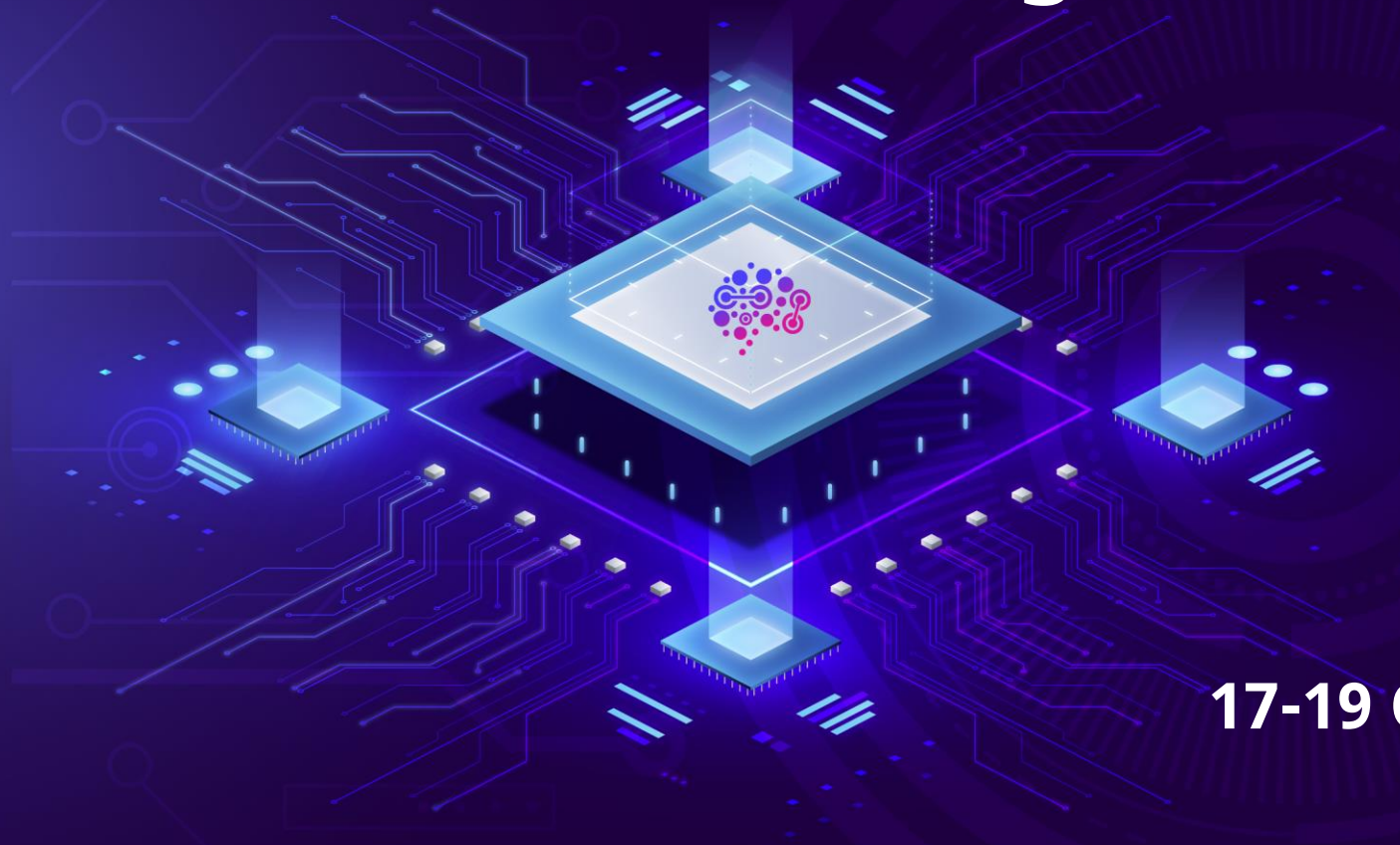
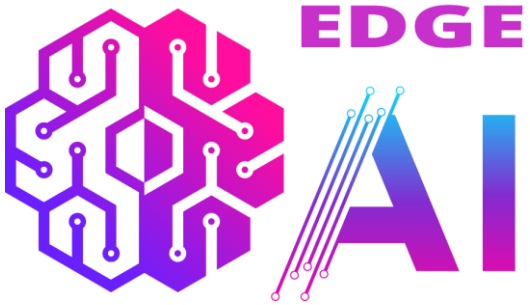
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI



Optimization and Compression of Computer Vision Models using Generative AI for Neuromorphic Accelerators

Tim Llewellynn, Bonseyes Community Association
Switzerland

17-19 October 2023 Athens, Greece

Copyright © 2023

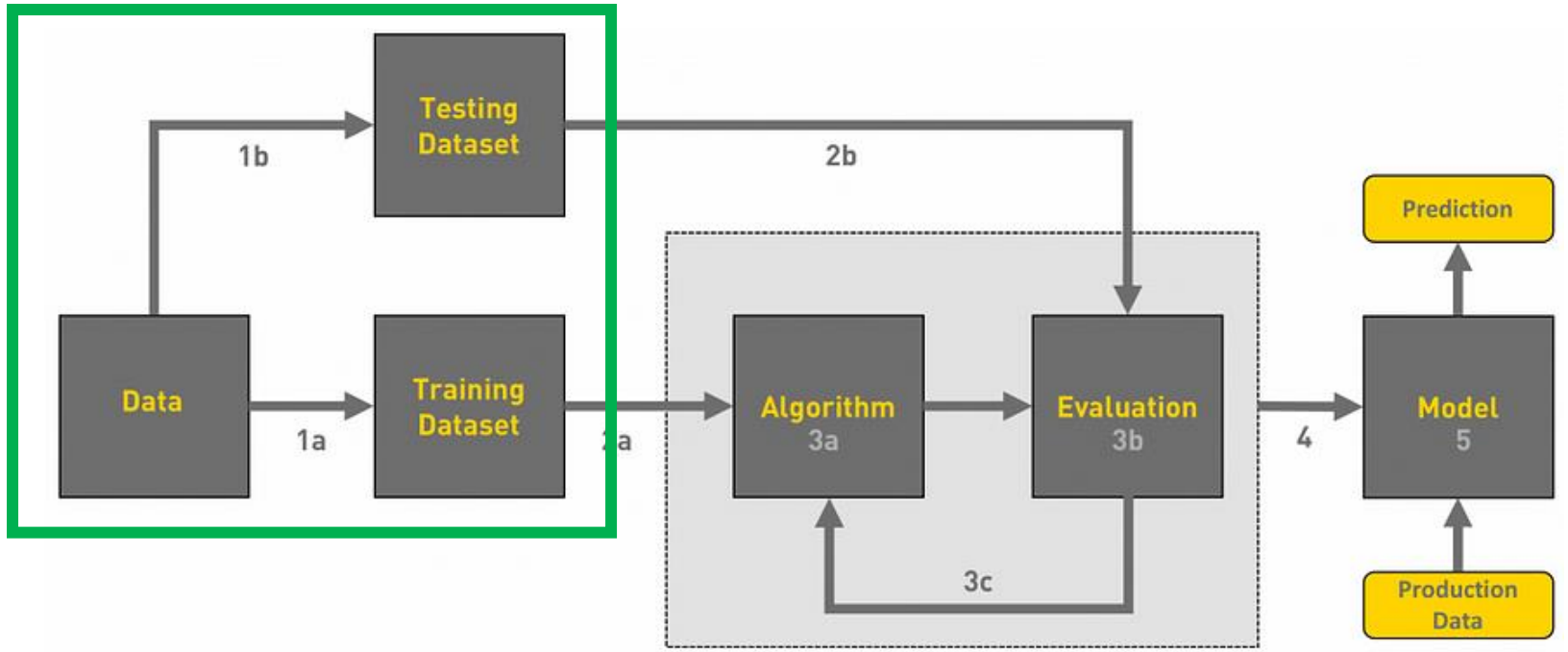
Presentation Outline

- Introduction
- State-of-the-art
- Research Goal
- Methodology
- Design and Implementation
- Experimental Results
- Future Research
- Conclusion

Copyright © 2023

Introduction

Where to Invest my Limited Resources?



- **Data-centric engineering** methods on the Data step of a classical ML workflow have typically been overlooked due to large manual data collection and labeling required.
- **Generative AI** models can alleviate this constraint, by automating the training and inference data workflows.
- Use cases where data is often scarce and costly to collect, such as **Edge AI applications**, stand to benefit greatly.

