CiRACORE
รางวัลพระราชทาน
นักเทคโนโลยีดีเด่น
ประจำปี 2562

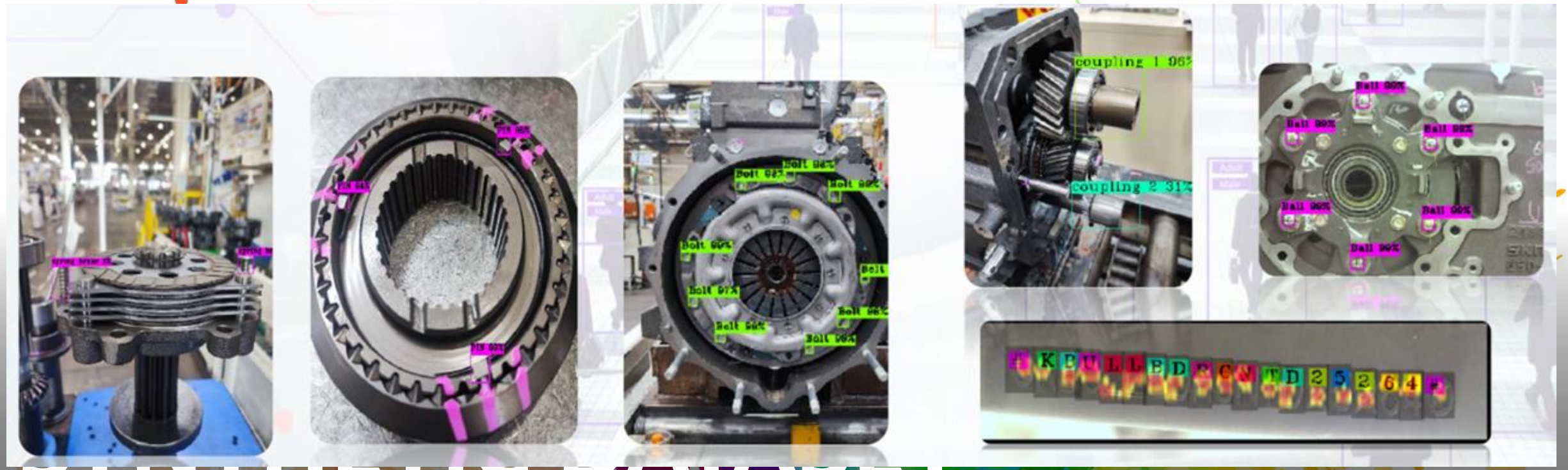
KMITL
พระจอมเกล้าลาดกระบัง

คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

GENERATIVE AI FOR INDUSTRIAL APPLICATIONS: SYNTHETIC DATASET

Associate Prof. Dr. Siridech Boonsang

Dean,
School of Information Technology
King Mongkut's Institute of Technology Ladkrabang

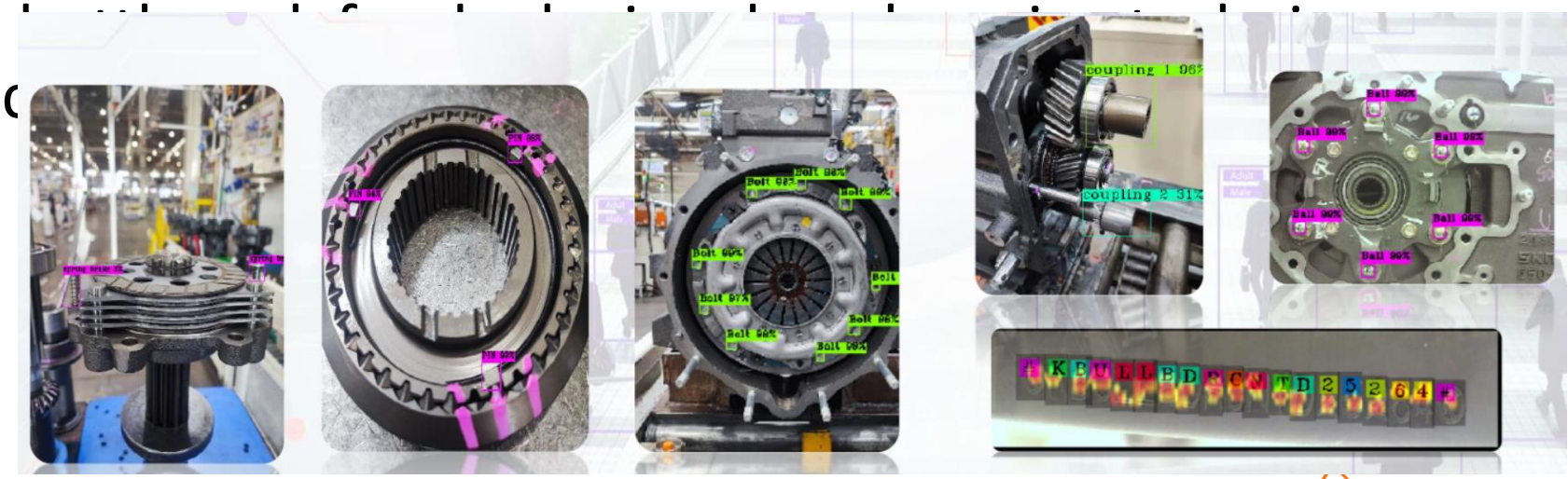


SYNTHETIC DATASET GENERATION

DEEP LEARNING IN INDUSTRIAL APPLICATIONS

SYNTHETIC DATASET GENERATION FOR OBJECT-TO-MODEL DEEP LEARNING IN INDUSTRIAL APPLICATIONS

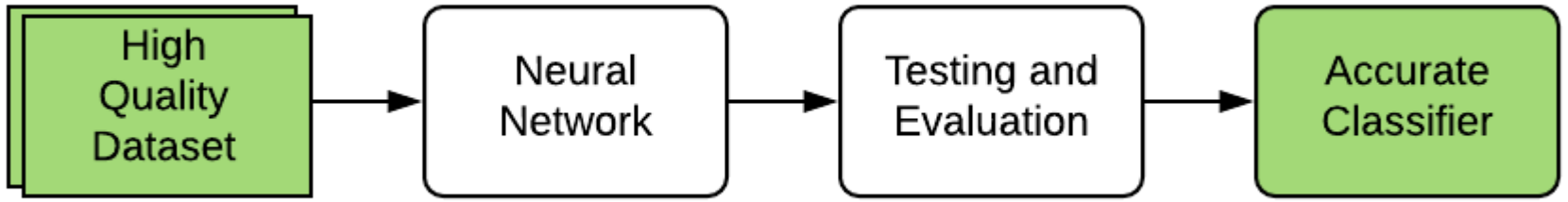
- - Availability of large image datasets has been crucial for the success of deep learning-based classification and detection methods.
- - Everyday object datasets are widely available, but specific industrial use-case datasets (e.g., identifying packaged products in a warehouse) are scarce.
- - In these industrial cases, datasets need to be created from scratch, which becomes a
in industrial applic

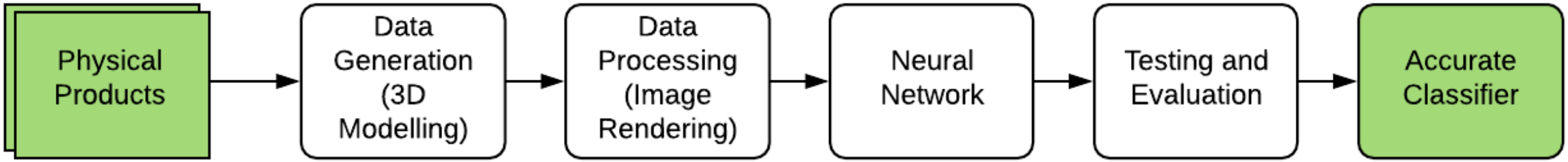


3D Modelling in Deep Learning

- 3D modeling has been traditionally used in computer vision research and is now being considered for deep learning-based image classification and object detection.
- Researchers have used 3D models in conjunction with CNNs to train networks for real image applications.
- Su et. al. [1] demonstrated the use of 3D models for viewpoint estimation by creating a database of rendered training images using CAD models, outperforming state-of-the-art methods on real image test sets.
- Peng et. al. [2] and Sakar et. al. [3] utilized a large number of 3D CAD models to render realistic training images and trained classifiers for classifying real-world images of the objects.
- Temblay et. al. [4] expanded on this approach by applying synthetic data generated from 3D CAD models to object detection. They employed domain randomization, varying parameters like lighting, pose, and object textures, and trained an object detection network on automatically generated non-photorealistic synthetic data.
- Wong et. al. proposed the method that use a synthetic dataset generated using photogrammetry techniques on real-world objects, consisting of 100k synthetic images. Training an InceptionV3 CNN on this dataset achieved 95.8% classification accuracy on real supermarket product images, and a one-stage RetinaNet detector trained on the synthetic, annotated images accurately localized and classified the products in real-time.

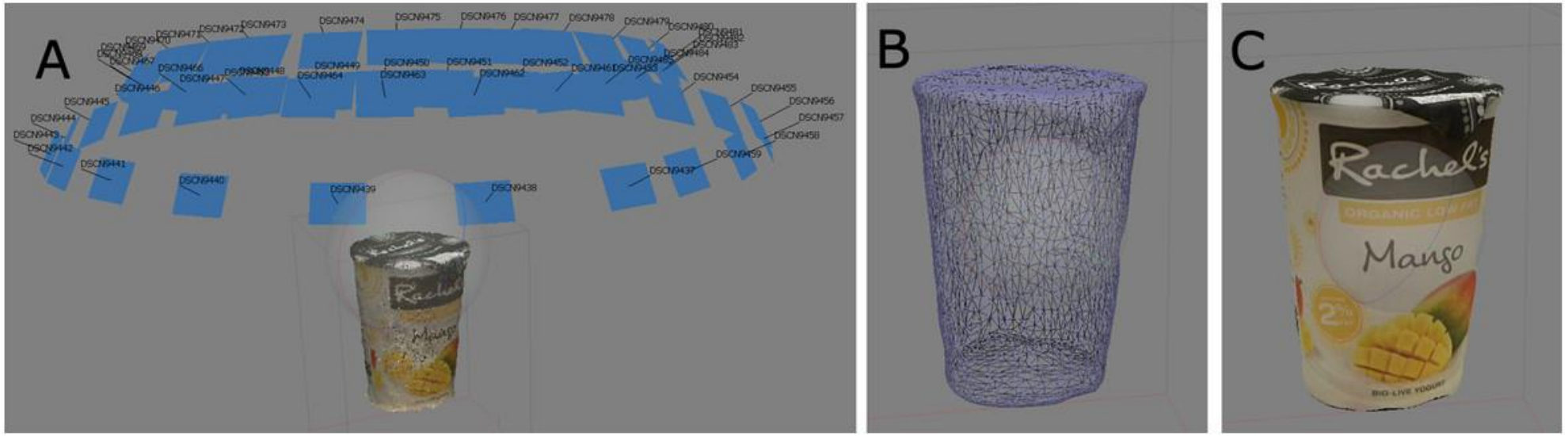
Standard Pipeline Design vs Custom Pipeline Design

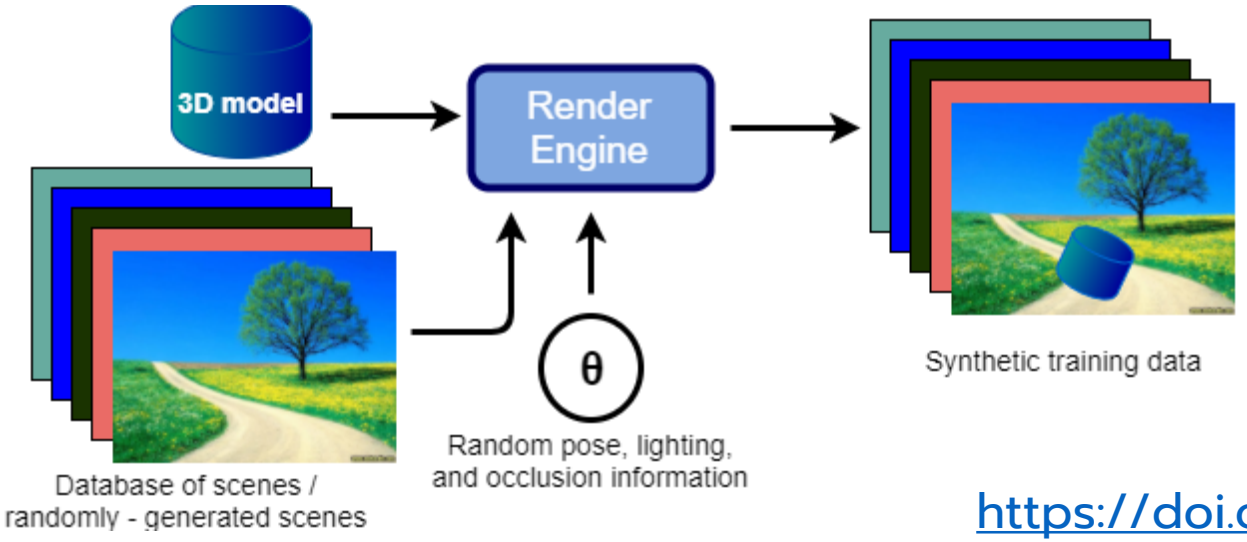




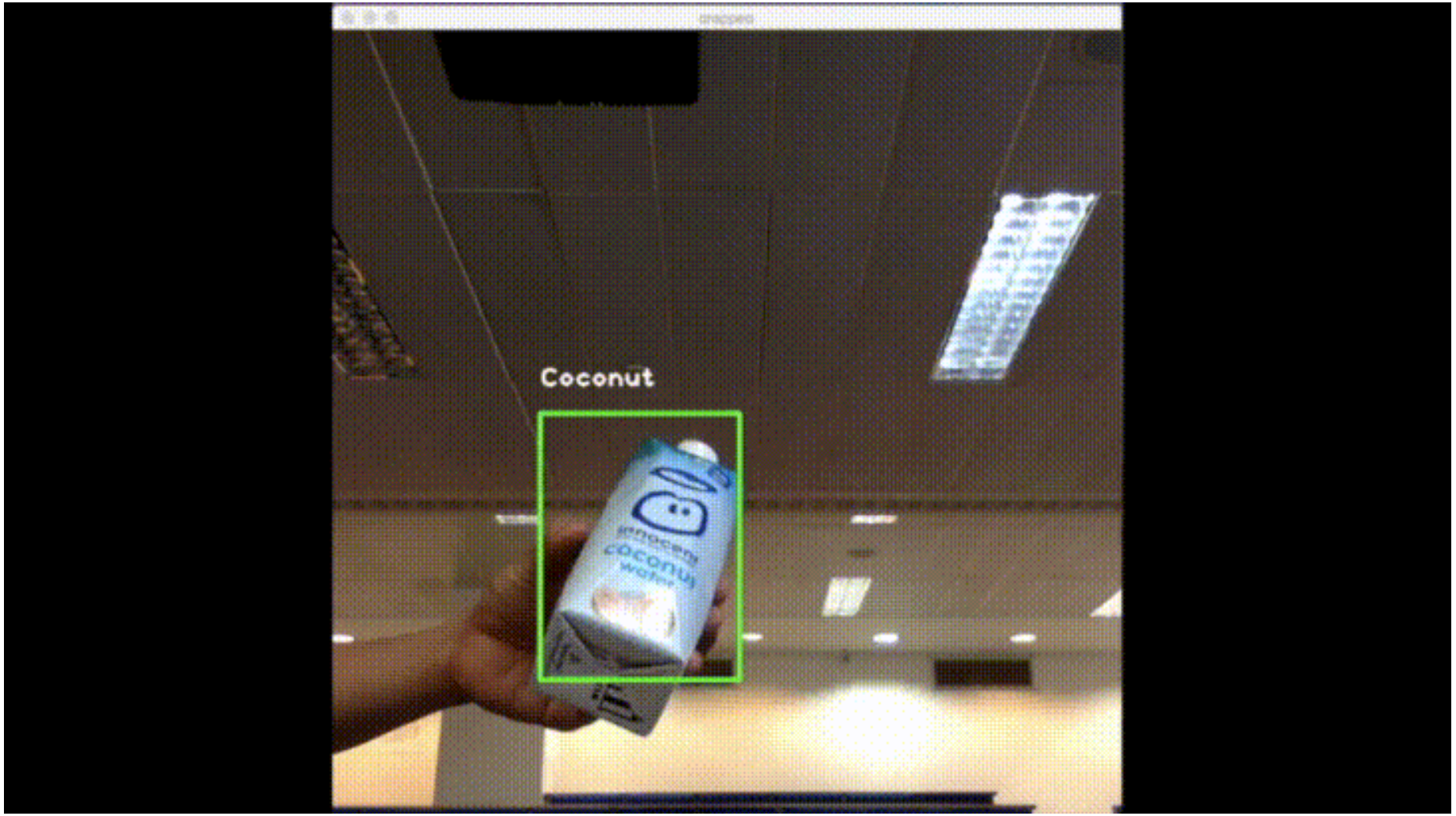
ACCEPTED FOR PUBLICATION IN PEERJ COMPUTER SCIENCE - SEPTEMBER 2019 https://doi.org/10.7717/peerj-cs.222

