Developments in the Built Environment 27 (2026) 100991

Contents lists available at ScienceDirect

Developments in the Built Environment

journal homepage: www.sciencedirect.com/journal/developments-in-the-built-environment

# Domain-adaptive generative data augmentation for building façade classification: Linking synthetic data to performance across urban contexts

Youngseo Hwang [a], Seunghyeon Wang [b], Namhyuk Ham [c], Sungkon Moon [a,*]

[a] Department of Civil Systems Engineering, College of Engineering, Ajou University, Suwon, Republic of Korea
[b] Department of Cognitive Convergence Science, Hanyang YK Intercollege, Hanyang University, Seoul, Republic of Korea
[c] Department of Architectural and Urban Engineering, Hanyang Cyber University, Seoul, Republic of Korea

ARTICLE INFO

*Keywords:*
Generative augmentation
Diffusion models
Low-rank adaptation
Synthetic data
Facade materials
Deep learning
Data classification

ABSTRACT

The limited availability of labeled data remains a major barrier to deploying computer vision models for building façade material classification. This study proposes a domain-adaptive generative augmentation framework that integrates architecture-aware prompt design with LoRA-based fine-tuning of a diffusion model. The framework generates synthetic façade images that preserve material-specific textures and region-dependent visual patterns. Unlike prior studies that mainly assess visual realism, this study links generative augmentation strategies directly to downstream classification performance. Experiments examined the effects of generation hyperparameters and LoRA training data volume, and the framework was validated using geographically distinct datasets from Korea and the England. Compared with conventional augmentation, the proposed method improved classification accuracy from 85.00% to 93.67% in Korea and from 71.67% to 83.67% in England. The results demonstrate that domain-adaptive generative augmentation reduces the synthetic-real data gap and improves practical classification robustness under data-scarce conditions for urban-scale built environment analysis.

## 1. Background and introduction

Exterior cladding material classification is a critical task in building façade image analysis because it enables the systematic understanding of building stock at an urban scale (Liu et al., 2017; Buyukdemircioglu and Kocaman, 2020). When automatically extracted from large-scale visual datasets, cladding information can support efficient urban-scale building analysis (Huang et al., 2023; Hwang et al., 2025). At present, building-level façade material information is commonly obtained through field surveys, expert interpretation of façade images, and fragmented administrative or property records. These approaches require substantial human effort, depend on specialist judgement and are difficult to apply consistently when thousands of buildings must be examined across extensive urban areas (Raghu et al., 2023). Aggregate material-production or sales statistics may describe regional trends, but they cannot identify which façade material is installed on each individual building or represent its current visual condition (Nasir et al., 2021).

Previous studies provide quantitative evidence of both the efficiency potential and the remaining limitations of image-based building surveys. As summarized in Table 1, remote façade surveying using street-view and aerial imagery has been reported to be approximately four times faster than conventional direct measurement. However, studies involving manual interpretation of street-view imagery reported audit times comparable with those of field surveys, excluding travel time. These findings indicate that remote imagery can reduce site-access and travel requirements; however, image acquisition alone does not eliminate the burden of human interpretation. Automated computer-vision analysis is therefore required to extend image-based surveys to large numbers of buildings without proportionally increasing the manual workload. Beyond technical performance, façade material information plays a critical role in supporting urban-scale applications in the built environment, including building-stock assessment, energy performance estimation, façade maintenance planning and digital twin development.

However, reliable façade classification remains constrained by the limited availability of diverse labeled datasets, whose collection and annotation require substantial time, cost, and domain expertise (Deng et al., 2009; Xiao and Kang, 2021). Conventional augmentation can improve model robustness but cannot introduce new material textures, architectural arrangements, or environmental conditions beyond those represented in the source data. Recent studies have therefore explored generative artificial intelligence to expand training datasets. In this

* Corresponding author.
*E-mail addresses:* dudtj7854@ajou.ac.kr (Y. Hwang), wsh1019@hanyang.ac.kr (S. Wang), nhham@hycu.ac.kr (N. Ham), skmoon@ajou.ac.kr (S. Moon).

https://doi.org/10.1016/j.dibe.2026.100991
Received 29 April 2026; Received in revised form 27 June 2026; Accepted 18 July 2026
Available online 20 July 2026
2666-1659/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

**Table 1**
Comparison of conventional field-based and remote or image-based survey approaches.

| Study | Survey Task | Compared Approaches | Reported Efficiency |
|---|---|---|---|
| Pittam et al. (Pittam et al., 2016) | Building façade geometry survey | Conventional: direct on-site measurement Remote: street-view and aerial-image-based measurement | Remote surveying was approximately four times faster |
| Steinmetz-Wood et al. (Steinmetz-Wood et al., 2019) | Urban streetscape audit | Conventional: field audit, 8.2 min per street segment excluding travel Remote: manual Google Street View audit, 9.8 min per street segment | Remote imagery avoided travel but did not reduce manual interpretation time |
| Roberge et al. (Roberge et al., 2022) | Neighborhood environment audit | Conventional: field audit, 59 min per neighborhood Remote: manual Google Street View audit, 57 min per neighborhood | Audit time decreased by approximately 3.4% |

study, domain-specific characteristics refer to class-discriminative material cues—such as brick bonding patterns, concrete surface roughness, stone granularity, color distribution, and surface aging—together with region-dependent visual attributes, including window arrangement, façade rhythm, exterior equipment, and building density. Generic text-to-image models may generate visually plausible building images without reliably preserving these task-relevant characteristics, thereby limiting the usefulness of the generated samples as training data for façade material classification (Sakaridis et al., 2018; Karras et al., 2019). Low-rank adaptation (LoRA), as a parameter-efficient fine-tuning method, provides a practical means of adapting a pretrained diffusion model to façade-specific visual distributions without retraining the complete model (Rombach et al., 2022).

This study addresses whether a limited set of real façade images can generate synthetic samples that preserve task-relevant material and regional characteristics, reduce the synthetic–real distribution gap, and improve downstream classification. To this end, we propose a domain-adaptive generative augmentation framework for façade material classification in data-scarce construction environments, combining LoRA-based parameter adaptation with structured prompt conditioning. The contributions are threefold. First, we establish a façade-specific augmentation workflow that integrates LoRA fine-tuning with a structured, architecture-aware prompt procedure. Second, we assess LoRA training-image volume, generation settings, and the individual contributions of LoRA fine-tuning and structured prompt conditioning through controlled baseline and component-ablation experiments. Third, we evaluate synthetic–real alignment and downstream classification on independent Korean and England datasets and verify the consistency of the augmentation trend with an additional classifier architecture; these constitute independent within-region validations rather than a cross-region transfer experiment. Accordingly, the contribution lies in the construction, formalization, and empirical evaluation of a façade-specific generative augmentation workflow, rather than in a new diffusion-model architecture or training objective.

## 2. Literature review

### 2.1. Computer vision and data augmentation for building façade analysis

Computer vision (CV) has increasingly been applied to automate the extraction of building information from image data, reducing reliance on subjective and inspector-dependent assessment procedures (Davtalab et al., 2022; Pandey and Mishra, 2025). However, when CV models are transferred from controlled datasets to real urban environments, their performance can be affected by variations in illumination, viewpoint, occlusion, and background complexity (Dong and Catbas, 2021). These challenges are particularly relevant to façade analysis, where exterior surfaces are frequently obscured by vegetation, vehicles, shadows, and surrounding urban objects.

The integration of deep-learning models with large-scale image-acquisition platforms, including unmanned aerial vehicles and Google Street View (GSV), has expanded the applicability of automated façade analysis. Wang et al. (Wang and Han, 2024) combined GSV imagery with deep-learning models to classify exterior cladding materials, and Kong et al. (Kong and Fan, 2020) used street-view images collected from Mapillary to segment façade components such as windows, walls, and doors. These studies demonstrate the feasibility of image-based façade analysis, although reliable performance remains dependent on diverse and representative training data.

The limited availability of high-quality labeled construction datasets remains a major constraint on the generalization of façade-analysis models. Manual image collection and annotation require considerable time, labor, and domain expertise, and uncontrolled field conditions make it difficult to construct consistent datasets that represent the full variability of real environments (Sami Ur Rehman et al., 2022; Laxman et al., 2023). Publicly available façade datasets are also limited, particularly for specialized tasks such as exterior cladding-material classification (Wang et al., 2024). Consequently, models trained on small or homogeneous datasets may show limited reliability under unfamiliar buildings or environmental conditions.

Conventional data augmentation has been widely adopted to mitigate data scarcity. Previous façade studies applied geometric and photometric transformations, including rotation, cropping, scaling, brightness adjustment, and weather-related modifications (Wang and Han, 2024; Wang, 2024; Wang et al., 2022). Although these methods can improve robustness, they modify existing samples without introducing fundamentally new material textures, architectural arrangements, or environmental conditions. Their effectiveness is therefore constrained when the original dataset lacks sufficient visual diversity.

These limitations have encouraged a transition toward generative approaches capable of producing new training samples beyond the visual scope of the original dataset (Goodfellow et al., 2014). Diffusion models have attracted particular attention because of their generation quality, training stability, and controllability through textual conditioning (Ho et al., 2020; Stability, 2024). In the construction domain, Ma et al. (Ma and Zheng, 2023) generated building images from textual descriptions, whereas Kim et al. (Kim and Yi, 2024) produced synthetic images of hazardous construction scenarios using a diffusion-based text-to-image framework. However, visually plausible synthetic images do not necessarily preserve the class-discriminative characteristics required for downstream recognition. Previous studies have emphasized that augmented data are useful only when the expanded training distribution adequately represents the variability of the target environment (Dunphy et al., 2022; Soans et al., 2025). Therefore, reducing the synthetic–real discrepancy remains essential for generative façade-material augmentation.

### 2.2. Parameter-efficient fine-tuning for domain-specific façade generation

Although pretrained diffusion models such as stable diffusion can generate diverse and visually plausible building images (Rombach et al.,

2022; Ma and Zheng, 2023), their generic training distributions may not sufficiently represent the fine-grained visual characteristics required for façade material classification (Park et al., 2025). In particular, material-discriminative attributes such as brick bonding patterns, concrete surface roughness, stone granularity, weathering, and region-dependent façade configurations may not be consistently preserved through zero-shot generation alone (Ma and Zheng, 2023). Adapting a pretrained diffusion model to these characteristics through conventional full fine-tuning would require updating and storing a large number of parameters, resulting in substantial computational and memory costs (Han et al., 2024; Rombach et al., 2021). These limitations motivate the use of parameter-efficient fine-tuning methods for domain-specific façade image generation.

Low-rank adaptation (LoRA), introduced by Hu et al. (2021), reduces the number of trainable parameters by inserting low-rank decomposed matrices into selected layers of a pretrained model while keeping the original model weights fixed. Rather than retraining the entire diffusion model, LoRA learns a compact set of task-specific parameters that modify the model's response to domain-relevant visual patterns. This approach reduces training time, memory consumption, and model-storage requirements while retaining the general generation capability of the pretrained model (Han et al., 2024; Hu et al., 2021). Previous studies have demonstrated that LoRA can effectively adapt generative models to specialized image domains using relatively small training datasets (Zhu et al., 2024; Chavan et al., 2023; Zhuang et al., 2024).

