CT2BSE: 3D BSE Microstructural Image Cross-Device Generation from μCT for Cement Hydration via Voxel Swin Transformer

Haozhong Gao[1,2], **Liangliang Zhang**[2,1*], **Yamin Han**[2], **Ruiqi Han**[1], **Lin Wang**[1,2] and **Bo Yang**[2,1]
[1]Shandong Key Laboratory of Ubiquitous Intelligent Computing, University of Jinan, Jinan 250022, China
[2]Quancheng Laboratory, Jinan 250100, China
ise_zhangll@ujn.edu.cn

Abstract

Acquiring three-dimensional (3D) microstructural images of cement hydration reveals critical microscale features essential for understanding hydration mechanisms and advancing material development. Micro-computed tomography (μCT) is widely used to capture these images due to its non-destructive, repeatable 3D imaging capability, despite at high operational cost. However, μCT suffers from limited resolution, weak texture, and missing phase information. In contrast, backscattered electron (BSE) provides high-resolution, phase-sensitive, and texture-rich images but are confined to 2D form. Inspired by the idea of fusing BSE style into μCT images for computationally imaging 3D BSE data, this paper proposes a cross-device generation method, termed CT2BSE, to construct 3D BSE microstructural images from μCT for cement hydration. A Voxel Swin Transformer is proposed to extract 3D μCT features by improving Video Swin Transformer with a Content-Aware Positional Encoding–3D, making encoder better suited for homogeneous images such as cement. Furthermore, a Holistic Style Injector–3D is proposed to fuse 3D μCT features with 2D BSE characteristics, thereby mitigating feature instability caused by the extreme intensity variations typically exhibited in cement images. Subsequently, the fused features are decoded into 3D volumes aligned with the BSE style while preserving original microstructure. Experimental results demonstrate that CT2BSE can generate high-fidelity 3D BSE images from μCT.

1 Introduction

3D microstructural images are widely used in the study of cement[Gallucci *et al.*, 2007] that is one of the most extensively used functional materials worldwide and the cornerstone of modern construction. They can intuitively characterize the spatial distribution and morphological characteristics of different phases at the micro-scale[Vigor *et al.*, 2022], which helps to understand hydration mechanisms deeply, thereby supporting the development and optimization of advanced cementitious materials. Higher-quality 3D microstructural images can provide more reliable structural and scale information for investigating complex hydration processes involving temporally evolving chemical reactions and physical changes. Consequently, the acquisition of high-quality 3D microstructural images for cement hydration is regarded as a critical prerequisite for cement research.

μCT is the most popular device for acquiring 3D microstructural images of cement hydration [Chung *et al.*, 2019]. The characteristic of this device is 3D imaging, and it does not require drying or vacuum dehydration of the sample, avoiding the destruction of samples. Moreover, μCT enables repeated scanning of the same sample owing to its non-destructive nature, making it particularly suitable for temporal research during cement hydration (up to 28 days).

However, the inherent flaws of μCT cannot be ignored, while μCT offers considerable advantages for 3D structure acquisition. Due to its imaging mechanism, μCT suffers from limited image quality, such as insufficient spatial resolution and a lack of texture-rich contrast, as well as a lack of explicit phase information, leading to incomplete 3D characterization.

In contrast, as another imaging device, BSE can provide high-resolution, texture-rich representations of the sample by reflecting the variations in the sample's ability to scatter the electron beam. More importantly, when combined with an energy-dispersive X-ray spectrometer (EDS), BSE allows the determination of phase distributions, but BSE typically provides only 2D images.

Considering the respective strengths of μCT in nondestructive 3D microstructure acquisition and BSE in producing high-quality 2D images with phase information, a question that arises here is whether we can fuse BSE style into μCT images and further cross-device generate 3D BSE images through computational imaging to compensate for the limitations of physical imaging from device itself. This would hold significant value for advancing research on cement hydration and, more broadly, for extending the imaging capabilities of devices through computational imaging.

Recent research has demonstrated that style transfer methods are capable of transforming images across different domains [Zhang *et al.*, 2025; Deng *et al.*, 2022; Gatys *et*

*Corresponding author: Liangliang Zhang

al., 2016]. However, most existing approaches are primarily designed for 2D content and style representations. Although video style transfer can be regarded as operating on 3D inputs [Wang *et al.*, 2020; Huang *et al.*, 2017; Ruder *et al.*, 2016], it remains fundamentally constrained by temporal inconsistency issues. Moreover, homogeneous images, such as cement microstructure, exhibit repetitive patterns and fractal-like self-similarity, which further complicate structural preservation and make model training challenging during the style transfer process. Therefore, overcoming dimensional limitations to achieve image-domain translation with structural preservation and without temporal inconsistency is of significant theoretical and practical importance.

This paper proposes a 3D BSE microstructural image cross-device generation method from μCT (i.e., CT2BSE) for cement hydration via Voxel Swin Transformer. Specifically, a Voxel Swin Transformer incorporating CAPE-3D is proposed to enable more effective feature extraction from 3D μCT volumes for cement hydration, in which CAPE-3D replaces the conventional 3D relative position bias. Subsequently, a Holistic Style Injector-3D (HSI-3D) is proposed to fuse the 3D μCT features with the 2D BSE characteristics. Finally, the fused features are decoded into 3D volumes that are aligned with the BSE style while preserving the original microstructure through the optimization of composite content and style losses. The main contributions of this paper are summarized as follows:

- For the first time a 3D BSE microstructural image cross-device generation method is proposed based on μCT and computational imaging.
- A Voxel Swin Transformer with CAPE-3D is proposed as an efficient 3D feature extractor, which addresses structural preservation in homogeneous images and alleviates the temporal inconsistency encountered in conventional frame-wise streaming processing.
- For the first time, 3D BSE images of cement hydration microstructure are constructed by transferring BSE style into μCT data, yielding high-fidelity and texture-rich 3D microstructural representations for cementitious materials.
- The HSI-3D is proposed to fuse multidimensional feature information and alleviate feature instability caused by the extreme intensity variations typically exhibited in cement images.

