Generative AI-Powered 3D-Content Creation for Military Training

Eduardo Barrera
Charles River Analytics, Inc.
Cambridge, Massachusetts
ebarrera@cra.com

Deepak Haste, Michael Renda, Sudipto Ghoshal
Qualtech Systems, Inc.
Rocky Hill, Connecticut
deepak@teamqsi.com, m.renda@teamqsi.com, sudipto@teamqsi.com

Jason H. Wong
Naval Information Warfare Center Pacific
San Diego, California
jason.h.wong.civ@us.navy.mil

ABSTRACT

The U.S. Marine Corps (USMC) has taken the initiative of introducing interactive learning experiences at its schoolhouses as a cost-effective and timesaving means to augment classroom instructions and physical equipment-training with immersive maintenance training and safety training in a simulated environment. However, the techniques for creating 3D models for immersive environments use Computer Aided Design (CAD) and graphics software, which demand significant manual effort, software skills, time, and financial investments. The USMC has recognized the need to rapidly build a repository of ready, reusable, and configurable 3D models of their assets in a scalable manner. Recent advances in generative artificial intelligence (AI) can fill this need by rapidly generating approximate but realistic 3D models from 2D pictures of equipment found in USMC training guides such as presentations and student handouts, thereby reducing program costs and accelerating student training.

In this paper, the research team presents a scalable and automated content-generation process that uses an ensemble of vision-based generative AI techniques, and a diverse web-sourced 3D-object dataset, to convert 2D images into 3D models with appropriate tradeoffs between desired quality and computational complexity. The research team will extend an existing foundational 2D-to-3D conversion-model trained with large and diverse web-scale data for "few-shot" transfer learning with domain-specific data. The 3D-content generation process will use open-source software and incorporate intuitive user-interfaces to minimize the need to learn machine learning (ML) or graphics programming. The resulting 3D objects can be imported into reusable libraries for use across various schoolhouse applications requiring immersive training content.

Furthermore, the team presents the results of performance experiments that convert images from a USMC schoolhouse course using techniques with varying degrees of complexity, and benchmark various vision-based AI/ML techniques for object fidelity and speed of conversion. The paper concludes with best practices and lessons learned from these content-conversion experiments.

ABOUT THE AUTHORS

Mr. Eduardo Barrera is a Software Engineer at Charles River Analytics where he leads the development and research efforts for various programs. He specializes in computer vision technologies and generative AI-driven solutions, and his work spans multiple product areas, including object detection and segmentation, 3D scene reconstruction, augmented reality / virtual reality (AR/VR) integration, and on-demand signal denoising. Mr. Barrera holds two B.S. degrees from Tufts University—in Mathematics and in Engineering Physics—and has previously conducted research in high-energy particle physics, quantum algorithm-accelerated computer vision, and computational quantum physics.

Mr. Deepak Haste is a Senior Director of Engineering at Qualtech Systems, Inc. (QSI), with a focus on customer-driven enhancements, productization, and commercialization of QSI's products through several key Small Business Innovations Research (SBIR) projects with NASA and DoD. His recent productization efforts include feature extensibility through plugin frameworks, prognostic capabilities, and embedding and interfacing of diagnostic software within onboard platforms. He is currently leading efforts to add AR/VR and Maintenance Training capabilities into QSI's product suite. Mr. Haste holds an M.S. degree in Electrical Engineering from Clemson

University and a Bachelor of Technology (B. Tech.) in Electrical Engineering from the Indian Institute of Technology (IIT), Bombay, India.

Mr. Michael Renda is a Software Engineer at QSI and is currently involved in enhancing its product suite with modern AR/VR capabilities. This work has manifested in QSI's AR/VR Designer, a tool that enables military instructors to author AR/VR scenes for training students in an immersive context. In addition, he has extended photogrammetry tools to develop a 3D-scanning app for iOS, allowing instructors to capture their machinery into 3D models and feature them in educational content created via the AR/VR Designer. Mr. Renda holds a B.S. in Computer Science and a B.S. degree in Economics from Wesleyan University.

Dr. Sudipto Ghoshal is a Vice President of Engineering at QSI, with experience in system diagnostics and prognostics, as well as design-time assessment and mitigation of operational risk and improvement of system availability. His research also involves the development of a failure-space modeling methodology leveraging general purpose system design languages such as SysML and adapting it for QSI's products. He has been a committee member of the IEEE SCC20 Diagnostics and Maintenance subcommittee since 1991. Dr. Ghoshal holds an MBA from the Kelley School of Business, Bloomington, Ph.D. and M.S. degrees in Biomedical Engineering from the University of Connecticut, and a B. Tech. in Electrical Engineering from IIT, Kharagpur, India.

Dr. Jason H. Wong is a Cognitive Scientist with the Naval Information Warfare Center Pacific, where he supports the Office of Naval Research in improving Warfighter decision-making and human-technology teams. His government career has included measuring cognitive performance for training and simulation at the Naval Undersea Warfare Center and fostering international scientific collaboration at the Office of Naval Research Global in Tokyo, Japan. Dr. Wong holds a Ph.D. in Human Factors and Applied Cognition from George Mason University and a B.S. in Psychology from the University of Illinois at Urbana-Champaign.