LoRA is particularly suitable for construction and façade-image applications because these datasets are typically limited in size, costly to label, and strongly influenced by project-, material-, and region-specific conditions (Asadi Shamsabadi et al., 2024; Choi et al., 2024; Li et al., 2025). In this study, the objective was not to reproduce a single building instance, but to adapt the diffusion model to class-level visual distributions associated with brick, concrete, and stone façades. LoRA therefore provides a practical balance between domain adaptation capability and computational efficiency. Compared with generic zero-shot generation, LoRA enables the model to learn façade-specific textures from real examples, while requiring substantially fewer trainable parameters than full model fine-tuning (Hu et al., 2021). More heavily conditioned alternatives, such as ControlNet, DreamBooth, and conditional diffusion, are discussed in Section 2.3 in terms of their input requirements, annotation effort, controllability, and suitability for limited façade datasets.

Accordingly, this study employs LoRA as a parameter-efficient mechanism for adapting stable diffusion to façade material distributions. The LoRA-based parameter adaptation is subsequently combined with structured architecture-aware prompt conditioning to generate synthetic images that reflect both material-specific textures and region-dependent architectural characteristics.

### 2.3. Positioning of domain adaptation and generative augmentation approaches

Synthetic-to-real adaptation has been studied in robotics, autonomous driving, biomedical imaging, and construction to reduce discrepancies between synthetic training data and real-world environments (Sankaranarayanan et al., 2018). In construction research, generative models have been applied to building-image generation and task-specific visual data synthesis (Wang et al., 2022; Goodfellow et al., 2014), and a transformer-based domain adaptation framework has been proposed for exterior cladding detection (Sami Ur Rehman et al., 2022). However, generative adaptation for façade material classification remains underexplored, particularly regarding the preservation of task-relevant material characteristics. Domain adaptation may be performed either within the downstream classifier or during training-data construction. Classifier-level approaches typically align feature representations between source and target domains (Ganin et al., 2016). In contrast, this study adopts a source-side approach by adapting the image-generation process before classifier training to better represent real façade textures and regional visual characteristics.

Several generative alternatives can be considered. Zero-shot text-to-image models require no task-specific training and can rapidly generate diverse architectural images (Rombach et al., 2022). However, prompt-only generations may not consistently preserve fine-grained façade characteristics such as brick bonding patterns, concrete roughness, stone granularity, weathering, and region-dependent configurations. Vision-language models can assist in identifying façade attributes and refining generation prompts, but they generally require integration with an image-generation model and may still provide limited control over fine-grained material textures without domain-specific adaptation (Varma et al., 2023). Control-based diffusion methods, including ControlNet, provide stronger spatial control through auxiliary inputs such as edge, depth, or segmentation maps (Zhang et al., 2023; Davari et al., 2025). However, these auxiliary inputs increase preprocessing and annotation requirements. DreamBooth can adapt a diffusion model using a small number of reference images (Ruiz et al., 2023), but its instance-oriented adaptation may be less suitable for class-level material diversity. Conditional diffusion methods similarly provide explicit attribute control but often require structured conditions and sufficient training samples.

### 2.4. Summary and research challenges

Although CV and generative models have shown potential for façade analysis, several challenges remain in applying synthetic data to façade material classification.

First, visually plausible building images do not necessarily preserve class-discriminative characteristics such as brick bonding patterns, concrete surface roughness, stone granularity, weathering, and region-dependent architectural configurations. Therefore, generative augmentation should reduce the synthetic–real distribution gap while retaining features relevant to downstream classification.

Second, prompt-based generation often depends on manually selected keywords, which limits reproducibility and makes it difficult to distinguish systematic domain conditioning from heuristic prompt tuning. A structured prompt-construction procedure based on material, architectural, environmental, and surface-condition descriptors is therefore required. In addition, the respective contributions of LoRA-based parameter adaptation and architecture-aware prompt conditioning should be isolated through component-wise ablation. Comparisons with no augmentation, conventional augmentation, and standard diffusion are also necessary to determine whether performance improvements arise from augmentation in general or from the proposed workflow.

Third, downstream classification performance alone is insufficient to characterize synthetic–real alignment. Generated images should be evaluated using distributional metrics, feature-space analysis, dimensionality-reduction visualization, and downstream classification performance. The influence of LoRA training-image volume also requires investigation under limited-data conditions. To address these challenges, this study evaluates LoRA training-image volume, generation hyperparameters, structured prompt conditioning, synthetic–real alignment, and classification performance within a unified framework. The workflow is independently evaluated using Korean and England datasets, and an additional classifier architecture is employed to examine whether the observed augmentation effect is specific to MobileNetV3.

## 3. Research scope and method

This study selects the England and Korea as representative cases to examine contrasting urban façade characteristics shaped by different development trajectories. The England exemplifies a context where

historical preservation, regulatory continuity, and material tradition have strongly influenced façade composition. In contrast, Korea represents a rapidly urbanized environment characterized by high-density development, frequent redevelopment, and the integration of advanced building technologies. By contrasting these two contexts, the study aims to capture a broad spectrum of façade typologies and operational conditions, thereby enhancing the generalizability of the analysis.

The objective of this study is to augment training data for building façade classification by applying LoRA-based stable diffusion. The research workflow was designed in four stages, as illustrated in Fig. 1. The methodology included constructing a real-world façade image dataset, training a LoRA-based stable diffusion model, generating synthetic façade images, and training and evaluating the classification model.

### 3.1. Real façade dataset construction

This study used building façade images collected using the GSV application programming interface (API) to construct the real-image datasets used for LoRA fine-tuning and façade classification. Mapo-gu, Seoul, Korea, was selected as the target area because it includes a combination of mature and newly developed urban districts and provides representative visual characteristics of Korean urban environments (Mo and Kang, 2022). A second façade dataset was collected from England, to evaluate the proposed framework in an independent regional setting. The Korean and England datasets were processed, trained, and evaluated independently; therefore, the experiment was not designed as a cross-region transfer test. The latitude and longitude coordinates of buildings within Mapo-gu were extracted using the Quantum Geographic Information System, and the images corresponding to these coordinates were retrieved through API requests. The camera parameters were set as follows: field of view at 20 to control the zoom level of the scene, pitch at 5 to define the camera's vertical angle relative to the Street View vehicle and heading at 90 to specify the horizontal viewing direction. All façade images were saved at a resolution of 600 × 600 pixels.

The collected images manually screened using predefined inclusion and exclusion criteria. As shown in Fig. 2, images were excluded if no buildings were present, if the façade was heavily occluded by objects such as trees or vehicles, if the resolution was too low to identify exterior materials, or if indoor scenes were captured. Images containing retrieval errors or façades whose primary cladding material could not be identified with sufficient confidence were also excluded. Images in which façade materials were clearly visible were classified as useable. The selected images were labeled according to façade material type. According to Oh et al. (2021), wet construction methods using brick and concrete and dry construction methods using stone are predominantly employed in Korea. Based on this classification, the primary categories used in this study were defined as brick, concrete, and stone. Labeling was performed according to visual characteristics such as color, texture, and pattern, and all images were assigned to one of the predefined classes. Following the criteria presented by Wang et al. (Wang and Han, 2024), the images were categorized as brick, concrete, or stone. Mixed-material façades were excluded/labeled according to the dominant visible material using a predefined criterion. Additionally, façade images collected from selected urban areas in the England were processed using the same screening, labeling, and categorization procedures to maintain dataset consistency. The finalized dataset was subsequently used for LoRA-based stable diffusion training and for training and validating the façade classification model. Following the screening process, 500 useable images were retained for each façade material class—brick, concrete, and stone—in each regional dataset. Consequently, each regional dataset contained 1500 façade images, resulting in a total of 3000 images across Korea and England.

### 3.2. Low-rank adaptation-based diffusion model fine-tuning

The stable diffusion model was fine-tuned using real façade images through LoRA. The base model was sd-v1-5-pruned-noema-fp16.