Copyright © 2023

State-of-the-art: Generative Models for Training Data

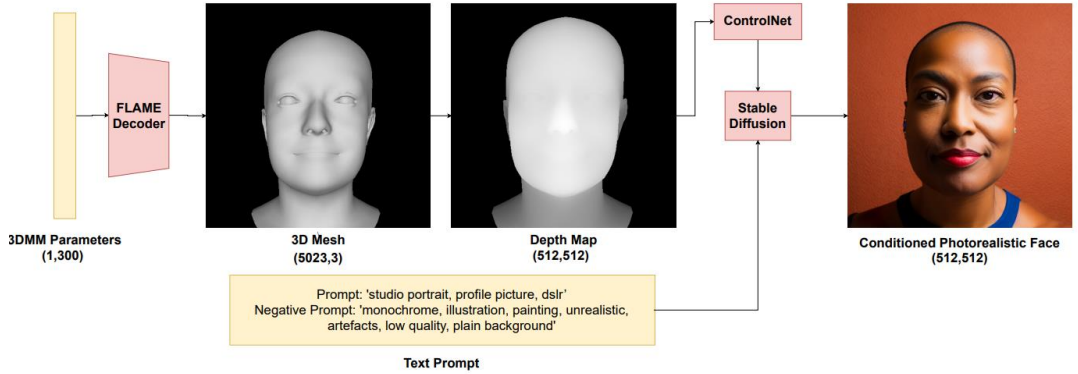
Fake It Without Making It: Conditioned Face Generation for Accurate 3D Face Shape Estimation
2307.13639.pdf (arxiv.org)

250K photorealistic images and corresponding 3DMM parameters.

Method	Median	Mean	Std
Deep3D [7]	1.286	1.864	2.361
DECA (detail)	1.19	1.469	1.249
DECA [10]	1.178	1.464	1.253
AlbedoGAN (detail)	0.95	1.173	0.987
MICA [39]	0.913	1.130	0.948
AlbedoGAN [19]	0.903	1.122	0.957
ControlFace (ours)	1.192	1.472	1.222

Table 1. Reconstruction error (mm) on the validation set of the NoW benchmark [25] in non-metrical reconstruction. Comparison results are presented from [19].

Backbone: ArcFace
Generative AI: SD1.5 + ControlNet



Revisiting DETR Pre-training for Object Detection
2308.01300.pdf (arxiv.org)

2.3M generated images from SDXL for training transformer based object detector. Outperforms Object365 in-the-wild dataset when training object detection tasks on COCO dataset.

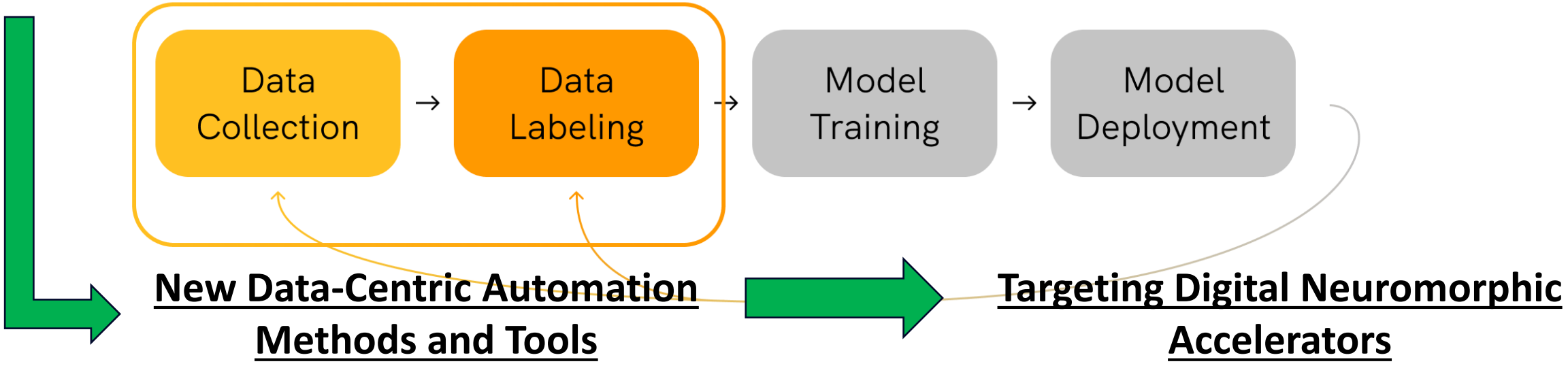
Backbone: ResNet50
Generative AI: SDXL + LLaVA Captions

