A Step from Virtual to Reality: Investigating the Potential of a Diffusion-Based Pipeline for Enhancing the Realism in Fully-Annotated Synthetic Construction Imagery

Sina Davari, Ali Tohidifar, and Daeho Kim

Department of Civil and Mineral Engineering, University of Toronto, Toronto, ON, Canada

sina.davari@mail.utoronto.ca, ali.tohidifar@mail.utoronto.ca, and civdaeho.kim@utoronto.ca

Abstract

In the rapidly evolving field of artificial intelligence (AI), synthetic data generation has become increasingly crucial, particularly in domains where real-world data is scarce, expensive, or sensitive. In this study, we introduce BCGen, a novel image realism enhancement pipeline that integrates our proprietary synthetic construction data generation and autonomous labeling engine, BlendCon, integrated with Generative AI. Leveraging the graphical capabilities of Blender and the deep learning prowess of the ControlNet model, BCGen represents a novel approach to synthesizing and enhancing construction site imagery. Our methodology narrows the reality gap, delivering images with increased realism and diversity while preserving the full annotations. The paper delineates our approach, methodology, and the broader implications of our findings. Through meticulous hyperparameter tuning and an innovative post-processing technique, we demonstrate the enhanced realism and diversity of the generated images, pointing towards the vast potential of synthetic data in visual AI applications within construction.

Keywords

Artificial Intelligence in Construction, Synthetic Data, ControlNet, Reality Gap

1 Introduction and Background

1.1 Overcoming Data Acquisition Challenges in Construction with Synthetic Solutions

Since 2017, the construction industry, known for its complex operations and dynamic environments, has increasingly harnessed Deep Learning (DL) to overcome unique challenges in data acquisition and processing across various domains, such as safety, road surveys, bridge inspection, and site operation monitoring [1]. Despite the significant role of DL in revolutionizing traditional practices, such as object detection, instance segmentation, and pose/activity recognition—often surpassing human capabilities— its integration into the construction sector faces hurdles, notably, the scarcity of fully-annotated data [2]. This scarcity stems from the high costs and time requirements for manual collection and labeling, logistical challenges in deploying sensors, and confidentiality issues, highlighting a critical barrier to leveraging data-hungry DL tools within the sector [3].

To address the limitation mentioned, synthetic data generation, bolstered by advances in rendering engines and GPUs, offers a promising solution, especially for tasks difficult to label in real-world settings. The adoption of Blender [4] and video game engines for data generation demonstrates this potential, particularly in facilitating the creation of human-centric images [5-7]. This strategy addresses the disparity in the volume of training images available in computer science versus construction studies, where the former often uses multi-million-image datasets while the latter struggles with far fewer images [8]. These rendering engines enable the generation of synthetic data, addressing data scarcity and streamlining the deployment of DL techniques in construction contexts more effectively.

1.2 An Introduction to our Synthetic Data Generation Engine, BlendCon

In a forthcoming publication, we introduce BlendCon, a computational framework that addresses the need for high-quality, diverse data in the construction industry, particularly for the advancement of Deep Neural Networks (DNNs)-based visual AI. The framework utilizes the graphics engine, Blender, to generate synthetic, fully-labeled data, which is a step forward in overcoming the limitations associated with manual data collection and privacy concerns. BlendCon creates a virtual construction site, producing diverse synthetic images by randomizing key elements of the scene. This approach has been validated in two key areas: trainability

and scalability. For instance, Yolov7 models trained with data generated by BlendCon showed comparable or superior performance to models trained with real data.

BlendCon distinguishes itself by integrating dynamic elements, such as mobile equipment and human workers, into synthetic image generation and by offering simulations from multiple perspectives, breaking away from the limitations of single-viewpoint models and enhancing diversity in synthetic data. Furthermore, it is equipped with multimodal label generation capability, producing various types of labels for each image, namely depth maps, semantic masks, and 2D and 3D bounding boxes (BBs) and key points for construction workers.

While BlendCon has proven effective in synthetic data generation and the trainability of DNNs, it still faces a pivotal challenge: the "reality gap." In the subsequent section, we delve deeper into this challenge and elaborate on how this gap, i.e., the divergence in the perceptual and contextual quality between synthetic and real-world imagery, is a crucial factor influencing the effectiveness of DNNs trained on synthetic data in real-world scenarios.