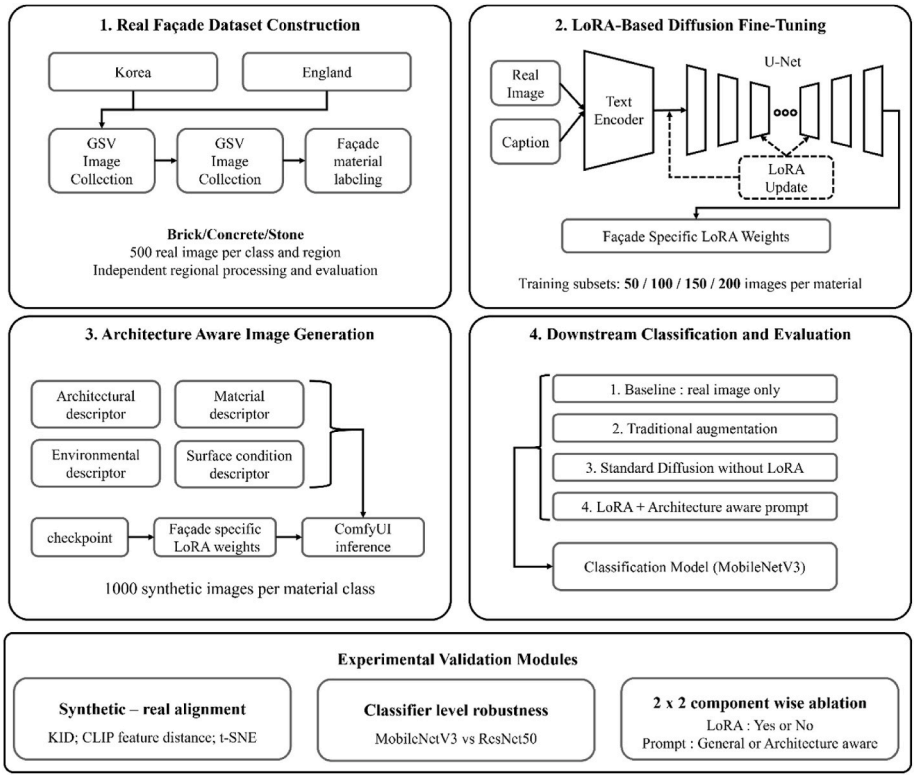



**Fig. 1.** Flowchart of LoRA-based Stable Diffusion and classification performance evaluation.

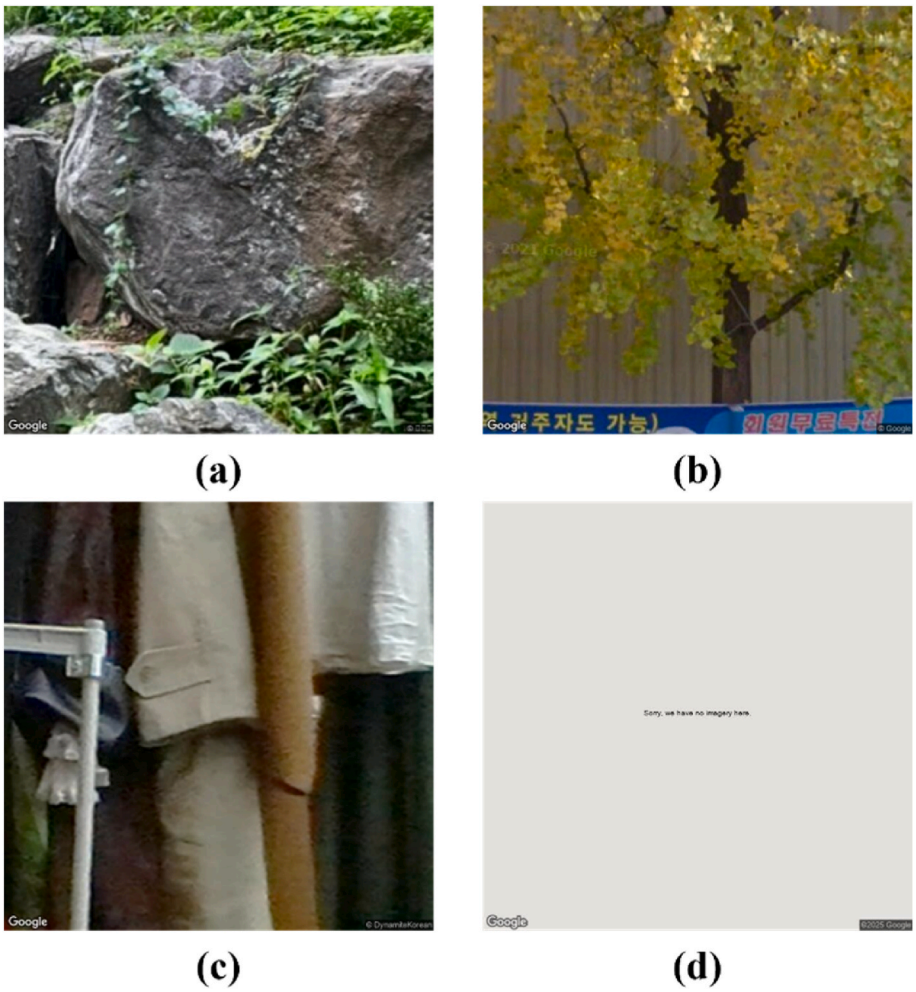



**Fig. 2.** Selected examples of unusable images; (a) image without buildings, (b) occluded image, (c) indoor image, (d) error image.

safetensors, and separate LoRA models were trained for each combination of regional dataset, façade material class, and training-image volume. The façade material classes consisted of brick, concrete, and stone, and each training image was paired with the caption ‘a building with a [cladding type] façade'. Only real images assigned to the training partition were used for LoRA fine-tuning.

LoRA training was implemented using the Kohya_ss toolkit (hollowstrawberr. kohya-colab) with the networks. lora module. The LoRA adapters were trained for both the U-Net and text encoder components using separate learning rates. The U-Net learning rate was set to 0.00005, and the text encoder learning rate was set to 0.0004. The LoRA network rank and alpha were set to 16 and 8, respectively. Training was conducted for 10 epochs with 10 repeats per image and a batch size of 4. The AdamW8bit optimizer and a cosine-with-restarts learning-rate scheduler with three cycles were utilized. The random seed was fixed at 42, and mixed-precision training was performed using FP16. All models were trained in a Google Colab environment using an NVIDIA A100 graphics processing unit.

To examine the sensitivity of LoRA adaptation to the number of available real images, separate models were trained using 50, 100, 150, and 200 images per façade material class. This range was selected based on previous research indicating that domain-relevant characteristics can be learned from approximately 100–200 training images under limited-data conditions (Zhuang et al., 2024). The four image-volume conditions were evaluated using the same training configuration to isolate the effect of training-data volume. Separate LoRA models were trained for the Korean and England datasets; therefore, the experimental design produced 24 models in total: two regions, three material classes, and four training-image-volume conditions.

### 3.3. Architecture-aware synthetic image generation

Synthetic façade images were generated by applying the fine-tuned LoRA weights to a stable diffusion model through a structured text-prompt system. The generation prompt was distinguished from the simple caption used during LoRA fine-tuning: training images were assigned the class-level caption ‘a building with a [cladding type] façade', and the generation prompt added descriptors controlling material appearance, regional architectural characteristics, environmental context, and surface condition. Accordingly, the architecture-aware prompt comprised four components—material, architectural, environmental, and surface-condition descriptors—following the structure ‘a building with a [color] [cladding type] façade, [regional architectural descriptor], [environmental descriptor], [surface condition]'. The material descriptor specified façade class and color; the architectural descriptor captured observable regional façade characteristics; the environmental descriptor described surrounding urban context and weather; and the surface-condition descriptor represented variations such as clean, aged, or weathered surfaces. These were derived from the real façade datasets and from previous studies on regional building and façade attributes (Roumpani et al., 2018; Mingi and Jungmin, 2025).

For the Korean dataset, the architectural descriptors included irregular window arrangements, small windows with security bars, aluminum window frames, exposed exterior equipment and pipes, dense

alley-scale environments, and closely adjacent buildings. For the England dataset, they included repetitive or symmetric window arrangements, bay windows, white window sashes, regular façade rhythms, and limited exterior equipment. The same descriptor categories and construction procedure were applied across all material classes for consistency and reproducibility, and negative prompts ('blurry, low-resolution, deformed, cartoon, fantasy, distorted, text, and watermark') were used to reduce common artifacts. For the component-ablation experiment, two prompt conditions were defined: a general-prompt condition using only 'a building with a [cladding type] façade', and an architecture-aware condition incorporating the four descriptor groups, enabling the contribution of structured prompt conditioning to be evaluated separately from that of LoRA fine-tuning.

Synthetic images were generated in ComfyUI using the Realistic Vision V6.0 B1 checkpoint (Kim and Yi, 2024). Because both the checkpoint and the trained LoRA weights were based on the stable diffusion 1.5 architecture, the LoRA adapters could be applied at inference. LoRA model strength and text-encoder strength were both set to 1.0, and images were generated at 600 × 600 pixels using the Euler sampler and normal scheduler. Randomized seeds promoted visual diversity, and all other settings were held constant across conditions. A predefined grid search was then conducted over the classifier-free guidance (CFG) scale, which controls adherence to the text prompt, and the number of sampling steps, which determines the denoising iterations and the quality–cost trade-off (Ji-Woo et al., 2024). CFG scales of 7, 8, and 9 and sampling steps of 20, 30, and 40 were evaluated, a range selected from preliminary tests and commonly used stable diffusion settings (Kim and Yi, 2024; Davari et al., 2025). If the CFG scale was too high, the image distortion became too severe, as shown in Fig. 3; therefore, a range between 7 and 9 was used. With the sampler, resolution, LoRA strength, and scheduler fixed, the search was intended to identify an appropriate setting within this range rather than to optimize all parameters exhaustively. The selected configuration was then applied consistently across all regional and material conditions, with each LoRA model generating 1000 synthetic images for its façade material class to construct the augmented datasets for the downstream classification and ablation experiments.

### 3.4. Downstream classification and performance evaluation

Synthetic façade images generated with the selected CFG-scale and sampling-step configuration were combined with real images to construct the augmented training datasets. Each training set comprised real images plus 1000 synthetic images, whereas the validation and test sets each contained 100 real façade images only and were kept identical across all experimental conditions (Bay et al., 2025). This design allowed every augmentation strategy to be evaluated under the same real-image validation and test conditions. MobileNetV3 was adopted as the primary backbone for its computational efficiency and suitability for image classification (Wang and Han, 2024), and ResNet-50 was additionally employed to test whether the augmentation effect was dependent on a specific classifier architecture. Both architectures used the same training, validation, and test partitions, and MobileNetV3 training dynamics were monitored over 200 epochs by recording the F1-score at each epoch.

Classification performance was evaluated using accuracy and F1-score, with class-wise prediction patterns examined through confusion matrices. Following hyperparameter tuning, the influence of LoRA training-data volume was assessed using LoRA models trained on 50, 100, 150, and 200 real images per material class to construct separate synthetic datasets; the corresponding classifiers used the same real-image partitions and an identical number of generated images so that this factor could be evaluated independently.

To separate the effect of augmentation in general from that of the proposed workflow, four baseline conditions were evaluated: (i) no augmentation (real images only); (ii) conventional augmentation using contrast adjustment, rotation, shear transformation, and brightness adjustment (Wang and Han, 2024); (iii) standard diffusion without LoRA fine-tuning; and (iv) the proposed LoRA-based diffusion with architecture-aware prompt conditioning. Conditions (ii)–(iv) each provided the same number of augmented or generated images. A component-wise ablation further crossed two LoRA conditions with two prompt conditions: standard diffusion or LoRA-based diffusion, each paired with a general or an architecture-aware prompt. Primary comparisons were conducted with MobileNetV3, and the principal conditions were re-evaluated with ResNet-50 for classifier-level robustness. Finally, synthetic-real distribution similarity was evaluated using the kernel inception distance, feature-space distance and cosine similarity, and t-distributed stochastic neighbor embedding (t-SNE) visualization, as detailed in the following evaluation section.

## 4. Implementation and experimental results

### 4.1. Quality and feature-space evaluation of generated images

The LoRA-based stable diffusion model generated synthetic façade images that visually reflected the material-specific and regional characteristics of the real datasets. Figs. 4 and 5 compare real and synthetic façade images from Korea and England, respectively. The generated images reproduced recognizable differences in brick bonding patterns, concrete surface roughness, stone granularity, color variation, and surface weathering. Regional characteristics, including façade composition and window arrangement, were also observed in the synthetic samples. These qualitative results suggest that LoRA fine-tuning improved the representation of façade-specific visual characteristics while maintaining variations in lighting, surface condition, and surrounding context.




**Fig. 3.** Representative façade images generated at excessively high CFG scales; (a) CFG 15, (b) CFG 17, and (c) CFG 20.

