applied sciences

MDPI

Article

StyleSPADE: Realistic Image Augmentation for Robust Infrastructure Crack Segmentation via Ensemble Learning

Jaeung Sim [1], Menas Kafatos [2], Seung Hee Kim [2,*] and Yangwon Lee [1,*]

[1] Major of Geomatics Engineering, Division of Earth Environmental System Sciences, Pukyong National University, Busan 48513, Republic of Korea; simjw0223@pukyong.ac.kr
[2] Institute for Earth, Computing, Human and Observing (ECHO), Chapman University, Orange, CA 92866, USA; kafatos@chapman.edu
* Correspondence: sekim@chapman.edu (S.H.K.); modconfi@pknu.ac.kr (Y.L.)

Featured Application

This research provides an advanced AI-based framework for automated structural health monitoring. The proposed StyleSPADE augmentation and ensemble learning strategy can be integrated into drone-based or robot-assisted inspection systems to accurately detect and analyze surface cracks in aging infrastructure such as bridges, tunnels, and roads, even with limited labeled datasets.

Abstract

The rapid deterioration of global infrastructure necessitates precise and automated crack detection technologies for proactive maintenance. However, deep learning-based segmentation models often suffer from a scarcity of diverse, high-quality labeled datasets. This study proposes StyleSPADE, a novel conditional image generation model that integrates semantic masks and style images to synthesize realistic crack data with diverse background textures while preserving precise geometric morphology. To validate the effectiveness of the generated data, we conducted extensive semantic segmentation tasks using Transformer-based (Mask2Former, Swin-UPerNet) and CNN-based (K-Net) models. Experimental results demonstrate that StyleSPADE-based augmentation significantly outperforms baseline models, achieving a Crack IoU of 0.6376 and an F1-score of 0.7586. Furthermore, we implemented a Stacking Ensemble strategy combining high-recall and high-precision models, which further improved performance to a Crack IoU of 0.6452. Our findings confirm that StyleSPADE effectively mitigates the data scarcity problem and enhances the robustness of crack detection in complex environmental conditions. This framework contributes to improving the efficiency and safety of infrastructure management by enabling reliable damage assessment in data-limited environments.

Keywords: infrastructure crack segmentation; StyleSPADE; data augmentation; conditional image generation; ensemble learning

Check for updates

Academic Editor: Zhengjun Liu

Received: 25 December 2025
Revised: 11 January 2026
Accepted: 13 January 2026
Published: 14 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid deterioration of major infrastructure—such as roads, bridges, and tunnels—has become a critical global challenge [1–3]. Numerous reports indicate that neglecting cracks in these structures leads to traffic accidents or catastrophic collapses, resulting in loss of life and a sharp increase in maintenance costs [4,5]. Therefore, accurately detecting

cracks—a key indicator of early structural damage—and securing technologies for rapid response are urgent tasks in structural maintenance [6,7].

Cracks are typically narrow, irregular in shape, and tend to propagate over time [8]. Such minute structural damage is difficult to identify in its early stages. Even when no significant external changes are visible, internal damage can progress, necessitating precise detection technology. While cracks are initially confined to small areas, they can expand under repeated external forces or environmental changes, significantly compromising structural stability. Conversely, detecting cracks before they expand allows for repair and reinforcement measures that extend a structure's lifespan at a lower cost, effectively preventing major accidents [9–11]. In this regard, early crack detection ensures long-term maintenance efficiency and safety.

Traditional crack detection methods have primarily relied on visual inspection or manual interpretation of photographs. However, these methods vary in accuracy depending on the inspector's skill level. Furthermore, because the captured data is typically interpreted by humans post-inspection, the process is not automated, making it inefficient for monitoring large areas. Additionally, physical inspections often require closing sections of roads or tunnels, which disrupts traffic flow and increases social costs and operational risks.

To overcome these limitations, active research is underway to utilize Unmanned Aerial Vehicles (UAVs), vehicle-mounted scanners, and LiDAR sensors to capture non-contact images for AI-based detection [12–16]. For example, Xiang et al. [17] proposed GC-YOLOv5s, a lightweight model for UAV-based road crack detection, which outperformed the original YOLOv5s. Ren and Zhong [18] developed the BCCD-YOLO model for concrete structures, improving precision by 3.3% through channel and spatial attention mechanisms. Meanwhile, Feng et al. [19] utilized MLS-based LiDAR data to convert road point clouds into graph structures and applied Graph Convolutional Networks (GCN) for detection.

While these studies demonstrate the technical feasibility of automated, non-contact damage detection, most rely on bounding box-based detection, which cannot precisely identify a crack's exact shape or area. LiDAR also faces constraints, including high equipment costs, a lack of color/texture information, and the requirement for manual post-processing. Furthermore, the performance of AI-based approaches depends heavily on large-scale labeled datasets. Collecting and labeling actual crack data is labor-intensive and costly. Moreover, limited diversity in crack shapes and backgrounds restricts the generalization capabilities of these models [20,21]. While geometric augmentations (rotation, scaling) and spectral adjustments (brightness, contrast) are commonly used [22–25], they are often too simple to reflect complex real-world backgrounds or structural diversity.

To address this, Generative Adversarial Network (GAN)-based image generation has been applied to synthesize crack shapes that are difficult to obtain in practice. For instance, Yu et al. [26] proposed VDCrackGAN, which combines VAE-GAN and Transformers, while Zhang et al. [27] used CrackGAN to generate high-quality images under incomplete labeling conditions. However, traditional GANs generate images from random noise vectors, making it difficult to control structural characteristics or reflect user intent, and they often require additional manual labeling.

More recently, diffusion-based generative models have been investigated for synthesizing realistic crack images and performing data augmentation. While Denoising Diffusion Probabilistic Models (DDPM) exhibit exceptional generative capacity and visual fidelity, they typically produce images in isolation without inherently generating corresponding pixel-wise annotations. Consequently, manual labeling remains a prerequisite before such synthesized data can be utilized for crack segmentation tasks [28].

Consequently, conditional image generation—which uses semantic masks as conditions—is gaining attention. Models such as SPADE (Spatially-Adaptive Normalization), Seman-

ticStyleGAN, and ControlNet enable precise synthesis by assigning specific visual characteristics to defined regions. For instance, Li et al. [29] employed the SPADE module to generate synthetic cardiac Magnetic Resonance Imaging (MRI) data, which led to a 5–10% improvement in segmentation performance. Filipovski and Gievska [30] demonstrated the effectiveness of SemanticStyleGAN for image inpainting tasks, utilizing StyleGAN's style-based generator to achieve higher quality image restoration. Similarly, Deng et al. [31] utilized ControlNet to synthesize images for weed detection, achieving a 1.4% increase in object detection accuracy compared to using real images alone.

However, directly applying these models to crack synthesis presents challenges: representing thin, low-contrast cracks is difficult, and models often produce noise or monotonous, repetitive backgrounds. This inability to simultaneously achieve accurate crack morphology and realistic background texture limits the model's generalization performance.

To address these limitations, this study proposes StyleSPADE, a conditional image generation model. StyleSPADE leverages both semantic masks and style images as inputs, facilitating the synthesis of highly realistic crack images across diverse background textures and color profiles. Crucially, the model is architecturally optimized for crack augmentation; by injecting style information through affine normalization parameters, it ensures that intricate, high-frequency crack geometries are preserved without distortion. This approach effectively mitigates the scarcity of labeled data by providing high-fidelity training samples that closely emulate real-world environments.

Furthermore, we empirically verify the effectiveness of StyleSPADE-based augmentation through semantic segmentation tasks. Unlike previous studies focused on bounding boxes, we perform detailed damage detection to precisely identify the shape and area of cracks. Finally, we analyze various segmentation models and backbone combinations and present an ensemble strategy to maximize detection performance. This study contributes to the field by providing an effective data augmentation technique and an ensemble strategy that enhances crack detection even in data-limited environments.

2. Materials and Methods

2.1. Overview

The overall research methodology is illustrated in Figure 1. We trained a conditional image generation model using a binary semantic mask—comprising background (0) and crack (1) regions—alongside actual crack images to generate synthetic data.

Conditional image generation models often exhibit reduced stability during training when the background is treated as a single class. This is because diverse backgrounds in the training data can cause boundary distortions or background noise in the synthesized images. Such issues arise because these models tend to rely more heavily on semantic information than on spatial adaptability [32]. Consequently, restricting the background of the training data to a specific, well-defined range can enhance the stability and quality of the generated output.

While this background restriction might potentially introduce a bias toward concrete-like textures during the generation phase, we explicitly addressed its generalization capability. This was achieved by evaluating downstream segmentation performance on an independent dataset characterized by heterogeneous surface conditions. Such an experimental design enables us to verify whether the advantages of stable, single-background generation effectively translate into enhanced robustness across diverse real-world environments.

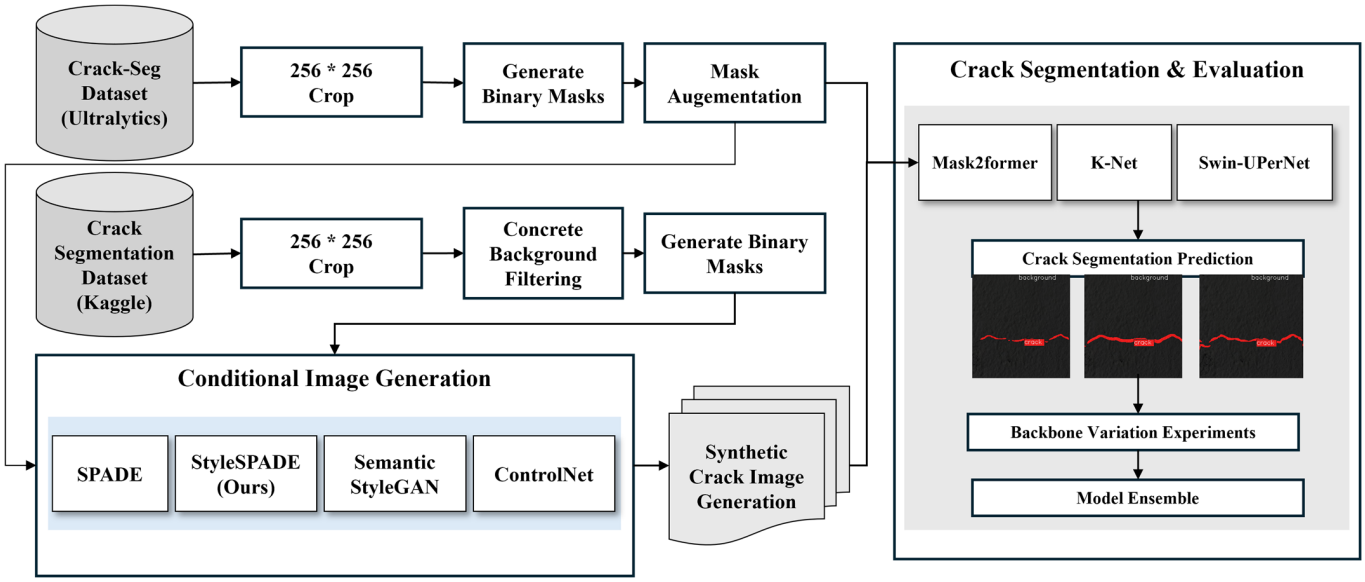

Figure 1. Overview of the proposed research pipeline. The framework comprises three primary stages: (1) Preprocessing: semantic mask extraction and background filtering from Kaggle and Ultralytics datasets; (2) Conditional Generation: synthesis of realistic crack images using StyleSPADE (Ours) alongside baseline models (SPADE, Semantic StyleGAN, and ControlNet); and (3) Evaluation: downstream crack segmentation performance assessment using Mask2former, K-Net, and Swin-UPerNet, followed by backbone variation analysis and model ensemble strategies.

