REALITY-ENHANCED SYNTHETIC IMAGE TRAINING DATASET FOR COMPUTER VISION CONSTRUCTION MONITORING

JUAN D. NÚÑEZ-MORALES, SHUN-HSIANG HSU, AMIR IBRAHIM, and MANI GOLPARVAR-FARD

Dept of Civil and Environmental Engineering, University of Illinois, Urbana-Champaign, USA

Recent growth in the availability of aerial and ground-level images from construction sites has created a surge in computer vision-based techniques for construction monitoring applications. To address the need for large volumes of visual data with ground truth for training the underlying machine learning models, the application of synthetic data, particularly from 3D/4D BIM, has gained traction in the research community. However, the existing domain gap between real and synthetic images, such as rendering noise from repeated material texture patterns, randomized positional camera parameters that collect the ground truth of BIM elements, and the poorly visibility of BIM elements to the camera viewpoints within these environments, has negatively impacted the potential use of these synthetic datasets as scale. To address these limitations, this paper presents a new synthetic image generation pipeline that optimizes and integrates camera extrinsic parameters with a novel synthetic image appearance enhancement technique to generate high volume of quality synthetic data with respective ground-truth. The use of synthetic datasets for training progress monitoring models is validated through several real-data segmentation cases that incorporate: 1) the automatic collection of synthetic images and ground-truth annotations from high-LoD BIM model disciplines, 2) optimization of positional camera parameters using element visibility metrics, and 3) the enhancement of realism of synthetic images using a patch-based generative approach. The benefits and the current limitations to automated progress monitoring, as well as robotic path automation and optimization for progress monitoring real data collection are discussed.

Keywords: Progress monitoring, Deep learning, BIM, Automation, Robotics, Synthetic data.

1 INTRODUCTION

Over the past decade, there has been a surge in the usage of camera-equipped drones as well as 360-degree cameras for reality capture and mapping as-built conditions on construction sites. Coupled with the growing usage of integrated Building Information Models (BIM) and schedule, these as-built vs. as-planned environments have fueled the development of a body of computer vision techniques to automatically track work-in-progress and check the quality of work-in-place. The underlying techniques used among many of the recent solutions is semantic segmentation of the states of work-in-progress directly from the reality capture data. Training, validating these techniques and using them to infer progress on new data, require comprehensive datasets of visual representation of various states of work-in-progress (e.g., aluminum/plywood stud vs.
insulation vs. sheetrock vs. plastering vs. painting for drywall installations) with relevant manually created ground truth. Creating such dataset is not easy. Many technology companies have services around their reality capture data to offer the required actionable insight to their customers while creating ground truth data. The main challenge associated with such approach is that a product-led growth of data annotation may not mature and scale to hundreds of states of work in progress that map to ASTM UnIFORMAT classification of construction objects. This problem is particularly exuberated when data hungry techniques such as deep learning semantic segmentation algorithms are used.

A recent trend in research has gravitated towards using synthetic data to solve the need for large volumes of data. Recent work has focused on inverse photogrammetry (Braun and Borrmann 2019), rendering of materials patches (Lin et al. 2019), 4D BIM simulations (Wei and Akinci 2022), and even style transfer generative models (Hong et al. 2020), however, the potential contribution of synthetic datasets remains untapped. The extensive work done in this field has allowed for the identification of the main challenges that make synthetic data unviable. These challenges of data quality achieved by current collection pipelines can be attributed to three items: 1) the need for a real-to-synthetic correspondence and geometrical alignment between BIM models and real photos limits the scalability of solutions, 2) the collection of exhaustive data points of a built environment generates views that do not reflect real-life captured scenes during progress data collection (e.g., indoor 360-walkthroughs), and 3) rendering engines produce scenes which lack realism needed to create models that can successfully detect repetitive and homogeneous material patterns on construction objects.

To overcome such challenges and provide high-volume and quality synthetic data for constructive elements, this study introduces an end-to-end synthetic data generation pipeline for creating effective simulated camera poses that create canonical views of 3D elements, as well as a deep learning-based realism enhancement process for minimizing the gap between synthetic and real data. The findings are discussed through an ablation study for several cases of semantic segmentation using Transformers-based models (Liu et al. 2021).

2 RELATED WORKS

The significant requirements for high-volume and high-quality images from deep learning models impose high labor and costs for the construction industry. As a workaround, research has developed several methods to leverage synthetic datasets from BIM models to train deep learning models for object segmentation tasks.