Image Rendering





https://doi.org/10.7717/peerj-cs.222



https://doi.org/10.7717/peerj-cs.222

Digital Twin based model of a “Pick and Place” robotic process based on the rotation classification pipeline

Procedia CIRP
Volume 103, **2021**, Pages 237-242

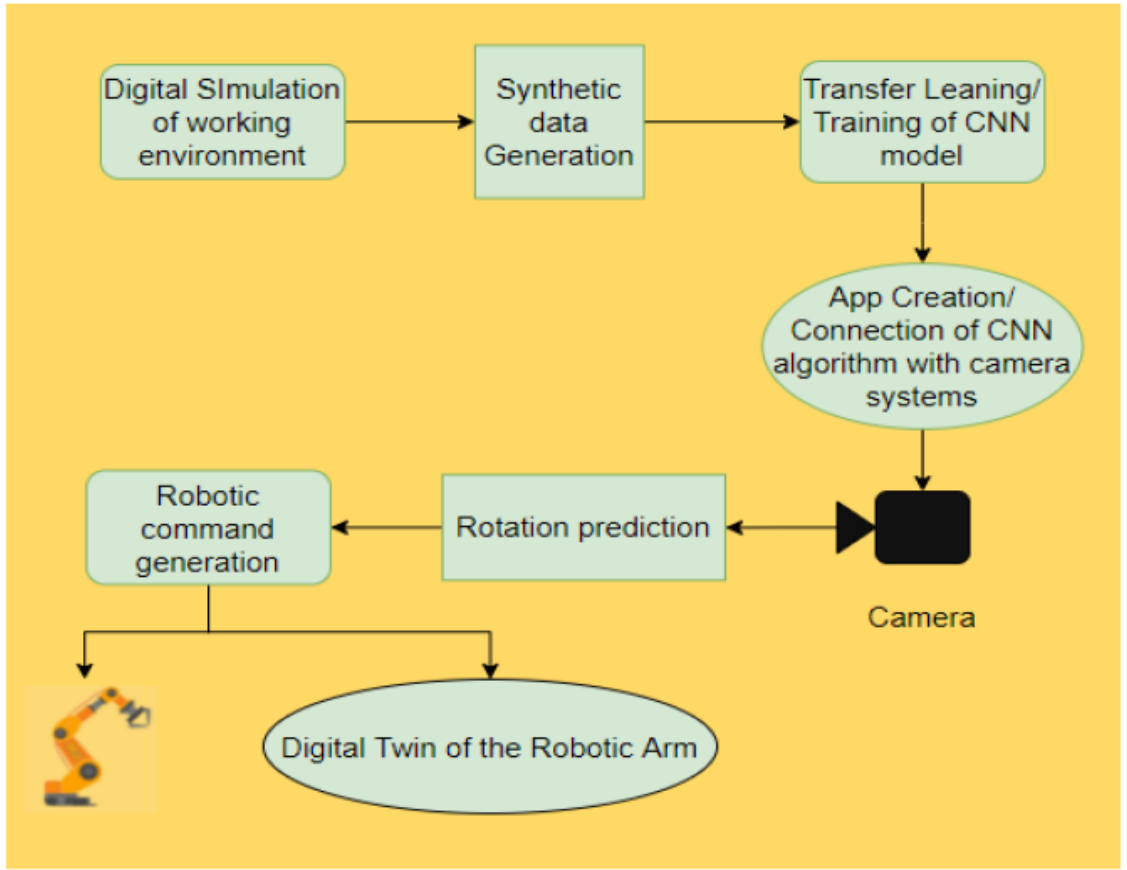


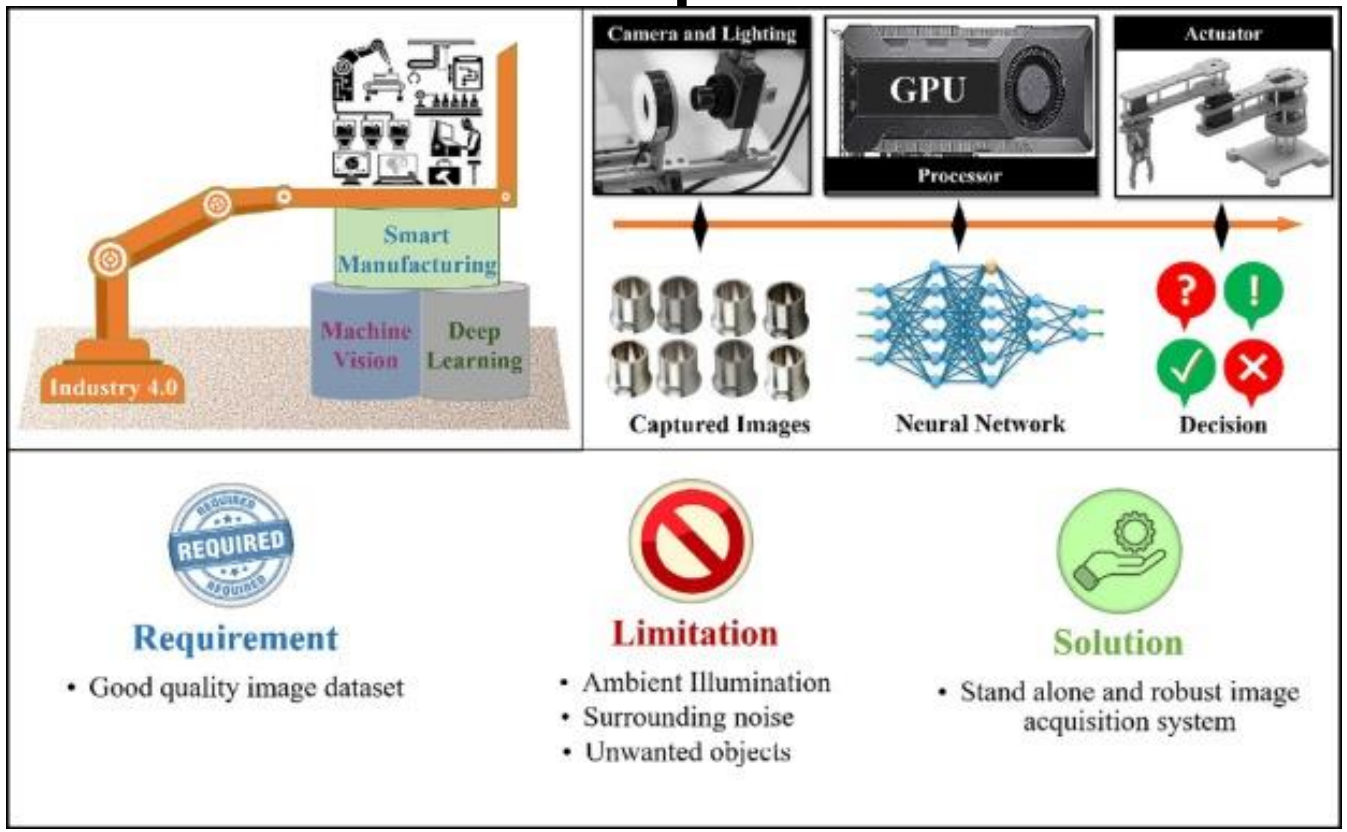
Fig. 1: Digital Twin-based implementation of the proposed synthetic-data based model

Fig. 2: Axes on part (1)

Fig. 3: Local and Global Coordinate Systems

https://doi.org/10.1016/j.procir.2021.10.038

Vision-based system for automated image dataset labelling and dimension measurements on shop floor

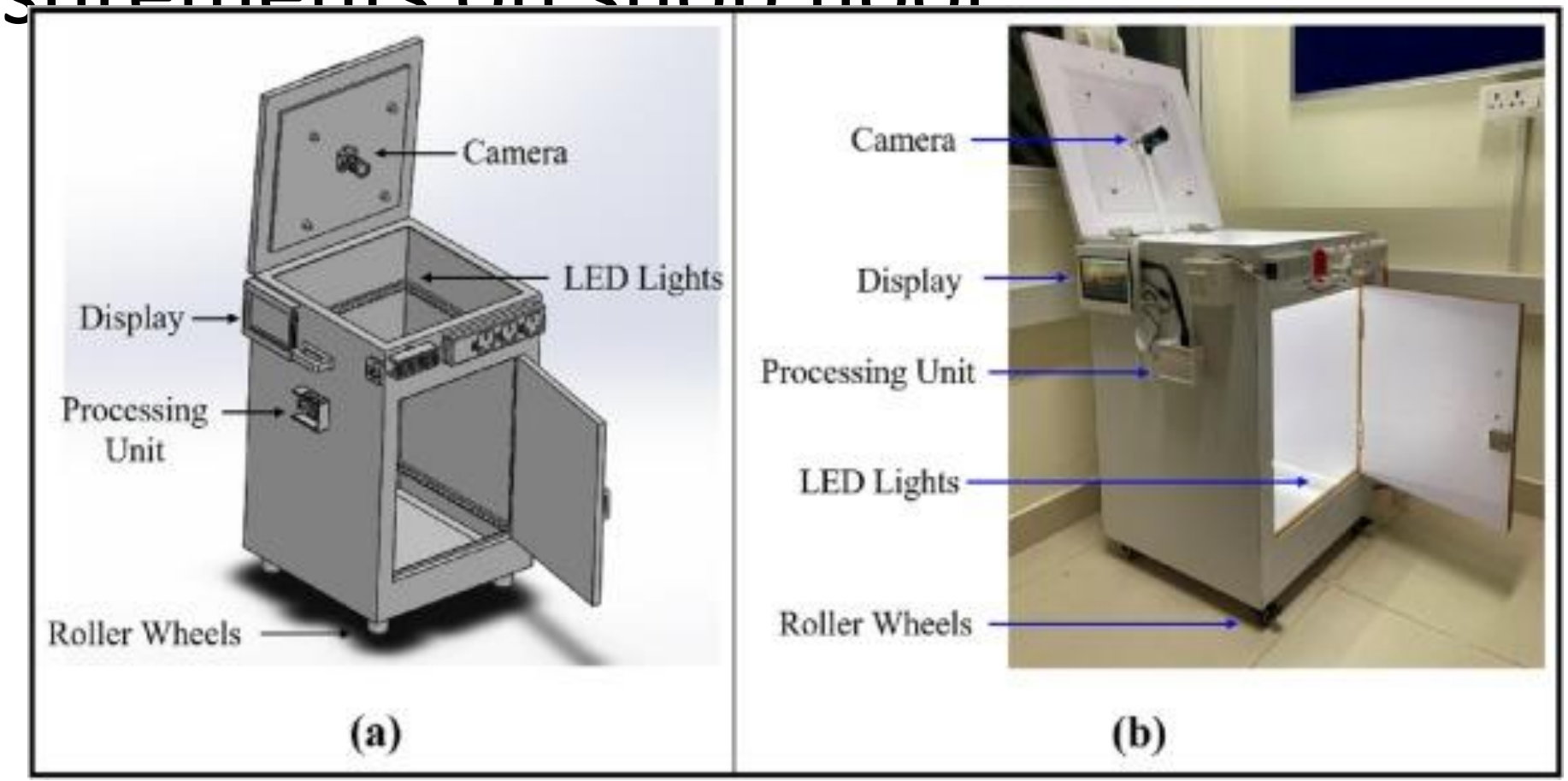


Measurement

Volume 216, July **2023**, 112980

https://doi.org/10.1016/j.measurement.2023.112980

Vision-based system for automated image dataset labelling and dimension measurements on shop floor



Measurement
Volume 216, July 2023, 112980

https://doi.org/10.1016/j.measurement.2023.112980

Vision-based system for automated image dataset labelling and dimension measurements on shop floor

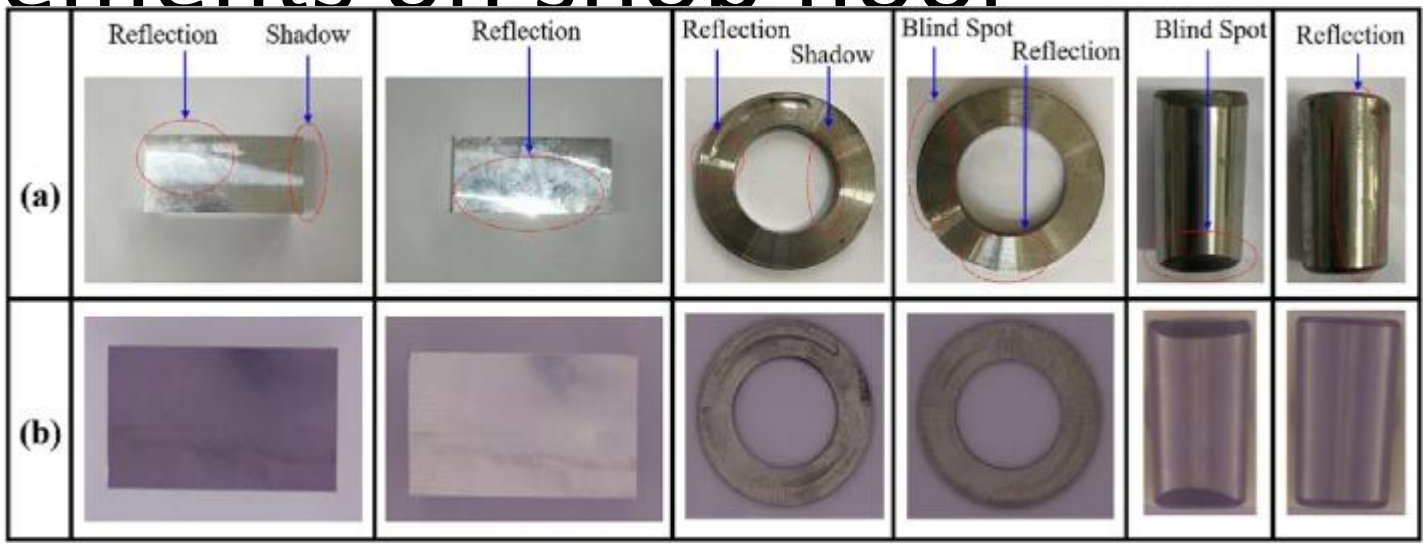
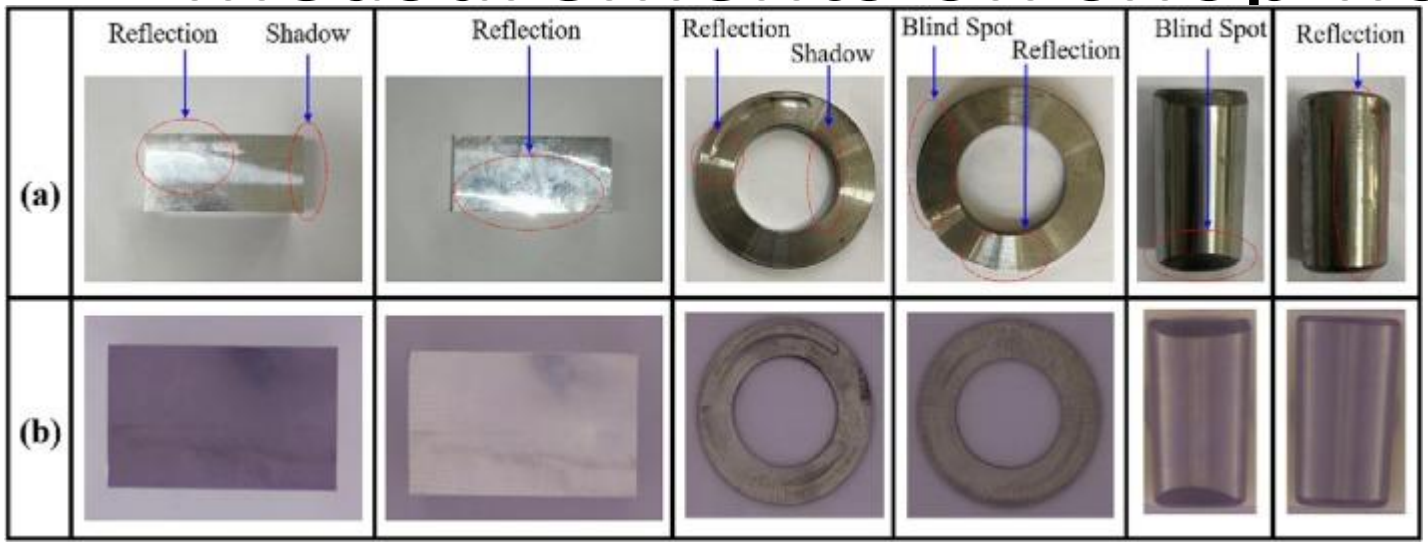


Fig. 5. Manufactured component images; (a) images directly acquired in the shop floor environment, (b) images acquired using the developed system.

