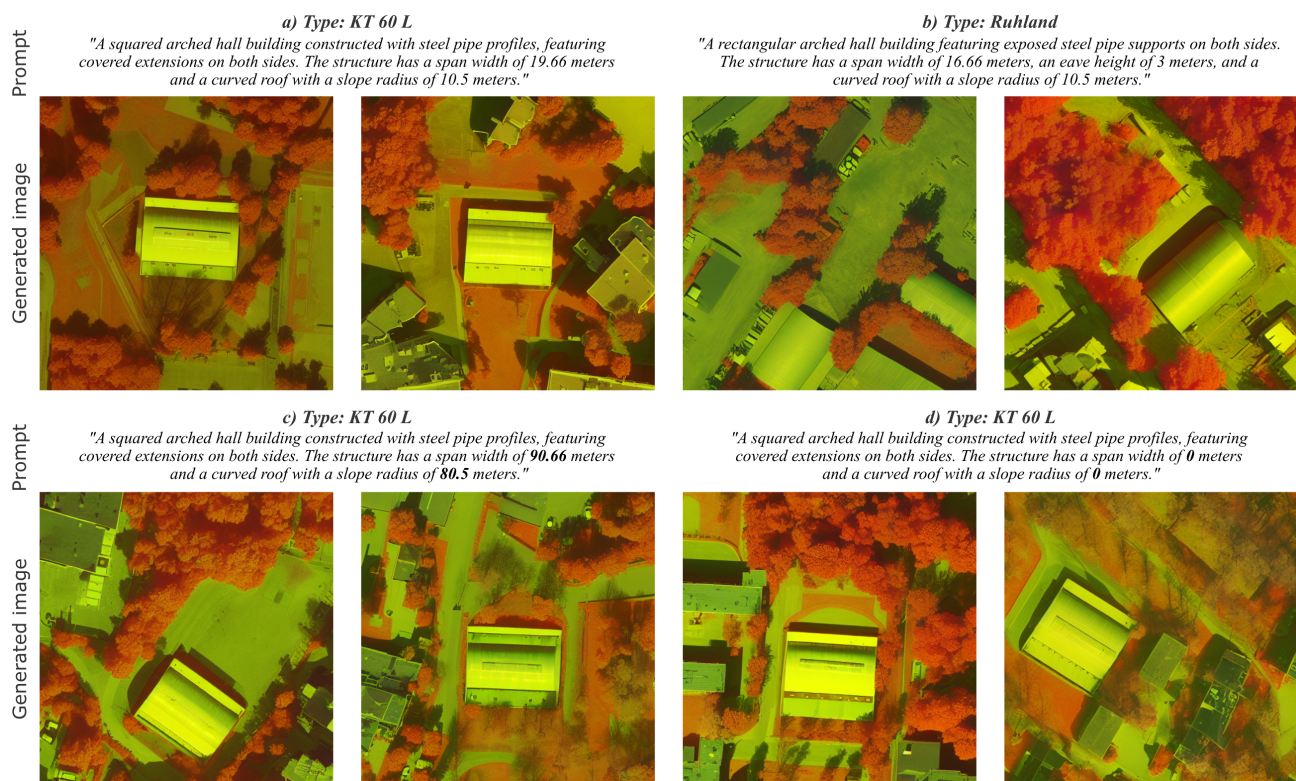


Figure 4. Samples of synthetic images generated after fine-tuning the Stable Diffusion v1.5 with LoRA. The prompt used to generate each image is shown at the top of each sample group. Two system hall types are presented: a) *KT 60 L* and b) *Ruhland*. c) and d) show cases where the prompt was changed using inaccurate values for specific geometric dimensions.

was executed under different randomly generated conditions (a set of masks for each system hall type) to produce synthetic samples. Then, a dataset containing synthetic images and their corresponding masks was created to train a building detection method using an instance segmentation algorithm. Finally, the synthetic images were assessed for usability by comparing the detection method's performance on real and synthetic images.

