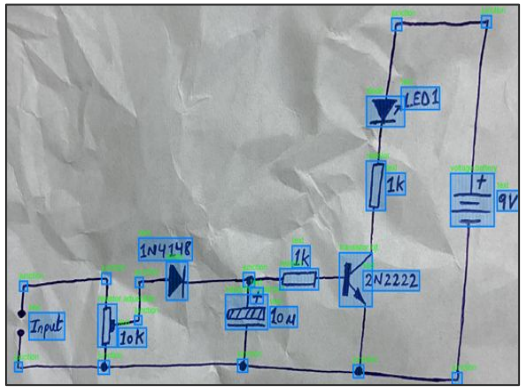
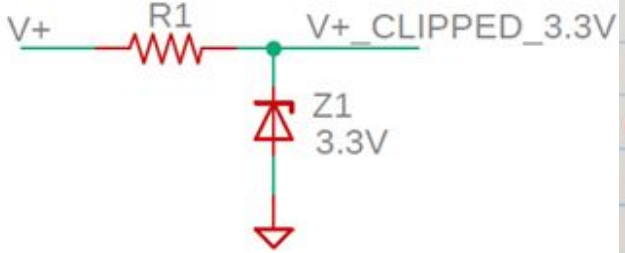
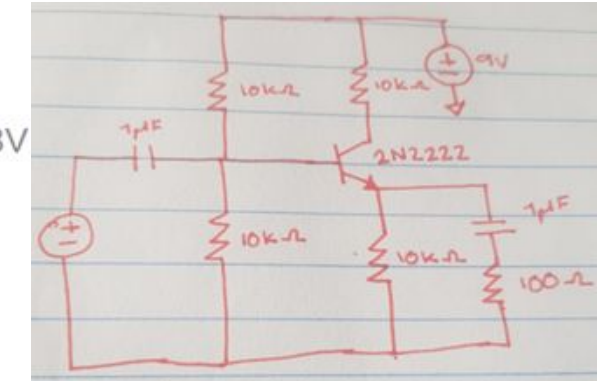
Master Thesis Machine Learning - Final Presentation





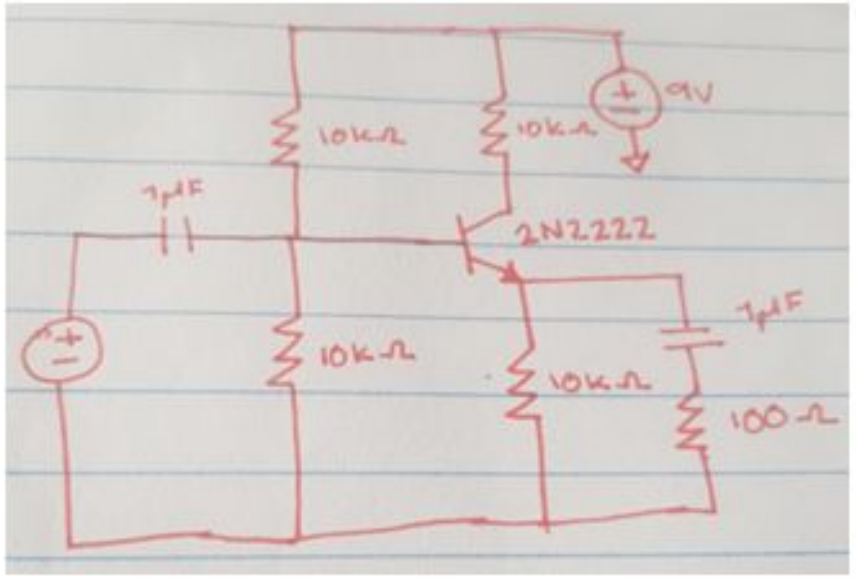
Multimodal Deep Learning for Automated Schematic Analysis

Vojtěch Sýkora

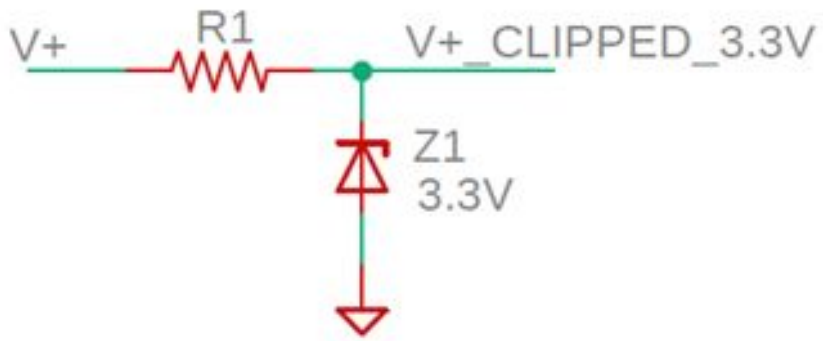
Problem & Summary

Short Summary

- Two datasets, both electric circuits
- Object detection of electric symbols
- I showed how targeted loss choices beat architecture tweaks for cross-domain symbol detection and why a smaller SOTA VLM is currently worse.
- The aim was a robust domain transfer of an existing pipeline.



(a) CGHD example (drafter_1/C2_D2_P1)

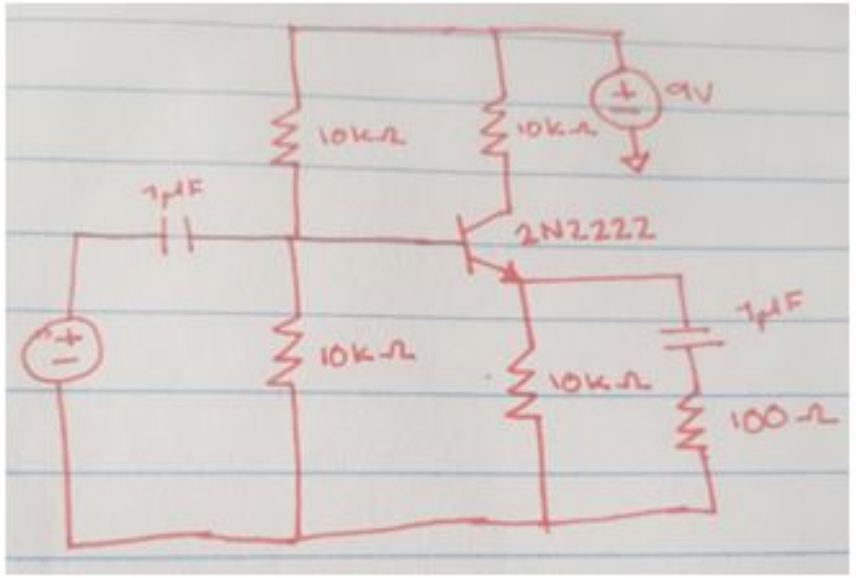


(b) Computer-generated schematic similar to RPi images[1]

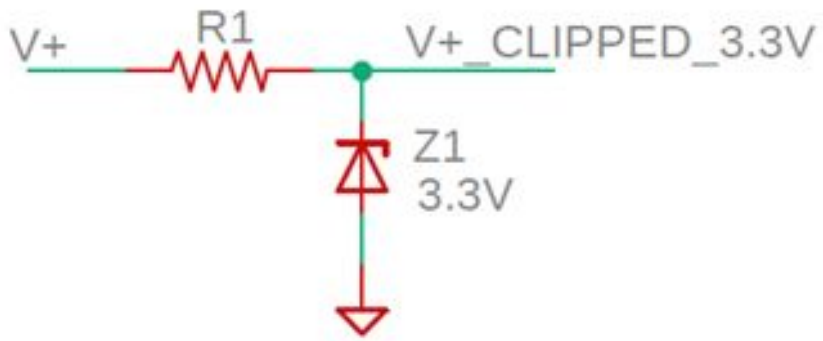
Problem

Object detection of electric symbols

- symbol class imbalance
- domain shift (hand-drawn → CAD)



(a) CGHD example (drafter_1/C2_D2_P1)



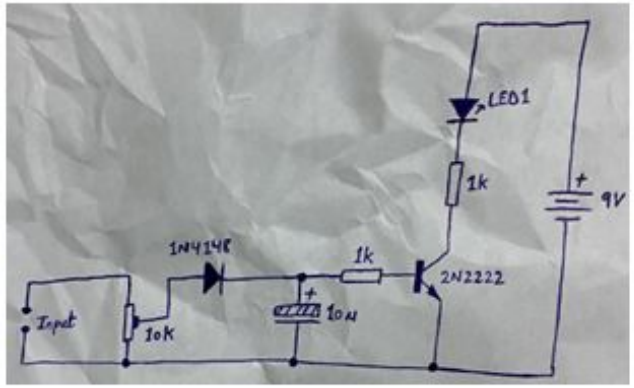
(b) Computer-generated schematic similar to RPi images[1]