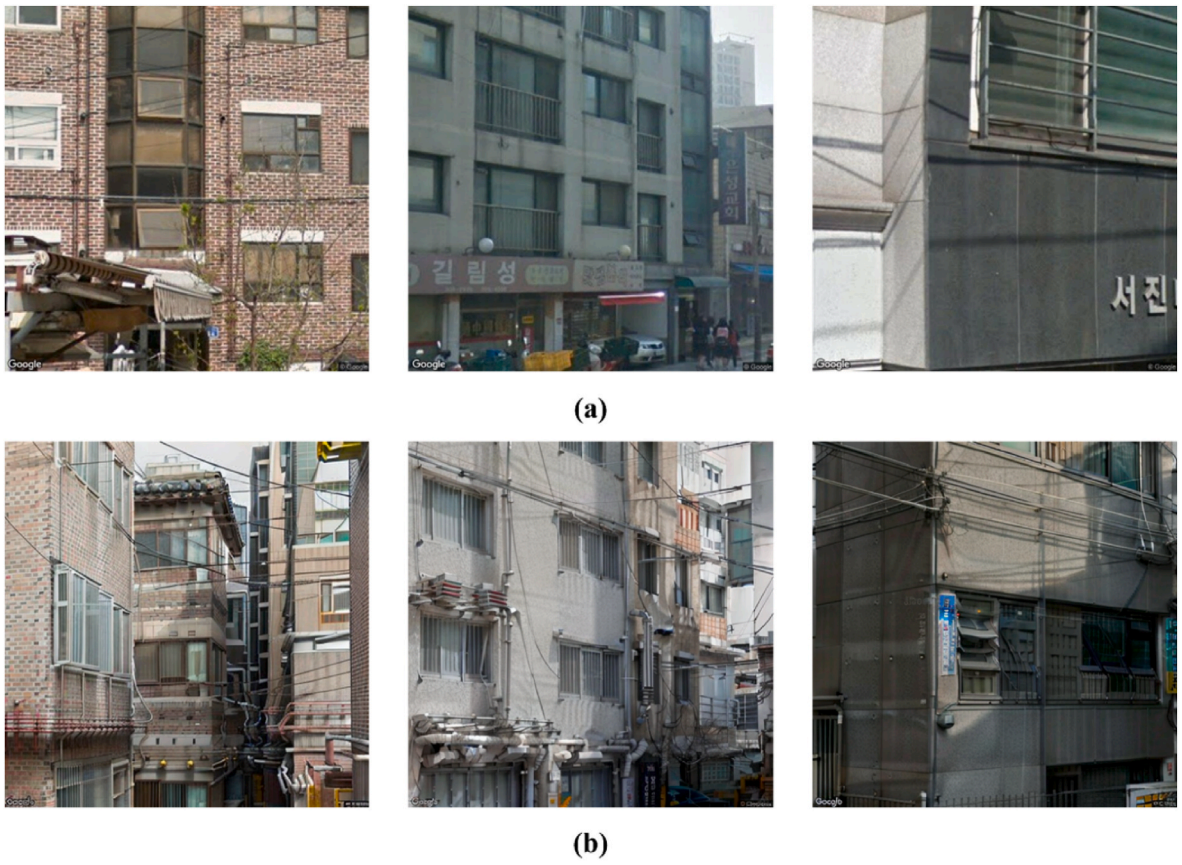



**Fig. 4.** Qualitative comparison of real and Stable-Diffusion-based augmented façade images in Korea; (a) real image data, (b) augmented image data (from left, brick, concrete, and stone).

Because visual inspection alone cannot fully determine synthetic–real similarity, the qualitative comparison was complemented by the KID, embedding-based feature-space analysis, and t-SNE visualization. The KID was used to evaluate class-wise distributional alignment, while feature-space distance and cosine similarity quantified the proximity of synthetic and real image representations. The KID was calculated from 2048-dimensional inception features using 100 subsets with a maximum subset size of 50, with lower values indicating closer agreement between the real and synthetic feature distributions. The feature-space distance was calculated as the Euclidean distance between the class-wise mean contrastive language-image pretraining (CLIP) embeddings of the real and synthetic images.

Table 2 presents the class-wise KID values obtained by comparing real façade images with synthetic images generated using standard diffusion and the proposed LoRA-based diffusion approach. Lower KID values indicate closer alignment between the feature distributions of real and synthetic images. The proposed approach achieved lower KID values across all three material classes. For brick façades, the KID decreased from 0.129 to 0.083, corresponding to a reduction of 35.77%. For concrete façades, the KID decreased from 0.145 to 0.074, representing the largest reduction of 48.71%. For stone façades, the KID decreased from 0.116 to 0.079, corresponding to a reduction of 31.44%. These consistent reductions indicate that LoRA-based adaptation improved the distributional similarity between synthetic and real façade images for all material classes, with the greatest improvement observed for concrete façades. Nevertheless, the KID measures feature-distribution alignment rather than perceptual realism alone; therefore, these results are interpreted together with the feature-space and downstream classification analyses.

Table 3 compares the class-wise feature-space distances between real façade images and synthetic images generated using standard diffusion and LoRA-based diffusion. Lower distance values indicate greater similarity between the feature representations of real and synthetic images. The LoRA-based diffusion reduced the feature-space distance from 0.782 to 0.482 for brick, from 0.769 to 0.490 for concrete, and from 0.766 to 0.491 for stone, corresponding to reductions of 38.34%, 36.30% and 35.74%, respectively. The consistent reductions across all three material classes indicate that LoRA-based adaptation generated synthetic images whose feature representations were closer to those of the real façade images than those generated using standard diffusion. These results support the KID findings and are further illustrated through the t-SNE visualization of the real and synthetic image distributions.

Fig. 6 visualizes the feature distributions of real, standard diffusion, and LoRA-generated façade images using t-SNE. For each material class and image domain, 100 images were randomly sampled, and penultimate-layer features were extracted using an ImageNet-pretrained MobileNetV3-Large model without task-specific fine-tuning. The extracted features were first reduced to a maximum of 50 dimensions using principal component analysis and then projected into a two-dimensional space using t-SNE. The visualization shows that the generated images formed class-dependent clusters, indicating that both diffusion approaches captured distinguishable material characteristics. However, the synthetic clusters remained partially separated from the real-image samples, suggesting that the synthetic–real feature gap was reduced but not eliminated. Because t-SNE primarily preserves local neighborhood relationships and its axis distances do not provide a direct quantitative measure of distributional similarity, the visualization was interpreted together with the KID and CLIP embedding-distance results. These complementary analyses provide quantitative evidence for whether the LoRA-generated samples were more closely aligned with the real façade-image distribution than those generated using standard diffusion.

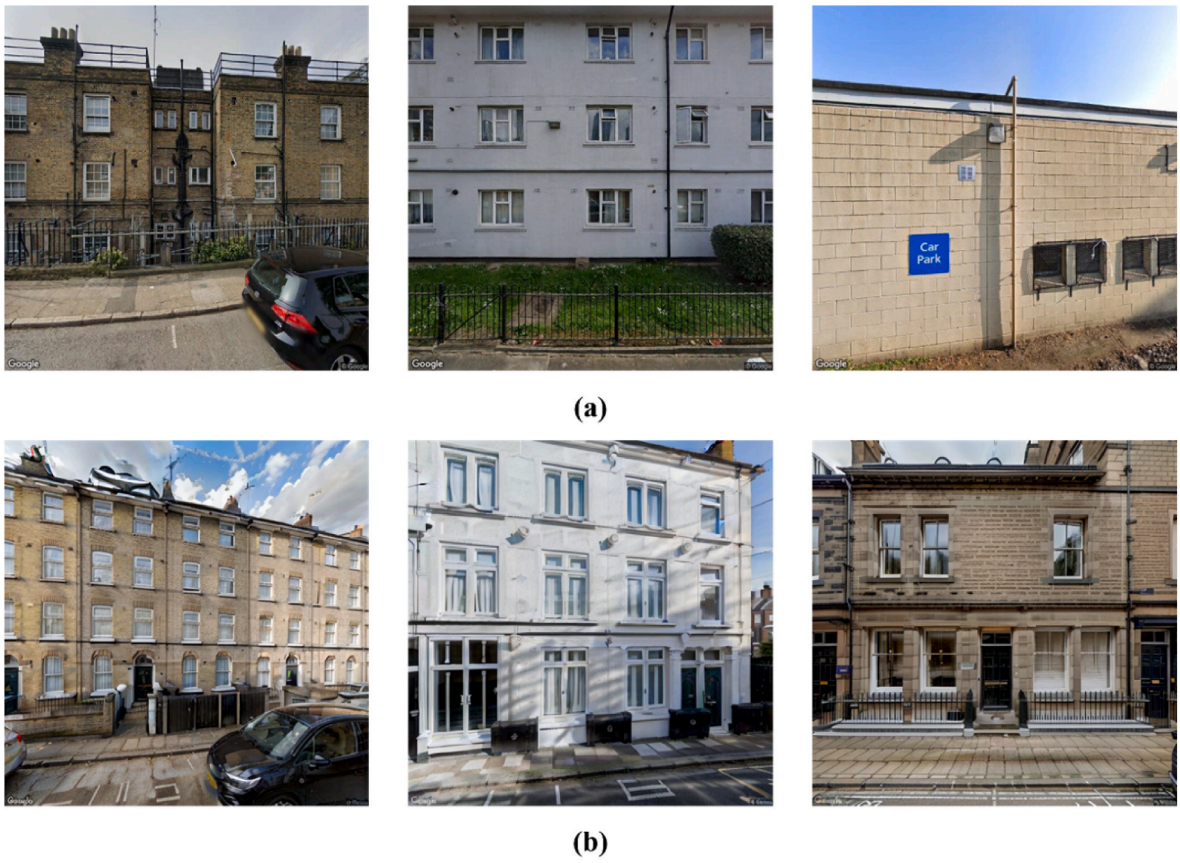



**Fig. 5.** Qualitative comparison of real and Stable-Diffusion-based augmented façade images in England; (a) real image data, (b) augmented image data (from left, brick, concrete, and stone).

**Table 2**
Comparison of KID between real and synthetic façade images generated by Standard Diffusion and LoRA-based Diffusion.

| Class | Standard Diffusion (No LoRA) KID ↓ | Proposed LoRA-Based Diffusion KID ↓ | Reduction |
|---|---|---|---|
| Brick | 0.129 | 0.083 | 35.77% |
| Concrete | 0.145 | 0.074 | 48.71% |
| Stone | 0.116 | 0.079 | 31.44% |

**Table 3**
Comparison of feature-space distances between real and synthetic façade images generated by Standard Diffusion and LoRA-based Diffusion.

| Class | Real-Standard Diffusion Distance ↓ | Real – LoRA Based Diffusion Distance ↓ | Reduction |
|---|---|---|---|
| Brick | 0.782 | 0.482 | 38.34% |
| Concrete | 0.769 | 0.490 | 36.30% |
| Stone | 0.766 | 0.491 | 35.74% |

### *4.2. Determination of training and generation settings*

Fig. 7 presents the F1-score recorded over 200 training epochs. The model showed a rapid performance increase during the initial training stage and reached a stable range of approximately 0.90–0.91 after the first 20 epochs. Although the highest individual F1-score was observed before epoch 50, subsequent values fluctuated within a narrow range without a consistent upward trend. The mean F1-score was 0.9047 during epochs 21–50 and 0.9066 during epochs 51–200, indicating only a marginal difference of approximately 0.0019. In addition, the learning rate had decreased from 0.001 to approximately $6.1 \times 10^{-8}$ by epoch 48, substantially limiting further parameter updates. Therefore, 50 epochs was selected as a conservative convergence-based training setting that provided stable performance while avoiding unnecessary computational cost, rather than as the epoch producing the single highest F1-score.

Table 4 presents the classification performance comparison across different synthetic datasets generated during the hyperparameter tuning process (CFG: 7_ Steps: 20 – CFG: 9_ Steps: 40), evaluated using Accuracy and F1-score. Classification performance varied depending on the combination of sampling steps and CFG scale applied during synthetic image generation. Among the tested configurations, the 9_20 setting achieved the highest performance, recording an Accuracy of 83.67% and an F1-score of 83.42%. In contrast, the 8_30 configuration yielded the lowest performance, implying that suboptimal hyperparameter combinations may introduce visual artifacts or insufficient texture fidelity, which negatively affect feature learning. Most configurations achieved performance values within a relatively narrow range of approximately 79%–81%. Accordingly, the 9_20 configuration was selected as the optimal hyperparameter setting, and all subsequent synthetic image generation processes were conducted using this configuration to ensure consistent data quality for classification model training.

