2024 EC³
2024 European Conference on Computing in Construction
Chania, Crete, Greece
July 15-17, 2024

DEVELOPMENT OF A DIGITAL TWIN-BASED SIMULATION SYSTEM AND A NOVEL SYNTHETIC VIDEO DATASET FOR ENHANCING COMPUTER VISION IN CONSTRUCTION SITE SAFETY

Zhengyu Wu, Yuxiang Feng, Yiannis Demiris, Panagiotis Angeloudis
Imperial College London, London, United Kingdom

Abstract

In this work, we introduce a novel construction simulation system and the first photo-realistic synthetic video dataset for the construction industry, named ConSynth. The system simulates vehicle dynamics, worker behaviours, and various environmental conditions (Figure 1), generating a dataset of 200 video sequences and 24,000 images with automated labels. Our experiments have shown that models trained on ConSynth achieve robust detection with less reliance on real-world data, indicating enhanced diversity, improved performance in underrepresented scenarios, and substantial generalisability. Our work promises significant advancements in construction safety monitoring through computer vision.

Introduction

Background

Construction sites are inherently hazardous. With the increasing application of deep learning techniques in safety monitoring on construction sites, accurate detection of construction equipment and workers has grown in importance. However, a significant barrier to this progression is the limited availability of labelled data, such as images or video frames, where each object (e.g. construction equipment or workers) is identified and annotated with a bounding box and a category. Currently, the two largest public datasets in this domain are the Detecting Moving Objects in Construction Sites (MOCS) dataset (Xuehui et al., 2021) and the Alberta Construction Image Dataset (ACID) (Xiao and Kang, 2021). Although models trained on such datasets have demonstrated the ability to effectively detect a variety of construction equipment or workers, a notable 'domain gap' often exists in their applications on real construction sites. This 'domain gap', a common issue in deep learning, signifies the discrepancies in sample distributions between the data that models are being trained on and the data in real-world applications. One common solution is to collect and annotate extra data from target construction sites and fine-tune the models. However, the manual annotation is time-consuming and error-prone, introducing noise and potential privacy concerns. Furthermore, the current focus of construction site datasets predominantly lies in 2D object detection, tracking, and instance segmentation. Other critical perception tasks witnessed in autonomous driving research, such as depth estimation or 3D object detection, remain largely unexplored in construction sites. This is mainly constrained by the significantly increased data collection and annotation cost associated with these tasks. To tackle these challenges, exploring synthetic datasets has emerged as a promising and vital research direction.

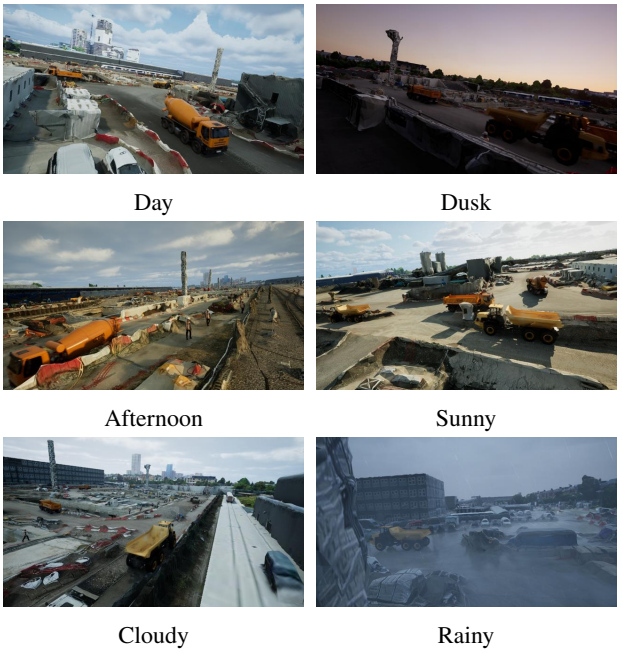


Figure 1: ConSynth contains video sequences initialised and rendered with different domain parameters

Research gaps

While synthetic construction datasets provide significant potential, investigations in current literature have revealed several limitations. One common method involves rendering 3D models of construction equipment or personnel and superimposing them onto backgrounds of construction sites, known as the crop and paste method (Lee et al., 2022). However, the generated images lack realism due to inconsistencies in camera perspective and lighting with the background images. Alternative methods attempt to craft a construction site manually using 3d modelling software and then render the models of construction objects within it (Soltani et al., 2016; Neuhausen et al., 2020; Lee et al., 2023). Yet, hand-crafting scenes fail to capture the details and complexity of real-world sites. Moreover, manual creation is often limited to specific research contexts and is difficult to adapt to general construction projects. Con-

sequently, models trained on these synthetic datasets tend to be constrained by substantial domain gaps, undermining their performance in actual application scenarios. In the field of autonomous driving, strategies such as transfer learning, domain randomisation, and domain adaptation are typically employed to bridge this gap. However, their application in construction site monitoring is still rare. Lastly, most datasets for the construction industry contain only images. These discontinuous images, sampled at different intervals, restrict the extraction of spatio-temporal features of the scene.