Model Architecture

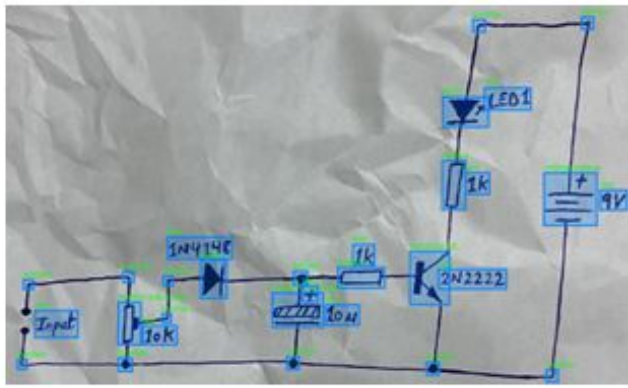
What I built on

- Modular Graph Extraction paper

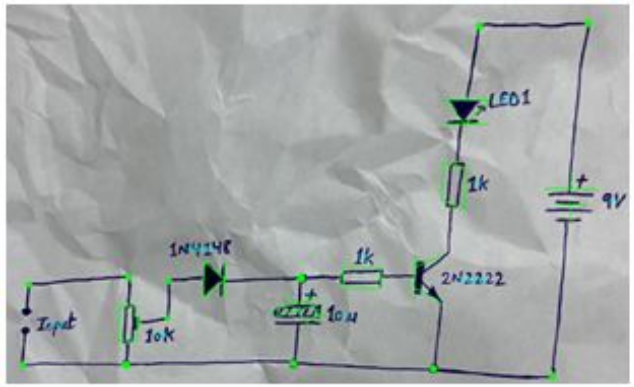
- Focus: object detection stage
- (Faster R-CNN + ResNet-FPN Backbone)
- Everything downstream benefits or suffers from its errors.



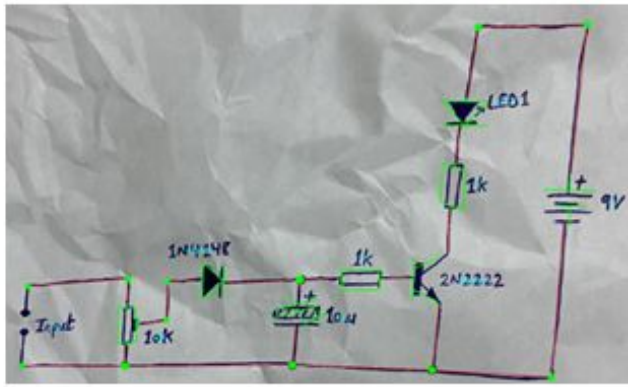
(a) Raw input image



(b) Object detection

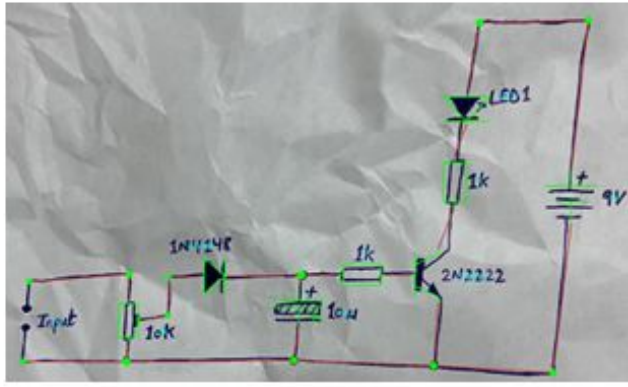


(c) Orientation and text recognition



(d) Edge extraction

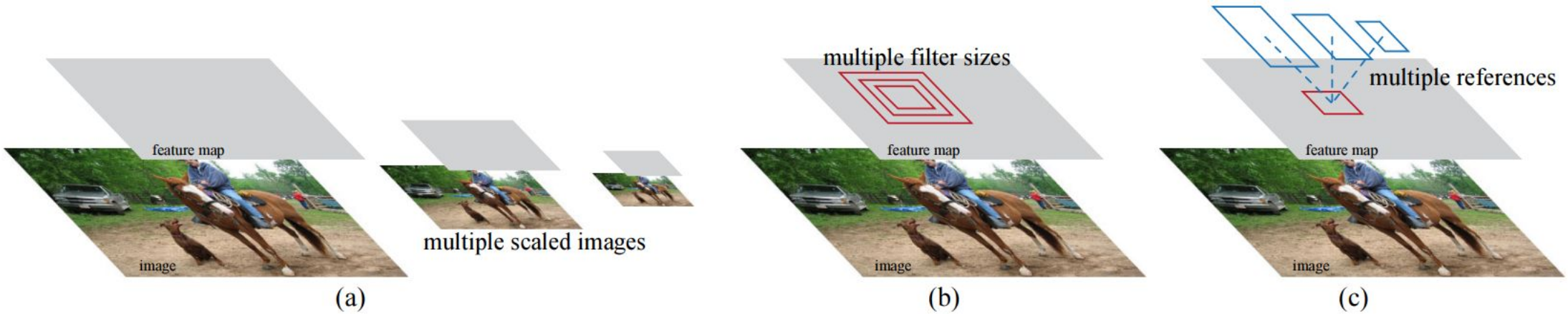
(e) Edge line segments



(f) Final graph rectification

Resnet-152-FPN Backbone: Object Detection

- **Backbone (Resnet-152-FPN)** builds multi-scale feature maps
  - called Feature Pyramid Network (FPN)



Faster R-CNN: Object Detection

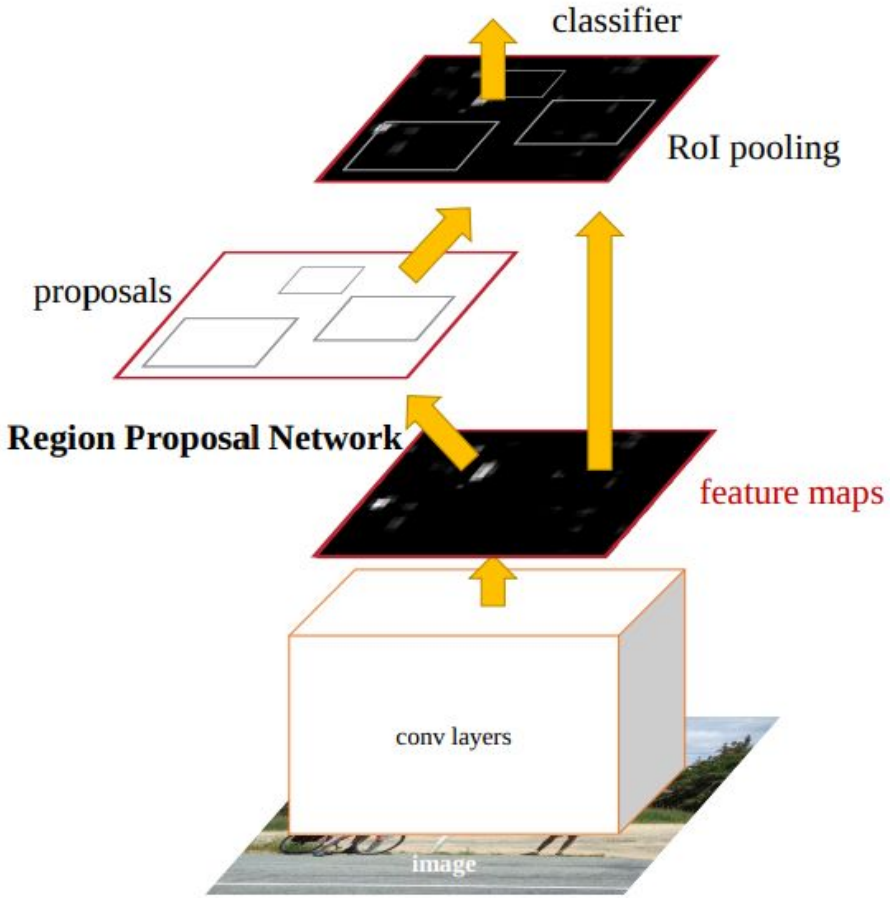
Two-stage detector

First Stage:

- **Region Proposal Network (RPN)** proposes candidate regions from every scale of FPN backbone

Second Stage:

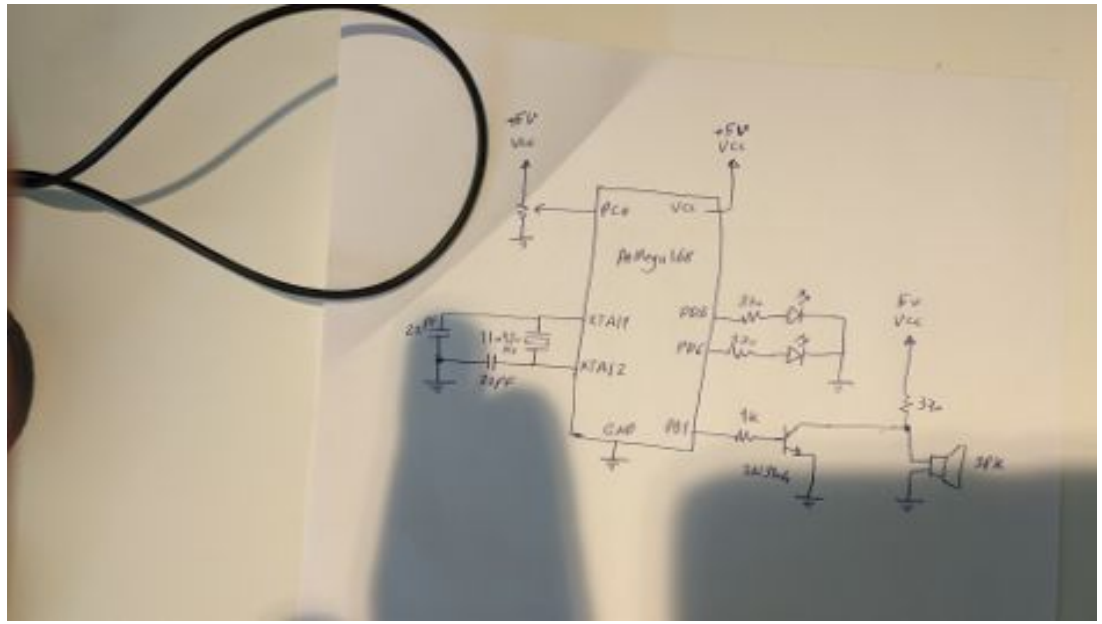
- **RoI pooling** crops fixed-size features from each proposal without misalignment.
- **Classifier** classifies each proposal and refines bbox coordinates.



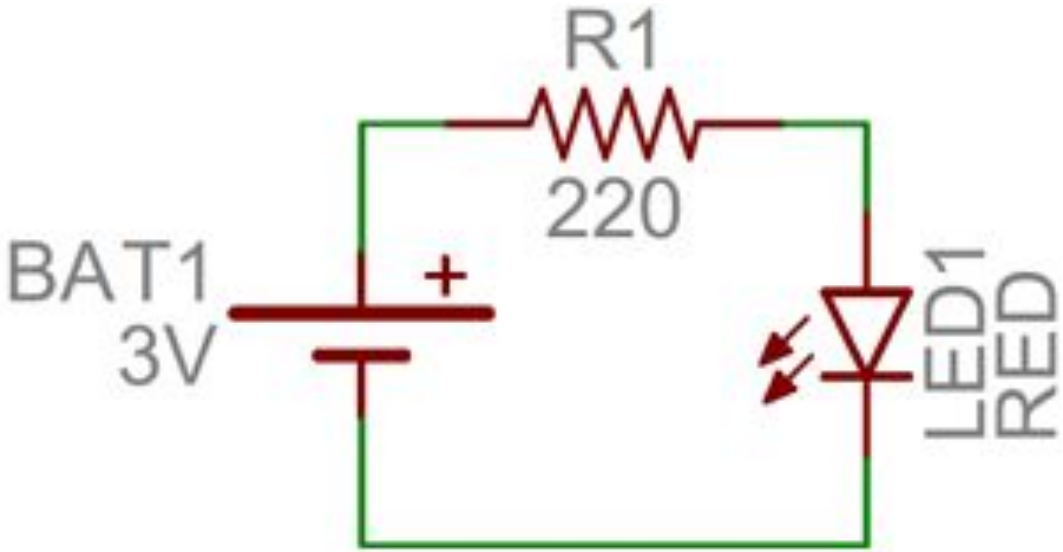
Datasets & Domain Gap

Datasets & Domain Gap

Circuit Graph Hand-Drawn Dataset (CGHD)

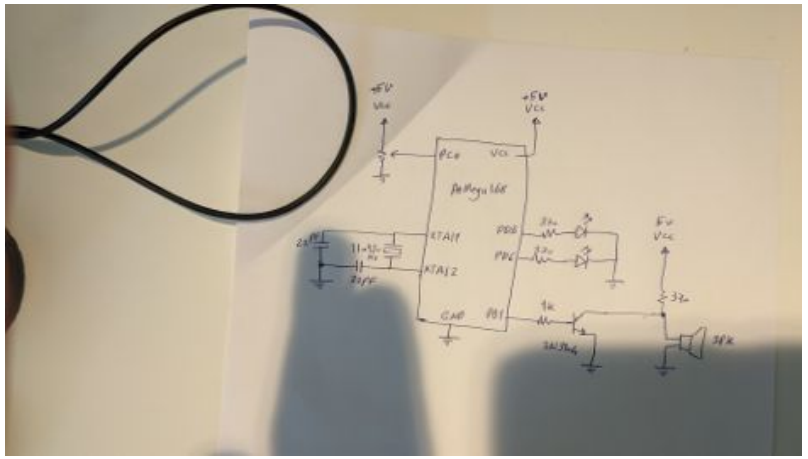


Raspberry Pi Dataset (RPi)