Caption	Gernerative Models	Pretraining Dataset	COCO		
			AP	AP_{50}	AP_{75}
-	-	O365	52.8	70.9	57.6
COCO captions	ControlNet	Control-COCO 2M	51.1	69.2	55.8
COCO captions	ControlNet	Control-COCO 2M	52.6	70.6	57.5
LLaVA captions	ControlNet	LLaVAControl-COCO 2M	50.7	69.6	55.4
LLaVA captions	ControlNet	LLaVAControl-COCO 2M	**52.9**	70.8	57.9
COCO captions	SDXL	SDXL-COCO 2M	52.5	70.7	57.3
LLaVA captions	SDXL	SDXL-COCO 2M	**52.9**	**71.0**	**58.0**

Copyright © 2023

Research Goal

- Our primary research objective is to leverage data-centric automation pipelines to enhance both the **quality and quantity of training data**, aiming to **significantly reduce** the associated **time and costs** for Edge AI development. Concurrently, we aspire to rectify the shortcomings inherent in current data techniques by utilizing generative models. This dual approach not only **aims for improved accuracy and efficiency** but also **addresses concerns** related to **bias and ethical considerations** in data handling and model training.



Copyright © 2023

Methodology: End-to-End Workflows

1. Automation of Data Pipelines

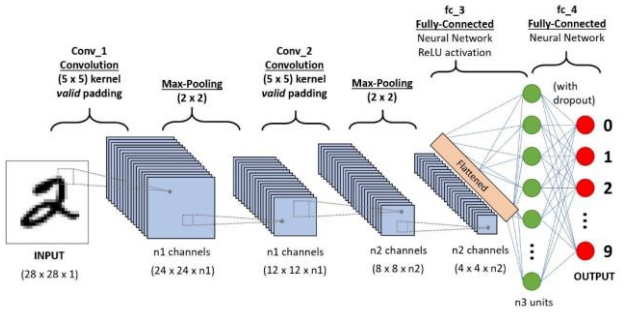
Techniques used in Generative AI (e.g., GANs, VAEs) and their adaptation for the target application.

➔ Speed of Generation / Quality
➔ Bias Considerations

2. Model Training

Model architecture selection and training for light-weight model suitable for neuromorphic AI accelerators.

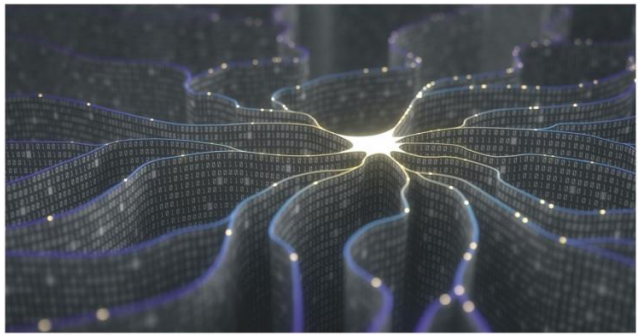
➔ Task Performance
➔ Memory Efficiency



3. Neuromorphic Accelerator Deployment

Integration approach with neuromorphic accelerators for performance evaluation and assessment.

