

ISC-Perception: A Hybrid Computer Vision Dataset for Object Detection in Novel Steel Assembly

Miftahur Rahman¹, Samuel Adebayo¹ *Member, IEEE*, Dorian A. Acevedo-Mejia², David Hester¹, Daniel McPolin¹, Karen Rafferty³, and Debra F. Laefer⁴

Abstract—The Intermeshed Steel Connection (ISC) system, when paired with robotic manipulators, can accelerate steel-frame assembly and improve worker safety by eliminating manual assembly. Dependable perception is one of the initial stages for ISC-aware robots. However, this is hampered by the absence of a dedicated image corpus, as collecting photographs on active construction sites is logistically difficult and raises safety and privacy concerns. In response, we introduce ISC-Perception, the first hybrid dataset expressly designed for ISC component detection. It blends procedurally rendered CAD images, game-engine photorealistic scenes, and a limited, curated set of real photographs, enabling fully automatic labelling of the synthetic portion. We explicitly account for all human effort to produce dataset, including simulation engine and scenes setup, asset preparation, post-processing scripts and quality checks; our total human time to generate a 10,000-image dataset was 30.5 h versus 166.7 h for manual labelling at 60 s per image (-81.7%). A manual pilot on a representative image with five instances of ISC members took 60 s (maximum 80 s), anchoring the manual baseline.. Detectors trained on ISC-Perception achieved a mean Average Precision at IoU 0.50 of 0.756, substantially surpassing models trained on synthetic-only or photorealistic-only data. On a 1,200-frame bench test, we report mAP@0.50/mAP@[0.50:0.95] of 0.943/0.823 . By bridging the data gap for construction-robotics perception, ISC-Perception facilitates rapid development of custom object detectors and is freely available for research and industrial use upon request.

Index Terms—ISC, robotics, structural steel assembly, automation, computer vision

I. INTRODUCTION

A. Background and Motivation

ROBOTIC manipulators have transformed factory-based manufacturing, yet their impact on construction remains modest. Unlike shop floors, building sites are unstructured, weather-exposed, and governed by stringent safety constraints, all of which complicate autonomous operation. Steel frame erection, which is one of the most labour-intensive, high-risk phases of a build, stands to benefit most from automation: cranes dominate the critical path while transporting heavy steel

items, bolting demands skilled crews working at height, and schedule delays cascade to all downstream trades [1]. A robot capable of recognising, grasping, and interlocking structural members in situ could shorten crane time, lower accident rates, and mitigate skilled-labour shortages.

In building construction, structural steel frames are composed of individual beams and columns that are typically assembled on-site due to the challenges associated with transporting large assemblies. This process involves several key steps: (1) identifying and lifting each structural element from the storage area, (2) transporting it to the installation location; (3) aligning it with the existing structure, and (4) fastening it to the structural frame using bolts or welds [1]. Although both methods are common for connecting steel members, bolts are generally preferred on-site due to their ease of installation, faster connection times, better quality control, and reduced inspection requirements. However, the extensive use of bolts in structural steel connections introduces additional challenges for the deployment of robots in the field.

The recently proposed Intermeshed Steel Connection (ISC) system eliminates most of the temporary bolts required by conventional moment or shear splices. The ISC can be manufactured using cutting-edge technologies such as high-density plasma cutting, water jet cutting, and laser cutting [2]. ISC has two types of components: ISC member and ISC connection plates. Initial design of ISC has 3 connection plates on one side (Fig. 1(a)) but the newer version requires only one connection plates on each side with fewer number of bolts (Fig. 1(b)). Precision-cut male–female tabs guide members into alignment so that only a handful of set-bolts are needed to secure the joint [2], [3]. By trimming cycle times and tolerating direct reuse, ISC reduces material waste and greenhouse-gas emissions while preserving structural capacity. These benefits align with industry trends towards design-for-manufacture-and-assembly (DfMA) [4] and circular construction. However, the ISC’s unconventional geometry poses fresh challenges for computer vision: the connection plates are compact, partially occluded once mated, and often coated with reflective galvanisation.

Robots cannot exploit ISC unless they can reliably identify connection plates, member ends, and mating features under dynamic lighting and clutter. Conventional weld or bolt heads provide distinctive geometry; ISC plates, by contrast, are largely planar and differ only by tab pattern. No public image corpus captures these subtleties, and collecting site photographs is fraught with *access restrictions, privacy regulations, and weather-dependent scheduling*. Consequently, vision models trained on generic construction datasets (e.g.,

¹School of Natural and Built Environment, Queen’s University Belfast, Belfast, United Kingdom .

E-mail: {miftahur.rahman, samuel.adebayo, d.hester, d.mcpolin}@qub.ac.uk

²School of Civil & Environmental Engineering, and Construction Management, University of Texas at San Antonio, USA.

Email: dorian.acevedo@utsa.edu

³School of Electronics, Electrical Engineering, and Computer Science, Queen’s University Belfast, Belfast, United Kingdom.

Email: k.rafferty@qub.ac.uk

⁴Department of Civil and Urban Engineering, New York University, USA. Email: debra.laefer@nyu.edu

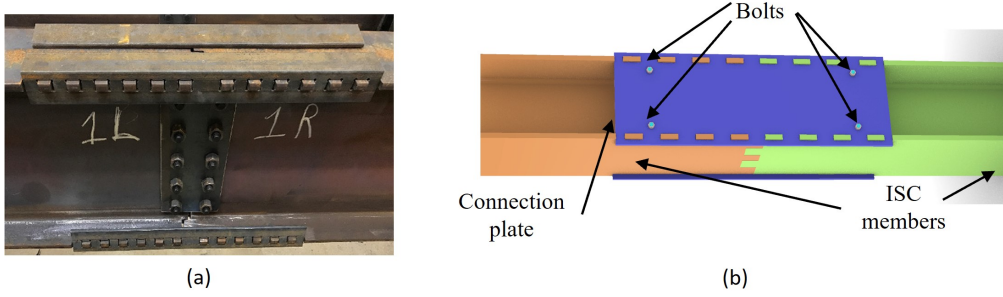


Fig. 1. Components of ISC beam-to-beam; (a) earlier version of fabricated ISC [3], (b) CAD drawing of ISC with single connection.

MOCS [5], SODA [6]) fail to generalise for robotic assembly tasks. Bridging this domain gap requires a *task-specific dataset* that mixes real jobsite context with photorealistic renders and procedurally generated scenes to achieve both scale and fidelity.

The economic stakes are high: the global structural-steel market exceeded USD 110B in 2023 and is projected to grow at approximately 5%, CAGR through 2030, driven by urban densification and industrial expansion [7]. While connection labour alone can account for up to 25% of erection cost [8], [9]. Hence, A vision-enabled robotic ISC assembly workflow has been proposed, which holds the promise of safer jobsites and substantially reduce cost and schedule savings [10].

B. Images as Fuel for Training Vision Models

Progress in deep learning has been driven less by algorithmic novelty than by the availability of vast, well-annotated image corpora. Construction-robotics vision imposes even steeper data demands: detectors must recognise partially occluded objects, track them under harsh illumination, and generalise across projects that vary by geometry, finish, and weather [11]–[16].

Image data plays a pivotal role by forming a foundation for training and testing computer vision and machine learning algorithms, providing essential visual information for tasks such as object detection, image classification, and semantic segmentation [17]. Without access to high-quality and diverse image datasets, the performance (accuracy and generalization) of these algorithms can be significantly hindered. As computer vision technology rapidly advances in the construction industry, the demand for accurate and comprehensive interpretation of construction site imagery has become increasingly urgent [18].

Generic datasets such as COCO or ImageNet misrepresent site reality: they lack steel members, cranes, PPE, and the dense clutter typical of erection yards. Direct transfer can depress mean Average Precision (mAP) by up to 40% when models are tested on construction imagery [18]. Building an in-domain corpus is equally fraught. Cameras are often barred by safety briefings, union rules, or privacy regulations; outdoor shoots hinge on weather windows; and pixel-accurate annotation of high-resolution frames can consume weeks of person-hours [19]. The hurdle is steeper still for bespoke components

such as the ISC, for which no archival photographs yet exist and whose galvanised surfaces frustrate automated labelling.

Data, not algorithms, has thus become the principal bottleneck. An effective remedy must supply (i) *scale* for deep networks, (ii) *fidelity* to capture ISC’s subtle tab geometry, and (iii) *diversity* in backgrounds, lighting, and occlusions, while curbing manual annotation cost. Section I-C surveys how synthetic and photorealistic imagery can satisfy those requirements and where current approaches fall short.

C. Synthetic and Photorealistic Data: Benefits and Pitfalls

A practical solution to address the challenges of limited access and varying construction site conditions is the creation of annotated synthetic image datasets to supplement real ones [18]. These synthetic datasets can be generated using computer graphics techniques, 3D modelling software or game engines enabling the simulation of diverse construction environments with different objects and backgrounds.

Computer vision models trained solely on synthetic images often perform worse than those trained on real images. For example, grocery item detection models trained on 400,000 synthetic images performed less effectively than models trained with only 760 real images [20]. Yet, combining just 76 real images with the synthetic images produced superior results compared to both models. Moreover, randomization techniques (such as lighting condition, weather condition, timing of the day, textures, camera perspective etc.) are used for generating synthetic images which reduce the sim2real gap and improve the diversity of the dataset [21] [22]. Therefore, a hybrid dataset that integrates real and synthetic images could be an effective approach for training computer vision models for construction applications [23]. However, obtaining sufficient real images for many construction scenarios or custom objects, such as the ISC, remains difficult. In such cases, computer-aided design tools can generate and render photorealistic models of custom objects in various settings, reducing the reliance on real images.

Additionally, ISC plates pose an additional hurdle: their galvanised coating creates specular highlights that shift with sun angle, and the laser-cut tab patterns differ by millimetres. Capturing these cues demands high-dynamic-range rendering plus fine surface normal maps—costly to generate at scale. Conversely, photographing ISC plates on active sites remains impractical, because the system is not yet widely deployed.

Hence a *hybrid* strategy [(i) auto-generates large volumes of domain-randomised synthetic frames, (ii) injects photorealistic ray-traced scenes for material fidelity, and (iii) enhances the mix with a small, curated set of real photographs] offers the best trade-off between cost and realism.

Research gap. To date, no public dataset combines these three modalities for steel-connection detection; existing construction corpora (MOCS [5], SODA [6]) neither model bespoke joints nor provide labels. Bridging this gap is therefore prerequisite to closing the perception loop for robotic ISC assembly.

D. Research Contribution

Existing vision datasets in construction focus on equipment or personnel safety. None address robotic assembly of structural steel components such as beams, columns, or ISC plates. We fill this void by devising and releasing *ISC-Perception*, a task-specific, hybrid corpus for object detection in robotic steel erection.

In summary, the main contributions of this paper are:

- A methodology for creating a hybrid dataset for ISC components using real, photorealistic, and synthetic images to tackle the scarcity of real images tailored for robotic assembly tasks, reducing human effort from 166h (manual 10,000 images at 60s per image) to 30.5h with our Unity-based pipeline (81.7%); see Table III
- The analysis of training performance of computer vision algorithms for different types of images and validation of the trained computer vision model in small-scale setup.