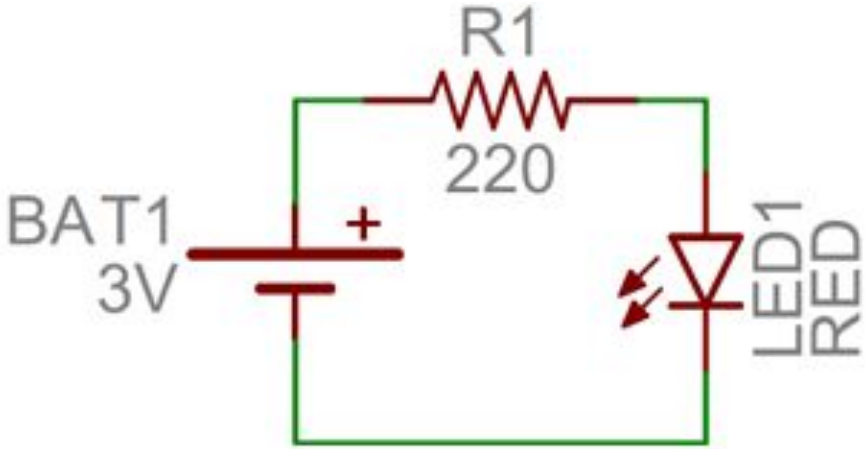


Datasets & Domain Gap

CGHD

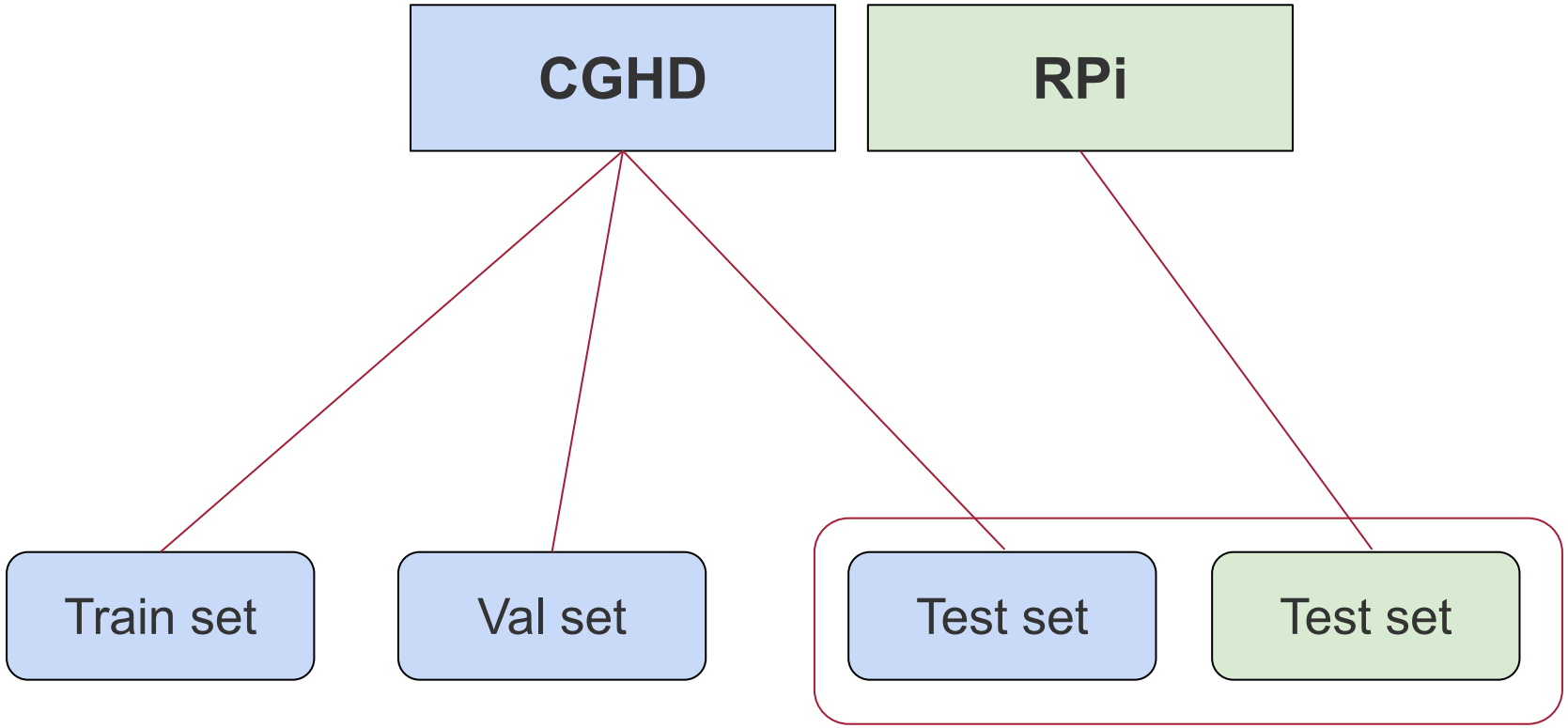


RPi



Dataset	Images	Boxes	Characteristics
CGHD	3,137	246,000	59 classes; large scale; class imbalance; real-world artefacts; mixed annotation standards
RPi	22	1,675	White backgrounds; coloured layers; class imbalance; dense text; massive stylistic gap compared to CGHD

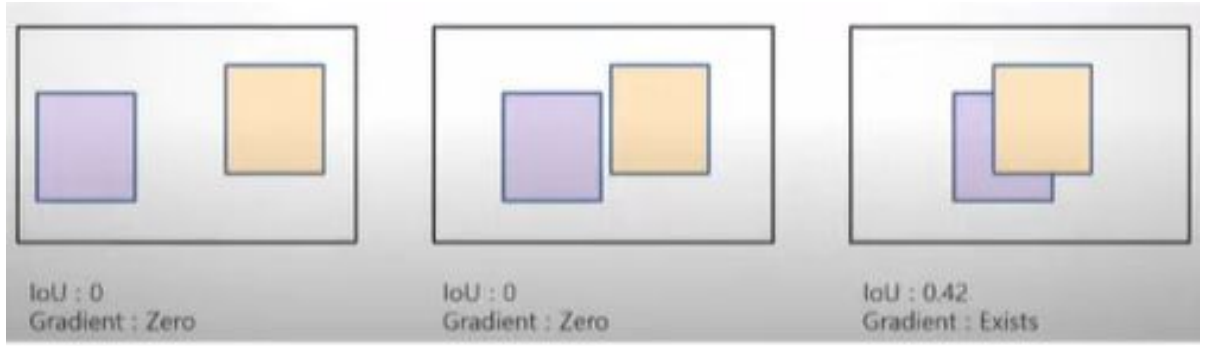
Datasets & Domain Gap



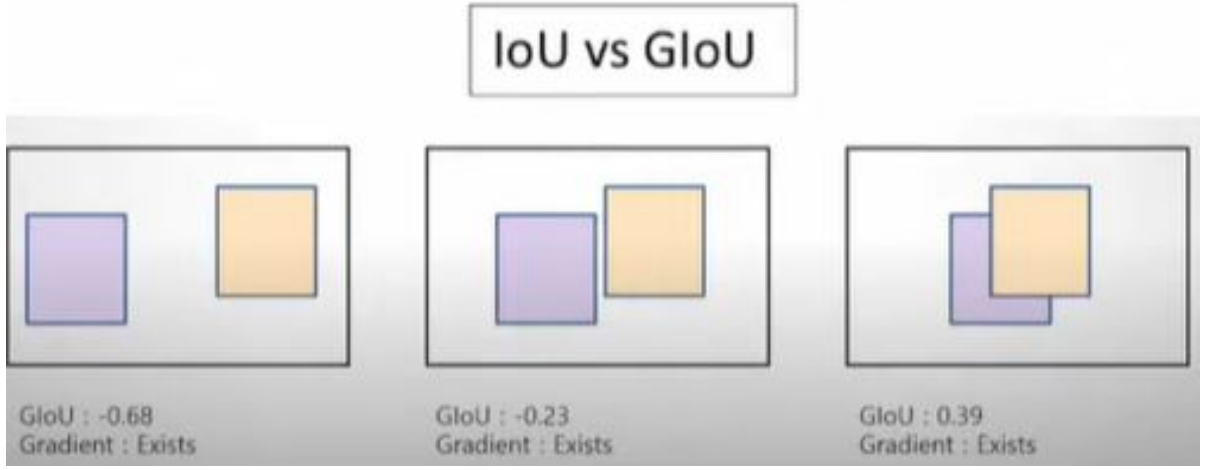
Model Enhancements

Faster R-CNN Enhancements (Losses)