As we delve into the realm of synthetic data generation and the pursuit of enhanced image realism, we arrive at an inquiry that guides our research trajectory: How can we further improve BlendCon? By enhancing image realism through generative AI, we propose the next generation of BlendCon, aimed at offering a scalable solution to the scarcity of labeled construction datasets and facilitating the adoption of deep learning within the construction industry.

1.3 Reality Gap in Synthetic Data Generation

The concept of the reality gap emerges from the inherent differences between the distributions of real and synthetic images. Synthetic images, while beneficial in overcoming data scarcity, often lack nuanced details and contextual variability present in real-world scenarios [9]. This disparity can lead to a decrease in the effectiveness of DNN models when applied to actual construction environments. Effectively bridging this gap is thus a critical objective in enhancing the utility of synthetic data for DNN training, a process referred to as domain adaptation in machine learning [10,11]. Two primary methods have emerged to address the reality gap: enhancing realism and domain randomization.

A) Enhancing Realism: This approach focuses on making synthetic images closely mimic real-world scenarios. Studies suggest that the heightened realism in synthetic data significantly improves the performance of DNNs, allowing them to better adapt to real-world applications [12]. By refining textures and lighting conditions, and incorporating real-world irregularities, this method aims to reduce the perceptual differences between synthetic and real imagery while preserving full annotations [13,14]. This approach is not without its limitations; significantly, the process of enhancing realism in synthetic images can be both time-consuming and costly, and the subjective nature of 'realism'—what is deemed realistic—remains ambiguously defined.

B) Domain Randomization: Contrasting with the pursuit of realism, domain randomization adopts an unconventional tactic –intentionally making synthetic images more abstract or less realistic. This method involves introducing a high degree of variability in the synthetic images, which paradoxically can lead to the development of more robust DNN models. The premise is that by exposing the DNN to a wide range of variations, the model learns to focus on the most critical features, becoming more adaptable to real-world variability. A notable study in this area demonstrated the effectiveness of this approach, where severely randomized images contributed to the training of robust models capable of bridging the reality gap [15]. This approach, however, requires significant computational resources and carries the risk of overgeneralizing DNN models, potentially leading to models that, while robust in handling diverse scenarios, may not perform optimally in specialized tasks, such as construction applications.

The choice between these approaches depends on the specific requirements of the application and the nature of the tasks the DNN models are expected to perform. This paper marks the beginning of our exploration into the first approach, laying the foundation for a future study that is planned to investigate both strategies with the objective of evaluating their impact and effectiveness in optimizing the utility of synthetic data for training DNN models in construction applications.

1.4 ControlNet for Enhanced Image Realism

In this study, we introduce BCGen, a pipeline for integrating the power of generative AI, specifically the Stable Diffusion (SD) model, ControlNet [16], into our automated construction image synthesis and labeling framework, BlendCon. This marks the first application of ControlNet in the construction domain for image-to-image translation, generating more realistic images from synthetic ones while maintaining full annotations.

ControlNet is designed to integrate spatial conditioning controls into large, pre-trained text-to-image diffusion models. By leveraging robust and deep encoding layers of these models and applying zero convolutions, it finetunes the imagery while avoiding the introduction of deleterious noise. Its ability to handle various inputs, such as edges and human poses, and its robust performance across datasets of differing scales, make ControlNet an essential tool for generating realistic images from synthetic data, particularly in applications where maintaining accurate annotations is as crucial as image quality itself [16].

ControlNet distinguishes itself by its ability to fine-

tune the realism of generated images while preserving annotations [16]. This capability is crucial in applications like construction site imaging, where maintaining the accuracy of annotations is as important as the visual realism of the images. By integrating ControlNet, we hypothesize that synthetic images could become more adaptable for real-world applications, thereby potentially bridging the reality gap. This assertion, however, remains to be validated in future phases of our research.

1.5 Related Work on Enhancing the Realism of Synthetic Images

Deep learning, particularly Generative Adversarial Networks (GANs) [17] and Diffusion Models [18] has been pivotal in enhancing image realism. GANs, by their adversarial nature, refine images to closely resemble real photographs. Diffusion Models, demonstrate remarkable capabilities in text-to-image generation and synthesizing photorealistic images, offering a significant leap in image quality and diversity [19]. GANs, while effective, may struggle with ensuring stability during training, producing artifacts [20]. Diffusion Models, however, characterized by their gradual process of image formation, offer higher stability and image quality, albeit at the cost of increased computational complexity.