To contextualise these contributions, the next section II, reviews the current state of the art in real and synthetic computer vision datasets. Subsequently, the section III discusses the procedural approach for generating the hybrid dataset. Section IV provides insight on the ISC-Perception dataset. Section V reviews the outcomes of the training and testing phases, followed by a discussion of the results and findings. Finally, Section VI of the paper summarizes the research results and their significant impacts on the construction industry.

II. LITERATURE REVIEW

This section provides an overview of prominent general purpose computer vision datasets (see II-A) and datasets specific to construction industry (see II-B).

A. Computer Vision Datasets

As this research focuses on generating an image dataset for ISC, this section provides a comprehensive overview of prominent computer vision datasets. Computer vision datasets can be broadly categorized into two main types: real-world datasets and synthetic datasets.

Real-world datasets consist of images captured from actual environments and are crucial for training and evaluating models across a range of tasks, including object recognition, object detection, segmentation, and scene understanding [24]. Numerous widely used datasets have been developed to support these tasks. Prominent examples include ImageNet [25], COCO (Common Objects in Context) [26], Pascal Visual

Object Classes [27], Open Images [28], Cityscapes [29], and KITTI [30]. Table I provides an overview of these key datasets, highlighting their specific features and contributions to the field.

Synthetic dataset generation in computer vision involves creating artificial images and annotations using tools such as rendering engines (e.g., Blender [31], Unity 3D [32], Nvidia Omniverse [33], Unreal Engine [34]), physics-based simulation software (e.g., Gazebo [35], Webots [36], CoppeliaSim [37], and generative AI like GANs. These datasets are particularly valuable for generating large-scale, cost-effective, and safe alternatives to real-world data collection. Examples include synthetic datasets derived from video games like Half-Life 2 [38], the SYNTHIA dataset for semantic segmentation [39], Hattori et al.'s 3D pedestrian models using Autodesk 3DS Max [40], and the Virtual-KITTI [41] and Virtual-KITTI 2 [42] datasets, which replicate urban driving scenes with automated annotations via Unity.

Synthetic datasets can be generated using various 3D CAD model rendering and visualization software, incorporating appropriate lighting and scene generation techniques. For example, Aubry et al. developed a dataset of 86,366 synthesized images by rendering each of 1,393 high-quality 3D chair models from 62 distinct viewpoints [43]. Additionally, Peng et al. explored the influence of pose, colour, textures, and background by training a deep convolutional neural network (CNN) using crowd-sourced 3D CAD models, highlighting the potential of synthetic data in improving model performance [44].

B. Computer Vision Dataset in Construction Industry

Vision technology has garnered significant interest across multiple sectors, including construction, where its application is transforming how visual data from construction sites is acquired and interpreted. This technology enables the extraction of valuable information such as progress monitoring, object detection, safety condition analysis, and quality control. Through the automated detection and tracking of workers, excavators, cranes, dump trucks, and other equipment, it is possible to efficiently identify unsafe conditions on construction sites [45], [46], [47], [48].

SODA [6], tailored for construction sites, contains 19,846 images of 15 object classes. The Moving Objects in Construction Sites (MOCS) dataset contains 41,668 images depicting 13 types of moving objects, including equipment and workers, commonly found on construction sites [5]. Those images were captured using a camera, UAV, and smartphone from 174 different construction sites of dam, bridge, building, tunnel and highway [5]. Del et. al created a small dataset of 1048 images comprising 08 different object classes for detecting construction equipment and human [49]. All these datasets were collected from real construction sites, carefully chosen and edited to remove any privacy information, and manually annotated, which is laborious and time consuming. In contrast, Barrera-Animas and Delgado proposed a method to generate synthetic data sets that closely resembles real-world conditions, using 3D models of construction machinery, workers,

TABLE I
OVERVIEW OF POPULAR DATASETS FOR COMPUTER VISION TASKS

Dataset	Purpose	Year	Classes	Images	Annotations	Domain
ImageNet	Object recognition/classification	2009	21,841	14,197,122	Bounding boxes	General
COCO	Object detection/segmentation	2014	80	328,000	Bboxes, masks	General
Pascal VOC	Object detection/classification	2005	20	11,530	Bounding boxes	General
Open Images	Object detection/classification	2016	19,958	9,011,219	Bounding boxes	General
KITTI	Autonomous driving	2012	9	7,481	3D/2D Bounding Boxes	Urban driving
Cityscapes	Semantic segmentation	2016	30	5,000	Segmentation masks	Urban

site environments and assets, combined with realistic lighting conditions in different seasons [50]. However, all of the mentioned datasets focus primarily on detecting construction equipment and workers to ensure safe operations, with none designed specifically for robotic assembly tasks.

III. METHOD OF GENERATING COMPUTER VISION HYBRID DATASET

This research aims to develop a dataset specifically for the robotic assembly of steel structures using ISC. A hybrid dataset will be created containing photorealistic images of ISC components, synthetic images from the simulation environment, and few real images and trained with the YOLOv8 algorithm for a robotic assembly application.

A. Dataset Composition

The creation of a robust computer vision dataset often begins with the selection of target objects for detection or segmentation. In this work, the dataset focuses on three main object classes: a) ISC member, b) ISC connection plate, and c) human; as illustrated in Fig. 1, the selection of these classes is driven by the requirements of future robotic assembly tasks. We envisage that the robot must accurately identify ISC components for assembly and detect humans to ensure safety compliance. While typical steel construction sites include equipment such as tower cranes, forklifts, and scaffolds, these are excluded from the dataset since the robot will not interact with them.

Given the novelty of ISC, real images of these components are limited in availability. Hence, to address this, the ISC-Perception dataset integrates three types of images from diverse sources:

- 1) Type 1: Photorealistic images from SolidWorks (SW) Visualize (category 1 or C1)
- 2) Type 2: Synthetic images from Unity
 - Built-in randomizers (category 2 or C2)
 - Custom randomizers (category 3 or C3)
- 3) Type 3: Real images
 - From previous project (category 4 or C4)
 - Human images from Roboflow Universe Public Dataset (category 5 or C5)

B. Image Generation

The image generation workflow is shown in Fig.2. Synthetic images in Unity (C2) sometimes suffer from jittering, motion blur, and unrealistic appearances (see Supplementary Fig. S1).

To overcome these limitations, custom randomizers (C3) were employed to generate images with enhanced variability across indoor and outdoor assembly scenes. Similarly, photorealistic images (C1) were generated with 3D CAD models in SolidWorks Visualize, incorporating diverse lighting and backgrounds. Finally, the dataset includes manually annotated real images of ISC components and humans, augmented through preprocessing techniques to bolster diversity. This hybrid composition ensures the dataset is both diverse and generalisable, and provides real-world authenticity to support the development of vision systems capable of detecting ISC components in complex assembly environments.

1) *Photorealistic Images from SolidWorks Visualize*: As previously established, real images of ISC components are scarce, hence we use CAD rendering software enabled by SolidWorks Visualize to generate high-quality photorealistic images to supplement the limited availability of real-world ISC data in ISC-Perception.

The first stage involves the creation of several 3D models of the two main ISC components using CAD software, as shown in Fig. 2. This is then followed by importing the models into SolidWorks Visualize for scene generation. During this stage, SW Visualize provides extensive randomization options to enhance dataset diversity. For randomization, SW Visualize pick from 9 total backgrounds and 3 model textures (1 metallic texture and 2 featuring rust), rotates between 0 to 360°, and varies the lighting conditions, see Table II. The output images from the Scene Generation stage are then annotated in Roboflow to get ground truth bounding boxes of ISC objects. Finally, the photorealistic images are augmented in Roboflow to add more variations to the dataset.

TABLE II
SUMMARY OF RANDOMIZATION OPTIONS USED IN SOLIDWORKS VISUALIZE

Background	Model Texture	Rotation	Lighting
Nine options:	Three textures:	0°–360°	Two options:
• Black background	• Cast carbon steel (red)		• Day
• Empty outdoor parking	• Metal rust 1		• Night
• Swiss snow	• Metal rust 2		
• Steel building site			
• Black/white background			
• Industrial lot (night)			
• Inside glass building			
• Empty indoor garage			
• Boiler room			

2) *Synthetic Images from Unity*: While the use of SW Visualize produces high-quality photorealistic images, the process of manual annotation is time-consuming. Unity with its Per-

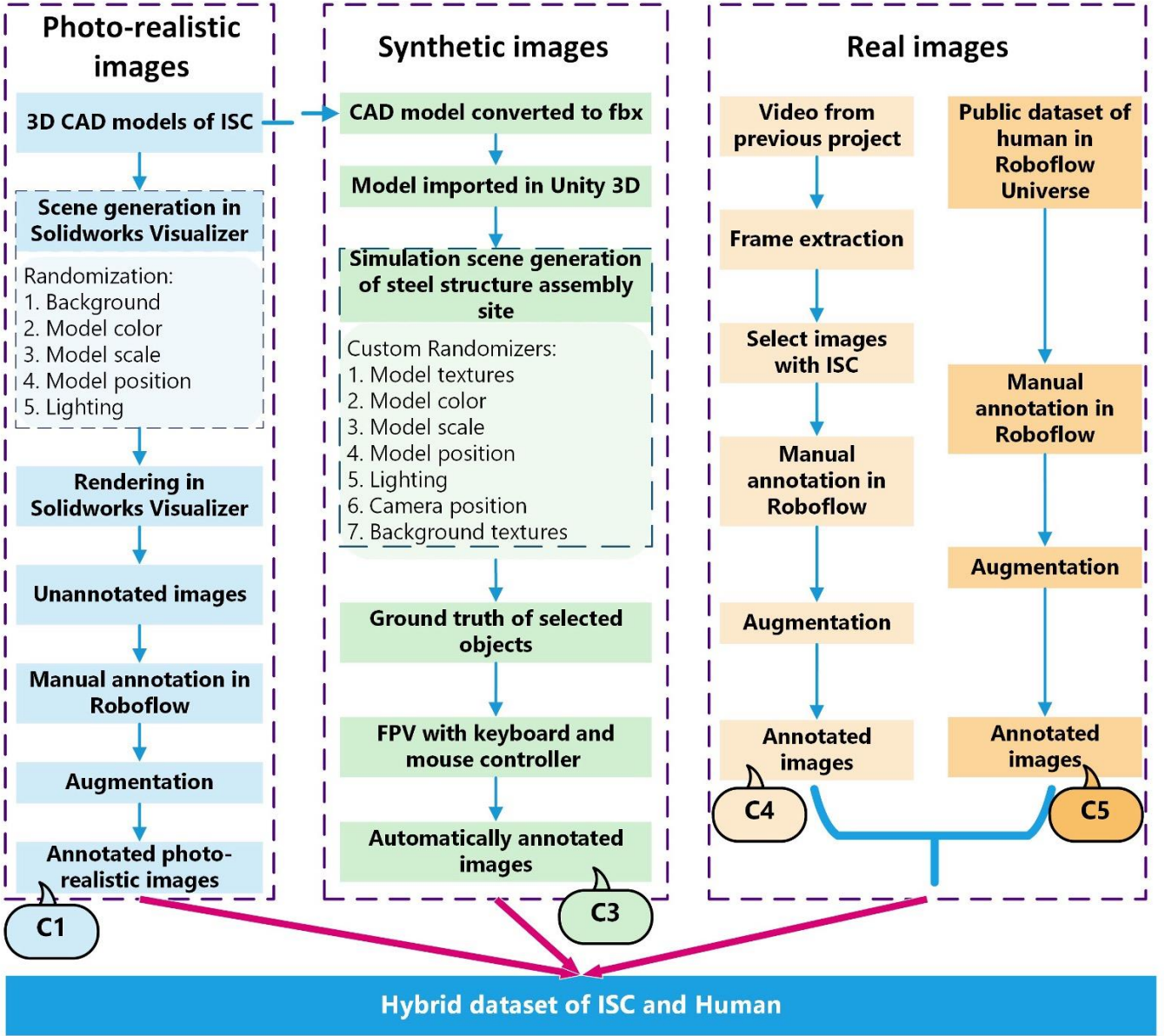


Fig. 2. Source of images and workflow for creating the hybrid dataset combining different types of images.

ception package, offers an efficient alternative for generating large volumes of automatically annotated synthetic images. C3 in Fig.2, shows the synthetic image generation process.

