Evaluating Video-to-3D Foundation Models for Wire Harness Assembly Verification

Master's thesis in Data Science and AI

MARTIN DIDERHOLM

Industrial and Materials Science

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Evaluating Video-to-3D Foundation Models for Wire Harness Assembly

MARTIN DIDERHOLM

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Industrial and Materials Science
Chalmers University of Technology
Gothenburg, Sweden 2026

Evaluating Video-to-3D Foundation Models for Wire Harness Assembly

MARTIN DIDERHOLM

© MARTIN DIDERHOLM, 2026.

Supervisor: Hao Wang, Department of Industrial and Material Science
Company Supervisors: Isak Ernstig and Patrick Andersson, Wiretronic
Examiner: Björn Johansson, Department of Industrial and Material Science

Master's Thesis 2026
Department of Industrial and Material Science
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: 3D point cloud reconstruction of an wire harness generated from 2D video footage using the Depth Anything V3 foundation model.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Evaluating Video-to-3D Foundation Models
for Wire Harness Assembly
MARTIN DIDERHOLM
Department of Industrial and Material Science
Chalmers University of Technology

Abstract

Automotive wire harness assembly is a heavily manual process, and a critical part of this workflow is Quality Control (QC). Automating this visual verification is challenging: traditional 2D computer vision struggles with spatial occlusions, and lacks essential geometric depth, whereas 3D data can capture a more comprehensive representation of the object. However, existing 3D machine learning methods depend on rigid scanning hardware and extensive training data. To bypass these technical bottlenecks and reduce the industry's critical reliance on human labor, this thesis evaluates emerging feed-forward "Video-to-3D" foundation models for automated assembly verification.

Using a custom dataset of smartphone videos, four state-of-the-art (SOTA) architectures (Pi3-X, VGGT, Depth Anything V3, and Map Anything) were evaluated for geometric fidelity and automated component classification. Results identify Depth Anything V3 demonstrated superior global spatial coherence and strong baseline counting capabilities, though high false-positive rates across the models indicate further refinement is needed. Pi3-x also proved to be a reliable architecture with competitive classification performance. Although current models lack the geometric precision for highly granular verification tasks, often exhibiting wire duplication or loss of detailed structural definition, they show high promise on broader classification tasks, such as distinguishing and counting clips versus terminal housings. While fully autonomous inspection requires further refinement to mitigate high false-positive rates, this flexible 3D approach establishes a step toward modernizing visual quality control of wire harnesses.

Keywords: Wire Harness Assembly, Assembly Verification, Video-to-3D, 3D Reconstruction, Foundation Models, Computer Vision, Quality Control.

Acknowledgements

I want to thank my academic supervisor, Hao Wang, for your expertise, patience, and constructive feedback throughout this thesis.
I am equally grateful to my company supervisors, Isak Ernstig and Patrick Andersson at Wiretronic, for providing such an interesting thesis project. Your practical insights, industry perspective, and resources were essential for this work.
Finally, I would like to thank my examiner, Björn Johansson, for your time and evaluation of this thesis.

Martin Diderholm, Gothenburg, June 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

2D	2-Dimensional
3D	3-Dimensional
CV	Computer Vision
DA3	Depth Anything V3
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
FP	False Positives
ML	Machine Learning
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QC	Quality Control
RGB	Red Green Blue
SAM3	Segment Anything Model 3
SOTA	State-of-the-art
TP	True Positives
VGGT	Visual Geometry Grounded Transformer

Contents

List of Figures

List of Tables

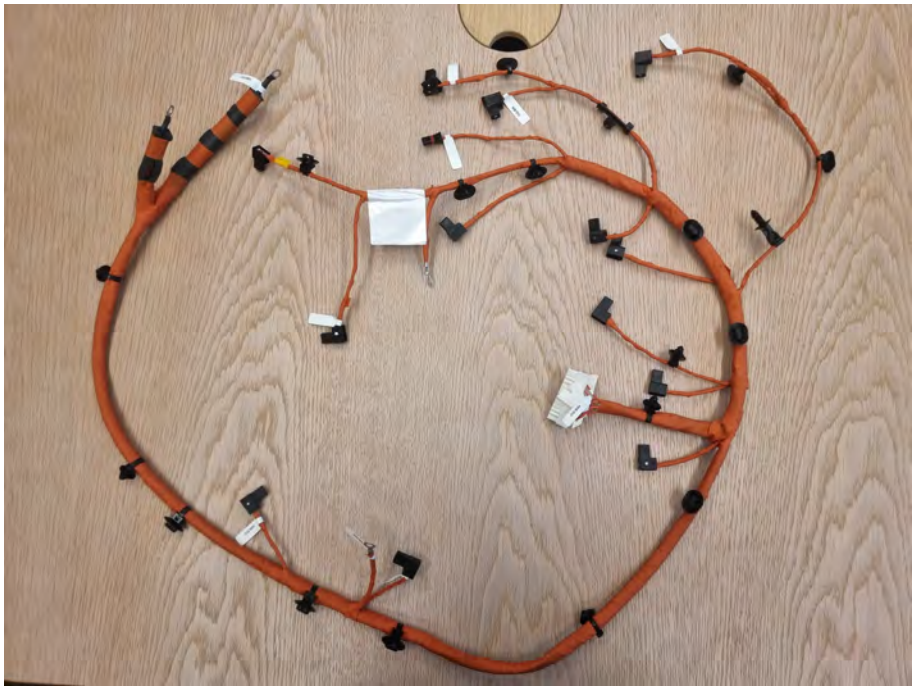
1
Introduction

1.1 Background

The automotive wire harness industry is vital in vehicles, managing functions from basic controls to advanced autonomous systems (see Figure 1.1) [1]. As vehicles become increasingly electrified, these harnesses grow more intricate, yet their production suffers from a low degree of automation. The industry's heavy reliance on manual assembly creates a hard-to-tackle bottleneck [2]. This manual reliance introduces vulnerabilities in product quality, ergonomics, and workforce safety [3]. Quality Control (QC) is an essential part of this process, yet it is currently constrained by these same manual limitations.

To address these limitations, Computer Vision (CV) has become a key enabler for improving the precision and speed of automated defect detection [4]. Empirical studies show that 2D CV techniques are highly prevalent and offer distinct benefits, such as high computational efficiency and proven effectiveness for flat, binary surface inspections [4]. However, when representing complex spatial data, 2D and 3D visualization techniques exhibit fundamentally different characteristics. While 2D imaging benefits from widespread, established use, to use too many overlapping elements into a single 2D image easily results in visual clutter, making it difficult to identify individual objects and general patterns [5]. Conversely, utilizing a 3D data enables an intuitive perception of 3D shapes, extents, and complex spatial distributions. By leveraging the third dimension, layout flexibility is increased, which actively helps reduce the occlusion of intersecting connections and nodes that typically occurs in flat projections [5].

Obtaining 3D spatial data for QC has traditionally relied on industrial scanners. Recently, "Video-to-3D" models, such as VGGT [6], have emerged as an technology capable of reconstructing high-fidelity 3D geometries directly from video footage.



(a) A wire harness.

(b) A mounting clip.

(c) A green terminal housing.

(d) A black terminal housing.

Figure 1.1: Wire harness and key components.

1.2 Problem Description

This thesis examines the problem of automating the visual assembly verification of wire harnesses. A competent automated method for ensuring that every physical component, such as housing and clips is present, accurately counted, and correctly seated before integration. However, the deformable and flexible nature of these components makes it difficult for conventional automated systems to accurately perceive and verify their configuration [2]. Currently, the deployment of computer vision in this domain is severely hindered by the lack of task-specific, annotated datasets [2]. Consequently, the industry faces a critical bottleneck, lacking a reliable automated method to verify assembly configurations without relying on manual labor. A computer vision model capable of extracting configuration information on a wire harness, from a video would bring value to wire harness production.

1.3 Aim and Purpose

The aim of this thesis is to evaluate the feasibility of using emerging "Video-to-3D" foundation models for the automated assembly verification of automotive wire harness. The purpose is to create utility in automating quality control tasks, which could ultimately provide a low-cost, flexible solution for assembly verification of wire harnesses.

To achieve this, the project investigates whether 3D machine learning reconstruction and classification utilizing handheld video footage from consumer grade devices can capture sufficient spatial accuracy to validate a harness configuration.

1.4 Research Questions

To systematically evaluate the use of "Video-to-3D" foundation models for wire harness assembly verification, this thesis addresses three core research questions. Starting with **RQ1** the identification of suitable State-of-the-Art (SOTA) architectures. Then **RQ2** assess the performance of these models in accurately reconstituting 3D geometries under real-world capture conditions. Finally, **RQ3** assesses the practical utility of these models by determining if the generated 3D spatial data can potentially support automated assembly verification.

- **RQ1: Identification and Selection of Foundation Models**
  What are the existing SOTA "Video-to-3D" foundation models, and which architectures demonstrate the theoretical potential to handle wire harnesses?
- **RQ2: Reconstruction Fidelity**
  To what extent can the identified reconstruction models restore the geometric and material characteristics of wire harnesses?
- **RQ3: Feasibility of 3D Data for Assembly Verification**
  Can the generated 3D spatial data be feasibly utilized to perform automated assembly verification tasks?

1.5 Scope and Delimitations

This thesis is strictly bounded to the visual inspection process within Quality Control (QC). To asses the 3D reconstructions, the scope differs between the qualitative and quantitative evaluations. Qualitatively, the models will be evaluated on their overall geometric and visual fidelity across the entire wire harness, exploring theoretical potential to support a broad range of visual QC tasks. However, the quantitative evaluation focuses on the automated assembly verification, restricted to determining the presence, total count and classification of attached small-scale components. Other processes, such as electrical testing, as well as the quantitative validation of wire routing and topology, tape wrapping errors, and micro-surface defects (e.g., scratches or bent pins), are outside the scope for this work.

The technical evaluation is limited to SOTA 3D reconstruction and depth prediction networks released since 2025. To assess these models, the project will solely utilize existing, pre-trained architectures; no models will be trained from scratch, nor will any fine-tuning be done. Furthermore, the resulting 3D reconstruction pipeline is not required to operate in real-time, as this study prioritizes geometric reconstruction quality and spatial fidelity over inference speed.

Data acquisition is restricted to handheld video recordings of static wire harnesses. These recordings will be conducted under controlled conditions, with a consistent background and lighting. The project is conducted in close collaboration with a single wire harness producer. Consequently, the data and component variation will be specific to their manufacturing.

1.6 AI Disclaimer

During the writing of this report Gemini 3 Pro was used exclusively for the purpose of spelling correction, suggesting reformulation and rephrasing to improve readability. Additionally, the AI-powered code editor Cursor was utilized to assist in the programming and implementation of the pipeline. No core concepts, data, or conclusions were generated by these models. The author takes full responsibility for the final content of this thesis.

2
Theory

2.1 Wire Harness

An automotive wire harness is an essential assembly of electrical cables used to connect a vehicle's electronic components, control units, sensors, and actuators [7]. Its core purpose is to guarantee a reliable electrical connection among these components throughout the vehicle's lifespan. Within a modern vehicle, several thousands of individual wires are bundled together using materials like cable ties, straps, tape and cable ducts to transmit signals and electrical power [7]. Given this complexity, the wire harness represents one of the most expensive parts manufactured for a car. The importance of the wire harness is escalating rapidly with current automotive trends shifting more towards electric cars and autonomous driving. These advancements demand high-voltage harnesses, higher data rates, and stricter requirements for the reliability and redundancy of electrical connections, while driving the need for smaller wire diameters and miniaturized terminal sizes to reduce vehicle weight [7].

2.1.1 Quality Control in Wire Harness Production

As these wire harnesses transmit electrical signals and power throughout the vehicle, rigorous QC is a paramount stage in production. The function of a wire harness is to guarantee reliable and an error-free connection over its entire lifespan, which is generally 15 to 20 years [7]. Any defect can lead to severe operational failures or electrical fires, making a mandatory 100% function and completeness test necessary for every completed harness [7].

Quality assurance protocols, such as the IPC/WHMA A-620 standard, dictate strict acceptability requirements for cable and wire harness assemblies [8]. To meet these standards, testing must identify and mitigate several common defects which arise. These common manual manufacturing flaws primary include open and short circuits, misconnections and mechanical integrity issues such as partially terminals [8].

To verify that quality standards are met, manufactures currently rely on a combination of manual and electrical inspection methods [9]. Operators routinely perform general visual and physical examinations to assess harness quality, checking for structural anomalies such as partially seated terminals [9]. These visual assessments are time-consuming and heavily dependent on the individual inspector's skill and subjective judgment. Because they are inherently prone to human error and

inconsistent evaluations, these manual checks frequently fail to detect subtle defects [9].

