

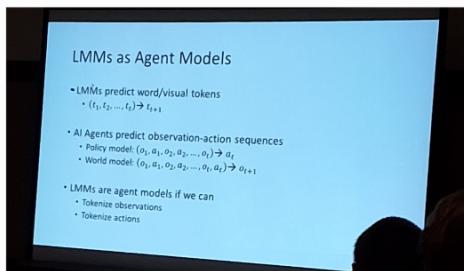
# [Computer Vision in the Wild]

## Opening remarks - speaker from Microsoft

- new benchmark in 2022 & 2023
- NO new dataset this year

## Theme: LMM in the wild

- AI agents



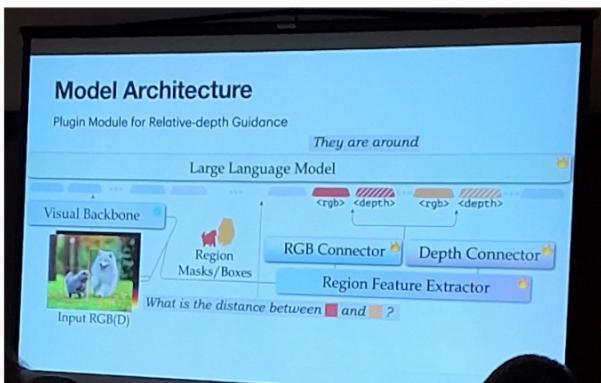
- Main point is to create specific tokenisers to accept the intended input

## Invited Talk : Spatial perception & control in the wild

- Spatial RGPT

<https://arxiv.org/abs/2406.01584>

- Talked about dataset creation
- Able to measure distance & size in 3D space



- mentioned training only depth connector first  
the depth + RGB, then the rest
    - AKA freeze others

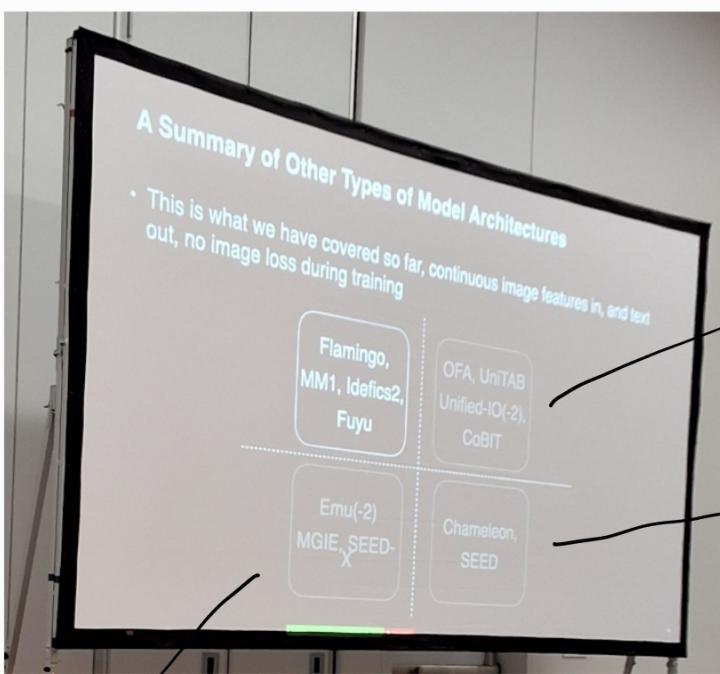
- Inject metric measurement into the training
  - Someone ask how robust w/ diff camera. And the training has the camera properties estimated.

- T no time for the spotlight talks

- Blink
  - multi agent VQA
  - what's in a name?
  - Vip Java

} the repos link can be found  
in the webpage

[Recent developments in vision Foundation Models] \* came in half way \* got reading. check webpage