Scene Simulation Generation: To start with, we used the models built from the CAD software during the generation of photorealistic images and imported those to Unity. Two steel structure assembly simulation scenes were created; one indoor and one outdoor (see Fig. 3 for sample views of robotic steel assembly). The indoor scene included a large workspace with walls displaying custom images to simulate construction environments. Fifty random construction site images were used as wall textures, changing every second to increase variation (Fig. 3). The scene was populated with objects such as concrete mixers, dump trucks, scaffolds, and ISC components, placed in various orientations to enhance generalisation. Lighting conditions included directional light mimicking sunlight, dynamically adjusted between -100° and

100° , and multiple indoor light sources for a realistic indoor setup.

The outdoor scene contained environmental elements such as trees and buildings alongside construction equipment, safety barriers, and ISC components placed on pallets or the ground. A directional light simulated sunlight, rotating to create shadows from objects like trees and buildings. ISC components were coloured with solid green, red, and white finishes to further diversify the dataset.

The use of custom randomizers addressed the limitations of Unity's built-in randomizers, which failed to generate realistic ISC environments. ISC components were rotated incrementally by 5° , and background objects were randomly rotated and translated using custom scripts. These randomisations ensured variability in the dataset. Ground truth labels were assigned using Unity's Perception package. Annotated images were recorded with a first-person camera capturing the scene

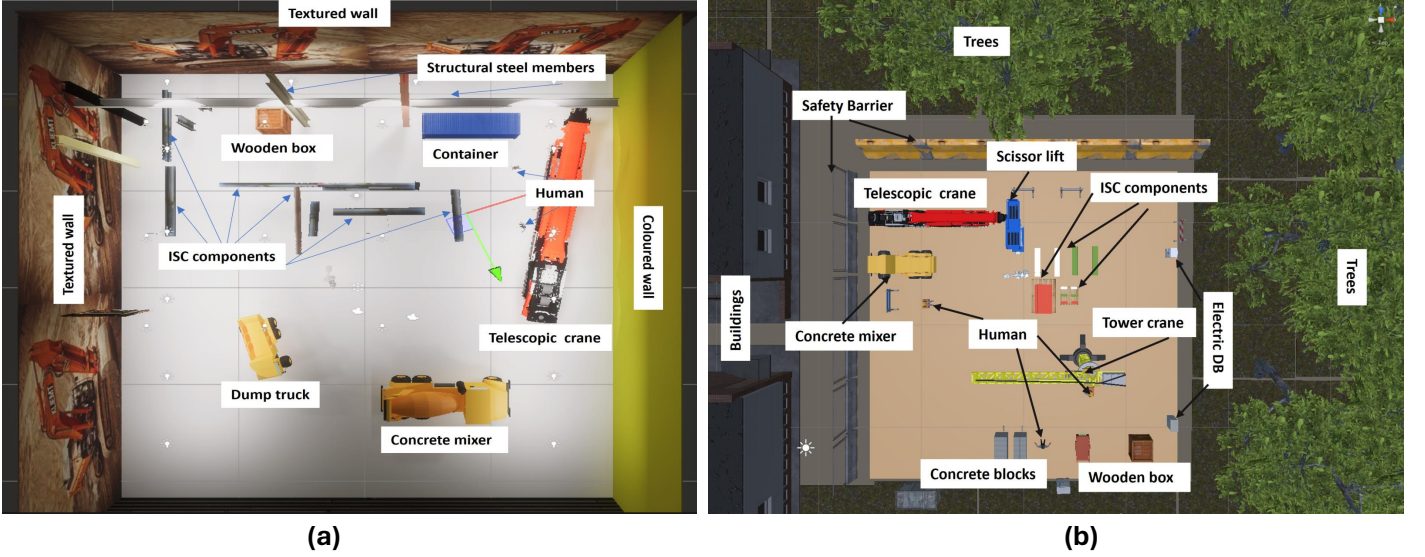


Fig. 3. View of Robotic Steel Assembly in Unity; (a) Outdoor Scene; (b) Indoor Scene

through a keyboard and mouse-controlled player.

This approach efficiently produced a diverse dataset of annotated synthetic images, complementing the photorealistic and real images, while addressing the limitations of Unity’s built-in randomizers in simulating real-world ISC environments.

3) *Real Images*: ISC-Perception includes real images to enhance the dataset’s authenticity and variability. However, with no publicly available ISC dataset, the real images were curated from multiple sources: still frames from ISC assembly videos, manually collected images, and publicly available datasets for humans from Roboflow Universe [51].

To incorporate ISC-related real images, still frames were extracted from previous project videos. Out of 29 extracted images, 16 were extracted for annotation while removing duplicates and closely matched frames. The collected images were then annotated using Roboflow’s annotation tool, followed by preprocessing steps, including brightness adjustments (-15% to $+15\%$) and rotations (-10° to $+10^\circ$). Additional augmentations such as 90° rotations, flips, and saturation adjustments generated 82 final images for the dataset.

Furthermore, small-scale ISC members and connection plates were fabricated for manual image collection with varying appearance, position, and rotation, resulting in a total of 207 annotated images using Label Studio [52].

Additionally, to address safety considerations, images of humans were included. Since numerous publicly available annotated datasets exist for humans, the Roboflow Universe dataset was utilized [51]. This dataset contains 235 images of individuals in various standing and sitting poses, with preprocessing effects such as colour, brightness, shear, and stretch. All human images were manually scrutinised to address privacy concerns before inclusion in the dataset.

IV. DATASET

ISC-Perception dataset integrates images from four primary sources as described in the section III-A. These sources

include; collected 13,399 images via **Unity with Custom randomizers (C3)**, 3,599 images via **SW Visualize (C1)**, 289 images of **Real Images from previous ISC project (C4)**, and 1,728 of **Human Images from Roboflow Universe (C5)**. The total number of images in the dataset is distributed across training, validation, and testing sets, ensuring a fair diversity and representation of different scenarios. This includes 15,974 images of training and validation images in dataset 3 and a test set comprising 3,087 images (16% of the total). Test images were manually selected to capture varied scenarios, ensuring robust evaluation. See Tables IV and V for the summary and distribution of the dataset.

A. Time-to-Dataset Accounting (Synthetic vs Manual)

Table III details the *human* effort for our Unity-based pipeline (30.5 h total; effective 11.0 s/image) and lists compute wall-clock separately (12 h). At our full dataset size ($N=15,974$), human time¹ is 30.5 h versus 266.2 h for manual labelling at 60 s/image (-88.5%). Compute wall-clock (render/export) scales linearly with N and is ≈ 19.2 h; we exclude this from human-time totals. All stages except quality assurance (QA) sampling are fixed one-off tasks; only the QA term scales with N (2% at 6 s/image). Hence 30.5 h at $N=10,000$ versus 30.7 h at $N=15,974$.

B. Statistics of the Datasets

To evaluate the model’s performance, three versions of ISC-Perception datasets were created, with a constant test set across all four versions. Table V provides a detailed distribution of images across datasets, while Fig. 4 illustrates key statistics.

- 1) *Image Distribution*: Dataset 3 (Hybrid Dataset) contains the largest number of images (15,974) from all sources,

¹Human time is dominated by fixed setup; the only N -dependent component is QA (2% at 6 s/image), which accounts for the 0.2 h increase from 10k to 15,974 images. Compute wall-clock (render/export) is reported separately and not counted as labour.

TABLE III

HUMAN TIME ACCOUNTING AT $N=10,000$ IMAGES. *Compute wall-clock* (GPU/CPU RENDERING/EXPORT) IS LISTED IN A SEPARATE COLUMN AND *not counted* AS HUMAN LABOUR. AT OUR FULL DATASET SIZE ($N=15,974$): MANUAL = 266.2 H, SYNTHETIC HUMAN = 30.7 H, COMPUTE $\approx$ 19.2 H.

Stage	Human time	Compute time	Notes
Unity setup & assets	6.0 h	–	project, import, materials
Scene/physics authoring	6.0 h	–	colliders, dynamics
Sensor/exporters	3.5 h	–	RGB, depth, masks, GT
Domain randomisation	3.5 h	–	poses, lights, textures
SolidWorks Visualize integration	3.0 h	–	mesh overlays, QA hooks
Render/export automation	2.0 h	12 h	batch scripts; GPU/CPU wall-clock
QA sampling (2% @ 6 s/image)	0.33 h	–	visual checks only
Final end-to-end checks	6.2 h	–	splits, metadata, hashes
Manual baseline (60 s/image)	166.7 h	–	single annotator; 5-instance pilot 60 s(max 80 s)
Synthetic total (human)	30.5 h	12 h	effective 11.0 s/imagehuman time

Note. Totals in the table correspond to $N=10,000$ (compute = 12 h).

At our full dataset size $N=15,974$, compute is $\approx$ 19.2 h; we report this separately in the text.