Based on the optimal hyperparameter configuration identified in the previous stage, this experiment further analyzed the influence of LoRA training data volume on classification performance using the Korean façade dataset. The results presented in Fig. 8 show that classification performance generally improved as the number of LoRA training images increased. When the number of training images increased from 50 to 150, Accuracy improved from 91.00% to 93.67%, and the F1-score increased from 91.02% to 93.68%, representing the highest performance observed in the Korean dataset. This trend indicates that expanding the LoRA training dataset enabled the generative model to capture façade-specific visual characteristics, such as material texture,

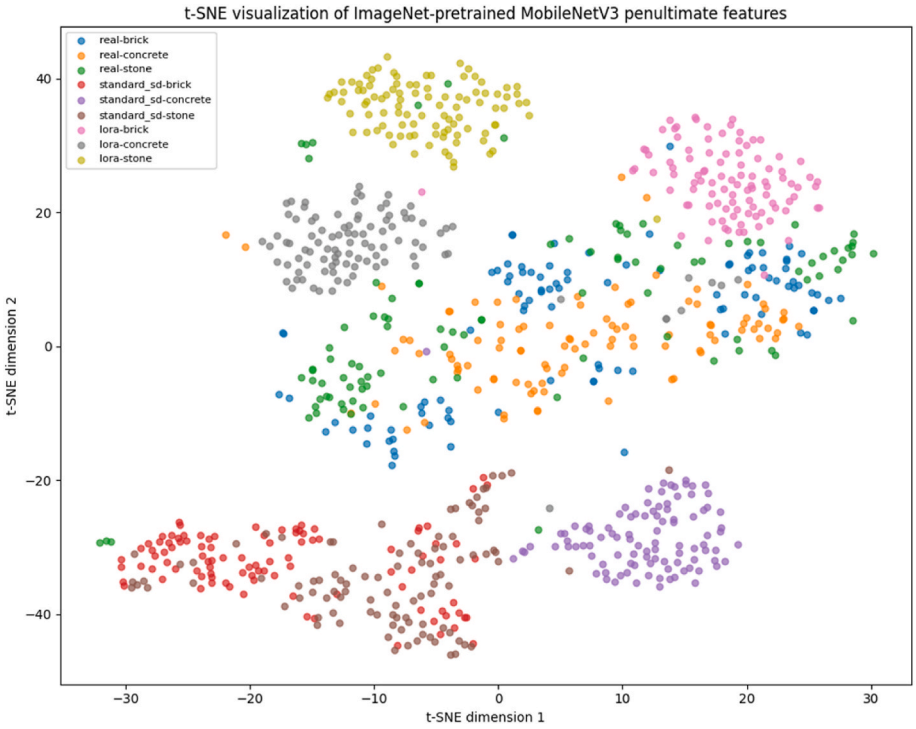



**Fig. 6.** t-SNE visualization of real and synthetic façade images using MobileNetV3 backbone features.

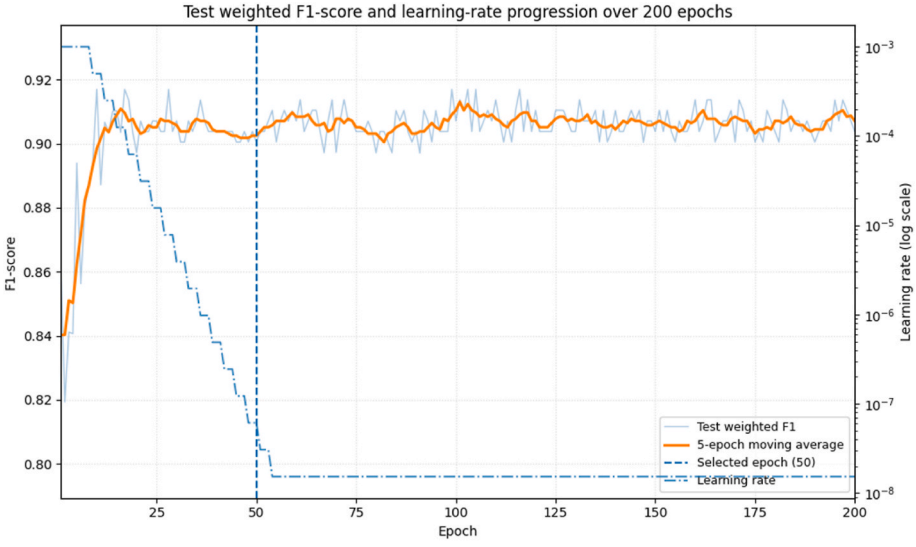



**Fig. 7.** F1-score and learning-rate progression over 200 training epochs.

**Table 4**
Accuracy & F1-score by CFG scale & sampling steps training images.

| CFG Scale | Sampling Steps | Accuracy (%) | F1-score (%) |
|---|---|---|---|
| 7 | 20 | 81.00 | 80.93 |
| 7 | 30 | 79.67 | 79.24 |
| 7 | 40 | 80.67 | 80.17 |
| 8 | 20 | 80.00 | 79.42 |
| 8 | 30 | 77.33 | 76.59 |
| 8 | 40 | 81.00 | 80.93 |
| **9** | **20** | **83.67** | **83.42** |
| 9 | 30 | 79.00 | 78.51 |
| 9 | 40 | 81.00 | 80.93 |

color distribution, and surface irregularities, with greater precision. Consequently, the classification model benefited from higher-quality synthetic images and achieved improved generalization performance.

However, when the number of training images was further increased to 200, both Accuracy and F1-score decreased to 90.00% and 90.10%, respectively. This decline suggests that excessive training data may reduce learning efficiency and introduce redundant or visually repetitive synthetic patterns, which can negatively affect feature discrimination. The results imply that there is an optimal range of LoRA training data volume, beyond which performance gains saturate or decline. Although the overall trend demonstrates improved performance up to 150 images, further class-wise analysis using confusion matrices was conducted to examine how the model distinguished between brick, concrete, and stone materials under different training data volumes.

Fig. 9 presents the class-wise prediction results for different LoRA

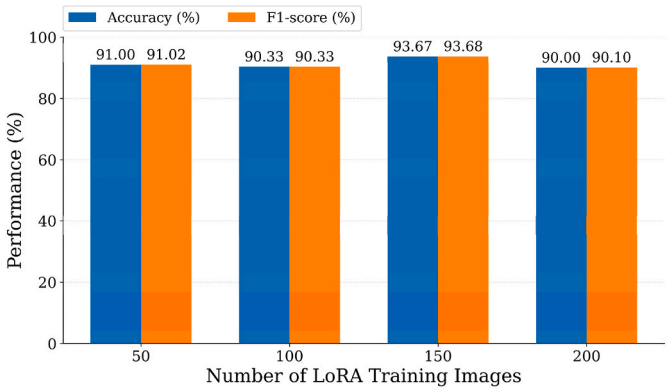



**Fig. 8.** Accuracy & F1-score by number of LoRA training images in Korea dataset.

training data volumes in the Korean dataset and provides a detailed examination of the model's generalization behavior across the three façade categories. The brick class demonstrated consistently strong performance across all experimental conditions. With 50 training images, 89% of brick samples were correctly classified, and this improved to 95% and 96% with 100 and 150 images, respectively. Even with 200 images, the accuracy remained high at 94%, indicating stable recognition performance. This robustness can be attributed to the distinct texture and repetitive structural patterns characteristic of brick façades. The concrete class showed relatively high but more fluctuating performance. The classification accuracy was 96% with 50 images, slightly decreased to 92% with 100 images, improved to 93% at 150 images, and dropped to 88% when 200 images were used. Misclassifications primarily occurred between concrete and stone, suggesting that their similar, gray-toned color distributions and surface textures posed challenges for feature discrimination. The stone class exhibited comparatively lower recognition accuracy due to its complex and irregular surface characteristics. The correct classification rates were 88%, 84%, 92%, and 88% for 50, 100, 150, and 200 images, respectively. Notably, the 150-image configuration achieved the highest stone accuracy (92%) while maintaining strong performance in the other two classes.

Overall, the model trained with 150 images per class achieved the most balanced performance across brick (96%), concrete (93%), and stone (92%) categories, with reduced misclassification rates. This trend is consistent with the F1-score results and confirms that 150 LoRA training images provide an optimal balance between data sufficiency and generalization stability in the Korean façade dataset.

Fig. 10 presents the results for the England dataset, showing that classification performance was sensitive to the number of LoRA training images. When 50 training images were used, the model showed moderate performance, with both evaluation metrics remaining in the mid-70% range. Increasing the number of LoRA training images to 100 led to a clear improvement, with Accuracy rising to 83.67% and F1-score to 82.91%. This result suggests that expanding the LoRA training dataset

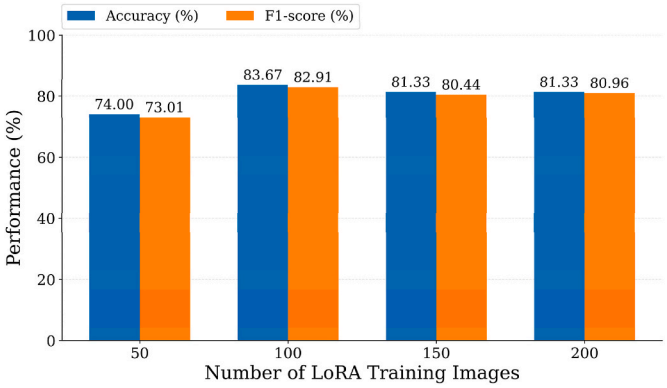



**Fig. 10.** Accuracy & F1-score by number of LoRA training images in England dataset.

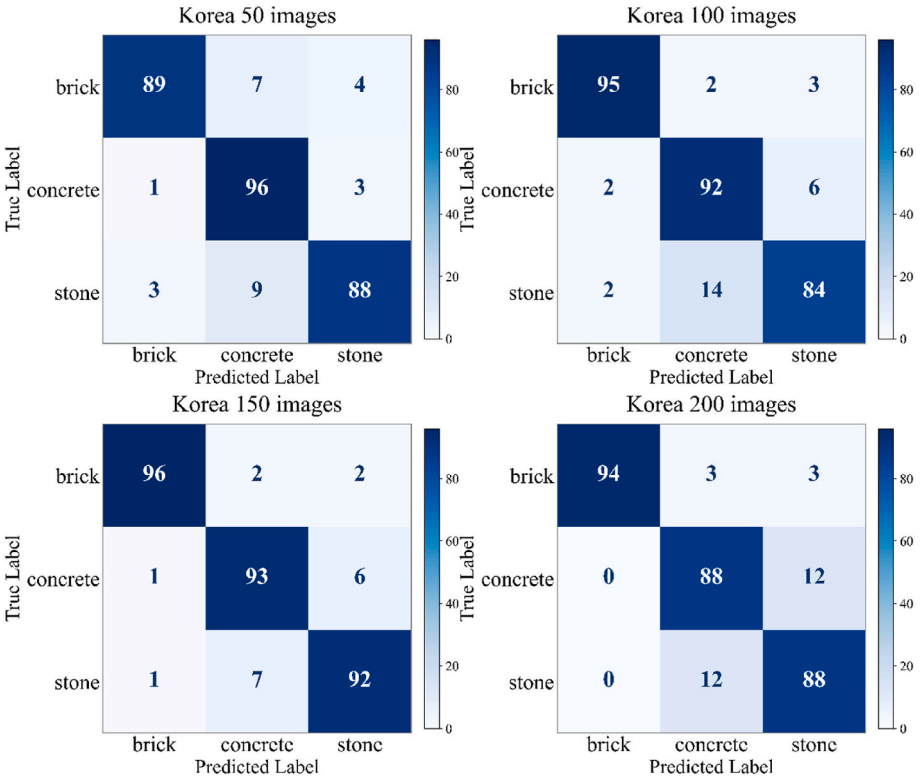



**Fig. 9.** Comparison of confusion matrices by training image quantity in Korea dataset.

enabled the generative model to better capture region-specific façade characteristics, such as masonry regularity, window alignment, and surface texture patterns commonly observed in English urban environments.