2 Related Works

2.1 Style Transfer for Image and Video

Since Gatys *et al.* proposed neural style transfer using convolutional neural networks[Gatys *et al.*, 2016], the field of arbitrary style transfer has experienced rapid development. A variety of generative-model-based approaches have subsequently been proposed, including methods based on variational autoencoders (VAEs) [Huang *et al.*, 2023], generative adversarial networks (GANs) [Azadi *et al.*, 2018; Xu *et al.*, 2021; Mukherkjee *et al.*, 2022], Transformers [Wu *et al.*, 2021; Deng *et al.*, 2022], and, more recently, diffusion models [Zhang *et al.*, 2023; Wang *et al.*, 2023; Chung *et al.*, 2024].

Ruder *et al.* first proposed video style transfer [Ruder *et al.*, 2016], which can be regarded as a natural extension of 2D style transfer to sequential data. However, the frame-by-frame processing method will lead to temporal inconsistency issues. Consequently, subsequent studies have focused on enforcing temporal consistency through optical flow, temporal regularization, or spatiotemporal modeling [Wang *et al.*, 2020; Huang *et al.*, 2017]. These approaches have demonstrated strong capabilities in modeling complex style distributions and generating visually plausible results across diverse image domains.

2.2 Image-Based Cement Research

Image-based approaches are extensively employed in cement research[Saseendran *et al.*, 2024; Kim *et al.*, 2024; Wang *et al.*, 2025], encompassing both μCT and BSE imaging modalities. BSE images are commonly used in conjunction with EDS to analyze phase information[Wang *et al.*, 2025; Mao *et al.*, 2025]. μCT images are used for in situ observation and performance evaluation[Miarka *et al.*, 2024; Kim *et al.*, 2024]. Both imaging modalities provide essential data foundations for cement research.

3 Method

3.1 Overall Framework

The overall architecture of CT2BSE is illustrated in Fig. 1, following an encoder–fusion–decoder paradigm. The μCT volume $I_c \in \mathbb{R}^{B\times C\times D\times H\times W}$ serves as the content input, with the Voxel Swin Transformer efficiently extracting volumetric content features. The BSE image $I_s \in \mathbb{R}^{B\times C\times H'\times W'}$ serves as the style input, and a Vision Transformer encoder is employed to obtain 2D style features.

For multidimensional feature fusion, the HSI-3D [Zhang *et al.*, 2025] is proposed. The extracted 3D feature map $F_c \in \mathbb{R}^{B\times C\times \frac{D}{8}\times \frac{H}{8}\times \frac{W}{8}}$ and the 2D feature map $F_s \in \mathbb{R}^{B\times C\times \frac{H'}{8}\times \frac{W'}{8}}$ are fused through this novel attention-based transformation module with linear computational complexity. This module aligns and merges the content and style representations, producing the fused feature $F_o \in \mathbb{R}^{B\times C\times \frac{D}{8}\times \frac{H}{8}\times \frac{W}{8}}$.

A 3D convolution-based decoder, equipped with volumetric upsampling, subsequently reconstructs the output at the original spatial resolution, enabling stable cross-device translation.

3.2 Voxel Swin Transformer

The 3D images of cement hydration microstructure exhibit two fundamental characteristics. First, their volumetric size is extremely large, which imposes substantial memory and computational burdens on models based on conventional 3D convolutions or attention mechanisms. Second, the inherent homogeneity within the hydrated cement microstructure increases the difficulty of effective feature extraction. Consequently, efficient 3D representation learning for such data

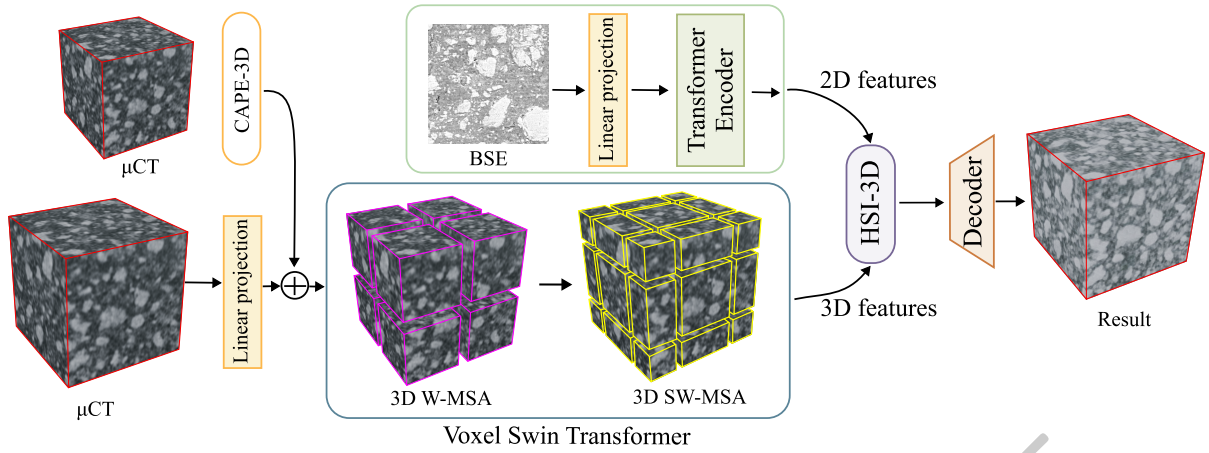


Figure 1: The overall architecture of CT2BSE. The 3D μCT volume is first linearly projected into an embedding space, after which CAPE-3D is added and the features are processed by the proposed Voxel Swin Transformer to extract volumetric features. Meanwhile, the 2D BSE image is encoded using a standard Vision Transformer. The resulting 3D and 2D features are then fused through the proposed HSI-3D, after which a decoder reconstructs the final output. The figure also highlights the most essential operations within the Voxel Swin Transformer block, namely the window partitioning and window shifting mechanisms.

must simultaneously ensure computational feasibility and robustness to structural redundancy.