4. Results

4.1 Experimental Protocol and Setup

The dataset described in Section 2, hereafter called "the real dataset," contained 412 patches for training and 303 patches for testing. The training set was augmented to 1800 samples using horizontal and vertical flips and rotations to increase the number of samples and reduce the class imbalance.

We fine-tuned SD (version 1.5) with LoRA³ to adapt SD to our domain. The input image size was 512×512 , and the network rank (*network_dim*) and scaling factor (*network_alpha*) were equal to 32 and 16, respectively. AdamW (Loshchilov and Hutter, 2017) as optimizer, and the UNet and text encoder learning rates were $1e-4$ and $1e-5$, respectively. The training batch size was 2, with a maximum number of steps of 5000 and two epochs.

To train ControlNet, the LoRA weights were merged into the SD base model. The inputs were the prompt (description of each system hall type) and condition (desired position of the

buildings). From SD fine-tuning, the prompt learns the appearance (style) of the buildings, whereas the condition sets the spatial structure of the buildings. For the condition, the RGB mask and its respective contour were concatenated to create a four-band tensor. As mentioned before, only the condition branch was trained, whereas the VAE, UNet, and text encoder components were frozen. The learning rate was $5e-6$, with a maximum number of steps of 20 K, a batch size of 8, and a gradient accumulation of 8. Once the generative model was trained, 400 images per system hall type were generated to create a dataset, hereafter called "the synthetic dataset".

To assess the usability of the synthetic dataset, we trained a Mask R-CNN (He et al., 2017) with a ResNet-50-FPN (Li et al., 2021) as the backbone for building instance segmentation considering the four system hall types. The input image size was 512×512 , batch size of 8, AdamW as optimizer, learning rate of $2e-4$, and cosine learning rate warm-up. Three different training approaches were evaluated: i) training and validation with the real dataset, ii) training and validation with the synthetic dataset, and iii) training with the synthetic dataset and validation using the real dataset. The following metrics were calculated to compare all models: Intersection over Union (IoU) (a measure of the level of spatial agreement between the predicted and reference polygons), Precision (to analyze the number of false positives), and Recall (to analyze the number of false negatives). All metrics were computed using the test set of the real dataset for a fair comparison between the models.

4.2 SD LoRA fine-tuning

Figure 4 presents samples of synthetic images generated after fine-tuning the SD with LoRA for two system hall types: a) *KT*

³ Based on Kohya scripts

60 L and b) *Ruhland*. The input prompts are shown at the top of each sample group.

In both cases, we observe a high similarity in terms of spectral appearance between the generated images and the real images from the real dataset (Figure 2) with reddish vegetation (as infrared is in the red channel) and yellowish-green roof tops. However, the spatial layout is not accurate, as there are roads and trees positioned at uncommon locations (Figure 4a) intersecting the building in a way that defies architectural logic. This suggests that while LoRA successfully captured the *style* and *texture* of the aerial sensor, it did not learn the *topological* constraints of an urban environment. Figure 4b, right side, shows background buildings with irregular shapes or a "melted" effect, which might be related to the size of the dataset used for fine-tuning, struggling to maintain the sharp and rectilinear edges of man-made structures.

Figure 4c and d show two cases in which the prompt was changed for the *KT 60 L* type. The geometric dimensions associated with this type were modified to analyze the dependency of the generated image with respect to the prompt. As expected, the model did not learn from the specific geometric dimensions given in the prompt, but from the entire sequence of words/tokens describing each type. The main reason for this is related to the number of samples for each type, because if we would like the model to learn from specific dimensions, we need to provide examples of many different combinations so that the model can infer that these geometric dimensions are a discriminative feature to be considered.

At this point, as the prompt is the unique input of the model, we cannot control the desired position of the building, which is introduced as a condition in the ControlNet model in the next section.

4.3 ControlNet

Figure 5 shows samples of the synthetic dataset generated after training the ControlNet with the fine-tuned SD with LoRA for the four system halls. LoRA still applies the texture style, but ControlNet handles the geometry. Below each synthetic image, the input condition is presented, which contains a black background with colored polygons on it. The location, rotation, and shape of the polygon define where and how the building should appear. The color of the polygon represents a different system hall type.