Subsequently, we performed a segmentation task to validate the generalization performance of the synthesized images. For this process, it is essential to train and test the segmentation model on a dataset that reflects diverse background conditions [33,34]. Specifically, if the dataset used to train the generation model is identical to the one used for the segmentation task, the synthetic images may overly resemble the evaluation data. This could lead to a data leakage effect, potentially overestimating the segmentation performance. To prevent this bias, we selected separate, independent datasets (1) for building the conditional image generation model (Kaggle Crack Segmentation Dataset) and (2) for training and evaluating the segmentation model (Ultralytics Crack-Seg Dataset).

2.2. Dataset

2.2.1. The Kaggle Crack Segmentation Dataset

First, we utilized the Kaggle Crack Segmentation Dataset [35] to build our conditional image generation model. This dataset is a comprehensive integration of 11 open-source crack image repositories, comprising approximately 11,200 images, including CrackTree200, the Crack Forest Dataset (CFD), DeepCrack, and Crack500. Approximately half of these images features concrete backgrounds.

Ham et al. [36] previously utilized this dataset to analyze tunnel cracks and reported that training a segmentation model exclusively on concrete backgrounds—matching the target tunnel environment—yielded superior performance compared to training on the entire, more diverse dataset. This finding suggests that narrowing the background scope can significantly enhance training efficacy.

Accordingly, we refined our training data by selecting only sub-datasets with concrete-related backgrounds (excluding "non-crack" or irrelevant backgrounds). This restriction ensures that the conditional image generation model can achieve high training stability, even when learning from a dataset with a unified background class. During preprocessing, the original 448 × 448 pixels were resized to 256 × 256 to match the model's input requirements.

Table 1 summarizes the names, image counts, and primary background types for each sub-dataset within the original Crack Segmentation Dataset. Based on these details, Table 2 presents the final sub-dataset composition and corresponding image counts after restructuring them according to the concrete background criteria.

Table 1. Summary of the Kaggle Crack Segmentation Dataset: distribution of images and surface types (concrete vs. asphalt) across the 11 sub-datasets.

Name of Sub-Dataset	Total Images	Background
CFD	118	concrete surface
Cracktree	206	concrete surface
Eugen Muller	55	concrete surface
Sylvie Chambon	185	concrete surface
Forest	118	concrete surface
Rissbilder	3822	concrete surface
Volker	990	concrete surface
DeepCrack	521	Some of the images are concrete; the others are not
Crack500	3363	asphalt surface
GAPS500	509	asphalt surface
noncrack	1411	concrete surface

Table 2. Reorganization of the Kaggle Crack Segmentation Dataset according to concrete background criteria to enhance the training stability of the conditional image generation models.

Re-Organized Dataset	Sub-Datasets	Number of Images
Original	All	11,298
Concrete background	CFD, Cracktree, Eugen Muller, Sylvie Chambon, Forest, Rissbilder, Volker, DeepCrack	5570

2.2.2. The Ultralytics Crack-Seg Dataset

To evaluate the generalization performance of our image generation model, we performed data augmentation on the Ultralytics Crack-Seg Dataset [37]. The augmented dataset was subsequently used for the training and testing of a segmentation model focused on infrastructure cracks. The Crack-Seg Dataset includes images of cracks on various structural surfaces, such as roads, sidewalks, and walls. It contains a total of 4029 images, which were partitioned into 3717 training, 200 validation, and 112 test images.

Notably, this dataset includes standard augmentation techniques, such as rotation and brightness adjustments, making it well-suited for benchmarking and validating segmentation model performance. Figure 2 illustrates examples of images from the Crack-Seg Dataset that have undergone these conventional augmentations.

The original images were 416 × 416 pixels in size. During the augmentation process, any margins created by rotation were removed. To maintain consistency with the model's input requirements, all images were center-cropped to 256 × 256. Furthermore, while the original masks were provided as coordinate-based polygons, we converted them into pixel-wise semantic masks and cropped them to match the corresponding image dimensions.

2.2.3. Augmentation and Redundancy Filtering of Crack Masks

The input masks for the conditional image generation model were synthesized by applying geometric augmentation techniques—including rotation, flipping, and translation—to the original masks from the Crack-Seg Dataset.



Figure 2. Visual examples of data augmentation applied to the Ultralytics Crack-Seg Dataset, including geometric (rotation) and photometric (brightness adjustment) transformations to improve model robustness.

To ensure dataset diversity and minimize redundancy, we implemented an overlap-based filtering process. During augmentation, the overlap ratio of crack pixels was calculated between the original and augmented masks, as well as among various augmented masks. Masks were excluded if their crack pixel overlap exceeded 15%; this was the threshold at which cases were deemed too similar to contribute meaningful variance. This threshold was determined through empirical sensitivity analysis: stricter criteria (<10%) significantly limited the yield of usable masks, while looser thresholds (>20%) resulted in geometrically and visually redundant patterns. Consequently, the 15% threshold struck a practical balance between dataset size and pattern diversity. This systematic selection secured a diverse set of masks while effectively minimizing redundancy. Figure 3 illustrates examples of the original crack semantic masks alongside augmented versions generated through conventional geometric transformations.

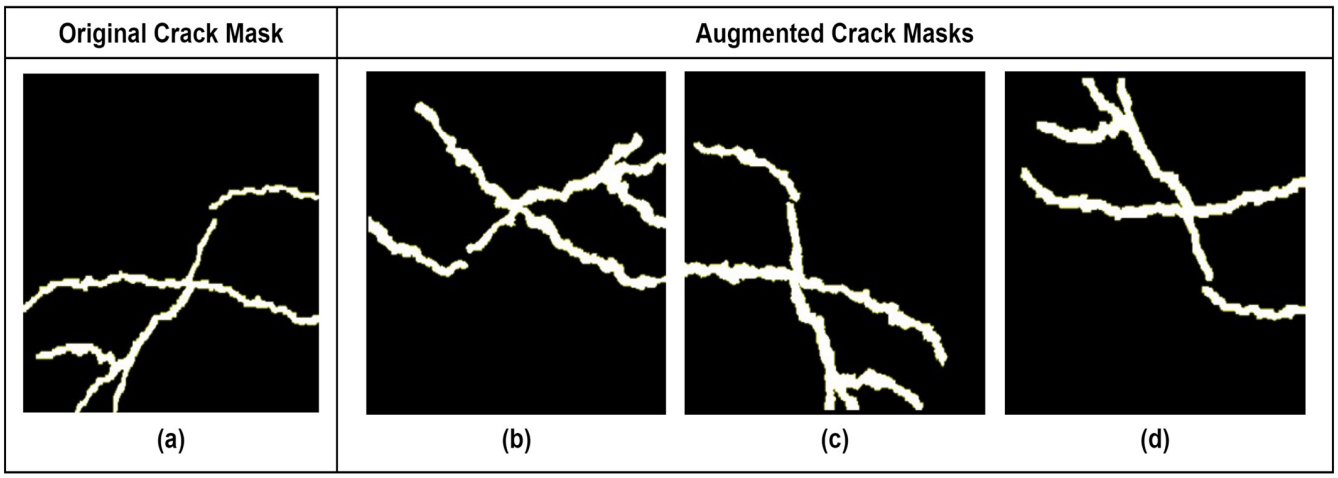

Figure 3. Examples of crack mask augmentation and redundancy filtering: (**a**) original crack semantic mask from the Crack-Seg dataset; (**b**–**d**) augmented masks generated through geometric transformations, ensuring a crack pixel overlap ratio of less than 15% to maintain data diversity.

2.3. Data Augmentation via Conditional Image Generation Models

Conditional image generation models are broadly categorized into GAN-based and Diffusion-based approaches. GANs facilitate learning through a competitive framework between a generator and a discriminator. The generator aims to synthesize images indistinguishable from real data to bypass the discriminator, while the discriminator evaluates

whether a given image is authentic or generated. This mutual competition enables rapid synthesis and high-resolution image generation once stable convergence is achieved.

However, GANs possess inherent limitations: if the generator converges on specific patterns that consistently succeed in bypassing the discriminator, it may repeatedly produce limited outputs, a phenomenon known as mode collapse. Furthermore, the adversarial training process is often unstable.

Consequently, diffusion-based image generation has recently gained significant traction. Unlike GANs, diffusion models do not utilize a competitive structure; instead, they operate by progressively removing Gaussian noise to restore and generate images. This approach ensures relatively stable training and high-quality results. However, the iterative sampling process is time-consuming, leading to slower generation speeds and high computational costs.

Given these characteristics, we selected a conditional image generation approach to synthesize crack images in a targeted format while enhancing diversity and realism. To evaluate the effectiveness of various architectures, we compared the following models:

- SPADE: A GAN-based model renowned for its layout preservation using semantic masks as input.
- SemanticStyleGAN: An extension of StyleGAN capable of reflecting diverse backgrounds even when trained on a single-class dataset.
- ControlNet: A model based on the large-scale pre-trained Stable Diffusion framework.
- StyleSPADE: The model proposed in this study.

These models are designed to generate high-quality synthetic images that closely resemble authentic data when guided by semantic masks, with each model exhibiting distinct structural characteristics based on its specific conditional injection method. To ensure fairness and reproducibility, we adopted the official training configurations provided by the original authors for all baselines, including SPADE, SemanticStyleGAN, and ControlNet. Due to hardware constraints (NVIDIA RTX 4090, 24 GB VRAM), we only modified the batch size and the total number of training iterations; all other hyperparameters—including optimizer settings, learning-rate schedules, loss weights, and architectural defaults—remained identical to the original implementations.

2.3.1. SPADE and StyleSPADE

SPADE is a conditional image generation model capable of synthesizing high-resolution images based on a specified semantic layout. Traditional normalization methods often struggle to preserve semantic information throughout deep network layers, as the normalization process tends to wash out the input mask's spatial details. To address this, SPADE learns to spatially adapt the affine transformation coefficients—specifically, scale (γ) and bias (β)—applied after normalization, directly using the semantic mask as a guide.