The Swin Transformer[Liu *et al.*, 2021] is an improved variant of the Transformer architecture that discards global self-attention in favor of computing attention within predefined local windows. Its central idea is to restrict self-attention to these partitioned windows rather than performing global computation, thereby substantially reducing the overall computational complexity. The introduction of shifted windows further enables information exchange across neighboring regions, allowing the model to approximate the effect of global attention while maintaining significantly lower resource requirements. The Video Swin Transformer expands the architecture into a 3D version, enabling the extraction of features from 3D inputs.

As illustrated in Figs. 2 and 3, the Voxel Swin Transformer has the following key characteristics. First, it adopts the 3D shifting-window mechanism from the Video Swin Transformer [Liu *et al.*, 2022], enabling self-attention computation within local 3D regions and significantly reducing computational cost. Second, CAPE-3D is designed to replace the original 3D relative position bias, thereby preserving the structural skeleton before and after the enhancement. Third, the patch merging layers used in the Video Swin Transformer are removed to ensure that the number of tokens remains constant throughout the entire feature extraction process.

$$\begin{aligned}
\hat{\mathbf{z}}^{l} &= \text{3DW-MSA}\left(\text{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1}, \\
\mathbf{z}^{l} &= \text{FFN}\left(\text{LN}\left(\hat{\mathbf{z}}^{l}\right)\right) + \hat{\mathbf{z}}^{l}, \\
\hat{\mathbf{z}}^{l+1} &= \text{3DSW-MSA}\left(\text{LN}\left(\mathbf{z}^{l}\right)\right) + \mathbf{z}^{l}, \\
\mathbf{z}^{l+1} &= \text{FFN}\left(\text{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1},
\end{aligned} \tag{1}$$

where Z^{l-1} denotes the input to the $(l-1)$-th Swin Transformer-3D block, LN represents Layer Normalization, and 3D W-MSA refers to the 3D window-based multi-head self-attention module. The 3D SW-MSA indicates the multi-head self-attention computed using shifted 3D windows. FFN denotes the feed-forward network.

3.3 CAPE-3D

Content-Aware Positional Encoding, originally introduced in StyTR-2 [Deng *et al.*, 2022], is a positional encoding mechanism that has been shown to effectively preserve structural layouts in various style transfer tasks. For cement hydration microstructure, which is characterized by strong homogeneity, training instability frequently occurs, including mode collapse and undesirable alterations of the underlying skeletal structure. Such deviations may result in enhanced images that are no longer regarded as originating from the same physical object. To address this issue, the 3D relative position bias employed in the original Video Swin Transformer is discarded in the Voxel Swin Transformer, while CAPE is extended to a 3D formulation and incorporated instead. The computation proceeds as follows:

$$\begin{aligned}
\mathcal{P}_{\mathcal{L}} &= \mathcal{F}_{\text{pos}}(\text{AvgPool}_{n\times n\times n}(\mathcal{E})), \\
\mathcal{P}_{\mathcal{CA}}(x,y,z) &= \sum_{j=0}^{s}\sum_{k=0}^{s}\sum_{l=0}^{s}(a_{jkl}\mathcal{P}_{\mathcal{L}}(x_j,y_k,z_l)),
\end{aligned} \tag{2}$$

where $\mathcal{E}$ denotes the input 3D tensor, $\text{AvgPool}_{n\times n\times n}$ represents the adaptive 3D average pooling operator, and $\mathcal{F}_{\text{pos}}$ denotes the 1×1 convolution used for positional feature transformation. The resulting positional encoding is subsequently upsampled to the original spatial dimensions through interpolation.

3.4 HSI-3D

Zhang *et al.* proposed the Holistic Style Injector [Zhang *et al.*, 2025], a 2D feature fusion mechanism that replaces the standard QK matrix multiplication with a Dual Relation Construction operator. This design mitigates over-attention to localized style regions, improves visual consistency, and reduces computational complexity to linear scale.