Measurement
Volume 216, July **2023**, 112980
https://doi.org/10.1016/j.measurement.2023.112980

Vision-based system for automated image dataset labelling and dimension measurements on shop floor



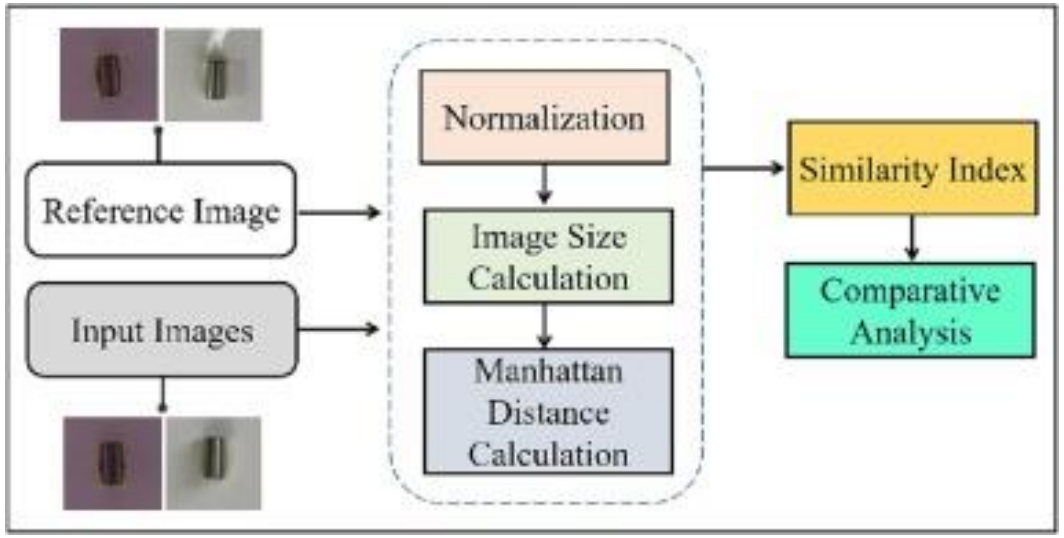
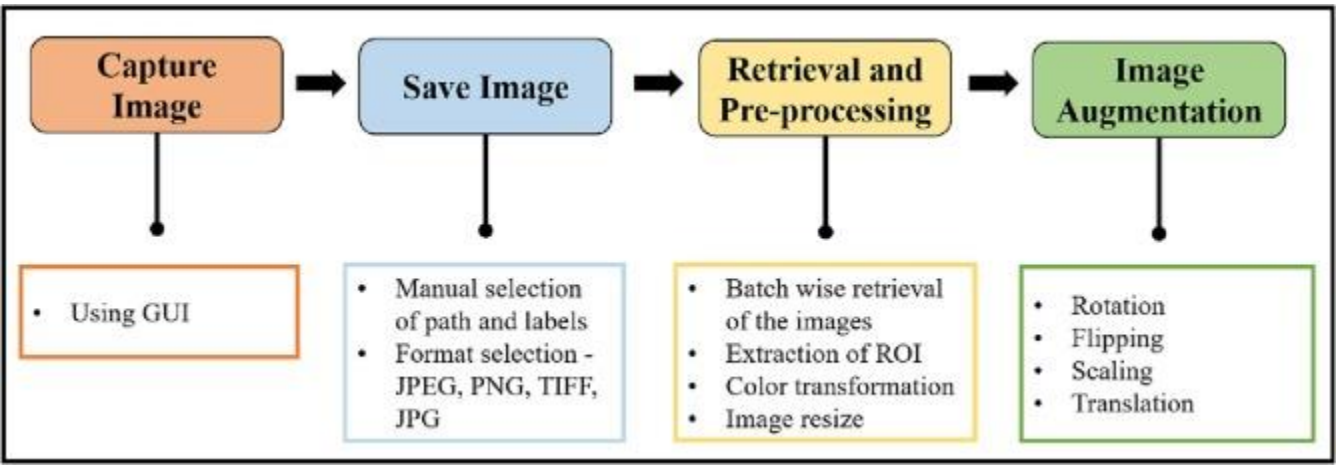


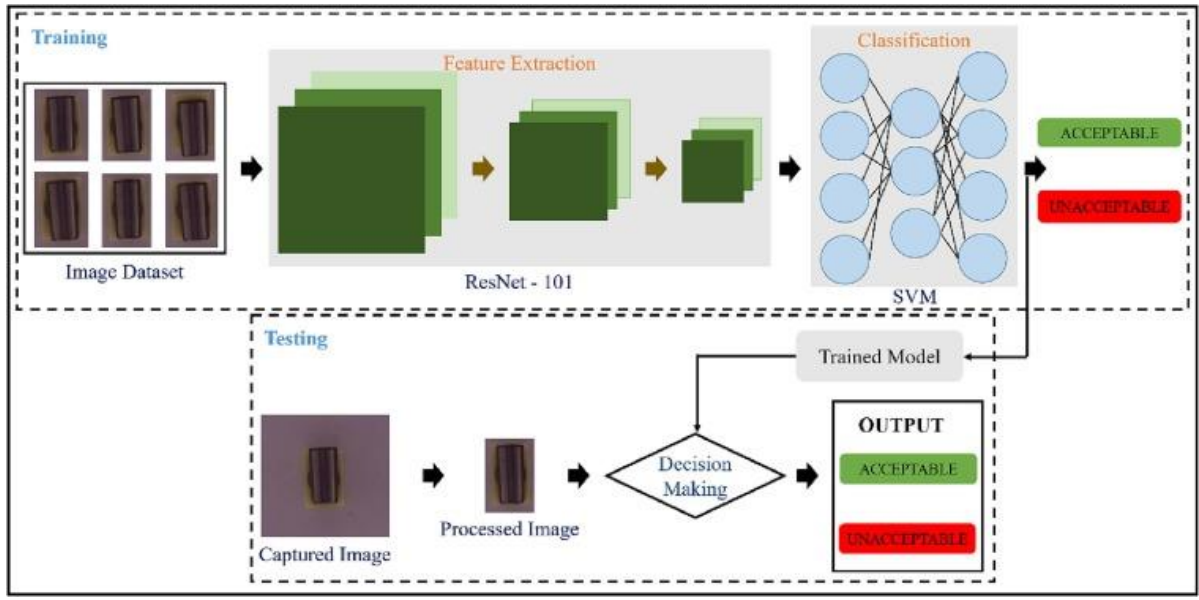
Fig. 5. Manufactured component images; (a) images directly acquired in the shop floor environment, (b) images acquired using the developed system.

Measurement
Volume 216, July 2023, 112980

https://doi.org/10.1016/j.measurement.2023.112980

Vision-based system for automated image dataset labelling and dimension measurements on shop floor





Measurement
Volume 216, July 2023, 112980

https://doi.org/10.1016/j.measurement.2023.112980

Generating Synthetic Data To Solve Industrial Control Problems By Modeling A Belt Conveyor

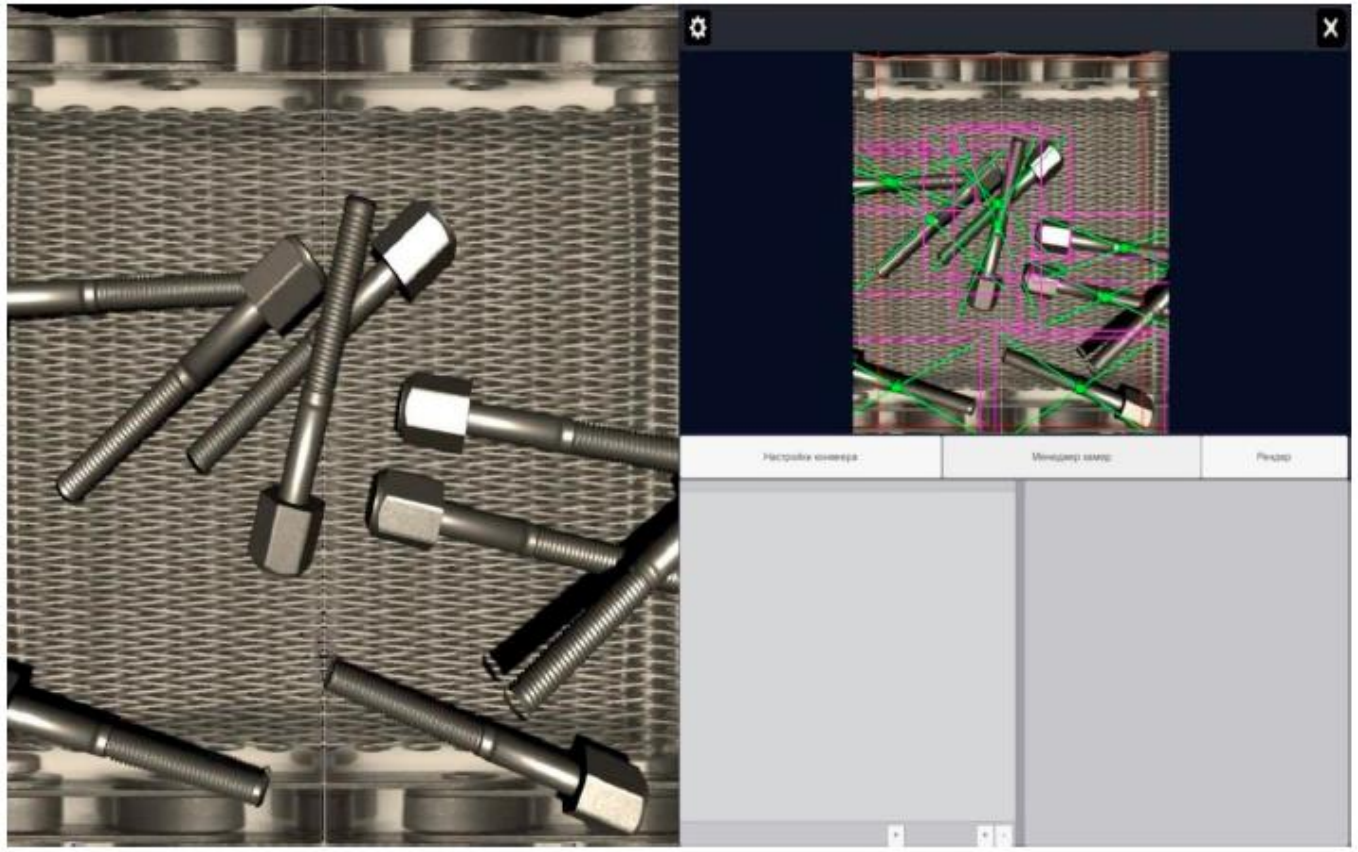


Fig. 3. Interface of the industrial belt conveyor model module

Fig. 4. Example of synthetic image with bolts

https://doi.org/10.1016/j.procs.2022.11.010

Procedia Computer Science

Volume 212, **2022**, Pages 264-274

Image-Bot: Generating Synthetic Object Detection Datasets for Small and Medium-Sized Manufacturing Companies

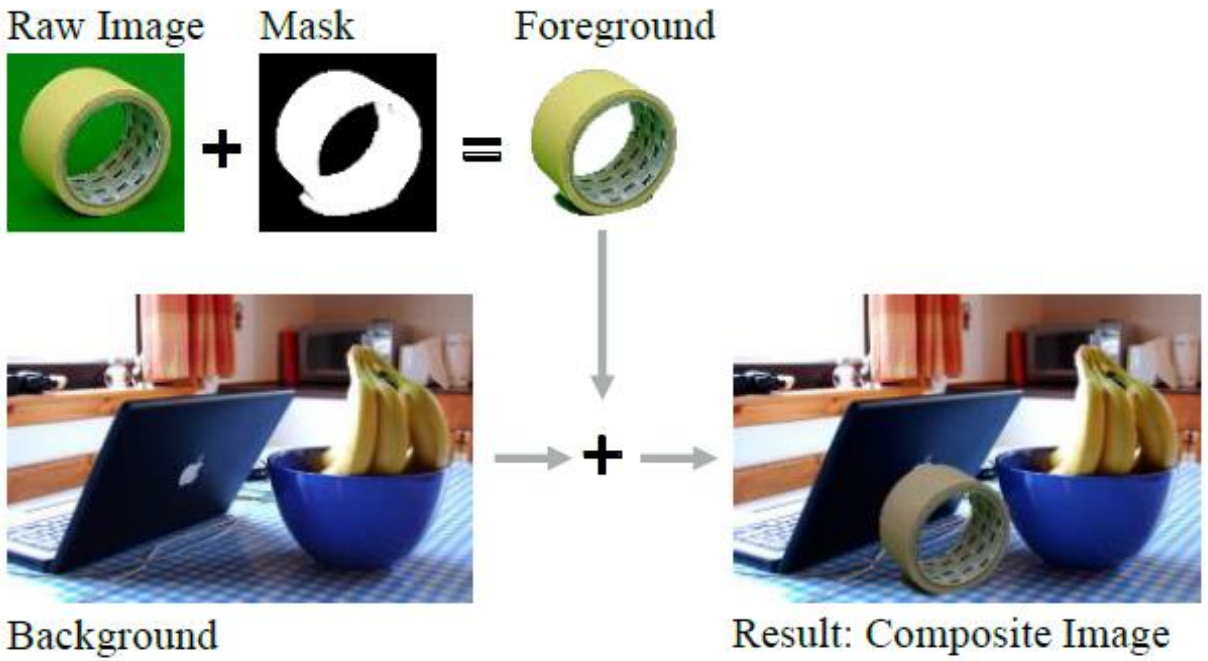


Fig. 1. Image masking and blending to insert a foreground image into a background image (background image from [20]).

Fig. 2. Physical setup to capture and process the images (personal computer is not on the picture).

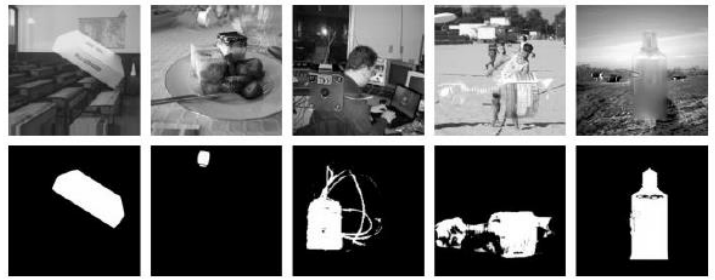


Fig. 6. Samples from the generated dataset (synthetic image and mask).

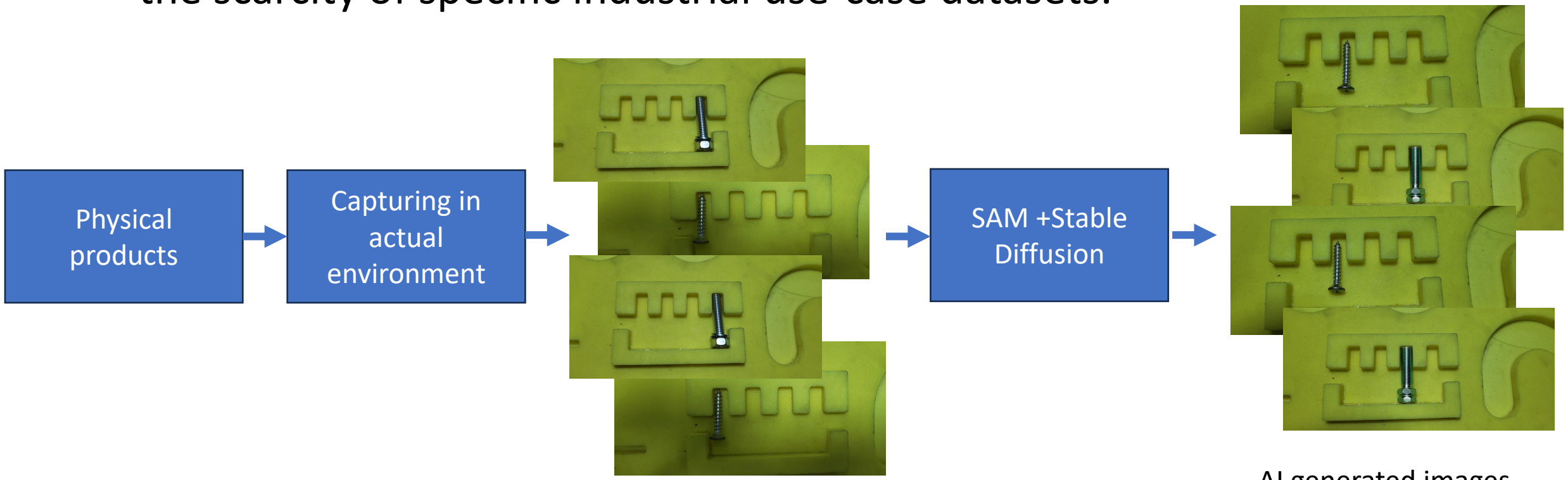
https://doi.org/10.1016/j.procir.2022.05.004

Procedia CIRP

Volume 107, **2022**, Pages 434-439

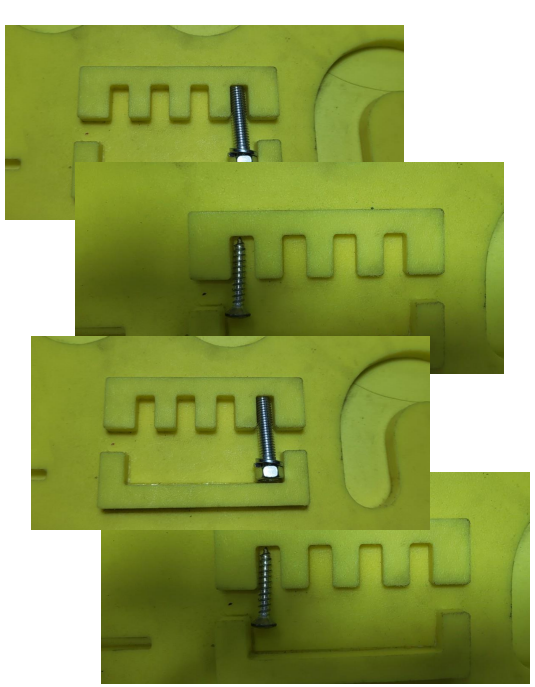
Proposed methodology

- Use of Generative AI (Stable Diffusion) to eliminate the problem of the scarcity of specific industrial use-case datasets.

