Strategies for Image Dataset Curation from High-Volume Industrial IoT Data

Dan Bricarello
Computer Vision Chapter Lead

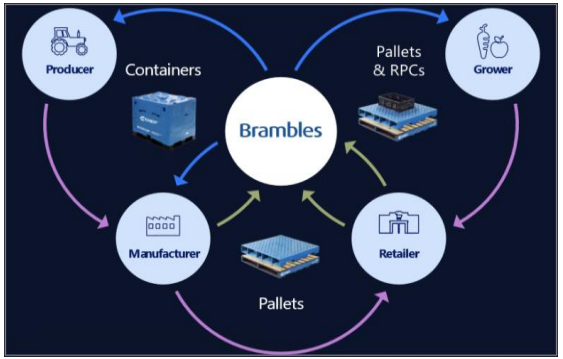
Apurva Godghase
Senior Computer Vision Engineer

Brambles

Brambles

Advancing the World's Supply Network, Together

- Brambles is a global provider of logistics solutions, connecting the world's supply network through its operations, people and technology
- As of 30 June 2024, Brambles:



Operated in ~60 countries

Owned ~347m pallets, crates and containers

Primary brand
CHEP
A Brambles Company

Employed ~13,000 People

Through 750+ service centres

Market leader
#1
market position in all regions

Asset Tracking at Brambles

Untracked assets can be reused without authorization

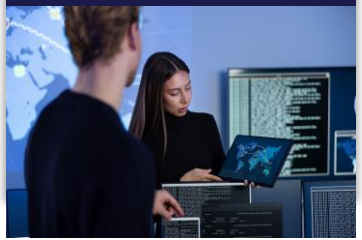
Misuse increases costs and consumption of natural resources (timber)

Helps productivity and sustainability goals

Hundreds of millions of wood pallets and other assets worldwide

Data collection

Data analysis



Action

Embedded Vision for Asset Tracking

Asset Tracking

Improving our understanding of the flow of every asset by uniquely identifying individual pallets through computer vision and QR codes

Serialize pallet pool with unique IDs

Embedded vision systems track assets

Support multiple tag formats

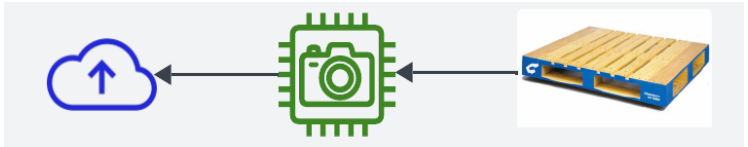
Current implementation uses QR tags

Reliable, scalable, mature system

High quality tracking data

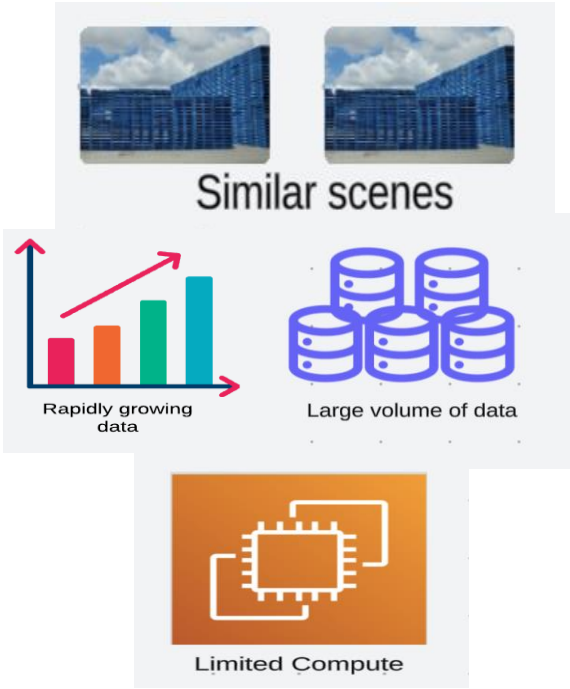
Vision Based Approach for Asset Tracking