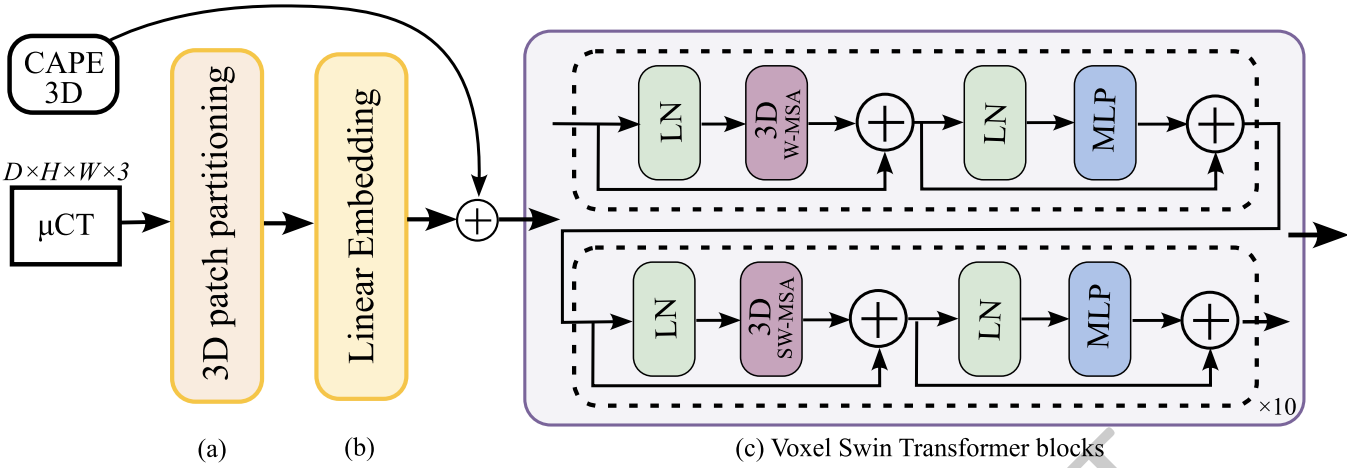


Figure 2: The architecture of the proposed Voxel Swin Transformer. The input 3D μCT volume is first divided into non-overlapping windows through 3D patch partitioning. Each patch is then linearly projected into tokens, after which the CAPE-3D is added. The encoded tokens are subsequently processed by ten layers of Voxel Swin Transformer blocks. Each block contains one 3D W-MSA and one 3D SW-MSA operation, enabling the model to approximate global attention at substantially reduced computational cost.

As illustrated in Fig. 3, the feature fusion requirement in this study involves aligning 3D content representations with 2D style representations. To this end, HSI is adapted into a 3D-compatible formulation, denoted as HSI-3D, with architectural modifications that enable its operation on volumetric inputs. Specifically, F_c denotes the 3D feature representation and F_s denotes the 2D feature representation. The query, key, and value tensors are constructed as $Q = \mathrm{Conv}(\mathrm{Norm}(F_c))$, $K = \mathrm{Conv}(\mathrm{Norm}(F_s))$, and $V = \mathrm{Norm}(F_s)$, respectively. Global context descriptors are obtained by $Q_c = \mathrm{AvgPooling}(Q)$ and (K_s, V_s) are produced by applying the global style aggregator GSA to K and V, respectively. The overall computation is formulated as follows:

$$F_{qk} = \lambda_g \times (Q_c \odot K_s) \oplus (1 - \lambda_g) \times (Q \odot K_s), \tag{3}$$

where $\odot$ and $\oplus$ denote element-wise multiplication and element-wise addition with broadcasting, respectively. Unlike the original HSI formulation, the definition of the fusion coefficient λ_g for integrating 3D features is modified as follows:

$$\lambda_g = \frac{\frac{Q_c \cdot K_s}{\|Q_c\| \|K_s\|} + 1}{2}. \tag{4}$$

In the Global Style Aggregation (GSA) module, four color moment statistics of the input features are computed and subsequently integrated through a dynamic network [Jia *et al.*, 2016], which aggregates these moment descriptors into a unified style representation.

3.5 Loss Function

Content loss and style loss are widely used objectives in 2D style transfer, typically constructed from feature activations extracted at multiple layers of a pre-trained VGG network[Simonyan and Zisserman, 2014]. In this study, the generated 3D output is randomly sliced along the three orthogonal axes. For each sampled slice, the content loss is computed against the corresponding slice of the μCT image, while the style loss is computed against the BSE image. The detailed formulation is given as follows:

$$\mathcal{L}_c = \frac{1}{N_l} \sum_{i=0}^{N_l} \|\phi_i(I_c) - \phi_i(I_o)\|_2, \tag{5}$$

where I_c denotes a slice extracted from the μCT volume, and I_o denotes the corresponding slice extracted from the generated 3D output. $\phi_i(\cdot)$ denotes features extracted from the i-th layer in a pretrained VGG and N_l is the number of layers.

$$\begin{aligned} \mathcal{L}_s = \frac{1}{N_l} \sum_{i=0}^{N_l} &\|\mu(\phi_i(I_s)) - \mu(\phi_i(I_o))\|_2 \\ &+ \|\sigma(\phi_i(I_s)) - \sigma(\phi_i(I_o))\|_2, \end{aligned} \tag{6}$$

where I_s denotes the 2D BSE image. $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and variance of extracted features, respectively.

In addition to the full-slice content and style losses, CT2BSE is designed to enhance the fidelity of fine-scale textural details in the generated results. To this end, for each randomly sampled slice, two additional cropped regions—corresponding to $\frac{1}{4}$ and $\frac{1}{8}$ of the original slice size—are extracted at randomly selected locations [Galerne *et al.*, 2025]. Content and style losses are then computed using the same cropped locations from the μCT slice and the BSE image. The overall loss formulation is summarized as follows:

$$\mathcal{L} = \lambda_i \sum_{i=0}^{n_i - 1} \mathcal{L}_s(I_o^{\downarrow 2^i}, I_s^{\downarrow 2^i}) + \mathcal{L}_c(I_o^{\downarrow 2^i}, I_c^{\downarrow 2^i}), \tag{7}$$

where λ_s denotes the weighting coefficient associated with each spatial scale, and $\downarrow 2^i$ indicates that the image is randomly cropped to a region whose size is reduced to $\frac{1}{2^i}$ of the original slice.