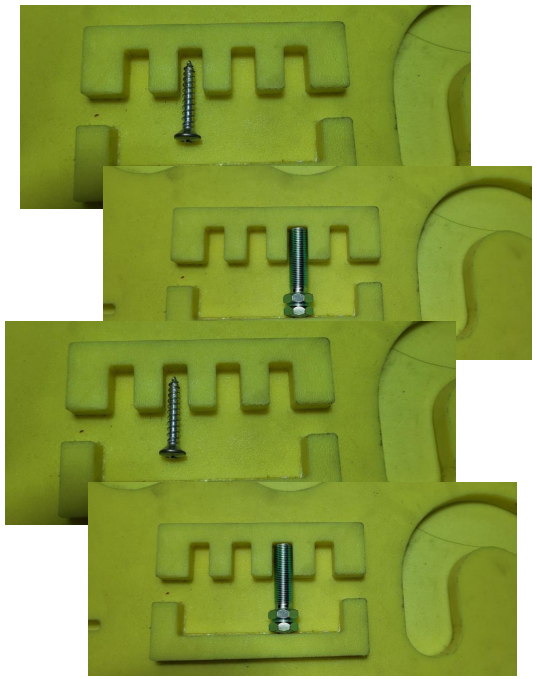


Proposed methodology

- With AI generated specific industrial use-case datasets, we fine tuned a pretrained distilled ViT large model to specific conditions



Captured images



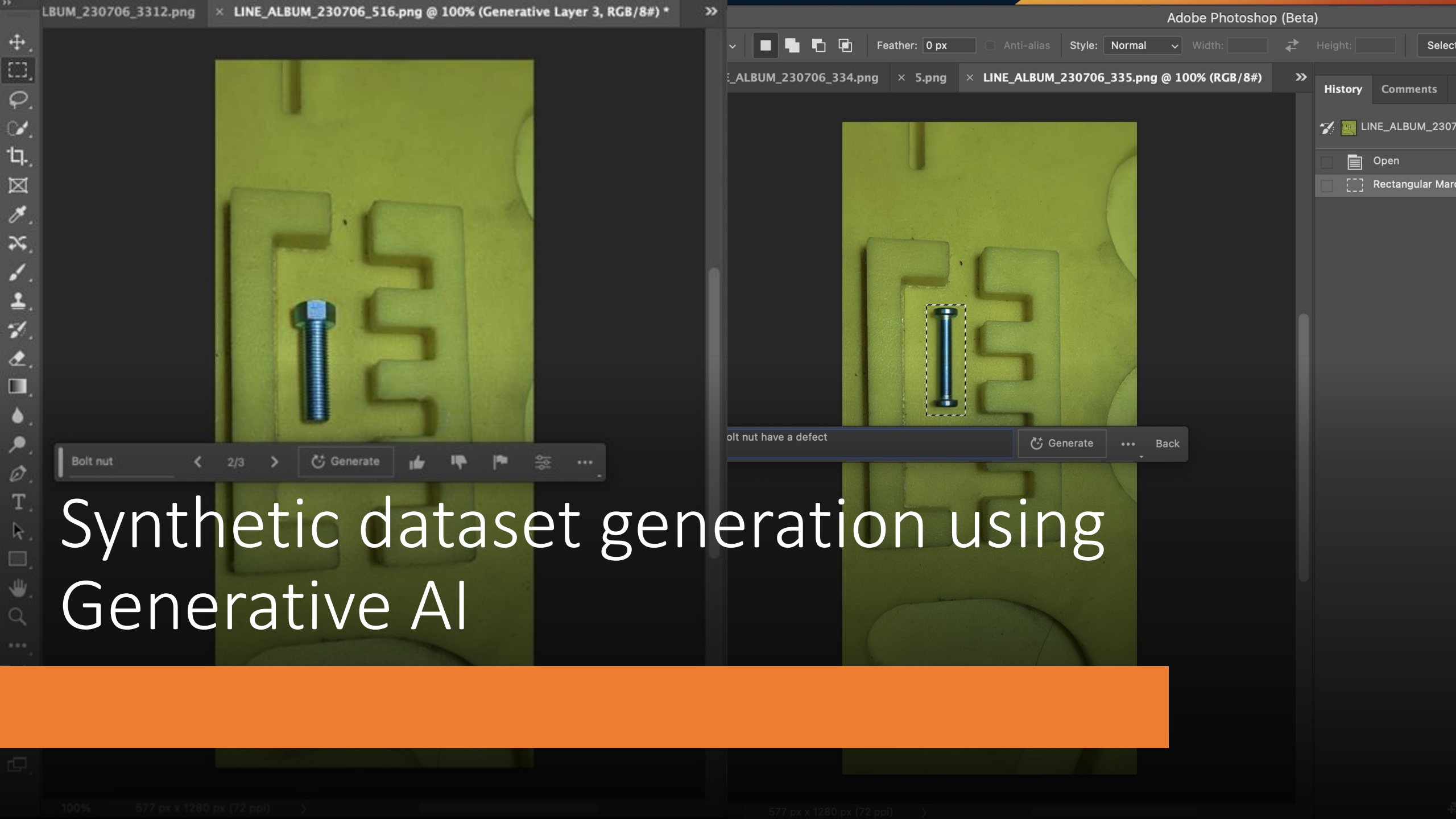
AI generated images

Fine tuning on

Pretrained distilled ViT foundation model

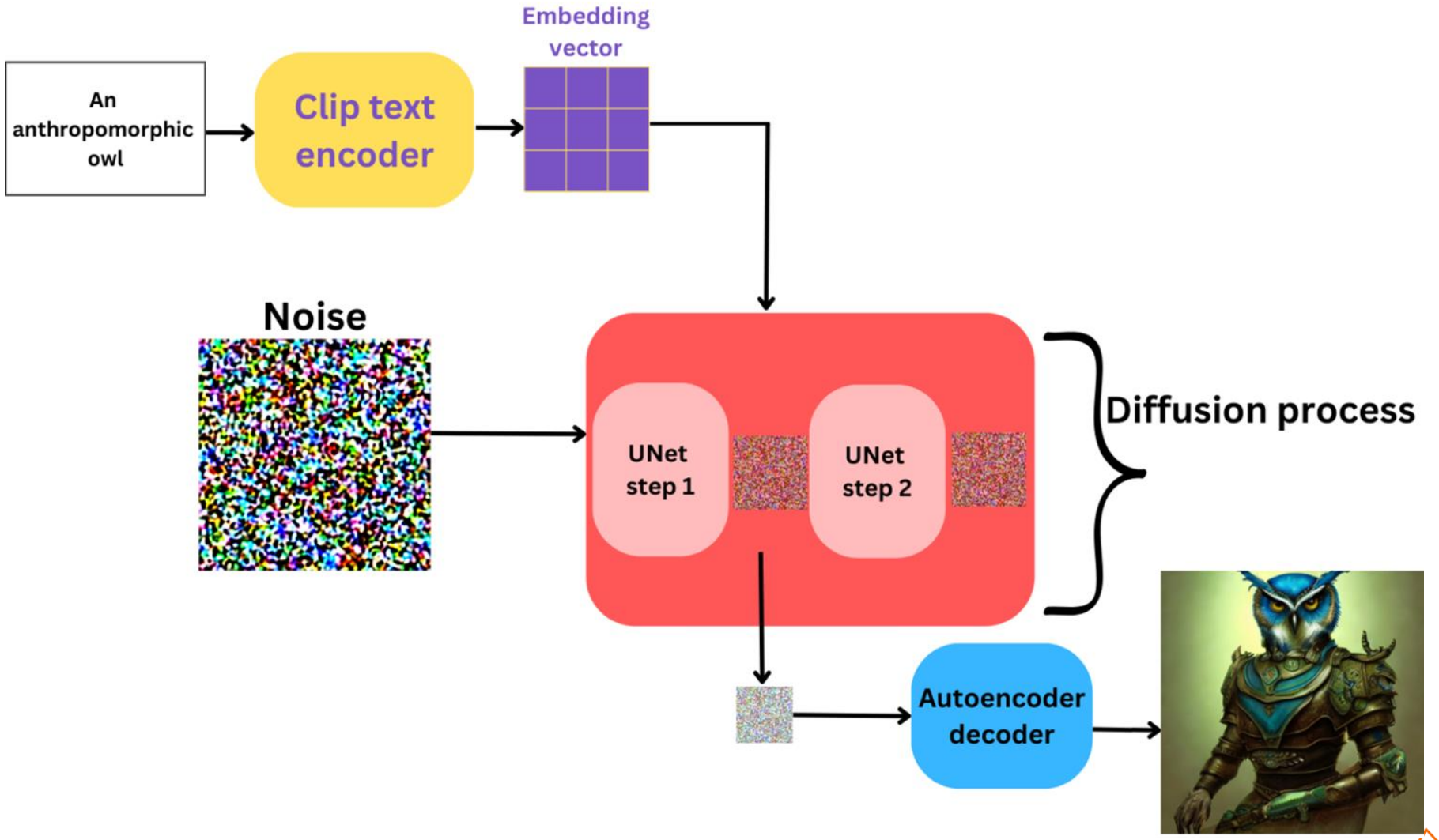
MLP

Product classes:
1.Bolt
2.Screw
3.Bolt with defect
...



Synthetic dataset generation using Generative AI

Stable Diffusion

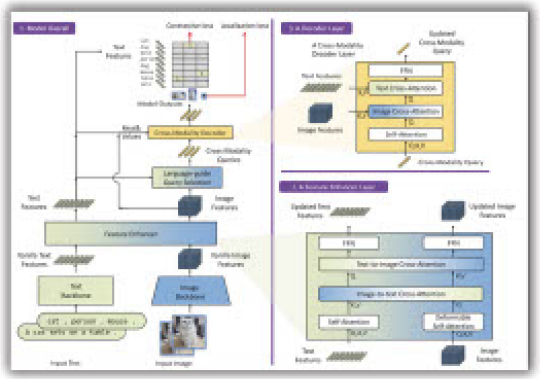


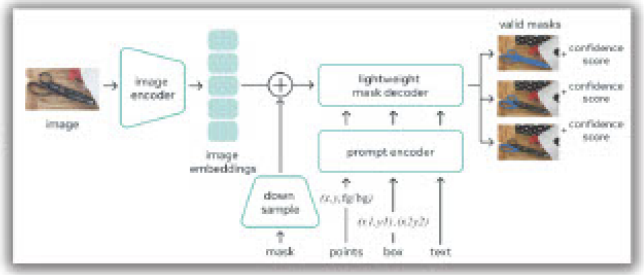
SAM + Stable diffusion

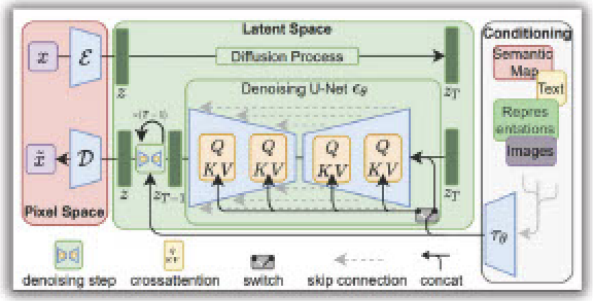
Grounding DINO

Segment-Anything

Stable Diffusion





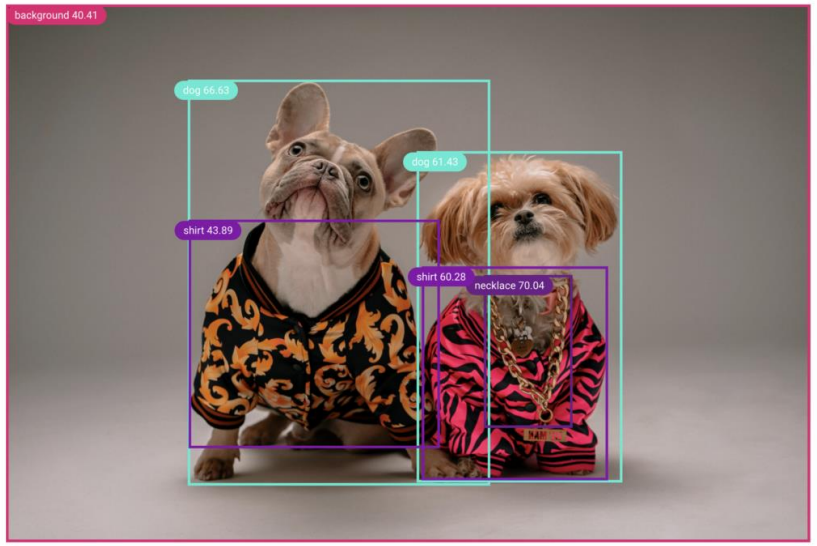


Text Input → Object Detection

Bounding Boxes → Segmentation Masks

Masks + Text Prompts → Generative Inpainting

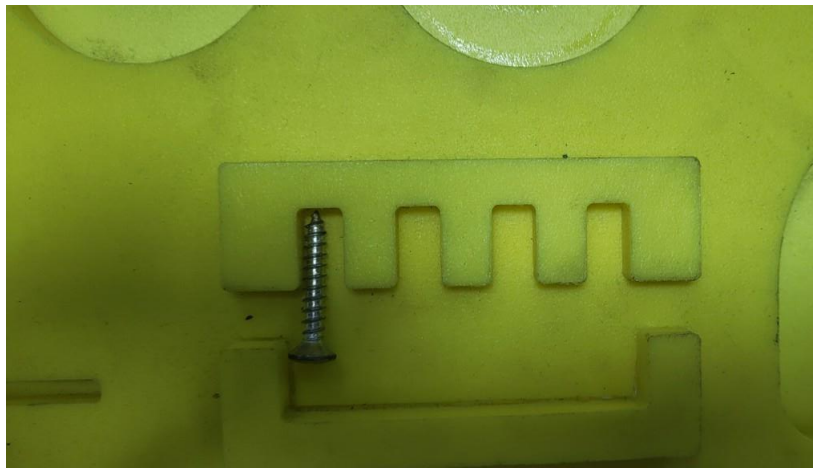
https://www.comet.com/site/blog/sam-stable-diffusion-for-text-to-image-inpainting/

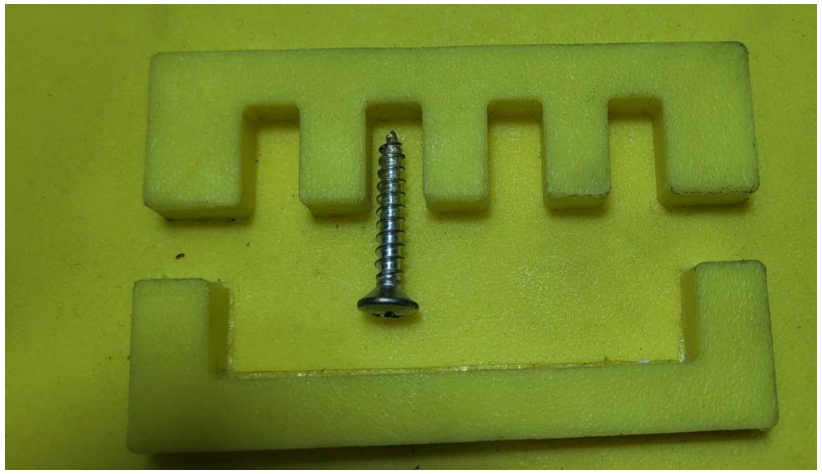


https://www.comet.com/site/blog/sam-stable-diffusion-for-text-to-image-inpainting/

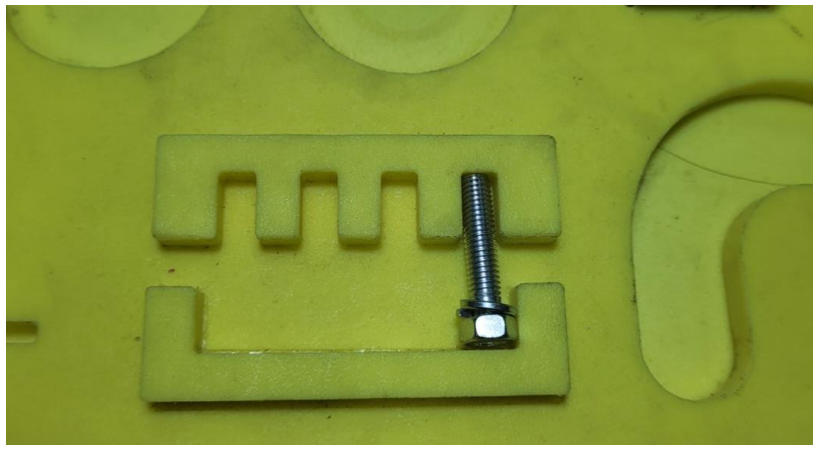
Captured image from real industrial environment