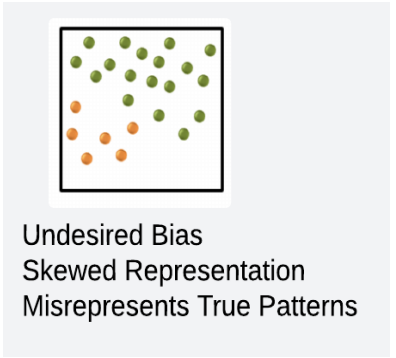
- Visibility into start and end of asset cycles
- Scans of QR tags used to serialize assets
- Embedded Vision system to detect and decode the QR codes
- Continually capture data



Scaling Embedded Vision in Industrial Environments

- Continuous capture by vision systems leads to many images having very similar scenes or irrelevant data
- Filtering and deduplication removes unintended biases caused by oversampling
- Large volume : (~10 devices capture ~1 Million images/day)
- Challenging for effective modeling and analysis.

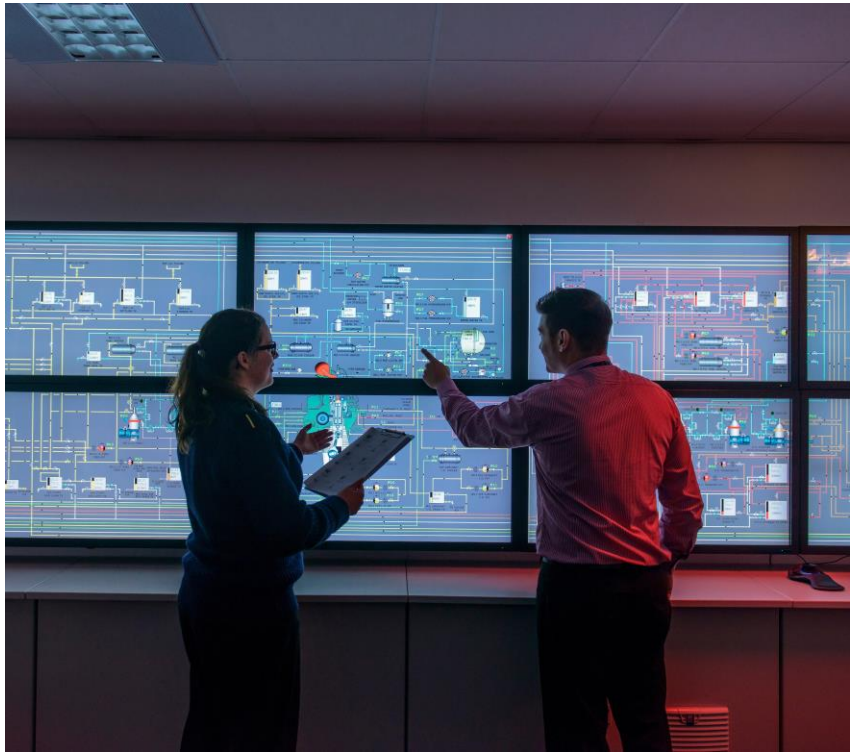




Data Curation

Objective

- High-quality subset selected from a huge image pool
- Reduce upload volume and focus on more meaningful data (Images with QR code scans)
- Build balanced, relevant, diverse training dataset of QR code images for training and monitoring various computer vision models (e.g., classification)



Filtering Data at the Edge (1/2)

Tag detection

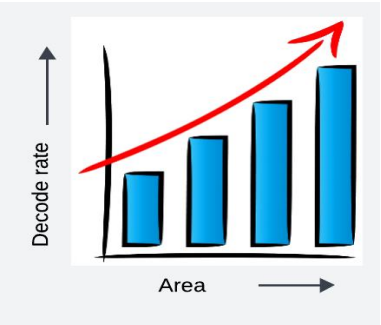
- QR Codes

Discard uninformative images

- Images with odd aspect ratio
- Tags at the edge of the frame

Filter on tag area

- QR code image crops below a certain size are less likely to decode



Filter Data at the Edge (2/2)