A recent study leveraged diffusion models, including ControlNet, to enhance the FFHQ-Aging dataset [21], producing synthetic images that exhibit a diverse array of facial expressions, ethnicities, and lighting conditions, thereby advancing the realism and quality of synthetic imagery for facial image augmentation [22]. Furthermore, in the medical domain, diffusion models have been extensively utilized for various applications, including realistic endoscopic image generation [23] and synthesizing MRI sequences and thoracic X-ray images [24].

Several of the mentioned studies have implemented established photorealism metrics to gauge the quality of the enhanced images. Metrics such as the Inception Score [25], Fréchet Inception Distance [26], Kernel Inception instance [27], Structural Similarity Index Measure [28], Learned Perceptual Image Patch Similarity [29], and Contrastive Language-Image Pre-training (CLIP) [30]-based metrics [31] have been pivotal in assessing the realism of synthetic imagery. However, despite such evaluations, there remains an underexplored area in the existing literature: assessing the cost-effectiveness of employing these advanced generative models. To the best of the authors' knowledge, a systematic examination of the cost-benefit analysis of utilizing such sophisticated techniques for enhancing synthetic image realism has not yet been documented.

2 Method

2.1 Architecture of BCGen: BlendCon with Integrated Generative AI

Our proposed framework, BCGen, leverages a three-part pipeline to enhance image realism. Figure 1 demonstrates the BCGen pipeline, encapsulating the end-to-end process from image synthesis with BlendCon, through realism enhancement via the ControlNet pipeline, to the final avatar cut and paste for anatomical accuracy, ensuring the retention of high-quality annotations.

To further elucidate, initially, BlendCon synthesizes RGB construction site images, along with their corresponding depth maps, semantic masks, and precise annotations of 2D and 3D bounding boxes and key points for construction workers, using inputs such as horizon, processed scenes, animated avatars, and lighting and camera configuration. The outputs are fed into the ControlNet pipeline, where the RGB images, depth maps, and semantic masks—alongside a text prompt—are processed by ControlNet to create more realistic images.

Our text prompt is "a high-quality, high-resolution

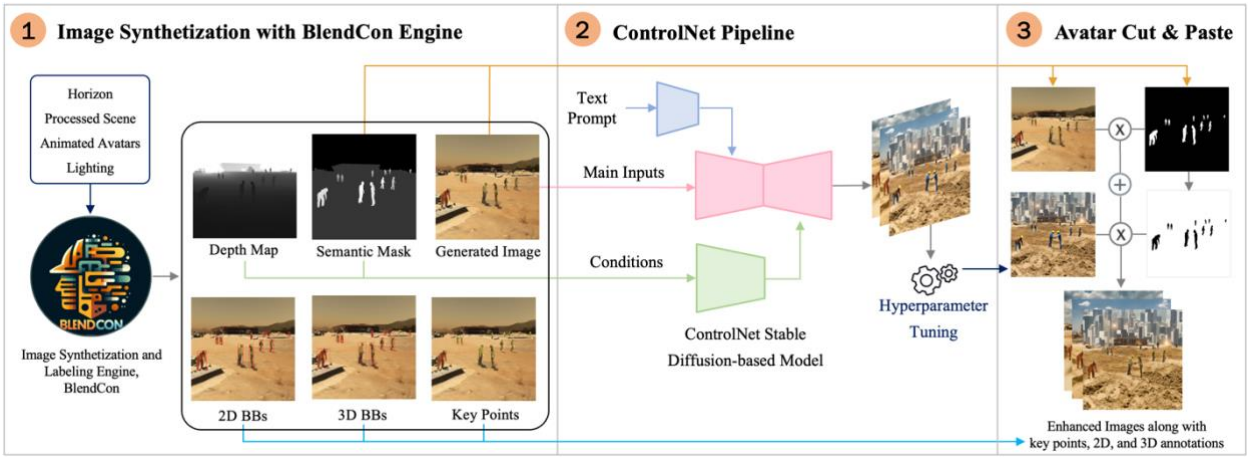

Figure 1. The proposed image realism enhancement pipeline

image of a construction site." We also pass "blurry, blurred, bad anatomy, low quality" as the negative prompt.