Generative AI-Powered 3D-Content Creation for Military Training

Eduardo Barrera
Charles River Analytics, Inc.
Cambridge, Massachusetts
ebarrera@cra.com

Deepak Haste, Michael Renda, Sudipto Ghoshal
Qualtech Systems, Inc.
Rocky Hill, Connecticut
deepak@teamqsi.com,
m.renda@teamqsi.com,
sudipto@teamqsi.com

Jason H. Wong
Naval Information Warfare Center Pacific
San Diego, California
jason.h.wong.civ@us.navy.mil

INTRODUCTION

The US Marine Corps (USMC) considers Augmented Reality (AR) and Virtual Reality (VR) to be effective and efficient technologies to improve the readiness of warfighting equipment via accelerated maintenance training in simulated environments, thereby improving safety protocols and enhancing troubleshooting procedures. In order to enhance the skills and abilities of technicians responsible for maintaining and sustaining military equipment, the USMC has taken the initiative to integrate immersive technologies into its schoolhouses. However, it is missing a critical building block of AR/VR-driven immersive environments, namely, access to reusable 3D content. The USMC has limited resources to generate 3D models internally using CAD and graphics software. Therefore, one approach would be to leverage existing training materials as sources of 3D assets. These materials, typically comprising of presentations, handouts, and manuals, often contain one-off images of USMC equipment. The team aims to employ state-of-the-art generative AI techniques to fill this niche use-case by developing a content transformation pipeline that can generate 3D models from these resources to facilitate immersive classrooms (Figure 1).

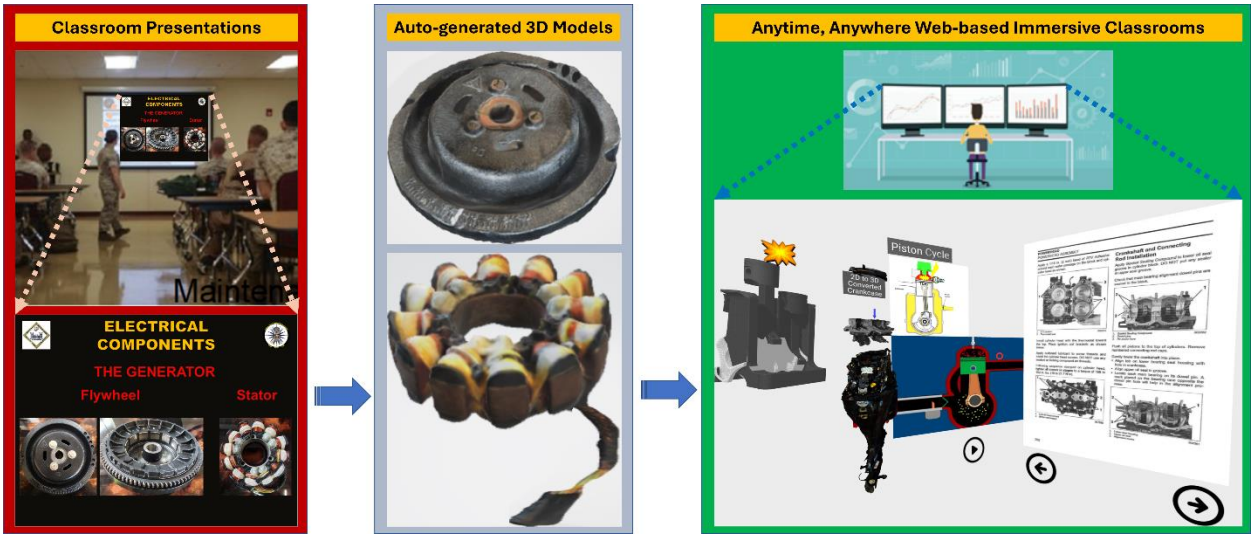

Figure 1: Transformation from Traditional Classroom Learning Model to Information Age Learning Model.

The research team's vision for modernizing traditional classroom learning entails extracting common pictures and graphics from course presentations and student handouts, and converting them into 3D models that can be used to create web-based immersive training material such as animations and practical applications.

BACKGROUND

Current State of Maintenance Training

The USMC has expressed the need for an Enterprise Ground Maintenance Training Simulator (EGMTS, 2023) to

modernize the state of ground maintenance training devices and provide maintainers with access to modern virtual training capabilities that provide up-to-date "anytime, anywhere" knowledge and expertise in remote environments outside the classrooms. Key enablers to this vision are a content ingestion and authoring system that allows for legacy content, such as 2D pictures and technical manuals, to be converted into 3D content for use in immersive training systems. The main challenges to fulfilling this vision is the labor-intensive nature of the process, requiring highly skilled graphic artists and many person-hours of effort to transform legacy data into immersive training content. Another challenge in realizing the vision of EGMTS is that the hardware-based high-fidelity trainers require specialized, and hence expensive, hardware to provide immersive training. Consequently, the USMC lacks the ability to provide access to immersive training via government-owned and personally-owned electronic devices.

Thus, an automated content transformation process with easily updateable 3D content that does not require expertise in 3D modeling will realize the vision of an enterprise-grade training platform. Having a web-based training platform will also make it accessible via remote-login capability and will allow maintainers to access immersive content from their work or personal computers. Furthermore, it will enable instructors to track the progress of the students undergoing training without the need to be physically present alongside the trainee. Student performance can also be tracked inside a learning management system (LMS) such as Moodle. The team has conducted prior research on the web-based pipeline to author immersive scenes, and a Moodle-based hosting platform for tracking student performance, fulfilling the goal of Immersive Information Age Training (Haste, 2022). In order to source the immersive content generation described in this paper, an equally seamless and streamlined pipeline of 3D model creation is needed. This research aims to fill that gap of rapid production of 3D models of various training equipment. The resulting pipeline will be used by various schoolhouses such as the Marine Corps Engineer School (MCES) to generate 3D models at scale for creation of varied immersive content in their curricula.

EGMTS defines a benchmark criterion where instructors are able to create a 3D model from legacy material and generate immersive content data within 4 hours. Furthermore, the Program Office Training Systems (PM TRASYS) has identified the main challenge in creating immersive content to be a lack of readily available 3D models in the schoolhouses. In order to achieve such a rapid turnaround in 3D model creation from common sources such as available 2D pictures in training presentations, an automated 3D generation process is needed at the schoolhouse location where the instructors can use the product instantly.