- **Focal Loss**
  - classification loss
  - for class imbalance

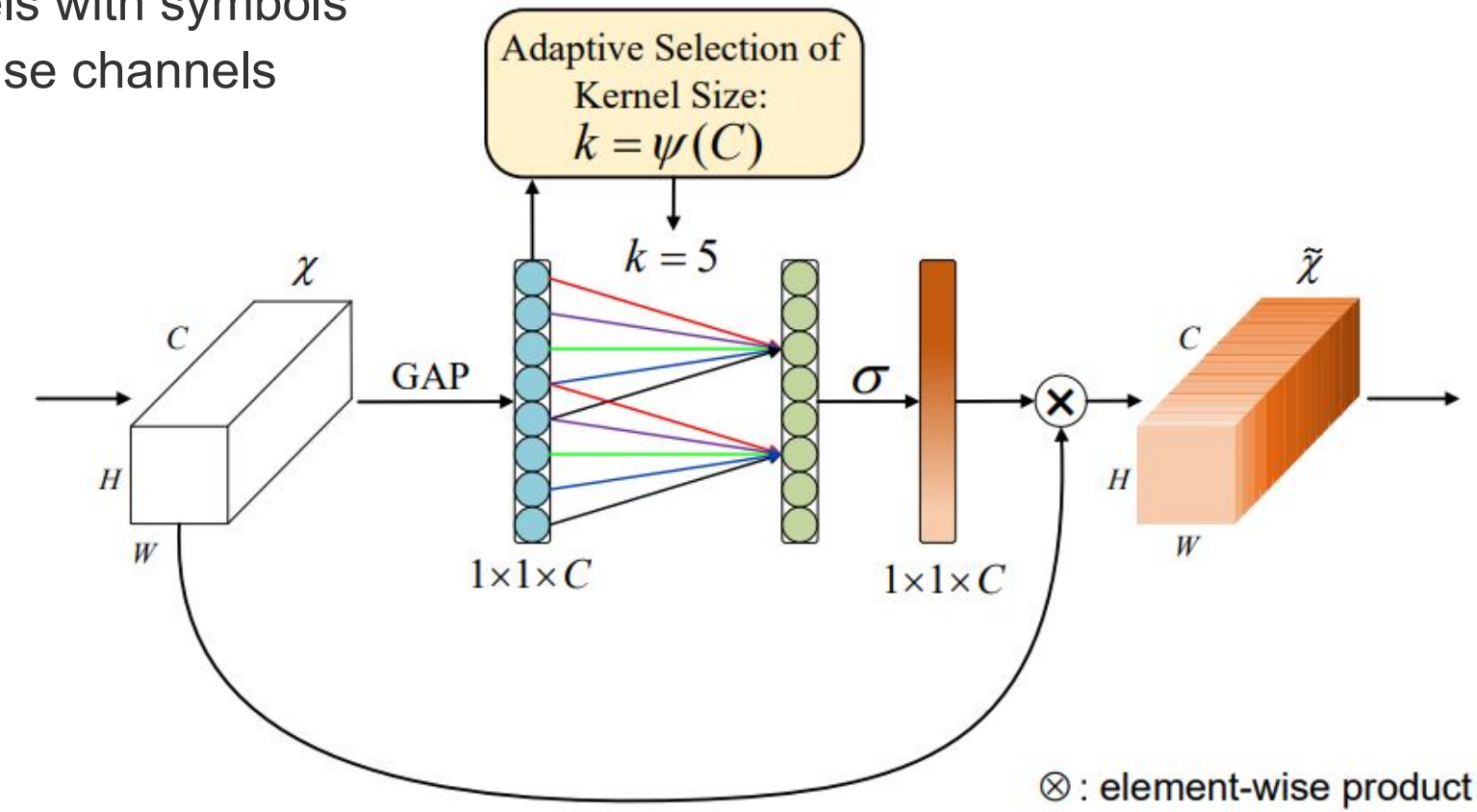


- **Generalized Intersection over Union Loss (GIoU)**
  - box regression loss
  - for localisation stability



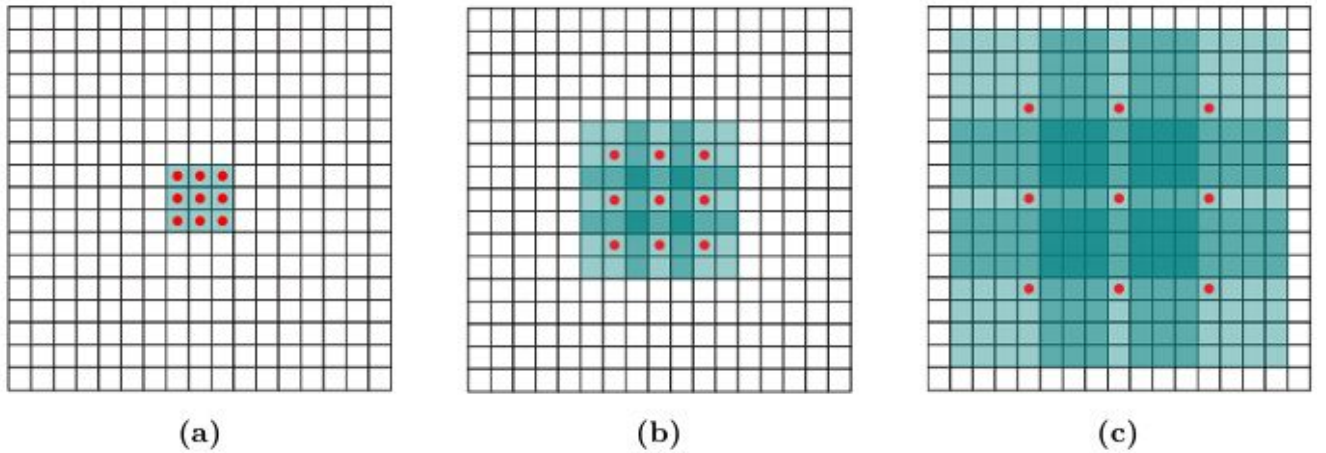
Faster R-CNN Enhancements (Architecture)

- **Efficient Channel Attention (ECA)**
  - emphasizes channels with symbols
  - avoiding domain noise channels



Faster R-CNN Enhancements (Architecture)

- **Dilated Convolutions**
  - you can see wider
  - without losing finer details



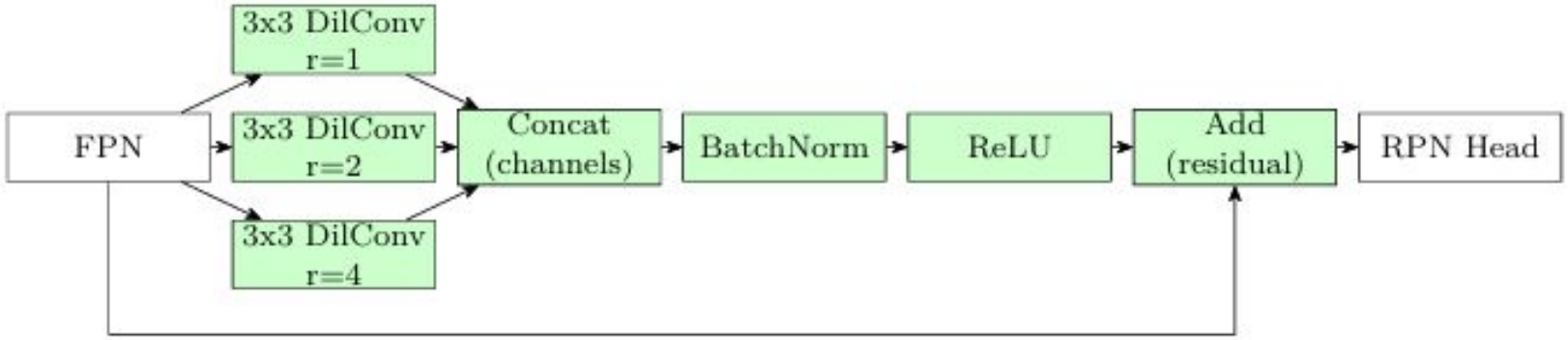
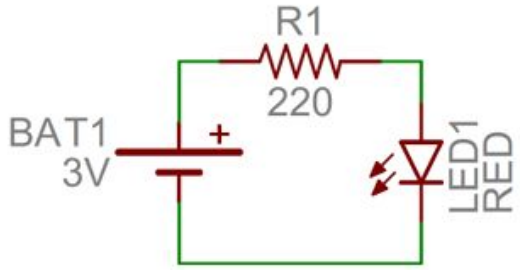


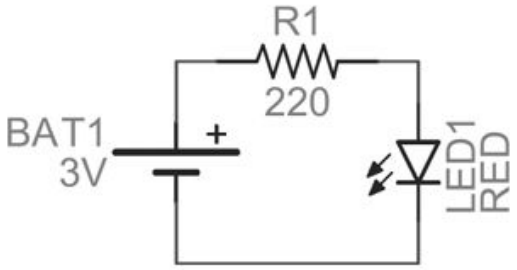
Figure 4.2.1.: All green blocks were added in our implementation of dilated convolutions addition to Faster R-CNN.