➔ FPS
➔ Accuracy Loss



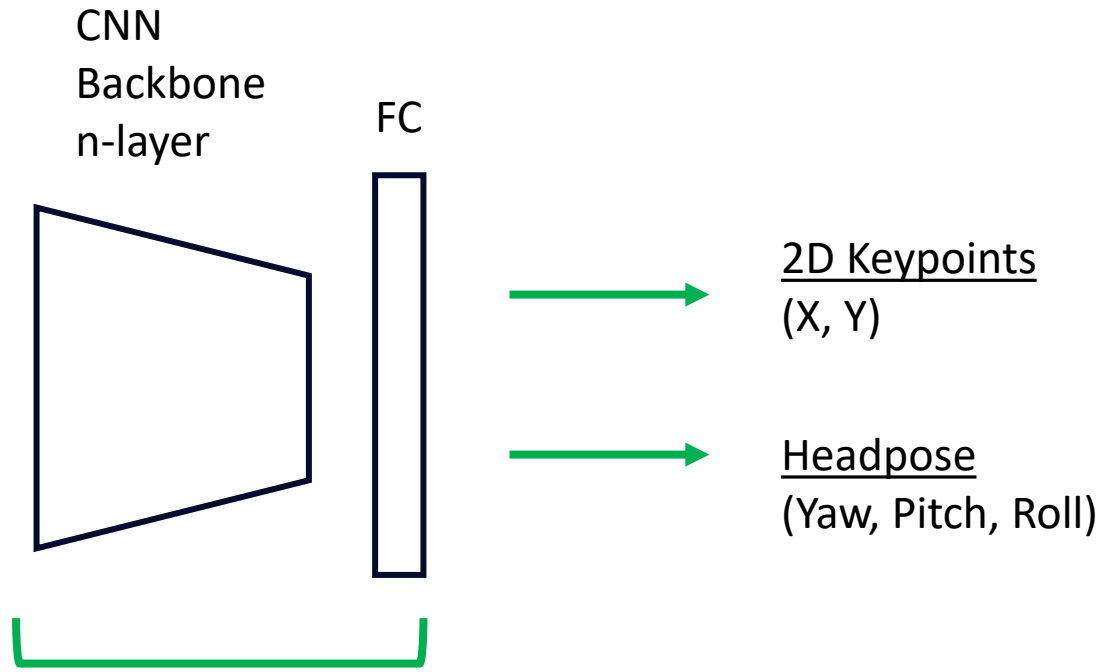
Copyright © 2023

Design and Implementation

Headpose and Keypoint Detection Task

Crop

Tile



Large Data Diversity

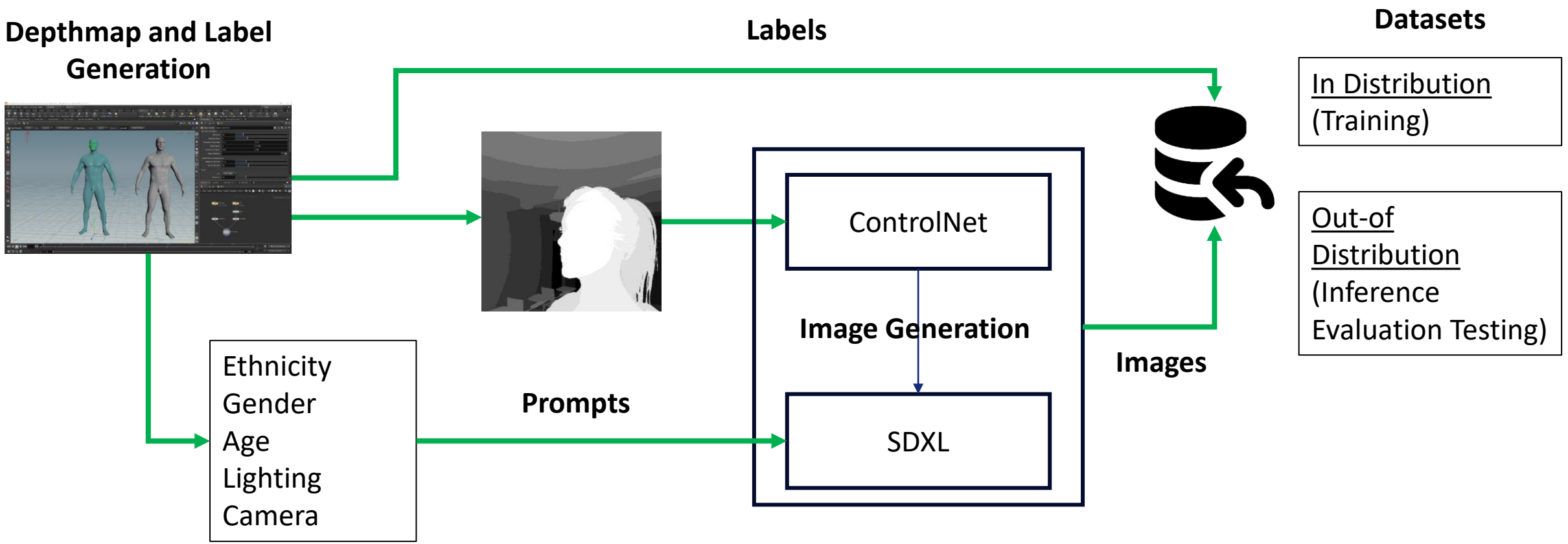
- Age
- Gender
- Ethnicity
- Cultural Background
- Expressions
- Lighting

- Lightweight feed-forward CNN
- Single head with multiple outputs
- Evaluation metrics
-- Landmark Error (NME) %
-- Headpose Angle Error (degrees)

Copyright © 2023

Design and Implementation

Automation of Training Data Pipeline



Copyright © 2023

Design and Implementation

Model Training and Neuromorphic Deployment

Training Dataset (SDXL-Headpose-100K)

100K Images

Evaluation Dataset (AFLW2000-3D)

2K Images

TensorFlow

Keras

Tensorflow Float32 Model
Quantize Tflite Model Int8
QAT with Int4

ONNX

Model Evaluation

Quantized and Compressed Model

Model Conversion

Fully-Digital Neuromorphic Accelerator

Copyright © 2023

Experimental Results

Automation of Training Data Pipeline

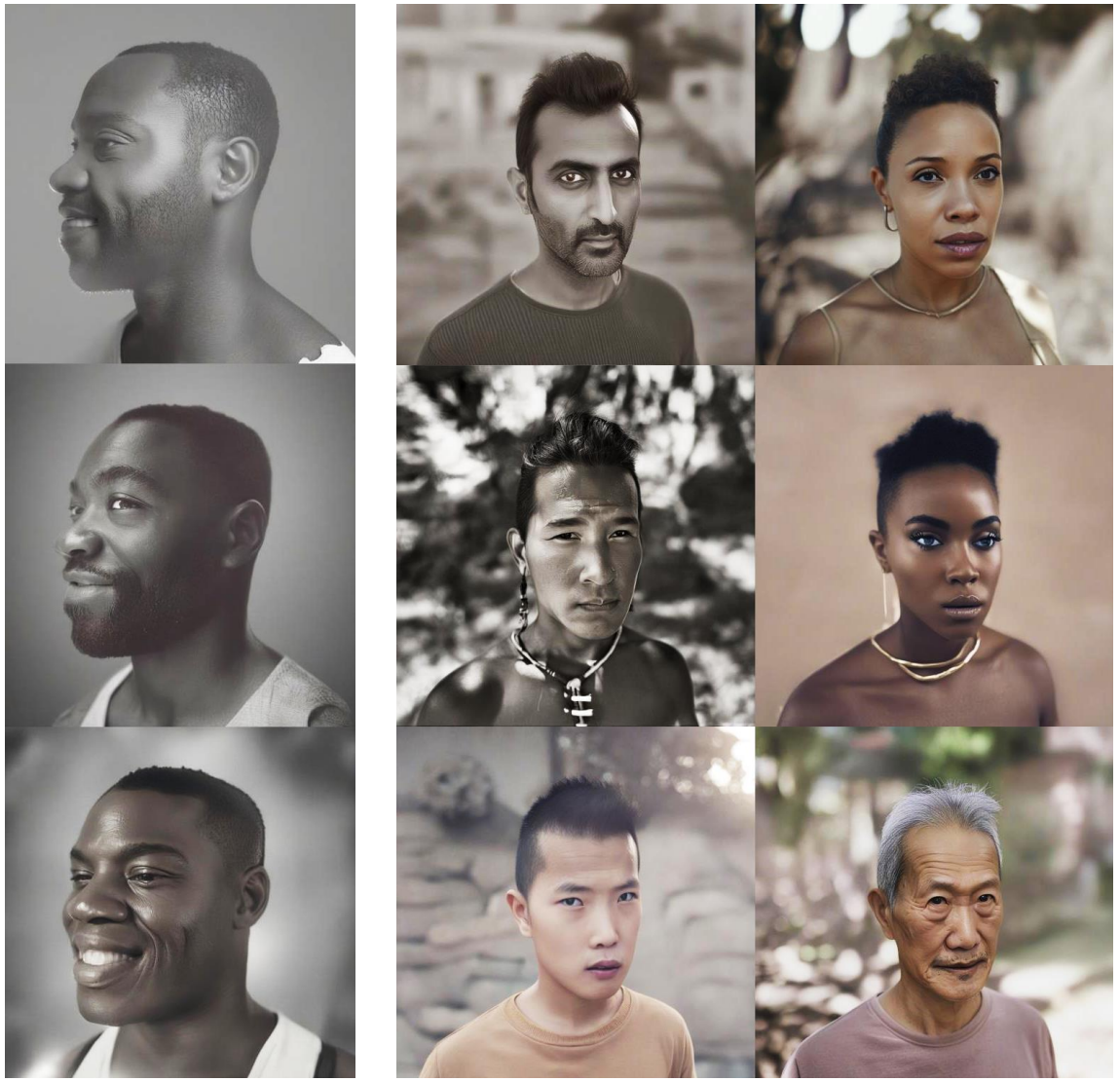
100K Images for Headpose using SDXL

- Age
- Gender
- Ethnicity
- Cultural Background
- Expressions

Large variance in image generation

Generation Time

GPU	Image	Cuda	Pytorch	GPU Memory	(576x768) (IT/s)
H100 PCIE	py21.07	12.2	2.1 Nightly	80GB	15,00
A100 PCIE	py21.07	12.2	2.1 Nightly	40GB	12,00
RTX 4090	py21.07	12.2	2.1 Nightly	24GB	11,00
RTX 3090 TI	py21.07	12.2	2.1 Nightly	24GB	6,19
RTX 3090	py21.07	12.2	2.1 Nightly	24GB	5,20



Copyright © 2023

Experimental Results

Model Training and Neuromorphic Deployment

CNN Model
- 5 layers
- ~250K Params
- 8bit = 250KB memory
- 4bit = 125KB memory

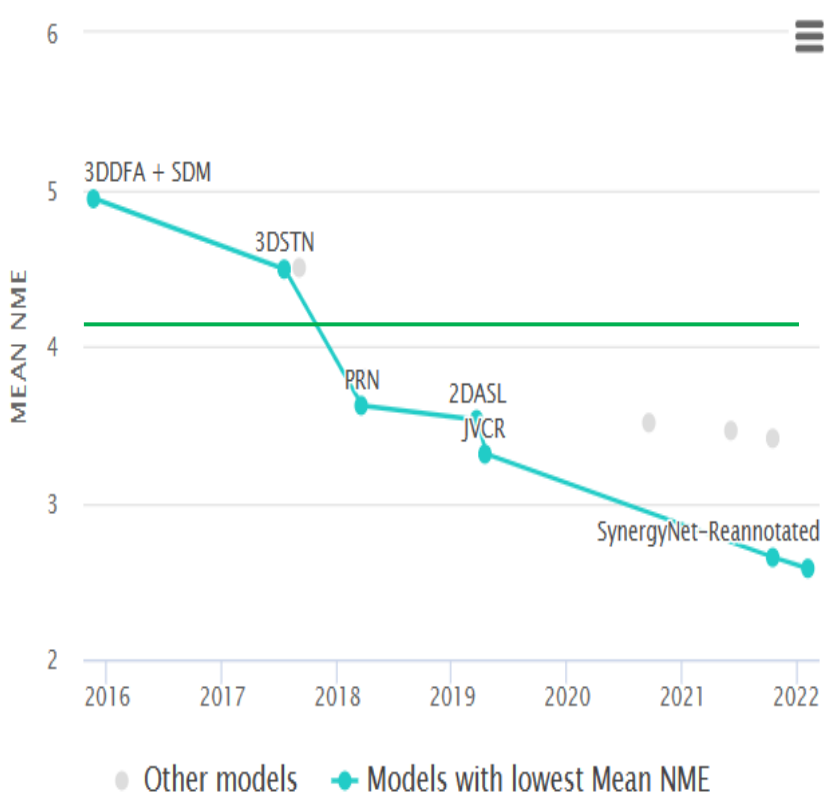
Accuracy on AFLW2000-3D

- Landmark NME % = 4.24
- Headpose MAE = 6.2

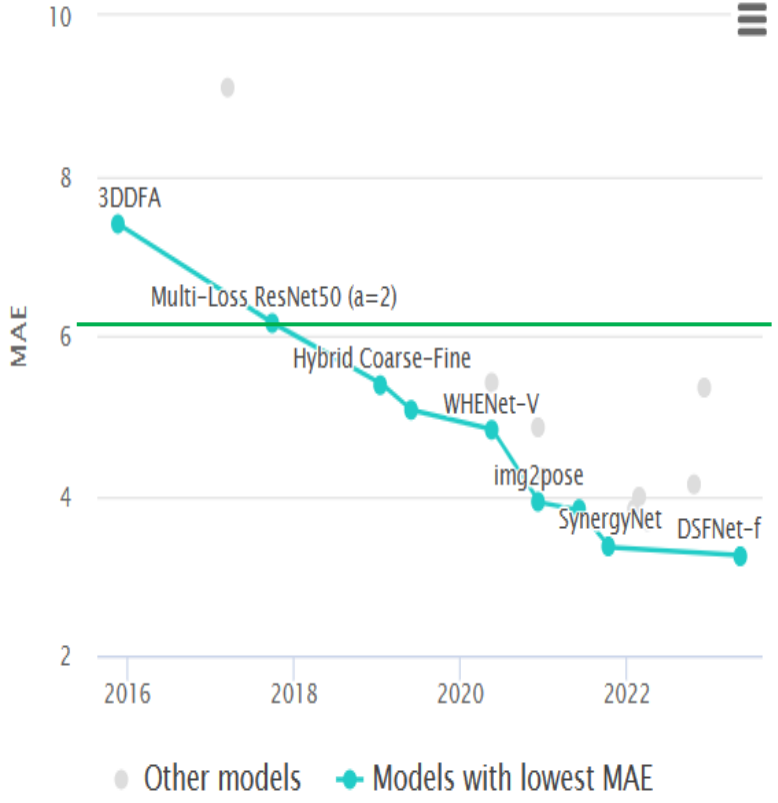
Latency

> 1000 FPS on fully digital neuromorphic accelerator

2D Landmark NME %



Headpose MAE



Copyright © 2023

Future Research

- **Temporal Consistency and Control of Occlusions from Accessories** – design for temporal consistency in video sequences as well as greater control of accessories that generate occlusions (clothes, hats, glasses, etc) through stacked control nets

- **Estimating Data Distribution and Bias Measurement from Prompts** – better understanding of bias from prompt generation and the distributed of demographics of generated data versus real-word data.

- **Automated Outlier and Duplicate Sample Detection** – detection of outliers (hallucinations) from generative models or poor-quality samples. Check for generated samples which appear duplicates.

- **Depthmaps from 3D Cameras** – compare depthmaps from real 3D cameras versus synthetics computer generated depthmaps

- **Self-supervised Learning with Foundational Models** – combine with transfer learning from foundation models such as DinoV2 as generic and scalable pretraining step for any computer vision task of small light-weight models for Edge AI applications.

- **Other Downstream Tasks** – extend to other downstream tasks such as human pose and hands to check generalization.

Copyright © 2023

Conclusion

The usage of Generative models offer a potential way to simplify edge AI model development through:

➔ increasing the quantity and quality of data for training and inference testing. Especially beneficial in applications where data is scarce or costly and/or time consuming to collect.

➔ reducing bias and alleviating ethical considerations as large open-source generative models can generate large variations while being able to operate on existing data sources.

➔ reducing the time and cost of data collection and labelling as pipelines can be automated, shared, reused, and versioned in the development process using containerization and GPU/HPC technologies.

➔ favoring the usage of simple feed-forward model architectures that support better low-bit quantization and can be converted to run on digital neuromorphic computing architectures without major rework of model architectures.

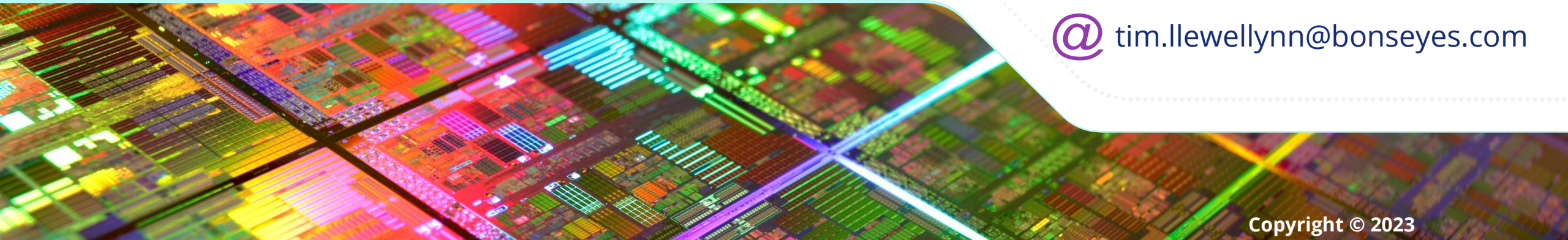
➔ promoting further research into data-centric methods and tools that offer substantial gains in improvement of the ML workflow for embedded systems. This is complimentary to model-centric methods which are already well established for edge AI applications.

Copyright © 2023

EDGE AI

Thank You

For your attention



@ tim.llewellynn@bonseyes.com

Copyright © 2023