Contributions

The contributions of this paper are two-fold. First, we propose a digital twin-based simulator that simulates realistic vehicle dynamics and worker behaviours. The simulator also adapts domain randomisation with the weather and lighting parameters to minimise domain gaps between reality and simulation. Second, our proposed synthetic dataset, 'ConSynth', is the first synthetic video dataset in the construction industry with the aim to broaden the scope of task applicability. This novel dataset provides continuous image sequences with precise labels for traditional computer vision tasks, such as 2D and 3D object detection, tracking, and instance segmentation, and enables additional tasks such as depth estimation, action recognition and behavioural analysis. Our research showcases that combining photo-realistic synthetic and real-world data significantly improves model robustness and accuracy in construction site scenarios, effectively reducing the reliance on extensive real-world data collection and manual annotation. Our findings highlight the potential of synthetic data in enhancing model robustness, offering a solution for deploying computer vision in dynamic and challenging construction environments.

Related Works

In this section, we review existing research on developing datasets for deep learning applications. Initially, it outlines the current state of real-world datasets in autonomous driving, underscoring the emergence of synthetic datasets as a promising alternative to manual data collection. Following this, the section discusses the development of synthetic datasets in the construction industry.

Real-world Datasets

Deep learning models are highly data-dependent. Over the past decade, the rapid advancement in autonomous driving research has been significantly fuelled by vast amounts of data from real-world sensors (Feng et al., 2021). The collaboration between the industrial and academic sectors has led to the annotation and release of several mature datasets, such as KITTI (Geiger et al., 2013), nuScenes (Caesar et al., 2019), and Argoverse (Chang et al., 2019). These datasets have catalysed the progress in autonomous driving research by supporting various tasks, notably 2D/3D detection and tracking, depth estimation, and stereo vision.

In the construction industry, researchers are also developing datasets to train deep-learning models. Xuehui et al. (2021) released the first large-scale construction site dataset, MOCS, which offers around 42,000 images covering 13 common construction equipment categories, supporting 2D object detection and segmentation. Similarly, the ACID dataset (Xiao and Kang, 2021) provides 10,000 images covering ten types of construction equipment, targeted primarily for 2D object detection. They have paved the way for deploying 2D detectors like Faster-RCNN, YOLO, and Mask RCNN in the industrial sector. However, compared with the autonomous driving sector, datasets for the construction industry are still relatively scarce and limited to a narrow range of tasks.

Synthetic Datasets in Autonomous Driving

The development of synthetic datasets is emerging as a solution to the challenges of data collection. In autonomous driving, a pioneering example is the dataset introduced by Richter et al. (2016), which utilises highly realistic imagery from commercial video games and intricate simulations of interactions between pedestrians and vehicles to provide pixel-level semantic segmentation labels. Combining synthetic with real datasets such as KITTI (Geiger et al., 2013) has enhanced model training accuracy while reducing reliance on expensive manual annotation of real-world data. Similarly, SYNTHIA (Ros et al., 2016) generates semantic segmentation labels for driving scenarios using a simulator built with the Unity game engine. Experiments in this work have demonstrated that models trained on SYNTHIA and fine-tuned with real-world datasets show improved accuracy. VIPER (Richter et al., 2017) extends the scope by providing annotations for multiple tasks, including 2D/3D object detection, multi-object tracking, and optical flow estimation. Additionally, it includes data from five different environmental conditions and diverse urban landscapes. AIODrive dataset (Weng et al., 2020) provides point cloud data from LiDAR sensors in varying densities, in addition to complete annotations for object detection, and covers diverse scenarios and weather conditions. CarlaScenes (Kloukiniotis et al., 2022) provides a practical solution for annotating real-world scene data for multiple challenging scenarios using the CARLA simulator (Dosovitskiy et al., 2017). SHIFT (Sun et al., 2022), another dataset based on the CARLA simulator (Dosovitskiy et al., 2017), encompasses the 13 most significant perception tasks in the domain of autonomous driving. The continuous domain shifts, a distinctive feature of SHIFT, enable the researchers to evaluate the 'domain gaps' and benchmark domain transfer algorithms.

Synthetic Datasets in Construction

To date, the synthetic approach has not been thoroughly explored in the research domain of construction. Most studies used synthetic datasets for specific tasks where manual data annotation is challenging. For example, stud-

(a) Crop and paste (Lee et al., 2022) (b) 3D rendering (Barrera-Animas and Delgado, 2023) (c) Digital twin-based simulation **(Ours)**

Figure 2: Comparison of methods used in generating construction synthetic dataset. Our simulation system generates more realistic images in terms of texture and environment.

ies by Soltani et al. (2016); Lee et al. (2022) rendered 3D models of excavators in various poses and superimposed them onto diverse construction site backgrounds to enhance the accuracy of detecting excavators in images, particularly smaller objects (Figure 2a). Similarly, Assadzadeh et al. (2022) synthesised images of excavators for training object detection models and extended the research to quantitatively analyse the role of domain randomisation in bridging the gap between synthetic and real-world images. Neuhausen et al. (2020) used manually created scenes and pre-captured motion animations to render 3D worker models under varying lighting and weather conditions, aiming to improve the tracking accuracy of construction workers. Additionally, Lee et al. (2023) employed a game engine to create a synthetic dataset to enhance the detection precision of workers' small personal protective equipment for safety monitoring. Barrera-Animas and Delgado (2023) expanded the research scope by covering multiple equipment categories, combining a series of 3D models, and rendering them under diverse lighting scenarios. However, the effectiveness of these datasets is still in question due to their severe lack of photorealism. Results indicated that models trained on these datasets performed poorly when tested on real-world data Barrera-Animas and Delgado (2023). Despite these limitations, these pioneering datasets have laid the groundwork for developing more complex, multi-task synthetic datasets in construction safety research.