However, further increasing the number of training images to 150 did not yield additional improvement, and a slight decrease in performance was observed. When 200 images were used, performance remained similar to the 150-image setting, indicating that the performance gains saturated beyond 100 training images. This pattern suggests that once the generative model sufficiently learned the dominant material and architectural features of the dataset, additional training data provided limited new discriminative information. Overall, the configuration using 100 LoRA training images achieved the highest performance and provided the most efficient balance between data volume and classification stability in the England façade dataset. To further investigate class-wise prediction tendencies and misclassification characteristics, confusion matrix analysis was subsequently conducted.

Fig. 11 presents the class-wise prediction results for different LoRA training data volumes in the England dataset and provides a detailed examination of the model's generalization behavior across the three façade categories. The brick class demonstrated relatively stable performance across all experimental conditions. With 50 training images, brick recognition reached 84%, which improved substantially to 95% when 100 images were used. When the number of training images increased to 150 and 200, performance slightly decreased to 89% and 87%, respectively. This trend suggests that approximately 100 LoRA training images were sufficient to capture the regular masonry patterns and structural repetition characteristic of English brick façades. The concrete class exhibited strong and progressively improving performance. The recognition rate was 88% with 50 images and increased to 95% at 100 images. The highest performance was observed with 150 training images, where concrete recognition reached 98%. Although a slight decrease to 93% was observed at 200 images, concrete remained the most consistently recognized façade type. Misclassifications in this category primarily occurred between concrete and stone, likely due to similarities in surface tone and texture.

In contrast, the stone class showed comparatively lower recognition accuracy. With 50 training images, stone recognition was 50%, indicating substantial confusion with brick. Increasing the training volume to 100 images improved stone classification to 61%, while 150 and 200 images resulted in recognition rates of 57% and 64%, respectively. Despite gradual improvement, confusion between stone and brick persisted, reflecting the heterogeneous and irregular texture characteristics of stone façades. Overall, the configuration using 100 training images achieved strong and balanced performance across brick (95%) and concrete (95%), while the 150-image configuration maximized concrete recognition (98%) and maintained stable performance in the other classes. These results indicate that approximately 100–150 LoRA training images provide an effective balance between data sufficiency and generalization stability in the England façade dataset.

### 4.3. Comparison of augmentation strategies and baselines

To demonstrate the effectiveness of the proposed method, both quantitative and qualitative evaluations were conducted by comparing the generative augmentation approach with conventional augmentation techniques in the Korean dataset. The generative augmentation model trained with 150 LoRA images achieved the highest generalization performance. In the quantitative evaluation, the LoRA-based fine-tuned generative augmentation method significantly outperformed the traditional augmentation approach. The conventional augmentation method achieved an overall accuracy of 85% and an F1-score of 84.82%. In contrast, the generative augmentation method achieved an accuracy of 93.67% and an F1-score of 93.68%, representing a substantial improvement in both evaluation metrics.

As shown in Fig. 12, the largest performance gains were observed in the concrete and stone classes. In the traditional augmentation setting, notable misclassification occurred between concrete and stone, particularly with stone samples being incorrectly predicted as concrete. By

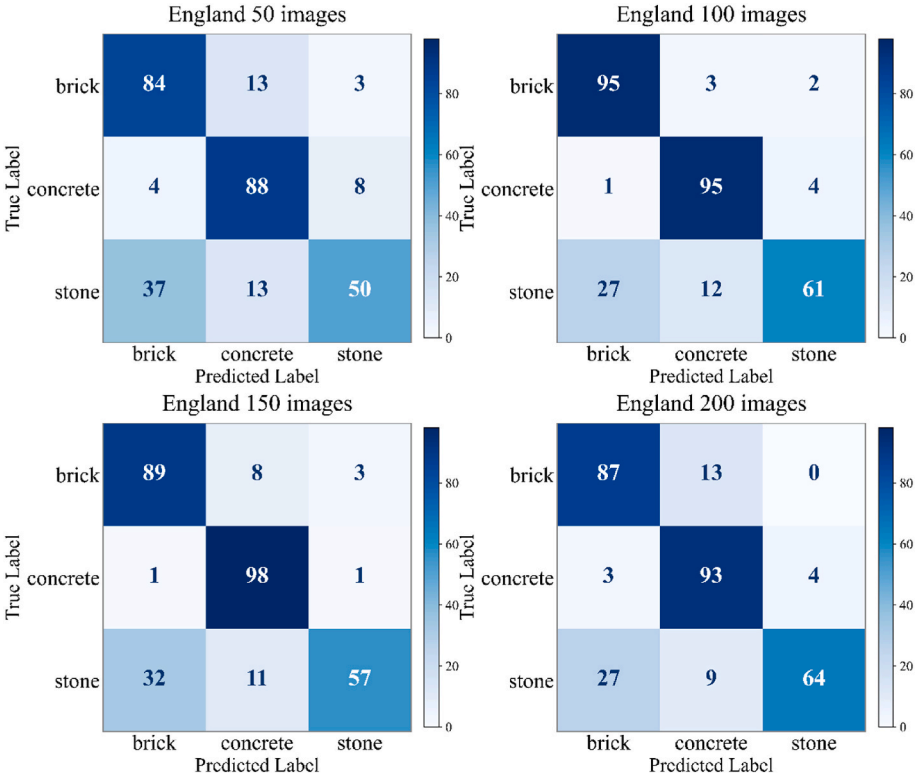



**Fig. 11.** Comparison of confusion matrices by training image quantity in England dataset.

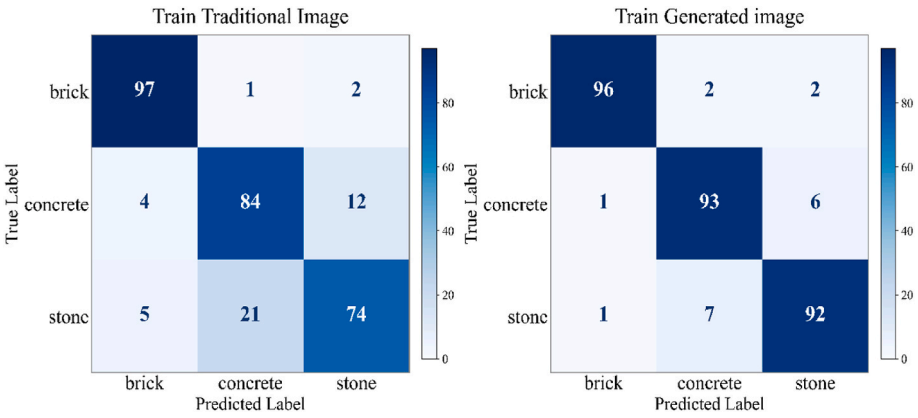



**Fig. 12.** Confusion matrices comparing traditional and generative augmentation methods in Korea dataset.

contrast, the generative augmentation approach substantially reduced these confusions, especially in the stone class, where correct predictions increased markedly. This improvement indicates that the LoRA-based synthetic images more accurately captured fine-grained material textures and surface characteristics, thereby strengthening the model's discriminative capability. Although brick recognition remained consistently high in both settings, the performance difference for this class was relatively small compared to concrete and stone. These findings highlight the limitations of conventional geometric and photometric transformations, which primarily expand data quantity without sufficiently enriching material-level texture diversity. In contrast, the generative augmentation approach better reflects real data distributions, leading to improved generalization and classification stability.

To evaluate the effectiveness of the proposed generative augmentation method, we compared it with conventional augmentation techniques using the England façade dataset. Based on the previous analysis of training data volume, the 100-image LoRA training configuration demonstrated the most stable and balanced classification performance in the England dataset. Therefore, this configuration was selected for comparison with the traditional augmentation approach. In the quantitative evaluation, the traditional augmentation method achieved an overall accuracy of 71.67% and an F1-score of 71.08%. In contrast, the generative augmentation method achieved an accuracy of 83.67% and an F1-score of 82.91%, indicating a substantial improvement in overall classification performance.

As shown in Fig. 13, the most significant performance improvement was observed in the concrete class. Under the traditional augmentation setting, a large portion of concrete samples were misclassified as stone, resulting in severe confusion between the two materials. The generative augmentation approach considerably reduced this misclassification and substantially improved concrete recognition. This suggests that the synthetic images generated by the LoRA-based stable diffusion model more effectively captured the subtle surface patterns and tonal differences distinguishing concrete from stone in English façades. However, a different trend was observed in the stone class. The traditional augmentation method achieved a higher correct prediction rate for stone compared to the generative augmentation approach. While generative augmentation improved overall accuracy and significantly enhanced concrete recognition, stone classification performance slightly decreased due to increased confusion with other façade categories. This indicates that the generative model introduced additional texture variations that partially overlapped with features of brick or concrete in the England dataset. Overall, the comparison demonstrates that generative augmentation substantially improved global classification performance, primarily driven by enhanced discrimination of concrete materials, although its impact varied across material types.