This integration of ControlNet paves the way for enhanced realism while meticulously preserving all annotations, including worker key points and 2D and 3D bounding boxes. Hyperparameter tuning is employed to refine this process and optimize the outputs, a topic we will explore in detail in the subsequent section. In the final stage, to ensure anatomical accuracy and maintain the integrity of our key point annotations, we initiated a process termed 'Avatar cut and paste,' which involves extracting the worker avatars from the initial BlendCon-generated images and superimposing them onto the images enhanced by ControlNet.

2.2 ControlNet Hyperparameter Tuning

Given ControlNet's extensive range of adjustable parameters, such as the degree of reliance on the input images, conditions, and text prompts, we embarked on a rigorous hyperparameter tuning exercise employing a grid search methodology. This allowed us to identify the most ideal settings for our particular use case. Moreover, it was during this tuning process that we encountered scenarios where, despite explicitly excluding poor anatomy and low quality in our negative text prompt, the output sometimes exhibited compromised structural integrity, especially in the anatomy of the construction workers, and as outlined previously, to counteract this, we resorted to avatar cut and paste.

We investigated six ControlNet hyperparameters, namely output image size, conditioning scales—which determine the weight of our conditions, i.e., depth maps and semantic masks, classifier-free guidance scale (CFG) —which dictates the influence of the text prompt on image generation, number of inference or denoising steps for the diffusion model, input image strength—which determines the input image weight, and the choice of diffusion model noise scheduler, across three different random seed initializations. The results were scrutinized, and the most effective hyperparameter combinations were selected through visual comparisons.

The initial phase of our study involved a qualitative assessment of the synthetic images generated by our pipeline, relying on visual observation to evaluate the quality. We considered any image unrealistic, blurry, distorted, or exhibiting anatomical inaccuracies and abstract backgrounds as unsatisfactory. Figure 2 showcases examples where the interplay of hyperparameters resulted in suboptimal results, such as compromised human anatomies, abstract backgrounds, and blurred images, underscoring the inherent challenges and complexities of synthetic data generation.

3 Results, Discussions, and Limitations

3.1 Hyperparameter Tuning Results

Our analysis underscored the significant impact of image size on output realism, diversity, and quality, investigating two sizes of 512 by 512 and 1280 by 1280, which revealed that larger images notably enhanced all aspects. Our investigation into the number of inference steps, specifically examining 40, 80, and 150 steps, revealed its critical significance: fewer than 50 steps often resulted in blurry and structurally unsound images, while exceeding 100 steps did not notably improve quality but extended runtime unnecessarily.

The investigated eight schedulers are linear multistep (LMSDiscrete), denoising diffusion implicit (DDIM), denoising diffusion probabilistic (DDPM), multistep diffusion probabilistic (DPMSolverMultistep), Euler (EulerDiscrete), pseudo numerical (PNDM), Euler with ancestral sampling (EulerAncestralDiscrete), and unified predictor-corrector scheduler (UniPCMultistep) [32]. The unified predictor-corrector noise scheduler was identified as the most effective for our task.

The interplay of CFG, Strength, and Conditioning Scalehyperparameters, and their impact on image realism was further explored. We charted the instances where these parameters harmonized to produce satisfactory outputs, yielding realistic images with minimum blurs, and anatomical inaccuracies, as demonstrated in Figure 3. The graph illustrates the frequency of the satisfactory outcomes across various configurations of CFG, strength, and conditioning scales, with marker size indicating the occurrence count.

We experimented with conditioning scales of [0.3, 0.8], [0.5, 0.5], [0.8, 0.3], and [0.8, 0.8], input image strengths of 0.5, 0.7, and 0.9, and CFGs of 5, 7.5, 10, and 12.5. Ultimately, we selected a conditioning scale of 0.8 for both depth maps and semantic masks, a strength of 90 percent, and a CFG of 12.5.

3.2 BCGen Results and Discussions

The application of ControlNet to the original synthetic images from BlendCon has resulted in enhancements in realism and diversity, as evidenced in Figure 4. The before-and-after comparisons illustrate the ControlNet-induced changes, with noticeable improvements in texture detail, lighting fidelity, and the incorporation of realistic environmental effects. These images not only demonstrate an enriched visual diversity but also indicate a substantial narrowing of the reality gap, affirming the potential of our approach in creating realistic images for use in AI training and other construction industry applications.