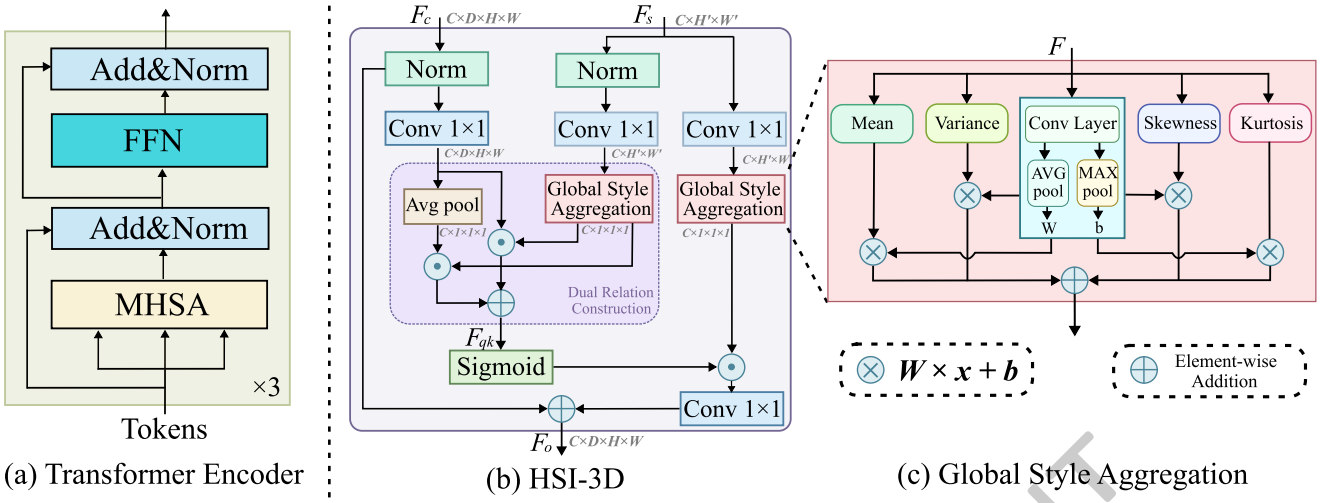


Figure 3: Subfigure (a) illustrates the conventional Transformer encoder architecture applied to the BSE input. Subfigure (b) presents the computational framework obtained by extending HSI into its 3D formulation. Subfigure (c) shows the Global Style Aggregation (GSA) module within HSI-3D, where a dynamic network is employed to combine the four color moments of the input features through adaptive mixing operations.

4 Experiments

4.1 Settings

The training and evaluation of the proposed model require two data modalities: μCT volumes and BSE images. In 2022, a set of high-quality μCT scans of cement hydration microstructure was acquired in the laboratory, with a voxel resolution of $3.5\,\mu\mathrm{m/pixel}$. Each volume has a spatial resolution of $1305 \times 1305 \times 1164$. Corresponding BSE surface images of the same cement specimens were also collected at magnification levels of $35\times$, $100\times$, $200\times$, $400\times$, and $600\times$.

During training, two μCT volumes were randomly selected and cropped into 128 sub-volumes of size $256 \times 256 \times 256$. Among them, three sub-images are randomly selected as test samples. For the BSE modality, all images captured at $200\times$ magnification were used, each cropped into patches of size 256×256. Three BSE patches covering a range from darker to brighter tonal characteristics were selected as the test set. The model was implemented using the PyTorch framework and optimized with the Adam optimizer. Training was performed on an NVIDIA A100 GPU (80 GB), requiring approximately 12 hours to achieve satisfactory convergence.

4.2 Results of CT2BSE on Different Samples

Fig. 4 presents the combinatorial results obtained from three μCT test samples and three BSE test samples. As expected, because the selected BSE images are arranged from dark to bright, the corresponding outputs exhibit a consistent progression in brightness. This demonstrates that the style characteristics inherent to the BSE modality are effectively transferred to the μCT images.

The enhanced results effectively mitigate common μCT artifacts, including local blurring and ambiguous phase boundaries, while substantially improving structural clarity and inter-phase contrast. Compared with the raw μCT images, the generated results exhibit clearer particle interiors, more

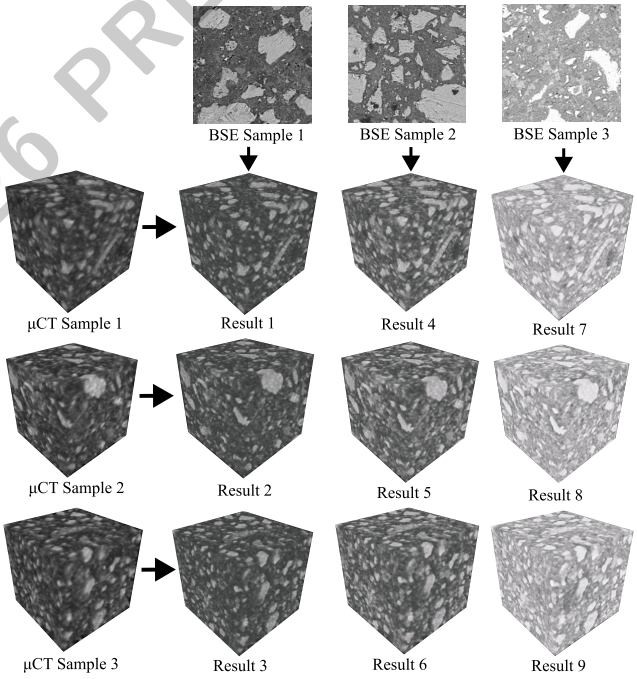


Figure 4: The results obtained by feeding all pairwise combinations of the three 3D μCT samples and the three 2D BSE images into CT2BSE. As the tonal intensity of the BSE images increases, the corresponding outputs derived from the same μCT sample exhibit a consistent brightening trend, demonstrating that the model effectively transfers the tonal characteristics of the BSE inputs. Moreover, the stylized outputs maintain strong alignment with the visual style of the respective BSE images. At the same time, the structural integrity of the μCT data is well preserved, with the underlying microstructural skeleton remaining highly consistent with the original volumetric input.