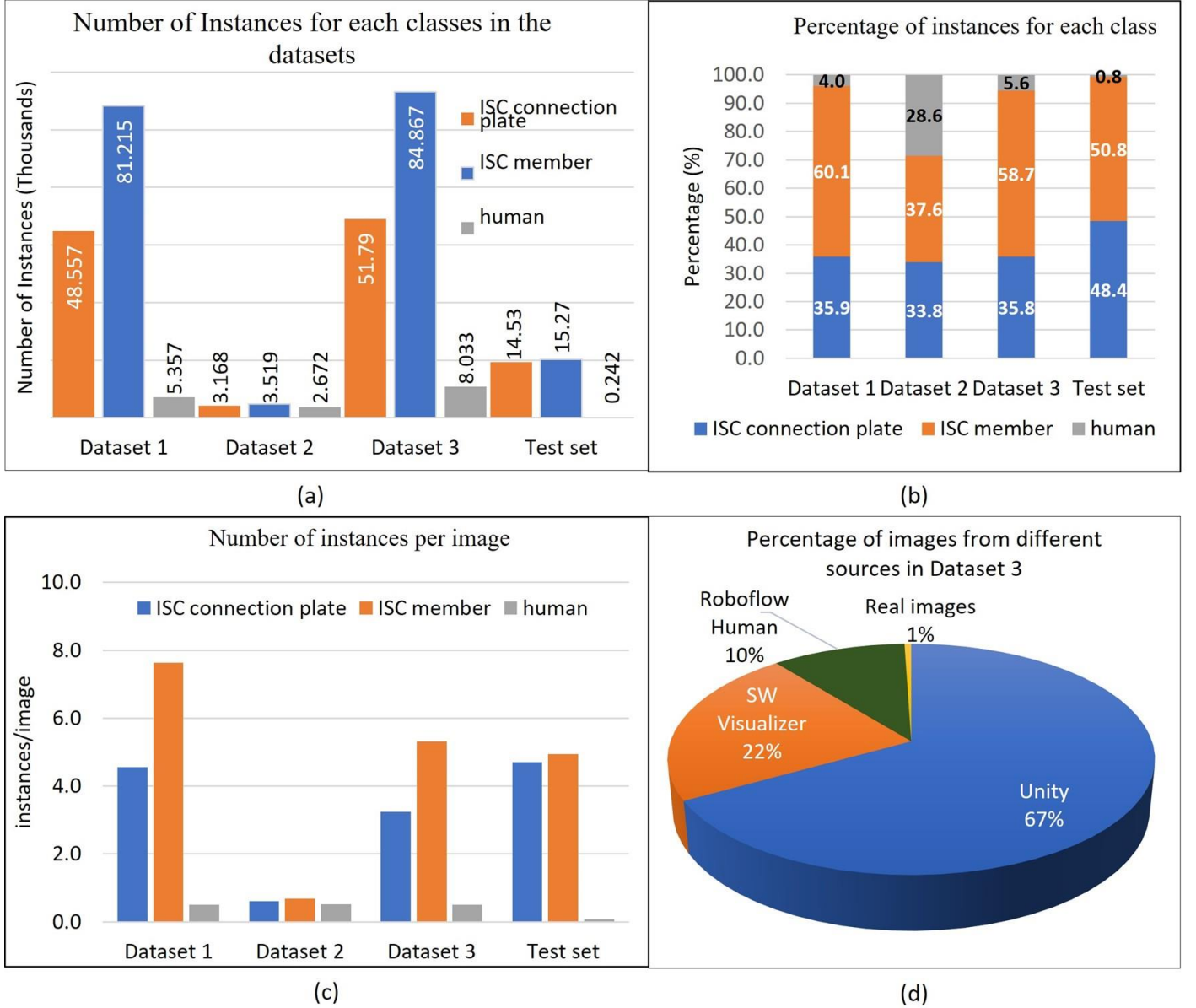


Fig. 4. Dataset statistics; (a) number of instances for each class and (b) percentage of instances in each dataset, (c) number of instances per image for each class, (d) percentage of images from different source in dataset 3.

while Dataset 2 (SW Visualize with Roboflow Human)

has fewer images (5, 195). Dataset 1 (Unity Custom ran-

TABLE IV
SUMMARY OF THE DATASETS FOR DETECTING ISC COMPONENTS

Dataset	Source of Images	Images
Dataset 1 (Custom randomizer dataset)	Unity (custom randomizer): indoor/outdoor scenes (C3)	10,651
Dataset 2 (SW visualize and Roboflow dataset)	SW Visualize + Roboflow human annotations (C1+C5)	5,195
Dataset 3 (Hybrid dataset)	Unity + SW Visualize + Roboflow + real images (C1+C3+C4+C5)	15,974

TABLE V
DISTRIBUTION OF IMAGES IN ISC-PERCEPTION ACROSS SOURCES

Source	Dataset 1	Dataset 2	Dataset 3	Test Set
Unity (Custom randomizer)	10,651	0	10,651	2,748
SW Visualize	0	3,551	3,551	48
Roboflow Human	0	1,644	1,644	84
Real Images	0	0	82	207
Total	10,651	5,195	15,974	3,087

domizer) focuses solely on synthetic data, with 10,651 images.

- 2) *Instances per Class*: Fig. 4(a) shows that ISC members dominate with 15,270 instances in the test set, followed by ISC connection plates (14,530) and humans (242). Dataset 1 has the highest average number of instances per image for each class, as shown in Fig. 4(c).
- 3) *Percentage of Sources*: In Dataset 3, 67% of images come from Unity, 22% from SolidWorks Visualize, 10% from Roboflow Universe, and 1% from real ISC images (Fig. 4(d)). ISC is a newly developed novel connection that has not yet been commercially adopted in construction, so it is difficult to get real images of ISC. Hence, only 82 real images of ISC could be collected and included in the dataset, as shown in Table V.

C. Example Images

Figure 5 showcases sample images from the dataset:

- 1) *Unity Synthetic Images*: Fig. 5(a) and Fig. 5(b) highlight images generated using built-in and custom randomizers, with varying lighting, object placements, and occlusion effects.
- 2) *Photorealistic Images*: Fig. 5(c) illustrates high-quality images from SolidWorks Visualize, featuring diverse scenes and objects, including grayscale and construction site settings.
- 3) *Real Images*: Fig. 5(d) shows human images from Roboflow Universe, while Fig. 5(e) presents real images from previous ISC assembly projects. Roboflow Universe aggregates community-contributed images from multiple providers (which may include stock-photography sources); we therefore cite the dataset as Roboflow Universe for panel (d).

V. PERFORMANCE ANALYSIS

Performance analysis of the datasets was divided into two categories. Initially, three different computer vision models

were generated by training the YOLOv8 algorithm with three different datasets (dataset 1, dataset 2 and dataset 3) from Table V. The training performance was initially analysed using several performance metrics. Finally, the trained model was applied to the test set from Table V to predict the desired object. The effect of different image types and datasets were analysed based on the prediction performance. We first report full-size training results, then a controlled size-matched comparison to disentangle composition from dataset size (Sec. V-D2).

A. Hardware configuration

To create the object detection model, the YOLOv8 algorithm was trained with all three versions of the dataset. An Alienware m16 laptop configured with core i9 13900HX processor, 32.0GB RAM and Nvidia GeForce RTX 4060 12GB GDDR6 graphics card was used to train the ISC components detection model. Each dataset was trained for a maximum 250 epochs, with 30 epochs patience as a stopping criterion. Other training parameters remained at default. The best model was saved for each version of the dataset.

B. Training settings

We trained YOLOv8n (Ultralytics v8.3.198) starting from pretrained model on COCO [53] using the Ultralytics trainer. We used an input size 640×640 , batch size 16, and random seed 42. We enabled a cosine learning-rate schedule and do not override the framework's base learning rate; other optimizer hyperparameters remain at defaults. For the main experiments, we used early stopping with a cap of 250 epochs and patience 30, selecting the best checkpoint by validation mAP. For the controlled size-matched comparison (Section. V-D2), we fixed the number of optimizer updates (no early stopping) and match augmentations, input size, batch size, and learning rate schedule across conditions.

C. Testing procedure

The best models trained were obtained from a collective of three distinct dataset configurations: the hybrid dataset (which includes custom randomizer, SW Visualize, Roboflow, and manually annotated images), a dataset using only the custom randomizer images, and a dataset using only the SW Visualize images. Tests were then conducted using two groups of test data:

- 1) *Complete test set*: This contains samples from all sources – custom randomizer, SW Visualize, Roboflow Human, and manually annotated images. This contains a total of 3087 images. Details of the number of samples from each source are presented in Table V.
- 2) *Small scale bench test*: This set comprises real-world samples using a multi-view, small scale experimental setup of robotic assembly of ISC, where synchronised cameras provided multiple perspectives of the same scene. This setup allows us to do continuous object detection and tracking of ISC components and humans.

To disentangle composition from dataset size, we also evaluated models trained on size- and class-matched subsets; see Section V-D2.

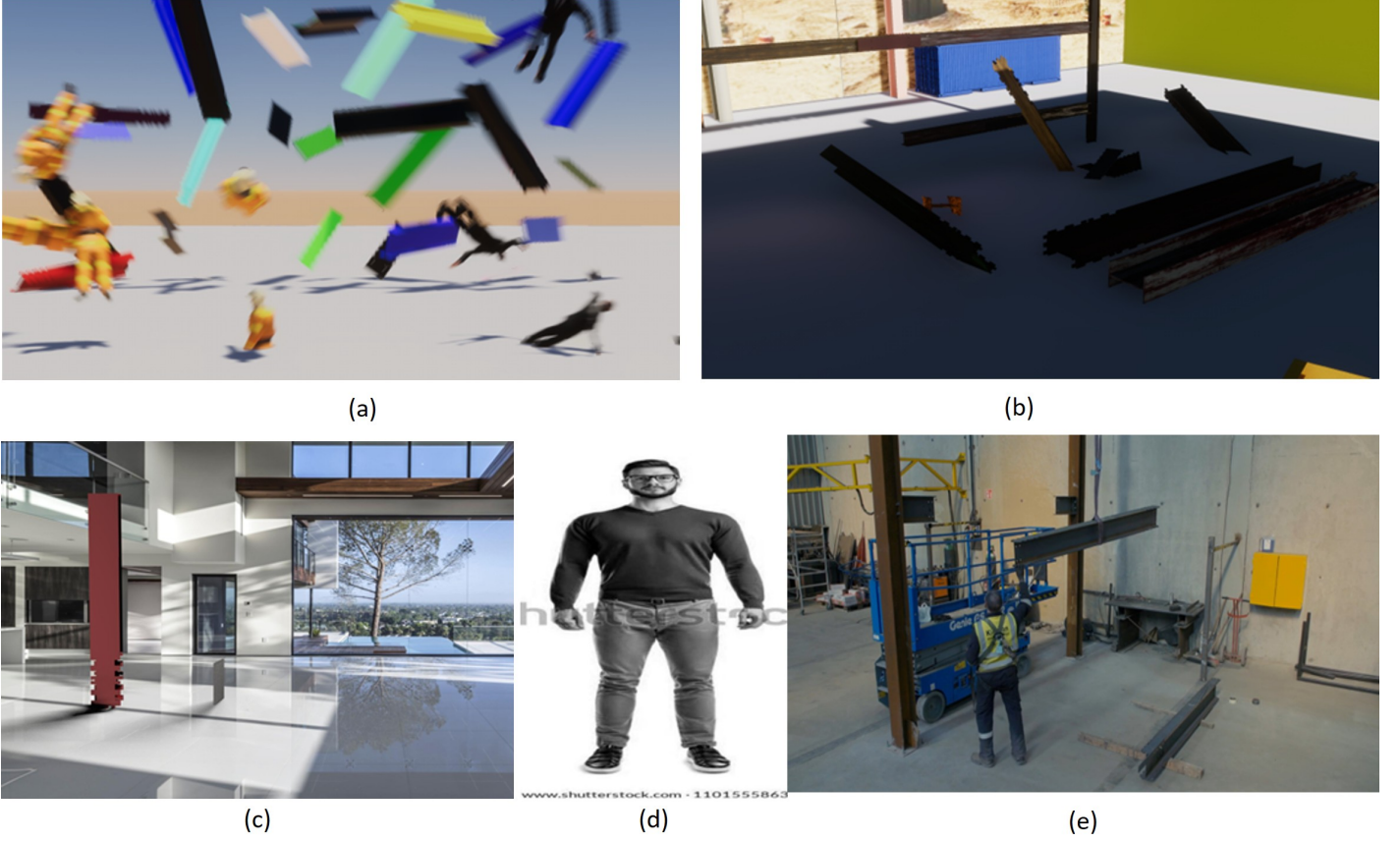


Fig. 5. Representative samples from ISC-Perception: (a) Unity (built-in randomizers, C2), (b) Unity (custom randomizers, C3), (c) SolidWorks Visualize photorealistic render (C1), (d) Human example from Roboflow Universe (C5), (e) Real ISC frame (C4). Roboflow Universe aggregates contributions from multiple providers (which can include stock libraries); we therefore cite Roboflow Universe as the source for (d).