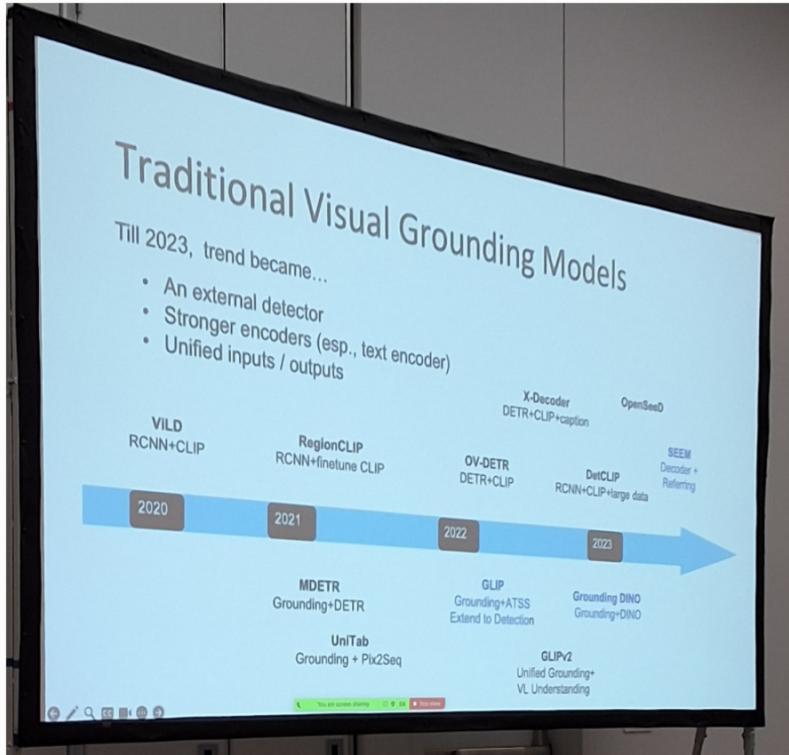
- (unifipol model example)
  - pro: unifipol archi.
  - cons: IPSS prep than diffusion
- order  
deorder
- Not very gd multimodal benchmark

not very gd multimodal benchmark

pro: strong image understanding

## LLM w/ fine-grained grounding - Apple Speaker

- current MM/LLM not perfect
  - ↳ 'object hallucinations'
  - ↳ not good spatial understanding
- use visual grounding to help
- visual language + grounding  $\Rightarrow$  better MM results
- GLIP / GLIP v2 < grounding DINO  
(2022) (2023)
- SPPM (2023)



- above is global image-perceiving, now we have fine-grained Region-level
  - ↳ training has additional tokens of region/pixels at input

<https://github.com/apple/ml-ferret>

↳ Basically saying this is good

<https://arxiv.org/abs/2404.07973v1> : FerretV2 (no repo currently)

- ↳ Resolution critical to fine-grained tasks
- ↳ They find that CLIP encode global patch better, DINOv2 encode local patch

### - Fine-grained pixel-level MLLMs

↳ Vista LLM (code coming soon)

↳ LISA

↳ GLaMM

<https://github.com/dvlab-research/LISA>

<https://github.com/mbzuai-oryx/groundingLMM>

### - Video based Region-level MLLMs

↳ PG-video - LLaVA

(or also aligned audio inputs (?)

<https://github.com/mbzuai-oryx/Video-LLaVA>

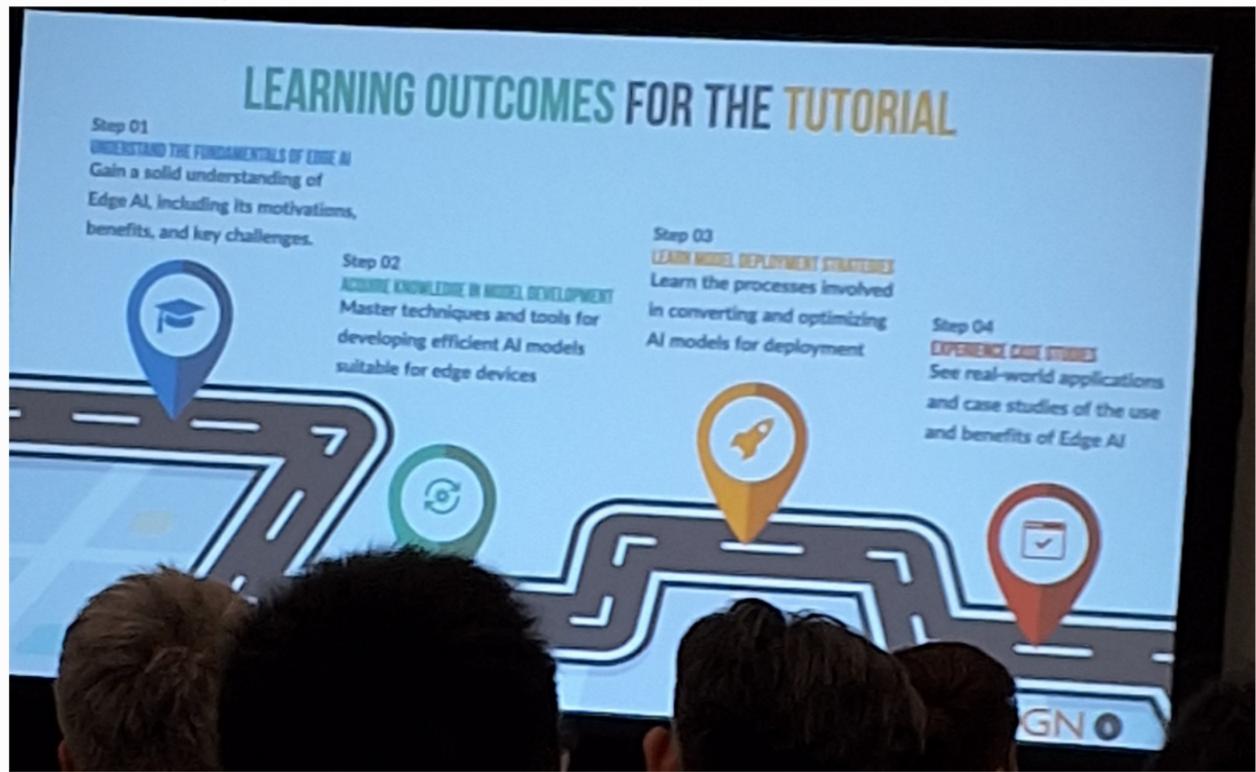
- I think good to watch the videos (not 3D based one)