While generative DL models today can produce hyperrealistic images, our contribution lies in the unique

capability of our pipeline to maintain original annotations.

The importance of annotated images cannot be overstated, as manual labeling of 2D, 3D, key points, semantic masks, and depth maps involves considerable costs, time, and potential for errors. Our effort ensures that the synthetic images generated are not only visually compelling but also maintain full annotations, making them immediately useful for DNN training and other applications within the construction industry.

3.3 Limitations of The Study

This research is a stepping stone, highlighting the necessity for verification and validation methods tailored to the unique requirements of enhanced synthetic image evaluation. The study presents several limitations that inform its theoretical implications:

A) Generalization of Results: The findings, although promising, are not yet generalizable across all potential input scenarios, indicating that further research is required to broaden the applicability of the results. For instance, the suboptimal results in indoor environments with a high degree of clutter, as seen in the last row of Figure 4, suggest that the model may struggle with overly complex indoor construction scenes. This issue may arise due to the lack of specific information in the textual prompt and a heavy reliance on it. This aspect will be further investigated in future studies.

B) Evaluation of Results: Established photorealism metrics could play a pivotal role in evaluating the quality of enhanced images, enabling the creation of business-oriented key performance indicators that measure the efficacy and cost-efficiency of synthetic data generation and enhancement. However, the investigation of these established metrics for evaluating the efficacy and cost-efficiency of the synthetic data enhancement process was not carried out in the current stage of our study.

As previously discussed in our methodology, the initial phase of our study was dedicated to a qualitative assessment of the synthetic images generated by our pipeline, where we relied on visual observation to determine the quality of the output. This subjective method highlights a limitation in our evaluation process, underscoring the need for developing objective criteria and metrics to assess realism and AI training applicability.

C) Computational Resources: The computational demand varies with the conditions set and image size, with the current setup requiring about 40 seconds per 1280 by 1280 image on two NVIDIA RTX 3090 GPUs, which could be a limiting factor for scalability.

D) Variability in Text Prompts: Our exploration of variability in text prompts was limited to a few variations of main and negative prompts, restricting our understanding of their precise impact on the results. Further detailed prompt engineering is designated for future research.