Following visual checks, harnesses are taken off the mounting board and placed onto specialized testing tables. operator must manually untangle the wires and fit the connectors and attachment parts into specific testing brackets. Once mounted, and electric function test and a completeness check are performed [7]. These automated checks primarily focus on continuity testing to systematically identify short circuits, open circuits, and miss-wired conductors throughout the assembly [8]. Additionally, if the harness utilizes plugs with secondary locks, the testing process also incorporates a physical check to verify a tight, secure fit of the locking mechanisms.

2.1.2 Automated Quality Control

To address the limitations of manual inspection, researchers are actively exploring automated quality control systems that leverage advanced sensor technologies and collaborative robotics. While the industry has not yet been able to fully of partially automate the wire harness assembly process [1], developing systems that can adapt to high physical variability of these components reamins a critical area of study.

Computer vision systems have been proposed for tasks such as automated defect detection and process verification [1]. In research settings, computer vision is being investigated to potentially monitor tasks and verify correct performance, such as checking if adhesive tape is correctly positioned during spot-taping operations [10]. Furthermore, vision-guided systems frequently utilize neural networks to classify different harness branches and identify specific terminals, which could eventually aid in alignment throughout the assembly sequence [10].

Beyond visual inspection, automated quality control relies on multi-model feedback to ensure mechanical integrity. Force-torque sensors are utilized alongside multi-camera systems to measure and verify the precise mating of electrical connectors [1]. Additionally, alternative automated feedback systems have been developed to verify mechanical seating, such as microphone setup designed to detect the specific "clicking" sound that confirms a successful connection [1].

While most previous research on vision-guided robotized assembly has focused on identification tasks using 2D images and basic image processing methods. However, recent advancements in imaging technology and computer vision research are enabling more possibilities for better vision systems [1]. Consequently, novel vision-based methods for 3D profile extraction have been developed to more reliably handle and verify these highly deformable components in an automated setting [10].

2.2 2D-to-3D Theory

To reconstruct a 3D physical environment from video, a vision system must mathematically reverse the image formation process. This requires mapping physical space to digital pixels across three primary coordinate systems:

- **World Coordinate System**: A fixed, absolute 3D reference frame for the global environment
- **Camera Coordinate System**: A local 3D frame originating at the camera's optical center
- **Pixel Coordinate System**: The 2D digital grid of the final captured image

(See Figure 2.1 for a visual representation of these coordinate systems and transformations.)

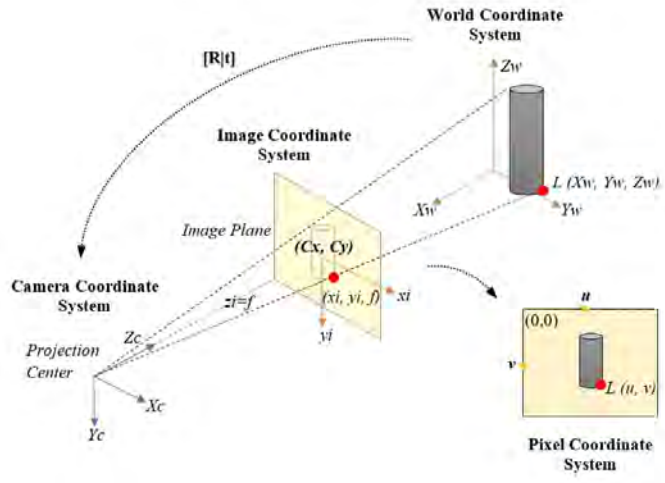


Figure 2.1: 3D-to-2D projection visualization. Adapted from Ortiz et al. (2017). Licensed under CC BY 4.0.

Camera Extrinsic and Intrinsics To mathematically project a 3D world point $L(Xw, Yw, Zw)$ onto a 2D image plane point $L(u, v)$ relies on a combination of extrinsic and intrinsics, also called camera parameters. This relationship will be described using the idealized pinhole camera model and a arbitrary scale factor Zc(representing the depth in the camera coordinate system). The projection is expressed with the following equation [11]:

$$Zc \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = A[R|t] \begin{bmatrix} Xw \\ Yw \\ Zw \\ 1 \end{bmatrix}$$

- **Extrinsics:** The Transformation from the global world coordinate system to the local camera coordinate system is handled by the extrinsic parameters, specifically a rotation matrix R and a translation vector t, denoted as $[R|t]$.
- **Intrinsics:** Once relative to the camera, the 3D point is projected onto the 2D sensor via the camera intrinsic matrix, denoted as A. The pinhole camera

model defines the camera's internal hardware through scaling factors for the horizontal and vertical pixel axes (f_x and f_y), as well as the coordinates for the principal point, which is the optical center of the image sensor (Cx, Cy). A fully comprehensive matrix can also include a skewness parameter (γ) to account for any any physical tilt between the pixel grids; however, under the idealized pinhole mode, the axes are assumed to be perfectly orthogonal, meaning this skewness parameter is set to zero.

The intrinsic matrix is formulated as:

$$A = \begin{bmatrix} f_x & \gamma & Cx \\ 0 & f_y & Cy \\ 0 & 0 & 1 \end{bmatrix}$$

While projection from 3D space to a 2D image is mathematically straightforward, while reversing the process is more challenging. When examining the projection equation $Zc \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = A[R|t] \begin{bmatrix} Xw \\ Yw \\ Zw \\ 1 \end{bmatrix}$, the scale factor Zc corresponds to the distance of the object from the camera. Because this depth value is flattened into the 2D plane during capture, it becomes a unknown when trying to invert the equation. Therefore, a single 2D pixel does not map back to a single 3D coordinate. Instead, it maps to an infinite 3D line, or "ray", originating from the camera's optical center and extending into the world. Any point along that ray will project onto the exact same pixel [12].

Because of this fundamental loss of depth, it is geometrically impossible to reconstruct a 3D structure from a single static image without prior knowledge. Traditionally, the system solves this by capturing the same global point from two or more different camera positions, projecting multiple rays into the 3D space to find where they intersect. However, before rays can be intersected, the system must first identify which pixels across different images represent the exact same physical point. Finding these correspondences is a highly complex challenge due to variations in lighting, occlusion, and perspective [13]. The method used to identify and mathematically describe these distinct points so they can be matched is known as feature extraction.

Alternatively, the missing scale factor Zc can be solved directly if the depth at each pixel is already known, bypassing need for multi-view intersection [12]. Consequently, both feature extraction and depth estimation for are both dedicated fields of research within computer vision.

To manage the geometric data resulting from these projections, *Open3D* is a open-source library for 3D data processing [14]. It provides the computational infrastructure to take parameters such as, camera intrinsics, extrinsics, and depth maps and project them into a 3D scene, also able to convert these scenes into point clouds.

2.2.1 Computer Vision

Computer vision is a multifaceted field that aims to replace aspects of human visual capabilities, enabling machines to perceive, interpret, and make decision based on visual data [15]. Where the primary objective is to create models that can from images extract relevant data to achieve "image understanding".

It integrates two foundational areas:

- **Image Processing:** Optimizes clarity and extracts features using computational transformations like contrast adjustment and sharpening.
- **Pattern Recognition:** Identifies and categorizes specific objects or patterns within the enhances images.

By leveraging optical sensors and artificial intelligence, computer vision analyzes large datasets for tasks like object identification and scene understanding. Consequently, computer vision is a highly versatile technology with a wide array of applications, ranging from automation in manufacturing and assembly lines to robotics and automated quality control [15].

Foundation Models

Foundation models represent a broad concept within artificial intelligence that spans across all sub-fields, including computer vision [16]. At their core, these are massive models trained on a huge amount of general data, usually without needing humans to manually label everything (a process called self-supervision) [16]. Because they serve as an incomplete but flexible starting base that can be adapted to a wide variety of specific tasks, they are called foundation models [16].

Their main two characteristics are: emergence and homogenization [16]. Emergence means that because these models are so large, they naturally develop new, unexpected skills on their own without being explicitly programmed to do so [16]. Homogenization means that a single, powerful model can be used as the building block for many different applications [16]. While this is incredibly efficient, it also means that any flaws in the foundation model will be passed down to all the systems built on top of it [16]. Within computer vision, foundation models are increasingly being trained on massive amounts of raw data from the web instead of carefully hand-labeled datasets [16].

2.2.2 Point Cloud Clustering

Once a 3D point cloud is generated, a way to cluster the unstructured spatial data is with: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [17]. Unlike partitioning algorithms (such as k-means) that require the total number of clusters to be predefined, DBSCAN groups points together based purely on local spatial density. It has two primary parameters: a search radius (ϵ) and a minimum number of points ($minPts$). The algorithm identifies core points that have at least $minPts$ within their ϵ-neighborhood and expands outward to form contiguous clusters [17].

Crucially, any points that fail to meet these density criteria and fall outside the dense regions are explicitly classified as spatial noise. This mechanism makes DBSCAN well-suited for isolating distinct physical components naturally filtering out noise.

2.3 State-of-the-Art 3D Reconstruction Architectures

Recent advancements in 3D computer vision have shifted away from slower methods that required multiple steps to build a 3D model [6]. Instead, SOTA architectures handle 3D reconstruction as a direct, single pass (known as a feed-forward prediction) prediction task [6, 18]. Even though these models might be designed for slightly different tasks—like estimating depth from a single image or building a scene from multiple views—they all share a very similar framework built on three main pillars:

2.3.1 Pre-trained Vision Transformer (ViT) Backbones

Rather than training feature extractions from scratch, these feed-forward 3D models leverage foundational Vision (ViT) (predominantly DINOv2) as their base [19, 20]. By dividing input images into small patches, called tokens (similar to LLMs). Because these transformers have already been trained on massive datasets, they provide the 3D model with a strong, general understanding of visual shapes and patterns right out of the box [20].

2.3.2 Alternating Self-Attention Mechanisms

To make sure the 3D reconstruction is accurate from every angle, the model cannot just look at images in isolation. Instead, it uses an "attention" mechanism that constantly switches between looking at patches within a single image, and comparing patches across all the different images. This allows the model to share spatial context and figure out how different viewpoints connect to each other. By doing this, the model effectively skips the traditional, difficult step of having to manually match features and calculate geometry, allowing it to easily handle any number of input images [6, 18].

2.3.3 Transformers-Based Dense Decoders

Once the model has compared all image patches and understands how they relate across different views, it must translate this abstract information back into physical 3D space. This is done using what is called a dense decoder [18, 20]. The decoder take the processed patches and maps them directly into practical 3D data, predicting things like the exact depth of every pixel, local 3D coordinates, and the position of the cameras [6, 19, 20].

3
Methodology

This chapter explains the methods used to test if 3D computer vision models can automatically inspect wire harnesses. First, it details the literature reviews conducted to find the best current methods and SOTA video-to-3D models. Next, it describes the experimental setup, which includes creating a custom video dataset and building a 3D processing pipeline using Segment Anything Model 3 (SAM3) [21]. The chapter then outlines how the results were evaluated, combining expert interviews to judge visual quality with an automated method to count wire components. Finally, it reviews the overall reliability and validity of the study.

3.1 Literature Review: 3D Computer Vision for Wire Harness Inspection

The aim of this systematic literature review is to analyze existing academic literature on computer vision-driven 3D reconstruction methods and the utilization of 3D data representations for automated inspection of wire harnesses. To achieve this aim and establish a comprehensive understanding of current technological capabilities, this review adheres to the methodological guideline outline in [22]. This referenced framework provides a widely recognized standard for conducting systematic review withing the domains of software engineering and applied computer science.

3.1.1 Literature Search

The literature search was conducted using Scopus. Scopus was selected as the search engine due its comprehensive coverage of peer-reviewed literature within engineering, computer science, and related multidisciplinary fields. While Google Scholar remains a highly comprehensive database that captures a significant amount of extra coverage, including non-peer-reviewed scientific papers, which Scopus excludes [23]. Compared to other academic search engines, Scopus provides advanced search functionalities that allow researchers more strict and reproducible searches.

The search query was constructed in three segments to capture the intersection of the domain, the task, and the technology. The first segment, '(wir* or cabl*) AND (harness* OR bundl*) AND assembl*, targets the core subject of the study: wire harness assembly. The second segment, (verif* OR "OC" OR "quality control" OR qualit* OR inspect* OR validate*), focuses the search strictly on QC and verifi-

cation processes. The final segment, encompasses a comprehensive list of terms to ensure the inclusion of methodologies utilizing 3D CV and spatial data representations. The search was done on the 19th of march and resulted in 46 documents.