Developing Construction Simulation System and Synthetic Video Dataset

In this paper, we created a novel construction simulation system to simulate the behaviours of various agents under different lighting and weather conditions. Within the system, we generated a novel synthetic video dataset, ConSynth, featuring annotations for various tasks. The overall process can be found in Figure 3.

Generation of the Digital Twin

As discussed in the previous section, existing methods generated images with a lack of realism. In this work, we propose using a digital twin model to preserve the details of actual sites. First, we selected an actual construction site in northwest London, UK, as the case study area. Then, a UAV was scheduled to follow a pre-designed route over the site, capturing images at a frequency of once per week. We estimated the three-dimensional structure using the Structure from Motion (SfM) algorithms based on images and the known camera parameters during capture. This process involved the use of Bentley ContextCapture software.

Although our digital twin model was regularly updated to reflect the continuous changes at the construction site, it only includes the case study site and a small surrounding area. A large proportion of the background was missing when we used cameras within the simulator to capture images at a flat angle. Tremblay et al. (2018) found that diversity in the background of synthetic datasets is crucial, as it encourages the model to learn the most representative features of the target objects, thereby distinguishing them from the complicated background. Expanding the UAV capturing area would be computationally costly and time-consuming. As an alternative, we integrated the 3D model from Google Earth to fill the distant scenery. After localising the scanned model in a geographic coordinate system, we removed the scanned region from Google's 3D model and overlaid our model. This approach ensures the most recent scanned model is used for the construction site area, while Google Earth, with less frequent changes, is used as the distant scenery.

Simulation of Site Logistics

The next step involves simulating the agents, such as the construction equipment and workers, with realistic behaviours. This simulation comprises three parts: (1) site road configuration, (2) generation of site agents, and (3) movement of each agent.

To accurately represent the site road configuration, our approach involves creating a vector map in alignment with the site's logistic plan (Figure 5). We utilised the map handling framework Lanelet2 (Poggenhans et al., 2018), a tool prevalent in the autonomous driving sector for map reading and trajectory planning. These vector maps, essential

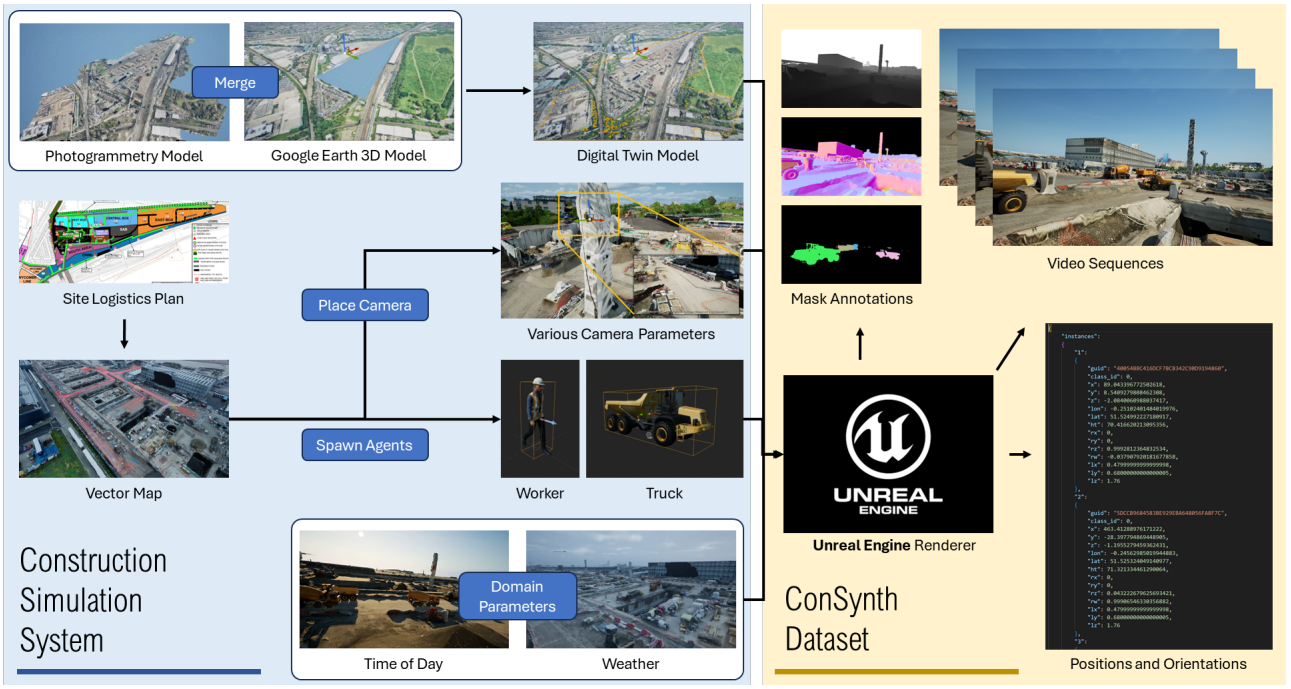


Figure 3: Diagram of the overall process to generate the synthetic video dataset within the construction simulation system