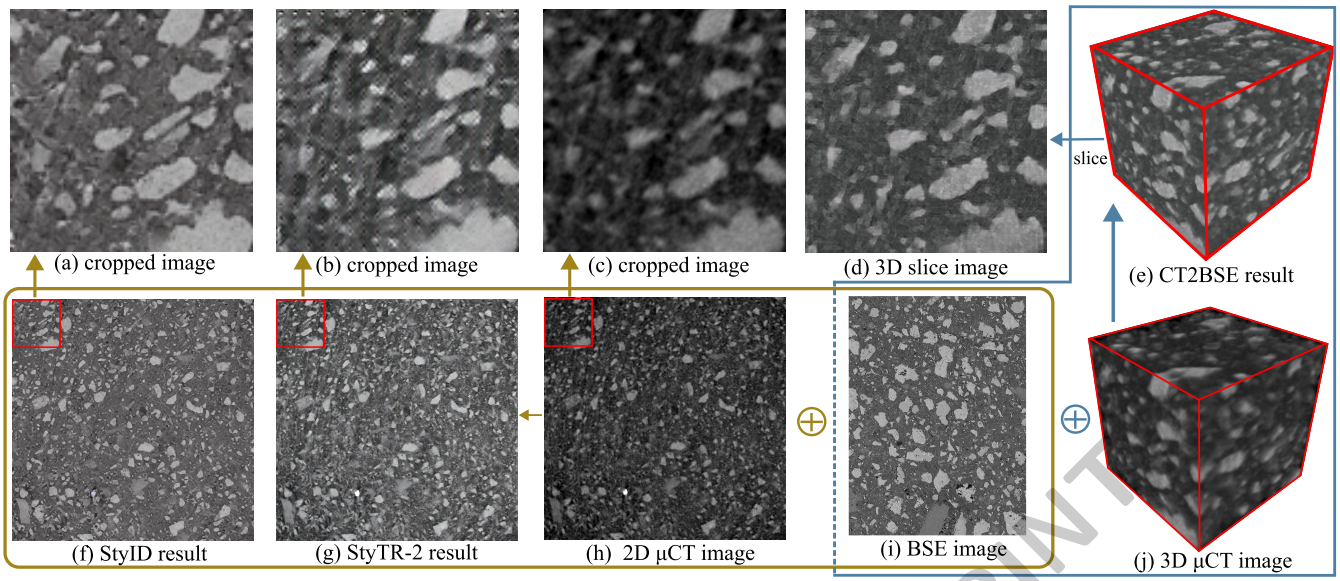


Figure 5: Comparison of 2D slices from CT2BSE with StyTR-2 and StyleID. Given the 2D μCT image in subfigure (h) and the BSE image in (i), the baselines generate the results in (f) and (g). In contrast, CT2BSE combines the 3D μCT volume shown in Fig. (j) with the same BSE image to generate the 3D volume illustrated in Fig. (e), from which a surface slice is extracted and presented in Fig. (d). Subfigures (a), (b), and (c) spatially correspond to the cropped regions in (f), (g), and (h), with (d) and the top-left region of (h) representing the same location.

pronounced unhydrated clinker phases, and more reliable phase discrimination. Meanwhile, the outputs closely resemble BSE images in tonal distribution and texture, consistently preserving characteristic granular features and bright-particle responses, leading to a more vivid and well-defined representation of the cement hydration microstructure.

4.3 Comparison with 2D Methods

In the 2D baseline setting, StyTR-2 and StyleID are selected. As shown in Fig. 5, StyTR-2 and StyleID take a single 2D μCT image as input, whereas CT2BSE processes a full 3D microstructural volume that contains the corresponding 2D surface. Using the same BSE image as the style reference for both methods, subfigures (f) and (g) present baselines stylized outputs and subfigures (e) present CT2BSE output. For a fair comparison, a cropped region from the upper-left corner is extracted from the 2D output, while for the 3D model the slice corresponding to the same 2D surface is extracted to form subfigures (a), (b) and (d). A comparison between subfigures (a), (b) and (d) reveals that, although the 2D method performs reasonably well at a global scale, its fine-scale structures degrade significantly after cropping, exhibiting noticeable detail collapse. In contrast, the 2D slice derived from our 3D model preserves substantially clearer and more coherent microstructural details. This demonstrates the superiority of the proposed approach, particularly in maintaining structural fidelity at microscopic scales.

4.4 Multi-perspective Analysis

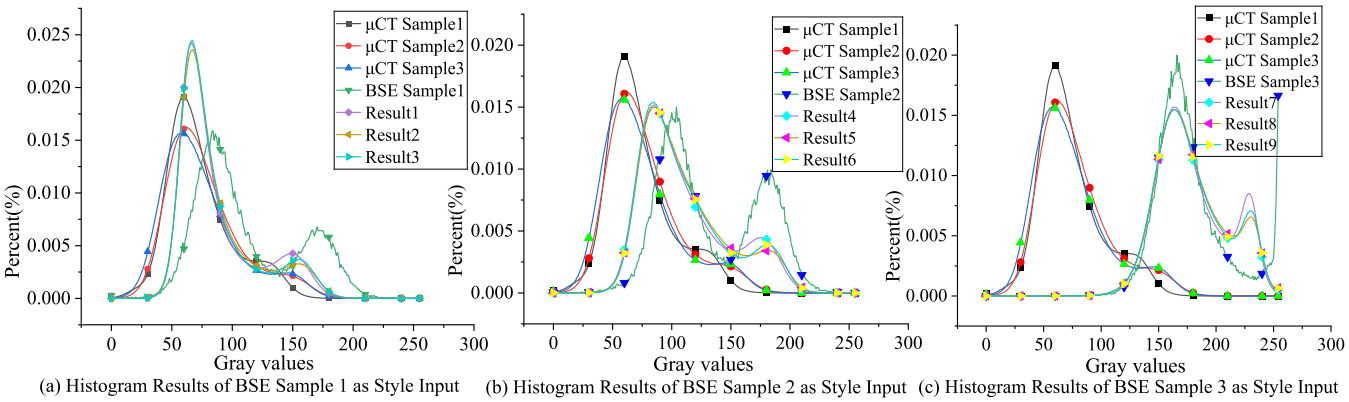
Traditional style-transfer pipelines often attempt to achieve 3D enhancement by independently applying 2D style transfer to each slice and subsequently stacking the outputs. However, this strategy inevitably introduces discontinuities between adjacent slices, leading to abrupt frame-to-frame variations that compromise the spatial coherence of the reconstructed 3D volume.