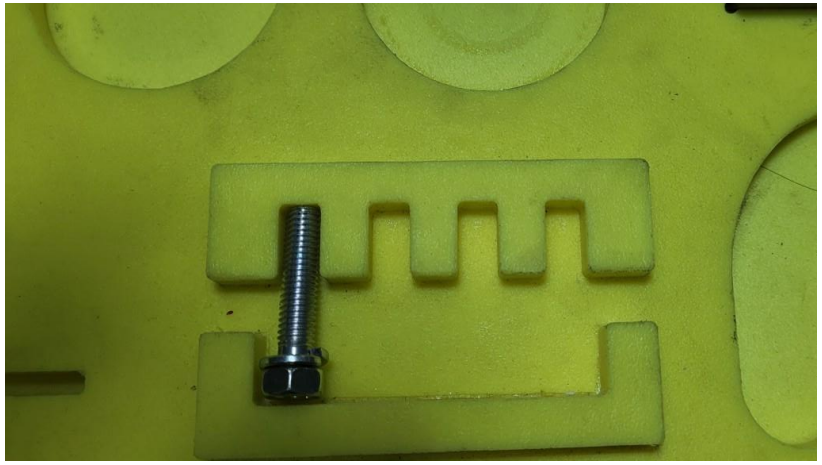
Screw



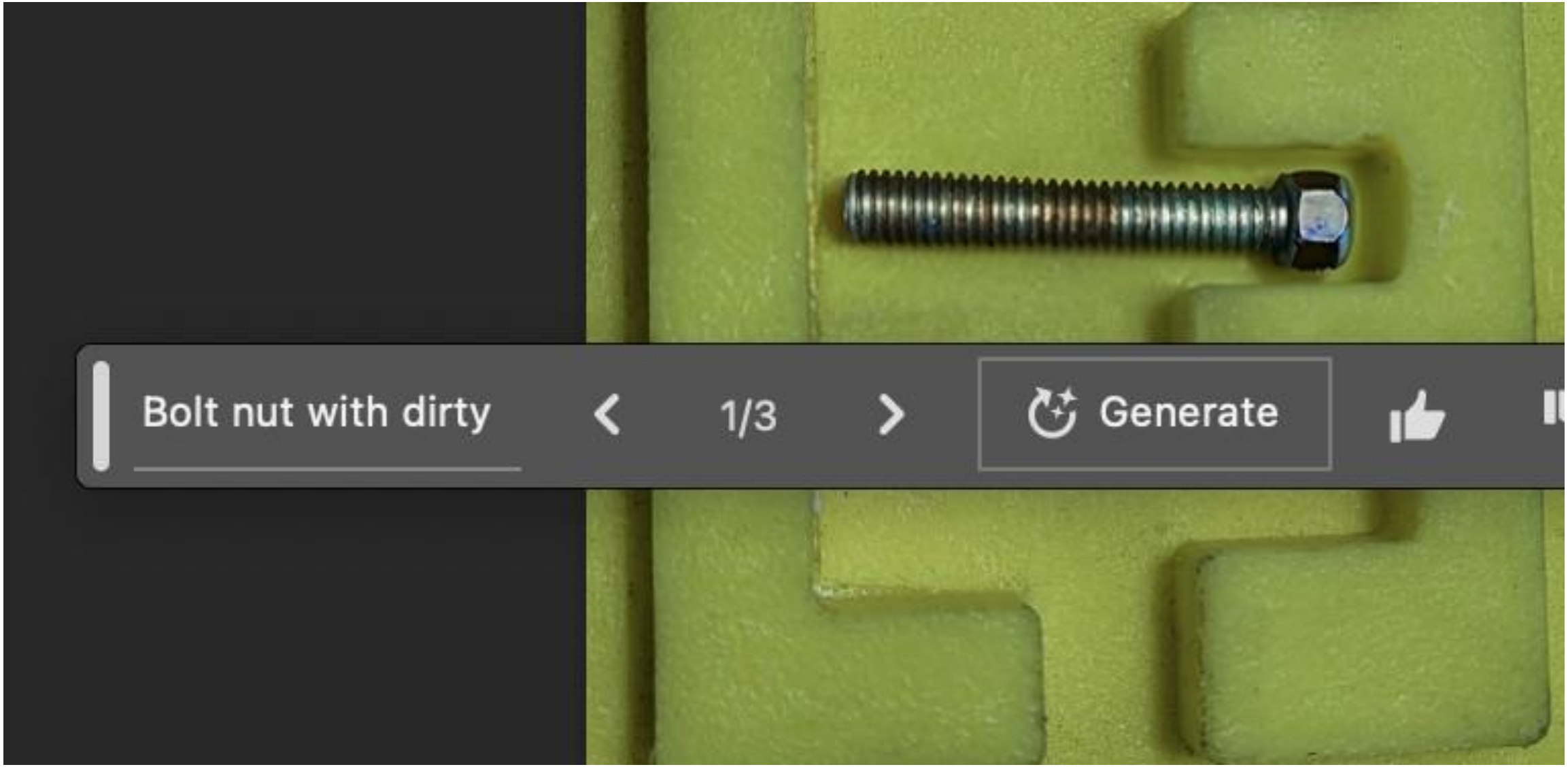


Bolt



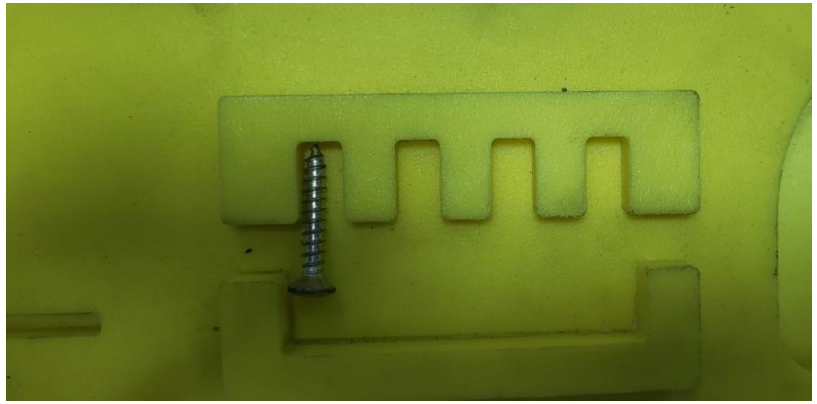


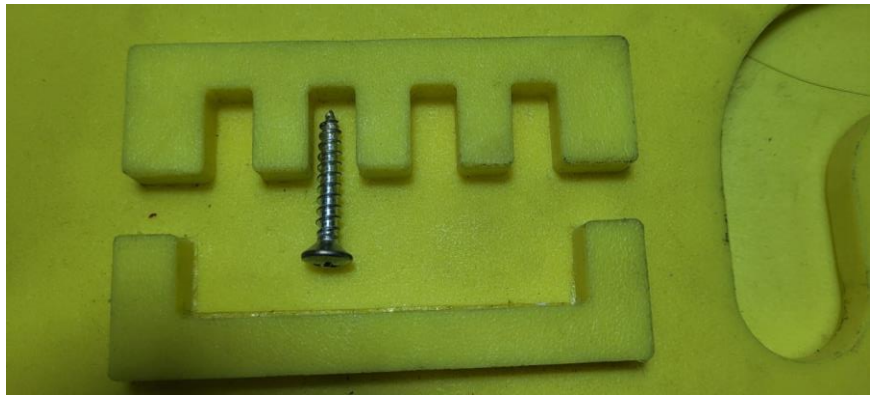
Generative inpainting



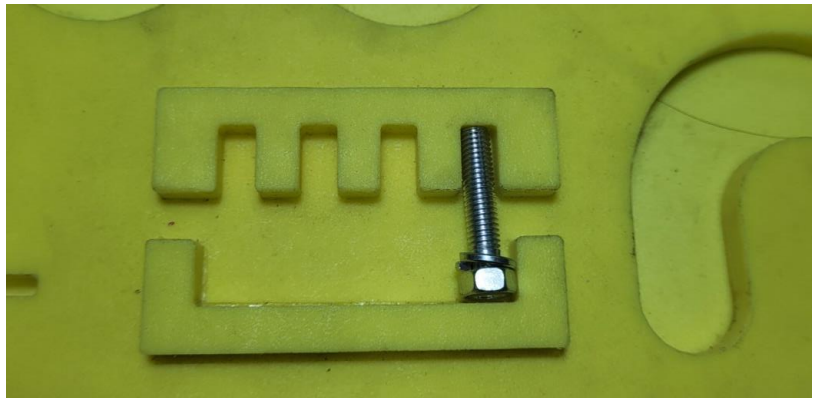
AI Generated images with *no additional* condition compared with real images