This approach mitigates the loss of semantic information inherent in conventional normalization techniques and ensures that the spatial structure of the input mask is retained during the image generation process [38]. The normalization operation in SPADE is formally defined as follows:

$$\hat{h}_{n,c,y,x} = \gamma_{c,y,x}(m) \cdot \frac{h_{n,c,y,x} - \mu_c}{\sigma_c} + \beta_{c,y,x}(m) \quad (1)$$

where γ and β are position-based modulation variables learned from the input semantic mask and are computed through a two-layer convolutional network. By utilizing this spatial modulation, SPADE overcomes the semantic information loss inherent in conventional normalization layers, enabling the generation of high-quality images that preserve complex semantic structures with fewer parameters. These characteristics are particularly

advantageous for synthesizing thin, intricate features such as crack morphologies. Consequently, SPADE's ability to effectively reflect semantic information while maintaining morphological stability led to its selection as a candidate model in this study.

However, SPADE exhibits a notable limitation: when the background of the input mask belongs to a single class, the synthesized background tends to be generated monotonically in repetitive shades of gray. To address this, we propose StyleSPADE, which extends the original architecture to incorporate external style images.

Building on research by Wang et al. [39], which demonstrated that VGG encoders can effectively extract stylistic features, we integrated a VGG-based style encoder to extract a latent style vector. This vector replaces the random noise typically used in the original generator.

The VGG-based style encoder extracts features from the relu5_3 layer of a pretrained VGG16 (the final convolutional block). These features undergo global average pooling and pass through two fully connected layers (512→256→128) to yield a 128-dimensional style vector that captures both textural and chromatic characteristics. This style vector is then integrated into each SPADE block via additive affine-parameter modulation. Specifically, two linear layers transform the style vector into channel-wise parameters (γ_{style}, β_{style}), which are subsequently added to the semantic-derived parameters:

$$\gamma_{final} = \gamma_{semantic} + \gamma_{style} \tag{2}$$

$$\beta_{final} = \beta_{semantic} + \beta_{style} \tag{3}$$

where $\gamma_{semantic}$ and $\beta_{semantic}$ represent spatially-adaptive parameters learned from the semantic mask, while γ_{style} and β_{style} facilitate global appearance modulation. This architecture enables decoupled control over the crack structure (via the semantic mask) and the background appearance (via the style vector). Unlike concatenation-based or multiplicative conditioning strategies, this additive modulation design ensures that style information does not overshadow the geometric constraints imposed by the semantic mask. Such a property is particularly critical for crack synthesis, where thin, high-frequency structures must be consistently preserved across diverse background textures.

The existing SPADE normalization mechanism, the internal structure of the SPADE block, and the architecture of the proposed StyleSPADE generator are detailed in Figure 4.

2.3.2. SemanticStyleGAN

SemanticStyleGAN is a conditional image generation model that utilizes semantic masks as input to synthesize high-quality images with independent control over specific regions. Unlike conventional StyleGAN-based models that rely on a single global latent code, SemanticStyleGAN partitions the latent space by semantic class, enabling region-specific control of structure and texture.

Specifically, the latent space is divided into a global latent code governing overall structure and background, and region-specific latent codes that modulate local texture and appearance [40].

Unlike SPADE, SemanticStyleGAN learns latent codes for different semantic regions independently before synthesizing them into a single coherent image. We selected SemanticStyleGAN as a comparative baseline because this decoupled design allows separate control of crack and background regions, enabling high visual diversity in background textures. Figure 5 illustrates the overall generation process.

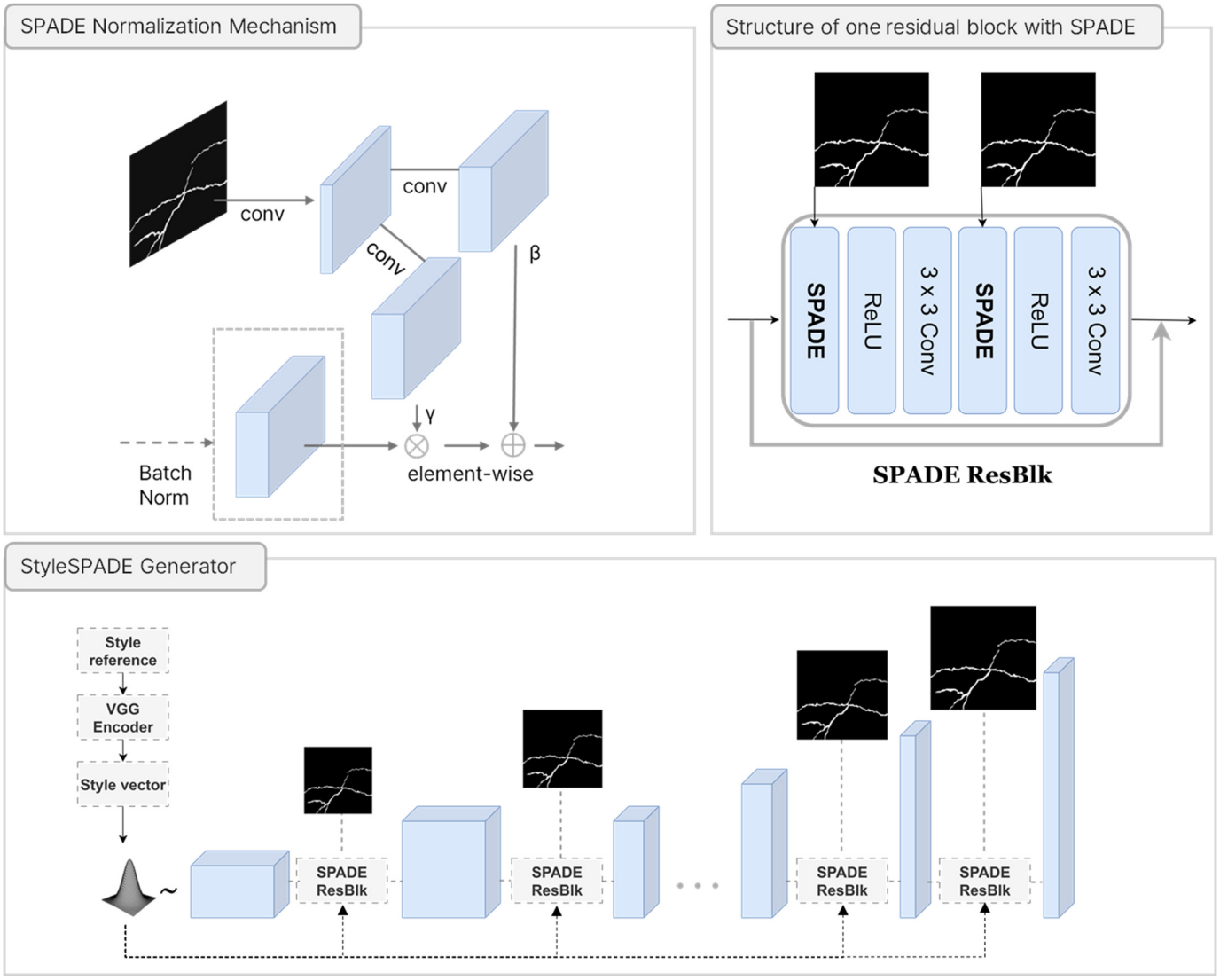

Figure 4. Detailed architecture of the proposed StyleSPADE: including the spatially-adaptive normalization (SPADE) mechanism, the internal structure of the SPADE residual block (ResBlk), and the global generator framework incorporating the VGG-based style encoder.

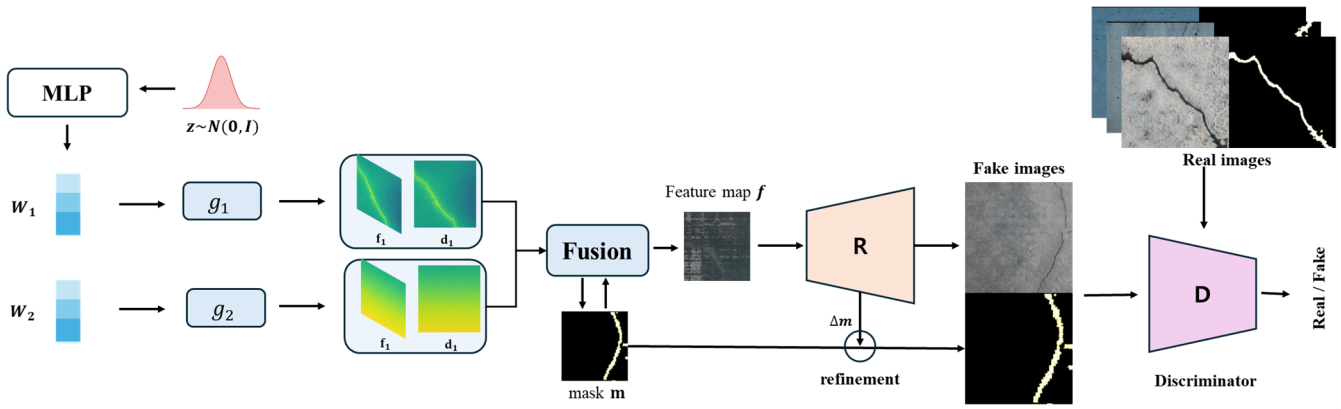

Figure 5. Architectural overview of the modified SemanticStyleGAN for crack synthesis. The framework utilizes decoupled latent codes (w_1, w_2) to modulate local region-wise styles through generators (g_1, g_2). These features are integrated via a Fusion module with the input semantic mask (m), followed by a refinement process (Δm) and a render network (R) to produce high-fidelity fake images, which are evaluated alongside real images by a discriminator (D).

2.3.3. ControlNet

ControlNet is a conditional image generation model built upon the Stable Diffusion framework, designed to enable precise spatial control by injecting external conditions such as semantic masks. The model introduces trainable copies of the pre-trained diffusion blocks, allowing conditional information to be incorporated while preserving the expressive power of the original Stable Diffusion model [41]. We selected ControlNet as a candidate model because it leverages the expressive capacity of pre-trained Stable Diffusion to generate realistic synthetic crack images and supports multimodal conditioning using semantic masks and text prompts, allowing independent control over cracks and backgrounds. Figure 6 illustrates the overall conditioning framework of ControlNet.