D. Results and Discussion: Testing with test set

1) *Results Overview:* The three models trained on Hybrid dataset, Custom randomizer dataset, and SW Visualize and Roboflow dataset – were all tested on a combined test set (Table V) containing 3,087 images- includes samples from all data sources; Unity custom randomizer, SW Visualize, Roboflow human images, and manually annotated real-world images. Four performance metrics (Precision, Recall, mAP@0.5 and mAP@[0.5:0.95]) were used to assess the trained performance of the model across the test data. We present the results of training in Table VI.

The model trained on the Hybrid dataset achieved the highest overall performance with a mAP (50-95) of 0.664 compared to 0.564 for custom randomizer dataset and 0.321 for SW Visualize and Roboflow dataset indicating that it is able to generalise across a diverse range of image types (see Table VI). The performance was particularly strong for human detection, where it achieved a mAP@[0.5:0.95] of 0.804 and high precision and recall scores. On the other hand, the model’s performance for ISC connection plates was lower with a mAP@[0.5:0.95] of 0.523 likely due to the complexity of detecting these components in varied real-world environments (Table VI).

2) *Controlled Size-Matched Comparison:* We trained the model with identical hyperparameters on size- and class-

matched subsets ($N=5,195$) of *Dataset-1*, *Dataset-2*, and the *Dataset-3*. Subsets are constructed by stratified sampling to preserve class priors (and, where available, instances-per-image bins); the test set is fixed (3,087 images). We equalised the training budget by using a fixed number of optimiser updates (no early stopping), and we match augmentations, input resolution, batch size, and learning-rate schedule across conditions as in the main training experiment. At fixed N , the hybrid composition achieves mAP@0.50 of 0.675 and mAP@[0.50:0.95] 0.549, exceeding Dataset-1 (0.546/0.430) and Dataset-2 (0.249/0.206), with higher precision (0.830 vs 0.776/0.649) and recall (0.625 vs 0.505/0.146). This corresponds to +0.129 mAP@0.50 (+23.6%) and +0.119 mAP@[.50:.95] (+27.7%) over Dataset-1, and +0.426 / +0.343 (+171% / +166%) over Dataset-2. These gains at constant size indicate the improvement stems from the *hybrid composition*, not merely dataset size. See Table VII.

3) *Confusion Matrix and Performance Curves:* As shown in the confusion matrix in Fig. 6 the model trained on hybrid dataset correctly identified a large proportion of ISC components and human instances compared to model trained on custom randomizer dataset and SW Visualize and Roboflow human dataset. However, all trained model exhibited some misclassification. For example, the trained model successfully detected 7530 connection plates while there are only 840 false

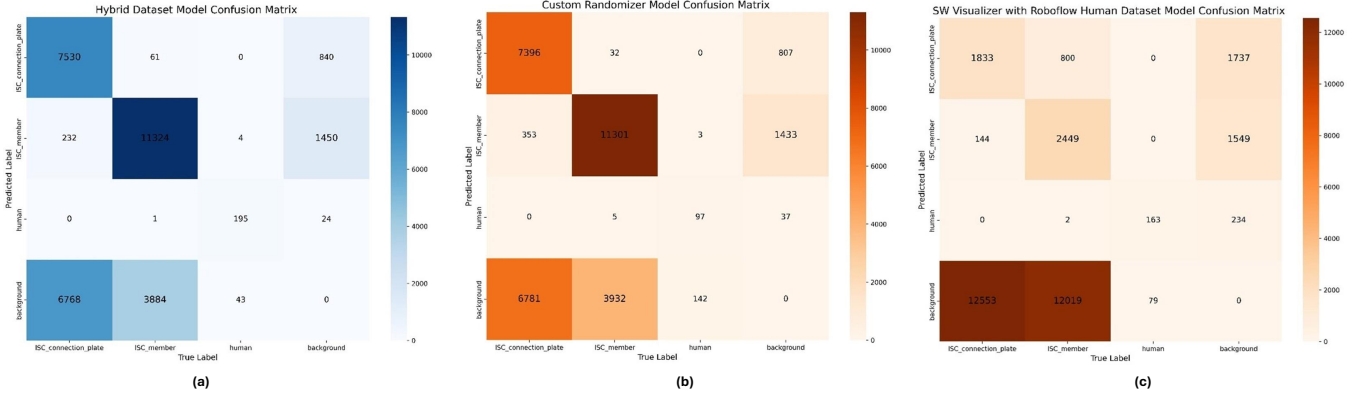


Fig. 6. Confusion Matrix plots of trained models on test Set; (a) on Hybrid Dataset (b) Custom Randomizer Dataset (c) SW Visualize with Roboflow Dataset

TABLE VI
PERFORMANCE OF TRAINED MODELS ON THE TEST SET

	Dataset	precision	recall	mAP@0.5	mAP@[0.5:0.95]
Overall	1	0.82	0.52	0.66	0.56
	2	0.40	0.32	0.39	0.32
	3	0.85	0.67	0.75	0.66
ISC connection plate	1	0.83	0.47	0.64	0.53
	2	0.36	0.21	0.21	0.18
	3	0.81	0.49	0.64	0.52
ISC member	1	0.80	0.71	0.75	0.67
	2	0.52	0.16	0.33	0.24
	3	0.80	0.73	0.76	0.66
Human	1	0.82	0.38	0.59	0.50
	2	0.33	0.67	0.61	0.54
	3	0.92	0.78	0.87	0.80

TABLE VII
SIZE-MATCHED COMPARISON. DATASET-2 USES ITS FULL TRAIN SET ($N=5,195$); DATASET-1 AND DATASET-3 ARE STRATIFIED-SAMPLED TO $N=5,195$. TEST SET FIXED (3,087). * SAMPLED; † FULL.

Train set (N)	mAP50	mAP50-95	Prec.	Rec.
Dataset-1* (5,195)	0.546	0.430	0.776	0.505
Dataset-2† (5,195)	0.249	0.206	0.649	0.146
Dataset-3* (hybrid, 5,195)	0.675	0.549	0.830	0.625

negatives for connection plates and the background and only 61 false negatives for connection plates and the ISC members (Fig.6(a)). Model trained on the hybrid dataset also correctly detected 11,324 instances of ISC members and 195 instances of human (Fig.6(a)). However, as per Fig.6(c), model trained on the SW Visualize and Roboflow human performed worst by correctly detecting only 1,883 instances of ISC connection plate, 2,449 instances of ISC member and 163 instances of human on the test set. As per the Fig.7(a), F1-Confidence Curve showing an overall F1 score of 0.74 at a confidence threshold of 0.377. Precision remained high for all classes, as demonstrated in the Precision-Confidence Curve Fig.7(c), although ISC connection plate detection showed a noticeable drop-off in recall confirming that the model sometimes missed these components in challenging scenes.

4) *Discussion on Model Performance Comparison: Hybrid dataset vs. Custom randomizer dataset vs. SW Visualize and Roboflow dataset:* The evaluation results of the model trained on the three distinct datasets are presented in Table VI, which summarise their performance metrics for each class.

Hybrid dataset model: The model trained on the Hybrid dataset (Table VI) demonstrated the highest overall performance across all object classes. It achieved a Box Precision of 0.846 and a Recall of 0.666, leading to an overall mAP@0.5 of 0.756 and a mAP@[0.5:0.95] of 0.664. The performance in detecting ISC connection plates was slightly lower, with a mAP@[0.5:0.95] of 0.523, indicating some difficulty in precise identification. However, the model excelled in ISC member detection, attaining a mAP@[0.5:0.95] of 0.664, and performed exceptionally well in detecting humans, with a mAP@[0.5:0.95] of 0.804 and a Recall of 0.777.

Custom randomizer dataset model: The Custom randomizer dataset model (Table VI) displayed a similar trend, but with a lower overall performance compared to the Hybrid dataset model. The Box Precision of 0.818 and Recall of 0.521 resulted in an overall mAP@0.5 of 0.659 and mAP@[0.5:0.95] of 0.564. For ISC connection plates, the performance was comparable to the Hybrid dataset model with a mAP@[0.5:0.95] of 0.523. However, the model exhibited reduced accuracy in detecting humans, with a mAP@[0.5:0.95] of 0.503, indicating limitations in handling more diverse human instances.

SW Visualize and Roboflow dataset model: The model trained on the SW Visualize and Roboflow dataset (Table VI) performed the weakest overall, reflecting its narrow focus on the data set. It achieved a much lower Box Precision of 0.404 and Recall of 0.321, resulting in an overall mAP@0.5 of 0.386 and mAP@[0.5:0.95] of 0.321. For ISC connection plates, the mAP@[0.5:0.95] was the lowest at 0.176, and ISC member detection also lagged, with a mAP@[0.5:0.95] of 0.244. While this model was relatively better at human detection with a mAP@[0.5:0.95] of 0.541, its overall ability to generalise to ISC components was clearly limited. From these results, it is evident that the Hybrid Dataset model provides the best performance across all object categories, particularly excelling in human detection and ISC member identification. The Custom Randomizer model, while decent, struggles with human detection and generalisation to real-world data. Lastly, the SW Visualize and Roboflow Human model shows clear limitations, particularly for ISC components, due to its narrow training focus.