To account for variations in terminology, verb endings, and plural forms, the asterisk operator (*) was utilized. this entire search string was matched against the fields of Article title, Abstract and Keywords. Table 3.1 outlines the complete search definitions including the inclusion and exclusion criteria.

Table 3.1: Overview of the Review Criteria

Review Criteria	**Details**
Database	Scopus
Search string	((wir* OR cabl*) AND (harness* OR bundl*) AND assembl* AND (verif* OR "QC" OR "quality control" OR qualit* OR inspect* OR validate*) AND ("video-to-3D" OR "3D" OR "3-D" OR "three-dimensional" OR "three dimensional" OR "computer vision" OR "machine vision" OR "point cloud*" OR "3D mesh" OR "depth map" OR "RGBD" OR "RGB-D" OR voxel* OR photogrammetr* OR "depth sensor*" OR depth OR "LiDAR" OR "structured light"))
Search field	Article title, Abstract, Keywords
Subject area	Engineering; Computer Science; Multidisciplinary; Physics and Astronomy; Mathematics; Decision Sciences
Article language	English
Inclusion criteria	Proposing or evaluating 3D computer vision, 3D reconstruction, or the use of any type of 3D data for the quality control, inspection, or assembly verification of physical wire harnesses. Must be a primary study (original research).
Exclusion criteria	Evaluating electrical functional testing or continuity; relying solely on 2D vision or manual inspection; focusing exclusively on CAD design/simulation without physical objects; review papers.
Search date	March 19, 2026

3.1.2 Literature Selection

The inclusion and exclusion criteria for the primary studies were systematically applied. Following an initial English-language filter, the screening was performed in two stages. The first stage assessed the titles and abstracts according to the criteria, discarding reviews or studies lacking relevance to wire harnesses and 3D data. The second stage involved a comprehensive full-text review fo the retained papers. In

this stage, papers that did not align with the research scope, such as those lacking actual 3D data utilization or focusing exclusively on robotic manipulation rather than inspection were discarded. Papers meeting all criteria were included. Finally, to mitigate potential selection bias, backward "snowballing" was conducted on these qualifying articles [23], yielding the final set of studies.

3.2 Model Selection

3.2.1 Initial Literature Search Strategy

To establish the current SOTA landscape regarding foundational models within 3D reconstruction and depth estimation, an initial systematic literature search was conducted on March 22, 2026. The primary objective was to identify comprehensive review articles, surveys, and SOTA overviews.

The search was conducted using the Google Scholar database. Because the field of generative and AI and foundation models advances rapidly, Google Scholar was specifically chosen for its inclusive, automated web-crawling approach and its ability to aggressively index pre-prints (e.g., arXiv) [23]. This approach prioritizes recency and circumvents the publication lag inherent in traditional, peer-reviewed databases, ensuring the most immediate technological developments are captured. To capture the most recent advancements, the search was restricted to articles published 2025 and filtered specifically of "Review articles." The search was executed across all text fields (Anywhere in the article) without subject area filtering, using the following Boolean search string:
"3D reconstruction" OR "depth estimation" foundation model" OR "zero-shot" review OR survey OR "state-of-the-art"

3.2.2 Selection criteria

To ensure relevance to the specific domain of generalized and fine-grained object reconstruction, strict inclusion and exclusion criteria were applied to the search results.

Inclusion criteria:

- Be a comprehensive review, survey, or SOTA overview paper
- Focus on or include the evaluation of foundational or zero-shot models for 3D reconstruction or depth estimation.
- Include discussion or benchmarks utilizing multi-view or video-to-3D/depth inputs.
- Demonstrate relevance to or capabilities regarding generalized environments, small objects, or deformable linear objects (DLOs), such as wires.

Exclusion criteria:

- Consisted of single-model proposal or individual case studies rather then aggregate reviews.
- Restricted to specific applications (e.g., autonomous driving, city-scale mapping, or medical applications) lacking applicability to smaller or fine-grained

objects.

3.2.3 Frontier Model Identification Strategy

Due to the rapid evolution of 3D reconstruction and depth estimation architecture, relying solely on published systematic reviews risks not including the most recent advancements. To ensure the methodology capture the current frontier of the field and bypasses the latency inherent in academic publishing, the initial literature search was supplemented with a retrieval strategy utilizing open-source pre-trained model registers. Hugging Face was selected as the repository for this search, as it is recognized as the most popular open model registry hosting the largest number of pre-trained model packages [24]

To systematically navigate this repository, models were filtered using the platform's native "Tasks" filter, specifically targeting the "Image-to-3D" and "Depth-Estimation" domains. Within these specialized categories, the last 30 days total download counts was utilized as the selection mechanism to identify the top models. The search was done 7th of April 2026.

While popularity does not guarantee performance, high community traction serves as an objective proxy for rapid adoption, with model performance acting as a plausible latent variable driving that popularity [24]. Furthermore, high download counts strongly correlate with superior documentation quality and community reuse [24], [25]. Therefore, the models acquired through this process represent tools a the AI community has recognized as capable and functional.

To complement this selection process, a recency constraint is applied. The foundation model landscape is highly volatile, as [24] observed, approximately half of the top 1,000 most downloaded models on Hugging Face are replaced within a 22-month period. Therefore, applying a recency filter to the repository search ensures the selection focuses strictly on more recent architectures.

Following the identification of candidate models through repository mining, a secondary verification was implemented. The selected candidate architectures were cross-referenced with their primary publications to evaluate them for documented, quantifiable claims of SOTA performance. While it is recognized that performance claims of a model's own publication are inherently subject to bias, this review was not intended as proof of performance. Instead, this verification was primarily conducted to establish the scientific inclination of the models creators, ensuring that the architectures gaining community traction are grounded in research methodologies rather than being undocumented implementations.

3.3 Experimental Study

Because it is impossible to predict how SOTA "Video-to-3D" foundation models will perform on the characteristics of wire harnesses without testing, an experimental study on a wire harness dataset was conducted. This phase of the research measures the practical capabilities of the selected models when applied to real-world assembly verification tasks. The research design for this study is structured into the following stages:

- **Dataset Creation**: A custom video dataset will be recorded using consumer-grade smartphone in a semi-controlled environment. It will feature manually labeled wire harness configurations to establish a ground truth of testing.
- **Model Implementation**: The selected models will run inference on the test set to generate 3D spatial data from the 2D videos, utilizing segmentation to isolate the harness geometry.
- **Evaluation Framework**: The 3D reconstructions will be evaluated qualitatively through human assessment to grade geometric fidelity, followed by quantitative evaluation measuring assembly verification accuracy against ground truth labels.

The objective of this experimental study is to propose a pipeline capable of generating 3D spatial representation of wire harnesses from standard video footage, and to determine if this spatial data can feasibly replace or enhance existing visual inspection methods for QC.

3.3.1 Dataset Creation

For this experimental study, two distinct wire harnesses with varying configurations provided by Wiretronic. To generate the dataset, these harnesses were video recorded using a Samsung S21 Ultra with a ratio of 1x1. This was done to simulate a flexible, low-cost capture workflow. All Videos were recorded against a consistent, standard background under static lighting conditions. During the capture process, each wire harness remained stationary while the camera was moved in a trajectory around the object, ensuring the multi-view perspectives. Following the data acquisition, the resulting videos were annotated in accordance to the present harness, detailing the component counts, such as terminal housings and clips.

3.3.2 Model Implementation

To execute and compare the selected SOTA foundation models, and end-to-end pipeline was developed. This pipeline established a standardized flow of data, from raw video input to isolated 3D geometry, ensuring that each model is subject to equal constraints. The architecture of this pipeline is divided into three primary stages: pre-processing, global scene reconstruction, and 3D segmentation.

Pre-processing Given the high compute needed to processing continuous video through large Machine Learning (ML) architectures, a frame extraction strategy was used. For each video in the dataset, 30 evenly spaced frames were sampled. To

ensure a fair comparative analysis, the same set of 30 extracted images from each video was utilized as the input for all four models.

Model Inference and Global Reconstruction

The pre-processed frame sequences were fed into the selected reconstruction architectures. When running the inference, each model generates a comprehensive spatial prediction of the entire scene, which includes camera parameters and pixel to world coordinates.

To ensure a fair comparison across models with different underlying architecture, all generated standardized 3D data. The Open3D [14] library was integrated into the pipeline to handle the structural output, converting the model predictions into a point cloud format. Consequently, the final output from every model was saved as a ‘.ply‘ point cloud file. This guarantees that subsequent processing and downstream tasks interact with the same data structure regardless of model. Alongside the .ply file, the pipeline also generates intermediate bridge files containing normalized per-pixel spatial data (XYZ maps), which are essential for subsequent segmentation.

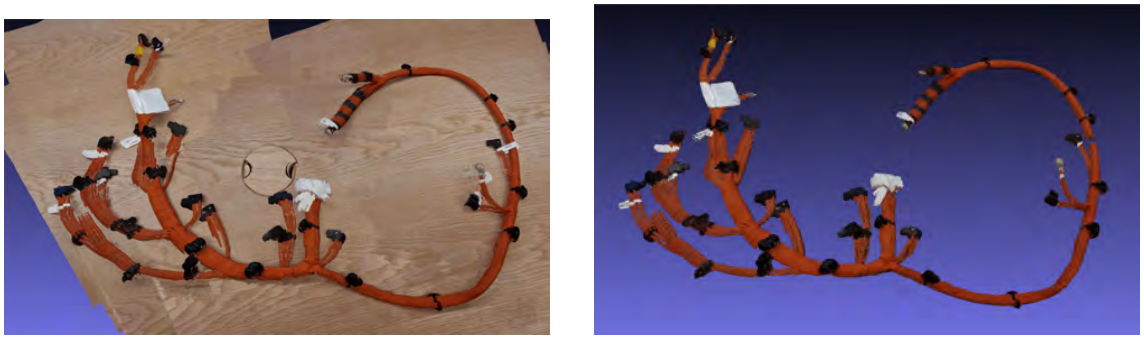
3D Segmentation

Because the global reconstruction captures the entire scene including the background and testing surface, the wire harness must be geometrically isolated. To achieve this, SAM3 [21] was integrated into the pipeline. A image segmentation model, being able to perform segmentation through text based prompts.

The pipeline utilizes SAM3 [21] in a post-reconstruction as a mask-guided filtering capacity. A SAM3 [21] inference on the 30 frames, utilizing the text-based prompt (‚such as "wire harness") to generate and propagate accurate object masks across all frames. After the 2D binary masks are created for all frames, the pipeline isolates the specific pixels representing the wire harness. These segmented pixels are then directly projected into global 3D coordinates using the spatial maps generated during inference. Finally, the system filters out all unmasked background geometry and saves only the projected 3D coordinates of the harness as a new point cloud, with Open3D [14]. Finally exports a ‘.ply‘ file containing only the reconstructed wire harness. See 3.1 for a visualization of the segmentation.

3.4 Evaluation Framework

A primary challenge in evaluating the geometric fidelity of the generated 3D reconstructions is the absence of ground truth 3D data. Because the wire harnesses are captured exclusively via video standard mathematical benchmarking (e.g., calculating spatial deviation or Chamfer distance) cannot be performed. Therefore, to assess the performance and of the model without a 3D baseline, a qualitative human



(a) Full 3D reconstruction **(b)** Segmented

Figure 3.1: 3D point cloud representation of the orange wire harness non-segmented and segmented views.

evaluation framework is used. This to understand the practicality of the models reconstruction. This is followed by a quantitative evaluation to measure task-specific accuracy by comparing the pipeline's automated component count against the physical ground truth of the wire harness.

3.4.1 Qualitative Evaluation

Knowledge of industry professionals regarding what constitutes a "successful" reconstruction, a qualitative human evaluation is conducted. This evaluation follows the methodological guideline fo the problem-centered expert interview outlined by [26]. Following this methodology, the interviews are structured into four sequential phases:

- **Phase 1. Individual Opening:** A brief introductory phase focusing on the expert's professional background, their specific role, and their primary focus when interacting with physical wire harnesses or CV systems.
- **Phase 2. Narrative Beginning:** An open-ended question where the expert is presented with the generated 3D point clouds alongside the physical wire harness reference. The expert is invited to freely explore the model and "think aloud," narrating their observations regarding fidelity, structural geometry, prominent flaws, and how they visually process the data for inspection.
- **Phase 3. General Explorations:** Follow-up question that delve into selected episodes from the expert's initial narrative. This phase investigates the general geometric usability of the models, asking the interviewee for example how well the 3D data captures difficult components (e.g., connectors and clips) and its suitability for building automated verification tools.
- **Phase 4. Specific Explorations:** Targeted questions that directly reference the interviewee's previous accounts to clarify specific meanings and technical requirements. This phase focuses on practical feasibility, asking the expert to specify exactly which geometric details are missing or not from the current results to constitute a reliable 3D data from automated QC in production.