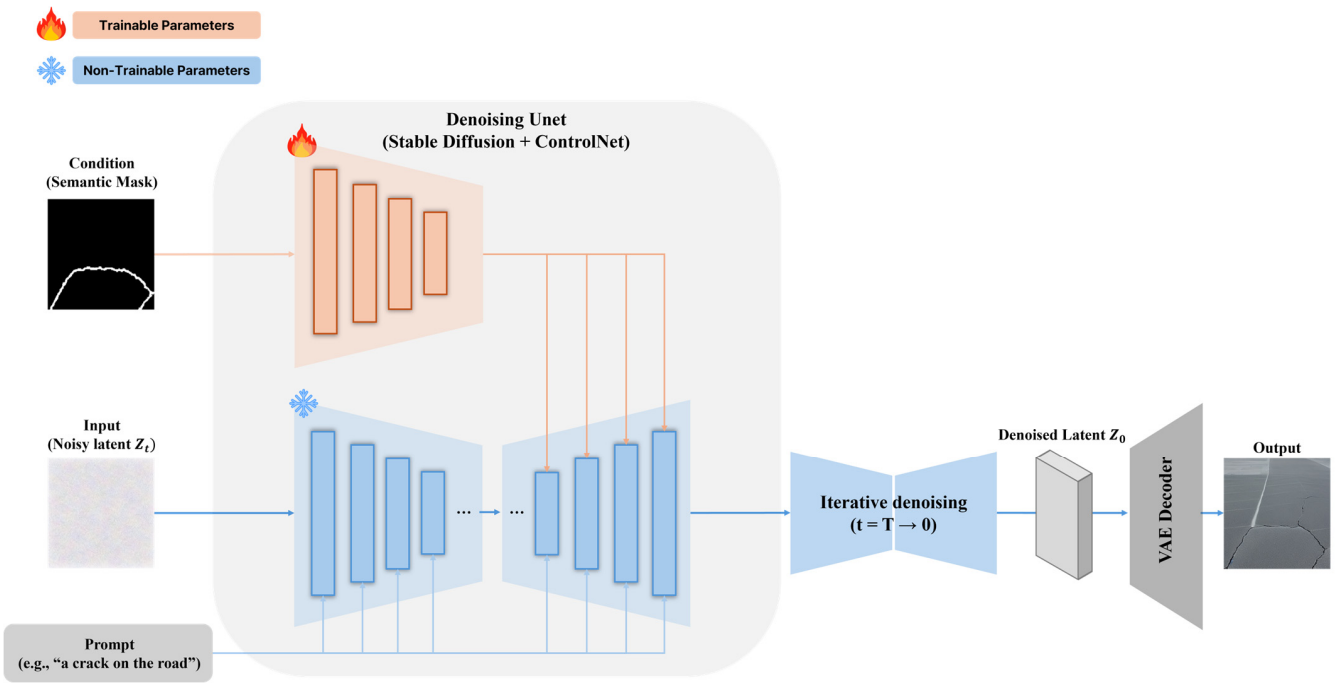

Figure 6. Architectural overview of the ControlNet-based synthesis framework. A frozen Stable Diffusion model is integrated with a trainable ControlNet to enable conditional image generation. Semantic masks are processed as spatial conditions and injected into the denoising U-Net, where the latent representation is iteratively refined and finally reconstructed into a high-fidelity image via a VAE decoder.

2.4. Segmentation Models

To validate the effectiveness of the synthetic images generated by the aforementioned conditional models, we conducted a series of semantic segmentation tasks. These experiments were designed to identify the most effective data augmentation technique and to provide a comparative analysis of performance across various backbones and segmentation architectures. All models were implemented and evaluated using the MMSegmentation framework.

For crack detection, our experimental setup included Transformer-based models, such as Mask2Former and Swin-UPerNet, alongside the CNN-based Kernel Network (K-Net). Given the architectural flexibility of Mask2Former and K-Net, we performed additional ablation studies by replacing their default ResNet-50 backbones with Swin-T and Swin-L, respectively. Furthermore, as Swin-UPerNet integrates the Swin Transformer with the UPerNet decoder as a unified architecture within the framework, we evaluated three backbone scales: Swin-T, Swin-B, and Swin-L. This comprehensive configuration allows

for a detailed analysis of how specific backbone-decoder combinations influence overall segmentation performance.

2.4.1. Mask2Former

Mask2Former is a universal Transformer-based framework designed to address various segmentation tasks, including semantic, instance, and panoptic segmentation, within a unified architecture. It consists of a pixel decoder that extracts multi-scale features and a Transformer decoder that predicts segmentation masks via learnable object queries.

A key characteristic of Mask2Former is its masked attention mechanism, which restricts cross-attention to regions relevant to the predicted masks. This design improves localization accuracy and computational efficiency, making the model particularly effective for segmenting irregular objects with complex boundaries [42].

In this study, Mask2Former was selected because its mask-centric attention mechanism is well suited for capturing thin, discontinuous structures such as surface cracks.

2.4.2. Swin-UPerNet

Swin-UPerNet is a robust segmentation framework that integrates a Swin Transformer backbone with a Unified Perceptual Parsing Network (UPerNet) decoder. The Swin Transformer utilizes a shifted window mechanism to capture both local and global contextual dependencies while maintaining linear computational complexity relative to image size. Its hierarchical structure enables the extraction of multi-resolution features, which is critical for representing objects at different scales [43].

Complementing this, the UPerNet decoder merges the architectural advantages of the Feature Pyramid Network (FPN) and the Pyramid Pooling Module (PPM). This combination facilitates the effective integration of rich semantic information across multiple scales, thereby enhancing overall decoding precision [44]. Swin-UPerNet was selected for this study because it synergistically combines the global expressive power of Transformers with the multi-scale integration characteristic of CNN-based decoders. This makes it exceptionally well-suited for the precise segmentation of irregularly shaped objects with significant scale variations, such as surface cracks.

2.4.3. K-Net

K-Net is a kernel-based segmentation framework that unifies semantic, instance, and panoptic segmentation within a single architecture by directly predicting masks through learnable kernels. Unlike traditional approaches that rely on bounding box heuristics, K-Net generates segmentation masks via dynamic convolution between feature maps and category-specific kernels [45].

This mask-centric formulation allows the model to effectively capture irregular object boundaries without explicit region proposals. In our experiments, K-Net was implemented with a fixed UPerNet segmentation head, and its performance was evaluated using different backbone architectures, including ResNet-50, Swin-T, and Swin-L.

The kernel-based representation and iterative refinement mechanism of K-Net are particularly suitable for segmenting thin and irregular structures, such as surface cracks, motivating its inclusion as a comparative model in this study.

2.5. Segmentation Model Training

All the models were trained on a high-performance workstation equipped with an Intel Core i9-13900K CPU, an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM), and 32 GB of RAM. The specific training hyperparameters for the segmentation models are summarized in Table 3. To ensure consistency across experiments, all models were trained for a total of 16,000 iterations with a uniform input image resolution. A default batch

size of 2 was employed; however, for the Mask2Former model family, the batch size was increased to 16 to optimize memory utilization and training stability. Optimization was performed using the AdamW optimizer, with the initial learning rate set to either 0.00006 or 0.0001 depending on the architecture. Learning rate scheduling was implemented via MultiStepLR, PolyLR, or LinearLR strategies as appropriate for each model's convergence characteristics. Additionally, weight decay was applied within a range of 0.0005 to 0.01 to prevent overfitting.

Table 3. Comprehensive summary of training hyperparameters for different segmentation architectures evaluated in this study, ensuring consistent experimental conditions across 16,000 iterations.

Method	Learning Rate	Optimizer	Batch Size	Max Iterations	Scheduler	Weight Decay
K-Net + UPerNet/Swin-T	0.00006	AdamW	2	16,000	MultiStepLR	0.0005
K-Net + UPerNet/Swin-L	0.00006	AdamW	2	16,000	PolyLR, MultiStepLR	0.01
K-Net + UPerNet/ResNet-50	0.0001	AdamW	2	16,000	MultiStepLR	0.0005
Mask2Former/Swin-T	0.0001	AdamW	16	16,000	PolyLR	0.05
Mask2Former/Swin-L	0.0001	AdamW	16	16,000	PolyLR	0.05
Mask2Former/ResNet-50	0.0001	AdamW	16	16,000	PolyLR	0.0005
Swin-UPerNet/Swin-T	0.00006	AdamW	2	16,000	LinearLR, PolyLR	0.01
Swin-UPerNet/Swin-B	0.00006	AdamW	2	16,000	LinearLR, PolyLR	0.01

2.6. Segmentation Model Ensemble

Based on the performance comparison across various segmentation models and backbones, we implemented an ensemble strategy to optimize the balance between Recall and Precision. Our primary approach combines the model with the highest Recall and the model with the highest Precision, aiming to simultaneously minimize false negatives (FN) and false positives (FP). To validate this design, we introduced two baseline configurations for comparison: a Recall-focused ensemble, which averages the two highest-recall models, and a Precision-focused ensemble, which averages the two highest-precision models. These baselines facilitate an explicit analysis of the performance trade-offs between ensembles biased toward minimizing either FN or FP. Subsequently, the most effective Avg-Logit configuration was further extended using a Stacking Ensemble approach. This step evaluates the efficacy of meta-learning–based fusion against simple averaging while maintaining computational efficiency by avoiding unnecessary complexity.

2.7. Segmentation Model Evaluation

To quantitatively evaluate the performance of the segmentation models, we utilized Precision, Recall, Intersection over Union (IoU), and the F1-score. These metrics provide a multi-faceted assessment of the spatial agreement between the predicted masks and the ground-truth annotations. Such a comprehensive evaluation is essential for crack segmentation, where models must accurately delineate irregularly shaped and fine-scaled structures.

Precision measures the proportion of true positive pixels relative to the total number of pixels classified as positive by the model. It serves as a critical indicator of the model's ability to suppress false positives (FP) and avoid over-detection in the background. Consequently, a higher Precision value signifies a lower rate of unnecessary or erroneous detections.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

Recall (or sensitivity) quantifies the proportion of actual ground-truth pixels correctly identified by the model. It serves as a primary metric for assessing the rate of false negatives (FN); a lower Recall indicates a higher frequency of omitted crack regions.

Precision and Recall typically exhibit an inverse relationship, forming a critical trade-off. In the context of crack segmentation, where the comprehensive detection of all structural defects is paramount to ensure safety, maximizing Recall is essential. However, to achieve a robust evaluation, both metrics must be considered simultaneously to ensure that high sensitivity does not come at the cost of excessive false positives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

IoU, also known as the Jaccard Index, is defined as the ratio of the intersection area between the predicted mask and the ground-truth mask to their total union area. It serves as a standard benchmark for evaluating the spatial overlap accuracy of a model. Unlike simple pixel accuracy, IoU provides a more robust assessment of segmentation performance by penalizing both over-segmentation (false positives) and under-segmentation (false negatives), making it particularly effective for measuring the localization precision of structural defects like cracks.

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{6}$$

The F1-score is the harmonic mean of Precision and Recall, providing a robust and comprehensive evaluation of the spatial agreement between predicted masks and ground-truth annotations. By integrating both FP and FN into a single metric, the F1-score effectively balances the penalties for over-segmentation and under-segmentation. This characteristic makes it particularly valuable for assessing crack detection performance, as it ensures that high sensitivity does not come at the expense of reliability. By utilizing these complementary metrics—Precision, Recall, IoU, and F1-score—this study provides a rigorous quantitative evaluation of the model's ability to achieve precise and stable segmentation of complex structural defects characterized by indistinct boundaries and irregular geometries.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

3. Results and Discussion

3.1. Comparative Analysis of Conditional Image Generation Models

This section evaluates the visual characteristics of synthetic images generated by SPADE, SemanticStyleGAN, ControlNet, and the proposed StyleSPADE. Additionally, semantic segmentation tasks were conducted using the Mask2Former architecture with a Swin-T backbone to quantitatively assess the impact of these synthetic datasets on model performance.