Recent implementations started leveraging additional scene information based on BIM geometry to improve segmentation accuracy. For example, Lin et al. (2019) extracted depth information of the BIM geometry and used as additional input in a segmentation inference pipeline. In such cases, a significant amount of ground-truth data was required and restricted to geometrical matching between the real data and BIM model.

With the introduction of generative models, many studies directly explored the generation of ground-truth synthetic data from BIM models. Relevant work from Hong et al. (2020) used GAN networks to perform a style transfer of BIM images to real-world image counterparts using adversarial training. Despite the improvements in the creation of purely synthetic data, such an approach was restricted to finished buildings and still required a relatively perfect correspondence between an existing real image and BIM model.

The scalability problem of real-to-synthetic image correspondences contradicted the purpose of using synthetic datasets. By contrast, most recent and novel work presented by Wei and Akinci (2022) proposed a pipeline for leveraging 4D BIM models using rendering engines. Such
a pipeline allowed for the control of the virtual environment to simulate different physical world conditions while being capable of providing realistic-looking images. However, shown experiments are still limited to the extent of the BIM elements shown in the scene and still need to explore potential accuracy improvements for segmentation models. To address the presented drawbacks, this paper presents a new pipeline decoupling the reality from the synthetic environment and improving the quality of generated synthetic images.

3 METHOD

The proposed method uses several BIM models with each model discipline at or higher than LoD 350. To prepare data, the elements within these models are augmented with a user-specified property at the element level to map to corresponding ASTM Uniformat classes. Figure 1 presents a new three-stage process to create high-volume and quality synthetic data from BIM models.

![Figure 1. Proposed synthetic data collection pipeline. The first stage collects rendered scenes as well as additional image modalities, the second stage makes use of the rendered scenes and image modalities in a patch-based mechanism to enhance the realism of rendered images. Finally, based on camera visibility metrics of resolution and distance, canonical views are selected.](image)

The first stage automatically designs a camera path for automated collection of synthetic data, mimicking a typical 360-camera walkthrough of a field engineer on a real project site. It is composed of: 1) a **BIM object metadata parser**, which collects the metadata corresponding to the element name, category, level, and coordinate boundaries of a bounding box corresponding to the mesh of a BIM element; 2) a **material and mask generator**, which uses object categories to determine a unique class color palette and the material type to be applied to corresponding elements for the final scene rendering, randomly sampling textures from a pool of collected materials patches; 3) a **lighting module** that iteratively simulates sun light at different intensities and positional parameters, similar to real project sites; and 4) a **camera creator** that uses floor
elements to compute an evenly spaced grid coordinate system for estimating camera positions, and eliminates cameras whose projected forward vector from the camera lens is within a pre-defined threshold of an object.

The second stage uses a realism enhancement mechanism adapted from the work proposed by Richter et al. (2023). First, provided G-Buffers from the Synthetic Data Pipeline (e.g., depthmaps and surface normals) are encoded at multiple resolutions for each element class. Second, these features are passed to residual blocks of an HRNet architecture, which performs inference to produce an enhanced synthetic image from which image patches are sampled. Third, patches are compared against a small dataset of real ground-truth annotated images. Patch matching is performed based on a Cosine similarity metric between encoded patch features passed through a VGG network. Finally, a PatchGAN determines between real and synthetic patches which is real and fake, using the Learned Perceptual Image Patch Similarity (LPIPS) as regularizer.

The third and final stage estimates the complete visual coverage of the simulated elements and redundant visibility of these elements in different rendered frames to assure all elements that are captured in the dataset have high viewpoint variability. Similar to Ibrahim et al. (2022), two metrics are devised to assess and improve the: 1) visual coverage, and 2) redundant observation of the elements in the collected data.

The calculation of the visual coverage employs a color-indexing process per visible BIM element during simulating the data collection process. The back-projected colors in each frame are used to deduce the number of pixels belonging to each element in each frame. Moreover, the redundant observation metric is calculated based on the number of frames an element is visible. Starting with a grid pattern data collection plan to sample waypoints in the BIM model, then, a greedy Next-Best-View (NBV) method, as proposed in Wong et al. (1998) is used to select which frames to include in the dataset.

4 EXPERIMENTS

The designed experiment considers a BIM model of a hospital construction, with MEP, complete drywall module sequences, and structural concrete for columns, beams, floor slabs, and load-bearing walls at LoD 400. Unreal Engine v4.26 is used as the rendering platform, where all the data collection process is automated, and images are collected at a 480x480 resolution. This experiment automatically models light intensities of 1000, 500, and 200 W/m², corresponding to clear sky, cloudy sky, and overcast sky intensities respectively, simulating the light intensities of typical real construction sites.