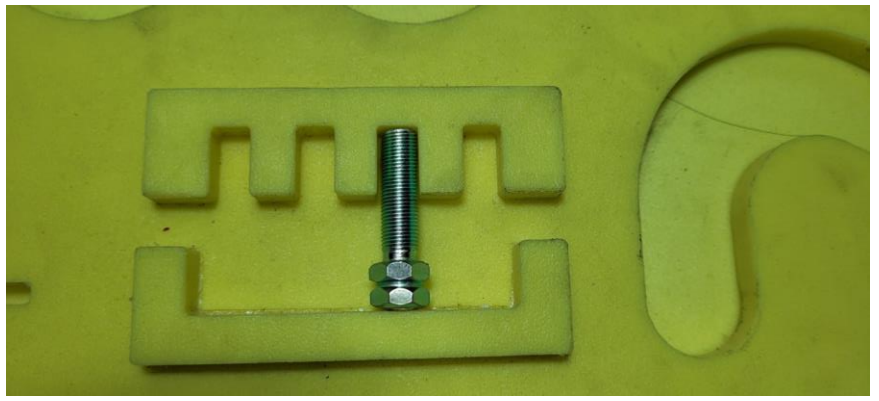
Screw





Bolt



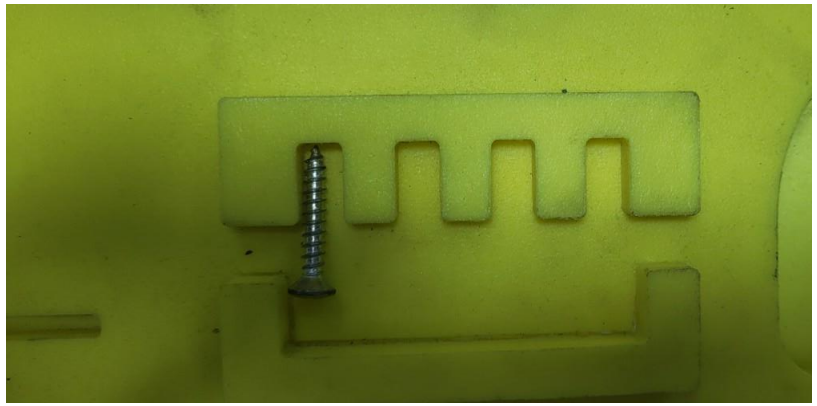


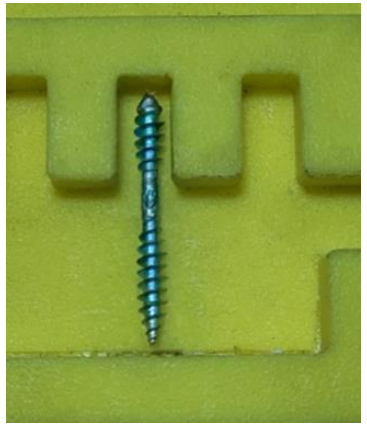
Real

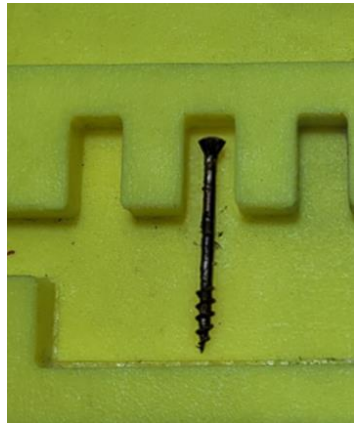
AI Generated

AI Generated images with dirt/defect conditions compared with real images

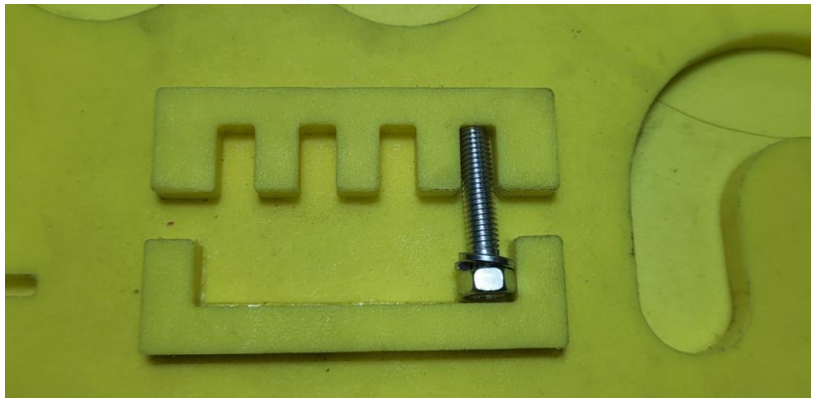
Screw

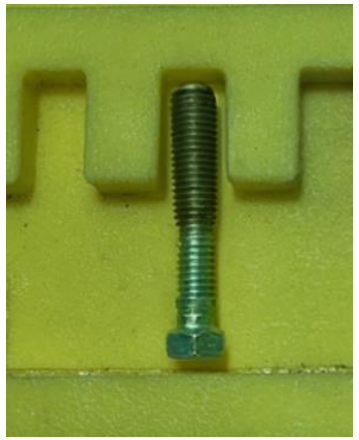


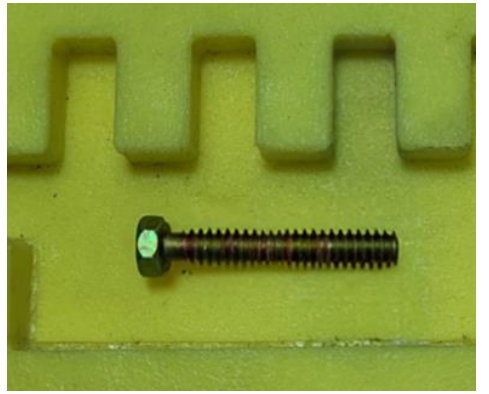




Bolt





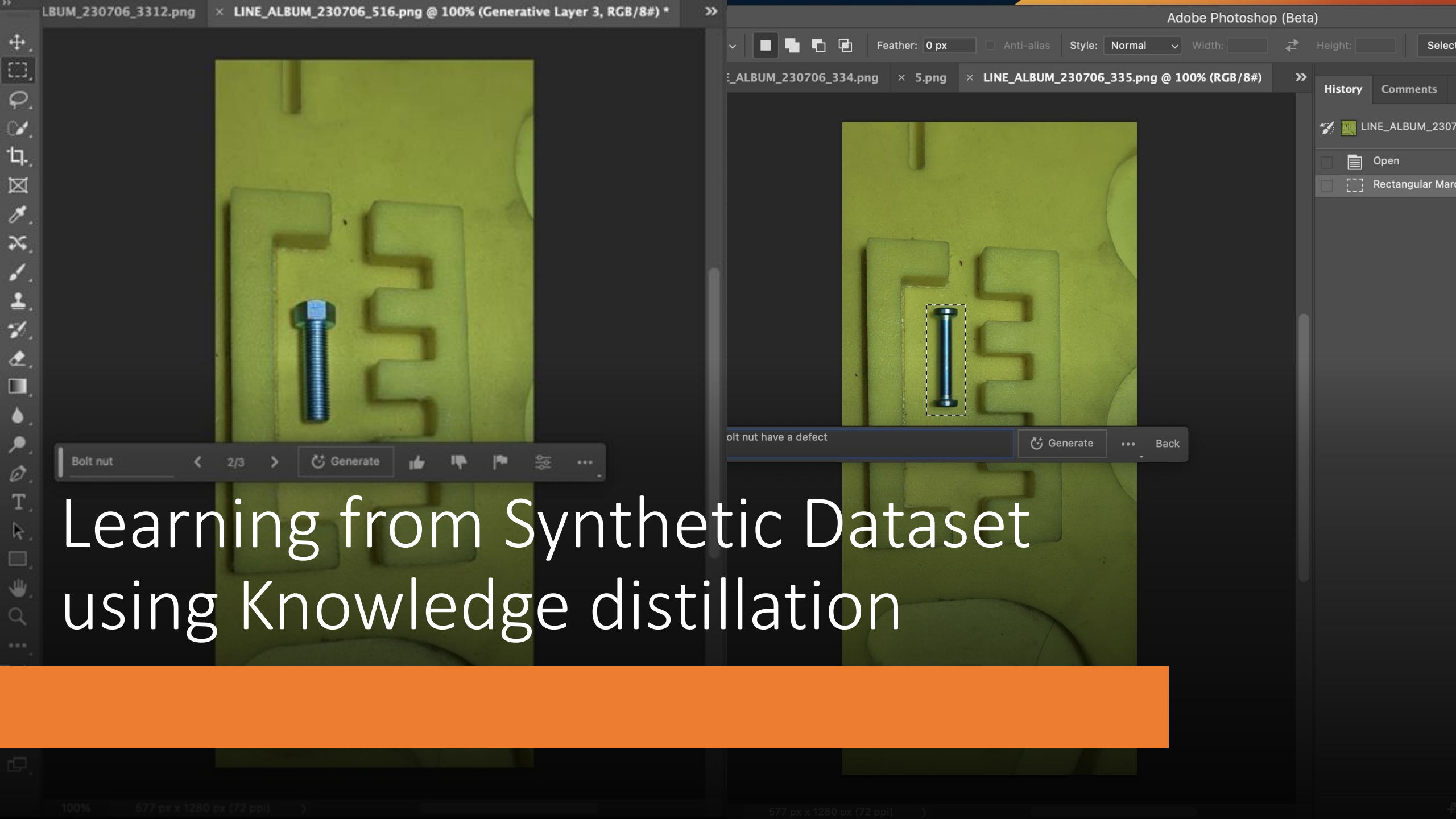


Defect

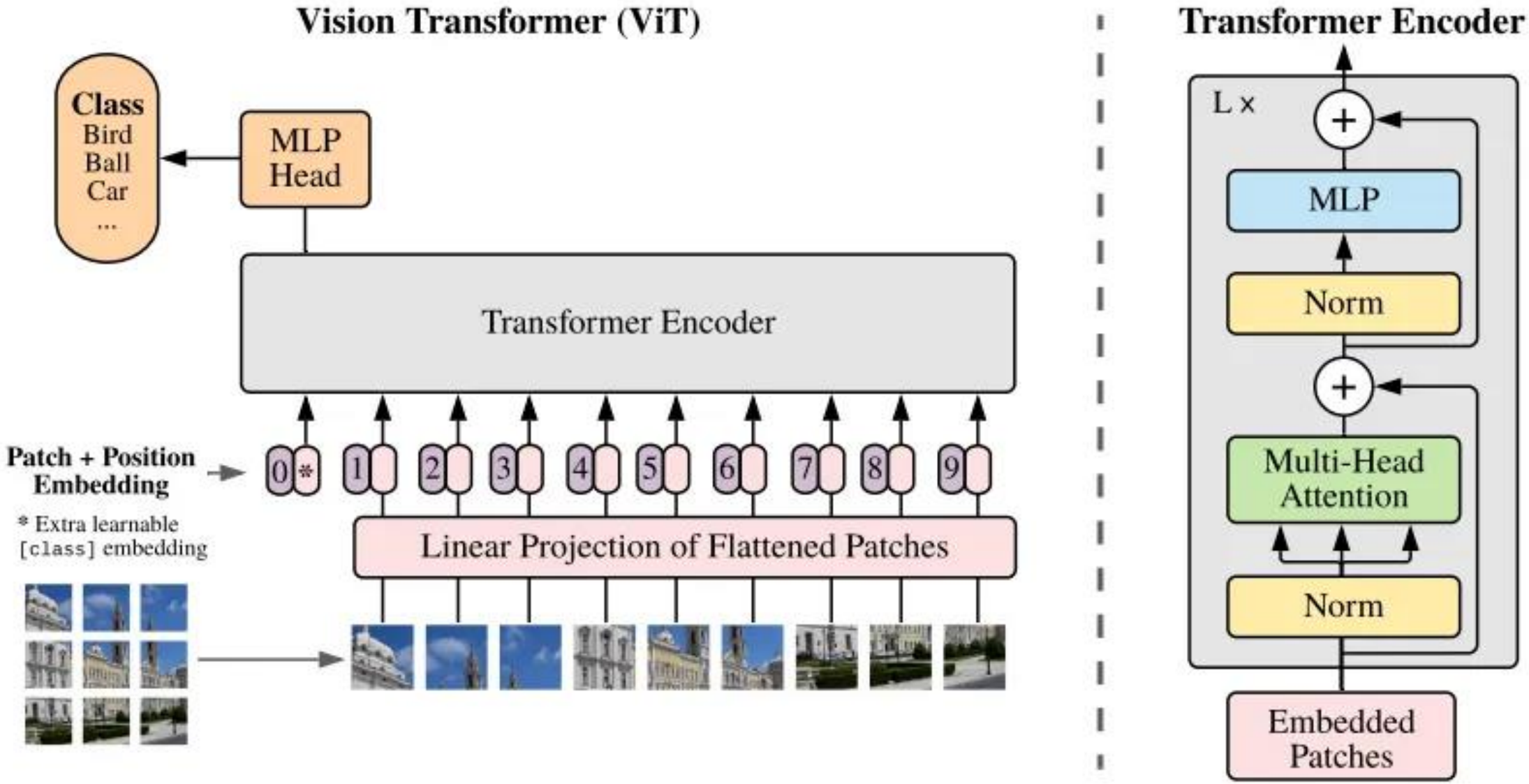
Dirt

Real

AI Generated

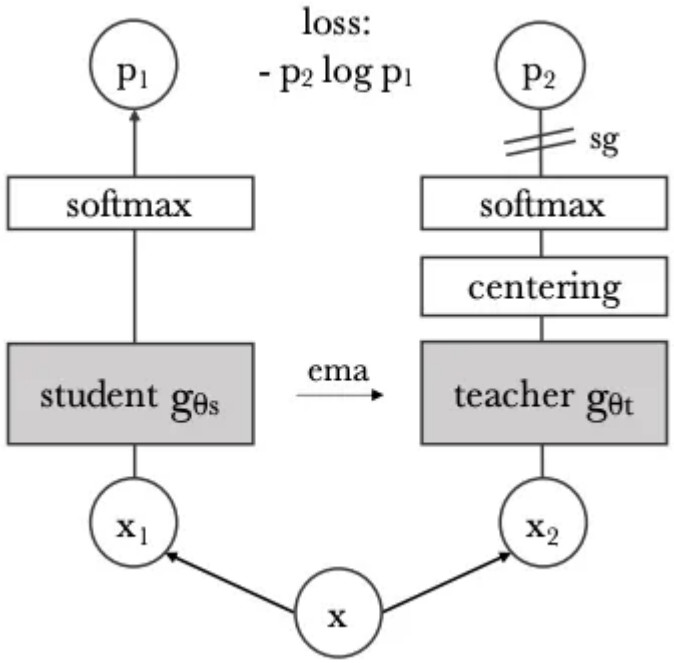


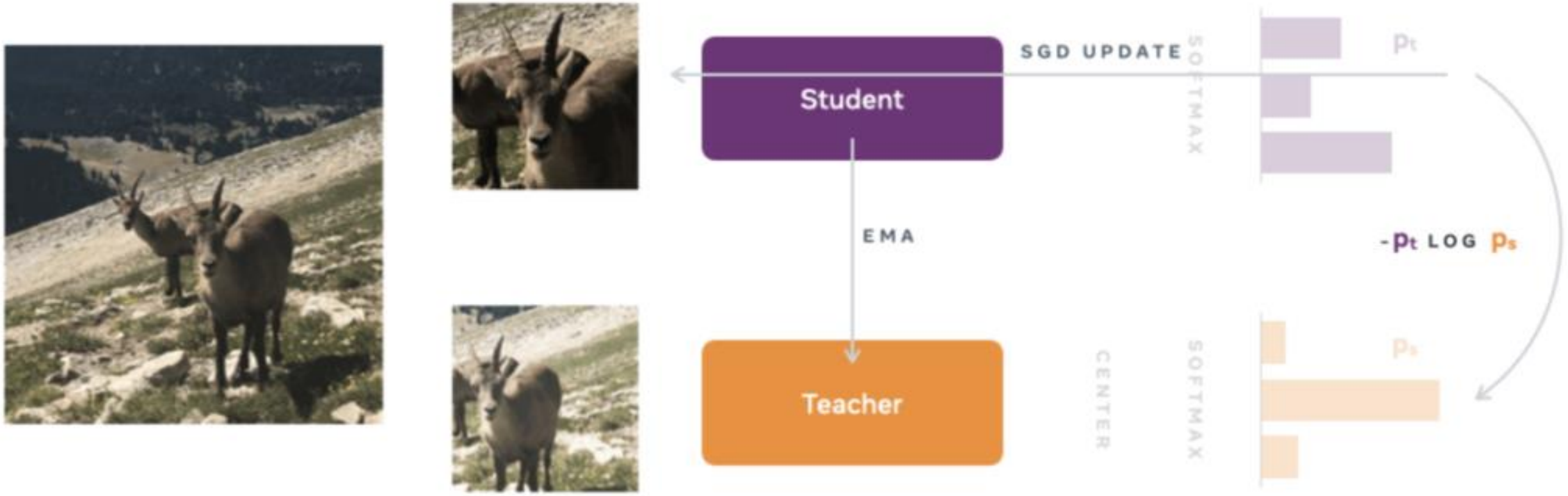
Learning from Synthetic Dataset using Knowledge distillation


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP
Head
Transformer Encoder
Patch + Position
Embedding
* Extra learnable
[class] embedding
0*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head
Attention
Norm
Embedded
Patches

DINO, an acronym for **DI**stillation of knowledge with **NO** labels

- DINOv2 combines elements from **DINOv1 and iBOT for improved performance.**
- Image-level objective: It uses a teacher network and a student network with the same architecture but different parameters.
- The student network **is trained using knowledge distillation to mimic the output of the teacher network**.
- In the first stage of training, global and local views of lower resolution are generated.
- The student network learns local to global correspondences by using all views as input while only the global views are used for the teacher network.
- The student network is **optimized using Stochastic Gradient Descent (SGD) to precisely copy the teacher network**.



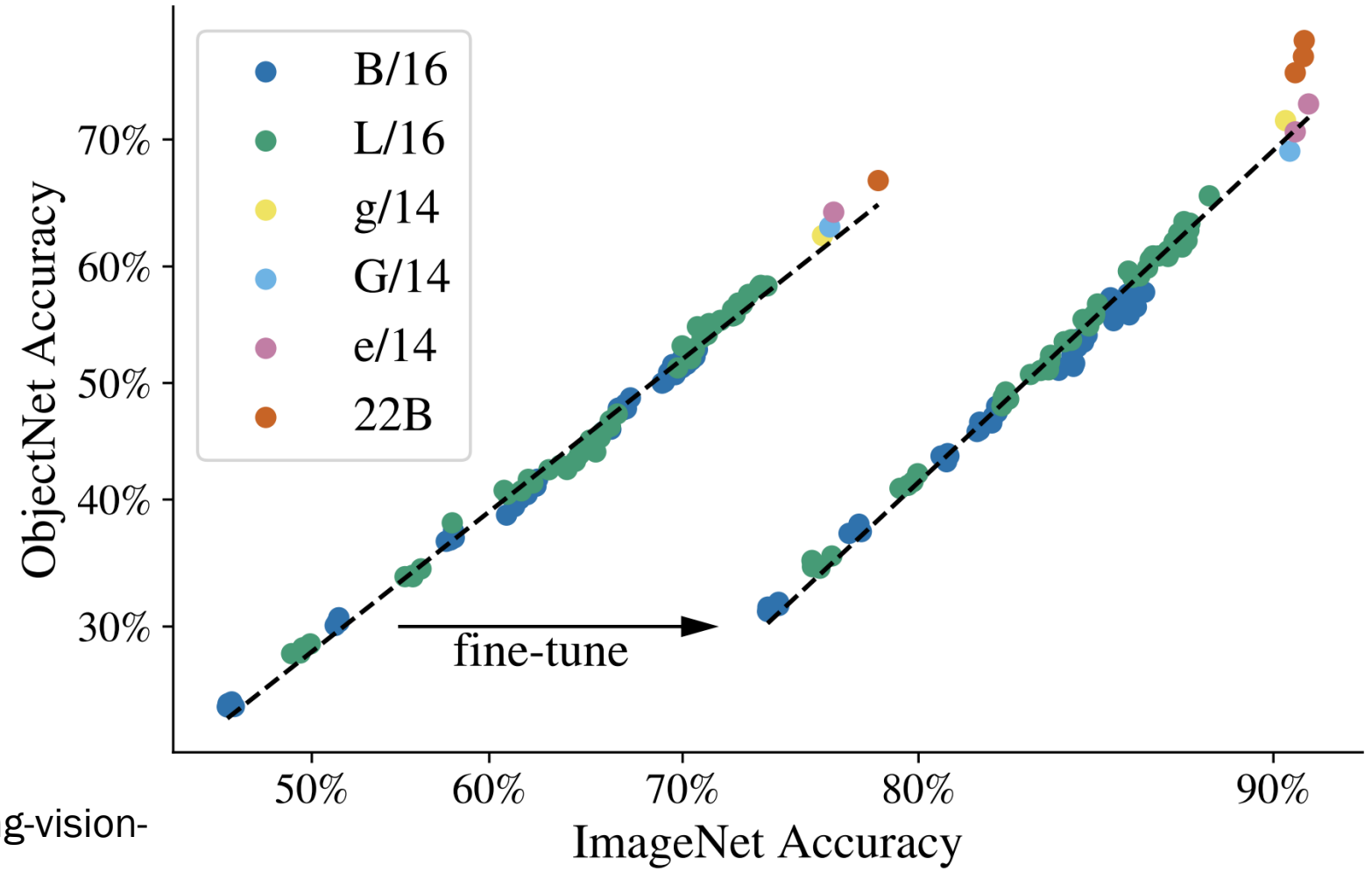


Knowledge Distillation approach, DINOv2 does not use a pre-existing teacher network. Instead, a self-supervised learning method is employed in which the teacher network is constructed from previous iterations of the student network using an exponential moving average (EMA).

https://blog.marvik.ai/2023/05/16/dinov2-exploring-self-supervised-vision-transformers/

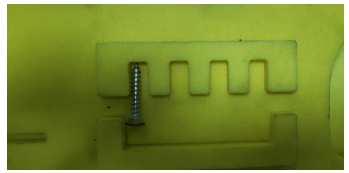
Pretrained models

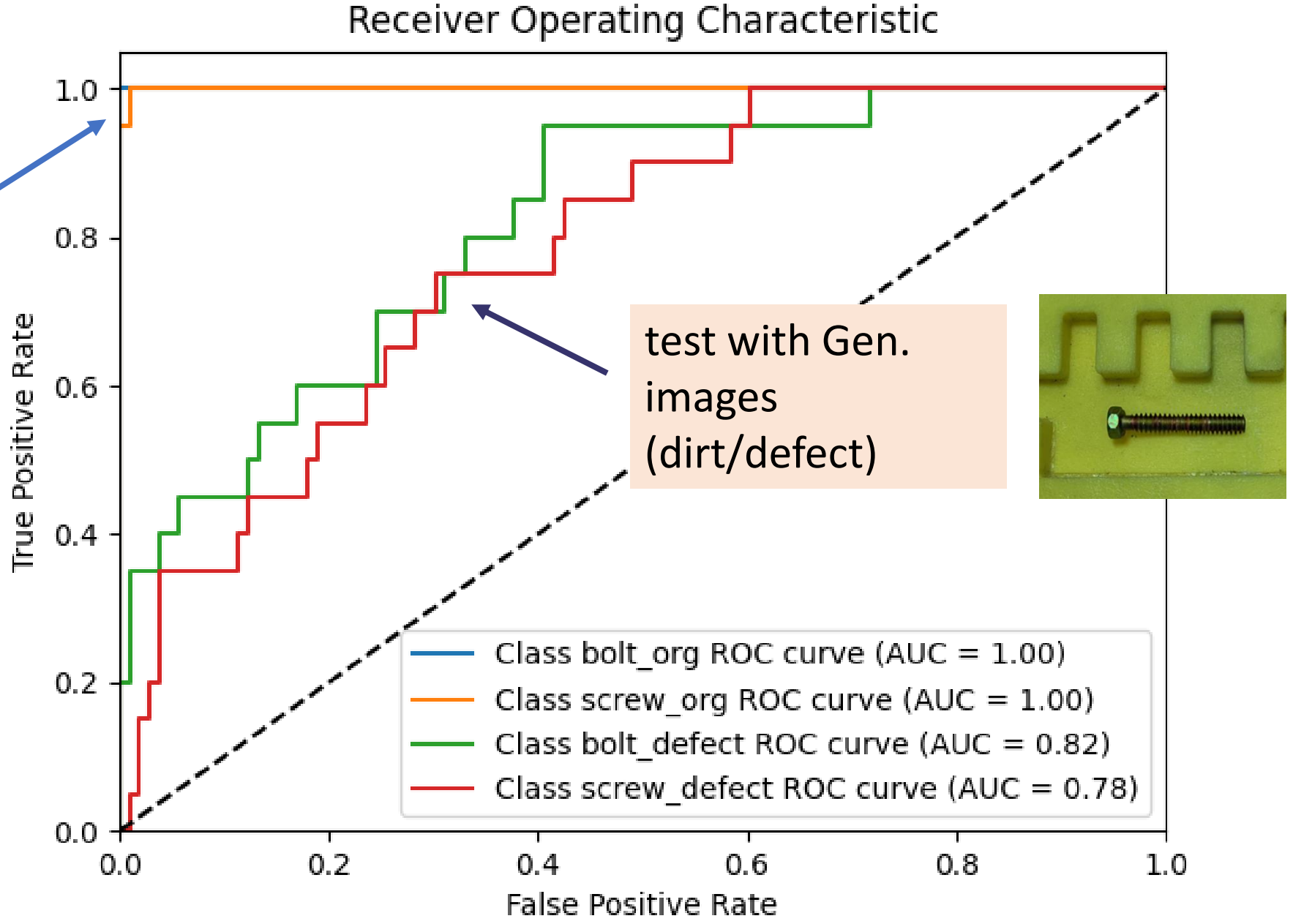
model	# of params	ImageNet k-NN	ImageNet linear
ViT-S/14 distilled	21 M	79.0%	81.1%
ViT-B/14 distilled	86 M	82.1%	84.5%
ViT-L/14 distilled	300 M	83.5%	86.3%
ViT-g/14	1,100 M	83.5%	86.5%