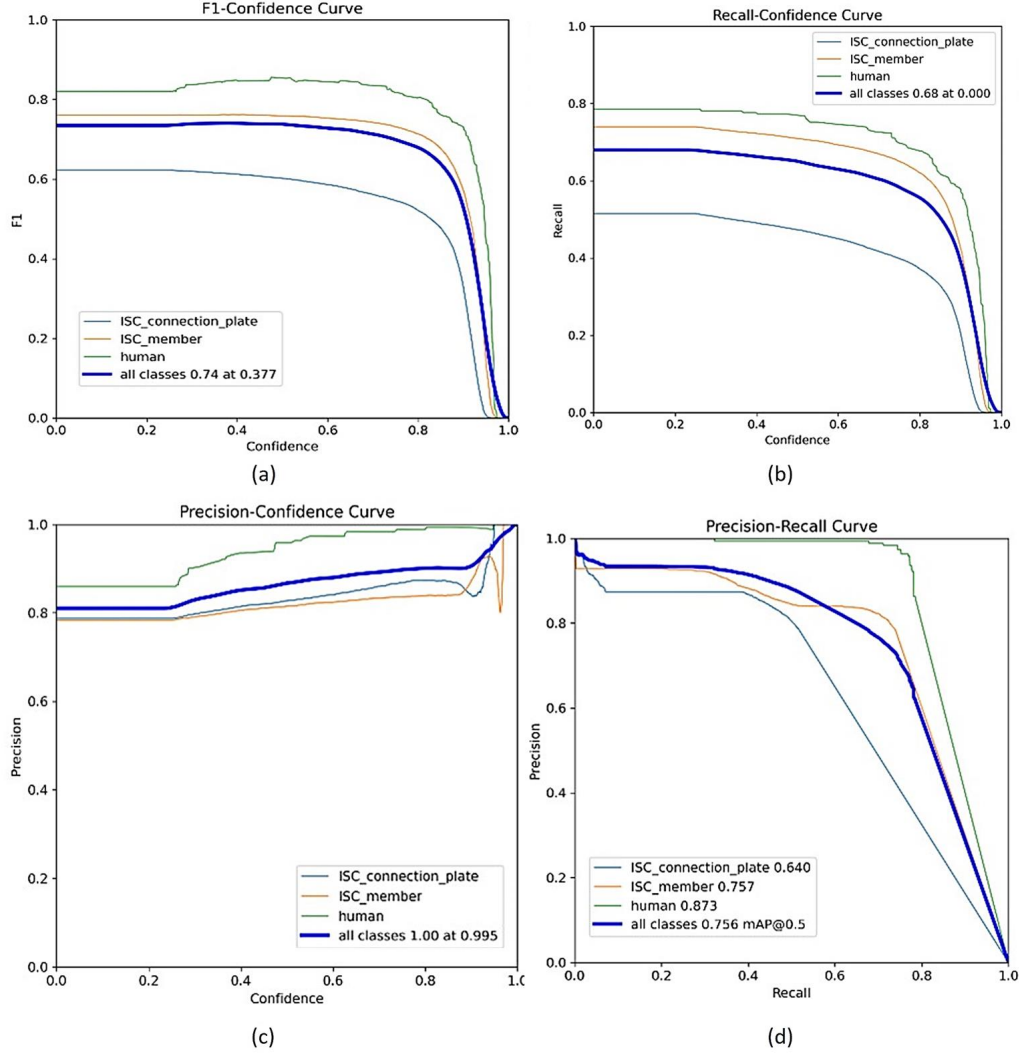


Fig. 7. Performance curves of the model trained on Hybrid Dataset Evaluated on Test Set; (a) F1 confidence curve, (b) recall-confidence curve, (c) precision-confidence curve, (d) precision-recall curve.

E. Small scale bench testing

The overall objective of our research is to develop and integrate the object detection model into a broader robotic assembly process. To assess the robustness and performance of the model in real-world conditions, we incorporated its output into the vision module of our assembly framework. The object detection model serves as an intermediary for subsequent vision-based tasks within this framework. Our setup, as shown in Fig. 8 consists of a multi-camera system, where synchronised views provide multiple perspectives to minimise occlusion and enhance overall detection robustness. From a 2 min, 60 fps bench-test video with two synchronised side views (Fig. 8), we temporally subsampled every 10th frame to reduce correlation and manually annotated the resulting $\sim 1,200$ frames for ISC components and humans. Using standard detection metrics, the detector achieved $\text{mAP}@0.50 = 0.943$, $\text{mAP}@[.50:.95] = 0.823$, precision = 0.951, and recall = 0.930. (See Fig. 9 for detailed detection result samples). However, the system was not without its challenges. Failure cases were primarily due to glare in the front-facing

camera view (Fig. 10), leading to occasional missed detections of connection plates; improving robustness to challenging lighting and appearance shifts is left for future work.

VI. CONCLUSIONS

This research demonstrated a procedural approach to generate a hybrid dataset for detecting ISC components in a steel-structure assembly site. As the collection of real images from the steel-structure assembly site is an arduous and unsafe task, synthetic and photorealistic images were created to compensate for the need for real images. Synthetic images from Unity 3D provide annotated images, while photorealistic images from SW Visualize require manual annotation. Multiple datasets were created using various types of images. Dataset 1 was created using only synthetic images generated with Unity’s custom randomizers. Dataset 2 was created using photorealistic images from SW Visualize and real human images from the Roboflow Universe public dataset. Dataset 3 was created from Unity’s custom randomizer, SW Visualize and Roboflow, and real-world images. Only 3 object classes

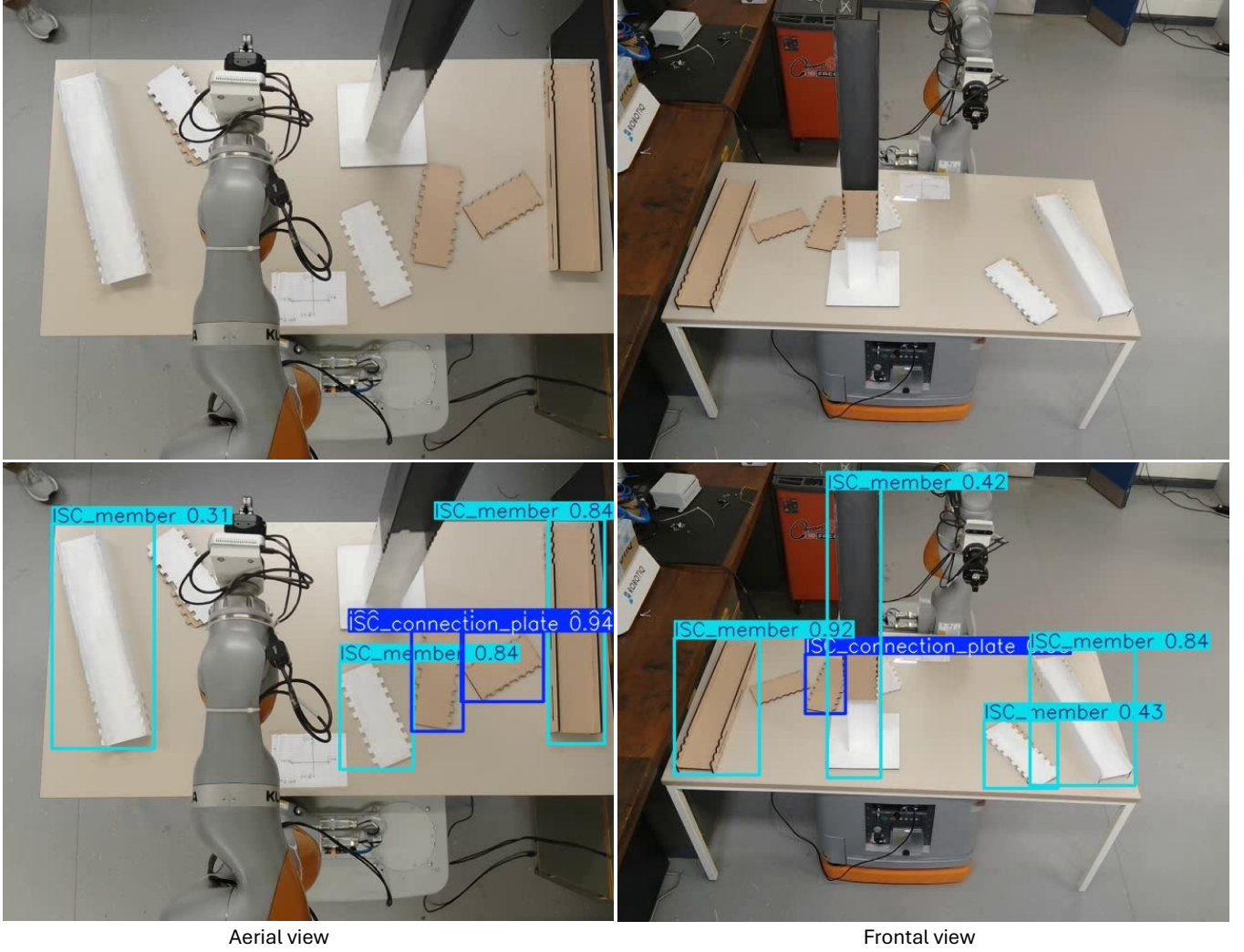


Fig. 8. Synchronized aerial- and frontal-camera views of the bench-top ISC assembly experiment, shown before (top row) and after (bottom row) YOLOv8 inference.

(ISC connection plate, ISC member, and human) were selected for object detection tasks. These were selected because the robot will only manipulate the ISC connection plate and ISC member, while the human was added to ensure the safety.

Testing results showed that models trained on the hybrid dataset (Dataset 3) outperformed those trained on either synthetic (Dataset 1) or photorealistic data alone (Dataset 2). The hybrid dataset model demonstrated superior precision and recall across all object classes (ISC connection plate, ISC member, and human) when tested on the complete test set. The custom randomizer (Dataset 1) model achieved reasonable performance in testing but still lagged behind the hybrid model. The model trained on SW Visualize and Roboflow Human images (Dataset 2) had the lowest performance, especially in detecting ISC connection plates and ISC members, highlighting the difficulty of generalising from such a limited dataset.