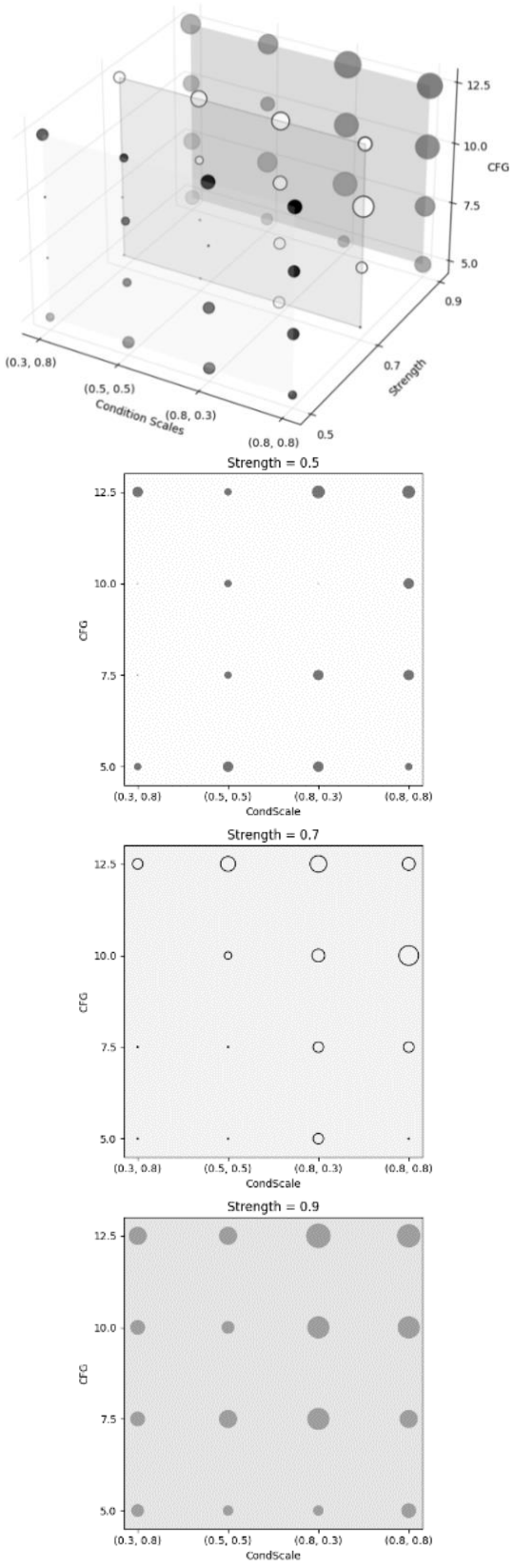

Figure 3. Visualization of ControlNet hyperparameter tuning through grid search

Figure 2. Examples of suboptimal image generations: (a): BlenCon's synthetic images, (b): a satisfactory result, (c) compromised human anatomy, (d): a generated image suffering from both blurriness and compromised human anatomy, (e): occurrence of both anatomical inaccuracies and an abstract background.

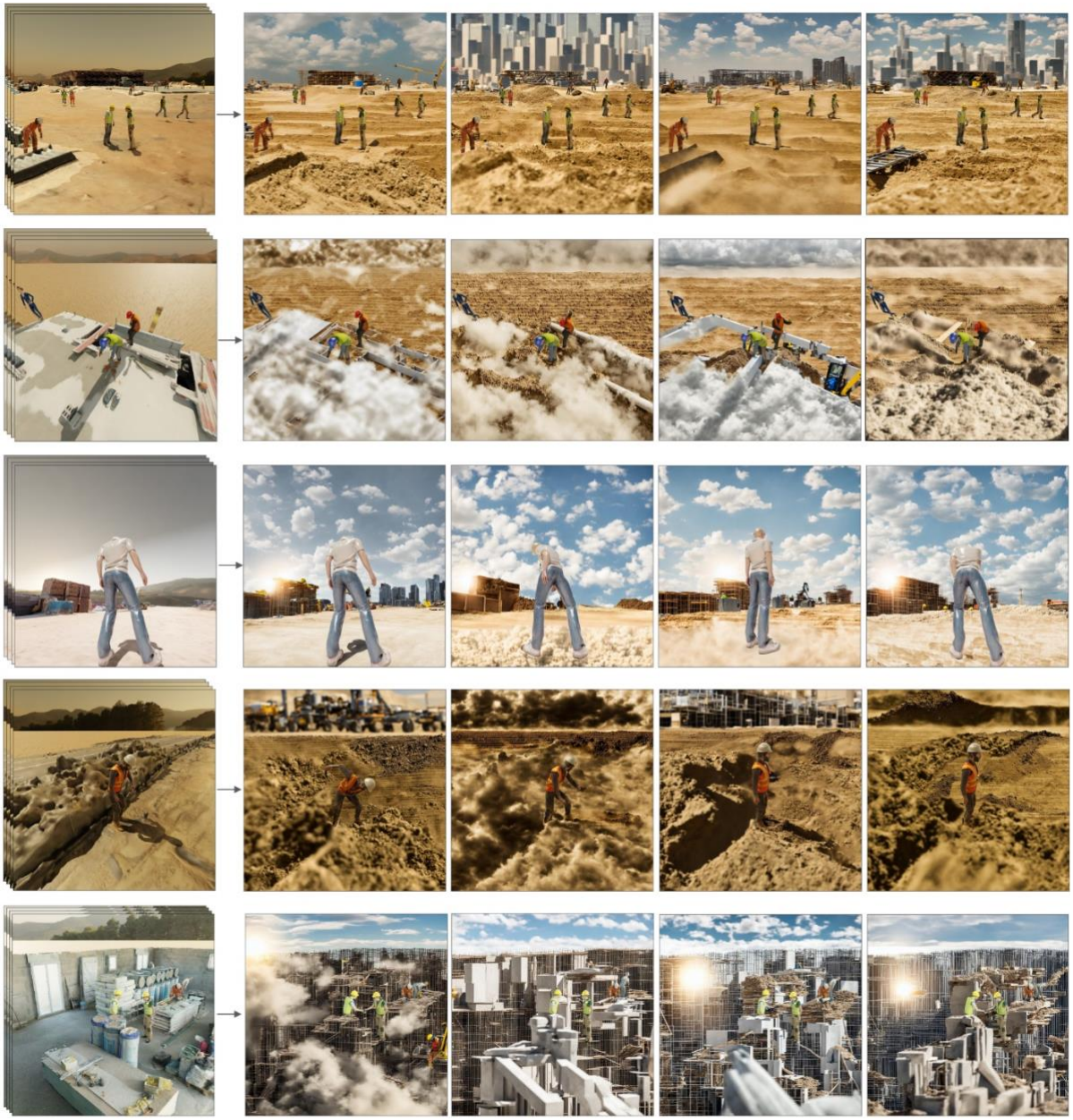

Figure 4. Contrast between the original synthetic construction site image from BlendCon (leftmost) and its enhanced iterations by ControlNet, showcasing diversity improvement and realism enhancements.