Fig. 14(a) presents the images produced through conventional augmentation, in which only simple visual changes, such as brightness adjustment, rotation, and contrast modification were applied, resulting in images that lacked a detailed surface texture and fine material granularity. By contrast, Fig. 14(b) shows the results of the generative augmentation method developed in this study, produced using the LoRA-based stable diffusion model. The synthetic images in Fig. 14(b) realistically reproduced the subtle surface roughness, shading, and reflectance characteristics of the actual façade materials. This difference indicates that the generative model learns visual patterns at the material level rather than relying on simple geometric transformations, thereby enabling more effective representation of texture-based attributes. Consequently, the performance improvements observed in the stone and concrete classes can be attributed to generative augmentation, which

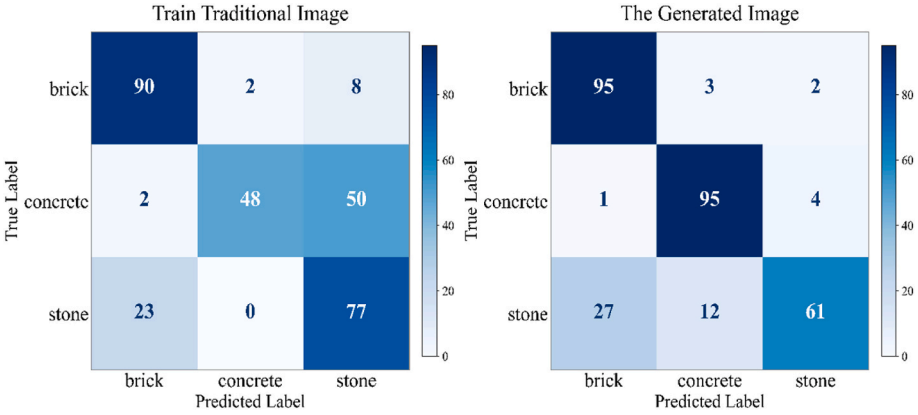



**Fig. 13.** Confusion matrices comparing traditional and generative augmentation methods in England dataset.




**Fig. 14.** Visual comparison of traditional and generative augmented images; (a) traditional augmented image, (b) generative augmented image.

provides richer texture-related visual information.

The comparative results in Table 5 present a unified evaluation of three data augmentation strategies—traditional augmentation, standard diffusion without LoRA, and the proposed LoRA-based stable diffusion method—across both the England and Korean façade datasets. The results reveal consistent performance trends in both regional contexts, while also highlighting differences in texture fidelity, data diversity, computational demand, and domain adaptability. The no-augmentation baseline, which used only the available real façade images, achieved an accuracy of 68.50% and a F1-score of 69.05% in England, and an accuracy of 73.80% and a F1-score of 74.70% in Korea. These results provide a reference for determining whether subsequent performance improvements were attributable to the addition of augmented or synthetic training data.

In terms of quantitative performance, traditional augmentation achieved an accuracy of 71.67% with an F1-score of 71.08% in the England dataset, and 85.00% accuracy with an F1-score of 84.82% in the Korean dataset. Although this method effectively increases dataset size through simple geometric and photometric transformations such as rotation and brightness adjustment, its limited ability to enrich material-level texture diversity restricts classification performance. The relatively lower F1-scores in both datasets indicate that basic transformations are insufficient to capture the complex surface characteristics of real façade materials.

Standard diffusion without LoRA improved classification results in both datasets, achieving 77.67% accuracy and an F1-score of 77.65% in England, and 87.67% accuracy with an F1-score of 87.73% in Korea. This improvement demonstrates that generative models introduce richer architectural textures and structural patterns compared to traditional augmentation. However, because the diffusion model was not fine-tuned to façade-specific domains, the generated images often reflected generic building appearances rather than material-specific visual cues, limiting domain adaptability. The proposed LoRA-based stable diffusion method achieved the highest performance in both datasets, with 83.67% accuracy and an F1-score of 82.91% in England, and 93.67% accuracy with an F1-score of 93.68% in Korea. By performing PEFT on façade-domain images, the model learned region-specific visual patterns and generated synthetic samples that closely resembled real material distributions. The advantages of the proposed approach are particularly evident in its ability to reproduce fine-grained surface roughness, reflectance properties, aging effects, and environmental variations such as weather and lighting conditions. This enhanced qualitative diversity enabled the classification model to learn more robust and discriminative representations across different regional façade contexts.

Overall, conventional augmentation provides a computationally inexpensive means of increasing sample variability but remains limited to geometric and photometric modifications of existing images. Standard diffusion introduces broader architectural and texture variations, although it does not explicitly learn region- and material-specific façade distributions. In contrast, LoRA-based diffusion better represents variations in surface texture, weathering, illumination, and façade composition, consistent with its improved synthetic–real alignment and downstream classification performance. However, this benefit requires domain-specific training images and additional computational resources, highlighting a practical trade-off between domain adaptability, performance improvement, and implementation cost.

**Table 5**
Comparative evaluation of data augmentation methods and Baselines.

| Description | Baselines | Traditional Augmentation | Standard Diffusion (No LoRA) | Proposed LoRA-Based Diffusion |
|---|---|---|---|---|
| | Only real façade images | Geometric & photometric transforms | Stable Diffusion generates generic façades | LoRA fine-tuning on façade domain for domain-specific synthesis |
| England Accuracy | 68.50 | 71.67 | 77.67 | 83.67 |
| England F1-score | 69.05 | 71.08 | 77.65 | 82.91 |
| Korea Accuracy | 73.80 | 85.00 | 87.67 | 93.67 |
| Korea F1-Score | 74.70 | 84.82 | 87.73 | 93.68 |
| Texture | - | Limited (generic patterns only) | Moderate (generic building textures) | High (domain-specific texture, material pattern) |
| Data Diversity | - | Low (synthetic transformations) | Moderate | High (weather, aging, lighting variation) |
| Computing Resource | - | Very low | Moderate | High (parameter-efficient LoRA) |
| Domain Adaptability | - | None | Weak | Strong (façade-specific fine-tuning) |
| Limitations | - | Cannot simulate material texture | Unrealistic façade patterns | High computing source |

### 4.4. Robustness and component-wise validation of the proposed framework

Table 6 compares the classification performance of MobileNetV3 and ResNet-50 under the baseline, standard diffusion, and LoRA-based diffusion training conditions. For MobileNetV3, standard diffusion augmentation increased accuracy from 68.50% to 77.67% and the F1-score from 69.05% to 77.65%. The LoRA-based diffusion condition achieved the highest performance, with an accuracy of 83.67% and an F1-score of 82.91%. Compared with the baseline, this represents improvements of 15.17% in accuracy and 13.86%in F1-score. The LoRA-based condition also exceeded standard diffusion by 6.00% in accuracy and 5.26% in F1-score.

ResNet-50 exhibited a performance trend similar to that observed for MobileNetV3. Standard diffusion improved the baseline accuracy from 58.28% to 64.25% and the F1-score from 59.44% to 64.58%, while LoRA-based diffusion further increased accuracy and the F1-score to 77.67% and 76.31%, respectively. Accordingly, both classifier architectures showed the same performance ranking: LoRA-based diffusion, Standard diffusion, and baseline. This consistent trend suggests that the effectiveness of the proposed augmentation approach is not specific to the MobileNetV3 architecture.

To isolate the effects of LoRA fine-tuning and architecture-aware prompt conditioning, four generation settings were compared. As shown in Table 7, replacing the general prompt with the architecture-aware prompt without LoRA increased accuracy from 75.67% to 77.67% and the F1-score from 75.02% to 77.65%. Applying LoRA with the general prompt produced a larger improvement, reaching 82.67% accuracy and an 82.04% F1-score. The combination of LoRA and the architecture-aware prompt achieved the highest performance, with 83.67% accuracy and an 82.91% F1-score. These results indicate that LoRA-based adaptation was the primary contributor to performance improvement, and structured prompt conditioning provided an additional complementary benefit.

## 5. Summarized findings and conclusion

### 5.1. Summary and discussion

The class-wise analysis revealed consistent material-dependent performance patterns across both the Korean and England façade datasets. In both regional contexts, the brick class exhibited consistently high recognition performance, whereas the stone class showed comparatively lower performance, indicating that the intrinsic visual complexity of façade materials influenced classification outcomes. These results indicate that materials characterized by structured and repetitive patterns are more easily captured by both the generative model and the classifier.

The concrete class demonstrated region-dependent behavior. In particular, the relative recognition performance of concrete varied across the two regional datasets, indicating that material classification was influenced not only by material texture but also by region-specific façade composition, surface condition, and surrounding visual context. This suggests that the generative model effectively learned subtle surface variations and tonal distributions characteristic of English concrete façades when trained with approximately 100 LoRA images. The stone class consistently presented the greatest challenge in both datasets. This trend reflects the heterogeneous and irregular surface properties of stone materials, which produce nonlinear texture patterns that are difficult for both generative and discriminative models to represent accurately. The confusion matrices clearly showed frequent misclassification between stone and concrete, likely because both materials share similar chromatic characteristics despite differing in micro-texture.

**Table 6**
Classification performance of MobileNetV3 and ResNet-50 under different augmentation conditions.

| Classifier | Training Setting | Accuracy | F1-score |
|---|---|---|---|
| MobileNetV3 | Baseline | 68.50 | 69.05 |
| MobileNetV3 | Standard Diffusion | 77.67 | 77.65 |
| MobileNetV3 | LoRA-based Diffusion | 83.67 | 82.91 |
| ResNet-50 | Baseline | 58.28 | 59.44 |
| ResNet-50 | Standard Diffusion | 64.25 | 64.58 |
| ResNet-50 | LoRA-based Diffusion | 77.67 | 76.31 |

**Table 7**
Component-wise ablation results for LoRA fine-tuning and architecture-aware prompt conditioning.

| Prompt Condition | LoRA fine-tuning | Accuracy | F1-score |
|---|---|---|---|
| General Prompt | No | 75.67 | 75.02 |
| Architecture-aware prompt | No | 77.67 | 77.65 |
| General Prompt | Yes | 82.67 | 82.04 |
| Architecture-aware prompt | Yes | 83.67 | 82.91 |

The analysis of LoRA training-image volume further showed that classification performance did not increase monotonically with the number of real images used for fine-tuning. The best-performing condition was obtained with 150 images per class in Korea and 100 images per class in England, indicating that the effective training-data volume depended on the regional dataset and its visual variability. Therefore, these values should be interpreted as the best-performing configurations within the investigated range rather than as universal optimal quantities.

When comparing augmentation strategies, the LoRA-based fine-tuned generative augmentation method consistently improved overall generalization performance relative to traditional augmentation in both datasets. The magnitude of improvement was especially notable in visually complex classes such as concrete and stone. In contrast, performance differences in the brick class were relatively small, as its repetitive structural patterns can already be partially diversified through simple geometric transformations. Standard diffusion also improved classification performance relative to the real-image-only baseline and conventional augmentation, demonstrating the general utility of generative data augmentation. However, the additional improvement achieved by LoRA-based diffusion indicates that façade-specific adaptation provided benefits beyond those obtained from generic synthetic image generation. The ablation results showed that LoRA fine-tuning provided the main performance gain, and architecture-aware prompting offered an additional complementary benefit. Their combination achieved the highest accuracy and F1-score.

From a qualitative standpoint, generative augmentation enhanced not only data quantity but also texture fidelity and environmental realism. This enriched visual diversity enabled the classification model to learn more robust discriminative features and reduce domain-specific confusion patterns. The KID and CLIP embedding analyses supported this interpretation: LoRA-generated images showed lower distributional distances and smaller feature-space gaps from the real images than standard diffusion samples across all material classes. The t-SNE visualization also showed class-dependent clustering and partial overlap between real and generated samples, although the remaining separation confirmed that the synthetic–real gap was reduced rather than eliminated.