for autonomous driving, offer detailed vector-based road information, including lanes, crossings, and intersections. Leveraging Lanelet2, we implemented the behavioural logic for agent movements. For instance, trucks are programmed to travel on vehicle lanes and navigate through intersections, while workers are set to follow pedestrian pathways and crossings. These movements are governed by the underlying logic provided by the Lanelet2 framework to ensure a more realistic simulation.

We then developed a rule-based system for generating agents based on the vector map. Initially, we tagged haul roads and one-way roads with a 'vehicle' tag and intersections with both 'vehicle' and 'intersection' tags. Pedestrian footways and crossings received a 'pedestrian' tag. The generation follows these rules:

1. Vehicles, such as dump trucks and mixer trucks, are only generated on lanes tagged with 'vehicle' and not 'intersection', with a minimum spacing of $S_{min,veh}$ and a maximum of N_{veh} vehicles on the map;
2. Pedestrians are generated only on lanes tagged 'pedestrian', with a minimum spacing of $S_{min,ped}$ and a maximum of N_{ped} pedestrians on the map;
3. For candidate lanes tagged 'pedestrian', we check if the lane intersects with any lane tagged with 'vehicle', identifying it as a pedestrian crossing.
4. For each lane tagged 'intersection', we use the $k-NN$ algorithm to cluster lanes belonging to the same intersection and generate a special agent as the controller: the intersection agent.

All agent movements are based on their respective lanes. Vehicles move on lanes tagged with 'vehicle', randomly choosing a connected lane at the end of their current lane. If the upcoming lane belongs to an intersection, the corresponding intersection agent will take over the control when a vehicle approaches the stop line. Similarly, pedestrians move only on pedestrian-tagged lanes and randomly select the next lane. If a pedestrian crossing is found to be chosen as the next lane, the agents check if the crossing holds a 'walkable' tag. If the tag is absent, they wait at the end of the crossing. Otherwise, they continue to walk through the crossing. Each intersection agent, simulating a traffic light, sequentially checks its managed lanes with a 'vehicle' tag. For each lane, if a vehicle is found waiting at a stop line, the agent allows it to pass through the intersection and assigns it to a lane at the other end. After a round of checks, a 'walkable' tag is added to all managed pedestrian crossings for a duration of t, allowing all pedestrian agents to cross the vehicle lanes freely.

Domain Randomisation

One major drawback of the current public datasets in construction is that most samples were collected under good illumination conditions. Consequently, models trained on these datasets struggle to handle adverse weather and lighting conditions effectively. To improve model robustness under varying conditions, such as night and rainy and foggy weather, we propose the use of domain randomisation. Specifically, we randomised the weather and time of day in the game engine to create scenes with diverse appearances (Table 1).

Camera Placement

We developed an algorithm for camera placement to ensure that they capture agent movements in the camera view rather than just the background. To begin with, we ran-

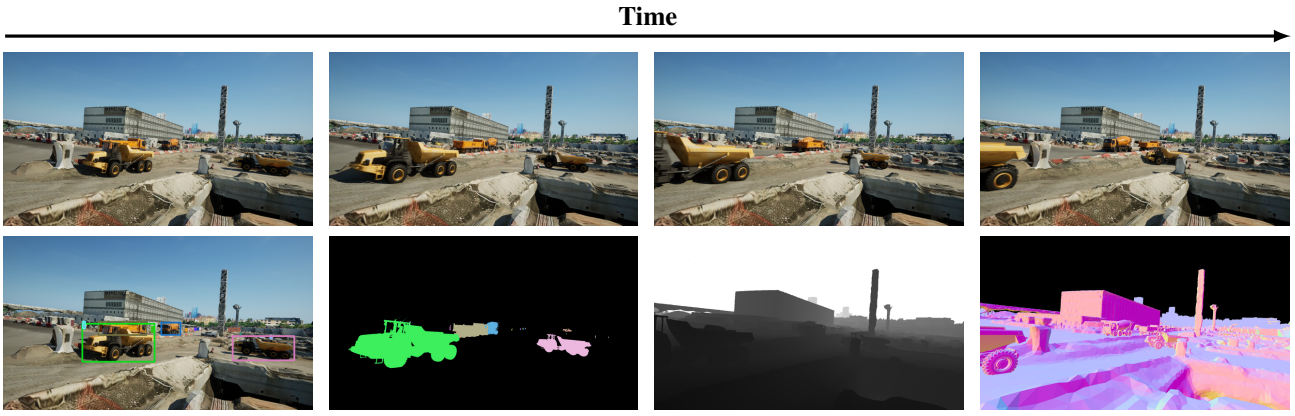


Figure 4: ConSynth provides continuous image sequences with precise labels of 2D bounding boxes, instance segmentation, depth map and normal map. The dataset supports both image-based and video-based computer vision tasks including object detection and tracking, instance segmentation, depth estimation, action recognition and behaviour analysis.