To ensure the interviewees can articulate with extra nuances and knowledge as naturally as possible, the interviews will be conducted in Swedish. However, resulting

data analysis will be translated to english and define the practical parameters of model success and identify critical failure modes of the models.

3.4.2 Quantitative Evaluation

To quantitatively assess the practical utility of the generated 3D reconstructions for automated assembly verification, a counting methodology was developed to measure task-specific accuracy. Specifically, the pipeline evaluates the ability to accurately count terminal connectors and clips on the reconstructed wire harnesses.

The initial stage leverages SAM3 [21] to generate semantic masks for each RGB frame in the multi-view sequence. The segmentation model isolates pixels corresponding to target components using text prompts. To determine the most robust segmentation strategy, a variety of naming conventions and descriptive prompts were evaluated for the components, of connectors and clips. The prompts were slightly altered depending on the specific wire harness being processed. For instance, the modifier "black" was added to the text prompt for the harness where the terminal housings were black.

Consistent with the 3D segmentation method detailed previously, the generated 2D binary masks of segmented pixels are subsequently mapped to their corresponding global 3D coordinates using the intermediate spatial XYZ maps. Following this projection, the 3D data extracted from every frame is aggregated into a single point cloud within the original reconstruction coordinate frame, with the help of Open3D [14]. Essentially creating a spatial representation of where the segmentation model estimates the components are located then projected into 3D space. With this a component segmented from varying perspectives maps to a consistent 3D location, if multiple images agree on a location, it creates a denser point cluster, while false positives remain sparse.

To extract the component count, the accumulated point cloud is processed using the DBSCAN algorithm. DBSCAN was selected because it does not require the number of clusters to be predefined, works with arbitrary cluster geometries, and removes noise by not including points that fail the density criteria. This noise modeling works in tandem with filtering out sparse points generated by false positives.

To evaluate the accuracy of the pipeline, the clustered point cloud was overlapped with the complete reconstructed wire harness. A correctly segmented and overlapping cluster was counted as a True Positive (TP), additionally each real component on the physical wire harness could only counted once, and each classified cluster could only be assigned to a single real component.

False positives were subsequently derived by subtracting the number of TP from the total number of generated clusters. Finally, these values alongside the real amount of components present on the physical wire harness, the precision and recall of the models were calculated to determine the pipeline performance.

3.5 Research Quality and Validity

To ensure internal validity during the literature reviews, clear inclusion and exclusion criteria were defined prior to the searches to minimize research bias and ensure relevant studies were evaluated. For the experimental study, internal validity was maintained by strictly standardizing the pipeline. All evaluated architectures processed the exact same 30-frame sequences. Additionally, uniform post-processing was applied to generate standardized point clouds across all tests, isolating the modes capabilities from external variables.

Regarding external validity, the finding of this study are intended to generalize within the domain of wire harness assembly and quality assurance. To ensure the evaluation is not limited to a single prototype, the experimental dataset included distinct wire harness configurations representing component variations found in the industry. Furthermore, by capturing data with a standard consumer-grade smartphone rather than specialized industrial scanners, and by evaluating open-source foundation models, the resulting pipeline demonstrates direct applicability to flexible, low-cost production setups. However, a notable limitation to the external validity is the small size of the dataset, which consists of only two distinct wire harnesses. While sufficient for a proof-of-concept experimental study, a larger and more varied dataset would be required to fully validate the pipeline.

To validate reliability, the literature searches can be reproduced on Scopus and Google Scholar using the documented Boolean strings and search criteria. Because no single database captures all literature, backward snowballing was utilized to mitigate publication bias. For the experimental phase, reproducibility is ensured with the deterministic nature of the pipeline. Because the study utilizes pre-trained foundation models running only running in inference, passing the same 30 images will yield the same 3D spatial data. All of the steps of the data processing steps are documented allowing for replication. However, a limitation to the study's overall reliability is that the literature screening and qualitative evaluations were conducted by a single researcher, which introduces the risk of unintentional selection bias.

4

Results

This chapter presents the findings from the evaluation of video-to-3D foundation models for the automated assembly verification of wire harnesses. First, it details the results of the systematic literature review. Next, it outlines the model selection process, which establishes the four chosen architectures for testing. The chapter then examines the visual reconstruction results of these models across two distinct wire harness configurations. Following this, it synthesizes qualitative evaluations from industry experts regarding the models' visual fidelity, spatial coherence, and practical feasibility for inspection tasks. Finally, it provides a quantitative assessment of the 3D processing pipeline's ability to segment and count terminal housings and mounting clips, highlighting the best-performing models and analyzing their common failure modes.

4.1 Literature Review: 3D Computer Vision for Wire Harness Inspection

The original database query yielded 46 articles. After filtering by language, 2 non-English documents were removed. The Remaining 44 papers underwent the first stage of screening, which involved reading the titles and abstracts to deduce whether the documents met the predefined inclusion and exclusion criteria. During this first screening, 28 articles were discarded as they did not align with the research scope. Therefore, 16 papers were advanced to the second stage of screening. In the second stage, the full texts of these 16 papers were read thoroughly, resulting in 6 primary studies being selected that fully satisfied all the inclusion and exclusion criteria. Finally, backward snowballing were conducted on these 6 studies, but no additional papers were included. Figure 4.1 shows a PRISMA flow chart of the reviewed stages.

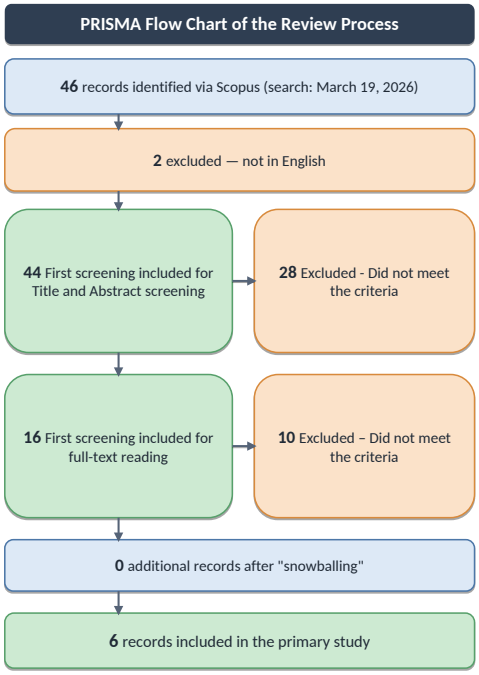


Figure 4.1: PRISMA Flow Chart of the Literature Review

4.1.1 Chosen papers

The table 4.1 presents an overview fo the primary studies gathered from in the systematic literature review. All selected papers focus on the application of 3D geometric data and machine learning techniques for the automated inspection or assembly verification of wire harnesses. The table 4.1 describes the core information extracted from each study, authors, the underlying methodology and 3D acquisition method (such as stereo cameras or 3D scanners), and the main findings. Specifically, the table highlights various approaches to 3D spatial data gathering and understanding.

4.1.2 Machine learning Approaches for creating 3D Geometric Data of Wire Harnesses

A primary bottleneck for deploying machine learning in wire harness inspection is the lack of large, high-quality 3D training data [27]. While researchers have utilized synthetic point clouds generated from 3D CAD models to automate labeling and overcome data scarcity, this approach introduces a "domain gap" between flawless simulated environments and noisy real-world sensor data [27]. Consequently, models trained purely synthetically struggle to generalize to physical deployments yielding accuracies as low as 28 % and still require a substantial blend of costly,

Table 4.1: Summary of the Selected Primary Studies

Study	Methodology & Technology	Main Findings
Kim et al. (2026)	**Tech:** Optical Coherence Tomography (OCT). **Method:** Converts 3D volumetric datasets into 2D depthwise images, applying median filtering and adaptive thresholding to measure terminal engagement depth.	Successfully detected subsurface defects (e.g., partial disconnections) missed by visual/electrical tests. Faster than 2D camera equivalents (0.56s vs 3.8s).
Okuyama et al. (2025)	**Tech:** Robot-mounted 3D stereo camera. **Method:** Extracts cable points by comparing pre- and post-routing point clouds, then interpolates points using B-Spline and Bézier curves.	Difference-based extraction effectively isolated cables. B-Splines yielded smoother paths, while Bézier curves captured sharp bends accurately (Chamfer distances 0.989–3.498 mm).
Nguyen et al. (2022)	**Tech:** 3D scanner and CAD/simulation software. **Method:** Uses PointNet++ for semantic segmentation, evaluating training strategies that mix synthetic and real point cloud data.	Pure synthetic training failed on real data (28.42% accuracy). A mixed dataset (70% synthetic, 30% real) achieved the best performance (94.62% mean IoU).
Ben Abdallah et al. (2020)	**Tech:** Robot-mounted 3D scanner. **Method:** Aligns one-shot real clouds to CAD models using SLAM and ICP, segmenting cables by iteratively fitting 3D subcylinders.	Reconstructed cables from sparse scans with 0.3 mm mean error. Successfully verified minimum bend radii and safe-distance interference.
Galassi et al. (2024)	**Tech:** Point cloud analysis with neural networks (PointNet++, RS-CNN). **Method:** Proposes a shared-encoder architecture with clip-specific decoders to detect misplaced wires during robotic routing.	Efficiently detected faults like wires routed outside clips. Achieved 96% mean accuracy while improving scalability and reducing training time.
Guo et al. (2022)	**Tech:** RGBD sensors. **Method:** Segments multi-branch wires using Cartesian distance, color, and bending continuity, applying a Gaussian Mixture Model to complete missing segments.	Distinguished overlapping and forking aviation wires with >96% accuracy. Effectively filled point cloud gaps caused by occlusion.

annotated real-world data to function reliably [27]. This persistent reliance on task-specific training datasets highlights a critical limitation in traditional deep learning architectures. It strongly motivates the exploration of emerging "Video-to-3D" foundation models, which possess generalized spatial reasoning out-of-the-box and offer a promising alternative for 3D geometric reconstruction without the need for specialized training data.

4.1.3 Processing and Utilizing 3D Geometric data

The systematic literature review reveals significant application within 3D computer for wire harness manufacturing. Studies have successfully tackled a wide array of inspection and state estimation tasks, by translating raw, noisy data into structured mathematical models. For example, [28] demonstrated that point cloud data extracted from stereo cameras can be smoothly interpolated using B-Spline and Bezier curves, providing a continuous mathematical description of flexible cables suitable for robotic planning. Similarly, [29] showed that sparse point clouds could be reconstructed into highly accurate volumetric models by iteratively fitting collection of 3D sub-cylinders. Furthermore, because wire harness frequently overlap occlude one another during assembly, [30] provided that probabilistic models, specifically Gaussian Mixture Models, can effectively estimate wire states and complete missing geometric data. On the QC front, these geometric reconstructions have been leveraged to automate verification tasks. Such as [9] utilized high-resolution volumetric data to achieve precise measurements of terminal engagement depth, successfully identifying subtle misalignments and partial disconnections that electrical tests often miss. To verify minimum bend radii and ensure safe and physical clearances between cables and surroundings components [29] utilized cylinder models. Additionally, [31] successfully detected wires being placed outside of their designated clamp, using a shared-encoder neural network architecture.

4.1.4 Conclusion of the Literature Review

The existing academic literature demonstrates that 3D CV and ML offer solutions of the automated inspection of wire harnesses. Researchers have successfully segmented overlapping cables, filled occlusion gaps, and reconstructed wires into continuous splines and 3D sub-cylinders. These geometric models enable the automated verification of terminal engagement depths, minimum bend radii, safe physical clearances, and specific routing errors. Finally, mixing synthetic and real data during training has been proven effective for vision models within the wire harness domain.

Despite these impressive advancements in processing and utilizing 3D data, a unifying limitation is present: a reliance on specialized 3D acquisition hardware. Every primary study identified in this review depend on physical depth-sensing equipment to capture the initial spatial data. This included the use of structured-light 3D scanner, OCT systems, and RGB-D stereo cameras. While highly effective for generating accurate point clouds, these methods tether the automated inspection process to expensive, rigid, and highly calibrated industrial-grade sensory equipment.