https://ai.googleblog.com/2023/03/scaling-vision-transformers-to-22.html

Fine tuned with 10 captured images and Test with captured images

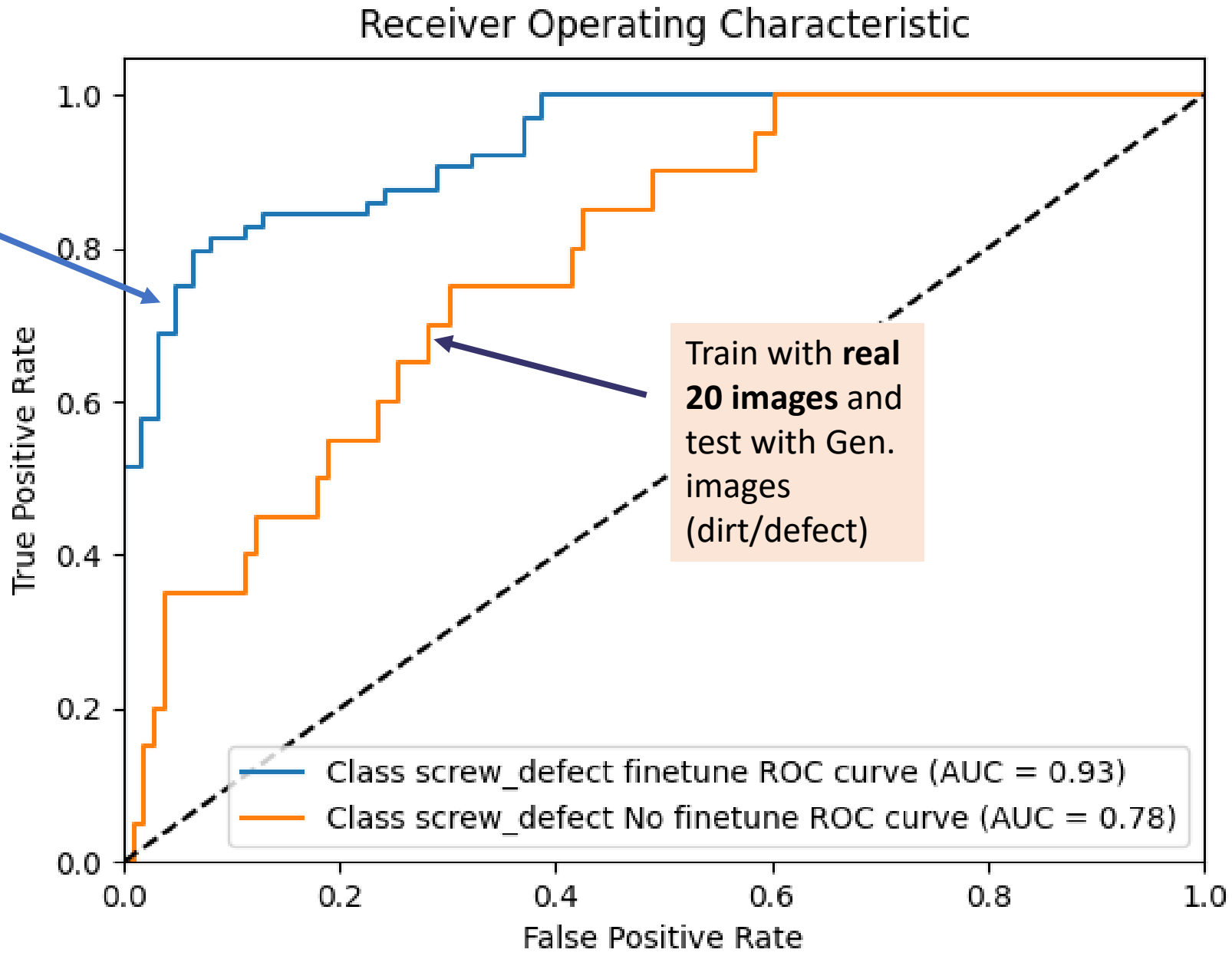




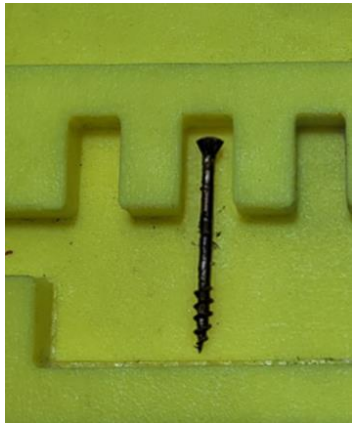
test with Gen. images (dirt/defect)

Fine tuning with 10 images of **Gen. images** (dirt/defect) and

Test with Gen. images (dirt/defect)

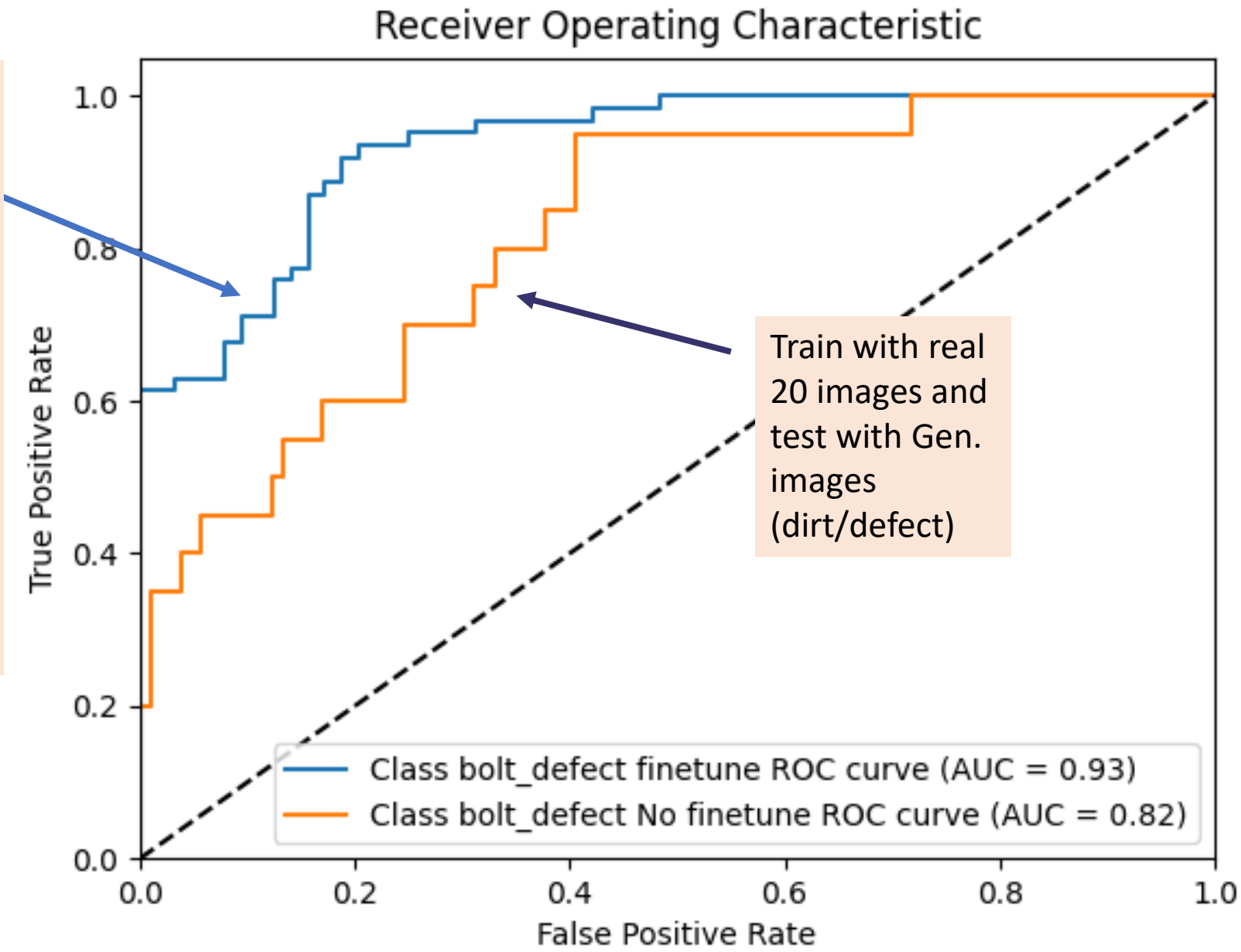


Screw



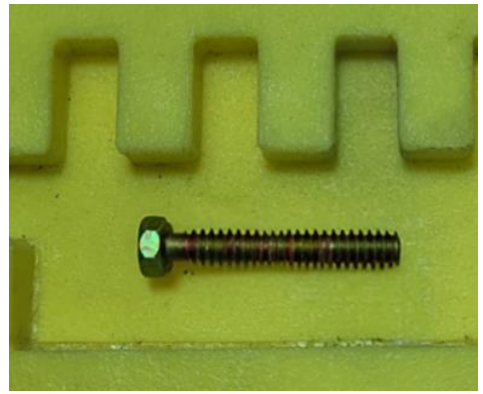
Fine tuning with 10 images of Gen. images (dirt/defect) and

Test with Gen. images (dirt/defect)

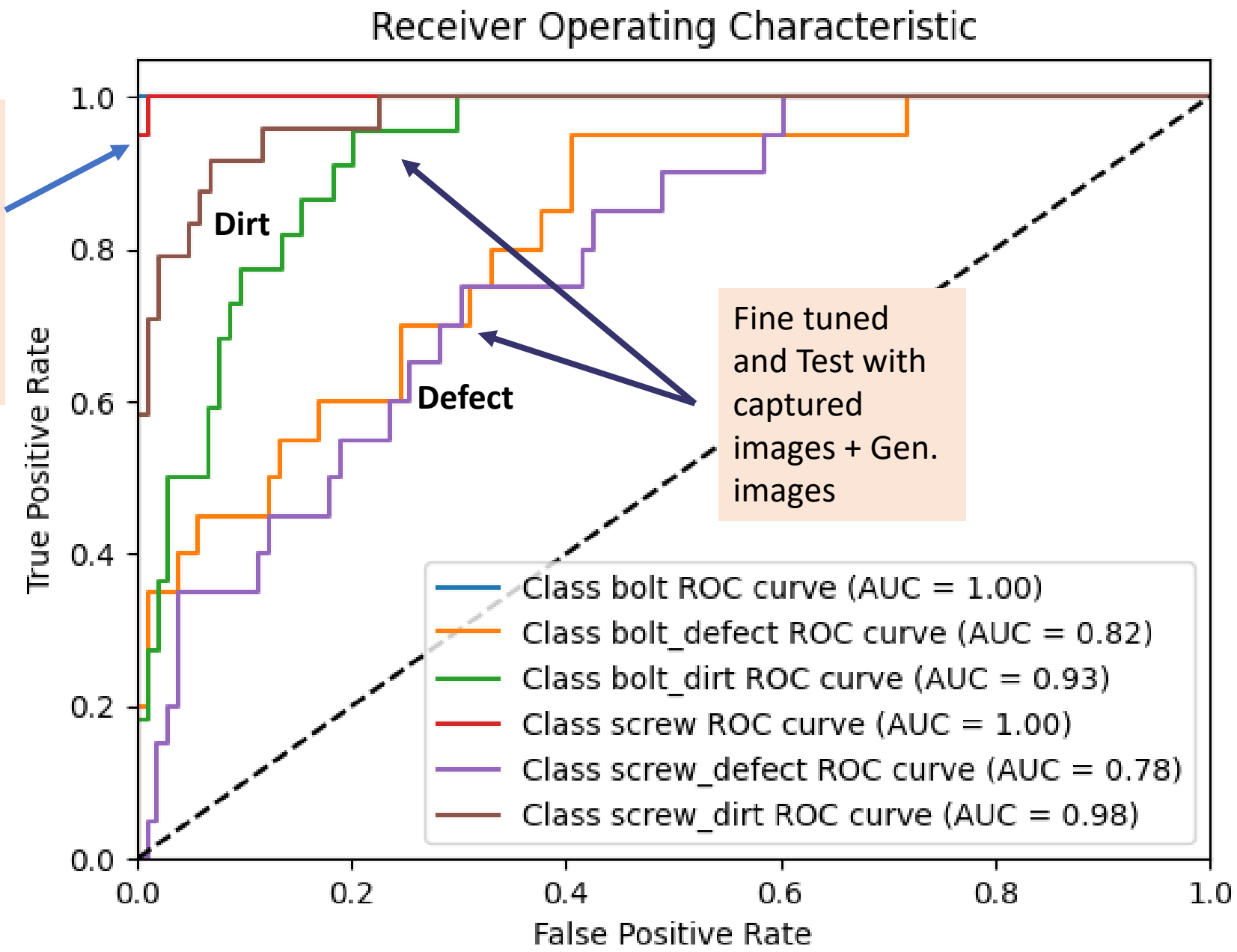


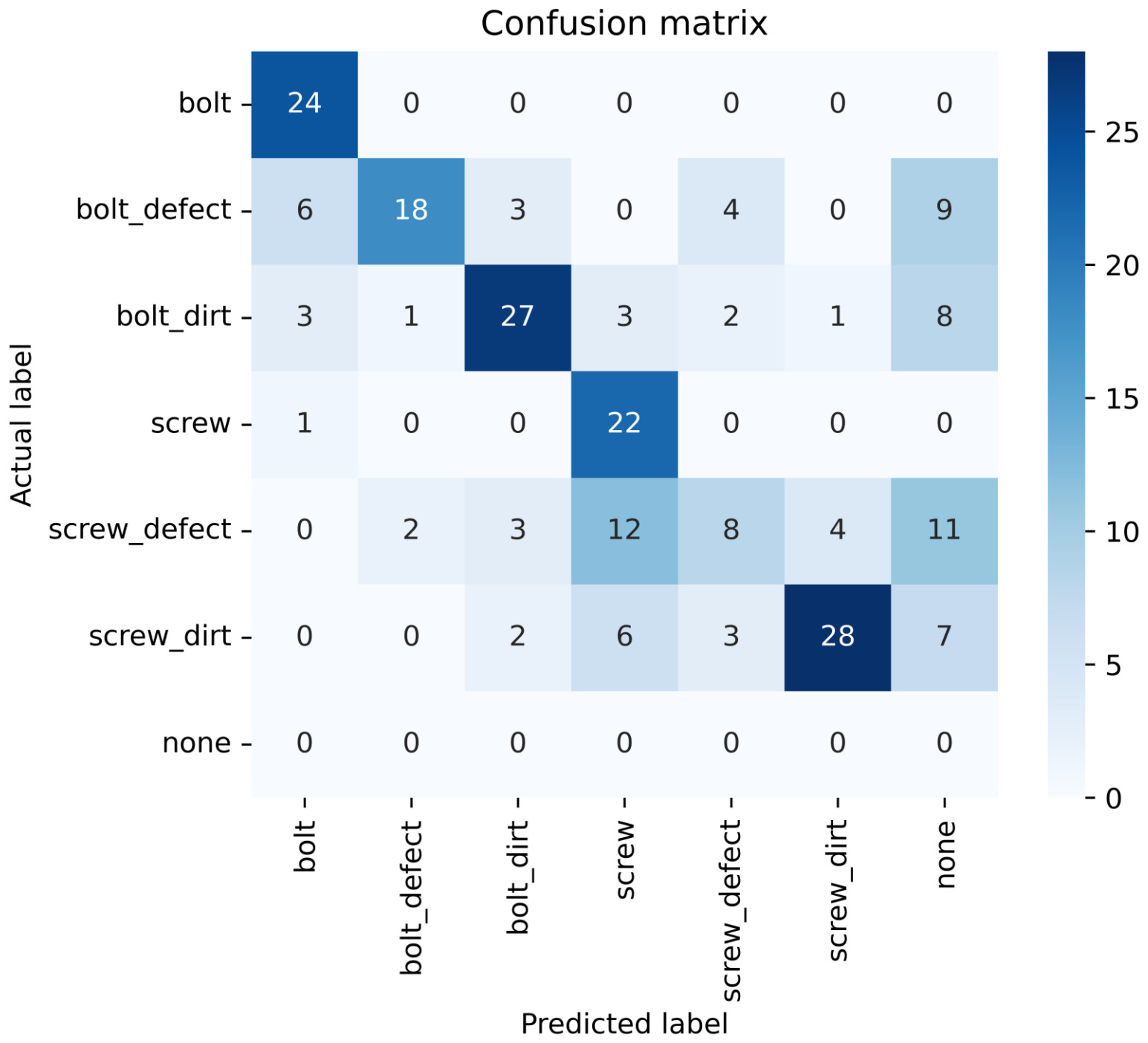
Train with real 20 images and test with Gen. images (dirt/defect)