Figure 5: Vector map for the case study construction site

Table 1: Distribution of domain randomisation parameters

Parameter	Variations	Probability
Weather	Clear sky	0.20
	Cloudy (Regular Cloud Cover)	0.10
	Cloudy (Partial Cloud Cover)	0.10
	Foggy	0.20
	Rainy (Light Rain)	0.10
	Rainy (Thunderstorm)	0.10
	Overcast	0.20
Time of Day	Daylight	0.75
	Night-time	0.25

domly selected a point on one of the lanes from the vector map. Then, a centre was set between z_{min} and z_{max} above this point, generating a semi-sphere with a radius between r_{min} and r_{max}. A point on the surface of this semi-sphere was chosen randomly as the initial camera position. We then performed ray tracing from the centre to this point to avoid obstructions and repeated the selection process until the view was unobstructed. A camera was then positioned at this point, oriented towards the centre. Finally, we randomised the camera's intrinsic and extrinsic parameters, including the field of view and orientation, to ensure our dataset covers a variety of perspectives.

Automated Annotation

We leveraged Unreal Engine 5 for rendering photo-realistic RGB images and mask annotations. In addition, we developed a C++-based tool for automated annotation that, for each rendered frame, collects 3D transformations of all agents in the scene, including relative coordinates (x, y, z), world coordinates $(latitude, longitude, height)$, rotation angles (r_y), and dimensions of 3D bounding boxes (l_x, l_y, l_z). It also gathers camera intrinsic and extrinsic parameters, exporting them along with the agent transformations. The segmentation annotations are initially exported as mask images, where a unique colour represents each agent. Following the generation, we utilised OpenCV (Bradski, 2000), a public computer vision library for image processing, to identify these colours and extract polygon-based segmentation data. Finally, the tool calculates the smallest enclosing rectangle from the segmentation as the 2D bounding box for each agent. The automated annotation system, illustrated in Figure 4, is of significance in reducing the time and resources required for the generation of extensive, high-quality and accurate labelled datasets.

Dataset Statistics

Our ConSynth dataset encompasses 200 video sequences, each uniquely produced within the digital twin environment using randomly varied parameters and captured from different locations (see Figure 1). Every sequence consists of 120 frames, recorded at 12 frames per second. Within the dataset, we focus on three primary categories: workers, dump trucks, and mixer trucks. The most populous category is the workers, with 237,215 instances, which is twice the number of dump truck instances. Mixer trucks, being the least common type, have 31,800 instances. Figure 6 details the distribution across these categories. No-

tably, the dataset totals 24,000 images, which results in an average of around 15 instances per image. This density of instances is vital for training robust models, offering a diverse and comprehensive range of data for each category.

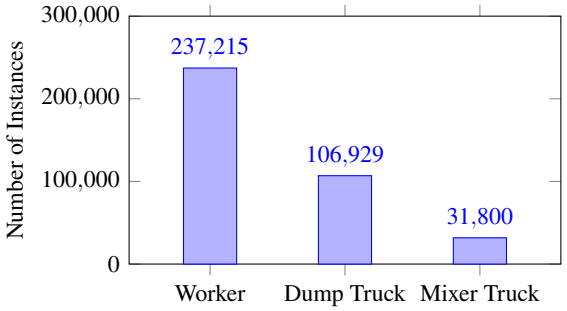


Figure 6: Distribution of different categories (workers, dump trucks, and mixer trucks) in the ConSynth dataset.

The images generated in the ConSynth dataset have a resolution of 1920×1080 pixels. We referred to the standards for object sizes set by the COCO dataset to analyse the distribution of objects of various sizes. According to COCO, in images of 640×480 resolution, objects with an area less than 32×32 pixels are classified as small objects, and those less than 96×96 pixels as medium objects, with the remainder being large objects. We scaled this standard proportionally to match our resolution, resulting in the distribution of small, medium, and large objects in our dataset being 63%, 24%, and 13%, respectively.

This significant presence of small objects in the dataset is a critical feature, as it aligns with real-world scenarios where computer vision models need to identify small yet significant objects on construction sites, commonly workers. To empirically evaluate the effectiveness of the ConSynth dataset in addressing real-world challenges, the next section will present a series of experiments. These experiments are designed to test the dataset's performance in enhancing object detection in real-world construction sites, particularly of small objects.

Experiments

In this section, we explore the effectiveness of using the synthetic ConSynth dataset in the field of object detection, with three focuses: enhancing real data with ConSynth to improve model performance, reducing the reliance on real data by leveraging ConSynth, and evaluating the performance of ConSynth in a case-study construction site.