One of the major hurdles in the generation of 3D models for the USMC is that a significant portion of the source material is unique and domain-specific. Hence, most of the popular foundational generative AI models are not pre-trained for the maintenance training domain for an individual schoolhouse. For instance, MCES trains its technicians on Marine engines, Marine Corps Combat Service Support Schools (MCCSSS) trains on combat vehicles, and MCCES trains its students on combat radio equipment with secret ratings. To address this unique challenge, the 3D content creation pipeline developed through this research makes provisions for in-house domain-specific training of custom AI models for generation of 3D models conditioned on 2D images using a local schoolhouse server.

Another unique challenge in the creation of 3D models for the USMC is that the source material taught at schoolhouses is often classified. Since a common industry practice is to use extensive cloud computing platforms (InfoWorld, 2024) for running generative AI for model training and 3D object creation, this would violate USMC cybersecurity policies. The unique offering of this research is its ability to generate AI models on local server hardware, thus addressing the data protection issue as well as affording cost and time savings due to the reduced hardware requirements.

Generative AI Method: Zero-1-to-3

The research team explored various deep neural networks to convert one-shot 2D photos into 3D models. Previous methods such as neural radiance fields (NeRFs) (Mildenhall, 2020) and graph-based convolutional neural networks (GNNs) (Zhou, 2020) lacked critical features such as fidelity and realism, often producing rough, grayscale, or untextured models.

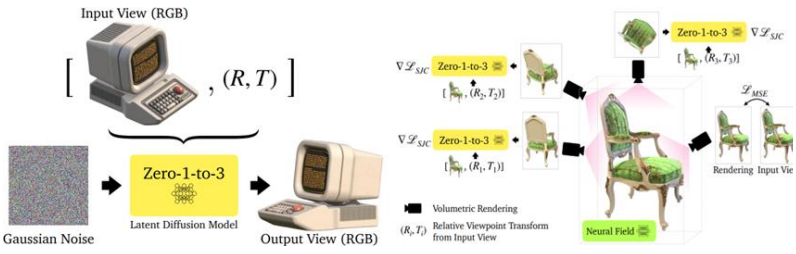

Figure 2: Viewpoint-conditioned image translation model using Conditional Latent Diffusion Architecture, then optimized 3D reconstruction via Zero-1-to-3 supervision (image from (Liu, 2023)).

Our choice, Zero-1-to-3, improves upon these by leveraging geometric priors from large-scale diffusion models to retain textures and enhance realism (Figure 2), leading to a more immersive learning experience for the USMC. Initially trained on the Objaverse dataset, which includes over 12 million detailed 3D models, Zero-1-to-3 benefits from rich geometry and fine details (Figure 3). However, the dataset's broad scope isn't tailored for specific applications such as USMC electromechanical objects.

Figure 3: The Objaverse dataset offers a collection of 12 million commonplace 3D objects across a range of domains (image from Liu (2023))

To address this, we integrated a human-in-the-loop system for domain-specific fine-tuning. Operators can curate a folder of relevant 3D objects to create a specialized training dataset, specifying types of components, detail levels, and contextual use within AR/VR training. This tailored approach enhances the model's efficiency and effectiveness for specialized applications, ensuring continuous adaptation and refinement as new requirements emerge, keeping Zero-1-to-3 at the forefront of 3D modeling technology for the USMC.

Industry Gaps and Optimizations: Tailoring Zero-1-to-3 to the USMC Use-Case

Our team made strides in tailoring the Zero-1-to-3 model for use in USMC schoolhouse environments, specifically in the areas of model sizing, efficiency, trainability, and dataset curation.

For example, to train the initial model, an end user would need eight NVIDIA A100 Tensor Core 80GB GPUs (A100, 2020). After customizing the model, it can be trained on a single A100 40GB GPU, requiring less than one-eight the compute power. As a result, this improvement reduces the approximate up-front cost to around $25K (for 1 A100 40GB and an associated server rack), compared to the original model, which required a capital investment of up to $1M+. Furthermore, the smallest version of the model was trained for 2,000 hours on the original hardware, and other versions with similar performance were trained for up to 6,000 hours (refer to section “BENCHMARK TESTING OF VARIOUS GENERATIVE AI MODELS”). By using these pre-trained weights, the team leveraged the highly general 3D semantic knowledge and geometric intuition already learned by the existing models, and only needed to fine-tune the model to perform well in the electro-mechanical domain, thereby drastically reducing the training resources needed while also gaining improved performance.

CONCEPT OF OPERATION (CONOPS)

To further customize the USMC use case, the team designed an automatic target-domain training dataset generator—the first of its kind for electromechanical objects and machinery. Since the Objaverse dataset consists of a wide range of common objects, a model trained on this dataset will perform well across generic domains, such as popular culture, cartoons, or related fields, but does not perform well on objects belonging to specialized domains. Our innovative custom dataset generator is extensible into new domains with user-defined input, enabling the training dataset to grow in secure environments after model delivery at USMC or other schoolhouses - a non-negotiable aspect in military applications. As such, this approach empowers in-house teams to curate sensitive datasets and models without needing to rely on external data sources.