3.1.1. Qualitative Comparison: SPADE vs. StyleSPADE

Figure 7 presents a visual comparison between synthetic crack images generated by the baseline SPADE model and the proposed StyleSPADE.

As illustrated in Figure 7b, which utilizes the semantic mask in (a) as input, StyleSPADE effectively captures natural textures and chromatic variations across the entire background. This addressing of the "monotonous gray background" issue—a critical limitation of the original SPADE—enables the reproduction of environmental contexts that closely resemble authentic concrete surfaces. Furthermore, StyleSPADE produces cracks with higher edge sharpness and distinct morphologies, significantly enhancing visual clarity.

In contrast, the synthetic image generated by SPADE (Figure 7c), while maintaining a discernible crack structure, exhibits lower sharpness compared to StyleSPADE. The background remains a uniform, featureless gray, failing to encapsulate the stochastic diversity of real-world environments. Such lack of background complexity not only diminishes visual

realism but also poses a significant risk of overfitting when these images are utilized as training data for segmentation models.

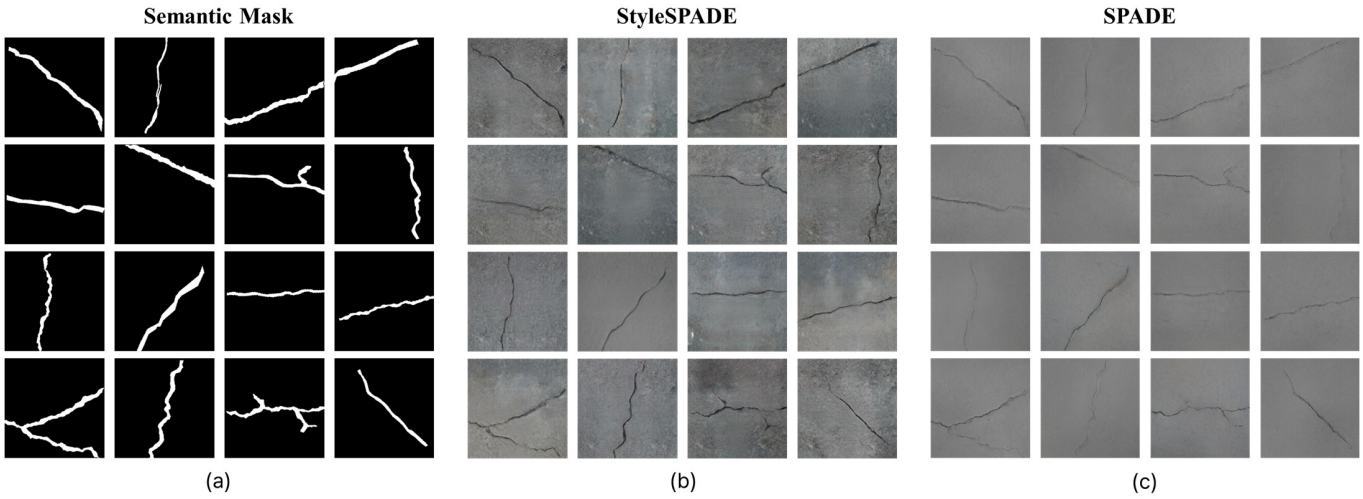

Figure 7. Visual comparison of synthetic crack images generated from identical semantic layouts: (**a**) input semantic masks; (**b**) results from the proposed StyleSPADE, demonstrating natural background textures and realistic chromatic variations; and (**c**) results from the baseline SPADE, showing the characteristic "monotonous gray background" limitation.

Consequently, StyleSPADE yields superior results by generating distinct crack shapes while simultaneously reflecting diverse background textures. These characteristics enhance the diversity of the synthetic dataset, which is expected to improve the robustness of downstream segmentation models and mitigate overfitting to simplified patterns.

3.1.2. Qualitative Evaluation of ControlNet

Figure 8 illustrates examples of synthetic crack images generated via ControlNet. The binary semantic masks used as conditional inputs are shown in Figure 8a,c, with their corresponding synthetic outputs presented in Figure 8b,d.

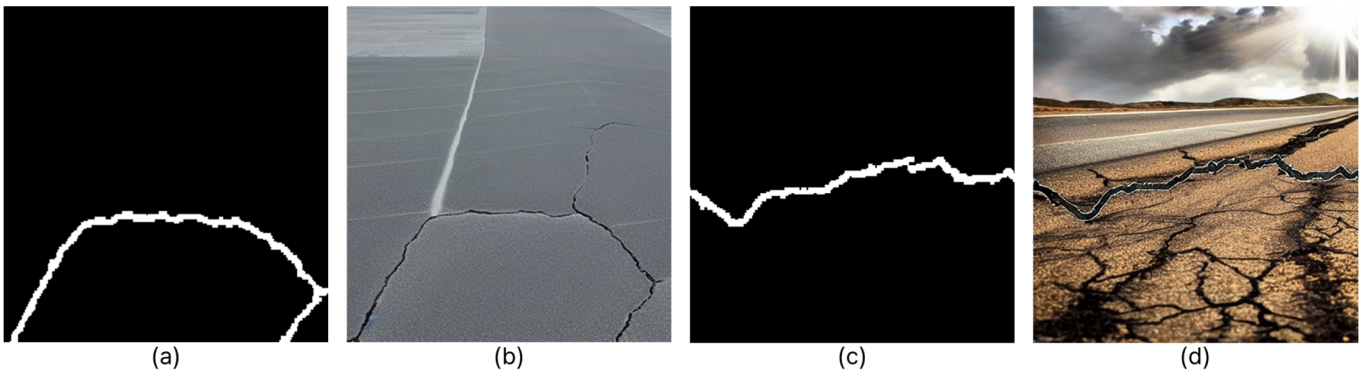

Figure 8. Synthetic crack images generated via ControlNet using the prompt "realistic crack on a road." (**a**,**c**) Input semantic masks; (**b**,**d**) Corresponding synthesized outputs. While the model adheres to the primary geometry of the input masks, it occasionally introduces auxiliary crack-like textures and complex background details to enhance visual realism, sometimes exceeding the explicit boundaries of the provided masks.

While ControlNet demonstrated the ability to synthesize high-quality images with diverse background textures, it exhibited a strong generative bias that diverged from the intended augmentation objective. Specifically, the model frequently introduced additional crack structures and artifacts that did not precisely align with the geometric constraints

of the input semantic masks. This tendency to prioritize generative diversity over strict conditional fidelity limits its utility for data augmentation, where precise pixel-level correspondence is required for training segmentation models. Due to this lack of spatial alignment between the conditional masks and the generated outputs, ControlNet was excluded from the final comparative analysis in this study.

Figure 9 provides a visual comparison between synthetic crack images generated by StyleSPADE and SemanticStyleGAN. The input semantic mask is presented in Figure 9a, with the corresponding synthetic outputs from SemanticStyleGAN and StyleSPADE shown in Figures 9b and 9c, respectively.

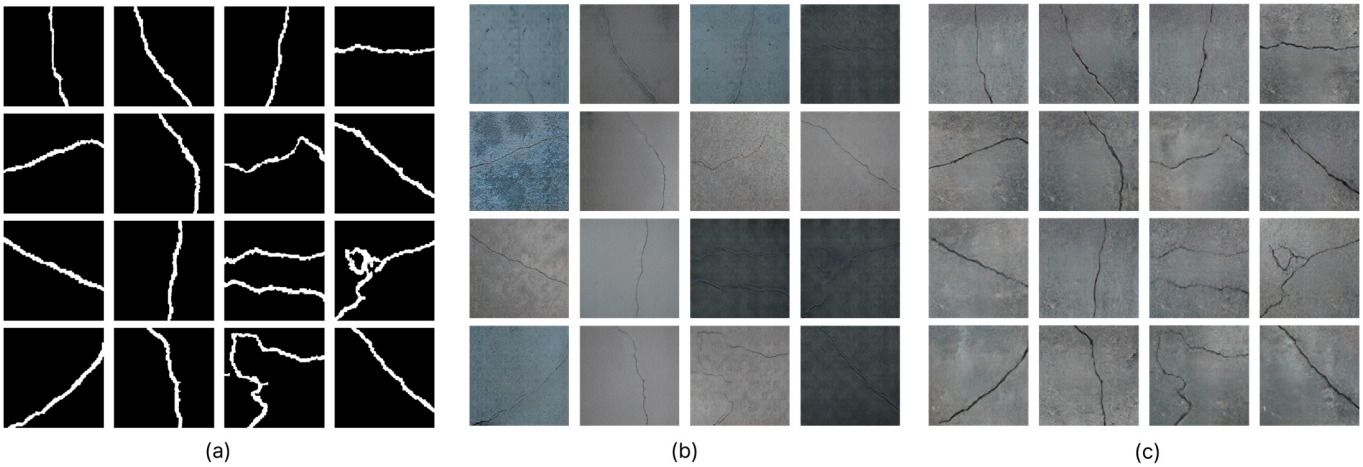

Figure 9. Visual comparison of synthetic crack images generated from identical semantic layouts: (**a**) input semantic masks; (**b**) results from SemanticStyleGAN, showing high background stylistic diversity but relatively indistinct crack features; and (**c**) results from the proposed StyleSPADE, demonstrating superior structural integrity and sharper crack morphologies.

SemanticStyleGAN synthesized images with a broader spectrum of chromatic and textural distributions compared to StyleSPADE, resulting in highly distinguishable background styles across different samples. However, the crack representations in the SemanticStyleGAN outputs were characterized by thin, blurred structures, which hindered visual identification—particularly against darker background regions. This suggests that while SemanticStyleGAN excels in region-specific style modulation, it possesses inherent limitations in reliably reproducing high-frequency crack details.

In contrast, while the background style variations produced by StyleSPADE were relatively more constrained than those of SemanticStyleGAN, the resulting crack morphologies were significantly sharper and more distinct. Furthermore, the transition between the crack boundaries and the surrounding concrete textures appeared more seamless and natural. These results indicate that StyleSPADE strikes a superior balance between background realism and the structural integrity of the target features.

3.1.3. Quantitative Evaluation of Conditional Generation Models

Table 4 summarizes the performance of segmentation models trained on synthetic data generated by each conditional image generation method. The baseline (Base) was established by training on the original 3717 images from the Crack-Seg dataset without augmentation. For comparative analysis, the training set was augmented with 3000 synthetic images from each respective method. To ensure a fair comparison, an identical augmentation budget was maintained across all models, thereby isolating the impact of generation quality and conditional fidelity from the confounding effects of dataset size. Furthermore, to account for training stochasticity, each experiment was repeated three

times under identical settings. Results are reported as mean ± standard deviation, with detailed trial-wise data provided in Appendix A (Table A1).