Training and Validation Setup

We employed the YOLOv8 model, a commonly used open-source one-stage object detector known for its anchor-free design, fast speed, and high accuracy. We used two datasets for training: a real dataset and a synthetic dataset (i.e. ConSynth). The real dataset combines the largest available construction site datasets, MOCS (Xuehui et al., 2021) and ACID (Xiao and Kang, 2021). For ACID, we split it into 80% for training and 20% for validation. The 80% from ACID was merged with the MOCS training set

(a) Real (MOCS + ACID) (b) Synthetic (ConSynth) (c) Case study

Figure 7: Examples from the experiment datasets. ConSynth dataset provides images similar to images from the case study dataset but targets adverse lighting and weather conditions.

to form our training dataset, and the 20% was combined with the MOCS validation set for our validation dataset. The final composition of the real dataset included 24,233 images in the training set and 5,219 images in the validation set. For the ConSynth dataset, encompassing 24,000 images, we similarly allocated 80% (19,200 images) for training and 20% (4,800 images) for validation.

We also created another dataset, the case study dataset, for validation independent from any of the training datasets. We collected 500,000 images from real cameras at the case study construction site, which covers various construction equipment, personnel, weather, and lighting conditions. From these, 10,000 images were randomly selected following a normal distribution. Using a MobileNetv2 model (Sandler et al., 2019) trained with the ImageNet dataset (Deng et al., 2009), we computed the visual embeddings for each image and selected the 100 most unique images based on these embeddings. These images were then manually annotated and reviewed by experienced engineers. Examples from these datasets can be seen in Figure 7.

The training was conducted on four RTX 4090 24G GPUs with a batch size of 16 over 100 epochs. The base learning rate was set to $lr_0 = 0.002$, employing an SGD optimiser with a learning rate factor of $lr_f = 0.01$. The size of the input image was maintained at 640×640. After the training, a key performance metric, mean Average Precision (*mAP*), was employed to evaluate each model's performance. The *mAP* is defined as the percentage of correct predictions, algebraically represented as:

$$mAP = \frac{TP}{TP + FP} \tag{1}$$

where TP denotes the number of true positives, FP represents the number of false positives.

Enhancing Model Performance with Synthetic Data

This section explores the impact of using the synthetic dataset ConSynth on model performance in three scenarios: (1) training separately on real and synthetic train sets and validating on their respective validation sets; (2) training separately on real and synthetic train sets and validating on the other validation set; (3) training on combined train sets and validating on combined validation sets. The results are detailed in Table 2.

Our research found that models trained on both the real and synthetic datasets performed well on their respective validation sets, indicating the effectiveness and domain-

Table 2: Performance Comparison of Models. Notably, models trained solely on the Consynth dataset improved the accuracy of detecting dump trucks by 20%. Models trained on the combined datasets improved the accuracy of detecting dump trucks by 40% and all classes by 13%.

Train	Validation	Worker	Dump Truck	Mixer Truck	All Classes
Real	Real	62.6	67.2	78.3	69.3
Synthetic	Synthetic	31.3	69.8	68.1	56.4
Real	Synthetic	6.1	11.4	10.2	9.3
Synthetic	Real	9.0	3.8	4.3	5.7
Real + Synthetic	Real	62.8	65.1	75.6	67.8
Real + Synthetic	Synthetic	25.9	64.7	61.9	50.8
Real	Case Study	10.4	6.4	13.1	9.9
Synthetic	Case Study	1.9	7.7 (+1.3)	3.1	4.2
Real + Synthetic	Case Study	**10.5** (+0.1)	**9.0** (+2.6)	**14.0** (+0.9)	**11.2** (+1.3)

specific robustness of the datasets. However, the performance of these models on the opposite validation sets was poorer. This performance drop is primarily attributable to the domain gap between the real and synthetic datasets. Despite our efforts to minimise this gap through photo-realistic rendering and domain randomisation during the data generation, a discernible difference remains. Another reason for the performance discrepancy is that the ConSynth dataset more closely resembles the data collected from the specific construction site used in our case study. The distribution of object types, appearances, and sizes in ConSynth differ from those in the public construction datasets, creating an inherent domain gap that will be further analysed in subsequent sections.

We also combined the real and synthetic datasets to bridge the gap between domains. Interestingly, the model trained on the combined dataset showed a slight decrease in performance metrics in each domain. However, this does not necessarily indicate a deterioration in model performance. Instead, it suggests that the model has adapted to data from multiple domains, leading to a more generalised model. This generalisation is a positive attribute, especially for practical applications in complex construction sites where conditions vary significantly.

To validate this generalisation, we evaluated all trained models on the dataset created and annotated from the case study construction site. As shown in Table 2, the model trained on the synthetic dataset has shown promising improvement in dump truck detection as the Mean Average Precision (*mAP*) improved by 20%. The model trained on the combined dataset outperformed the other two models in all aspects. Notably, the improvement in dump truck detection was significant, jumping from 6.4 to 9.0 by 40% while the *mAP* for all classes was improved by 13%. However, despite these improvements, all three models' *mAP* was generally low. This is attributed to the high resolution of site cameras and the distance of these cameras from the targets, resulting in small object sizes in the images. The down-sampling of high-resolution images from 1920×1080 pixels to 640×360 pixels for convolution operations in the backbone likely leads to a loss of critical features, intensifying the challenge of detecting small objects. Yet, with the presence of a large number of small objects in the dataset, the ConSynth dataset helps enhance the model's performance in detecting small objects on construction sites.