For the *KT 60 L* type, with the red polygons in the mask, we observed that the buildings are located in the desired position and orientation, even in cases where the polygon is complex with an irregular shape (first column) because of occlusion with trees. Moreover, the generated shadows are consistent in terms of direction based on the position of the sun and the location of objects, such as trees and other buildings. Regarding the spectral appearance, the generated textures and colors are according to the *KT 60 L* type, except for the last column, where the building roofs appear darker and with an unusual texture.

For the *Ruhland* type, with the green polygons in the mask, we can see the model's ability to handle multiple objects of the same type in one image (second and third columns), even if these objects are too close to each other (last column). The generated buildings have the typical rounded (arched hall) appearance of the *Ruhland* type. Similar to the previous case, the buildings are located precisely at the exact angles and positions

given in the mask. In the last column, two buildings are adjacent, as seen in the mask, where a single pixel separates them, while the generated image managed to create the two buildings, but the separation is not accurate, making one of the buildings larger than expected.

For the *GT 60 L* type, with the blue polygons in the mask, the generated buildings followed a strict rectangular form with sharp edges provided in the mask. Moreover, we can distinguish specific textures for this type, which are different from the other types, demonstrating that the use of different colors for each type correctly triggers different LoRA-learned visual characteristics.

For the 42×42 type, with the yellow polygons in the mask, we observed that the model generated buildings with varying orientations, textures, and appearances, including incomplete buildings (third column). The last column shows a different roof pattern from the others, with the roof split into three parts rather than two, which is a characteristic of this type.

In general, the addition of a colored mask as a condition to train the ControlNet successfully overcame the limitations found after fine-tuning the SD with LoRA. The mask color and polygon shape provide explicit, pixel-perfect constraints, eliminating the need to guess the geometry from the prompt. Moreover, the spatial layout consistency improved, where trees and roads were located around the desired building accurately, without irregular shapes, and preventing hallucinations. In conclusion, ControlNet complemented the LoRA-learned features (textures and appearance) with realistic and coherent geospatial structures.

4.4 Building instance segmentation

Table 1 summarizes the results obtained after training the Mask R-CNN for instance segmentation using the real and synthetic datasets to assess the usability of the synthetic dataset. Both datasets were split into three sets: training, validation and test. The models were trained on the training sets, where the validation set were employed to select the best model. The IoU, Precision and Recall were computed using the test samples from the real dataset. Three approaches were employed: i) training and validation using the real dataset, ii) training and validation using the synthetic dataset, and iii) training from the synthetic dataset and validation from the real dataset.

In general, as expected, the model trained using the real dataset yielded better results than the model trained using the synthetic dataset. However, the difference between these two models is not too high, with cases where the synthetic dataset produced better results, such as the Precision values for *KT 60 L* ($93.0\% > 88.8\%$) and *Ruhland* ($93.0\% > 85.3\%$), and IoU values for *Ruhland* ($81.9\% > 79.7\%$).

The addition of real samples for validation during training had a huge impact on *GT 60 L* and 42×24 , which are types with only a few known instances. For the *GT 60 L* type, there were gains in terms of IoU ($+43.7\%$) and Recall ($+57.3\%$), with a reduction in Precision (-7.3%), which means that the number of false negatives was reduced drastically, while the number of false positives slightly increased. For the 42×24 type, there were gains in terms of IoU ($+26.1\%$) and Precision ($+29.4\%$), with a small reduction in Recall (-1.5%), which means that the number of false positives decreased significantly, whereas the number of false negatives did not change significantly.