The collected images make use of the enhancement pipeline, using 200 ground-truth real-world image annotations to correctly map known real elements to the ones annotated from the synthetic data collection pipeline. The basic model training hyperparameters are the default ones proposed in Richter et al. (2023). Visibility metrics are computed based on manually set pixel and frame thresholds. To meet the required visual coverage in each frame, an element is set to meet a minimum of 625 pixels per frame. To meet the frame redundancy metric, the threshold of each element being visible across frames is set to a minimum of five frames.

Finally, five different image segmentation model cases are trained to understand the effectiveness of each of the synthetic data pipeline steps on a test set. This work employs a SWIN-b transformer model (Liu et al. 2021) and measures the Intersection over Union (IoU) of segmentation on a test set of 500 real-ground truth images. Each trained case makes use of a pretrained model on the ADE20K dataset and is then finetuned during 100k iterations.
5 RESULTS

This section includes the intermediate results of each of the steps of the proposed pipeline. Figure 2 shows an example for the automated synthetic data collection pipeline, and each of the image modalities collected. Focusing on semantic segmentation, the pipeline can collect fully annotated images for each of the constructive element class. Moreover, additional image information such as surface normals and depthmaps are collected simultaneously, which are later used for the enhancing of image realism and camera positional parameter selection.

![Images](a) (b) (c) (d)

Figure 2. A set of automatically collected synthetic image modes, with (a) rendered images, (b) ground-truth annotations, (c) surface normals, and (d) depthmaps.

Figure 3 shows qualitative results of the realism enhancement, minimizing noise image information and smoothing repetitive and homogeneous surface patterns. Once cameras positional parameters are selected based on their visibility metrics, remaining ones are pruned, creating a separate dataset containing the necessary views for optimal performance of the segmentation network. From a total of 775 synthetic images provided, optimization retrieves only 435 scenes, providing a 44% in data reduction.

Based on the explored experiments, the largest increase in segmentation performance is given by the case of training a model with purely synthetic data without enhancements, nor camera selection, increasing IoU by 12.85%. Moreover, model performance results reported in the case of camera selection through optimized positional parameters show a comparable increase despite eliminating half of the data points. Results for all cases are shown in Table 1.

![Images](a) (b)

Figure 3. A sample of an enhanced synthetic image using the proposed pipeline. Image on the left correspond to a scene directly rendered from the graphics engine, while the image on the right shows the enhanced version of the rendered images.

<table>
<thead>
<tr>
<th>Trained Model Case</th>
<th>No. Real Training Images</th>
<th>No. Synthetic Training Images</th>
<th>Test IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWIN-B</td>
<td>2,000</td>
<td>0</td>
<td>61.29</td>
</tr>
<tr>
<td>SWIN-B + No Refinement</td>
<td>2,000</td>
<td>765</td>
<td>74.14</td>
</tr>
<tr>
<td>SWIN-B + Enhancement</td>
<td>2,000</td>
<td>765</td>
<td>70.01</td>
</tr>
<tr>
<td>SWIN-B + Camera Selection</td>
<td>2,000</td>
<td>435</td>
<td>70.85</td>
</tr>
<tr>
<td>SWIN-B + Full Pipeline</td>
<td>2,000</td>
<td>435</td>
<td>70.9</td>
</tr>
</tbody>
</table>

Table 1. Performance results for all segmentation model cases using different steps of the pipeline for synthetic data collection. Detected classes include ceiling, wall, floor, person, window, pipe, and duct.
6 DISCUSSION AND CONCLUSION

This paper presented a complete pipeline for generating high-volume and high-quality synthetic ground-truth dataset for training segmentation models needed for computer vision progress tracking use cases. Results were shown on how the realism of synthetic images can be enhanced to eliminate repetitive patterns from material texture patches, and automatically alter the images for more realism. Camera visibility metrics are leveraged for selecting canonical views of elements, thus filtering non-contributing images from the segmentation pipeline. Performance improvement was evaluated through a dataset ablation study, demonstrating that when selecting best camera views, the synthetic data provides significant increase in segmentation accuracy, despite pruning close to half of the synthetic dataset. Future work focuses on investigating better pruning and enhancement techniques for synthetic data at larger scales to achieve increased machine learning performance. The proposed pipeline of methods can provide a path refinement technique for planning unmanned robotic vehicle for reality capture on real-world jobsites as well. Designing and implementing a research experiment to collect such data autonomously and validate the usage for autonomous progress tracking is also considered as part of the future work.

References