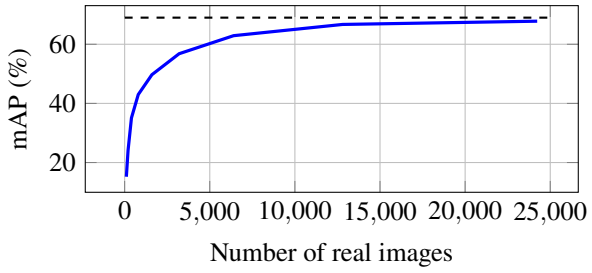
Finally, we conducted a quantitative analysis of video footage from our case study construction site to investigate how the ConSynth dataset enhances model performance in specific construction scenarios. The results, depicted in Figure 8, compare the performance of models: (a) trained solely on the real dataset and (b) trained exclusively on the ConSynth dataset, with the latter having no exposure to real dataset images. The findings demonstrate that the model can effectively learn features of different classes from the ConSynth dataset and make accurate predictions, even without real data. In this comparison, the model trained with ConSynth data notably reduced false positives in the worker category and improved the detection of overlapping dump trucks, thereby reducing false negatives. However, it struggled to detect workers partially obscured by railings, a limitation likely due to a lack of representative samples in the synthetic dataset. This issue can be mitigated by combining the ConSynth dataset with real data. Overall, the ConSynth dataset showed promising results, effectively reducing reliance on real data and demonstrating significant potential.

(a) Model trained on real dataset (b) Model trained on ConSynth dataset

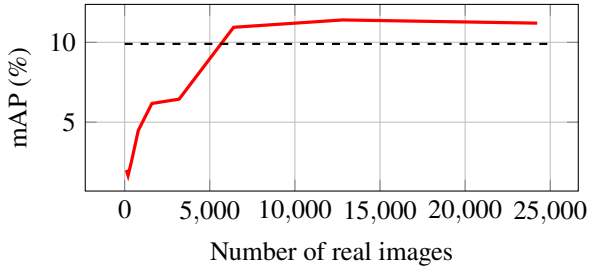
Figure 8: Models trained solely on the ConSynth dataset yield better accuracy in detecting dump trucks and comparable accuracy in detecting workers.

Relieving the Burden of Manual Labelling

In this section, we primarily explore how synthetic datasets can alleviate the reliance on real-world data. We randomly selected k samples from a real training set comprising 24,233 images. The sampling followed a normal distribution, where $k \in [100, 200, 400, 800, 1600, 3200, 6400, 12800]$. We combined the selected samples with the entire synthetic dataset to form eight distinct training sets. These datasets were then used for training with the same hyper-parameters from the previous section.



(a) Model performance (mAP) on the real dataset



(b) Model performance (*mAP*) on the case study dataset

Figure 9: Comparison of model performance in response to varying proportions of real data in training. Reliance on real data can be significantly reduced with synthetic data, especially in underrepresented scenarios targeted by synthetic data.

Validation was performed on the real dataset's validation set, and the results are depicted in Figure 9a as a solid blue line. The dashed line represents the baseline ($mAP = 69.0\%$), which is the performance using only the real dataset for training and validation. The graph indicates that as the proportion of real data in the mixed dataset increases, the model's performance on the real training set improves, signifying the learning of domain-specific features. While the mixed dataset did not surpass the baseline, it is notable that when the real data was reduced by 47.2% (from 24,233 to 12,800), the model's performance remained close, with *mAP*s of 67.8 and 66.7, respectively. Further reduction of real data to 26.4% (from 24,233 to 6,400) still resulted in a significant *mAP* (62.9%). This demonstrates that mixing synthetic with real data for training substantially decreases the dependency on real data.

We also verified the performance of models trained on the mixed datasets using data collected from a case study construction site, as shown in Figure 9b. The dashed line represents the performance of models trained on real data for this case study dataset. The trend observed was similar to the previous experiment, where reducing the real data in the dataset by approximately 75% still yielded comparable performance. Notably, the model exceeded the baseline performance on this dataset, indicating the use of synthetic data in improving the model's robustness. This can be attributed to using a digital twin model closely resembling the actual construction site in the synthetic data generation process, which aids in improving the model's performance in underrepresented site scenarios.

Conclusions

In this work, we have introduced a digital twin-based simulation system for generating synthetic datasets for construction site scenarios. Our method outperforms previous approaches by producing photo-realistic images to enhance the overall model performance in various construction scenarios. Moreover, we have introduced ConSynth, a novel synthetic dataset. The dataset is the first video dataset in construction that supports common perception tasks and provides valuable annotations for potential tasks which have not been extensively applied in the industry.