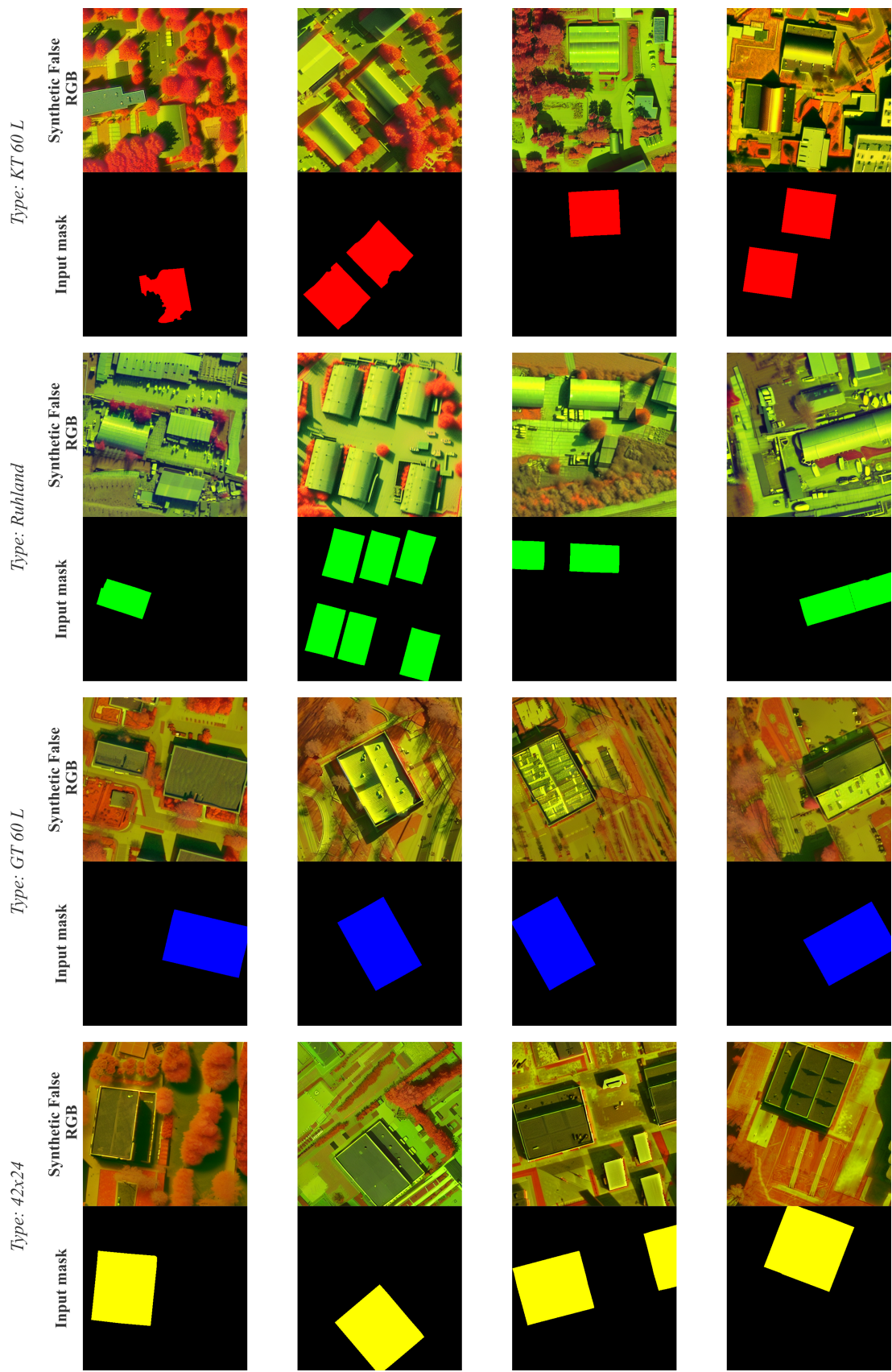


Figure 5. Samples of the synthetic images generated after training the ControlNet with the fine-tuned Stable Diffusion model with LoRA for the four system hall types: *KT 60 L*, *Ruhland*, *GT 60 L*, and *42 × 24*. The input conditions (masks) are presented below each image, where each color is associated with a specific system hall type.