Table 4. Quantitative performance evaluation of crack segmentation models across different data augmentation methods, measured by IoU, Precision, Recall, and F1-score. Results are reported as mean ± standard deviation over three independent runs.

Setting	IoU	Precision	Recall	F1-Score
Base (No augmentation)	0.6149 ± 0.0043	0.7211 ± 0.0320	0.8377 ± 0.0293	0.7432 ± 0.0037
SemanticStyleGAN	0.6161 ± 0.0016	0.7580 ± 0.0197	0.7897 ± 0.0198	0.7418 ± 0.0014
SPADE	0.6224 ± 0.0026	0.7802 ± 0.0417	0.7833 ± 0.0541	0.7471 ± 0.0025
StyleSPADE (Ours)	0.6360 ± 0.0022	0.7617 ± 0.0214	0.8183 ± 0.0208	0.7584 ± 0.0013

Augmentation using SemanticStyleGAN yielded only marginal improvements in Crack IoU and F1-score compared to the baseline, with relatively small performance gains and comparable variability across repeated runs. This limited efficacy is attributed to the dilution of structural information during the fusion of crack and background latent representations, which often resulted in visually indistinct crack morphologies—particularly under dark or low-contrast background conditions—thereby constraining the model's ability to learn discriminative features.

In contrast, both SPADE and StyleSPADE augmentations resulted in substantial and consistent performance improvements. The SPADE-based model achieved a Crack IoU of 0.6224 ± 0.0026 and an F1-score of 0.7471 ± 0.0025, outperforming SemanticStyleGAN while maintaining relatively stable performance across repeated runs. This improvement is primarily attributed to SPADE's spatially adaptive normalization architecture, which effectively preserves geometric precision derived from semantic masks.

The proposed StyleSPADE-based augmentation achieved the highest performance across nearly all evaluation metrics, with a Crack IoU of 0.6360 ± 0.0022 and an F1-score of 0.7584 ± 0.0013. Notably, the low standard deviation across repeated trials indicates that the observed performance gains are robust and not attributable to random training fluctuations. Although StyleSPADE slightly lagged behind SPADE in Recall, its superior Precision and IoU demonstrate its ability to generate high-fidelity crack structures that enable reliable boundary delineation and morphological preservation. By overcoming the "monotonous gray-scale background" limitation inherent in the original SPADE, StyleSPADE introduces essential textural diversity, which effectively enhances generalization performance and reduces background-to-foreground confusion.

Figure 10 illustrates the qualitative performance of the segmentation models across different augmentation strategies, compared against the ground truth (GT) masks. Each column presents the original image, the GT mask, the Base model output (trained without augmentation), and the results from models augmented with SemanticStyleGAN, SPADE, and StyleSPADE.

In Case (a), where the GT contains a discontinuous crack structure, only the models augmented by SemanticStyleGAN and StyleSPADE accurately captured the discontinuities; other models tended to over-segment, producing erroneous continuous lines. In Case (b), the Base model exhibited poor structural continuity, omitting the central crack section. While the SPADE-augmented model showed minor over-detection along the boundaries, both the SemanticStyleGAN and StyleSPADE models maintained high fidelity to the original crack thickness and morphology.

In Case (c), although all models showed a general tendency toward over-segmentation, the StyleSPADE-augmented model was notably more effective at suppressing false positives and point-like noise. This trend continued in Cases (d) and (e), where models using

SemanticStyleGAN and SPADE exhibited severe over-detection of background textures unrelated to the actual cracks. In contrast, StyleSPADE significantly mitigated these artifacts, yielding only minimal, isolated false detections.

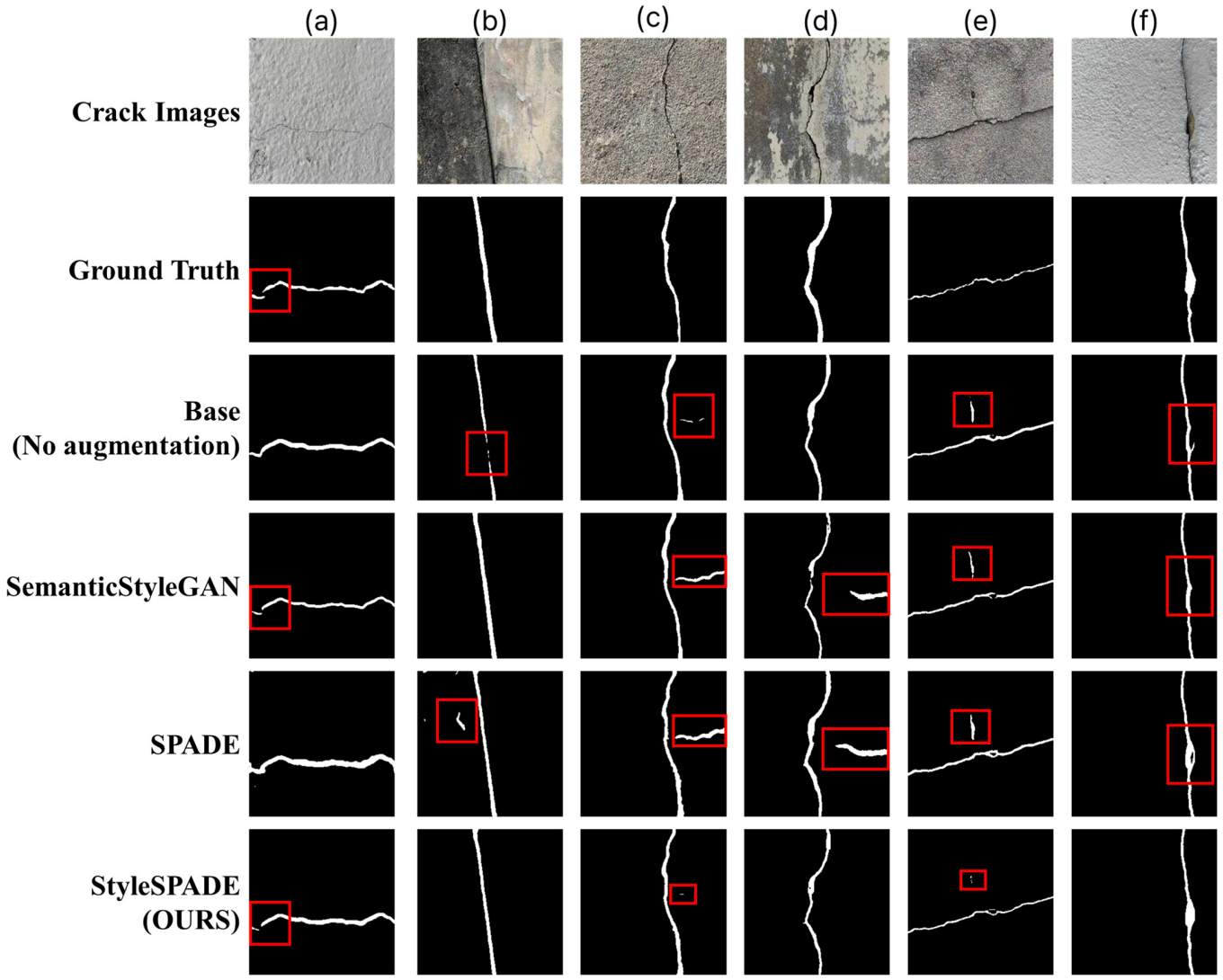

Figure 10. Qualitative comparison of segmentation performance across various augmentation strategies. Columns (**a**–**f**) represent distinct crack scenarios. From top to bottom, the rows display the input images, ground truth masks, and predictions from the Base model, SemanticStyleGAN, SPADE, and StyleSPADE. Red boxes highlight specific regions where the models show significant performance variations or refinement details.

Case (f) highlights a deep crack with a substantial width. The Base and SemanticStyleGAN models failed to represent the crack's volumetric features, treating it as a thin line. While SPADE partially captured the width, the StyleSPADE-augmented model achieved the highest correlation with the GT in terms of both shape and scale.

The experimental results demonstrate that the proposed StyleSPADE augmentation provides superior performance in capturing crack discontinuity, representing thick morphologies, and suppressing false positives. Consequently, StyleSPADE was selected as the optimal augmentation technique for this study. This performance advantage is attributed to StyleSPADE's ability to precisely preserve crack geometry via spatially adaptive normalization while simultaneously providing robust control over background variability through style vector injection.

3.2. Comparative Performance Analysis of Segmentation Architectures

Following the implementation of the StyleSPADE-based augmentation, which exhibited the most significant performance gains, we evaluated various segmentation models by varying their backbone architectures and applying ensemble strategies. The quantitative results are summarized in Table 5, while qualitative predictions are illustrated in Figure 11. Regarding the K-Net architecture, the ResNet-50 backbone yielded the lowest performance (Crack IoU: 0.4917), indicating structural limitations when pairing a CNN-based backbone with a CNN-based head for complex crack detection. However, performance improved substantially when utilizing the Swin Transformer backbone, with the Swin-L configuration achieving the highest overall results (Crack IoU: 0.6397, F1-score: 0.7607). Mask2Former also exhibited enhanced performance when transitioning from ResNet-50 to the Swin Transformer; notably, the Swin-T configuration achieved a Crack IoU of 0.6376, outperforming the more complex Swin-L variant. Swin-UPerNet maintained consistent performance across all scales, with the Swin-B configuration peaking at 0.6325 IoU.

Table 5. Quantitative comparison of ensemble configurations to validate the proposed design. The results evaluate the performance trade-offs between recall-focused, precision-focused, and balanced Avg-Logit strategies, highlighting the effectiveness of the balanced approach in optimizing the IoU and F1 score.

Ensemble Method	IoU	Precision	Recall	F1-Score
Recall-focused (Avg-Logit)	0.6406	0.8113	0.7768	0.7628
Precision-focused (Avg-Logit)	0.6412	0.8214	0.7629	0.7602
Balanced (Avg-Logit)	0.6437	0.8020	0.7904	0.7635

These comparative results indicate that increasing backbone scale (parameters) does not inherently improve the detection of thin, irregular crack patterns. This suggests that for crack segmentation, structural suitability and the capacity to extract fine-grained local features are more critical than raw model complexity.

To evaluate their respective performance characteristics, three Avg-Logit ensemble configurations—recall-focused, precision-focused, and balanced—were analyzed. This comparison facilitates a systematic assessment of the trade-offs between false negative and false positive suppression, depending on the ensemble bias. The quantitative details of these comparisons are documented in Table 5.