Data Sampling

- Systematic sampling to reduce volume of data

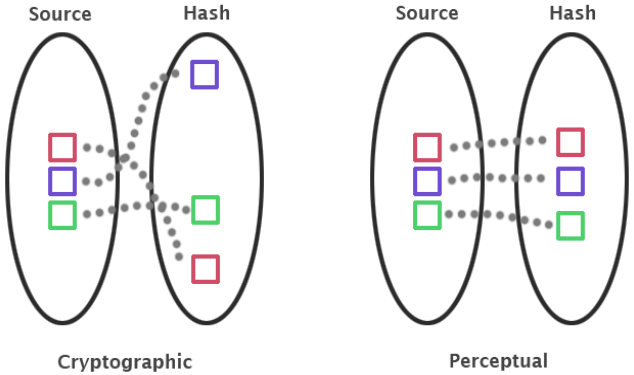
Data Compression

- JPEG/ PNG instead of raw image data

Store Selective Data

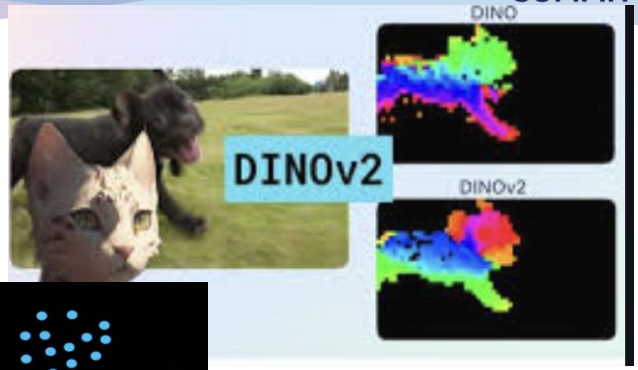
- Selective data is stored to the cloud
- Used for various modeling tasks

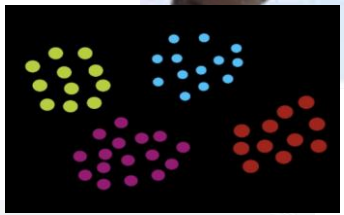
Deduplication



Method 1: Perceptual Hashing

- Easy, fast
- Sensitive to transforms





Method 2: DINOv2 based embeddings

- Semantic understanding
- Robust to transforms
- Complex

Method 3: Custom embeddings

(Preferred approach)

- Good for domain specific data
- Dependency on other models to extract embeddings

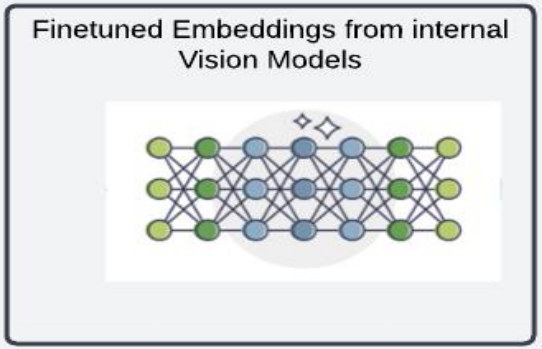


Image Quality Assessment

Overall Perceptual Quality

BRISQUE

- Measure of Natural Scene Statistic
- Compute intensive

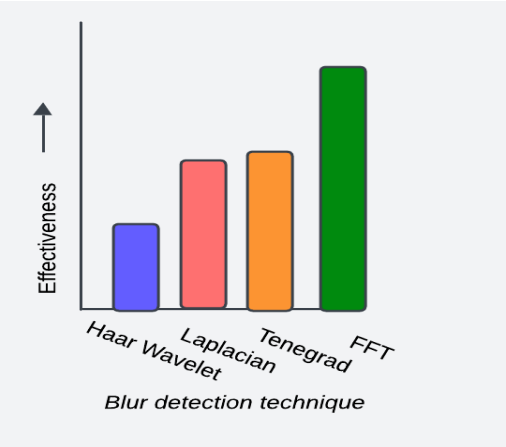
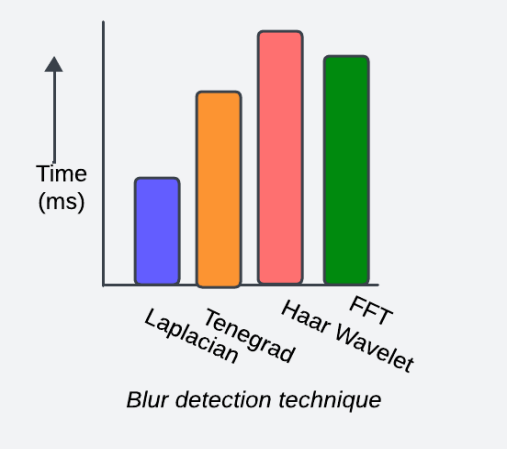