Figure 4 shows the concept of operations (CONOPS) of the content generator. The CONOPS comprises four stages:

Custom Domain-Specific Dataset Generator

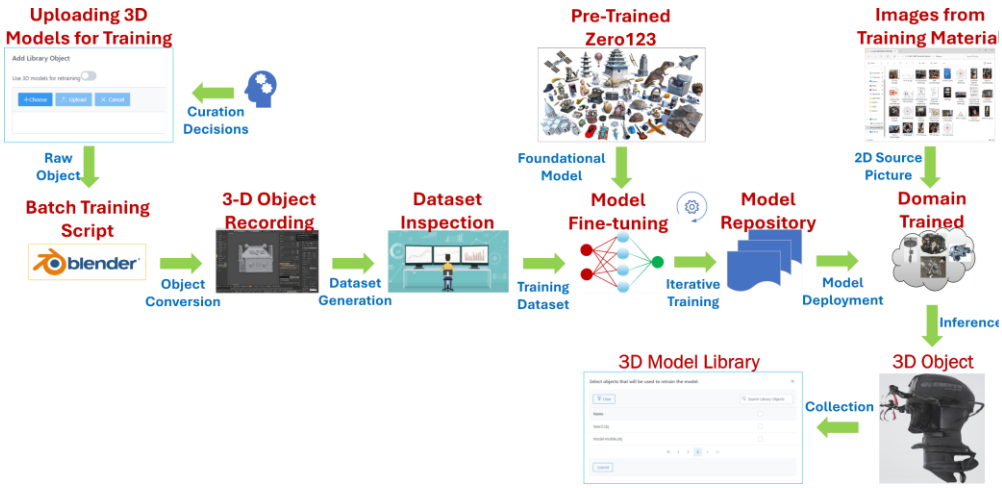

Figure 4: 3D Model Training Overview

The first step in the 3D content creation pipeline involves selecting a target domain for inference. In our case, the focus is on electro-mechanical objects that a typical student in a USMC schoolhouse might encounter in an AR/VR training simulation. Once the domain is selected, the Zero-1-to-3 model needs be fine-tuned to the USMC use case. The research team developed a streamlined custom dataset generator that converts user-defined and domain-relevant 3D objects into a suitable format for fine-tuning. This pipeline enables the pre-trained, non-domain model to be tailored to specific domain data with minimal incremental training, requiring only 5 to 10 objects. To achieve this, the schoolhouse course creators upload 3D models to the web-based Immersive Designer, where they have the option to mark them for training. The Immersive Designer then sends these 3D models to a curated folder containing target objects for the domain in Wavefront Object (.OBJ) file format (standard 3D file format that defines the geometry of 3D objects). Thereafter, a daemon, running a headless shell script, initiates a dataset generation process, ingesting the folder containing user-curated objects and converting these .OBJ files into a multi-view format, illuminating the object and capturing images along with their corresponding camera extrinsic matrices that the model expects. The script uses Blender's (Blender, 2022) open-source software development kit to apply transformations and various image capture techniques to the objects in the origin folder in an automated manner to generate the target domain training dataset. Once the dataset has been processed, the instructors can inspect the multi-view images to ensure that they align with their expectations.

Domain Model Training

After the dataset is ready, the user begins fine-tuning an existing Zero-1-to-3 benchmark using the training pipeline to tune the model to the domain-specific use-case. The research team fine-tuned three popular pre-trained model weights for Zero-1-to-3:

1. **Zero123-Base**: The weights in (Liu, 2023) trained on Objaverse 1.0. Fine-tuned to **Zero123-Base_USMC.**
2. **Zero123-XL**: The updated weights from (Liu, 2023) after the public dataset Objaverse received an update, increasing its size from 800,000 objects to 12 million objects. Fine-tuned to **Zero123-XL_USMC.**
3. **Zero123-Stable**: An updated weights file that was fine-tuned from Zero123-XL by the StabilityAI team and includes improved model conditioning hyperparameters. Fine-tuned to **Zero123-Stable_USMC.**

Using the custom dataset generated for the USMC use case, the team tailored the fine-tuning process to enhance the model's performance in generating 3D models of electro-mechanical objects. This fine-tuning process involves iterative adjustments to hyperparameters such as learning rate, batch size, and training steps, ensuring the model converges to its optimal performance in the USMC domain. By using publicly available pre-trained weights, the team saved time and resources, enabling us to evaluate the model's convergence and performance without requiring thousands of hours like in (Liu, 2023). This approach not only improves the model's performance in specialized applications but also enables continuous adaptation and refinement as new requirements emerge, ensuring that the fine-tuned Zero-1-to-3 model remains at the forefront of 3D modeling technology for the USMC.

3D Asset Generation Process

After deployment, the fine-tuned, domain-trained Zero-1-to-3 model uses its inferencing capabilities on the schoolhouse source material, such as 2D photos extracted from student handouts and presentations, to generate 3D assets with original image textures in a batch format. The 3D model generation process combines open-source 3D models, state-of-the-art generative neural networks, and commercial hardware to streamline the pipeline of immersive maintenance training experiences without the burden of manual 3D modeling on the instructors or reliance on outside contractors. There are several Zero-1-to-3 variants available, and in this paper, the team focused on three popular model checkpoints, namely, the original model in (Liu, 2023), the "XL" version trained on the updated Objaverse dataset, and a high-performance StabilityAI version. The 3D model creation pipeline supports legacy schoolhouse materials such as training manual graphics, technical diagrams, and pictures in formats such as Portable Network Graphics (PNG), Joint Photographic Experts Group (JPEG), and Scalable Vector Graphics (SVG). The fine-tuning process generated a series of models with varying degrees of performance on the validation dataset, and the team selected the model that best suited its needs for production.

Integration of 3D Models with the Immersive Content Creation Pipeline

The resulting 3D models are then automatically pushed to a reusable library where the objects are cataloged and made available to the Immersive Designer, allowing instructors to author AR/VR simulations using the WebXR framework (WebXR, 2022), providing a realistic and immersive AR/VR experience for the students.