Faster R-CNN Enhancements

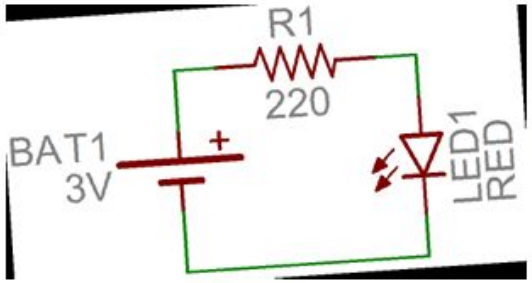
Image Transformations



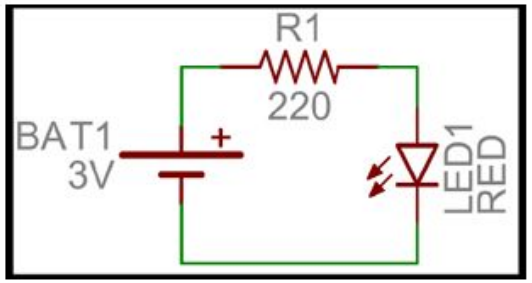
(a) Original Image



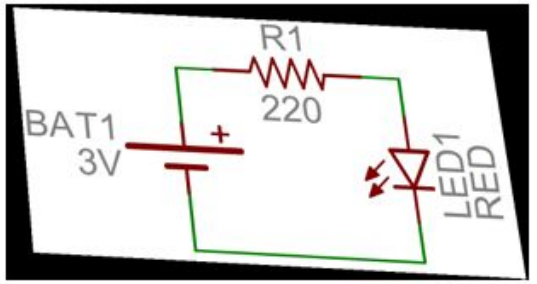
(b) Grayscale



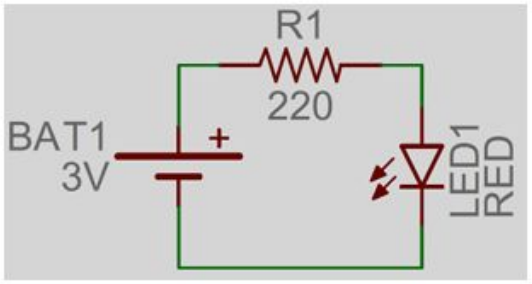
(c) Rotation



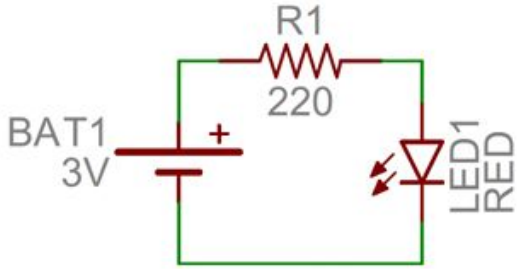
(d) Scaling



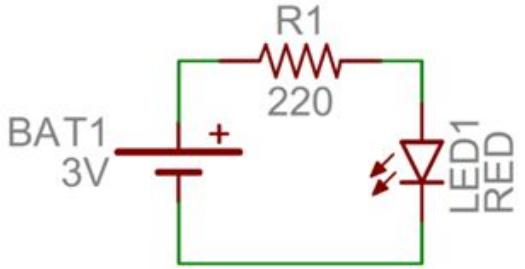
(e) Perspective



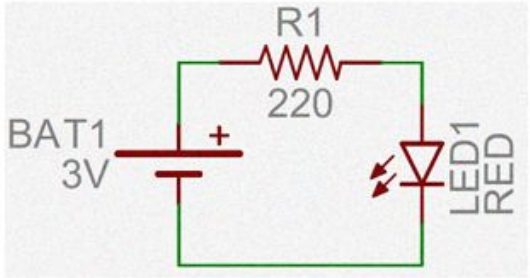
(f) Colour Jitter



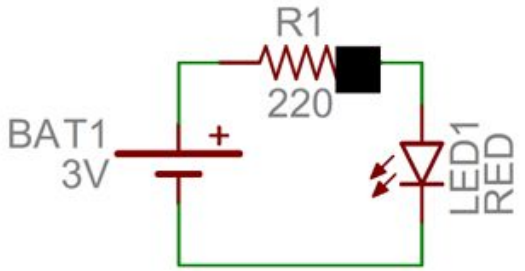
(g) Auto Contrast



(h) Gaussian Blur



(i) Noise Addition



(j) Random Erasing

Faster R-CNN Experiments & Results

Faster R-CNN Results (Maximum Bounding Boxes)

- Maximum boxes in one image
  - CGHD … 542
  - RPi … 436

- We tested limiting the number during inference

Max BBox	CGHD val
50	33.9
100	42.5
max	**44.3**
max+50	**44.3**

Faster R-CNN Results (Losses & Architecture)

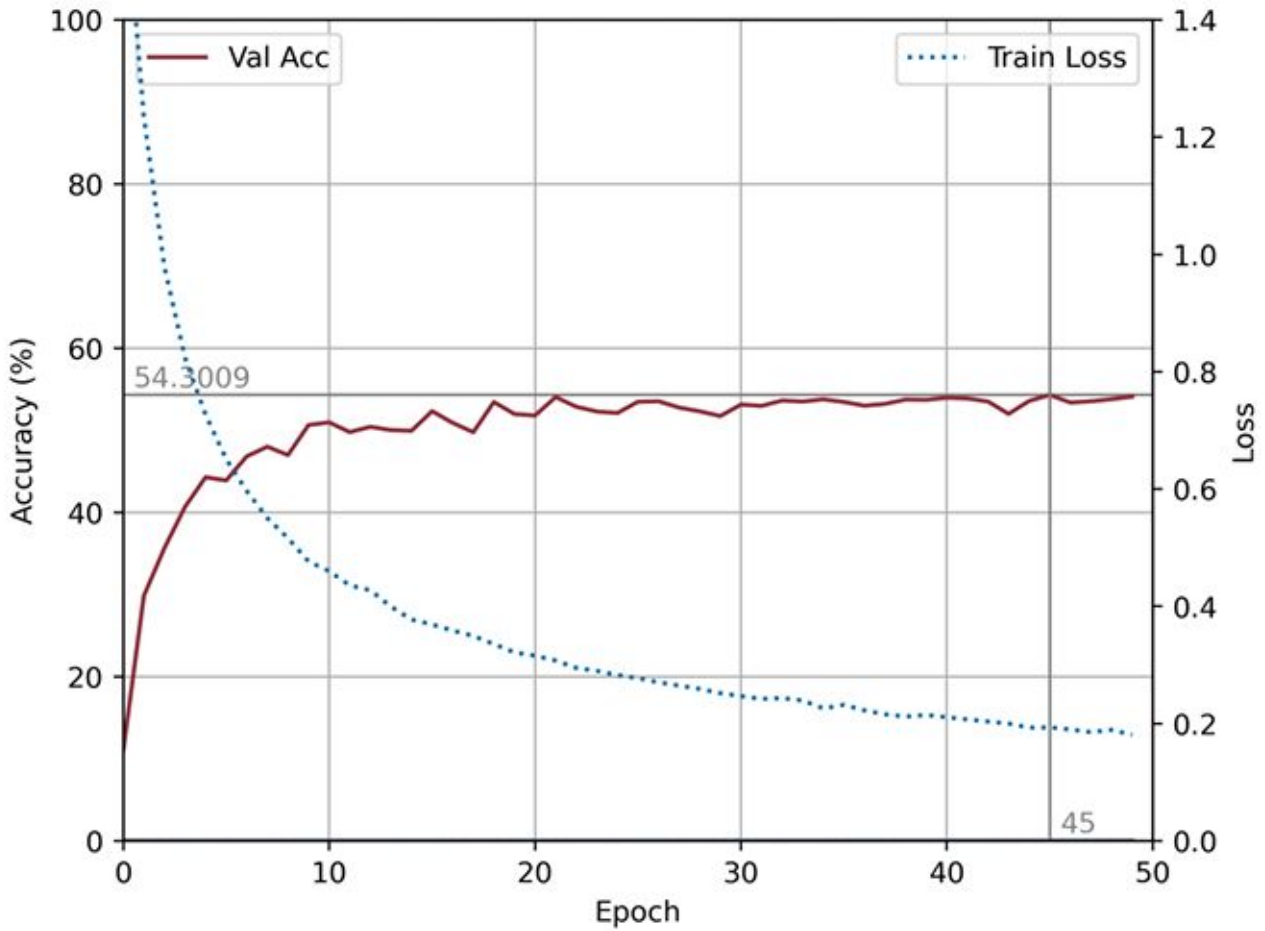
- 10% increase in validation mAP over baseline
- Losses outperformed architecture changes

Configuration	CGHD val mAP (%)
Baseline	44.3
Focal loss	54.2
GIoU loss	54.2
Focal + GIoU	**54.3**
ECA	53.3
Dilated convolutions	52.2
Focal + GIoU + ECA	53.4
Focal + GIoU + Dilated	53.1

Faster R-CNN Results (Losses & Architecture)

- Best configuration: Focal + GIoU

Parameter	New Value	Old Value
learning rate	0.0035	0.01
momentum	0.9	0
gamma	0.98	0.99
batch size	1	3
box detections	542	500



Faster R-CNN Results (Image Transformations)

- Training … (Focal + GIoU) + Image Transformation

- 11% increase over baseline
- 1% Increase because of ColorJitter Transformation

Configuration	CGHD val mAP (%)
Baseline	44.3
Focal + GIoU	54.3
ColorJitter	**55.3**
RandomErasing	54.7
RandomGrayscale	54.6
RandomAutocontrast	54.5
GaussianBlur	54.2
GaussianNoise	53.6
RandomPerspective	47.2
RandomRotation	45.3
RandomAffine	43.2