Image clarity

Blur Detection (Preferred approach)

- Measure of sharpness
- Address blurry images



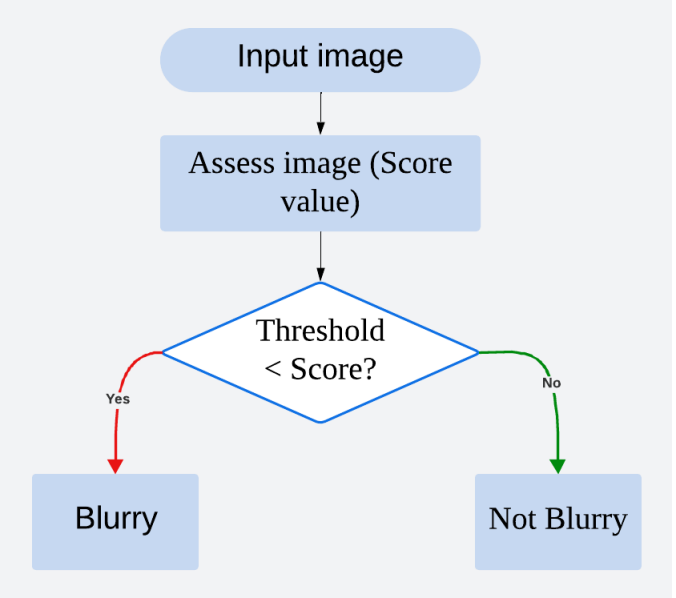


Image Quality Assessment : Illumination

Average Pixel Intensity

- **Overexpose**

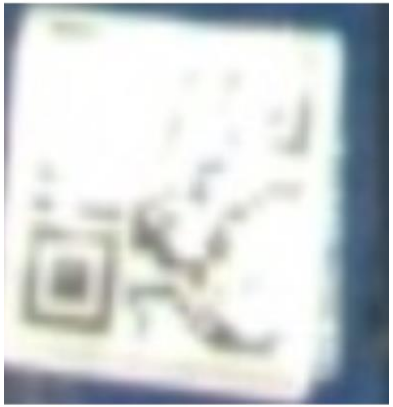
  Losing detail in bright areas

- **Underexposure**

  Losing detail due to insufficient light

- **Well-lit**

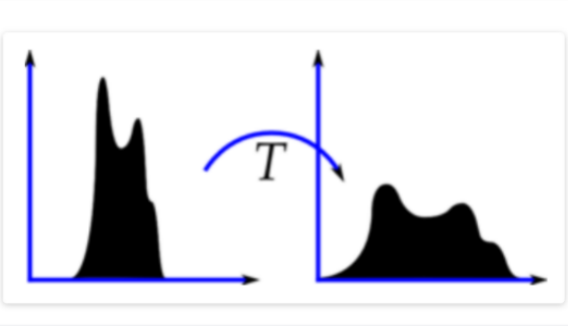
  Sufficient lighting



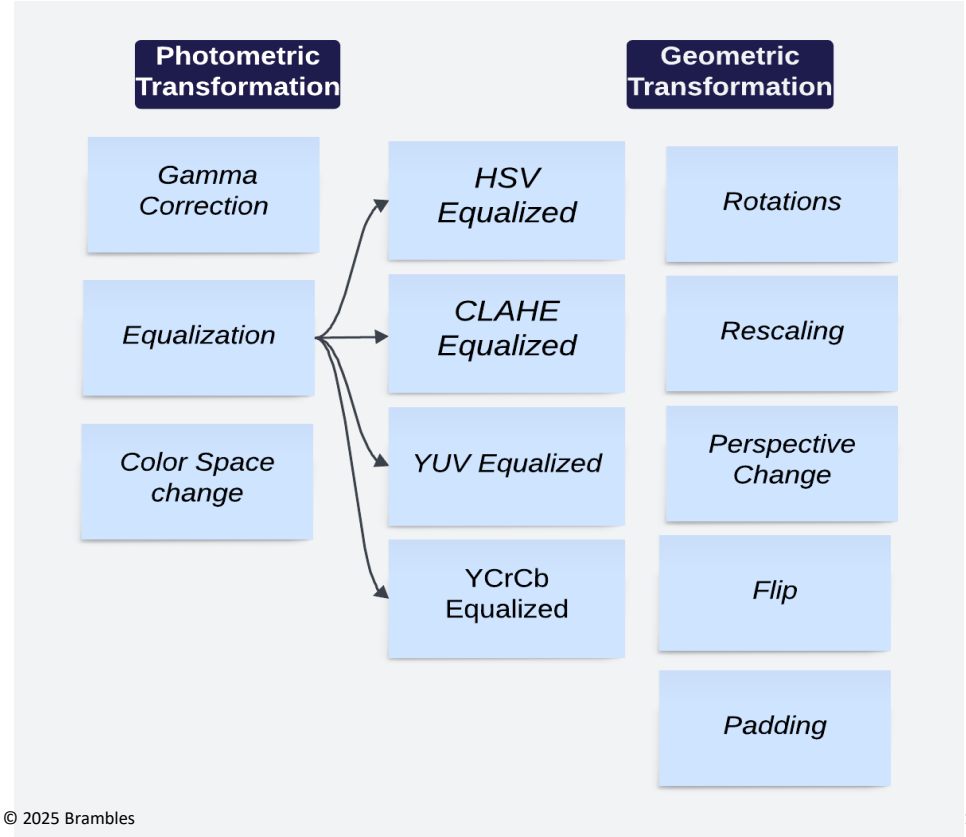
Over exposed	well-lit	Under exposed

Image Enhancement

- Transformed images improve human annotation accuracy and can be used for data augmentation