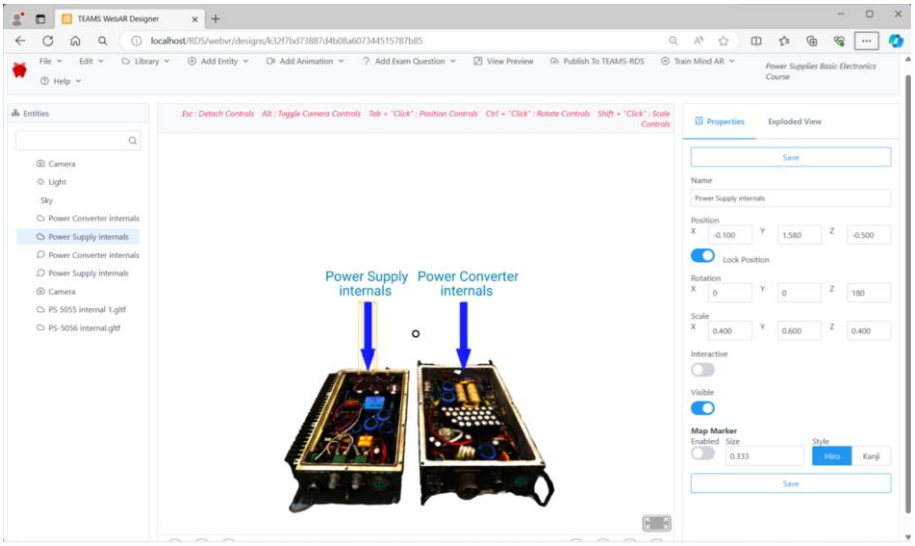

Figure 5: Web-based Immersive Designer

In a prior effort (Haste, 2022), the research team developed an immersive scene editor, based on the WebXR technology, that enables authoring of AR/VR scenes over the network without requiring 3D programming and graphical design skills, or any specialized software such as Unity (Unity, 2022), or hardware. The editor has an interface to pull pictures, libraries of 3D objects and other reusable content to sequence 3D animations. The immersive editor supports authoring of complex immersive scenes, component identification tests and practical applications (Figure 5) that can be hosted in an LMS such as Moodle (Moodle, 2022), which is the USMC's LMS of choice. The embedded AR/VR scenes serve as immersive training content for students.

The resulting end product is a reconfigurable "anytime anywhere" training platform for students and Warfighters. The research team integrated the 3D model generation pipeline with the AR/VR scene authoring pipeline to enable an end-to-end toolchain to generate immersive content from training material such as classroom presentations and handouts.

This approach ensures that the Zero-1-to-3 model is specifically tuned to meet the unique requirements of the USMC schoolhouse environments, enhancing the realism and relevance of the generated 3D models for maintenance training.

BENCHMARK TESTING OF VARIOUS GENERATIVE AI MODELS

The research team conducted benchmark testing on the three original Zero-1-to-3 models and their fine-tuned counterparts for the generative AI-based asset conversion utility described earlier (see section "CONCEPT OF

OPERATION (CONOPS)").

Experimental Setup

The benchmarking exercise for the 3D model generation and immersive content creation process was conducted using the Program Of Instruction (POI) material, supplied by MCES, for the Evinrude 55 HP Outboard Engine taught in the Small Craft Mechanics Course (SCMC).

The SCMC training material used for capturing immersive maintenance training content included: (a) Field Technical Manuals (PDF); (b) Student Handouts (Word); (c) Lecture Presentation (PowerPoint); (d) Videos; and (e) Exams.

The research team worked with the USMC instructors to assess the usability and scalability of various 3D model generation techniques.

Performance of Fine-Tuned and Original Models

The fine-tuning process was tailored through iterative adjustments to hyperparameters such as learning rate, batch size, training steps, and many others, ensuring that the model converged to its optimal performance in the USMC domain. By using publicly available pretrained weights, the team saved time and money, enabling us to evaluate the model's convergence and performance without requiring thousands of hours like in (Liu, 2023). Figure 6 shows the different models' convergence and loss reduction using three different pretrained weights benchmarks. The fine-tuning process illustrated significant improvement in loss reduction across all three benchmarks, with a consistent decrease in loss as the number of epochs progressed. As shown in the loss curves, each pretrained benchmark initially exhibited admirable but lackluster performance in our USMC domain, but through iterative fine-tuning, the loss converged toward optimal levels. This convergence indicates the model's ability to adapt and refine itself to specific domain data, showcasing robustness and resilience in learning complex geometric intuition for the electro-mechanical object domain for the USMC.

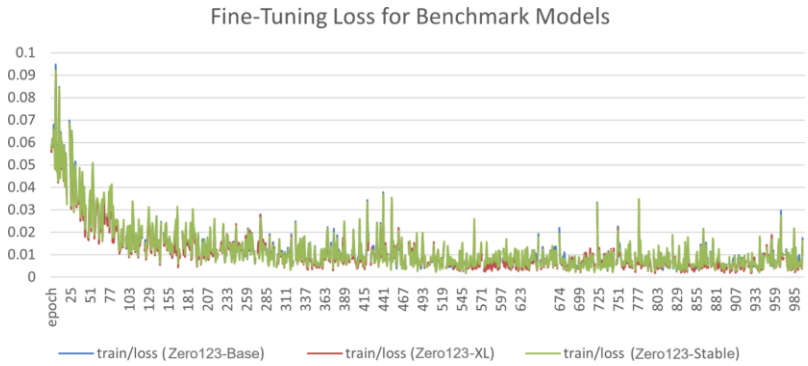

Figure 6: Fine-tuning loss curves for three different benchmark Zero-1-to-3 models (Zero123-Base, Zero123-XL, and Zero123-Stable).