To improve the impact of the synthetic images in the hybrid dataset, better quality and realistic simulation scenarios will be created in the future work. With proper computer aided design tools and graphics software, accurate colour, shape

and scale will be created for the objects in the steel structure assembly site scene in Unity. Moreover, the manual annotation of the photorealistic images will be converted to an automatic annotation process to save time and create more variations. More natural lighting and other environmental effects will be introduced for both Unity and the SW visualize scene to enhance the photorealism. While this process of generating a hybrid dataset is not completely automatic, this method provides a procedure to generate a hybrid dataset where the collection of real images is very difficult, and the 3D model of the target object is readily available. Overall, the results of this work reinforce the importance of using datasets for robust object detection in real-world industrial settings and provide a foundation for future research in automating complex tasks in the construction and assembly industries. This method provides a scalable approach that can further be adapted to other robotic applications, particularly in challenging environments where human access is seen as restricted such as nuclear sites, tunnels, or remote workspaces. In addition, it has been robustly shown that hybrid datasets can provide a rich way to train computer vision models especially for emerging applications

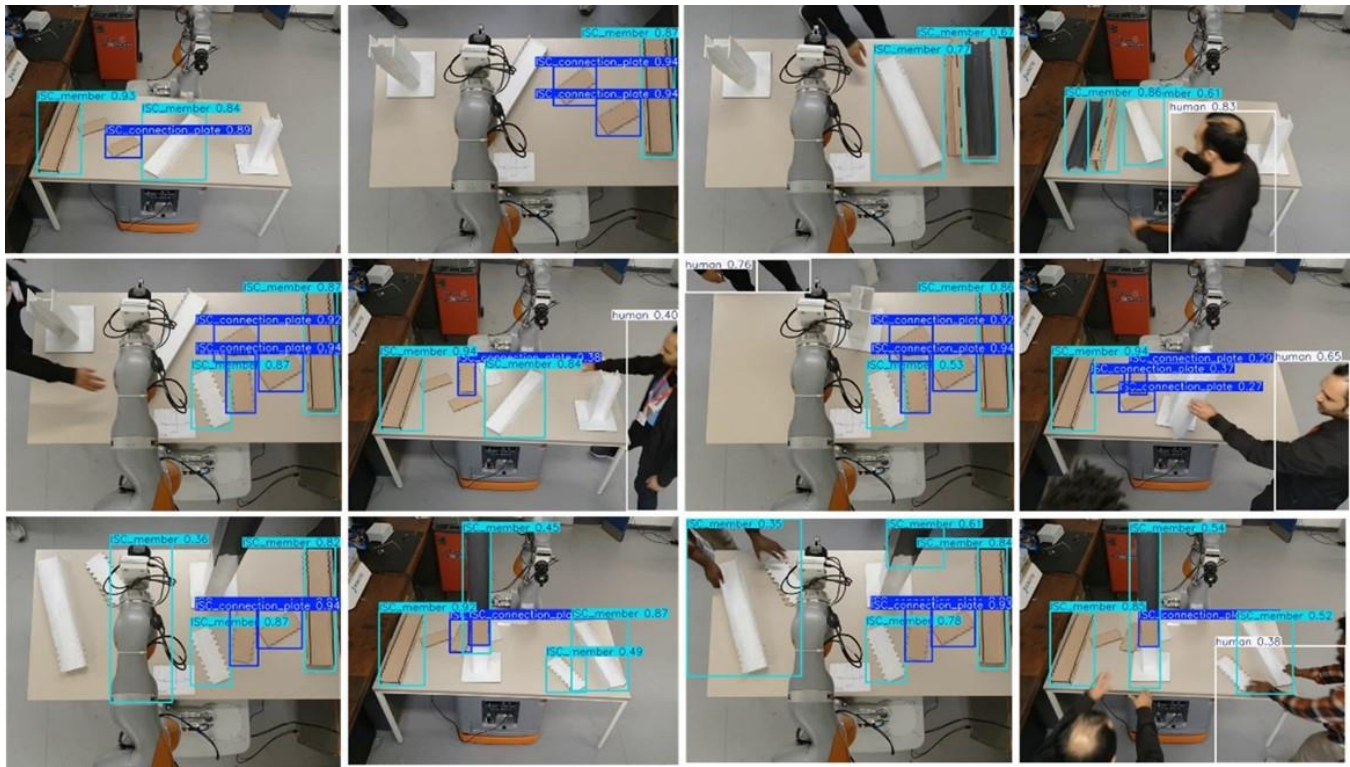


Fig. 9. Real-Time Object Detection Tracking Performance on ISC Objects, Connection Plates, and Human Workers.

were limited relevant public datasets are available.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by the National Science Foundation (NSF) award 2222815, the Science Foundation Ireland award 21/US/3797, and the Department for the Economy (DfE), UK, through grant USI-218.

REFERENCES

- [1] C.-J. Liang, S.-C. Kang, and M.-H. Lee, "RAS: a robotic assembly system for steel structure erection and assembly," *International Journal of Intelligent Robotics and Applications*, vol. 1, no. 4, pp. 459–476, 2017. [Online]. Available: <https://doi.org/10.1007/s41315-017-0030-x>
- [2] M. E. Shemshadian, R. Labbane, A. E. Schultz, J.-L. Le, D. F. Laefer, S. Al-Sabah, and P. McGetrick, "Experimental study of intermeshed steel connections manufactured using advanced cutting techniques," *Journal of Constructional Steel Research*, vol. 172, p. 106169, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0143974X19313616>
- [3] S. Al-Sabah, D. F. Laefer, L. Truong Hong, M. Phuoc Huynh, J.-L. Le, T. Martin, P. Matis, P. McGetrick, A. Schultz, M. E. Shemshadian, and R. Dizon, "Introduction of the intermeshed steel connection—a new universal steel connection," *Buildings*, vol. 10, no. 3, 2020. [Online]. Available: <https://www.mdpi.com/2075-5309/10/3/37>
- [4] S. Montazeri, Z. Lei, and N. Odo, "Design for manufacturing and assembly (DfMA) in construction: A holistic review of current trends and future directions," *Buildings*, vol. 14, no. 1, 2024. [Online]. Available: <https://www.mdpi.com/2075-5309/14/1/285>
- [5] A. Xuehui, Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, and L. Zhiwei, "Dataset and benchmark for detecting moving objects in construction sites," *Automation in Construction*, vol. 122, p. 103482, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0926580520310621>
- [6] R. Duan, H. Deng, M. Tian, Y. Deng, and J. Lin, "Soda: A large-scale open site object detection dataset for deep learning in construction," *Automation in Construction*, vol. 142, p. 104499, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0926580522003727>
- [7] Precedence Research, "Structural steel market revenue to attain usd 177.97 bn by 2033," <https://www.precedenceresearch.com/insights/structural-steel-market>, 2025, accessed: 2025-05-29.
- [8] C. J. Carter, *Practical Information for Designers: Economy in Steel*. American Society of Civil Engineers, 2012, pp. 1–8. [Online]. Available: <https://ascelibrary.org/doi/abs/10.1061/40700%282004%29174>
- [9] R. Chambers, "Connecting the costs of a steel frame," <https://www.ellandsteel.com/connecting-the-costs-of-a-steel-frame/>, 2022, accessed: 2025-05-29.
- [10] M. Rahman, S. Adebayo, D. Hester, D. McPolin, K. Raferty, I. Awolusi, and D. F. Laefer, "A proposed strategy for automating intermeshed steel connection assembly using robotics," in *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, vol. 42. IAARC Publications, 2025, pp. 477–484.
- [11] Z. Jiang and J. I. Messner, "Computer vision applications in construction and asset management phases: A literature review," *Journal of Information Technology in Construction (ITcon)*, vol. 28, no. 9, pp. 176–199, 2023. [Online]. Available: <http://www.itcon.org/paper/2023/9>
- [12] Y. Li and Y. Zhang, "Application research of computer vision technology in automation," in *2020 International Conference on Computer Information and Big Data Applications (CIBDA)*, 2020, pp. 374–377.
- [13] M. Nain, S. Sharma, and S. Chaurasia, "Safety and compliance management system using computer vision and deep learning," *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, p. 012013, mar 2021. [Online]. Available: <https://dx.doi.org/10.1088/1757-899X/1099/1/012013>
- [14] B. H. Guo, Y. Zou, Y. Fang, Y. M. Goh, and P. X. Zou, "Computer vision technologies for safety science and management in construction: A critical review and future research directions," *Safety Science*, vol. 135, p. 105130, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925757320305270>
- [15] J. Seo, S. Han, S. Lee, and H. Kim, "Computer vision techniques for construction safety and health monitoring," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 239–251, 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1474034615000269>

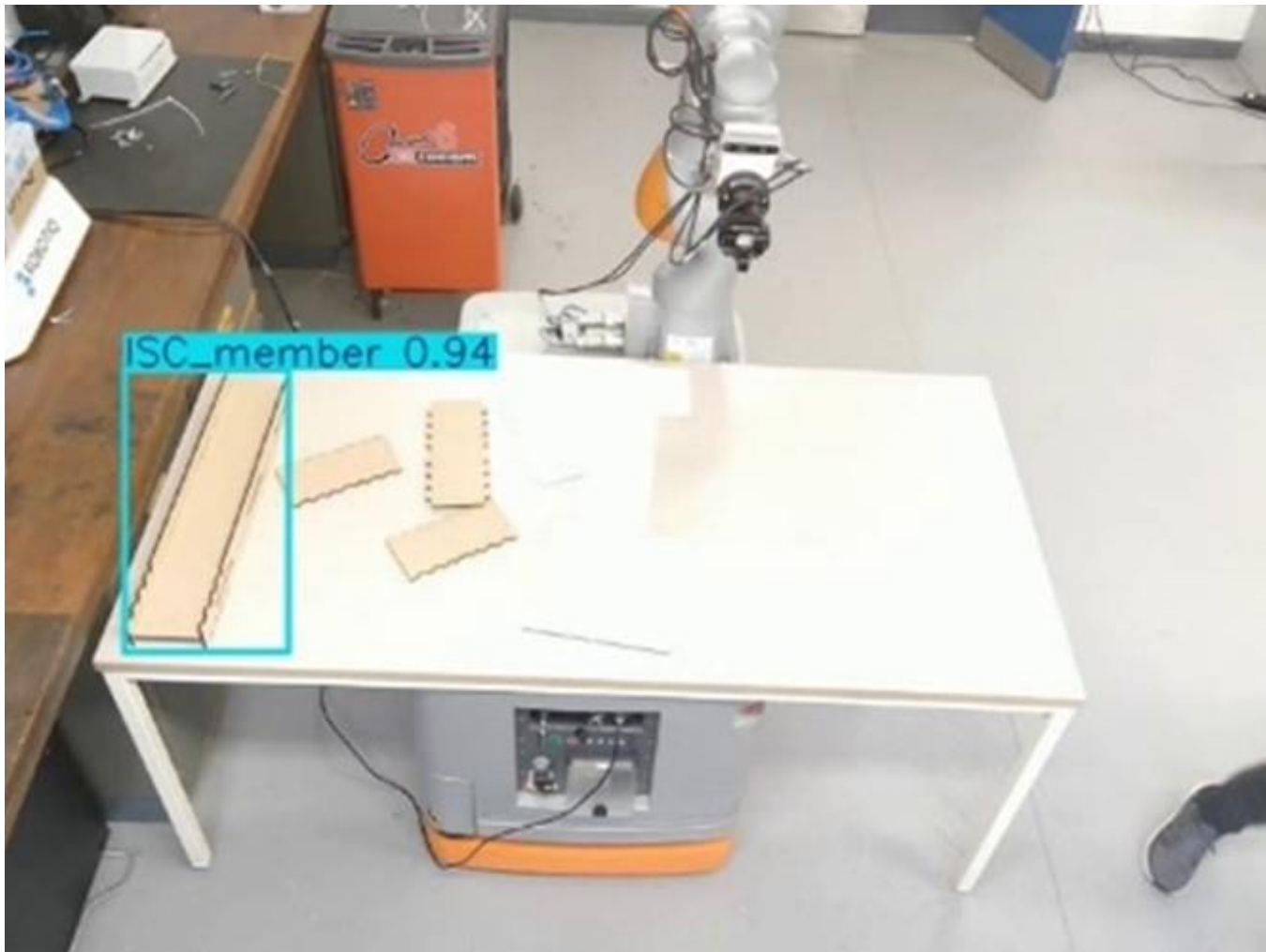


Fig. 10. Glaring in the frontal view impacts detection performance.