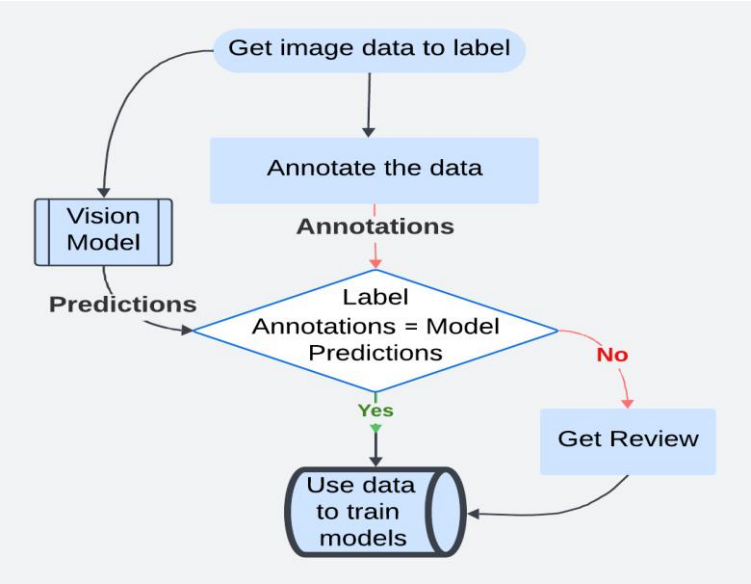
Histogram Equalization

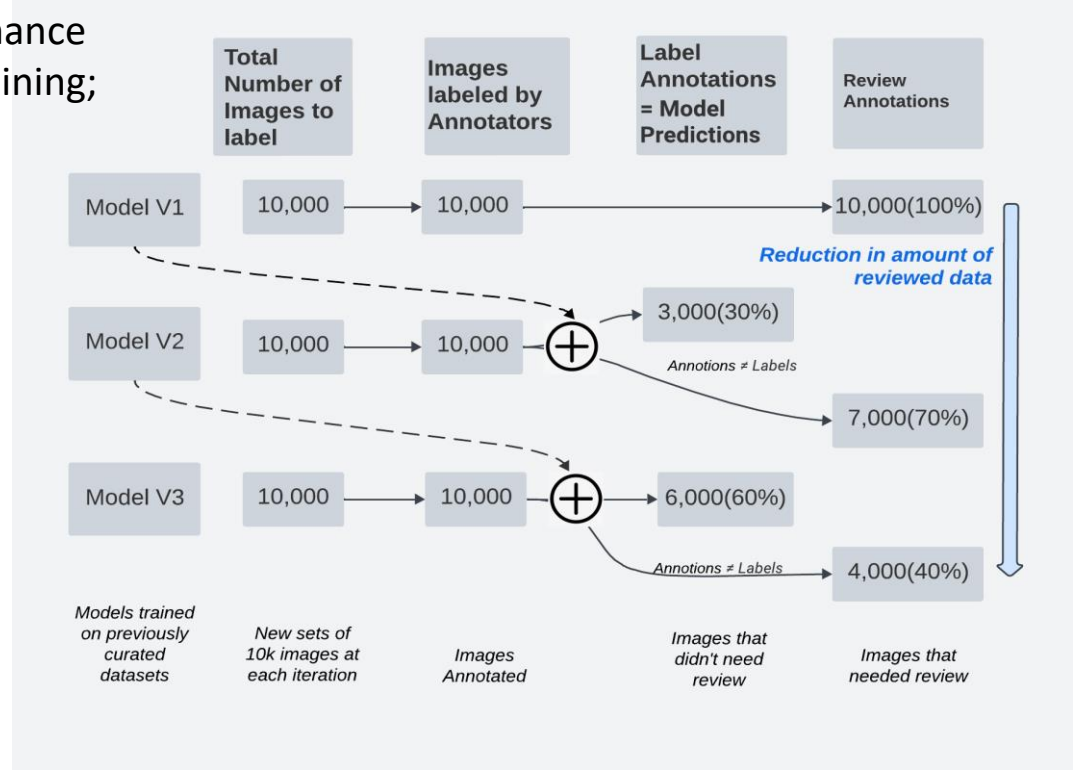


Annotation & Training

Data Labeling and Review

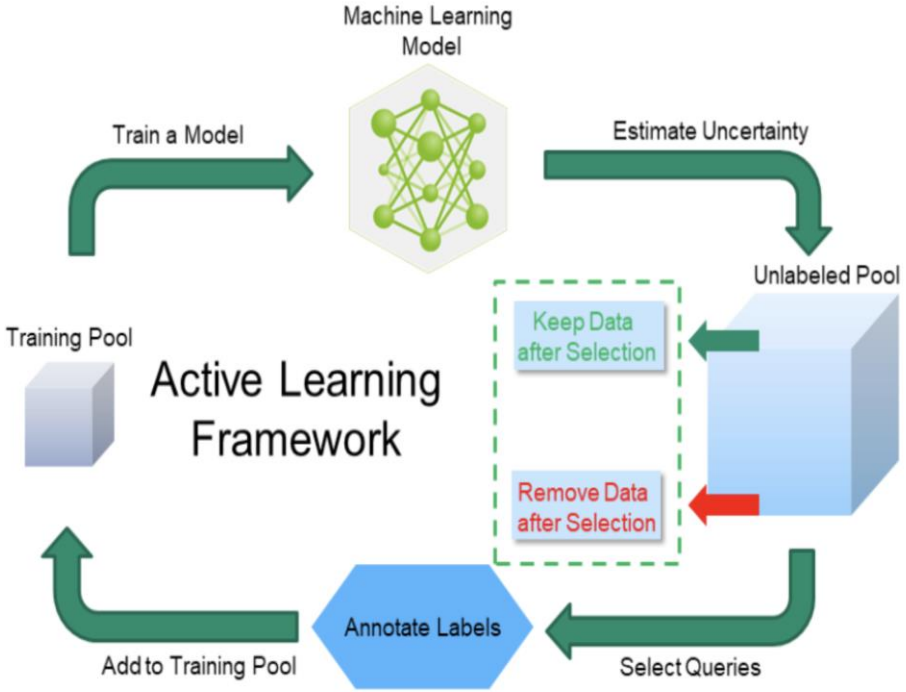
- Annotations may contain errors
- Incorrect labels can impact model performance
- Quality checks are essential for reliable training; labeling is expensive