Among the three Avg-Logit ensemble strategies, the balanced configuration achieved the best overall performance, yielding a Crack IoU of 0.6437 and an F1-score of 0.7635. Based on this result, the balanced Avg-Logit ensemble was further extended using a Stacking Ensemble approach to assess the effectiveness of meta-learning–based fusion beyond simple averaging. The stacking ensemble achieved superior performance across all metrics (Crack IoU: 0.6452, Precision: 0.8056, Recall: 0.7968, F1-score: 0.7668), which is attributed to the U-Net meta-learner's ability to integrate complementary feature representations from the probability maps of the base models.

In conclusion, the stacking ensemble strategy proved more effective than simple averaging in mitigating the Precision–Recall trade-off and enhancing the overall robustness of crack detection. In the context of infrastructure maintenance (e.g., bridges and tunnels), even marginal gains in segmentation accuracy are meaningful. As shown in Table 6, the stacking ensemble achieved the highest overall results across all key metrics, including an F1-score of 0.7668. In safety-critical inspections, such consistent improvements are vital for reducing missed structural defects or false alarms.

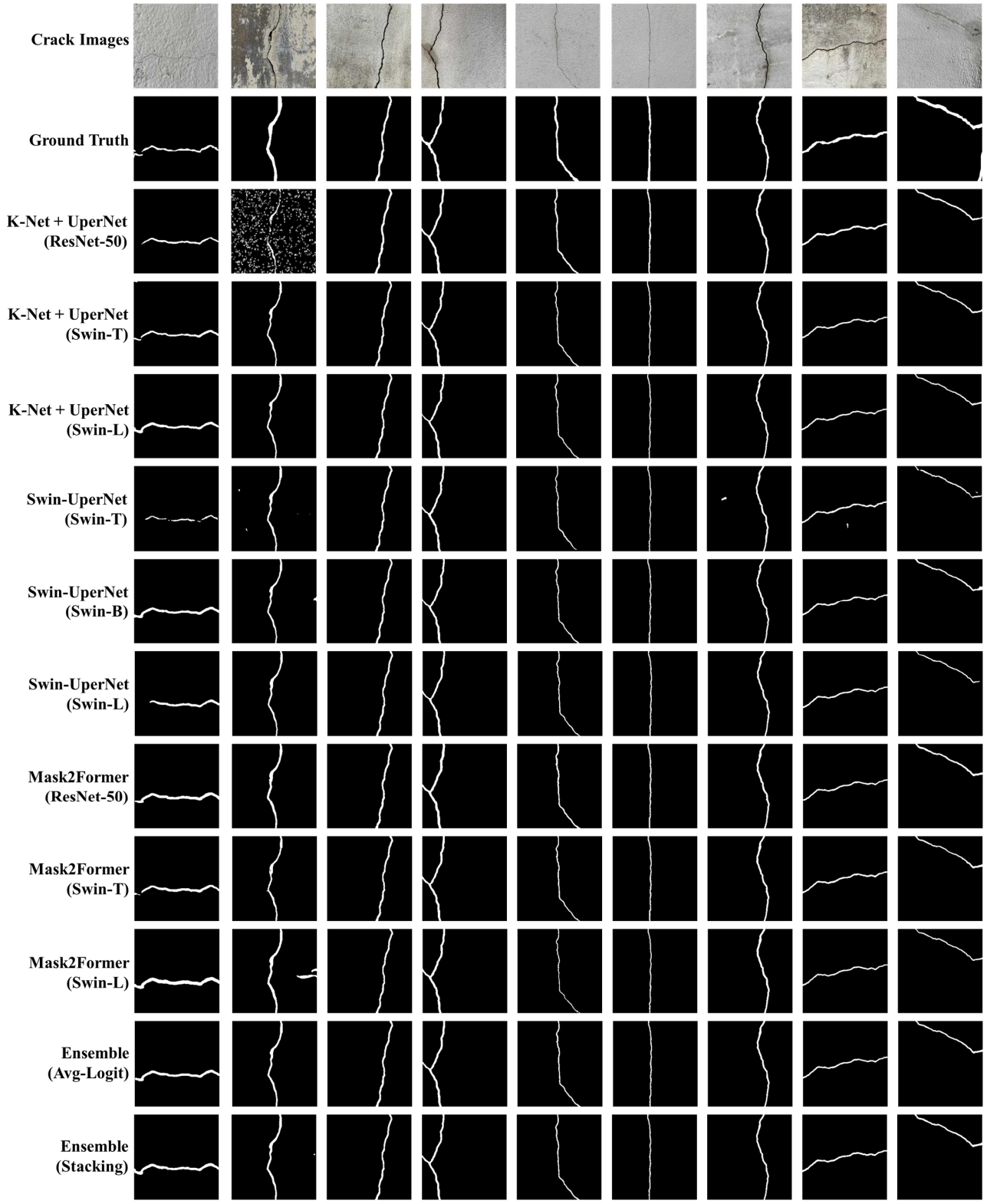

Figure 11. Qualitative comparison of crack segmentation results across individual models, various backbones, and ensemble strategies. The rows display predictions from single architectures (K-Net, UPerNet, Mask2Former) with varying backbones (ResNet, Swin), followed by the proposed Avg-Logit and Stacking ensembles. The Stacking ensemble demonstrates superior connectivity and boundary precision, particularly in complex crack patterns, compared to individual backbone-based models.

Table 6. Quantitative comparison of crack segmentation performance across various architectures and ensemble strategies. All models were trained using StyleSPADE-augmented data, with the Stacking Ensemble achieving the highest overall performance across most metrics, including IoU (0.6452) and F1-score (0.7668).

Method	Backbone	IoU	Precision	Recall	F1-Score
Knet + UPerNet	ResNet-50	0.4917	0.6350	0.6524	0.5944
UPerNet	Swin-T	0.6012	0.7823	0.7557	0.7296
Mask2Former	ResNet-50	0.6210	0.7560	0.8211	0.7471
UPerNet	Swin-L	0.6318	0.7957	0.7769	0.7543
UPerNet	Swin-B	0.6325	0.8022	0.7709	0.7553
Mask2Former	Swin-L	0.6331	0.7776	0.8052	0.7586
Knet + UPerNet	Swin-T	0.6351	0.8158	0.7599	0.7551
Mask2Former	Swin-T	0.6376	0.7773	0.8041	0.7586
Knet + UPerNet	Swin-L	0.6397	0.8208	0.7641	0.7607
Ensemble (Avg-Logit)	-	0.6437	0.8020	0.7904	0.7635
Ensemble (Stacking)	-	0.6452	0.8056	0.7968	0.7668

3.3. Computational Cost and Practical Deployment Considerations

From a practical deployment perspective, the proposed framework is intentionally designed to decouple computationally intensive components from inference-time operations. Since StyleSPADE is utilized exclusively for offline data augmentation and synthetic image generation during the training phase, it introduces no additional computational overhead during deployment. Once the models are trained, the inference pipeline remains identical to that of conventional single-model segmentation approaches. Regarding the ensemble strategies, the computational cost varies by the chosen fusion scheme. The Avg-Logit ensemble requires only simple probability averaging, incurring marginal overhead proportional to the number of base models. In contrast, the stacking ensemble incorporates a U-Net–based meta-learner, adding complexity equivalent to the meta-network itself. While stacking yields the highest accuracy, it remains an optional component primarily intended to demonstrate the performance upper bound under offline or batch-processing conditions. Representative computational costs for each stage are reported in Table 7, illustrating that the framework offers flexible trade-offs between accuracy and efficiency to accommodate diverse operational constraints.

Table 7. Computational complexity analysis of the proposed framework. The table reports measured training and inference costs for StyleSPADE, various segmentation backbones, and the meta-learner on an NVIDIA RTX 4090. These metrics demonstrate the feasibility of offline augmentation and efficient real-time inference for practical deployment.

Component	Phase	Metrics	Measured Cost
StyleSPADE	Training	Time per iteration	0.194 s
StyleSPADE	Generation	Time per image	0.067 s
K-Net/Swin-L	Training	Time per iteration	0.1308–0.1328 s
Mask2Former/ResNet-50	Training	Time per iteration	0.157–0.163 s
K-Net/Swin-L	Inference	Time per image	0.087 s
Mask2Former/ResNet-50	Inference	Time per image	0.185–0.198 s
U-Net (meta-learner)	Training	Time per iteration	0.185–0.198 s
U-Net (meta-learner)	Inference	Time per image	0.357 s

4. Conclusions

This study presents a data augmentation framework based on a conditional image generation model and a segmentation ensemble strategy designed to enhance crack detection in civil infrastructure. Traditional conditional generative models often struggle with representing fine crack morphologies, experiencing artifacts around mask boundaries

and producing monotonous backgrounds. To overcome these limitations, we proposed StyleSPADE, which integrates semantic masks and style images as dual inputs. By leveraging Spatially Adaptive Normalization (SPADE), StyleSPADE ensures the precise preservation of crack geometry while synthesizing realistic backgrounds with diverse textural and chromatic characteristics.

The segmentation model trained with StyleSPADE-augmented data achieved a Crack IoU of 0.6376 and an F1-score of 0.7586, outperforming both the baseline SPADE and SemanticStyleGAN. Significant improvements were observed in several critical areas, including the detection of discontinuous structures, the representation of deep/thick cracks, and the effective suppression of false positives. Furthermore, a comparative analysis of architectural configurations revealed that the K-Net/Swin-L combination achieved the highest individual performance (Crack IoU: 0.6397), confirming that backbone selection is a critical factor in segmentation accuracy. Finally, the implementation of a Stacking Ensemble strategy—combining high-recall and high-precision models—resulted in superior overall metrics (Crack IoU: 0.6452, F1-score: 0.7668), demonstrating its ability to mitigate the precision-recall trade-off.

The primary contributions of this research are fourfold:

- Task-Specific StyleSPADE Framework: We developed a style-conditioned adaptation of SPADE optimized for crack synthesis, demonstrating that additive affine-parameter modulation effectively balances geometric preservation with background diversity.
- Performance Validation: We empirically demonstrated that the proposed synthetic data significantly enhances the performance of downstream segmentation models.
- Optimal Configuration Identification: Through a systematic comparative study, we identified the most effective backbone-decoder combinations tailored for crack detection tasks.
- Ensemble Optimization: We proposed a practical stacking ensemble approach that simultaneously minimizes false negatives and false positives, achieving superior segmentation accuracy.

These findings suggest that robust crack detection is achievable even in data-constrained environments, potentially facilitating the early detection of structural damage and improving maintenance efficiency. While the proposed augmentation strategy has improved generalization across heterogeneous surface conditions, future research will extend this framework by training generative models on broader background subsets beyond concrete, such as asphalt and pavement. Additionally, we plan to conduct a systematic cross-domain evaluation to quantify transferability between material domains and identify optimal augmentation strategies for diverse infrastructure types. Furthermore, we intend to develop lightweight segmentation models for real-time applications and integrate automated crack-severity assessment. Finally, an ablation study on synthetic data volume (e.g., 1k–10k) will be performed to analyze performance saturation trends, providing practical guidance for selecting the optimal augmentation scale under varying computational budgets.