Currently, ConSynth includes three different categories of construction personnel equipment. Future work could expand the dataset's scope by incorporating additional construction equipment, such as excavators, loaders, and cranes, into the simulation. This expansion would further enhance the dataset's applicability and relevance to real-world construction scenarios. In addition, the models evaluated in this study exhibit a relatively low mean Average Precision (mAP) for detecting small objects in high-resolution images. Future work should conduct extensive experiments to explore improvements in mAP, potentially through methods like data augmentation and advanced algorithms, including the Slicing Aided Hyper Inference (SAHI) (Akyon et al., 2022).

We anticipate that our approach will not only improve the accuracy and robustness of computer vision applications in dynamic and challenging construction environments but also significantly reduce the costs associated with data collection and labelling. With a richer dataset, computer vision models are expected to not only accurately locate key targets and detect potential hazards but also contribute to the proactive enhancement of safety monitoring on construction sites, thereby mitigating risk and ensuring the well-being of personnel. In light of this, the ConSynth dataset will be made available at `https://github.com/tsl-imperial/ConSynth`.

Acknowledgments

The authors wish to extend their sincere gratitude to their industry collaborator, Balfour Beatty VINCI SYSTRA Joint Venture (BBVS JV), for their invaluable support and collaboration in this research. The provision of funding and access to the case study construction site for data collection were integral to the success of this study. Their contributions have been fundamental in advancing the research presented in this paper.

References

Akyon, F. C., Altinuc, S. O., and Temizel, A. (2022). Slicing aided hyper inference and fine-tuning for small object detection. 2022 IEEE International Conference on Image Processing (ICIP), pages 966–970.

Assadzadeh, A., Arashpour, M., Brilakis, I., Ngo, T., and Konstantinou, E. (2022). Vision-based excavator pose estimation using synthetically generated datasets with domain randomization. Automation in Construction, 134:104089.

Barrera-Animas, A. Y. and Delgado, J. M. D. (2023). Generating real-world-like labelled synthetic datasets for construction site applications. Automation in Construction, 151:104850.

Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2019). nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027.

Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., and Hays, J. (2019). Argoverse: 3d tracking and forecasting with rich maps.

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

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. (2017). CARLA: An open urban driving simulator. In Levine, S., Vanhoucke, V., and Goldberg, K., editors, Proceedings of the 1st Annual Conference on Robot Learning, volume 78 of Proceedings of Machine Learning Research, pages 1–16. PMLR.

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Gläser, C., Timm, F., Wiesbeck, W., and Dietmayer, K. (2021). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems, 22(3):1341–1360.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR).

Kloukiniotis, A., Papandreou, A., Anagnostopoulos, C., Lalos, A., Kapsalas, P., Nguyen, D.-V., and Moustakas, K. (2022). CarlaScenes: A synthetic dataset for odometry in autonomous driving. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE.

Lee, H., Jeon, J., Lee, D., Park, C., Kim, J., and Lee, D. (2023). Game engine-driven synthetic data generation for computer vision-based safety monitoring of construction workers. Automation in Construction, 155:105060.

Lee, J. G., Hwang, J., Chi, S., and Seo, J. (2022). Synthetic image dataset development for vision-based construction equipment detection. Journal of Computing in Civil Engineering, 36(5).

Neuhausen, M., Herbers, P., and König, M. (2020). Using synthetic data to improve and evaluate the tracking performance of construction workers on site. Applied Sciences, 10(14):4948.

Poggenhans, F., Pauls, J.-H., Janosovits, J., Orf, S., Naumann, M., Kuhnt, F., and Mayr, M. (2018). Lanelet2: A high-definition map framework for the future of automated driving. In Proc. IEEE Intell. Trans. Syst. Conf., Hawaii, USA.

Richter, S. R., Hayder, Z., and Koltun, V. (2017). Playing for benchmarks. In 2017 IEEE International Conference on Computer Vision (ICCV). IEEE.

Richter, S. R., Vineet, V., Roth, S., and Koltun, V. (2016). Playing for data: Ground truth from computer games.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2019). Mobilenetv2: Inverted residuals and linear bottlenecks.

Soltani, M. M., Zhu, Z., and Hammad, A. (2016). Automated annotation for visual recognition of construction resources using synthetic images. Automation in Construction, 62:14–23.

Sun, T., Segu, M., Postels, J., Wang, Y., Van Gool, L., Schiele, B., Tombari, F., and Yu, F. (2022). Shift: A synthetic driving dataset for continuous multi-task domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 21371–21382.

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., and Birchfield, S. (2018). Training deep networks with synthetic data: Bridging the reality gap by domain randomization.

Weng, X., Man, Y., Cheng, D., Park, J., O'Toole, M., and Kitani, K. (2020). All-In-One Drive: A Large-Scale Comprehensive Perception Dataset with High-Density Long-Range Point Clouds. arXiv.

Xiao, B. and Kang, S.-C. (2021). Development of an image data set of construction machines for deep learning object detection. Journal of Computing in Civil Engineering, 35(2).

Xuehui, A., Li, Z., Zuguang, L., Chengzhi, W., Pengfei, L., and Zhiwei, L. (2021). Dataset and benchmark for detecting moving objects in construction sites. Automation in Construction, 122:103482.