Type	Data	↑ IoU	↑ Precision	↑ Recall
<i>KT 60 L</i>	Train: Real, Val: Real	86.9	88.8	97.6
	Train: Synthetic, Val: Synthetic	83.2	93.0	88.8
	Train: Synthetic, Val: Real	82.6	91.7	89.3
<i>Ruhland</i>	Train: Real, Val: Real	79.7	85.3	92.3
	Train: Synthetic, Val: Synthetic	81.9	93.0	87.3
	Train: Synthetic, Val: Real	76.8	86.7	87.1
<i>GT 60 L</i>	Train: Real, Val: Real	48.2	93.4	49.9
	Train: Synthetic, Val: Synthetic	39.7	91.8	41.2
	Train: Synthetic, Val: Real	83.4	84.5	98.5
<i>42×24</i>	Train: Real, Val: Real	72.7	73.8	98.1
	Train: Synthetic, Val: Synthetic	62.0	63.8	95.6
	Train: Synthetic, Val: Real	88.1	93.2	94.1

Table 1. Comparison between the results obtained after training the Mask R-CNN for instance segmentation using the real and synthetic datasets. All metrics were computed using the test samples from the real dataset. Three approaches are presented: i) training and validation from the real dataset, ii) training and validation from the synthetic dataset, and iii) training from the synthetic dataset and validation from the real dataset.

These results demonstrate that the trained generative model successfully produced accurate synthetic samples, as shown in Figure 5. These synthetic samples can be used to train a building detection method and accurately generalize to real samples according to the metrics presented in Table 1. Moreover, the detection model can be enhanced using a few real samples for validation during training. This analysis focused on the usability of the synthetic dataset; however, it is also interesting to verify whether the pixel values from the synthetic images have the same semantic meaning as the real images. For instance, in the infrared band, the vegetation must appear brighter with an irregular texture. In addition, in the nDSM band, trees and buildings must be brighter because they are taller than other objects, whereas buildings must have a soft texture because of the roof material.

Figure 6 shows a real (left) and synthetic (right) image with their respective bands: Infrared, Gray, and nDSM. In the Infrared band, a high similarity can be observed between the appearance and texture of the trees in both images. The Gray band exhibits similar behavior. In the nDSM band, we can clearly note that the building in the synthetic image does not have a soft texture, as in the real image. Based on this, we can infer that even if the generative model can produce synthetic images useful for training building detection methods, the bands of the synthetic images do not have the same semantic meaning as the real ones, especially the nDSM. One solution would be to force the model to generate a realistic nDSM by introducing a reconstruction loss focused on this band.

5. Conclusions

In this work, we proposed a methodology to generate synthetic samples of different system hall types as an alternative for training building detection methods. The generative approach was based on Stable Diffusion and ControlNet, where Stable Diffusion was fine-tuned with LoRA for better adaptation to the building domain. The inputs for ControlNet were a prompt describing the specific system hall type and condition with a colored mask containing polygons to indicate the location and orientation of the desired buildings. To assess the synthetic dataset generated using ControlNet, a Mask RCNN with a ResNet-50-FPN backbone was trained for building instance segmentation.

The Stable Diffusion model fine-tuned with LoRA successfully learned the appearance and texture of the images. However, the spatial layout was not optimal, and hallucinations produced background buildings with irregular shapes. ControlNet managed to overcome these limitations by exploiting the location, rotation, and shape of the polygon to define where and how the building should appear. The addition of this condition complemented the LoRA-learned features with realistic and coherent geospatial structures.

The synthetic dataset generated by ControlNet was useful for training a building detection model, which achieved metrics