4 Looking Ahead: Future Directions

Building upon the work presented in the previous section, the development of business-oriented performance indicators metrics for measuring synthetic data quality and degree of enhancement is essential. These metrics should not only assess the visual fidelity of the images but also quantify the cost-benefit of enhancing realism within a corporate context. Upcoming studies will focus on validating DNN trainability, performance, and affordability with enhanced images, crucial for practical AI applications.

Further exploration into additional input modalities for ControlNet, such as human poses or edge maps as conditions or main inputs, is planned. Leveraging human key point detection models, such as OpenPose [33], could allow for a more nuanced representation of worker anatomies. This aligns with methods like Control-GPT [19], which combines programmatic sketches with text-to-image generation, a technique that could be adapted to enrich our dataset diversity and control [19]. Additionally, we plan to explore the implementation of alternative generative models besides ControlNet to assess their performance and facilitate comparative analysis.

By experimenting with multiple prompts and conditions, we aim to refine our generative model's output further, ensuring that the synthetic images not only serve the construction industry's current needs but also pave the way for emergent AI-driven solutions.

In future research, we aim to employ photorealism metrics for developing business-oriented key performance indicators, thereby measuring the efficacy and cost-efficiency of synthetic data generation and enhancement. This initiative seeks to standardize the validation of synthetic image quality within the AI field.

Additionally, we will evaluate the performance of DNNs trained on limitedly available real-life construction datasets, synthetic data generated by state-of-the-art models, such as Midjourney [35], and domain-randomized synthetic imagery. This evaluation is crucial to verify our method's cost-effectiveness and practicality in real-world applications, shedding light on the economic viability of leveraging such advanced techniques in the construction industry.

5 Conclusion

In this study, focusing on the investigation of the reality gap in synthetic data generation, we introduced a pipeline incorporating the stable diffusion-based model, ControlNet, within our synthetic construction data generation and labeling engine, BlendCon. This pipeline paves the way for generating a diverse range of enhanced synthetic images, while preserving their full annotations, i.e., depth maps, semantic masks, and 2D and 3D bounding boxes and key points for construction workers. Our investigation reveals that ControlNet's hyperparameters critically influence the enhancement of realism, prompting us to conduct a thorough search across over 2,300 hyperparameter combinations, evaluating them through visual observation. This exhaustive process underscored the significance of quantifying realism via photorealism metrics and highlighted the need to balance the costs associated with synthetic image generation and enhancement. In conclusion, our study demonstrates that cutting-edge, controllable diffusion-based generative models hold significant potential for the construction industry, enabling the creation of realistic, fully annotated synthetic imagery by narrowing the reality gap.

References

[1] Xu Y., Zhou Y., Sekula P., and Ding L. Machine learning in construction: From shallow to deep learning. *Developments in the Built Environment*, 6:100045, 2021.

[2] Gupta S., Agrawal A., Gopalakrishnan K., and Narayanan P. Deep Learning with Limited Numerical Precision. In *Proceedings of the International Conference on Machine Learning*, pages 446–454, 2015.

[3] Liu J., Luo H., and Liu H. Deep learning-based data analytics for safety in construction. *Automation in Construction*, 140:104302, 2022.

[4] Blender Foundation. Blender. On-line: https://www.blender.org, Accessed: 22/02/2024.

[5] Tang H. and Jia K. A New Benchmark: On the Utility of Synthetic Data with Blender for Bare Supervised Learning and Downstream Domain Adaptation. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 15954–15964, 2023.

[6] Neuhausen M., Herbers P., and König M. Using Synthetic Data to Improve and Evaluate the Tracking Performance of Construction Workers on Site. *Applied Sciences*, 10(14):4948, 2020.