Faster R-CNN Results (Evaluating on both datasets)

- Better at validation set does not mean better at test set
- Large variety of artefacts in the CGHD dataset

Model	CGHD test mAP (%)	RPi mAP (%)
Baseline	41.3	28.3
Focal + GIoU	**48.6**	**32.0**
Focal + GIoU + ColorJitter	48.0	28.6
Improvement (Focal + GIoU)	+7.3	+3.7

Vision Language Model: Molmo

Vision Language Model (VLM): Molmo-7B-D

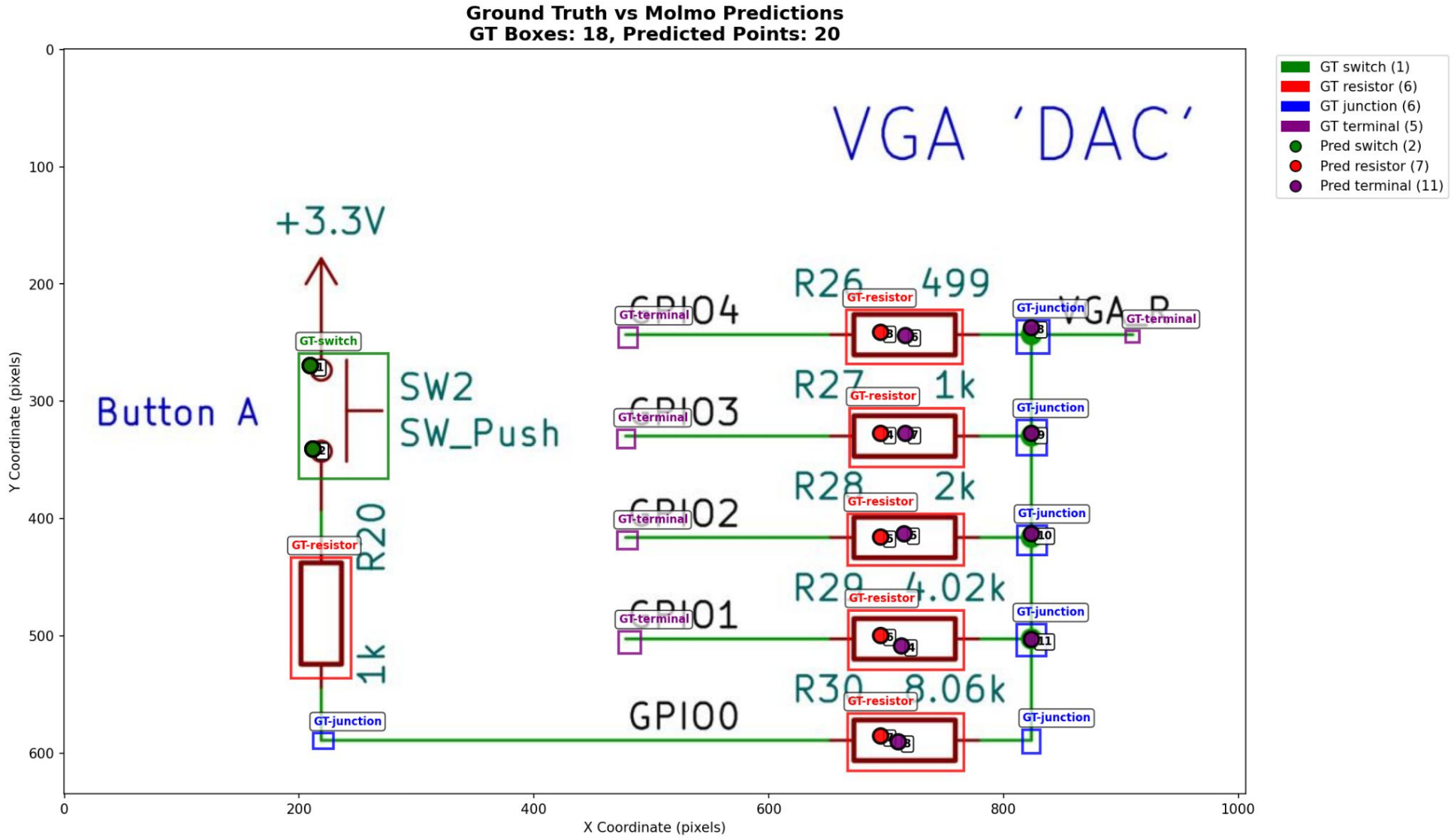
- open VLM from the Allen Institute of AI
- SOTA on par with GPT-4o, Gemini 1.5 Pro, or Claude 3.5 Sonnet
- Locally run 7 billion parameter model

Molmo-7B-D Results

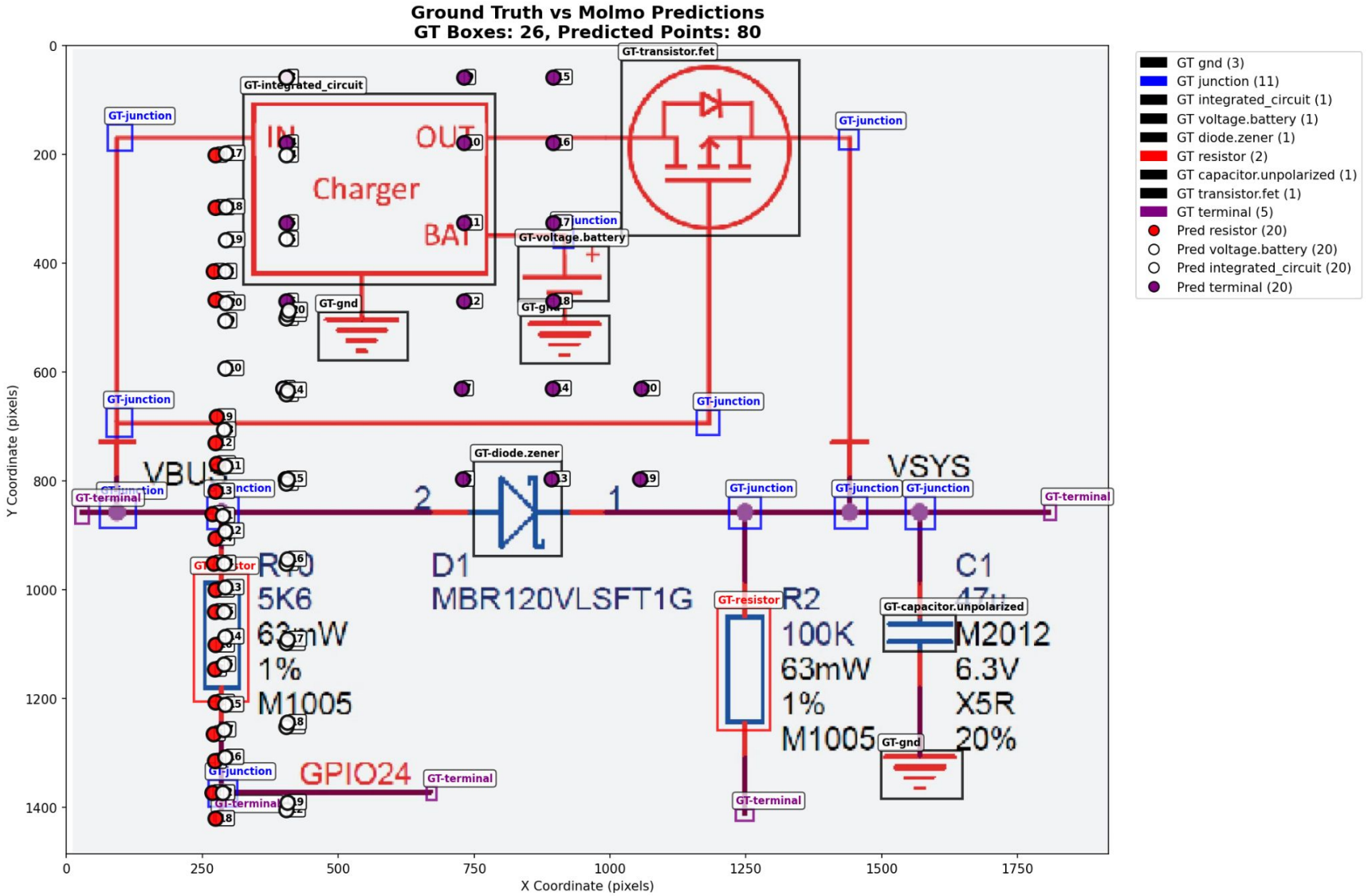
- Metric 1: accuracy of predicting symbol classes found in image (classification)
- Metric 2: accuracy of localising bounding boxes of symbols (localisation)
- tested on RPi dataset