In contrast, CT2BSE directly models the 3D input as a volumetric entity, thereby eliminating inter-slice inconsistency. As illustrated in Fig. 7, the three orthogonal views of Result 1 consistently preserve clear and continuous structural details across axes, without exhibiting the slice-jumping artifacts commonly observed in 2D-based approaches.

4.5 Quantitative Evaluation

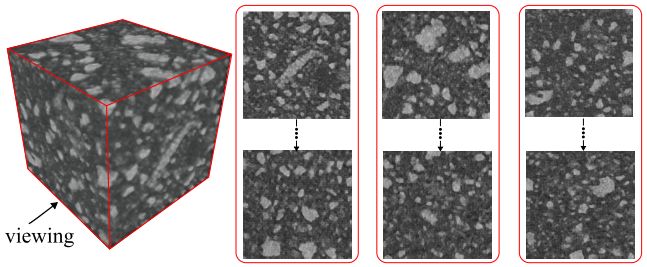
In typical cement images, the histogram often exhibits two dominant peaks corresponding to the unhydrated cement particles and the hydration products, respectively. By examining the horizontal positions of these peaks, one can infer the degree to which the stylization process reflects the characteristics of the target modality.

As shown in Fig. 6, subfigures (a), (b), and (c) illustrate the device-transfer results under BSE Samples 1, 2, and 3, respectively. Across all three BSE styles, the enhanced images show a clear shift of their peak positions toward the corresponding peaks of the BSE histograms. While the shift in (a) is relatively modest, the results in (b) and (c) achieve near-complete alignment: the second peak representing hydration products in (b), and the first peak corresponding to unhydrated cement particles in (c), both match closely with the peak positions of the respective BSE references. This alignment indicates that the device-transfer outputs successfully capture the statistical characteristics associated with the BSE imaging modality.



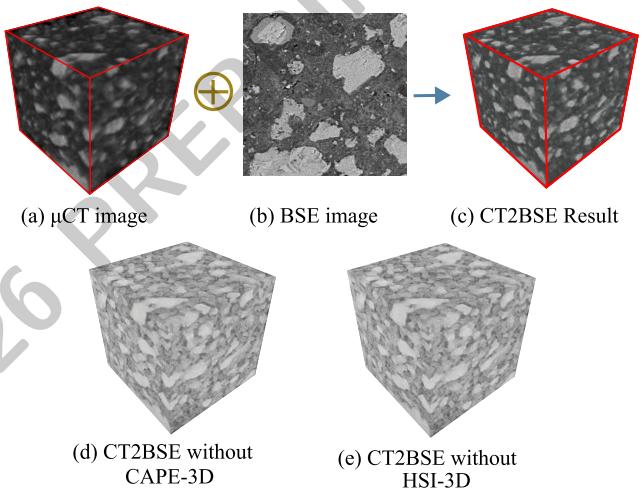
(a) Histogram Results of BSE Sample 1 as Style Input (b) Histogram Results of BSE Sample 2 as Style Input (c) Histogram Results of BSE Sample 3 as Style Input

Figure 6: Subfigures (a)–(c) show the histogram results obtained when BSE Samples 1–3 are used as style inputs. Across all three cases, the peaks of the enhanced results consistently shift toward the corresponding peaks of the BSE distributions. In subfigure (a), only a minor peak displacement is observed, whereas in subfigures (b) and (c), the first and second peaks exhibit overlap with those of the BSE images.



(a) Result 1 (b) right to left (c) top to bottom (d) front to back

Figure 7: The orthogonal slices of Result 1 along the three principal axes. The consistently coherent appearance across all views indicates that the proposed method maintains stable volumetric behavior without exhibiting slice overlap or frame-to-frame discontinuities. Subfigure (a) shows the reconstructed 3D result, while subfigures (b)–(d) display representative slices from right to left, top to bottom, and front to back, respectively.



(a) μCT image (b) BSE image (c) CT2BSE Result

(d) CT2BSE without CAPE-3D (e) CT2BSE without HSI-3D

Figure 8: The ablation study of the CT2BSE model. Subfigures (a) and (b) show the inputs to the model, while subfigure (c) displays the output of the full CT2BSE. Subfigure (d) illustrates the result obtained by replacing CAPE-3D with the relative position bias, and subfigure (e) shows the result of substituting HSI-3D with AdaIN.

4.6 Ablation Experiment

To further validate the effectiveness of the proposed HSI-3D and CAPE-3D modules, we design two complementary ablation experiments. Specifically, we compare the proposed HSI-3D with the classical AdaIN module by replacing HSI-3D with AdaIN, and we also evaluate the impact of retaining the 3D relative position bias instead of using CAPE-3D.

For HSI-3D, as shown in subfigure (e) of Fig. 8, this variant fails to preserve the style characteristics of the BSE images. The underlying reason is that BSE images may contain large bright regions, resulting in extreme mean and variance values during training. These unstable statistics propagate through the decoder, causing undesirable and inconsistent stylization artifacts.

For CAPE-3D, as shown in subfigure (d) of Fig. 8, replacing CAPE-3D with the original 3D relative position bias of Swin Transformer leads to degraded stylization quality. This observation further confirms the necessity of CAPE-3D in maintaining stable and coherent style transfer within the 3D setting.

5 Conclusion

This paper proposes CT2BSE, which can cross-device generate 3D BSE microstructural images from μCT. A Voxel Swin Transformer equipped with CAPE-3D is proposed to efficiently extract 3D features while preserving the original microstructural content of cement hydration. Furthermore, a multidimensional feature fusion module is proposed to mitigate feature instability caused by the extreme grayscale variations inherent in cement images. Ultimately, high-fidelity 3D BSE images of cement hydration are cross-device generated from μCT to BSE using computational imaging paradigm.