Author Contributions: Conceptualization, J.S., S.H.K. and Y.L.; methodology, J.S., S.H.K. and Y.L.; formal analysis, J.S.; data curation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, M.K., S.H.K. and Y.L.; supervision, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant (2021-MOIS37-002) from the Intelligent Technology Development Program on Disaster Response and Emergency Management funded by the Ministry of Interior and Safety (MOIS, Korea). This research was supported by the Regional Innovation System & Education (RISE) program through the Institute for Regional Innovation System & Education in

Busan Metropolitan City, funded by the Ministry of Education (MOE) and the Busan Metropolitan City, Republic of Korea (2025-RISE-02-001-009). This work was supported by the National Research Foundation (NRF), Korea, under project BK21 FOUR.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Detailed trial-wise results of crack segmentation across three independent runs for each data augmentation method. This table provides the underlying data used to calculate the mean and standard deviation reported in the main text, demonstrating the performance consistency of the proposed StyleSPADE method compared to other baselines.

Setting	Trial	IoU	Precision	Recall	F1-Score
Base (No augmentation)	Trial-1	0.6166	0.7544	0.8061	0.7426
	Trial-2	0.6181	0.7182	0.8429	0.7472
	Trial-3	0.6100	0.6906	0.8641	0.7399
SemanticStyleGAN	Trial-1	0.6171	0.7481	0.8042	0.7430
	Trial-2	0.6170	0.7807	0.7671	0.7421
	Trial-3	0.6143	0.7451	0.7978	0.7402
SPADE	Trial-1	0.6222	0.7326	0.8450	0.7488
	Trial-2	0.6199	0.8099	0.7436	0.7442
	Trial-3	0.6250	0.7982	0.7614	0.7483
StyleSPADE (Ours)	Trial-1	0.6376	0.7773	0.8041	0.7586
	Trial-2	0.6335	0.7373	0.8421	0.7570
	Trial-3	0.6368	0.7706	0.8086	0.7595

References

1. Alexander, L. *Safety Protocols in the Operation of Aging Infrastructure*; U.S. Department of Transportation, Federal Highway Administration: Washington, DC, USA, 2025.
2. Han, D. Heterogeneous deterioration process and risk of deficiencies of aging bridges for transportation asset management. *Sustainability* **2021**, *13*, 7094. [CrossRef]
3. Kim, H.J.; Seong, Y.H.; Han, J.W.; Kwon, S.H.; Kim, C.Y. Demonstrating the test procedure for preventive maintenance of aging concrete bridges. *Infrastructures* **2023**, *8*, 54. [CrossRef]
4. Gan, L.; Liu, H.; Yan, Y.; Chen, A. Bridge bottom crack detection and modeling based on faster R-CNN and BIM. *IET Image Process.* **2024**, *18*, 664–677. [CrossRef]
5. Harb, S.; Achanccaray, P.; Maboudi, M.; Gerke, M. Multi-temporal crack segmentation in concrete structure using deep learning approaches. *arXiv* **2024**, arXiv:2411.04620. [CrossRef]
6. Chen, B.; Zhang, H.; Wang, G.; Huo, J.; Li, Y.; Li, L. Automatic concrete infrastructure crack semantic segmentation using deep learning. *Autom. Constr.* **2023**, *152*, 104950. [CrossRef]
7. Fan, C.; Ding, Y.; Liu, X.; Yang, K. A review of crack research in concrete structures based on data-driven and intelligent algorithms. *Structures* **2025**, *75*, 108800. [CrossRef]
8. Qayyum, W.; Ehtisham, R.; Bahrami, A.; Mir, J.; Khan, Q.U.Z.; Ahmad, A.; Özkılıç, Y.O. Predicting characteristics of cracks in concrete structure using convolutional neural network and image processing. *Front. Mater.* **2023**, *10*, 1210543. [CrossRef]
9. Bah, A.S.; Zhang, Y.; Sasai, K.; Conciatori, D.; Chouinard, L.; Zufferey, N.; Power, G.J.; Sanchez, T.; Chen, X. Bridge service life and impact of maintenance events on the structural state index. *Case Stud. Constr. Mater.* **2025**, *22*, e04766. [CrossRef]
10. Hopper, T.; Manafpour, A.; Radlinska, A. *Bridge Deck Cracking: Effects on In-Service Performance, Prevention, and Remediation*; Pennsylvania Department of Transportation: Harrisburg, PA, USA, 2015.

11. Tomaszkiewicz, K.; Owerko, T. A pre-failure narrow concrete cracks dataset for engineering structures damage classification and segmentation. *Sci. Data* **2023**, *10*, 925. [CrossRef]
12. Kaveh, H.; Alhajj, R. Recent advances in crack detection technologies for structures: A survey of 2022–2023 literature. *Front. Built Environ.* **2024**, *10*, 1321634. [CrossRef]
13. Yuan, Q.; Shi, Y.; Li, M. A review of computer vision-based crack detection methods in civil infrastructure: Progress and challenges. *Remote Sens.* **2024**, *16*, 2910. [CrossRef]
14. Mondal, T.G.; Chen, G. Artificial intelligence in civil infrastructure health monitoring—Historical perspectives, current trends, and future visions. *Front. Built Environ.* **2022**, *8*, 1007886. [CrossRef]
15. Hamishebahar, Y.; Guan, H.; So, S.; Jo, J. A comprehensive review of deep learning-based crack detection approaches. *Appl. Sci.* **2022**, *12*, 1374. [CrossRef]
16. König, J.; Jenkins, M.; Mannion, M.; Barrie, P.; Morison, G. What's cracking? A review and analysis of deep learning methods for structural crack segmentation, detection and quantification. *arXiv* **2022**, arXiv:2202.03714.
17. Xiang, X.; Hu, H.; Ding, Y.; Zheng, Y.; Wu, S. GC-YOLOv5s: A lightweight detector for UAV road crack detection. *Appl. Sci.* **2023**, *13*, 11030. [CrossRef]
18. Ren, W.; Zhong, Z. Building construction crack detection with BCCD YOLO enhanced feature fusion and attention mechanisms. *Sci. Rep.* **2025**, *15*, 23167. [CrossRef] [PubMed]
19. Feng, H.; Li, W.; Luo, Z.; Chen, Y.; Fatholahi, S.N.; Cheng, M.; Wang, C.; Junior, J.M.; Li, J. GCN-based pavement crack detection using mobile LiDAR point clouds. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 11052–11061. [CrossRef]
20. Perez, F.; Vasconcelos, C.; Avila, S.; Valle, E. Data augmentation for skin lesion analysis. In Proceedings of the International Workshop on Computer-Assisted and Robotic Endoscopy, Granada, Spain, 16 September 2018; pp. 303–311. [CrossRef]
21. Forest, F.; Porta, H.; Tuia, D.; Fink, O. From classification to segmentation with explainable AI: A study on crack detection and growth monitoring. *Autom. Constr.* **2024**, *165*, 105497. [CrossRef]
22. Wang, Z.; Yang, J.; Jiang, H.; Fan, X. CNN training with twenty samples for crack detection via data augmentation. *Sensors* **2020**, *20*, 4849. [CrossRef]
23. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *Information* **2020**, *11*, 125. [CrossRef]
24. Tang, J.; Sharma, M.; Zhang, R. Explaining the effect of data augmentation on image classification tasks. *J. Imag.* **2020**, *6*, 62.
25. Taylor, L.; Nitschke, G. Improving deep learning with generic data augmentation. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Bengaluru, India, 18–21 November 2018; pp. 1542–1547. [CrossRef]
26. Yu, G.; Zhou, X.; Chen, X. VDCrackGAN: A generative adversarial network with transformer for pavement crack data augmentation. *Appl. Sci.* **2024**, *14*, 7907. [CrossRef]
27. Zhang, K.; Zhang, Y.; Cheng, H.D. CrackGAN: Pavement crack detection using partially accurate ground truths based on generative adversarial learning. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 1306–1319. [CrossRef]
28. Yao, H.; Liu, Y.; Besklubova, S.; Brilakis, I.; Guo, M.; Wang, J.; Wang, M. A data augmentation method for pavement crack detection based on super-resolution and denoising diffusion probabilistic models. *Comput. Aided Civ. Infrastruct. Eng.* **2025**, *40*, 4212–4225. [CrossRef]
29. Li, J.; Ma, X.; Shi, Y. Multi-Label Conditioned Diffusion for Cardiac MR Image Augmentation and Segmentation. *Bioengineering* **2025**, *12*, 812. [CrossRef] [PubMed]
30. Filipovski, D.; Gievska, S. SemanticStyleGAN: Generative Image Inpainting Using Style-Based Generator. In Proceedings of the International Conference on ICT Innovations, Skopje, North Macedonia, 12 April 2022; pp. 38–51. [CrossRef]
31. Deng, B.; Lu, Y. Weed image augmentation by controlnet-added stable diffusion for multi-class weed detection. *Comput. Electron. Agric.* **2025**, *232*, 110123. [CrossRef]
32. Tan, Z.; Chen, D.; Chu, Q.; Chai, M.; Liao, J.; He, M.; Yuan, L.; Hua, G.; Yu, N. Efficient semantic image synthesis via class-adaptive normalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4852–4866. [CrossRef]
33. Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A.; Keutzer, K.; Gong, B. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2100–2110. [CrossRef]
34. Peng, D.; Lei, Y.; Liu, L.; Zhang, P.; Liu, J. Global and local texture randomization for synthetic-to-real semantic segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 6594–6608. [CrossRef] [PubMed]
35. Crack Segmentation Dataset. Available online: https://www.kaggle.com/datasets/lakshaymiddha/crack-segmentation-dataset (accessed on 26 July 2025).
36. Ham, S.; Bae, S.; Kim, H.; Lee, I.; Lee, G.; Kim, D. Training of semantic segmentation model for crack analysis of tunnel concrete lining. *J. Korean Tunn. Undergr. Space Assoc.* **2021**, *23*, 549–558. [CrossRef]
37. Crack Computer Vision Model. Available online: https://universe.roboflow.com/university-bswxt/crack-bphdr (accessed on 26 July 2025).

38. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic Image Synthesis with Spatially-Adaptive Normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2337–2346. [CrossRef]
39. Wang, P.; Li, Y.; Vasconcelos, N. Rethinking and Improving the Robustness of Image Style Transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 124–133. [CrossRef]
40. Shi, Y.; Yang, X.; Wan, Y.; Shen, X. SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11244–11254. [CrossRef]
41. Zhang, L.; Rao, A.; Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 3836–3847. [CrossRef]
42. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-Attention Mask Transformer for Universal Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1280–1289. [CrossRef]
43. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022. [CrossRef]
44. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified Perceptual Parsing for Scene Understanding. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11209, pp. 432–448. [CrossRef]
45. Zhang, W.; Pang, J.; Chen, K.; Loy, C.C. K-Net: Towards Unified Image Segmentation. In *Advances in Neural Information Processing Systems*; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 10326–10338.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.