Furthermore, while existing research has successfully addresses routing paths, microscopic defects, internal terminal engagement, there is an absence of studies examining the geometric relationship and classification between the mounting clips and the harness terminals themselves. Verifying that the correct terminals or clips are present and on the proper positions on the wire harness. As this would be a strong documentation tool within QC.

4.2 Model Selection - Results

4.2.1 Literature Retrieval and Selection

The initial systematic search targeting literature published in the year 2026 yielded a total of 144 review articles and surveys. Upon conducting a detailed screening against the predefined inclusion and exclusion criteria, it was determined that none of these retrieved papers sufficiently addressed the criteria.

Following the methodological pivot to primary search, the search parameters were expanded to only include literature published in 2025, which generated an initial pool of 358 papers. Applying the targeted, non-exhaustive review strategy, approximately 200 of these papers were systematically screened. This process yielded a relevant research paper that fulfilled all criteria. This paper provides comparative benchmarking data.

4.2.2 Comparative Benchmark Performance

The selected primary research paper [32] provides a comprehensive review of recent advances in feed-forward 3D reconstruction and view synthesis, categorizing SOTA models by their underlying architectures and evaluating them across various 3D perception tasks. While the paper covers multiple domains the data extracted for this study focuses strictly on the comparative benchmarks for video depth estimation. This focus directly aligns with the predefined inclusion criteria and the overarching goal of the project.

The video depth estimation benchmarking evaluates a range of recent feed-forward models across standard datasets. The data measures relative error to determine depth precision [32]. The data demonstrated a clear hierarchy in modeling capabilities. The pi3 model emerges as the best-performing architecture overall, consistently recording the lowest absolute relative errors and the highest inlier percentages across the tested datasets. VGGT demonstrates highly robust performance as a close second, matching or closely trailing pi3 across all key metrics. Because they outperformed on the comparative benchmarks for video-to-3D geometry, pi3 and VGGT were extracted as two SOTA models for this study.

4.2.3 Frontier Model Retrieval Results

To complement the established SOTA models and capture more recent models, the supplementary retrieval strategy was used targeting open-source repositories and pre-print archives.

Repository Mining and Community Metrics

An analysis of the Hugging Face repository was conducted utilizing the domain tags "Depth Estimation" and "Image-to-3D". Following the established methodology, models were ranked strictly by absolute download counts within their respective filters, one model per domain tag was chosen.

- **Depth-Estimation:** The model with the highest absolute downloads was excluded as it failed to satisfy the recency constraint. The subsequent highest-ranked model, Depth Anything V3 (DA3), was therefore selected.
- **Image-to-3D:** The selection process for this category required filtering out several top-ranking models. Candidates were sequentially skipped due to failing the recency rule, having restrictive licensing conditions, or having already been selected in the previous category. Map Anything emerged as the highest-ranked valid candidate meeting all criteria.

Benchmark Claim Verification and Scientific Inclination

To verify the scientific inclination of these candidates, DA3 and Map Anything were cross-referenced with their primary pre-print publications. The analysis of their self-reported performance claims revealed a highly interconnected benchmarking among the four prominent architectures (DA3, Map Anything, Pi3, and VGGT):

- The DA3 pre-print explicitly claims SOTA performance in zero-shot depth estimation, documenting significant, quantifiable improvements over its predecessor Depth Anything V2, Pi3, and VGGT [20].
- The Map Anything pre-print documents strong multi-view capabilities, claiming to match the metrics of Pi3 while outperforming VGGT [18].

Notably, three of these model papers (DA3, Map Anything, and Pi3) actively benchmark against the others or their immediate predecessors, such as Depth Anything V2, when establishing what they state is SOTA architectures. VGGT stands as the sole exception, note that it was the first model released among the group.

4.2.4 Final Model Selection Pool

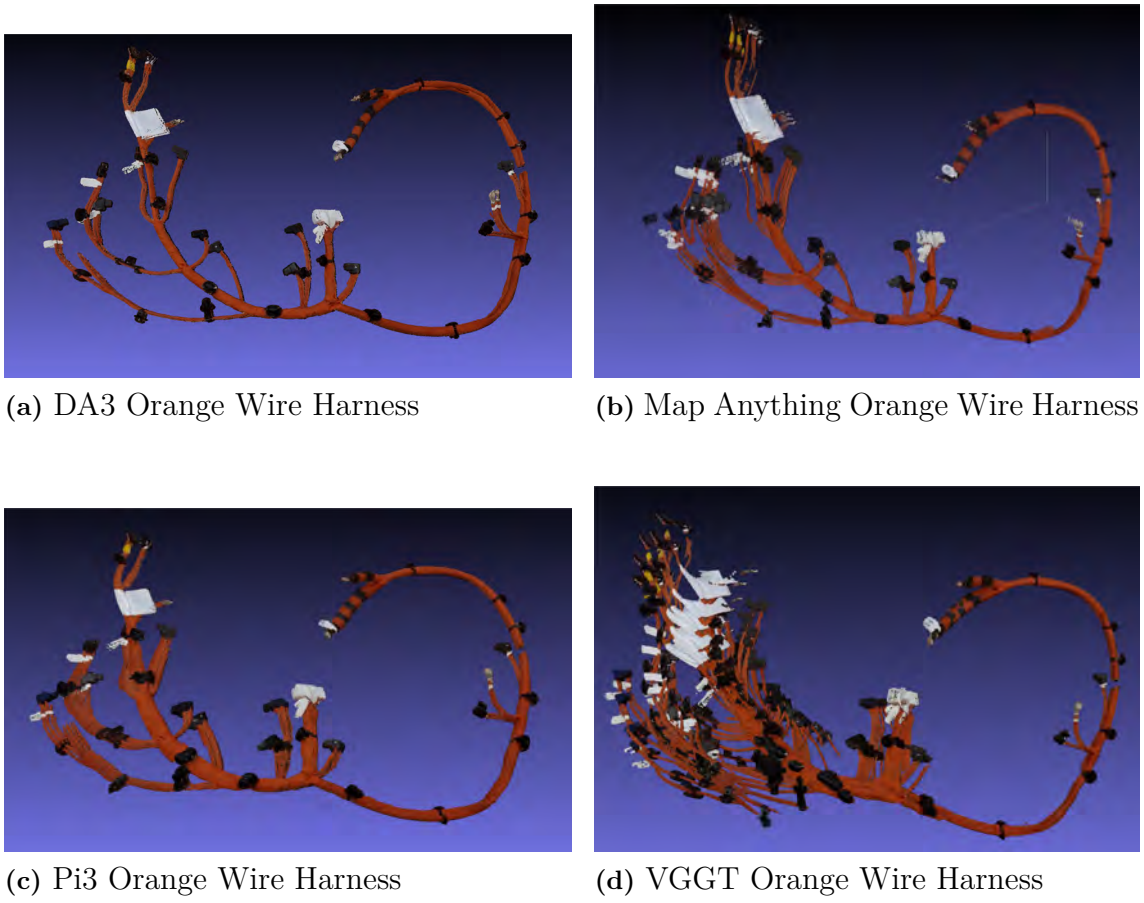
Following the conclusion of both primary literature benchmark extraction and the supplementary frontier retrieval strategy, the final pool of models selected for comparative evaluation in this study was established. The final selection consists of four architectures and the specific version used:

- **Depth Anything V3**: "depth-anything/DA3-GIANT-1.1"
- **Map Anything**: "facebook/map-anything"
- **Pi3-x**: "yyfz233/Pi3X"
- **VGGT**: "facebook/VGGT-1B"

4.3 Model Evaluation

4.3.1 Visual Reconstruction Results

This section presents the 3D point clouds generated from the selected foundation models. The following figures show the two wire harness configurations for this study, the orange wire harness and the black wire harness. While this section strives to showcase the most prominent characteristics of the models, a more exhaustive set of images can be found in the appendix A.1.



(a) DA3 Orange Wire Harness

(b) Map Anything Orange Wire Harness

(c) Pi3 Orange Wire Harness

(d) VGGT Orange Wire Harness

Figure 4.2: A top-down view of the orange wire harness on from all 4 models.

Figure 4.2 provides an overview of the global structure of the orange wire harness across the four models. This shows the models capabilities in maintaining creating continuity across the entire object. Figure 4.3 and 4.4 display Map Anything and DA3 for similar reason, with the black wire harness a top-down view and lower frontal view.

Figure 4.3: 3D reconstruction from Map Anything of the black wire harness

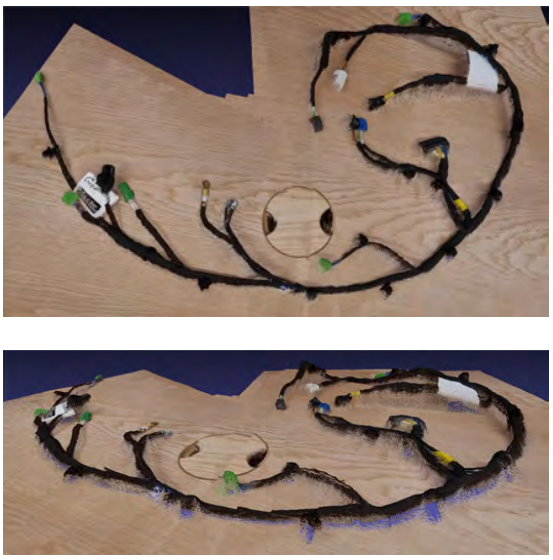
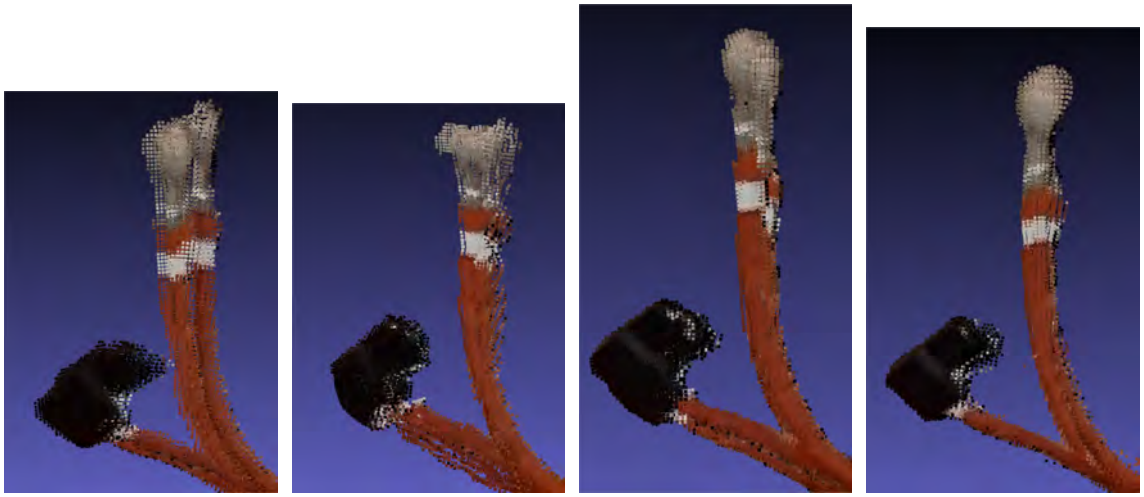


Figure 4.4: 3D reconstruction from DA3 of the black wire harness



(a) DA3 Orange **(b)** Map Anything **(c)** Pi3-x **(d)** VGGT

Figure 4.5: Zoomed in image of a reconstructed grounding ring

Beyond the global structure, the models have differences in resolving high-frequency details. Figure 4.5 demonstrates this by displaying the same grounding ring from all four architectures. Figure 4.6 and 4.7 highlight the main areas of instability of the reconstruction. The first illustrates the prominent failure mode of duplication (or "ghosting") prevalent at complex branching points, while the figure contrasts point density and fidelity of terminal housings between two models.

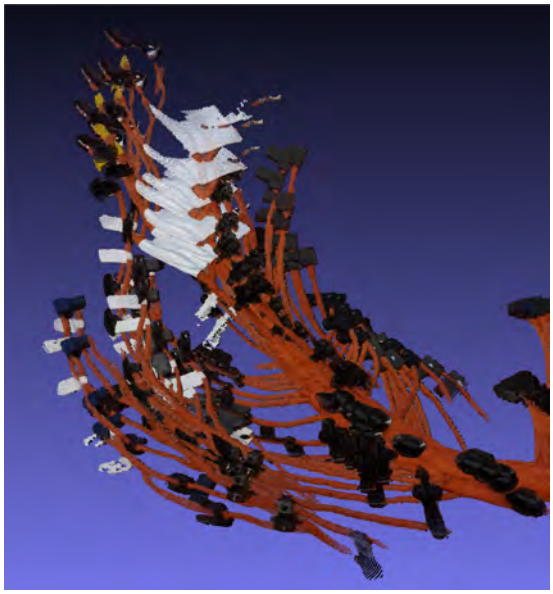


Figure 4.6: Reconstruction from VGGT displaying the duplication issue around heavy branching area.