However, CT2BSE requires substantial GPU memory, with the training process consuming 79,013 MiB out of 81,920 MiB. In future work, we plan to reduce the model size through knowledge distillation to make it more adaptable to a wider range of scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62403209); the Shandong Provincial Natural Science Foundation (Grant No. ZR2024QF021); the Research Project of Quancheng Laboratory under (Grant No. QCL20250304, QCL20250303); the Youth Innovation Team Program of Shandong Higher Education Institutions, China (Grant No. 2024KJH104); the University of Jinan Young Faculty Interdisciplinary Convergence Development Project2025 (XKJC-202507); the National Natural Science Foundation of China (Grant No. 61872419); the Shandong Provincial Natural Science Foundation (Grant Nos. ZR2022ZD02, ZR2023LZH015); "New 20 Rules for University"Program of Jinan City (Grant No. 202534127); the Natural Science Foundation of Shandong Province Computational Intelligence (Grant No. SDCI202404); the Key Research and Development Program of Shandong Province (Grant No. 2024CXPT084); the National "Challenge-Based" Major Science and Technology Project in the Building Materials Industry (Project No. 2023JBGS11-03); and the High-performance Computing Platform at University of Jinan.

References

[Azadi *et al.*, 2018] Samaneh Azadi, Matthew Fisher, Vladimir G Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7564–7573, 2018.

[Chung *et al.*, 2019] Sang-Yeop Chung, Ji-Su Kim, Dietmar Stephan, and Tong-Seok Han. Overview of the use of micro-computed tomography (micro-ct) to investigate the relation between the material characteristics and properties of cement-based materials. *Construction and Building Materials*, 229:116843, 2019.

[Chung *et al.*, 2024] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8795–8805, 2024.

[Deng *et al.*, 2022] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022.

[Galerne *et al.*, 2025] Bruno Galerne, Jianling Wang, Lara Raad, and Jean-Michel Morel. Sgsst: Scaling gaussian splatting style transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26535–26544, 2025.

[Gallucci *et al.*, 2007] E Gallucci, K Scrivener, A Groso, M Stampanoni, and G Margaritondo. 3d experimental investigation of the microstructure of cement pastes using synchrotron x-ray microtomography (μct). *Cement and Concrete Research*, 37(3):360–368, 2007.

[Gatys *et al.*, 2016] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[Huang *et al.*, 2017] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 783–791, 2017.

[Huang *et al.*, 2023] Siyu Huang, Jie An, Donglai Wei, Jiebo Luo, and Hanspeter Pfister. Quantart: Quantizing image style transfer towards high visual fidelity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5947–5956, 2023.

[Jia *et al.*, 2016] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016.

[Kim *et al.*, 2024] Se-Yun Kim, Donghwi Eum, Hoonhee Lee, Kyoungsoo Park, Kenjiro Terada, and Tong-Seok Han. Evaluating tensile strength of cement paste using multiscale modeling and in-situ splitting tests with micro-ct. *Construction and Building Materials*, 411:134642, 2024.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2022] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[Mao *et al.*, 2025] Li-xuan Mao, Fuqiang He, Lihui Li, Wei Xu, Yong Wang, and Qing-feng Liu. A quantitative study of phase assemblage in cement-fly ash-slag ternary systems using machine learning-assisted bse-eds image analysis. *Construction and Building Materials*, 498:143712, 2025.

[Miarka *et al.*, 2024] Petr Miarka, Daniel Kytỳř, Petr Koudelka, and Vlastimil Bílek. Damage localisation in fresh cement mortar observed via in situ (timelapse) x-ray μct imaging. *Cement and Concrete Composites*, 154:105736, 2024.

[Mukherkjee *et al.*, 2022] Debadyuti Mukherkjee, Pritam Saha, Dmitry Kaplun, Aleksandr Sinitca, and Ram Sarkar. Brain tumor image generation using an aggregation of gan models with style transfer. *Scientific reports*, 12(1):9141, 2022.

[Ruder *et al.*, 2016] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *German conference on pattern recognition*, pages 26–36. Springer, 2016.

[Saseendran *et al.*, 2024] Vishnu Saseendran, Namiko Yamamoto, Peter J Collins, Aleksandra Radlińska, Sara Mueller, and Enrique M Jackson. Unlocking the potential: analyzing 3d microstructure of small-scale cement samples from space using deep learning. *npj Microgravity*, 10(1):11, 2024.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Vigor *et al.*, 2022] James E Vigor, Susan A Bernal, Xianghui Xiao, and John L Provis. Time-resolved 3d characterisation of early-age microstructural development of portland cement. *Journal of Materials Science*, 57(8):4952–4969, 2022.

[Wang *et al.*, 2020] Wenjing Wang, Shuai Yang, Jizheng Xu, and Jiaying Liu. Consistent video style transfer via relaxation and regularization. *IEEE Transactions on Image Processing*, 29:9125–9139, 2020.

[Wang *et al.*, 2023] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7677–7689, 2023.

[Wang *et al.*, 2025] Chunfu Wang, Yuling Wang, and Yunfeng Pan. The fractal characteristics of cement-based materials with grinding aid via bse image analysis. *Scientific Reports*, 15(1):37717, 2025.

[Wu *et al.*, 2021] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14618–14627, 2021.

[Xu *et al.*, 2021] Wenju Xu, Chengjiang Long, Ruisheng Wang, and Guanghui Wang. Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6383–6392, 2021.

[Zhang *et al.*, 2023] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023.

[Zhang *et al.*, 2025] Shuhao Zhang, Hui Kang, Yang Liu, Fang Mei, and Hongjuan Li. Hsi: A holistic style injector for arbitrary style transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23433–23442, 2025.