The reduction in final loss values across the benchmarks underscores the model's potential for delivering accurate and high-fidelity 3D reconstructions for immersive military training scenarios and opens the field of fine-tuning to other complex domains. Figure 7 illustrates the progression from an original image to its 3D reconstruction using six different models. These include three original models (Zero123-Base, Zero123-XL, and Zero123-

Figure 7: Original image of Evinrude engine compared against pictures of generated 3D engine models from fine-tuned and original models.

Stable) and their corresponding domain fine-tuned versions. The original models, trained on non-domain datasets, provide a strong foundation for general 3D object reconstruction. However, the fine-tuned models, specifically adapted for electro-mechanical objects using our custom dataset generator, show significant improvements in the quality of the 3D reconstructions. These enhancements are evident in the finer feature details, more accurate textures, and overall realism of the generated models. Specifically, the team saw the improved prediction performance in the fine-tuned Zero123-Stable model. The graphic clearly demonstrates that our fine-tuning process leads to superior 3D asset outputs, making them better suited for specialized applications in USMC training environments.

Model Benchmarking and Statistics

In order to analyze the performance of different models for generating 3D content, we conducted a benchmarking study that includes various metrics, including the training dataset, specialized equipment, training time, object generation speed, and output quality. The comparison aims to clarify how each model was originally trained and their required training conditions. Table 1 presents a detailed comparison of the three models trained in this study.

Table 1: Comparison of Various 3D Model Creation Methods

	Zero123-Base (Liu, 2023)	**Zero123-XL (Updated Objaverse)**	**Zero123-Stable (StabilityAI Weights)**
Training Dataset	The standard Obajverse dataset of 800,000 objects	Updated 2024 ObjaverseXL dataset of 12 million objects	Updated 2024 ObjaverseXL dataset of 12 million objects
Fine-Tuning Dataset	In-House Electro-mechanical objects dataset	In-House Electro-mechanical objects dataset	In-House Electro-mechanical objects dataset
Specialized Equipment/Environment	NVIDIA A100 40 GB Tensor Core GPU	NVIDIA A100 40 GB Tensor Core GPU	NVIDIA A100 40 GB Tensor Core GPU
Training Time on 100 Objects	1 – 2 hours	12 – 24 hours	12 – 24 hours
Object Generation Speed	10 - 30 minutes	10 - 30 minutes	10 - 30 minutes
Output Quality	Low	Medium	Medium

COST BENEFIT ANALYSIS OF VARIOUS 3D MODEL CREATION PIPELINES

At the conclusion of the benchmarking survey, the research team researched comparative analyses and cost benefit analyses of three different approaches for the creation of 3D models for use in a 3D asset library. The term “content creation” has many meanings with regards to modernizing maintenance training. In this paper, content creation includes a whole gamut of generation of 3-D models of the to-be-maintained equipment to the creation of digital lessons where these 3D models are an integral part of lessons throughout the entire USMC curriculum.

Generative AI

At one end of the spectrum are the generative AI methods described in this paper, that convert 2D pictures found in student handouts to 3D models. The team found that the quality of the 3D conversion is the lowest, with the expectation that it will be used to create 3D models of common tools and parts that do not require high-fidelity representation. Since the conversion is fully automated, the 3D asset generation speed is governed by the processing power of the AI hardware. With the GPU setup described previously (see section “Industry Gaps and Optimizations: Tailoring Zero-1-to-3 to the USMC Use-Case”), the duration averaged about 10 to 30 minutes per 3D model to deploy in the AR/VR library. The primary capital investment is in the procurement of a capable GPU. As of writing of this paper, the NVIDIA A100 40 GB Tensor Core GPU, that was used for the experiment, retailed for around $19K (Cisco, 2024), with an additional $5K - 6K to purchase a customized server rack (Rack, 2024) to run the ML algorithms. However, after this initial investment, zero subsequent investment is needed to process any number of images. Hence, this method is suitable for generating 3D models en masse for common use in a schoolhouse learning curricula.

3D Scanning/Photogrammetry

In the middle of the range is 3D-scanning using a customized iOS app built using Apple's RealityKit (Apple AR Kit, 2024) API. The iPhone has Light Detection and Ranging (LiDAR) that enables the app to automatically scan and 3D-synthesize physical objects during walkarounds. The 3D scanning app runs on Apple's iPhones 13 Pro or later. Instructors intending to generate 3D scans of physical training objects can run the 3D scanning app on their iPhone, and walk around the object with the iPhone pointed towards it. The app will instruct the user to take walkaround scans of the object at various heights, after which it automatically synthesizes the 3D model of the object in an .OBJ format. It generates a detailed surface scan of the objects with medium quality that is suitable for equipment taught in an immersive class such as a marine engine. The typical time to scan an object is 5 minutes, with an additional 10 to 15 minutes needed to generate a 3D model for use in a library. The only investment is the procurement of an iPhone, which typically costs $1K. After the initial investment, zero subsequent investment is needed to 3D-scan any number of objects.

Specialized 3D Modeling Contractors

At the top of the stack are high-fidelity CAD models generated by vendors who create not just CAD models, but entire training content pipelines (differing from the generative AI and 3D scanning/photogrammetry options described above). A survey of eleven such vendors developing maintenance training solutions conducted by US government civilians in 2024 revealed a wide array for content creation strategies. Five of the eleven vendors offered hybrid solutions which varied broadly. Some products allow for customers to bring their own CAD files or 3D objects and build lessons, while others only allowed for minor modifications of vendor-created content (e.g., changing the text associated with a particular maintenance step). The remaining six vendors only allowed for vendor-authored content, where they employ their team of developers, graphic artists, user experience designers, and instructional systems designers to develop the 3D models and surrounding lessons. Most vendors noted that the Government would retain ownership of the resulting models, though some mentioned that the data provided would be the models paired with the lesson materials stored in a proprietary format. Vendors cautioned against mixing company and customer-created content as well, as there is often a mismatch in graphic quality and model fidelity that degrades training efficacy.