(a) Pi3-x

(b) VGGT

Figure 4.7: Close up of two terminals and two clips reconstructed.

4.3.2 Qualitative Interview Results

To asses the practical capabilities and geometric fidelity of the generated 3D reconstructions, qualitative interviews were conducted with three industry professionals: two CV experts and one Quality Manager. The interviewees evaluated the generated point clouds based on their correspondence to reality, handling of details, overall spatial coherence and practical viability. The following section synthesizes their observations, highlight the core themes that emerged from the interviews.

Visual Fidelity

A critical metric for one of the experts was the models ability faithfully reproduce the physical reality of the wire harness, specifically focusing on visual accuracy, color representation, and the consistency of measurable 3D points. When evaluating geometric fidelity, another descried a hierarchy of detail. Capturing the basic cable structure server as the essential baseline. From there, finer features include accurately rendering electrical contacts, ring terminals, and mounting clips.

In terms of visual fidelity and detailed geometry, DA3 emerged as the top performer and most consistent, with VGGT also being capable of producing superior close-up reconstructions, though it proved to be significantly less consistent overall. Both models demonstrated superior capability in capturing small details with greater accuracy than the other architectures. Conversely, Map Anything was ranked the lowers in terms of fidelity, struggling to retain details and completely failing to

reconstruct connectors in certain views.

Global Coherence and Topology

Spatial coherence is defined as the models ability to logically stitch the topology together without placing objects incorrectly within the global 3D space.

For coherence, the CV experts noted that DA3 provided the most robust and consistent representation of the overall wire harness topology, surpassing the other models. Pi3-x was also identified as a reliable architecture for coherence, whereas specifically VGGT but also Map Anything struggled with coherence.

Common Failure Modes

Throughout the evaluation, the experts highlighted to primary failure modes across the models:

1. **Duplication at Branch Splits:** A prevalent issue was duplication or "ghosting" of wires. The interviewees observed that this occurred most frequently at areas where the wire harness branched or split into multiple paths. This failure is theorized from one of the experts to stem from the models internal feature matching struggling to differentiate between visually similar, repeating cables routed in close proximity. When the model fails to separate these features, it computes subpar camera placements, resulting in duplicated or misaligned geometry.
2. **Loss of Geometric Definition and Micro-feature Loss:** While overall visual fidelity exists on a spectrum, the failure to capture details becomes a critical barrier for inspection tasks. Rather than maintaining sharp structural boundaries, the reconstructions have a loss of geometric precision. This lack of definition makes it difficult to differentiate between similarity shaped and colored components, such as distinguishing two types of clips.

Feasibility for Automated Assembly Verification

A primary objective of the qualitative evaluation was to determine if this generated 3D data could be reliably utilized for automated assembly verification tasks.

Feasibility for Micro-Level Tasks: All three experts agreed that none of the current foundational models are ready for highly granular verification tasks. The models lacks a consistent geometric precision required to, for example count individual wires inserted into a specific terminal housing, or to verify if a ring terminal is correctly seated. Current noise and unreliability of fidelity makes this automated micro-level tasks unfeasible.

Feasibility for Macro-Level Classification: Despite micro-level limitations, the interviewees confirmed that the 3D data holds potential for macro-level classification and verification. The geometry and The geometry and topological routing are cohesive enough to build automated systems capable of distinguished between distinct

components. An automated system could realistically classify a clip versus a terminal housing by analyzing both its reconstructed geometric shape and its topological context, such as recognizing that clips are typically situated along the wire, whereas terminals cap the ends of cables.

Intermediate Use-Cases: To bridge the gap between current technical capabilities and factory implementation, the QC manager suggested adapting the verification workflow itself. Rather than attempting to scan and verify a complex, fully assembled harness all at once. The system could use layered assembly scanning, verifying the harness in stages e.g., scanning first when only terminal housings are attached, then after clips are added. Furthermore, to address current technical limitations, the CV experts stated that hybrid approaches—such as combining the dense prediction of these models with traditional feature matchers (e.g., COLMAP) to correct camera poses, which might be required to achieve fully autonomous verification.

Advantages Over 2D Vision

The experts highlighted that this 3D representation offers concrete advantages over traditional 2D inspection. Unlike a static images a 3D model allows a human inspector to navigate the spatial data, essentially taking a "virtual camera" down to the table view or inspecting a contact housing from multiple angles.

Ultimately, the interviewees state that while hybrid approaches such as combining the dense prediction of these models with traditional feature matcher (e.g., COLMAP) to correct camera poses, might be required for fully autonomous verification of the current models.

4.3.3 Quantitative Evaluation

To quantitatively assess the practical use of the generated 3D reconstruction for automated assembly verification. The pipeline was evaluated on its ability to accurately segment and count terminal housings and clips. Table 4.2 present the True Positives (TP), False Positives (FP), Recall, Precision, and F1-Scores across the two wire harness and the 4 models. The Ground Truth (GT) for the orange harness consists of 19 clips and 15 terminals housings (including one white terminal). The black harness contains 13 clips and 12 terminals housings (note that two additional terminals without housing were excluded from the count).
The results show that component counting with the top-performing architectures segment and count the majority of components. On the larger orange harness, DA3 emerged as the most robust model, achieving an F1-score of 78.8% and 76.0% for the terminal housing and clips respectively, which is visualized in figure 4.8. While the figure 4.9 shows the Pi3-x reconstructed model being the best performer on the black smaller wire harness. The following figures display the 3D wire harness reconstructions with the pipeline's predicted clusters overlapped. Each visualization isolates a specific component class (either terminal housings or clips), distinct colors to differentiate individually classified instances.

Table 4.2: Model Performance Metrics by Harness Color

Model	Component	TP	FP	Recall	Precision	F1-Score
Orange Harness — GT: 15 Terminal Housings, 19 Clips						
DA3	Terminal Housings	13	5	86.7%	72.2%	**78.8%**
	Clips	19	12	**100.0%**	**61.3%**	**76.0%**
MAP	Terminal Housings	14	14	**93.3%**	50.0%	65.1%
	Clips	18	23	94.7%	43.9%	59.9%
PI3	Terminal Housings	12	8	80.0%	60.0%	68.6%
	Clips	18	12	94.7%	60.0%	73.5%
VGGT	Terminal Housings	10	2	66.7%	**83.3%**	74.1%
	Clips	11	7	57.9%	61.1%	59.5%
Black Harness — GT: 12 Terminal Housings, 13 Clips						
DA3	Terminal Housings	10	0	83.3%	**100.0%**	90.9%
	Clips	12	2	**92.3%**	**85.7%**	**88.9%**
MAP	Terminal Housings	11	0	**91.7%**	**100.0%**	**95.7%**
	Clips	12	4	**92.3%**	75.0%	82.8%
PI3	Terminal Housings	11	0	**91.7%**	**100.0%**	**95.7%**
	Clips	12	2	**92.3%**	**85.7%**	**88.9%**
VGGT	Terminal Housings	8	11	66.7%	42.1%	51.6%
	Clips	9	13	69.2%	40.9%	51.4%

(a) Classifying **clips**.

(b) Classifying **terminals housings.**

Figure 4.8: Clustering of components on the orange wire harness using DA3 reconstruction.

(a) Classifying **clips**.

(b) Classifying **terminals housings.**

Figure 4.9: Clustering of components on the black wire harness using Pi3-x reconstruction.

Despite the high recall rates across most models, the main issue for the pipeline is the high number of False Positives, as all models except DA3 and Pi3-x as had a precision score of below 50% on one of the wire harnesses. Meaning the pipeline frequently count more components than are actually present.

Figure 4.10 shows the examples with many False Positive often by the MAP and VGGT architectures. These primary occur because the models struggles to coherently place components in a single, stable location within the 3D space. This instability leads to the pipeline projecting and clustering multiple copies hence multiple DBSCAN classifications of the exact same physical component.

(a) Classifying **terminal housings** with a Map Anything reconstruction.

(b) Classifying **terminal housings** with a VGGT reconstruction.

Figure 4.10: Clustering of components on reconstructed wire harnesses.

5

Discussion

5.1 Contextualizing the Study within Existing Literature

While existing studies on automated wire harness inspections successfully utilize 3D data for verifications, they are constrained by a strict reliance on specialized hardware (such as 3D scanners) or a dependence on large datasets of annotated training data. Furthermore, current research is very focused on robotic implementation rather than the quality control aspect of the wire harness assembly process.

This reveals a research gap for verification systems that do not depend on these hardware and data bottlenecks. Consequently, this thesis explores the viability of emerging "Video-to-3D" foundation models. Because these architectures possess generalized, out-of-the-box spital reasoning, they offer a promising alternative for 3D reconstruction with low or no task-specific training. By utilizing flexible, consumer-grade 2D video, this study demonstrates how 3D quality control can potentially decouple from expensive industrial sensors and eliminate reliance on annotated datasets.

5.2 Model Selection Method and Results

Identifying state-of-art architectures for video-based 3D reconstruction presents a clear challenge. The field of spatial foundation models moves faster than the traditional academic publishing cycle. Relying on systematic reviews especially peer-reviewed risks missing current SOTA, as these reviews take time to publish and are not conducted often this specific metic.

To achieve this, the primary literature search was supplemented by retrieving models from the largest open-source registry [24]. Using download metrics is not a flawless method for determining true performance, as a model's popularity can be influenced by external factors like institutional backing or marketing. Therefore, this approach should not be viewed as a definitive method of finding SOTA models.

However, within the limited scope of this study, community traction serves as a practical proxy. Broad adaptation correlates with functional reliability [24] and available documentation. This practical approach identified Depth Anything V3,

Map Anything, Pi3 and VGGT. Cross-referencing these models with their respective publications showed a interconnected network of benchmarking. The creators of these architectures frequently test against one another to establish SOTA claims, confirming that the selection found models in active research.

5.3 Experimental Study Method

To ensure an equal evaluation across the different architectures, the implementation used a standardized pipeline built around the Open3D library. The pipeline was purposely not built to accommodate each models individually, as tailoring the workflow for each model would introduce bias into the comparison. However, the downside fo this standardization is that the models might not reach their full potential. Because the pipeline treats all inputs and outputs to and from the model the same. It cannot adapt to unique architectural strengths of a given model. Additionally, model inference time and compute was not a focus of this study the practical limits of these required a restriction on the input data. This led to 30 images being processed for each 3D reconstruction.

The downstream tools used in this framework, specifically SAM3 and DBSCAN, played a major role in the quantitative performance scores. It is important to clarify how the pipeline divides the workload. The classification between terminal housings and clips was done by SAM3 which did this with the original 2D images. However, the counting of these components relied on the generated 3D data with the use of DBSCAN. This shows that while the 3D reconstruction models did not classify the objects, their spatial accuracy was necessary to locate and count them.

A clear limitation of this experimental setup is the small dataset, which only included two distinct wire harnesses. Because of this limited sample size, the results are hard to broadly generalize to all wire harness configurations found in the industry. Due to this constraint, this study should be viewed as a proof of concept.

5.4 Nuances in Model Performance and Implications

When examining the 3D reconstructions, there is a clear overall winner among the models Depth Anything V3. This superiority is most present when comparing the architectures ability to handle spatial coherence. For instance, in Figure 4.2, DA3 is the only model that successfully avoids major duplication or "ghosting" on the heavily branched left side of the wire harness, where as the others do not. Furthermore, DA3 demonstrated strong visual fidelity when rendering close-up views of most components.

However, despite this no model preformed the best across the whole wire harness. DA3, for example did not perform the best in all scenarios. On the rightmost wire of

the orange wire harness (as seen in figure 4.2), DA3 generated a significant duplication of the wire which was not seem by the other models. Similarly, when examining the close-up rending of the grounding ring in Figure 4.5, VGGT arguably produced the sharpest visual fidelity, and Pi3-x also outperformed DA3 in capturing these specific components. This highlights a primary trade-off seen between the models: visual fidelity versus coherence. VGGT illustrates this contrast well, while it produced some of the sharpest close-up details Figure 4.5, it simultaneously suffered from the most extreme cases of duplication seen in figure 4.6, also leading to subpar close ups in some instances such as Figure 4.7b. Pi3-X offered a middle ground, providing reliable global coherence that was second only to DA3, while maintaining acceptable visual fidelity.