- [16] J. Teizer, "Right-time vs real-time pro-active construction safety and health system architecture," *Construction Innovation*, vol. 16, no. 3, pp. 253–280, 07 2016. [Online]. Available: <https://doi.org/10.1108/CI-10-2015-0049>
- [17] B. Zhong, H. Wu, L. Ding, P. E. Love, H. Li, H. Luo, and L. Jiao, "Mapping computer vision research in construction: Developments, knowledge gaps and implications for research," *Automation in Construction*, vol. 107, p. 102919, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0926580519303875>
- [18] M. M. Soltani, Z. Zhu, and A. Hammad, "Automated annotation for visual recognition of construction resources using synthetic images," *Automation in Construction*, vol. 62, pp. 14–23, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0926580515002071>
- [19] K. Mostafa and T. Hegazy, "Review of image-based analysis and applications in construction," *Automation in Construction*, vol. 122, p. 103516, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0926580520310967>
- [20] S. Borkman, A. Crespi, S. Dhakad, S. Ganguly, J. Hogins, Y.-C. Jhang, M. Kamalzadeh, B. Li, S. Leal, P. Parisi, C. Romero, W. Smith, A. Thaman, S. Warren, and N. Yadav, "Unity perception: Generate synthetic data for computer vision," *arXiv*, 2021. [Online]. Available: <https://arxiv.org/abs/2107.04259>
- [21] W. Remmas, M. Lints, and J. J. Uudmäe, "PCGOD: Enhancing object detection with synthetic data for scarce and sensitive computer vision tasks," *IEEE Access*, vol. 13, pp. 91 325–91 333, 2025.
- [22] G. Wang, H. Li, P. Li, X. Lang, Y. Feng, Z. Ding, and S. Xie, "M4SFWD: A multi-faceted synthetic dataset for remote sensing forest wildfires detection," *Expert Systems with Applications*, vol. 248, p. 123489, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417424003543>
- [23] E. Bayraktar, C. B. Yigit, and P. Boyraz, "A hybrid image dataset toward bridging the gap between real and simulation environments for robotics," *Machine Vision and Applications*, vol. 30, no. 1, pp. 23–40, 2019. [Online]. Available: <https://doi.org/10.1007/s00138-018-0966-3>
- [24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. [Online]. Available: <https://doi.org/10.1007/s11263-015-0816-y>
- [25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision (ECCV)*. Springer International Publishing, 2014, pp. 740–755.
- [27] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010. [Online]. Available: <https://doi.org/10.1007/s11263-009-0275-4>
- [28] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, 2020. [Online]. Available: <https://doi.org/10.1007/s11263-020-01316-z>
- [29] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset

- for semantic urban scene understanding,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.
- [30] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. [Online]. Available: <https://doi.org/10.1177/0278364913491297>
- [31] S. Basak, H. Javidnia, F. Khan, R. McDonnell, and M. Schukat, “Methodology for building synthetic datasets with virtual humans,” in *2020 31st Irish Signals and Systems Conference (ISSC)*, 2020, pp. 1–6.
- [32] Unity Technologies, “Unity Perception package,” <https://github.com/Unity-Technologies/com.unity.perception>, 2020.
- [33] C. A. Akar, J. Tekli, D. Jess, M. Khoury, M. Kamradt, and M. Guthe, “Synthetic object recognition dataset for industries,” in *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2022, pp. 150–155.
- [34] W. Qiu and A. Yuille, “Unrealcv: Connecting computer vision to unreal engine,” in *Computer Vision – ECCV 2016 Workshops*. Cham: Springer International Publishing, 2016, pp. 909–916.
- [35] N. Koenig and A. Howard, “Design and use paradigms for gazebo, an open-source multi-robot simulator,” in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, vol. 3, 2004, pp. 2149–2154 vol.3.
- [36] O. Michel, “Cyberbotics Ltd. webots™: Professional mobile robot simulation,” *International Journal of Advanced Robotic Systems*, vol. 1, no. 1, p. 5, 2004. [Online]. Available: <https://doi.org/10.5772/5618>
- [37] E. Rohmer, S. P. N. Singh, and M. Freese, “Coppeliasim (formerly v-rep): a versatile and scalable robot simulation framework,” in *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2013, www.coppeliasim.com.
- [38] G. R. Taylor, A. J. Chosak, and P. C. Brewer, “Ovrv: Using virtual worlds to design and evaluate surveillance systems,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
- [39] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “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)*, 2016, pp. 3234–3243. [Online]. Available: <http://ieeexplore.ieee.org/document/7780721/>
- [40] H. Hattori, V. N. Boddeti, K. Kitani, and T. Kanade, “Learning scene-specific pedestrian detectors without real data,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3819–3827. [Online]. Available: <https://ieeexplore.ieee.org/document/7299006>
- [41] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, “VirtualWorlds as proxy for multi-object tracking analysis,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4340–4349. [Online]. Available: <https://www.computer.org/csdl/proceedings-article/cvpr/2016/8851e340/12OmNzayNkI>
- [42] Y. Cabon, N. Murray, and M. Humenberger, “Virtual kitti 2,” 2020. [Online]. Available: <https://arxiv.org/abs/2001.10773>
- [43] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, “Seeing 3d chairs: Exemplar part-based 2d-3d alignment using a large dataset of CAD models,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3762–3769. [Online]. Available: <https://ieeexplore.ieee.org/document/6909876>
- [44] X. Peng, B. Sun, K. Ali, and K. Saenko, “Learning deep object detectors from 3d models,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1278–1286. [Online]. Available: <http://ieeexplore.ieee.org/document/7410508/>
- [45] S. Du, M. Shehata, and W. Badawy, “Hard hat detection in video sequences based on face features, motion and color information,” in *2011 3rd International Conference on Computer Research and Development*, vol. 4, 2011, pp. 25–29. [Online]. Available: <https://ieeexplore.ieee.org/document/5763846>
- [46] E. Rezazadeh Azar and B. McCabe, “Automated visual recognition of dump trucks in construction videos,” *Journal of Computing in Civil Engineering*, vol. 26, no. 6, pp. 769–781, 2012. [Online]. Available: <https://ascelibrary.org/doi/10.1061/%28ASCE%29CP.1943-5487.0000179>
- [47] S. Chi and C. H. Caldas, “Automated object identification using optical video cameras on construction sites,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 26, no. 5, pp. 368–380, 2011. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8667.2010.00690.x>
- [48] M.-W. Park and I. Brilakis, “Construction worker detection in video frames for initializing vision trackers,” *Automation in Construction*, vol. 28, pp. 15–25, 2012. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0926580512001136>
- [49] A. Del Savio, A. Luna, D. Cárdenas-Salas, M. Vergara, and G. Urday, “Dataset of manually classified images obtained from a construction site,” *Data in Brief*, vol. 42, p. 108042, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352340922002530>
- [50] A. Y. Barrera-Animas and J. M. Davila Delgado, “Generating real-world-like labelled synthetic datasets for construction site applications,” *Automation in Construction*, vol. 151, p. 104850, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0926580523001103>
- [51] tank detect, “Person dataset dataset,” <https://universe.roboflow.com/tank-detect/person-dataset-kzsop>, jul 2025, visited on 2025-09-02. [Online]. Available: <https://universe.roboflow.com/tank-detect/person-dataset-kzsop>
- [52] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, “Label Studio: Data labeling software,” 2020–2024, open source software available from <https://github.com/HumanSignal/label-studio>. [Online]. Available: <https://github.com/HumanSignal/label-studio>
- [53] G. Jocher, A. Chaurasia, and J. Qiu, “YOLO by ultralytics,” original-date: 2023-01-18T08:53:27Z. [Online]. Available: <https://github.com/ultralytics/ultralytics>



Miftahur Rahman is a Lecturer in Robotics at Kingston University London. His research focuses on autonomous robotic systems, mobile manipulators, sensor fusion, construction automation, and computer vision. He holds a Ph.D. in Manufacturing from Cranfield University and has contributed to several interdisciplinary projects and peer-reviewed publications in intelligent inspection and repair robotics.



Samuel Adebayo (Member, IEEE) received the Ph.D. degree in Machine Learning from Queen's University Belfast, U.K., in 2024. He is a Research Fellow in Computer Vision at Queen's University Belfast. His research spans deep learning, causal machine learning, conformal prediction and the integration of psychological principles; perception, cognition, and emotion into computational intelligence.



Dorian A Acevedo-Mejía is a Ph.D. Candidate in Civil and Environmental Engineering at The University of Texas at San Antonio. His research focuses on the development of self-centering horizontal structural systems to enhance the seismic performance of steel structures. He has published in Q1 journals, including the *Journal of Constructional Steel Research*, and in high-impact international conferences such as the World Conference on Earthquake Engineering and the World Conference on Seismic Isolation. He has over 18 years of professional experience as a structural engineer, working on infrastructure and mining projects around the world. His research interests include earthquake engineering, structural dynamics, and the design of resilient steel systems.



David Hester received the B.Eng. degree in Civil Engineering and the Ph.D. degree in Bridge Structural Health Monitoring from University College Dublin, in 2000 and 2012, respectively. He is currently a Senior Lecturer in Structural Engineering with Queen's University Belfast. His main research interests include structural dynamics and bridge structural health monitoring.



Daniel McPolin received the Ph.D. degree in Structural Engineering from Queen's University Belfast. He is a Senior Lecturer in the School of Natural and Built Environment at Queen's, researching engineered-timber composites, high-performance cementitious materials and immersive digital technologies for construction. Dr McPolin gained international recognition for the Guinness-record "world's largest Meccano bridge" outreach project and is a Co-Investigator on the ARISE steel-assembly robotics programme.



Karen Rafferty is currently the Head of the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast. She has over fifteen years' experience working within the fields of software engineering, sensor fusion, and real-time software development, and over ten years' experience within the areas of virtual and augmented reality and multi-sensorial systems. Her research interests include the application of tools and technologies to lead new disruptive practices and systems for many application areas, with a main focus on Health

and Training, and Industry and Automation.



Debra F Laefer received her Ph.D. degree in Civil Engineering from the University of Illinois Urbana-Champaign in 2001. She is a Full Professor of Urban Informatics in NYU's Department of Civil and Urban Engineering and Center for Urban Science + Progress. Prof. Laefer's work bridges geotechnical engineering, remote sensing and urban informatics, emphasising protection of historic fabric during subterranean construction and ultra-dense aerial LiDAR for city-scale modelling. She has published over 160 papers and holds 4 patents. She is co-inventor of the

ISC with Dr. Salam Al-Sabah