The process, cost, and time required for vendors or other third-parties to create content varies widely. Significant factors include the amount of already-available data, the complexity of the system to be modeled, the amount of detail required in the model, and the number of procedures to be demonstrated in the lesson plan. More than one vendor began the content creation process by ingesting CAD files and porting them into their environment of choice - Unity was commonly cited. One vendor estimated the cost to implement high-complexity procedures to be approximately ~$50K, requiring up to a month of development time. Additionally, any content changes must be done by the contractor; the Government would have no ability to make their own changes.

In a typical scenario, it can take weeks or months to generate content, depending on the complexity of the system. For example, a large server cabinet can take months to turn into training content. When told that the USMC oftentimes does not own the CAD models (or other components of the Technical Data Package), some companies responded that they could create those models as well. They described a process where people (oftentimes contractors, but it could be active-duty personnel or DoD civilians) will take photos of the equipment from all required angles (this can add up to thousands of pictures taken over many days). Then, graphic designers will create a scale-model of the equipment for ingestion into the training environment. Creating a digital representation of a system like this from scratch (i.e., with no available 3D CAD model to use) can take months of work and can cost hundreds of thousands of dollars. One vendor noted that a server rack required nearly a week of scanning and nearly two months of modeling by multiple graphics designers. This is a process that would have to be undertaken for each server rack, vehicle, and component that needs to be turned into digital content for maintainer training.

Even implementing a modification to the model (i.e., a new hardware iteration) can take weeks or months and thousands of dollars. This is because new photos and source material must be acquired, and then models need to be tweaked by contractor-employed graphic designers. Minor modifications can still take weeks due to the process of going back and forth with a contractor, while large modifications require large work orders.

Table 2: Cost Benefit Analysis of Various 3D Model Creation Methods

	Generative AI-powered 2D image to 3D model conversion	3D Scanning/Photogrammetry	Outside contractors specializing in 3D Modeling
Source material	Common images found in USMC schoolhouse material such as presentations and student handouts	Physical training objects found in a schoolhouse	Thousands of high-quality photos taken over several days. Graphic designers process photos into a scale-model.
Specialized equipment/environment	NVIDIA A100 40 GB Tensor Core GPU	iPhone 13 Pro or later	Unity
Cost	Initial investment of $25K for Deep Learning server. $0 subsequently.	$1K for iPhone 13 Pro or later. $0 subsequently.	Hundreds of thousands of dollars to implement each high-complexity procedures
Time to create 3D model for use in a library	10 - 30 min	15 min	1 month of development time per equipment
3D object quality	Low-Medium	Medium	High
Scalability	High. Ability to batch-covert several images at once.	Medium. Proportional to the extent of the effort put in 3D-scanning the equipment.	Low. Vendor will have to be contracted for each equipment.
Customization potential	Medium. Users can pick and choose pictures of desired physical objects from course material.	High. Ability to organically generate 3D models from desired physical objects.	Low. Limited ability to supply own 3D content (mostly CAD). Tens of thousands of dollars and up to weeks and months for modifications.

Usability Feedback

The research team is in the process of integrating the generative AI-based 2D picture to 3D model conversation pipeline into the immersive training platform that is currently being used by USMC schoolhouses such as the MCES. The SCMC chief instructor at MCES, who has decades of experience in schoolhouse instructions, was briefed on this capability and commented that this capability has the potential to save time (see Table 2) for the team of course creators. The chief instructor mentioned that the course creator(s) are currently using the in-house 3D scanning app (see Section "3D Scanning/Photogrammetry") for generating their 3D assets, and while this affords significant time and cost savings compared to relying on outside contractors, the generative AI-powered automated conversion of the 3D models will provide even further time savings. One of the features that the chief instructor would like to see improved is precision while extracting the 3D model from its background and removing the extraneous features.

CHALLENGES AND STRATEGIC INSIGHTS

Downsizing the Model

The original Zero123-Base model was trained on an enormous dataset over a long period of time, resulting in a substantial monetary cost. To make our solution viable for the USMC, the research team needed to achieve significant optimizations without compromising performance. As such, downsizing the model to fit on a single A100 40 GB GPU required several strategic insights to reduce its size and computational demands while maintaining its performance.

Among the strategies used by the team were advanced memory management techniques, like gradient checkpointing, which optimized the training loop to handle memory more efficiently. Additionally, the team customized the training regimen by carefully tuning hyperparameters, including learning rate schedules and batch sizes, to ensure efficient training and convergence within the available computational resources. By integrating these techniques, the team

successfully adapted the Zero-1-to-3 model to run efficiently on a single A100 40GB GPU, ensuring high performance while significantly reducing hardware requirements and costs. This optimization makes the model practical for deployment in environments with limited resources, without requiring the purchase of high-performance hardware.

Data Reliability and Accuracy

To ensure the dataset generator produced high-quality data, the team incorporated rigorous validation procedures at various stages of the data generation process. This included automatic checks for data consistency, such as verifying that all generated images were correctly paired with their corresponding camera matrices, and ensuring that the transformations applied were accurate and representative of real-world scenarios. Additionally, the research team conducted manual reviews of subsets of generated data to catch any logic issues that automation may have introduced, and adjusted the dataset generator accordingly.

The team implemented comprehensive preprocessing steps to clean the data, removing any anomalies, duplicates, or irrelevant information that could potentially skew the training process. This preprocessing phase included standardizing formats and ensuring consistent labeling across the dataset. To maintain high standards of data quality, the team used automation tools and scripts to perform initial checks, followed by manual inspections to catch any subtle issues that the automated processes might miss.