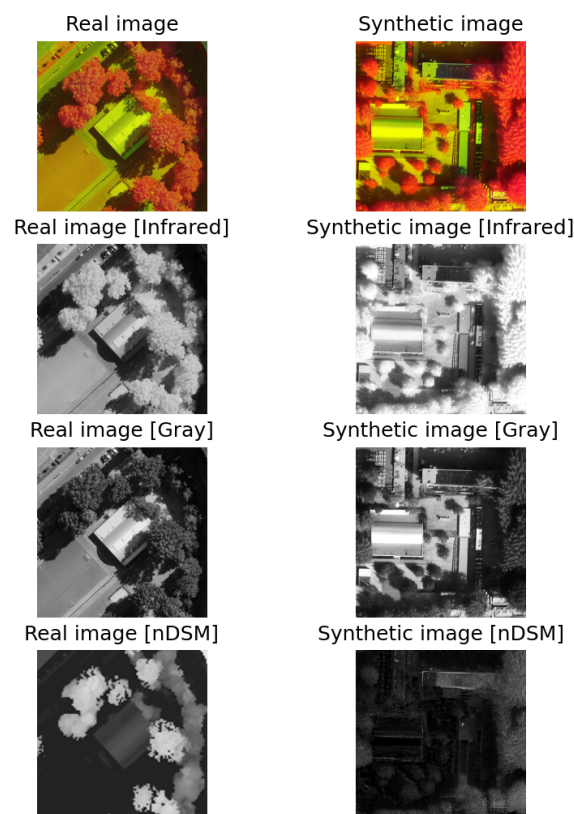


Figure 6. Comparison between a real and synthetic image, including their corresponding bands: Infrared, Gray, and nDSM.

very close to those of a model trained on a real dataset. Moreover, introducing a few images for validation during training significantly improved the results, particularly for system hall types with only a few known instances. This demonstrates the usability of the synthetic dataset for training detection models in the absence of enough real samples.

Further research will focus on analyzing the effect of the number of real training samples required to train the generative model and produce accurate synthetic samples. Moreover, more recent generative methods will be explored, including those that operate in pixel space rather than latent space.

Acknowledgments

We acknowledge the federal states where data was obtained from their geoportals: Berlin, Brandenburg, Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt, and Thuringia. This project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project Monument mass phenomenon system hall, project id 525823382.

References

- Achanccaray, P., Gerke, M., Wesche, L., Hoyer, S., Thiele, K., Knufinke, U., Krafczyk, C., 2023a. Automatic detection of specific constructions on a large scale using deep learning in very high resolution airborne imagery: the case of steel construction system halls of the high modernism period. *PFG-Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 91(3), 189–209.
- Achanccaray, P., Gerke, M., Wesche, L., Hoyer, S., Thiele, K., Knufinke, U., Krafczyk, C., 2023b. On the assessment of instance segmentation for the automatic detection of specific constructions from very high resolution airborne imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-1/W2-2023, 1303–1309. <https://isprs-archives.copernicus.org/articles/XLVIII-1-W2-2023/1303/2023/>.
- Chen, S., Yan, Q., Qu, Y., Gao, W., Yang, J., Deng, F., 2025. Ortho-NeRF: generating a true digital orthophoto map using the neural radiance field from unmanned aerial vehicle images. *Geo-spatial Information Science*, 28(2), 741–760. <https://doi.org/10.1080/10095020.2023.2296014>.
- Goktepe, M., hossein Shamseddin, A., Uysal, E., Monteagudo, J. M., Drees, L., Toker, A., Asseng, S., Von Bloh, M., 2025. EcoMapper: Generative Modeling for Climate-Aware Satellite Imagery. *Forty-second International Conference on Machine Learning*.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. *Communications of the ACM*, 63(11), 139–144.
- He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, 2961–2969.
- Heinrich, A., Mende, V., Wesche, L., Achanccaray, P., 2024. Database of recorded serial manufactured MLK-buildings (GDR) (Release 2). doi.org/10.24355/dbbs.084-202403130624-0.
- Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A., 2017. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.
- Li, Y., Xie, S., Chen, X., Dollar, P., He, K., Girshick, R., 2021. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*.
- Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Mirza, M., Osindero, S., 2014. Conditional Generative Adversarial Nets. *CoRR*, abs/1411.1784. <https://arxiv.org/abs/1411.1784>.
- 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.
- Wesche, L., Achanccaray Diaz, P. M., Hoyer, S., Knufinke, U., Gerke, M., Krafczyk, C., Thiele, K., 2023. Dataset of german steel system halls from the period of high modernism. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project Denkmal Massenphänomen Gewerbehalle? within the SPP2255 Kulturerbe Konstruktion, project id 442249122.
- 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*, 3836–3847.