Conversely, there was no clear "worst" model. While VGGT struggled with severe ghosting despite retaining detail, Map Anything showed the weakest close-up fidelity of all models, often struggling to generate geometric shapes of smaller details such as components, while demonstrating poor overall coherence. Ultimately, these results highlight that the models handle 3D space differently, and a single model's strengths can shift depending on the specific area of the object.

The counting pipeline offers an objective measure of how well these models organize 3D space, as coherent geometric reconstructions theoretically provide clearer point cloud groups for the DBSCAN algorithm to cluster. However, the pipeline struggled more with the orange wire harness, generating a high number of false positives across the models. This could stem from the more frequently branching of wires, but could also originate from SAM3's reliance on visual contrast during the initial 2D classification. Unlike the black harness, which has multi-colored terminals and all black clips, while the orange wire harness features all black clips and terminal housings (except for one white housing).

The qualitative evaluation reveals a lack of fine geometric detail in the 3D reconstructions. Which means these models are currently incapable of replacing human inspectors for visual quality control. Especially tasks requiring high precision, such as verifying whether individual wires are correctly inserted or ensuring that a ring terminal is correctly seated, remain unachievable for these architectures.

Despite limitations, the best-performing models demonstrate potential for broader verification tasks. As substantiated by the expert interviews, the generated 3D data retains sufficient geometric accuracy to show promise determining the presence, total count, and classification of distinct components like mounting clips and terminal housings. Furthermore, this 3D spatial representation could offer more immediate utility as a documentation tool. Allowing inspectors to navigate the data from multiple angles. While the segmentation and DBSCAN clustering method proposed in this thesis does not possess the consistency and accuracy required for production, it serves as an effective proof of concept. Ultimately, this is just one method of extracting verification data, and there remains other ways these 3D reconstructions could be utilized to perform these tasks in the future.

5.5 Answers to the Research Questions

Research Question 1

What are the existing SOTA "Video-to-3D" foundation models, and which architectures demonstrate the theoretical potential to handle wire harnesses?
Based on the selection methodology, the "Video-to-3D" foundation models identified for this study are Pi3-x, VGGT, Depth Anything V3 (DA3) and Map Anything. These were selected by combining recent academic benchmarks with community traction metrics. While relying on download counts as a proxy means these architectures may not represent the definitive SOTA, they capture capable, frontier models currently in active research. They demonstrate strong theoretical potential for wire harness inspection due to their generalized nature.

Research Question 2

To what extent can the identified reconstruction models restore the geometric and material characteristics of wire harnesses?
The results demonstrate that the identified foundation models successfully restore the overall topological structure and visual appearance of wire harnesses, but they lack the ability to generate fine geometric characteristics. While the performance of reconstruction varies across the architectures, Depth Anything V3 (DA3) emerged as the most robust overall, having high visual fidelity with the most consistent global coherence out of all the models. Despite this, all models exhibited two main limitations in their geometric reconstruction: duplication (or "ghosting") and a general loss of sharp definition of details. Consequently, while the models accurately capture material colors and general shapes well enough to be identifiable from a distance, this lack of reconstructed detail makes it difficult to reliably differentiate between structurally very similar small components upon closer inspection, such as distinguishing between two similar types of clips.

Research Question 3

Can the generated 3D spatial data be feasibly utilized to perform automated assembly verification tasks?
To address this question, the feasibility of utilizing generated 3D spatial data for automated assembly verification depends on the granularity of the task. As highlighted in the qualitative interviews, the current foundation models are not ready for verification tasks in need of high detail. Due to the lack of consistent geometric precision, it is unfeasible to use this data to perform things, such as determining if individual wires are correctly inserted or if a ring terminal is fully and correctly seated. However, the experts confirmed that the 3D data holds significant potential for less high-fidelity verification tasks. The overall geometry and topological routing are cohesive enough that verifying larger, distinct structural differences, such as classifying the presence of a clip versus a terminal housing is feasible. Ultimately,

while the generated 3D spatial data provides information needed for these broader automated checks, its practical success relies on the implementation of a competent automated downstream system capable of accurately interpreting its spatial context.

6
Conclusion

This thesis evaluated the feasibility of utilizing emerging "Video-to-3D" foundation models for the automated visual assembly verification of automotive wire harnesses. The aim was to explore a flexible alternative to specialized industrial scanners to address the industry's high reliance on manual labor within quality control. By evaluating four SOTA architectures (DA3, Map Anything, Pi3-x, and VGGT) using standard smartphone video, this research demonstrated that these models can reconstruct the global topological structure of wire harnesses without task-specific training. DA3 (DA3) proved the most robust, offering the best balance of global coherence and visual fidelity, whereas Map Anything lacked close-up fidelity and VGGT suffered from severe geometric duplication.

Prevalent failure modes included wire duplication at frequently branching sections and a loss of fine-feature definition. While an automated component counting method showed promising results, it suffered from high false-positive rates. Consequently, while the models capture general shapes and colors well, they currently lack the geometric precision required for highly detailed tasks, like verifying individual wire insertions or ring terminal seating. Despite these limitations, the generated 3D data holds significant potential for broader verification tasks. The spatial coherence is sufficient to classify and verify larger, distinct components, such as distinguishing mounting clips from terminal housings. This presents a path toward automated quality control that bypasses the need for massive annotated datasets and specialized equipment.

Bibliography

[1] O. Salunkhe, W. Quadrini, H. Wang, J. Stahre, D. Romero, L. Fumagalli, and D. Lämkull, "Review of current status and future directions for collaborative and semi-automated automotive wire harnesses assembly," *Procedia CIRP*, vol. 120, pp. 696–701, 2023.

[2] B. L. Žagar, A. Caporali, A. Szymko, P. Kicki, K. Walas, G. Palli, and A. C. Knoll, "Copy and paste augmentation for deformable wiring harness bags segmentation," in *2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pp. 721–726, IEEE, 2023.

[3] H. Wang, O. Salunkhe, W. Quadrini, D. Lämkull, F. Ore, M. Despeisse, L. Fumagalli, J. Stahre, and B. Johansson, "A systematic literature review of computer vision applications in robotized wire harness assembly," *Advanced Engineering Informatics*, vol. 62, p. 102596, 2024.

[4] F. d. O. Marinho, T. T. d. S. Zomer, L. F. C. d. S. Durão, P. A. Cauchick-Miguel, and E. Zancul, "Computer vision in quality 4.0: empirical insights from industrial demands," *Production*, vol. 36, 2026.

[5] S. Dübel, M. Röhlig, H. Schumann, and M. Trapp, "2d and 3d presentation of spatial data: A systematic review," in *2014 IEEE VIS International Workshop on 3DVis (3DVis)*, (Paris, France), pp. 11–18, IEEE, Nov 2014.

[6] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotnỳ, "Vggt: Visual geometry grounded transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[7] J. Trommnau, J. Kühnle, J. Siegert, R. Inderka, and T. Bauernhansl, "Overview of the state of the art in the production process of automotive wire harnesses, current research and future trends," *Procedia CIRP*, vol. 81, pp. 387–392, 2019.

[8] P. M. P. Khouri, M. J. Walsh, C. Brandl, and M. Rybachuk, "Design and automation of electrical cable harnesses testing system," *Microelectronics Reliability*, vol. 120, p. 114097, 2021.

[9] H. Kim, J. Bok, S. Han, M. Jeon, and J. Kim, "Advancing automotive wire harness inspection via optical coherence tomography coupled with a robust defect detection algorithm," *Sensors and Actuators A: Physical*, vol. 398, p. 117319, 2026.

[10] G. E. Navas-Reascos, D. Romero, J. Stahre, and A. Caballero-Ruiz, "Wire harness assembly process supported by collaborative robots: Literature review and call for rd," *Robotics*, vol. 11, no. 3, p. 65, 2022.

[11] Z. Zhang, "A flexible new technique for camera calibration," Tech. Rep. MSR-TR-98-71, Microsoft Research, 1998.

[12] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer, 2nd ed., 2022.

[13] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd ed., 2004.

[14] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," *arXiv preprint arXiv:1801.09847*, 2018.

[15] W. Gendy and D. Patel, "Advancements in computer vision: A comprehensive survey of image processing and interdisciplinary applications," *Academic Journal of Science and Technology*, vol. 13, no. 2, pp. 28–34, 2024.

[16] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[17] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, pp. 226–231, 1996.

[18] N. Keetha, N. Müller, J. Schönberger, L. Porzi, Y. Zhang, T. Fischer, A. Knapitsch, D. Zauss, E. Weber, N. Antunes, *et al.*, "Mapanything: Universal feed-forward metric 3d reconstruction," *arXiv preprint arXiv:2509.13414*, 2025.

[19] Y. Wang, J. Zhou, H. Zhu, W. Chang, Y. Zhou, Z. Li, J. Chen, J. Pang, C. Shen, and T. He, "π^3: Permutation-equivariant visual geometry learning," *arXiv preprint arXiv:2507.13347*, 2025.

[20] H. Lin, S. Chen, J. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang, "Depth anything 3: Recovering the visual space from any views," *arXiv preprint arXiv:2511.10647*, 2025.

[21] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.

[22] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Tech. Rep. EBSE-2007-01, Keele University and Lincoln University, July 2007.

[23] A. Martín-Martín, M. Thelwall, E. Orduna-Malea, and E. Delgado López-Cózar, "Google scholar, microsoft academic, scopus, dimensions, web of science, and opencitations' coci: a multidisciplinary comparison of coverage via citations," *Scientometrics*, vol. 126, no. 1, pp. 871–906, 2021.

[24] J. Jones, W. Jiang, N. Synovic, G. K. Thiruvathukal, and J. C. Davis, "What do we know about hugging face? a systematic literature review and quantitative validation of qualitative claims," in *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2024.

[25] P. Kadasi, S. Reddy, S. V. Chaturvedula, R. Sen, A. Saha, S. Sikdar, S. Sarkar, S. Mittal, R. Jindal, and M. Singh, "Model hubs and beyond: Analyzing model popularity, performance, and documentation," *arXiv preprint arXiv:2503.15222*, 2025.

[26] S. Döringer, "'the problem-centred expert interview'. combining qualitative interviewing approaches for investigating implicit expert knowledge," *Interna-*

tional Journal of Social Research Methodology, vol. 24, no. 3, pp. 265–278, 2021.

[27] H. G. Nguyen, R. Habiboglu, and J. Franke, "Enabling deep learning using synthetic data: A case study for the automotive wiring harness manufacturing," *Procedia CIRP*, vol. 107, pp. 1263–1268, 2022.

[28] T. Okuyama, P.-C. Chien, H. Tsukida, Y. Kato, and J. Ohya, "A two-stage approach for wire harness cable description using 3d point clouds for robotic manufacturing," in *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pp. 689–695, SCITEPRESS, 2025.

[29] H. Ben Abdallah, J.-J. Orteu, I. Jovančević, and B. Dolives, "Three-dimensional point cloud analysis for automatic inspection of complex aeronautical mechanical assemblies," *Journal of Electronic Imaging*, vol. 29, no. 4, p. 041012, 2020.

[30] J. Guo, J. Zhang, Y. Gai, D. Wu, and K. Chen, "Visual recognition method for deformable wires in aircrafts assembly based on sequential segmentation and probabilisitic estimation," in *2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC)*, IEEE, 2022.

[31] K. Galassi, A. Caporali, G. Laudante, and G. Palli, "Scalable shared encoding architecture for learning-based error detection in robotic wiring harness assembly," in *2024 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, IEEE, 2024.

[32] J. Zhang, Y. Li, A. Chen, M. Xu, K. Liu, J. Wang, X.-X. Long, H. Liang, Z. Xu, H. Su, C. Theobalt, C. Rupprecht, *et al.*, "Advances in feed-forward 3d reconstruction and view synthesis: A survey," *arXiv preprint arXiv:2507.14501*, 2025.

A

Appendix 1

A.1 Figures of the 3D Reconstruction

This section provides additional images of the 3D reconstructed wire harnesses. Specifically, it presents 3D segmented views of the four foundation models: DA3, Map Anything, Pi3, and VGGT. For each model, both the orange and black wire harness configurations are illustrated from three perspectives for a better overview.