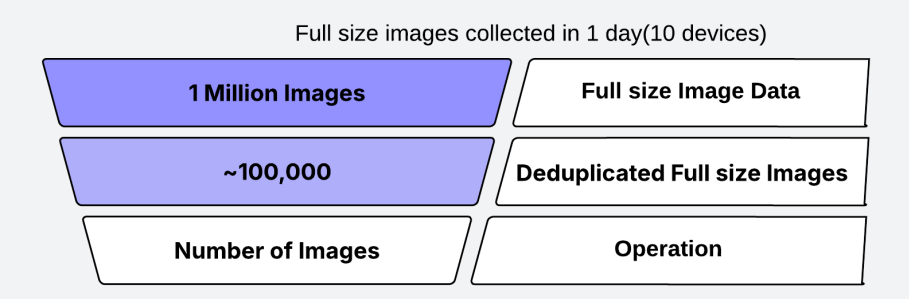


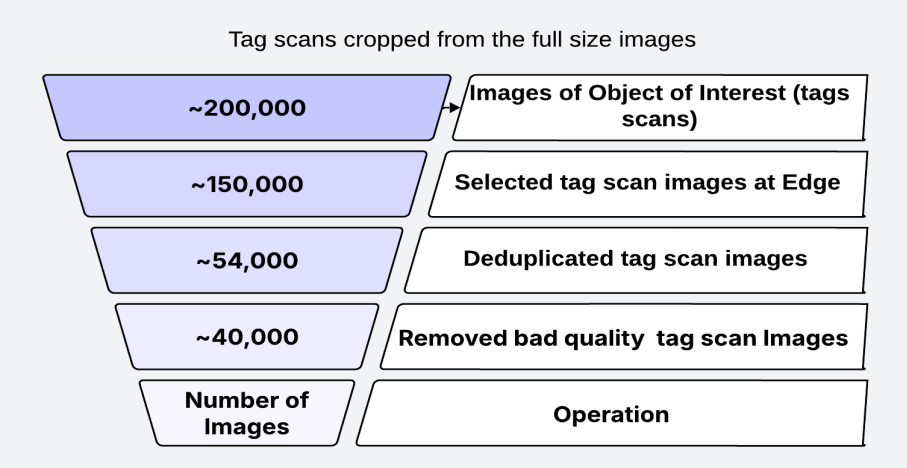
Active Learning

- **Label efficiency** – fewer labels needed
- **Cost savings** – reduce annotation effort
- **Targeted improvement** – focuses on hard cases
- **Faster convergence** – improves model quicker
- **Human-in-the-loop** – more control & feedback
- **Works with rare data** – finds edge cases
- **Requires uncertainty estimation** – can be tricky
- **Complex pipeline** – harder to implement
- **Sampling bias risk** – if query strategy is flawed



Data Reduction Metrics





Strategy	Images to label	Selection based on active learning
Active Learning	40,000	12,000(30%)
Random Selection	40,000	20,000 (50%)

Conclusion

Efficient Data Management

- Systematic selection of relevant images helped in optimizing storage and processing
- Reduced redundancy while preserving valuable information for modeling

Improved Model Performance

- Curating high-quality and diverse datasets enhanced the accuracy and generalizability
- Aids in mitigating class imbalance and sheds light on areas of improvement

Scalability & Practical Impact

- Enables scalable solutions for real-time monitoring and predictive analytics

Future Work & Opportunities

- Exploring automated selection techniques using AI-driven active learning.
- Enhancing real-time edge processing for continuous model updates.

References

- https://arxiv.org/pdf/2304.07193
- https://arxiv.org/pdf/2212.08035
- https://www.mathworks.com/help/images/ref/brisque.html
- https://images.app.goo.gl/NDZGEoWUii2p4weo9
- https://ieeexplore.ieee.org/abstract/document/5217220

Questions?