[7] Fabbri M., Brasó G., Maugeri G., Cetintas O., Gasparini R., Ošep A., Calderara S., Leal-Taixé L., and Cucchiara R. MOTSynth: How Can Synthetic Data Help Pedestrian Detection and Tracking?. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 10849–10859, 2021.

[8] Krizhevsky A., Sutskever I., and Hinton G. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[9] Gamal O., Rameshbabu K., Imran M., and Roth H. Bridging the Reality Gap: Investigation of Deep Convolution Neural Networks Ability to Learn

from a Combination of Real and Synthetic Data. In *Proceedings of the 17th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2020)*, pages 446–454, 2020.

[10] Kouw W. and Loog M. An introduction to domain adaptation and transfer learning. *Machine Learning*, 2019.

[11] Lemberger P. and Panico I. A primer on domain adaptation. *Machine Learning*, 2020.

[12] Wood E., Baltrušaitis T., Hewitt C., Dziadzio S., Cashman T. J., and Shotton J. Fake It till You Make It: Face Analysis in the Wild Using Synthetic Data Alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[13] Shen T., Zhao G., and You S. A Study on Improving Realism of Synthetic Data for Machine Learning. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, 2023.

[14] Atapattu C. and Rekabdar B. Improving the realism of synthetic images through a combination of adversarial and perceptual losses. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, pages 14–19, 2019.

[15] Tremblay J., Prakash A., Acuna D., Brophy M., Jampani V., Anil C., To T., Cameracci E., Boochoon S., and Birchfield S. Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Zhang L., Rao A., and Agrawala M. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, 2023.

[17] Ozair S., Courville A. C., Bengio Y. Generative Adversarial Networks. Machine Learning, 2014.

[18] Ho J., Jain A., Abbeel P. Denoising Diffusion Probabilistic Models. *Machine Learning*, 2020.

[19] Zhang T., Zhang Y., Vineet V., Joshi N., Wang X. Controllable Text-to-Image Generation with GPT-4. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, 2023.

[20] Saxena D. and Cao J. Generative Adversarial Networks (GANs Survey): Challenges, Solutions, and Future Directions. *Machine Learning*, 2023.

[21] Or-El R., Sengupta S., Fried O., Shechtman E., Kemelmacher-Shlizerman I. Lifespan Age Transformation Synthesis. In *Proceedings of the Computer Vision and Pattern Recognition*, 2020.

[22] Zhang J., Hsiung L., and Hsu G. Unleashing Stable Diffusion for Multi-Modal Facial Image Augmentation. In *Proceedings of the International Conference on Advanced Robotics and Intelligent Systems (ARIS)*, pages 1–5, Taipei, Taiwan, 2023.

[23] Kaleta J., Dall'Alba D., Płotka S., and Korzeniowski P. Minimal data requirement for realistic endoscopic image generation with Stable Diffusion. *International Journal of Computer Assisted Radiology and Surgery*, 19:531–539, 2023.

[24] Kazerouni A., Khodapanah Aghdam E., Heidari M., Azad R., Fayyaz M., Hacihaliloglu I., and Merhof D. Diffusion Models for Medical Image Analysis: A Comprehensive Survey. *Image and Video Processing*, 88:102846, 2023.

[25] Salimans T., Goodfellow I. J., Zaremba W., Cheung V., Radford A., and Chen X. Improved techniques for training GANs. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[26] Heusel M., Ramsauer H., Unterthiner T., Nessler B., and Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2018.

[27] Binkowski M., Sutherland D., Arbel M., and Gretton A. Demystifying MMD GANs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[28] Wang Z., Bovik A., Sheikh H., and Simoncelli E. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[29] Zhang R., Isola P., Efros A., Shechtman E., and Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[30] Radford A., Kim J., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., and Sutskever I. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, 2021.

[31] Hessel J., Holtzman A., Forbes M., Le Bras R., and Choi Y. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the Computer Vision and Pattern Recognition*, 2022.

[32] Hugging Face. Schedulers. On-line: https://huggingface.co/docs/diffusers/en/api/schedulers/overview, Accessed: 22/02/2024.

[33] Cao Z., Hidalgo Martinez G., Simon T., Wei S., and Sheikh Y. A. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019.

[34] Midjourney, Inc. Midjourney. On-line: https://docs.midjourney.com/, Accessed: 24/02/2024.