	Metric 1: Symbols found (%)		Metric 2: Locations Correct (%)	
	Normal	Few-shot	Normal	Few-shot
Total avg	16.7	**29.6**	**9.1**	5.4
Avg small	11.3	33.8	2.4	10.0
Avg mid	15.1	23.2	20.2	6.3
Avg big	23.7	31.7	4.6	0.0

Molmo-7B-D results (Good)



Molmo-7B-D results (Bad)



Conclusion & Future Work

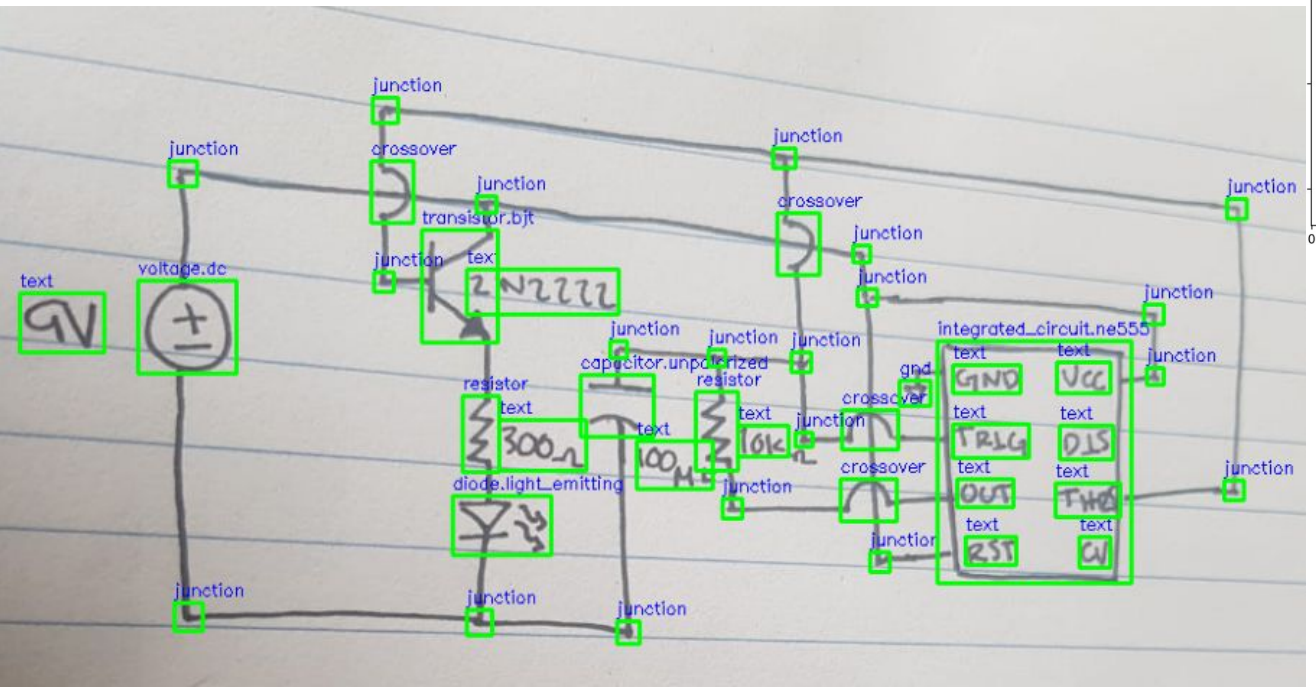
Conclusion

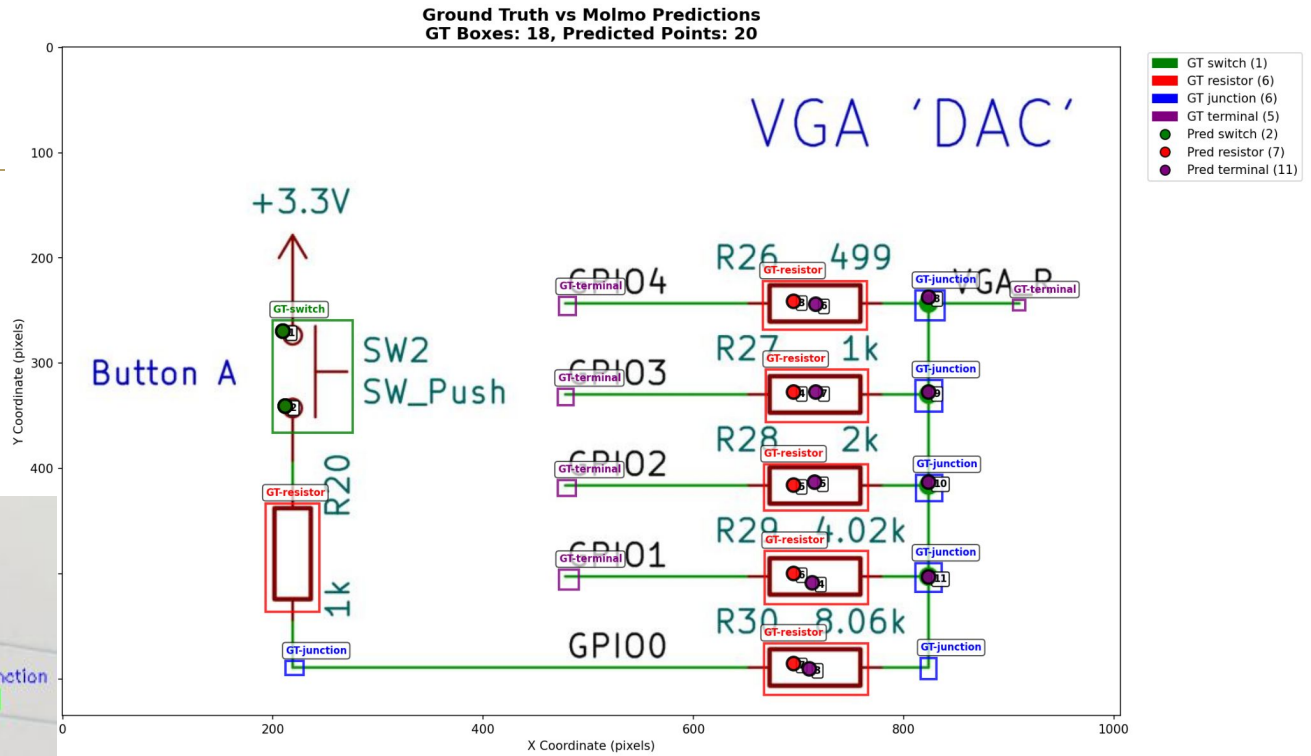
- Created a **new Raspberry Pi dataset.**
- Specialized **Faster R-CNN** performed better (trained on domain data).
- **Molmo-7B-D** showed some understanding of the task.
- **Increased performance** by significant **7.3%** (CGHD) and **3.7%** (RPi).
  - Showed domain transfer.
  - Targeted losses outperform architecture changes and image transformations.

Model	RPi accuracy (%)
Faster R-CNN Baseline	28.3
Faster R-CNN (Focal + GIoU)	32.0
Molmo-7B-D	12.9
Molmo-7B-D (Few-shot)	17.5

Future work

1. More **variety** and more **quality** of labelled **data.**
2. VLMs: **Larger** models, **agentic** workflows, image embeddings.
3. **Hybrid** combination of precise Faster R-CNN and context understanding of a VLM.
4. **Symbol standard normalisation** (mapping IEEE↔IEC) pre-detector.





Thank you for your attention.

References

[1] F. Thoma, J. Bayer, and Y. Li, A public ground-truth dataset for handwritten circuit diagram images, 2021. arXiv: 2107.10373 [cs.CV].

[4] J. Bayer, L. van Waveren, and A. Dengel, Modular graph extraction for hand written circuit diagram images, 2024. arXiv: 2402.11093 [cs.CV].

[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks”, in Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[22] M. Deitke, C. Clark, S. Lee, et al., “Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models”, arXiv preprint arXiv:2409.17146, 2024, Submitted: 25 September 2024; Revised: 5 December 2024; Accessed: 2 July 2025.