Bolt



Fine tuned and Test with captured images





Conclusion

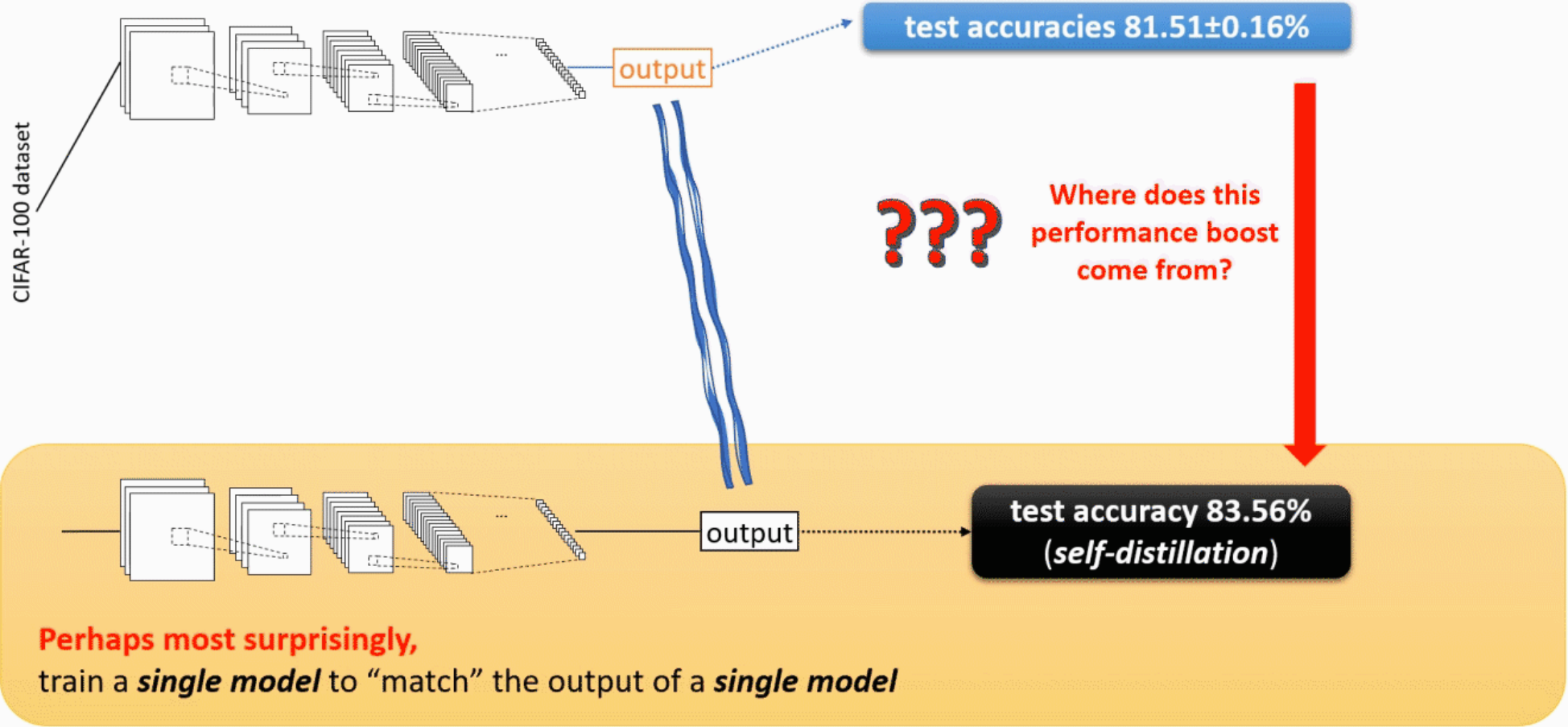
- We have achieved a remarkable feat by showcasing the remarkable capabilities of Generative AI in synthesizing highly targeted datasets for specific industrial use-cases. Through the application of this innovative technique, we have effectively addressed the long-standing challenge of limited availability and scarcity of industrial datasets within the factory environment.
- The Pretrained distilled ViT foundation model serves as an exceptional starting point, as it possesses a comprehensive understanding of various visual features and patterns. This pre-existing knowledge allows us to accelerate the fine-tuning process and optimize the model's performance for detecting and classifying rare industrial defects or contamination instances.

Thank you

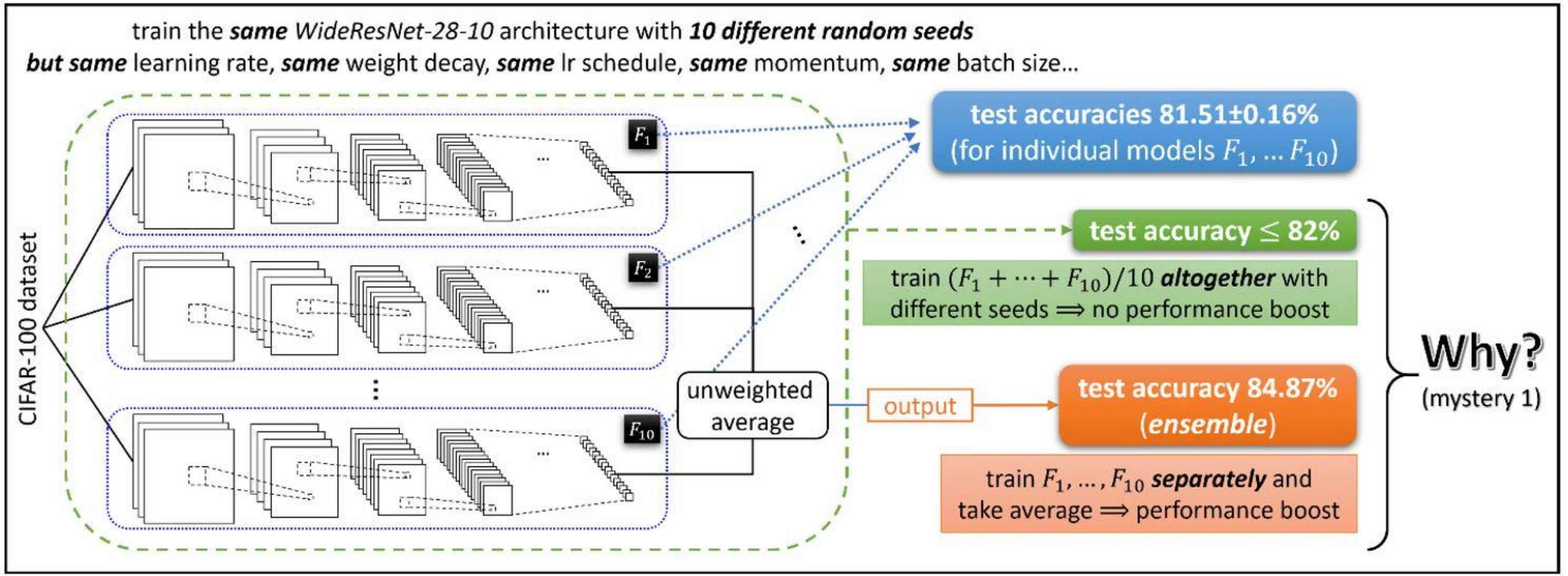
Associate Prof. Dr. Siridech Boonsang

Dean, School of Information Technology, KMITL

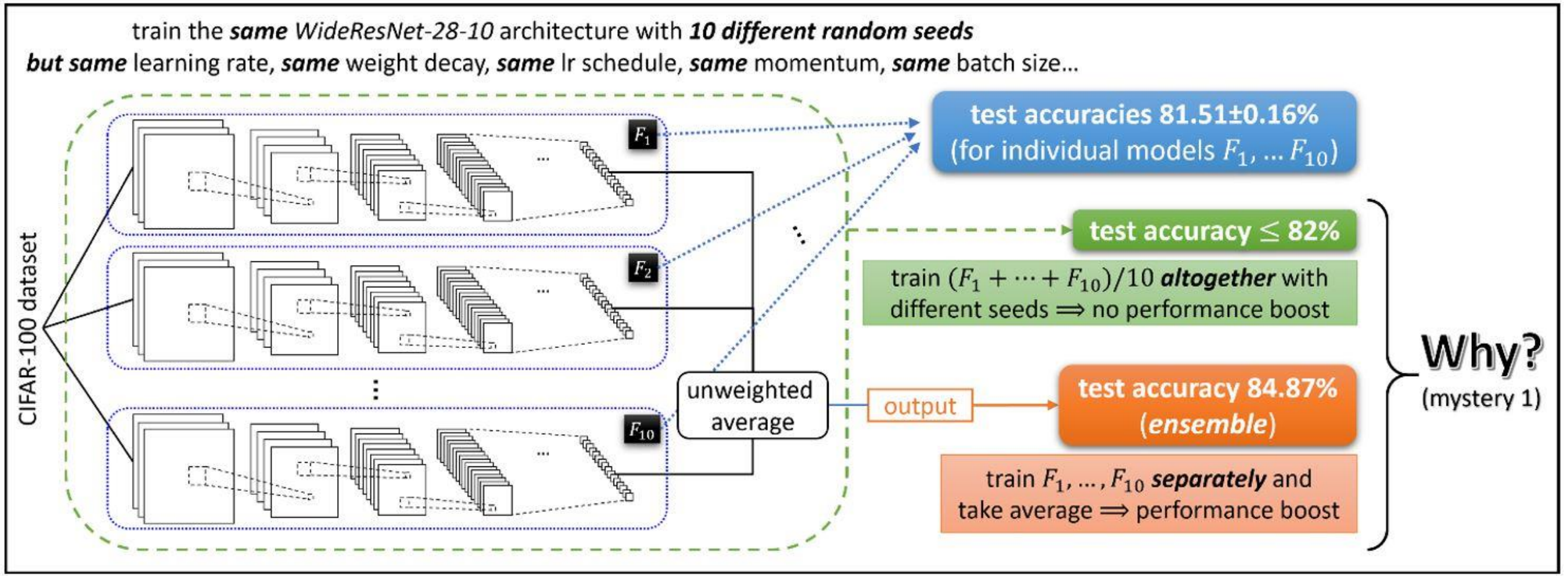
Knowledge distillation.



Ensemble.


train the *same* *WideResNet-28-10* architecture with ***10 different random seeds***
but same learning rate, ***same*** weight decay, ***same*** lr schedule, ***same*** momentum, ***same*** batch size...
CIFAR-100 dataset
F_1
F_2
F_{10}
unweighted average
output
test accuracies 81.51±0.16%
(for individual models $F_1, \dots F_{10}$)
test accuracy ≤ 82%
train $(F_1 + \cdots + F_{10})/10$ ***altogether*** with different seeds ⟹ no performance boost
test accuracy 84.87%
(***ensemble***)
train $F_1, \dots, F_{10}$ ***separately*** and take average ⟹ performance boost
Why?
(mystery 1)

MYSTERY 2: KNOWLEDGE DISTILLATION.



Mystery 2: Knowledge distillation.

- Ensemble models improve test-time performance but become 10 times slower during inference due to the need to compute outputs from multiple neural networks.

- Knowledge distillation is a technique proposed to address this issue by training another individual model to match the output of the ensemble.

- Knowledge distillation involves matching the ensemble's output, also known as **"dark knowledge,"** which may include probabilities for multiple classes, with the true training label.

Mystery 2: Knowledge distillation.

- The individual model trained through knowledge distillation can achieve similar test-time performance as the larger ensemble.

- Matching the outputs of the ensemble during knowledge distillation leads to better test accuracy compared to matching the true labels, although the reasons for this improvement are not fully understood.

- It is possible to perform ensemble learning over the models trained through knowledge distillation to further enhance test accuracy.

MYSTERY 2: KNOWLEDGE DISTILLATION.

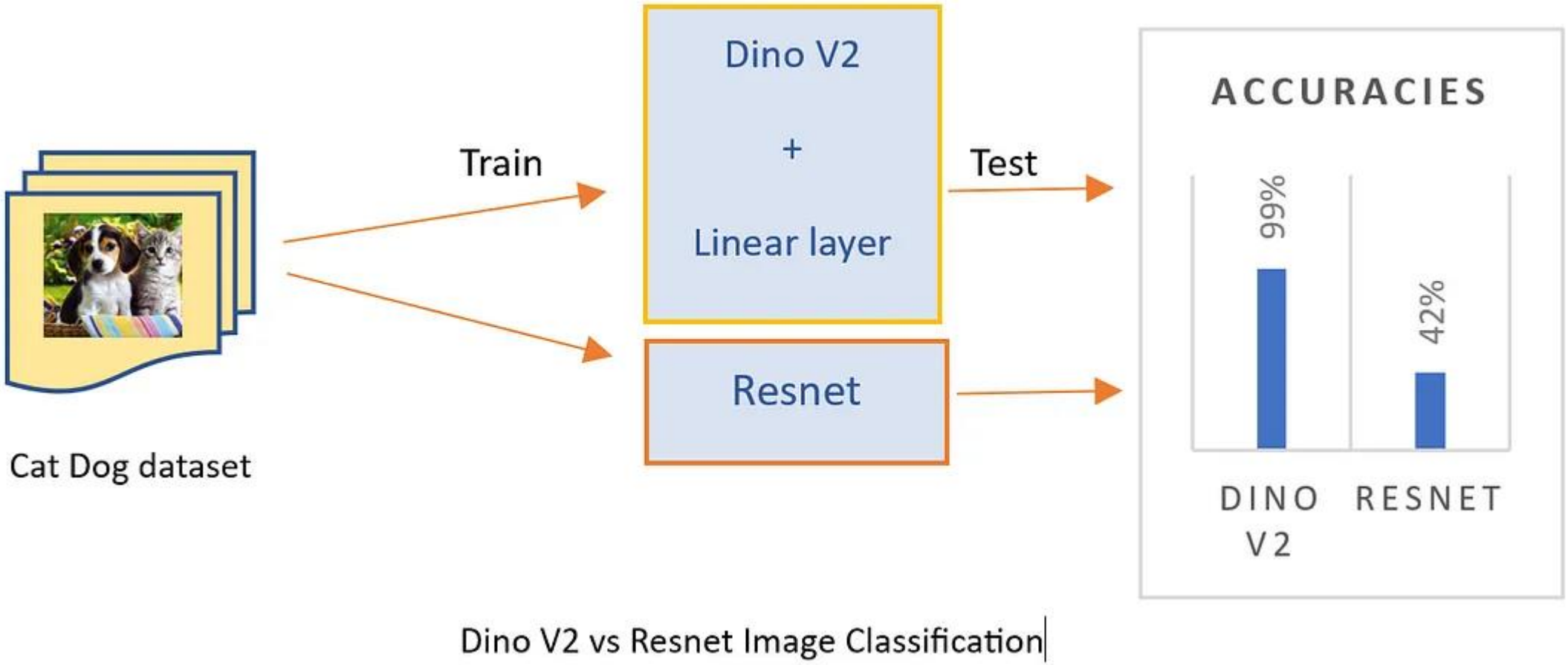
- knowledge distillationwas proposed to address a problem with ensemble models in deep learning.
- Ensemble models are made up of multiple individual models that work together to make predictions. However, these models can be computationally expensive and difficult to train.
- Knowledge distillation involves training another individual model to match the output of the ensemble. This model is often called the student model, while the ensemble is called the teacher model.
- The output of the ensemble on a given input (such as an image of a cat) is sometimes referred to as the dark knowledge. This output may include probabilities for different classes, such as "80% cat + 10% dog + 10% car."
- The true training label for the input is known, such as "100% cat." The goal of knowledge distillation is to train the student model to match the dark knowledge output as closely as possible, while still predicting the correct label.
- The student model can be much smaller and faster than the ensemble, but still achieve similar performance at test time. This is because it has learned to mimic the behavior of the ensemble, which has already learned to recognize patterns in the data.
- Overall, knowledge distillation is a useful technique for reducing the computational cost of ensemble models, while still maintaining their accuracy.

MYSTERY 3: SELF-DISTILLATION.

- The concept of knowledge distillation is introduced, which involves using a teacher ensemble model to improve the performance of a student individual model.
- The teacher ensemble model has a test accuracy of 84.8%, which means that the student individual model can achieve 83.8% accuracy by learning from it.
- The phenomenon of self-distillation is then introduced, which involves using an individual model of the same architecture as the teacher to improve its own performance.
- This is surprising because if an individual model only achieves 81.5% test accuracy, it is not clear how using the same model as a teacher can consistently boost the accuracy to 83.5%.
- The image in Figure 2 provides a visual representation of this process, which involves training the same model again using itself as the teacher.
- The concept of self-distillation is important because it provides a way to improve the performance of individual models without relying on an external teacher ensemble model.
- This is particularly useful in scenarios where it may not be feasible to use a teacher ensemble model, such as in resource-constrained environments.
- The mystery of how self-distillation works is still not fully understood, and further research is needed to uncover the underlying mechanisms behind this phenomenon.

Vision Transformer

- The input image is divided into patches of size P×P
- Each patch is flattened and transformed into a D-dimensional vector.
- Position embeddings from the original transformer and class tokens are added to the patch embeddings.
- The position is represented as a single number instead of a 2D position embedding based on x, y positions.
- The patches are converted into tokens.
- The token input is treated the same as regular NLP tasks.
- No modifications are made to the encoder transformer model.



Dino V2 vs Resnet Image Classification

https://purnasaigudikandula.medium.com/dinov2-image-classification-visualization-and-paper-review-745bee52c826