The alternative-classifier experiment provided an additional robustness check. ResNet-50 exhibited the same performance ranking as MobileNetV3—LoRA-based diffusion, standard diffusion, and baseline—suggesting that the observed benefit of the proposed augmentation strategy was not specific to a single classifier architecture. Nevertheless, the absolute performance differed between the two classifiers, indicating that augmentation effectiveness and classifier capacity should be considered jointly when applying the framework.

These findings imply that future research on construction vision systems should move beyond generic data augmentation strategies and focus on domain-informed data generation mechanisms that explicitly

account for material properties, environmental conditions, and regional architectural characteristics. However, the present findings are limited to three façade-material classes, two independent regional datasets, and a stable diffusion-based generation framework. Further validation using additional cities, façade materials, diffusion models, and cross-region transfer experiments is required to assess broader generalizability.

### 5.2. Conclusion and reflection

The proposed LoRA-based generative augmentation framework produced synthetic façade images that more closely represented the material-specific and regional characteristics of the real datasets. KID, feature-space, and classification results showed that LoRA-based diffusion reduced the synthetic–real discrepancy and outperformed no augmentation, conventional augmentation, and standard diffusion. The component-wise ablation results further showed that LoRA-based parameter adaptation and architecture-aware prompt conditioning provided complementary benefits, with their combination achieving the highest downstream classification performance. The consistent performance trend across MobileNetV3 and ResNet-50 further suggests that the augmentation benefit was not limited to a single classifier architecture. However, the remaining feature-space separation and material-dependent errors indicate that the synthetic–real gap was reduced rather than eliminated.

The framework provides a practical approach to expanding façade-image datasets under limited-data conditions and may support building-stock assessment, maintenance planning, and urban digital-twin applications. Nevertheless, the study was limited to three material classes, two regional datasets, and one diffusion-model family. Future work should therefore examine additional materials, cities, and generative architectures, as well as cross-region transfer settings, to assess the generalizability and robustness of the proposed framework.

## CRediT authorship contribution statement

**Youngseo Hwang:** Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Seunghyeon Wang:** Conceptualization, Data curation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Namhyuk Ham:** Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Sungkon Moon:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (Ministry of Science and ICT) (RS-2025-21063090); (RS-2025-00557724).

## Data availability

Data will be made available on request.

## References

Asadi Shamsabadi, E., Erfani, S.M.H., Xu, C., Dias-da-Costa, D., 2024. Efficient semi-supervised surface crack segmentation with small datasets based on consistency regularisation and pseudo-labelling. Autom. ConStruct. 158, 105181.

Bay, M., von Marcard, T., Fazlija, D., 2025. The Impact of Synthetic Data on Object Detection Model Performance: a Comparative Analysis with Real-World Data. arXiv e-prints arXiv:2510.12208.

Buyukdemircioglu, M., Kocaman, S., 2020. Reconstruction and efficient visualization of heterogeneous 3D city models. Remote Sens. 12, 2128.

Chavan, A., Liu, Z., Gupta, D., Xing, E., Shen, Z., 2023. One-For-All: Generalized Lora for parameter-efficient fine-tuning. *arXiv preprint arXiv:2306.07967*.

Choi, S.-M., Cha, H.-S., Jiang, S., 2024. Hybrid data augmentation for enhanced crack detection in building construction. Buildings 14, 1929.

Davari, S., Kim, D., Tohidifar, A., 2025. ControlNet-based domain adaptation for synthetic construction images via graphical simulation and generative AI. Autom. ConStruct. 180, 106562.

Davtalab, O., Kazemian, A., Yuan, X., Khoshnevis, B., 2022. Automated inspection in robotic additive manufacturing using deep learning for layer deformation detection. J. Intell. Manuf. 33, 771–784.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.

Dong, C.-Z., Catbas, F.N., 2021. A review of computer vision–based structural health monitoring at local and global levels. Struct. Health Monit. 20, 692–743.

Dunphy, K., Fekri, M.N., Grolinger, K., Sadhu, A., 2022. Data augmentation for deep-learning-based multiclass structural damage detection using limited information. Sensors 22, 6193.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17, 1–35.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Han, Z., Gao, C., Liu, J., Zhang, J., Qian Zhang, S., 2024. Parameter-Efficient fine-tuning for Large Models: a Comprehensive Survey. arXiv:2403.14608. Available: https://ui.adsabs.harvard.edu/abs/2024arXiv240314608H.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. 33, 6840–6851.

hollowstrawberr. kohya-colab. (2025, 2025.05.30). Available: https://github.com/hollowstrawberry/kohya-colab.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al., 2021. LoRA: Low-Rank adaptation of large Language models. arXiv 2106, 09685. Available: https://ui.adsabs.harvard.edu/abs/2021arXiv210609685H.

Huang, Y., Zhang, F., Gao, Y., Tu, W., Duarte, F., Ratti, C., et al., 2023. Comprehensive urban space representation with varying numbers of street-level images. Comput. Environ. Urban Syst. 106, 102043.

Hwang, D., Kim, J.-J., Moon, S., Wang, S., 2025. Image augmentation approaches for building dimension estimation in street view images using object detection and instance segmentation based on deep learning. Appl. Sci. 15, 2525.

Ji-Woo, P., Woo-Jin, S., Chan-Sun, P., Hyun-Jun, P., 2024. Analysis of image generation accuracy based on hyperparameters in stable diffusion. Journal of the Korea Institute of Information and Communication Engineering 28, 1275–1281.

Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410.

Kim, H., Yi, J.-S., 2024. Image generation of hazardous situations in construction sites using text-to-image generative model for training deep neural networks. Autom. ConStruct. 166, 105615.

Kong, G., Fan, H., 2020. Enhanced facade parsing for street-level images using convolutional neural networks. IEEE Trans. Geosci. Rem. Sens. 59, 10519–10531.

Laxman, K., Tabassum, N., Ai, L., Cole, C., Ziehl, P., 2023. Automated crack detection and crack depth prediction for reinforced concrete structures using deep learning. Constr. Build. Mater. 370, 130709.

Li, R., Zhao, L., Wei, H., Hu, G., Xu, Y., Ouyang, B., et al., 2025. Multi-defect type beam bridge dataset: GYU-DET. Sci. Data 12, 1101.

Liu, L., Silva, E.A., Wu, C., Wang, H., 2017. A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. Comput. Environ. Urban Syst. 65, 113–125.

Ma, H., Zheng, H., 2023. Text Semantics to image generation: a method of building facades design base on stable diffusion model. In: The International Conference on Computational Design and Robotic Fabrication, pp. 24–34.

Mingi, J., Jungmin, N., 2025. Analysis of the facade design elements for low-rise residential areas - comparative analysis of buildings featured in architectural magazines with general buildings since 2015. KIEAE Journal 25, 51–64.

Mo, J., Kang, Y., 2022. Understanding Seoul's Mapo·Yongsan·Seongdong (Mayongseong) Phenomenon: the Rise of LWP Centers in Seoul. Korean journal of Urban Studies, pp. 93–139.

Nasir, U., Chang, R., Omrany, H., 2021. Calculation methods for construction material stocks: a systematic review. Appl. Sci. 11, 6612.

Oh, J.-H., Kim, H.-G., Ahn, H.-S., Jeon, Y.-C., Kim, S.-S., 2021. An Analysis on Characteristics of Existing Public Buildings Envelope in View of Green Remodeling, pp. 352–353.

Pandey, V., Mishra, S.S., 2025. A review of image-based deep learning methods for crack detection. Multimed. Tool. Appl. 1–43.

Park, J., Hong, S., Choo, S., 2025. A generation method and evaluation of architectural facade design using stable diffusion with LoRA and ControlNet. Journal of the Architectural Institute of Korea 41, 85–96.
Pittam, J., O'Sullivan, P.D., O'Sullivan, G., 2016. A remote measurement and mapping technique for orderly rapid aggregation of building stock geometry. Autom. ConStruct. 71, 382–397.
Raghu, D., Bucher, M.J.J., De Wolf, C., 2023. Towards a 'resource cadastre' for a circular economy – Urban-scale building material detection using street view imagery and computer vision. Resour. Conserv. Recycl. 198, 107140.
Roberge, J.-B., Contreras, G., Kakinami, L., Van Hulst, A., Henderson, M., Barnett, T.A., 2022. Validation of desk-based audits using Google Street view® to monitor the obesogenic potential of neighbourhoods in a pediatric sample: a pilot study in the QUALITY cohort. Int. J. Health Geogr. 21, 2.
Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2021. High-Resolution image synthesis with latent diffusion models. arXiv:2112, 10752. Available: https://ui.adsabs.harvard.edu/abs/2021arXiv211210752R.
Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695.
Roumpani, F., Hudson, P., Hudson-Smith, A., 2018. The use of historical data in rule-based modelling for scenarios to improve resilience within the building stock. Hist. Environ.: Policy & Practice 9, 328–345.
Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K., 2023. Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22500–22510.
Sakaridis, C., Dai, D., Van Gool, L., 2018. Semantic foggy scene understanding with synthetic data. Int. J. Comput. Vis. 126, 973–992.
Sami Ur Rehman, M., Shafiq, M.T., Ullah, F., 2022. Automated computer vision-based construction progress monitoring: a systematic review. Buildings 12, 1037.
Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R., 2018. Learning from synthetic data: addressing domain shift for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3752–3761.
Soans, R., Masuda, R., Fukumizu, Y., 2025. A multi-task deep learning framework for road quality analysis with scene mapping via sim-to-real adaptation. Appl. Sci. 15, 8849.
Stability, A.I., 2024. Stable Diffusion 3.5 Prompt Guide.
Steinmetz-Wood, M., Velauthapillai, K., O'Brien, G., Ross, N.A., 2019. Assessing the micro-scale environment using Google Street View: the virtual systematic tool for evaluating pedestrian streetscapes (virtual-STEPS). BMC Public Health 19, 1246.
Varma, M., Delbrouck, J.-B., Hooper, S., Chaudhari, A., Langlotz, C., 2023. Villa: Fine-grained vision-language representation learning from real-world data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22225–22235.
Wang, S., 2024. Evaluation of impact of image augmentation techniques on two tasks: window detection and window states detection. Results Eng. 24, 103571.
Wang, S., Korolija, I., Rovas, D., 2022. Impact of Traditional Augmentation Methods on Window State Detection.
Wang, S., Park, S., Park, S., Kim, J., 2024. Building façade datasets for analyzing building characteristics using deep learning. Data Brief 57, 110885.
Wang, S., Han, J., 2024. Automated detection of exterior cladding material in urban area from street view images using deep learning. J. Build. Eng. 96, 110466.
Xiao, B., Kang, S.-C., 2021. Development of an image data set of construction machines for deep learning object detection. J. Comput. Civ. Eng. 35, 05020005.
Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847.
Zhu, Y., Shen, Z., Zhao, Z., Wang, S., Wang, X., Zhao, X., et al., 2024. Melo: Low-rank adaptation is better than fine-tuning for medical image diagnosis. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5.
Zhuang, Z., Zhang, Y., Wang, X., Lu, J., Wei, Y., Zhang, Y., 2024. Time-Varying LoRA: towards effective cross-domain fine-tuning of diffusion models. Adv. Neural Inf. Process. Syst. 37, 73920–73951.