Future Work

Our current model is trained primarily on datasets comprising single objects against plain backgrounds. While this approach has yielded strong generalization, addressing the challenge of generalizing our model to scenes with complex backgrounds remains an essential area for future research. Future work will focus on enhancing the model's capability to accurately reconstruct and generate objects among varied backgrounds, thereby increasing its applicability and robustness in more diverse environments.

This may include improving the model's architecture but also expanding the training datasets to include a wider variety of scenes and backgrounds. The team anticipates that these improvements will significantly enhance the quality and realism of the generated objects, making our model more versatile and effective for a broader range of applications.

CONCLUSIONS

In this paper, the research team demonstrated the concept of an automated web-based 3D content-generation pipeline that provides savings in terms of: (a) time spent in manually creating the models and (b) expenses related to acquiring licenses of professional CAD modeling tools and/or hiring outside contractors to professionally scan 3D models. Moreover, this process does not require any 3D modeling or graphical design skills. The 3D models are seamlessly prepopulated into an AR/VR content library that can be sourced for generation of immersive content using the AR/VR Designer. All these capabilities lead to minimizing the need for physical training assets that take up schoolhouses' valuable resources, thus easing the staffing and logistical burden on the schoolhouses. The resulting immersive training can help the USMC achieve the strategic goal of training technicians to be efficient maintainers.

The following key findings emerged from the analysis of the generative AI scalability testing. First, using the fine-tuned Zero-1-to-3 model significantly reduced the time required to produce usable 3D models, with the process averaging just a few minutes per object. This represents a significant improvement over traditional methods, which could take several days for a single model produced by an artist. Second, the cost-benefit analysis revealed that the upfront investment in the AI model and computing hardware could be offset by eliminating the recurring expenses associated with professional CAD software licenses and specialized labor. Our generative AI-based 3D modeling approach is advantageous when asset quality is not of great importance; conversely, when asset quality is critical, the traditional methods of photogrammetry or specialized labor may be more appropriate.

The following key findings emerged from the comparison of 3D model creation pipelines. First, our comparison highlighted the difference in quality between the predecessor NeRF models and the state-of-the-art Zero-1-to-3 model. The Zero-1-to-3 model performed admirably even prior to fine-tuning on electro-mechanical objects and exhibited a remarkable understanding of complex geometric intuition and components characteristic of electro-mechanical

equipment. After fine-tuning with minimal domain training input, the model's performance in generating high-fidelity models was exceptional.

Future work in 3D object generation could involve daisy-chaining text-to-image models developed using deep learning methodologies that generate digital images from natural language descriptions, with the image-to-3D model research conducted in this paper. In this manner, users could potentially generate text-to-3D models just by entering textual "prompts."

ACKNOWLEDGEMENTS

The research team would like to thank Dr. Peter Squire from ONR for supporting and funding the research, and Dr. Jason Wong for his guidance during the development phase. The team would like to thank Mr. James (Scotty) Moore, Capabilities Branch Head at MCES, for providing access to the SCMC training material, and for facilitating numerous touchpoints with the Marines at the MCES. The team would also like to thank GySgt Craig Ruhnke at the MCES, and CWO Micah Soboleski and CWO Cervantes at MCCSSS, for providing access to their training material and facilities, and well as providing valuable feedback on the training potential of various 3D content-generation pipelines.

REFERENCES

A100 (2020). *Blender.org — NVIDIA A100 Tensor Core GPU*. *https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf*

Apple AR Kit (2024). *RealityKit.* *https://developer.apple.com/augmented-reality/realitykit/*

Blender (2022). *Blender.org — Home of the Blender project—Free and Open 3D Creation Software.* *https://www.blender.org/*

Cisco (2024). *NVIDIA A100 Tensor Core 40GB PCIe GPU Card*. *https://www.cdw.com/product/cisco-nvidia-a100-tensor-core-40gb-pcie-gpu-card/6546028*

EGMTS (2023) *Enterprise Ground Maintenance Training Simulator*. *https://www.tecom.marines.mil/Units/Divisions/Range-and-Training-Programs-Division/EGMTS/*

Haste, D, Ghoshal, S, Yerdon, V, Hidalgo, M, Beaubien, J, & Wong, J. (2022). *Immersive Content Creation Pipeline for Information Age Training, Interservice / Industry Training, Simulation and Education .Conference (I/ITSEC) 2022 Paper No. 22242*

InfoWorld (2024). *The temptation of AI as a service*. *https://edt.infoworld.com/q/15b35J71QoikijVNcZq2k5dAB/wv*

Mildenhall, B, et al (2020). *NeRF - Representing Scenes as Neural Radiance Fields for View Synthesis.* *https://arxiv.org/pdf/2003.08934*

Moodle. (2022). *Moodle—Open-source learning platform*. *https://moodle.org/*

Liu, R., et al (2023). *Zero-1-to-3: Zero-shot One Image to 3D Object.* *https://arxiv.org/pdf/2303.11328*

Liu, R., et al (2024). *Allen Institute for AI, Objaverse-XL: A Universe of 10M+ 3D Objects*. *https://objaverse.allenai.org/objaverse-xl-paper.pdf*

Rack (2024). *VBOZ 42U E Series VE6X4208 Server Rack*. *https://rack.com.sg/product-category/d-series/37u-server-rack-d-series/*

Unity (2022). *Unity Real-Time Development Platform*. *https://www.unity.com/*

WebXR (2022). *Immersive Web Developer Home*. *https://immersiveweb.dev/*

Zhou, J., et al (2020). *Graph neural networks: A review of methods and applications. AI Open, 1, 57-81*. *https://doi.org/10.1016/j.aiopen.2021.01.001*