A.1.1 DA3

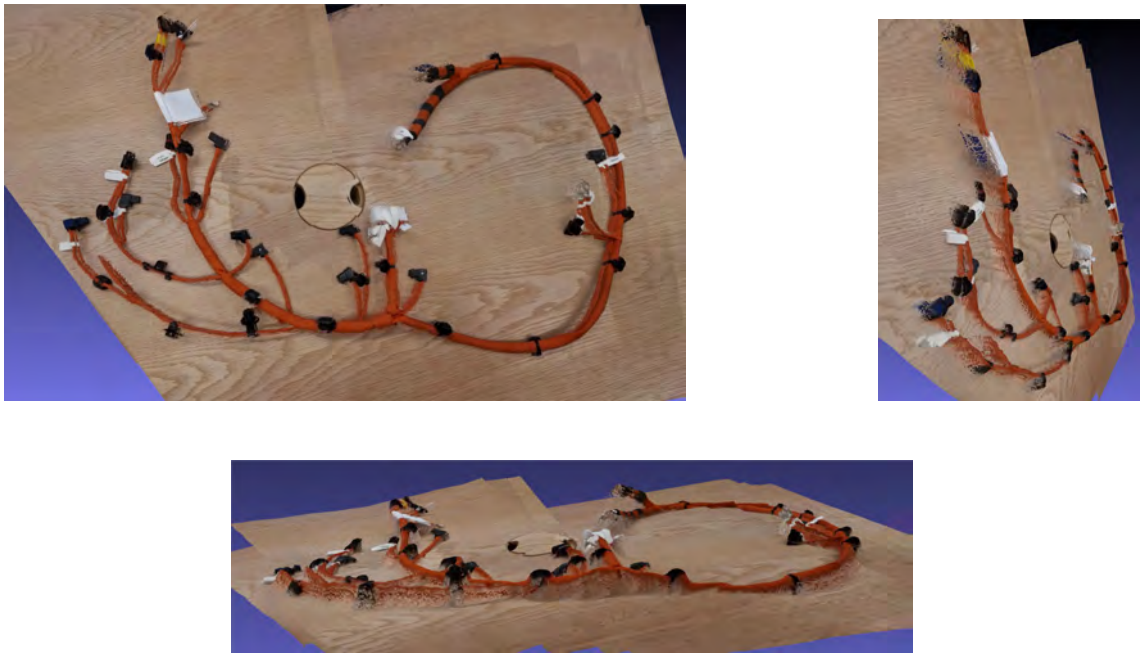


Figure A.1: 3D segmented views of the orange wire harness for DA3

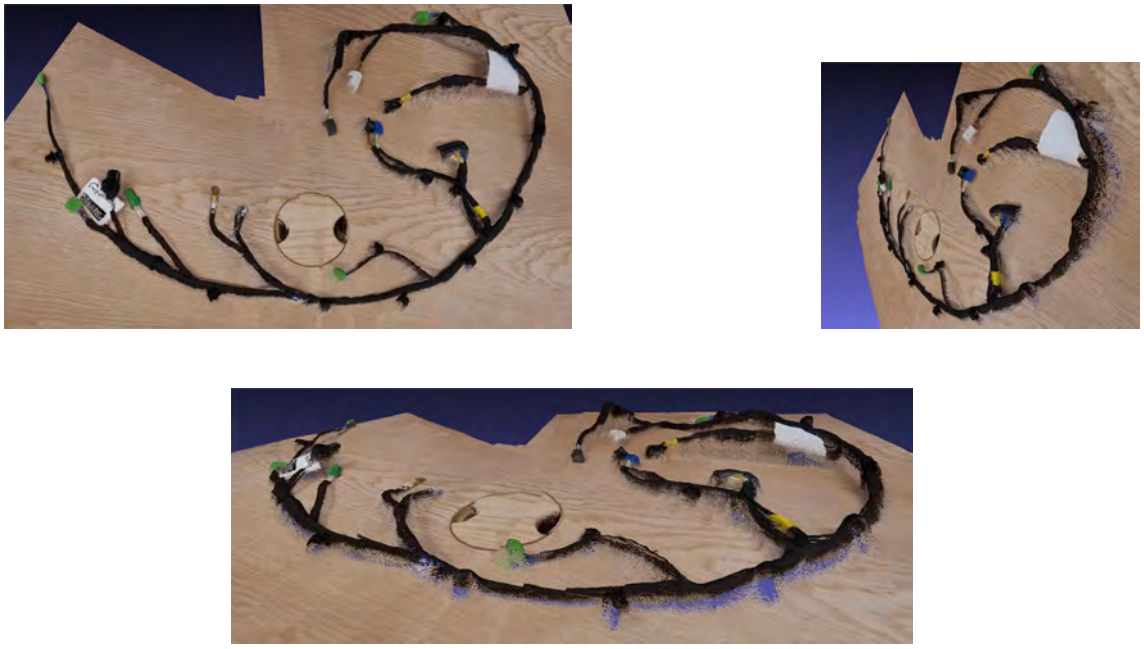


Figure A.2: 3D segmented views of the black wire harness for DA3

A.1.2 Map Anything

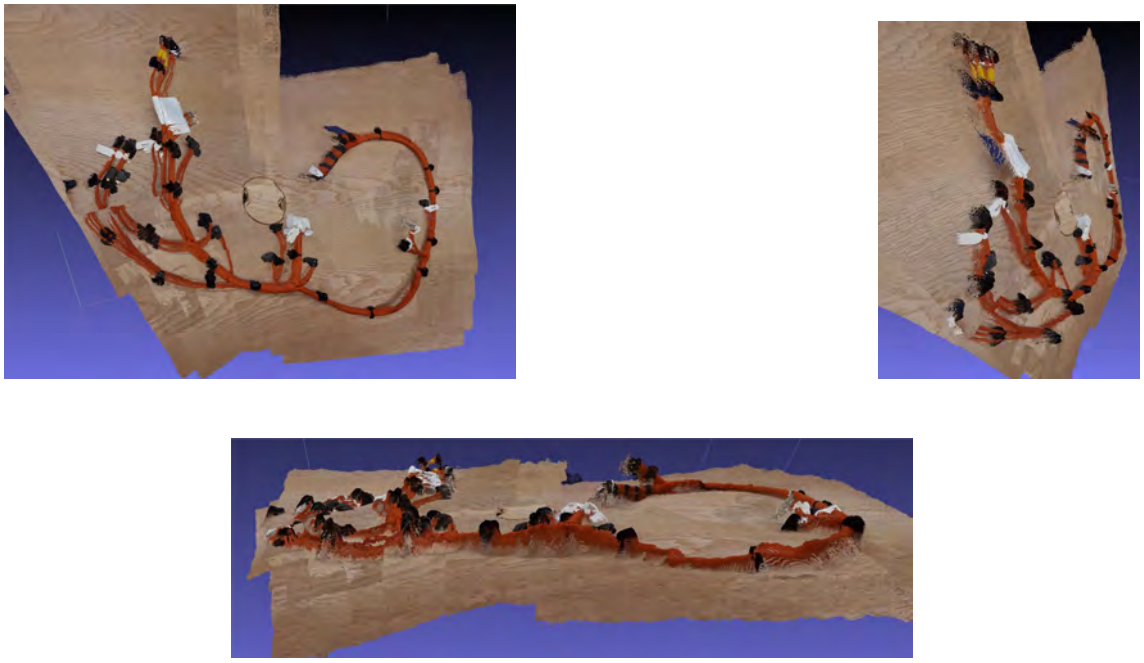


Figure A.3: 3D segmented views of the orange wire harness for Map Anything

Figure A.4: 3D segmented views of the black wire harness for Map Anything

A.1.3 Pi3

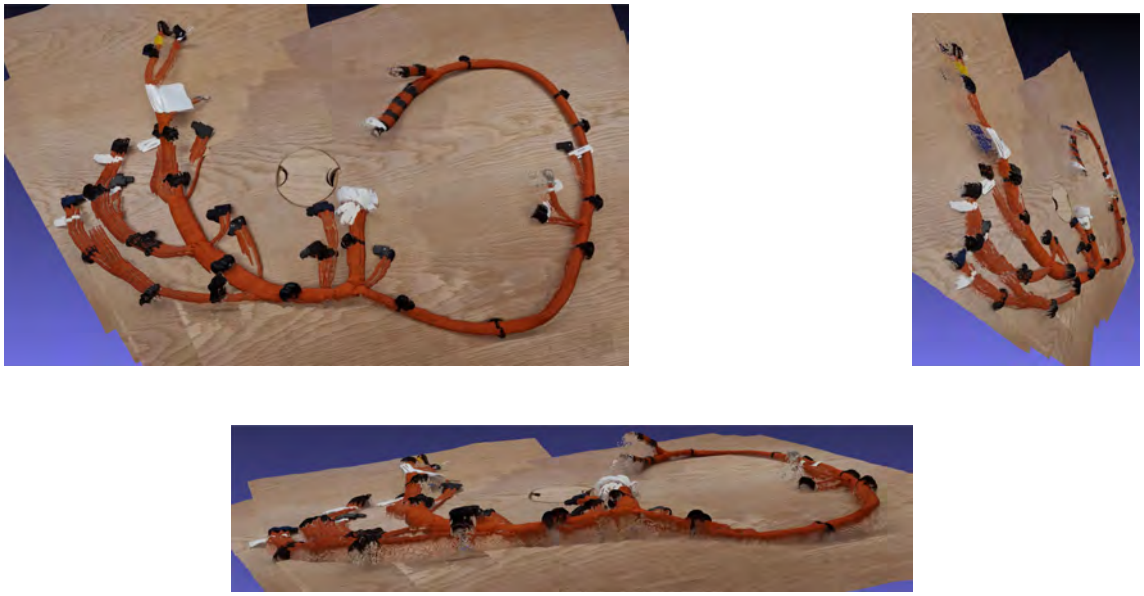


Figure A.5: 3D segmented views of the orange wire harness for Pi3

Figure A.6: 3D segmented views of the black wire harness for Pi3

A.1.4 VGGT

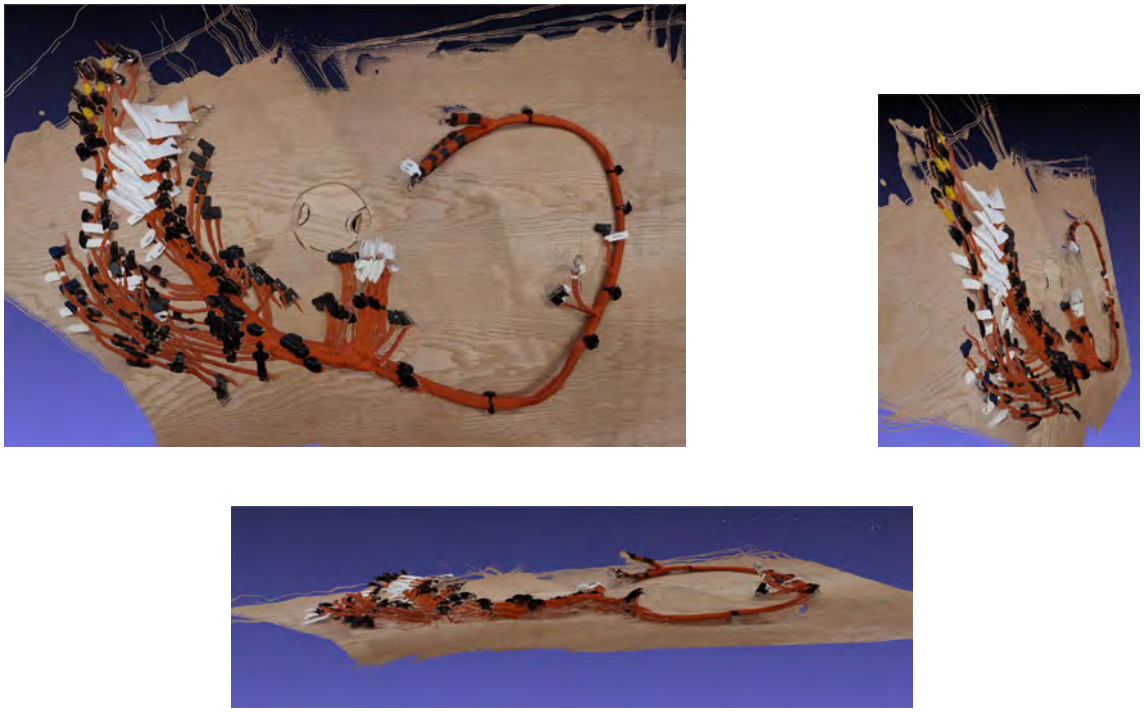


Figure A.7: 3D segmented views of the orange wire harness for VGGT

Figure A.8: 3D segmented views of the black wire harness for VGGT

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY