Sandia National Laboratories

Exceptional service in the national interest

UNSUPERVISED MULTI-MODAL MACHINE LEARNING FOR IN-PROCESS MONITORING OF ADDITIVE MANUFACTURING:

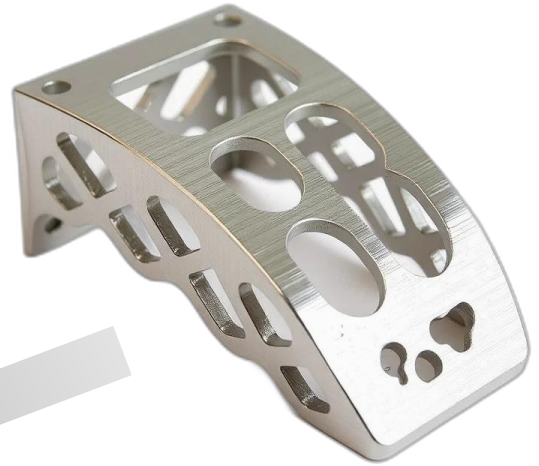
Anthony Garland, Matthew McKinney, Dale Cillessen, Jesse Adamczyk, Dan Bolintineanu, Michael Heiden, Elliot Fowler, Brad Boyce

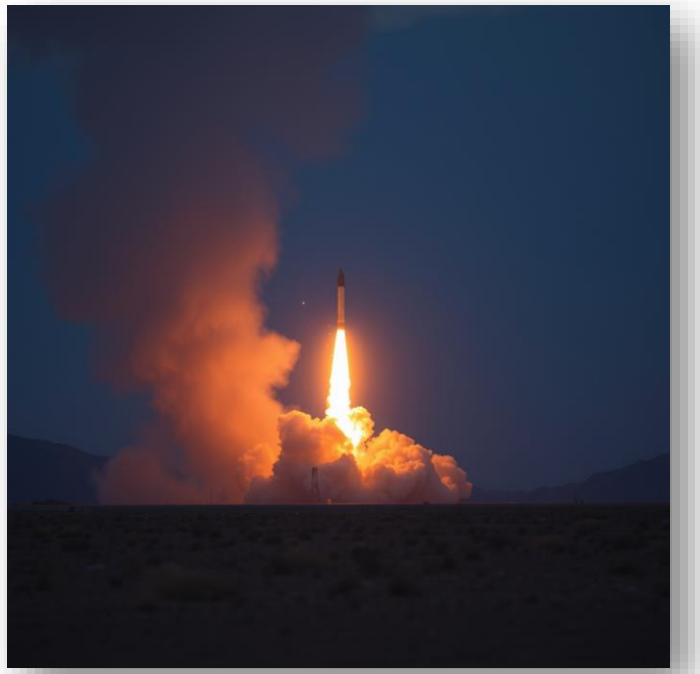
SAND2024-10348C

SAND2024-10348C

OUTLINE

- Introduction and Objectives
- Methods
  - Datasets
  - Model architecture
  - Training process
- Results
  - In-situ LPBF
  - Lattice Dataset
- Analysis of Results
- Conclusion





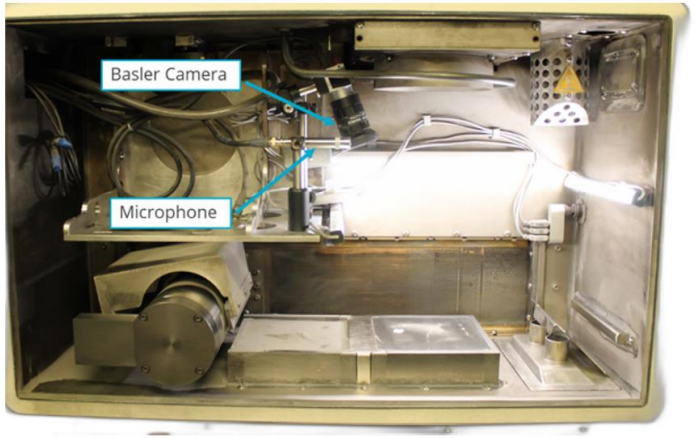
INTRODUCTION

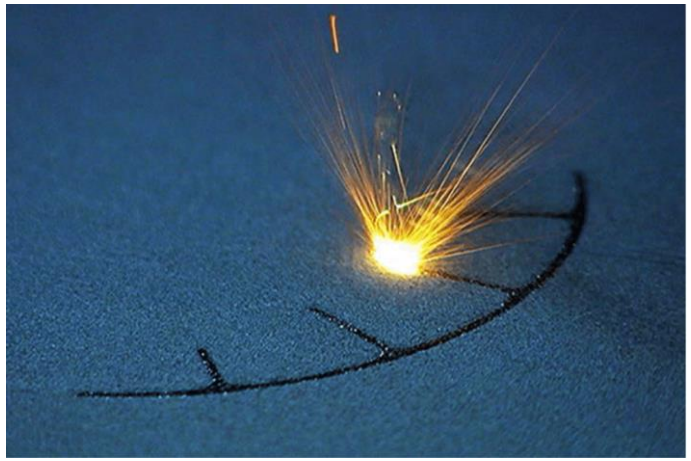
Additive manufacturing

- **Layer Formation**: A thin layer of metal powder is evenly spread across the build platform, creating a uniform layer of material.
- **Selective Melting**: A high-powered laser selectively melts and fuses the metal powder together, following a predetermined pattern generated from a 3D CAD model.
- **Layer-by-Layer Build**: Steps 1-2 are repeated, with the build platform lowering and a new layer of powder being applied, until the entire part is built layer by layer, resulting in a fully dense and functional component.

Challenges

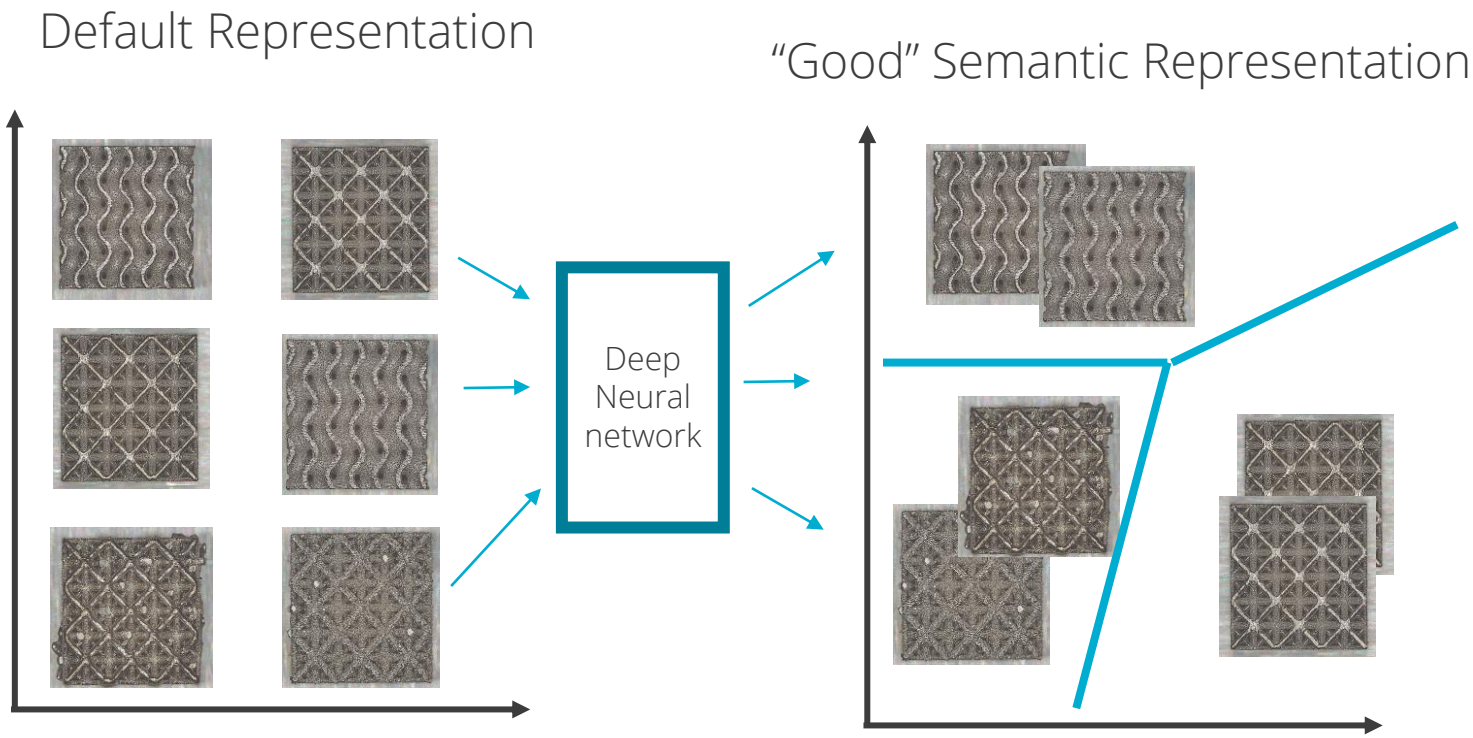
- **Ensuring Print Consistency**: Guaranteeing that the 3D printing process yields consistent results across multiple builds
- **Defect Detection and Identification**: Identifying and characterizing defects such as pores and other
- **In-Situ Data Interpretation**: Developing effective methods for interpreting data generated during the printing process to detect and identify defects.





OBJECTIVES

- Intelligently fuse multi-modal in-situ monitoring data.
- Minimize the labeling burden no humans.
- Limit a-prior assumptions
- Generate “good” representations
- Use unsupervised machine learning
- Why?
  - Uncover hidden fingerprints in our data.
  - Generate representations for down stream tasks



A good representation is one that makes a subsequent learning task easier. The choice of representation will usually depend on the choice of the subsequent learning task [Bengio, Goodfellow ..]

How to generate this neural network?

METHOD

- Inspired by CLIP (Contrastive Language-Image Pre-training)
  - One of the most cited multi-modal ML papers ever
- CLIP trained encoders for text and images
- Contrastive loss (no labels needed, just tuples)
- Generates ‘good representations
- Challenges
  - Adapting for manufacturing-specific sensor modalities
  - Will it even work?
  - What encoder architectures will work?

Learning transferable visual models from natural language supervision

A Radford, JW Kim, C Hallacy, A Ramesh, G Goh, S Agarwal, G Sastry, A Askell, P Mishkin…

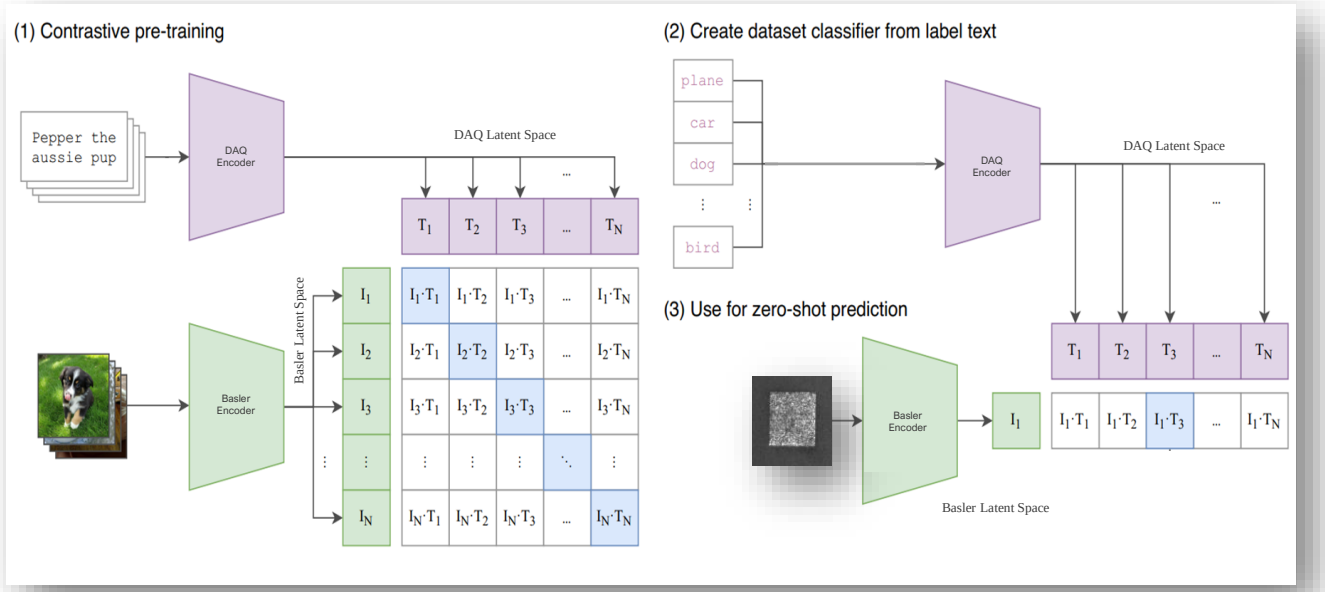
International conference on machine learning, 2021 · proceedings.mlr.press

Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations

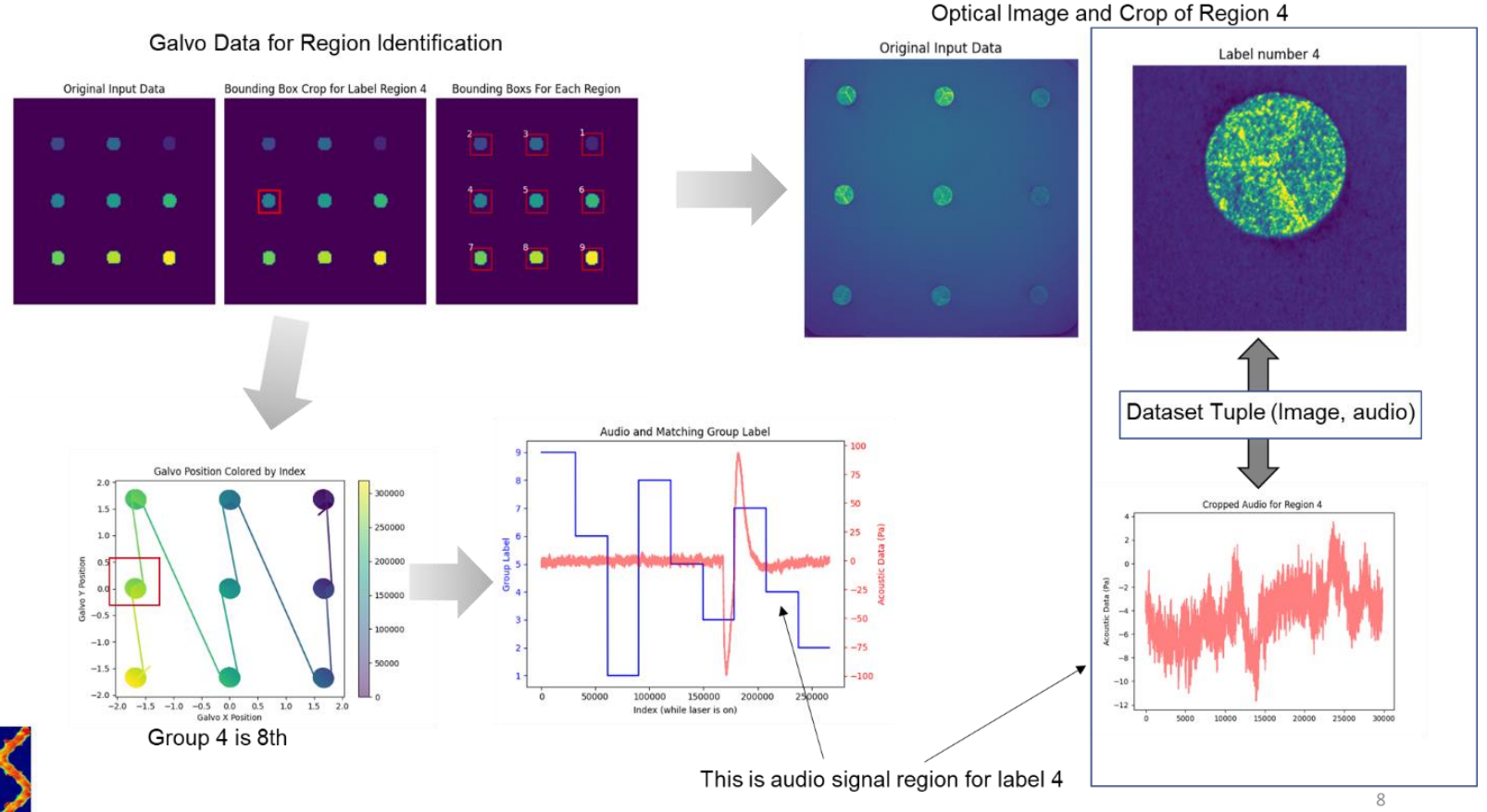
SHOW MORE

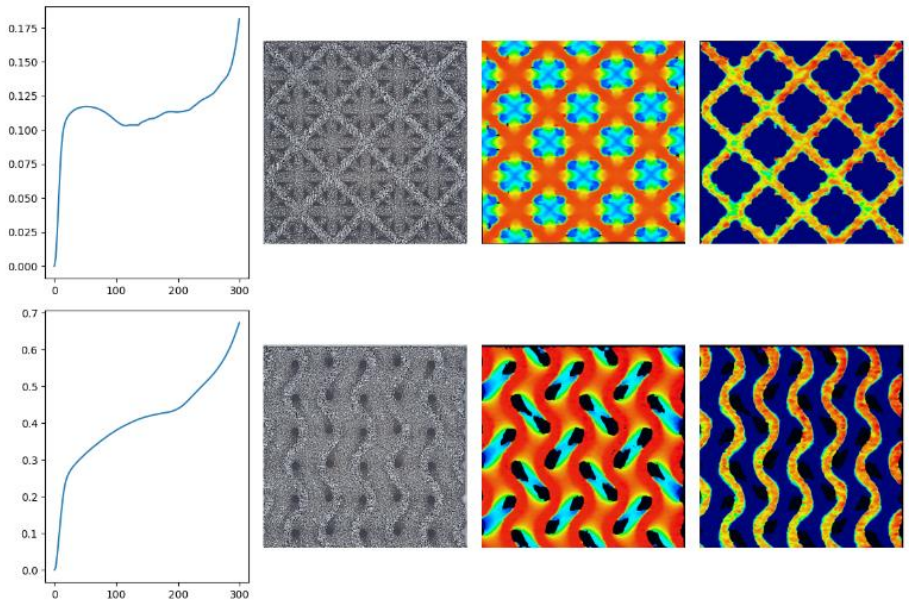
☆ Save Cite Cited by 12391 Related articles All 17 versions



DATASETS

- In-situ
  - Generate data tuples from in-situ Laser Powder Bed Fusion (LPBF) process data
  - Multiple modalities: images, audio, DAQ signals



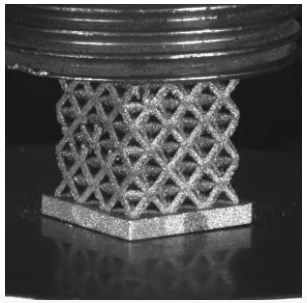


- Lattice
  Post build information
  Optical, height maps, forces displacement curves.

DATASET VARRIATION

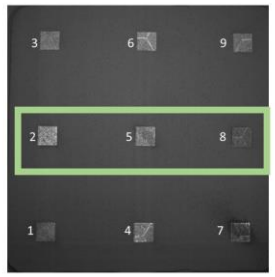
- 4 builds
- Cubes nominal
- Cubes with 6 off-nominal
- Cylinders nominal
- Cylinders with 6 off nominal

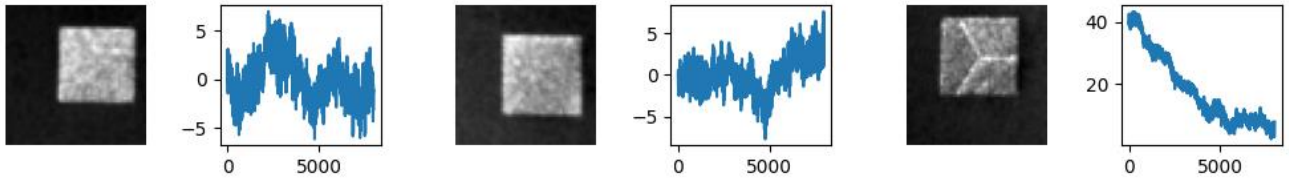
- Gyroids and Octets
- Variety of print settings

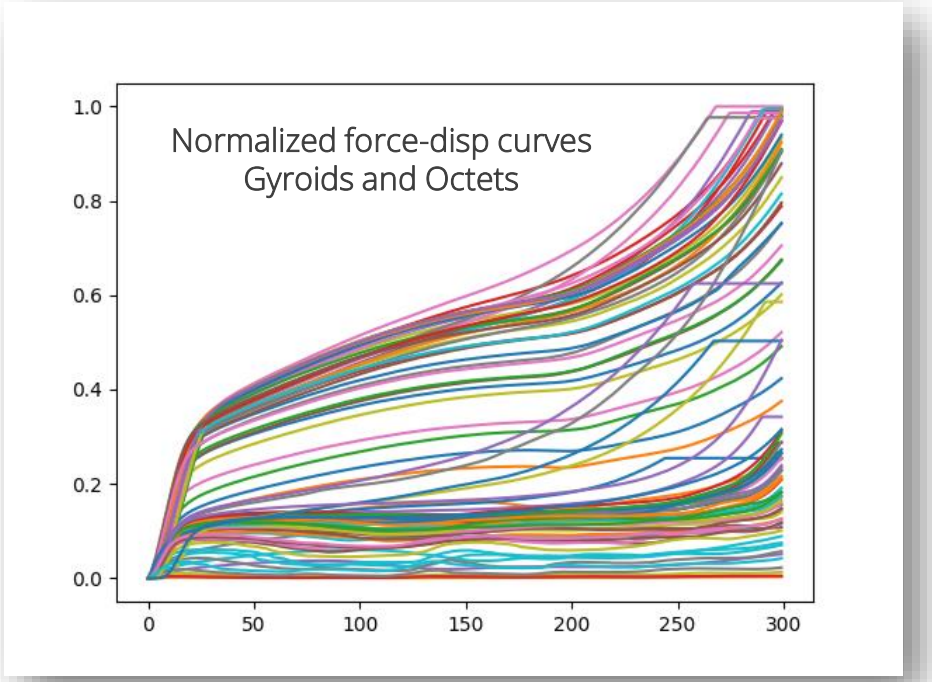


#	Laser Power (W)	Laser Speed (mm/s)
1	107	1400
2	113	1400
3	119	1400
4	102	1400
5	113	1400
6	124	1400
7	113	1330
8	113	1400
9	113	1470

#	Laser Power (W)	Laser Speed (mm/s)
1	250	2300
2	113	1400
3	100	2000
4	200	1800
5	113	1400
6	150	2400
7	150	1000
8	113	1400
9	225	2800

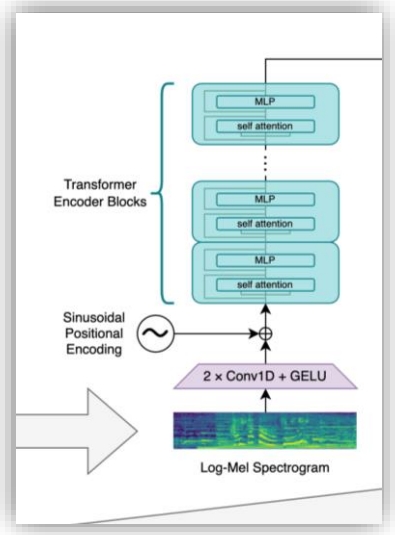


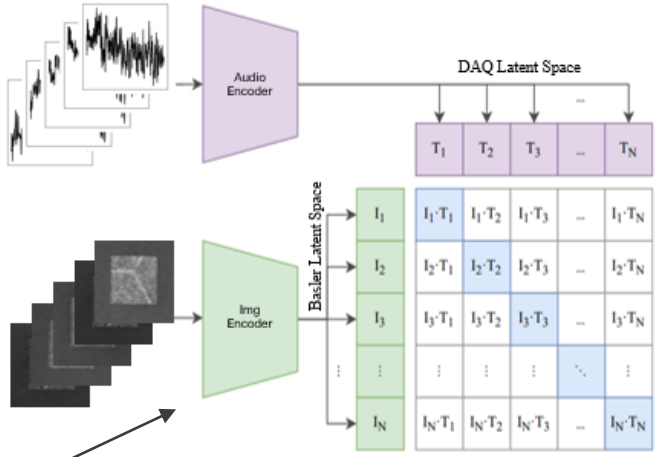


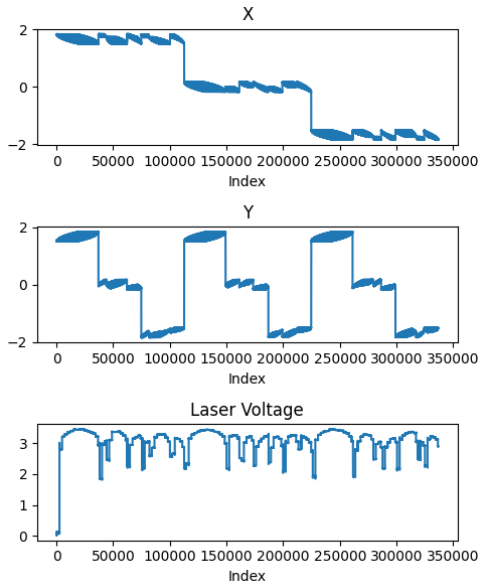


MODEL ARCHITECTURE FOR IN-SITU

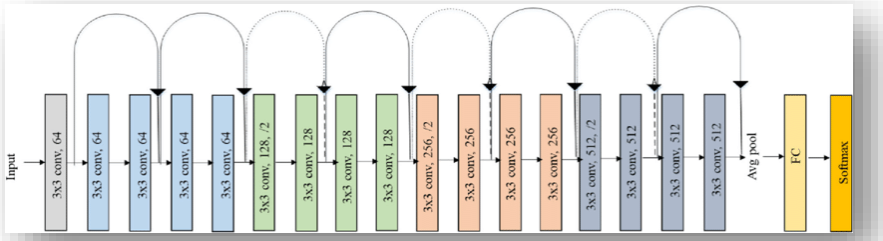
Whisper (Audio encoder)



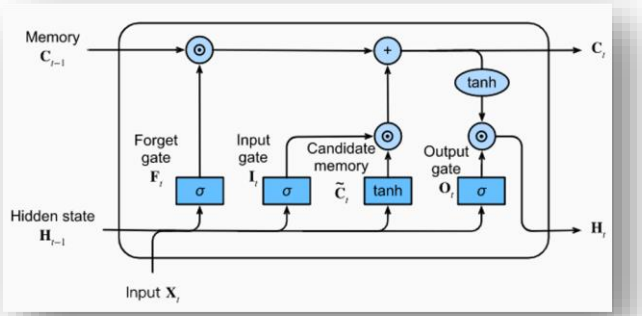




Resnet18

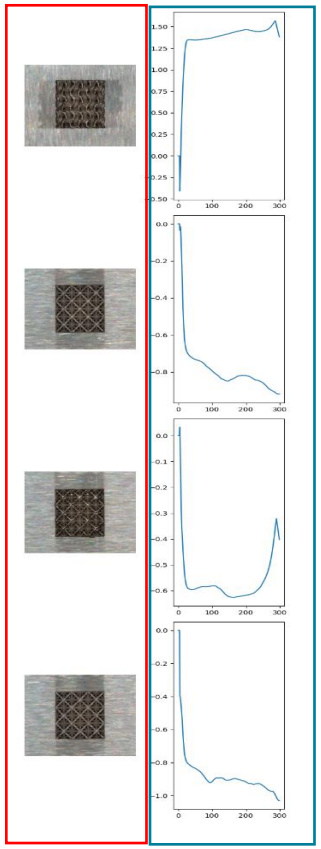


Multi-resolution LSTM

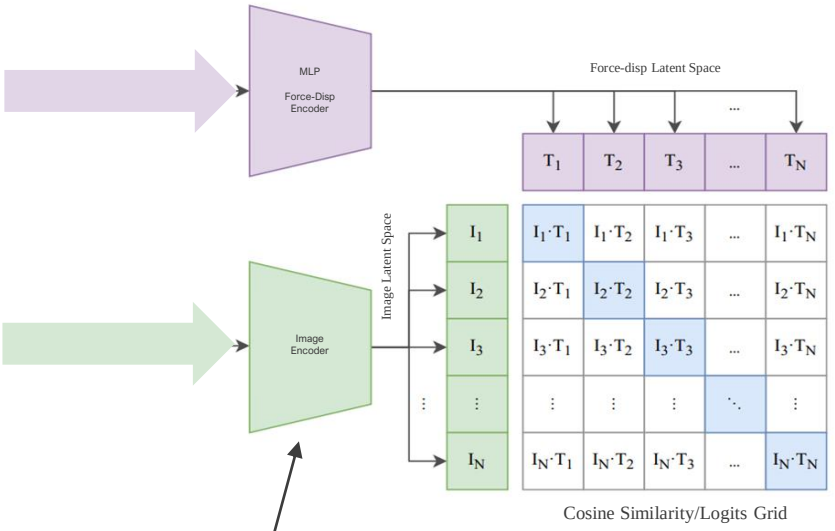


wikipedia

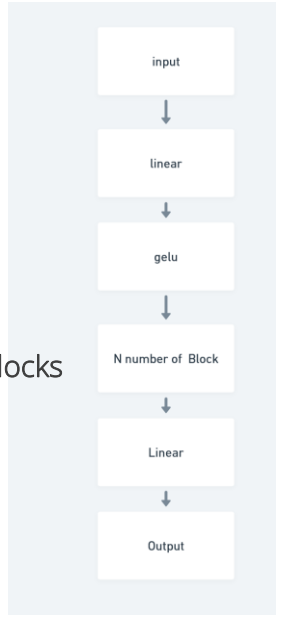
MODEL ARCHITECTURE FOR LATTICES



Resnet like MLP

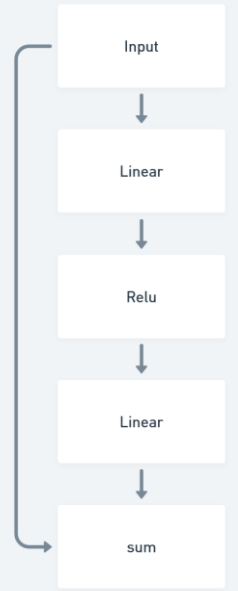


MLP

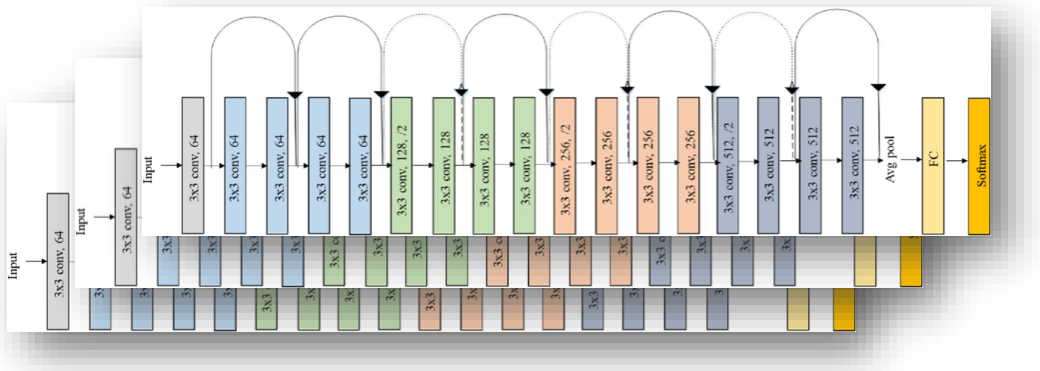


5 blocks

MLP ResBlock



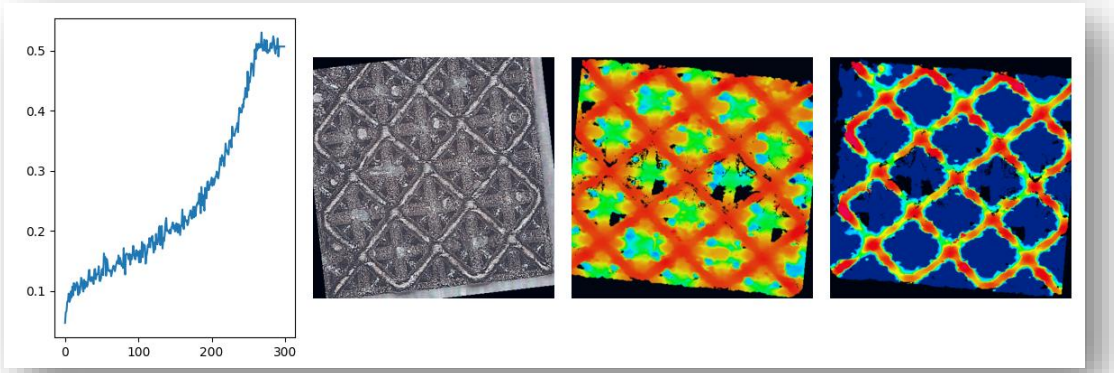
3X Resenet18



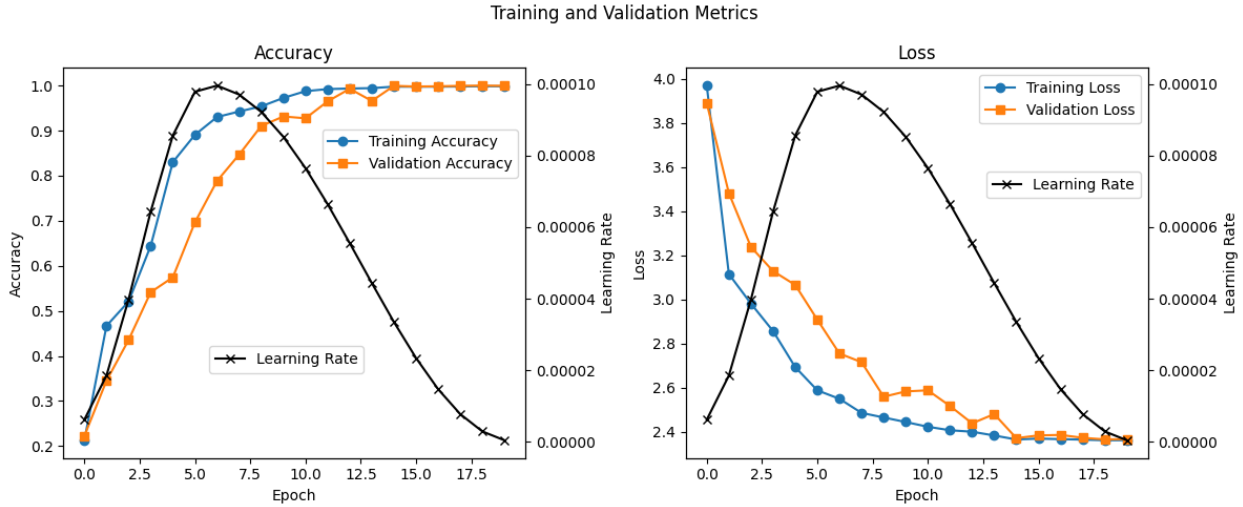
TRAINING PROCESS

- Contrastive loss function
  - Matching tuples pushed together
  - Dissimilar tuples pushed apart
  - Accuracy = the most similar vector from the other modality is the match
- Data augmentation
  - Standard augmentation methods for Images
  - Audio. Add noise
  - Force-disp curves. Noise and shift curve by N idxs
- Optimization details
  - One cycle annealing lr
  - Limited hyper parameter turning since it was working so well

Lattice tuple with augmentation



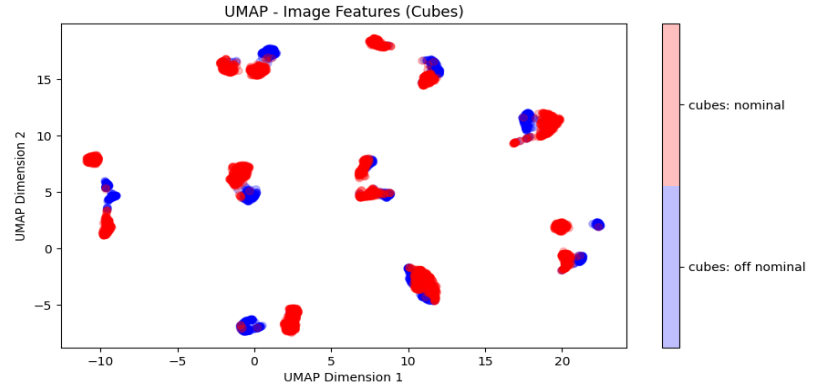
In-situ CLIP training metrics

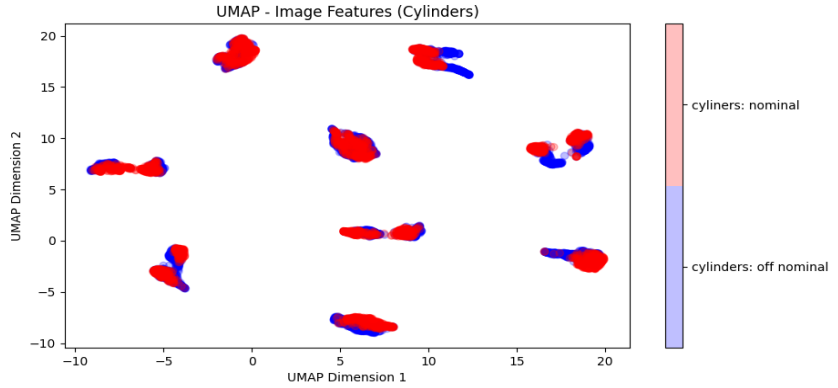


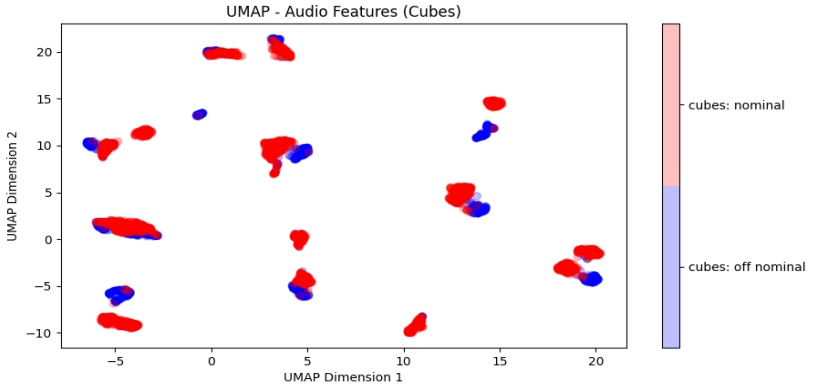
Loss goes down consistent
Accuracy maxes out at 100% for training and valuation

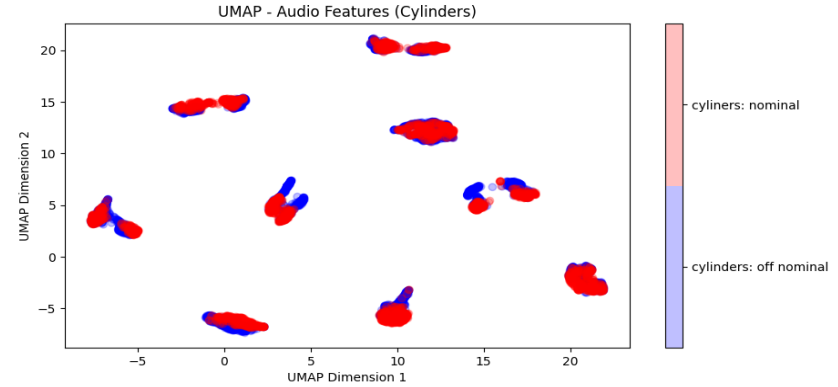
RESULTS IN-SITU

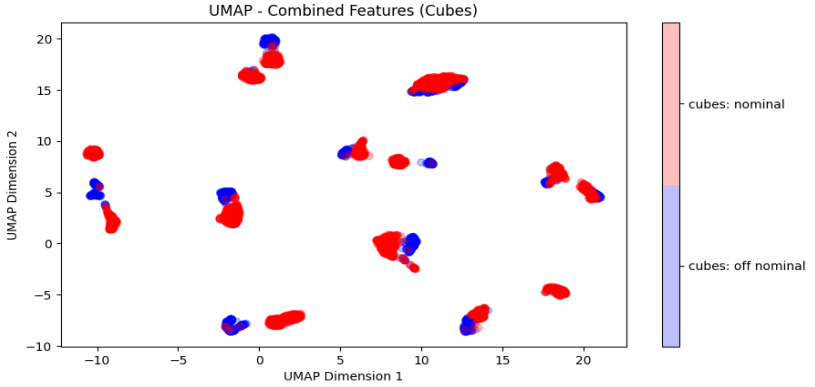
- Learned latent space
- Umap for dim reduction
- Each point = data tuple
  - Image, audio, DAQ
- 9 groups = 9 locations
- Blue
  - From off-nominal build
  - 3 are still nominal
- Red = nominal build
- Cubes have more process variation than cylinders.

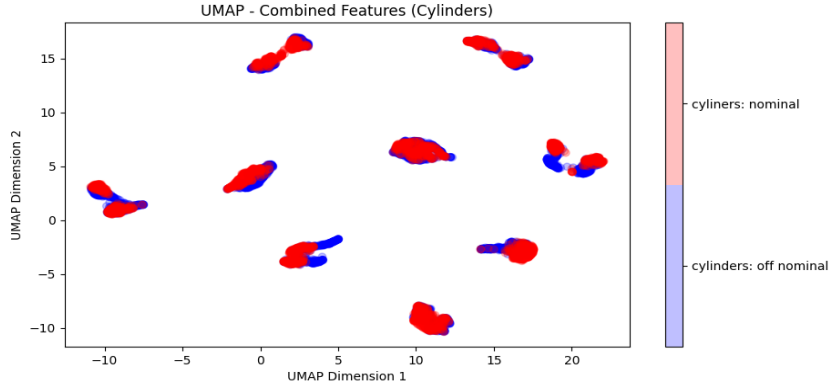






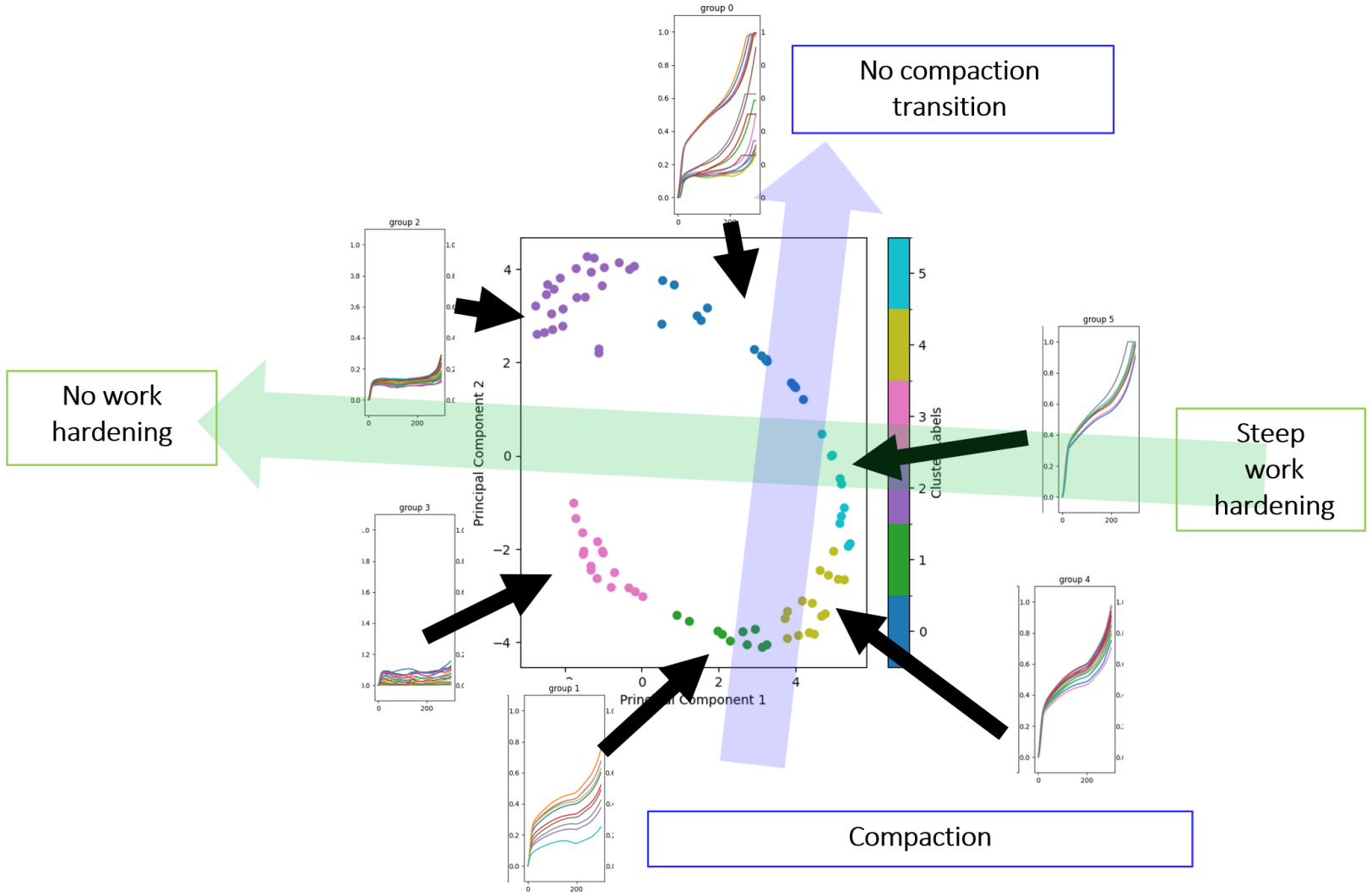






Items with more variation in the process parameters are further apart in latent space. Why? The signals are some how different AND the model learned to discriminate

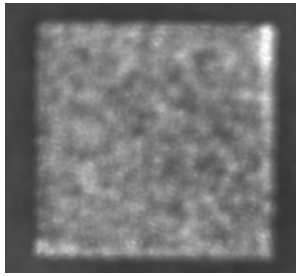
RESULTS: LATTICE D



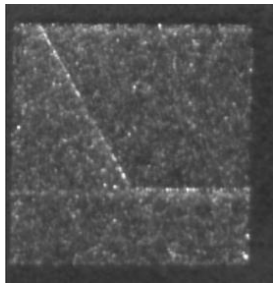
Lattice vectors lie on a 2D manifold
The points have a relatively smooth transition (therefore clustering has little meaning)
The ‘directions’ on the manifold have physical meaning.

CHALLENGES AND LIMITATIONS

- In-situ
  - Data quality and quantity
  - Angle of camera results in close and far items being out of focus
  - Microphone has limitations in quality and quantity
  - More data would be helpful
    - More of the same to help further confirm results
    - Diverse data. Our builds are ‘simple’
      - (do overhangs ‘sound’ different?)
    - Photodiode data would be helpful.
    - Often data tuples are often almost identical
- Lattices
  - Limited dataset size
  - Interpretation of Umap 2D output is hard

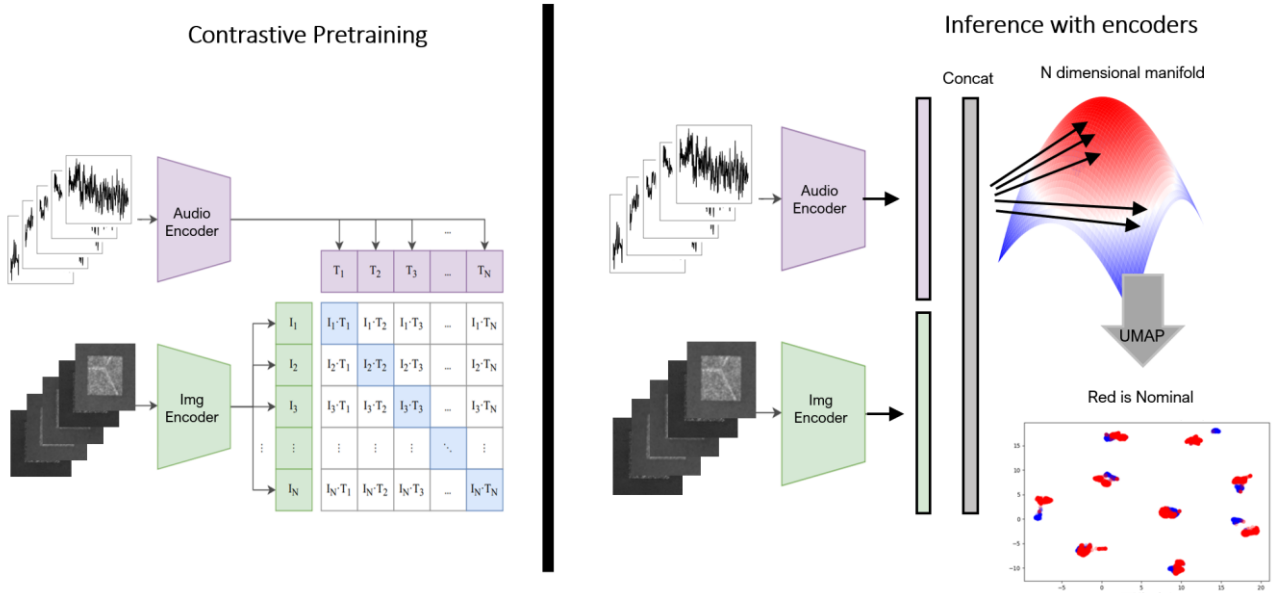


Vs.



CONCLUSION

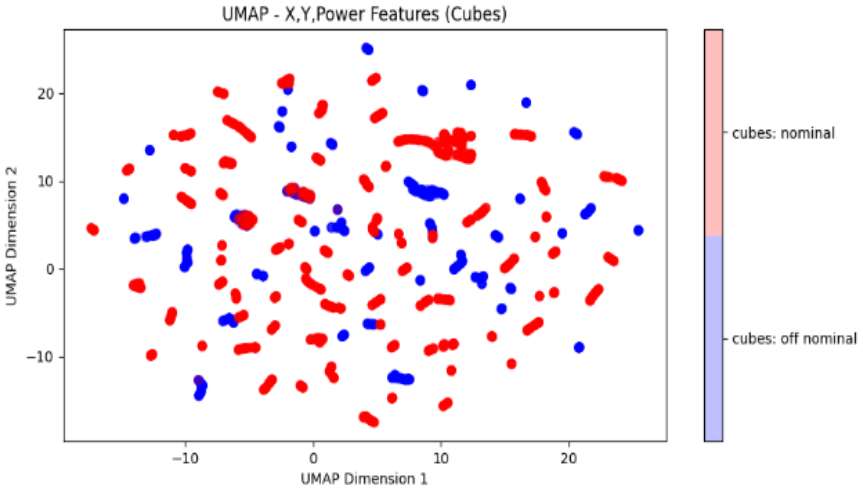
- Used multimodal data to generate meaningful representations of the data
- Used unlabeled data and contrastive loss
- Correlated expensive (images) with cheap measurement (audio)
- Adapted CLIP to work on manufacturing data
- Extended CLIP beyond 2 modalities
- Encoder architecture agnostic.
  - Used a variety of encoders (Resnet18, Whisper,LSTM, MLP)
- Efficient and stable training -> The technique scales easily to bigger datasets
- The generated representations can be used for down stream tasks
  - Process consistency checks
  - Physical insight (hidden fingerprints)

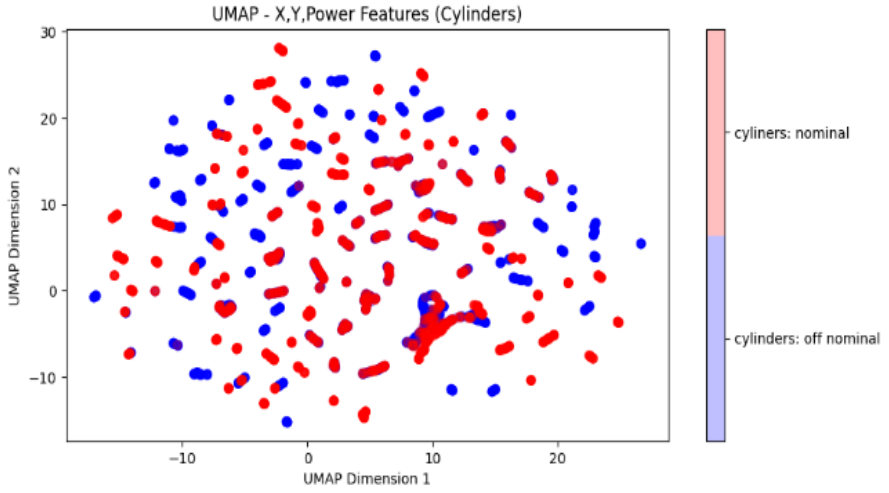


Q&A / THANK YOU

BACKUP

DAQ LATENT SPACE -> UMAP





MULTI-SCALE LSTM

$$y = f(x) = W_2 \cdot ReLU(W_1 \cdot concat\left([h_i]_{\{i=0\}}^{\{N-1\}}\right) + b_1) + b_2$$

where:

$$h_i = LSTM_i(x[::2^i])_T, i \in \{0, 1, \ldots, N-1\}$$

and x is the input sequence, N is the number of LSTM heads , h_i is the final hidden state of the i-th LSTM head, $x[::2^i]$ denotes the input sequence subsampled by a factor of 2^i , $LSTM_i(\cdot)_T$, represents the last output of the i-th LSTM , W_1, W_2 are weight matrices and b_1, b_2 are bias vectors of the fully connected layers, $concat[\cdot]$ represents the concatenation operation.

CLIP LOSS

For a batch of N image-audio pairs, the loss function can be expressed as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\left(\frac{exp\left(\frac{sim(x_i, y_i)}{\tau}\right)}{\sum_{j=1}^{N} exp\left(\frac{sim(x_i, y_j)}{\tau}\right)}\right) + \log\left(\frac{exp\left(\frac{sim(y_i, x_i)}{\tau}\right)}{\sum_{j=1}^{N} exp\left(\frac{sim(y_i, x_j)}{\tau}\right)}\right)\right]$$

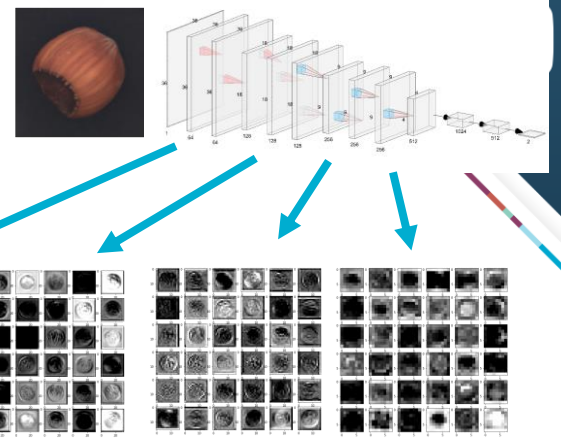
Where $sim(u, v) = \frac{u^T v}{||u||||v||}$ is the cosine similarity between vectors u and v, τ is a temperature parameter that controls the softmax distribution, and x_i and y_i are the i-th image and audio inputs, respectively.

FADS

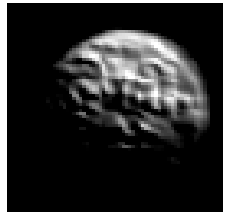
- Given a pretrained NN, and 'nominal' images X
- Observe the 2D convolutional activation maps caused by the nominal images X
- Summarize the 2D activation map using some function f_a (e.g. min, max, avg)
- Do for all 'training' images. Calculate the average and std conv filter 'activation'

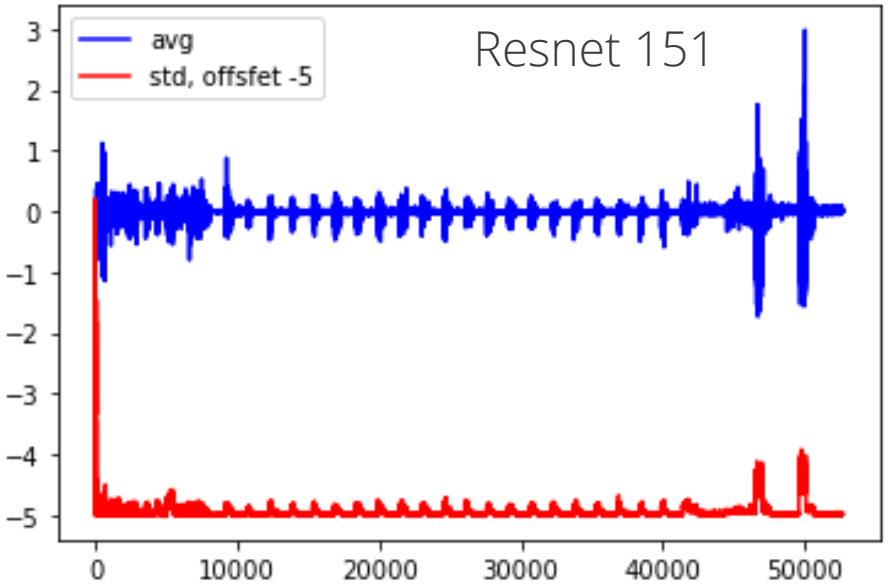
$$\phi_{i,\bar{x}} = mean$$
$$\phi_{i,\sigma} = \text{standard deviation}$$

- $\phi_{i,\bar{x}}$ and $\phi_{i,\sigma}$ are then a 'fingerprint' of the nominal case

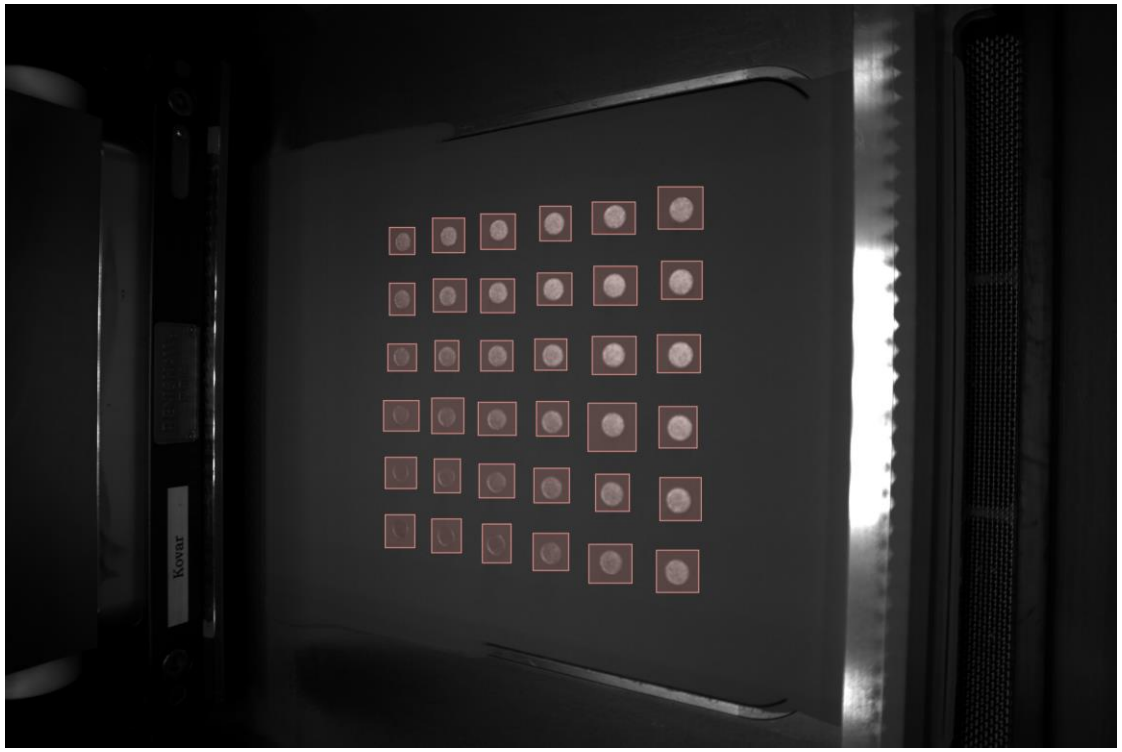


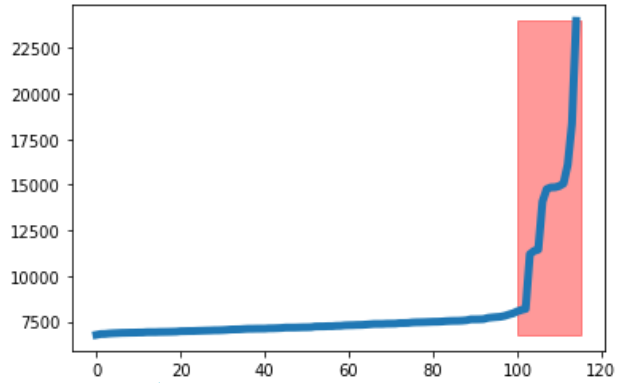
f_a summarize conv activation map to single number.

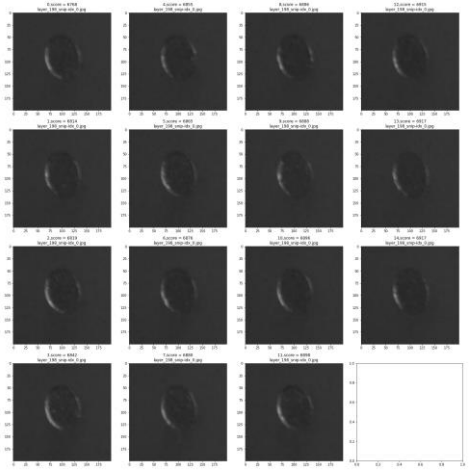
$5 = f_a($  $)$

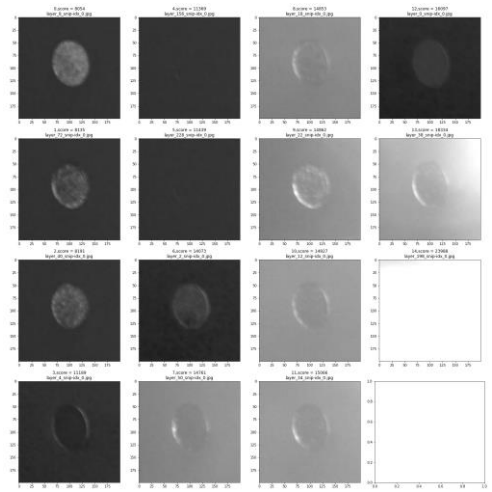


ADDITIVE MANUFACTURING IN-PROCESS MONITORING + FADS









Question: Does this print look like all the ones before?
Answer: We can identify any layers that looked unusual using FADS.

Setup:

- Some layers were printed with incorrect process settings.

Result:

- We identified all bad layers.