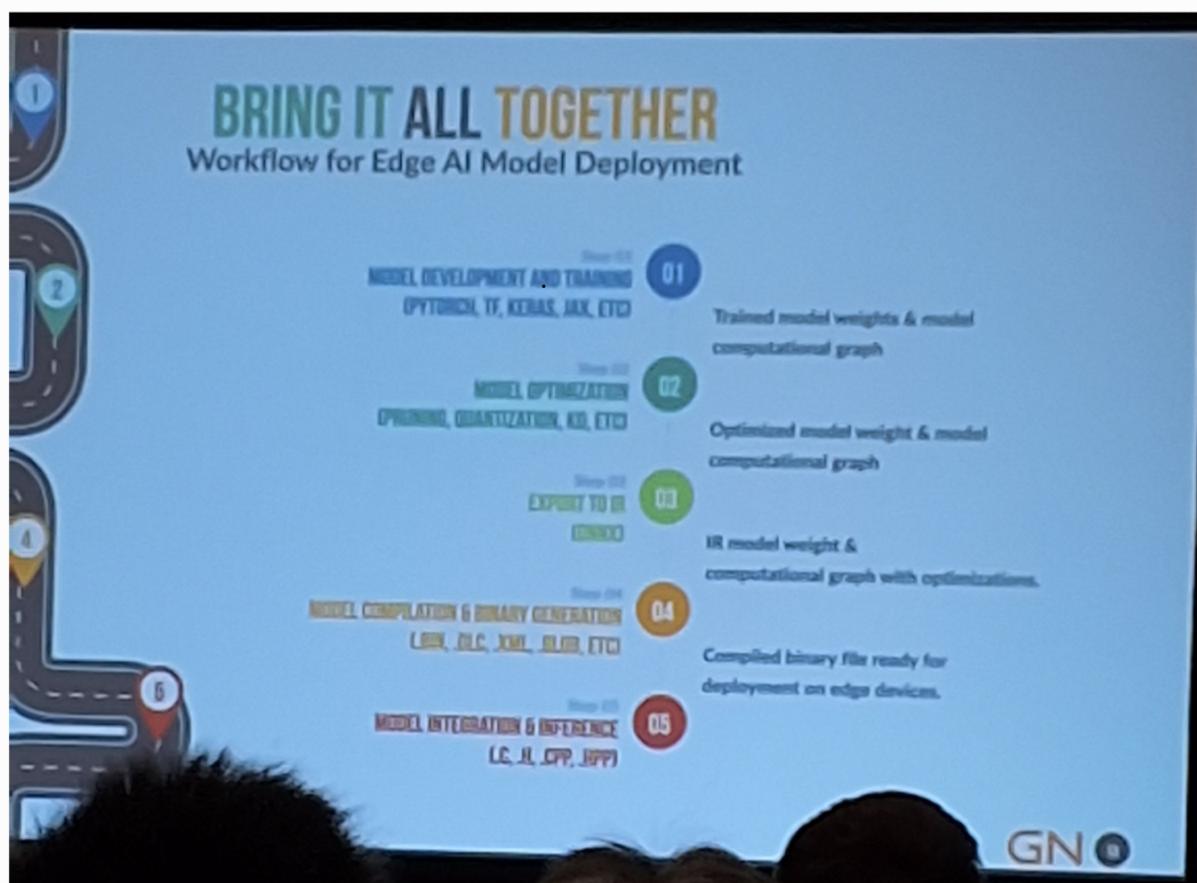
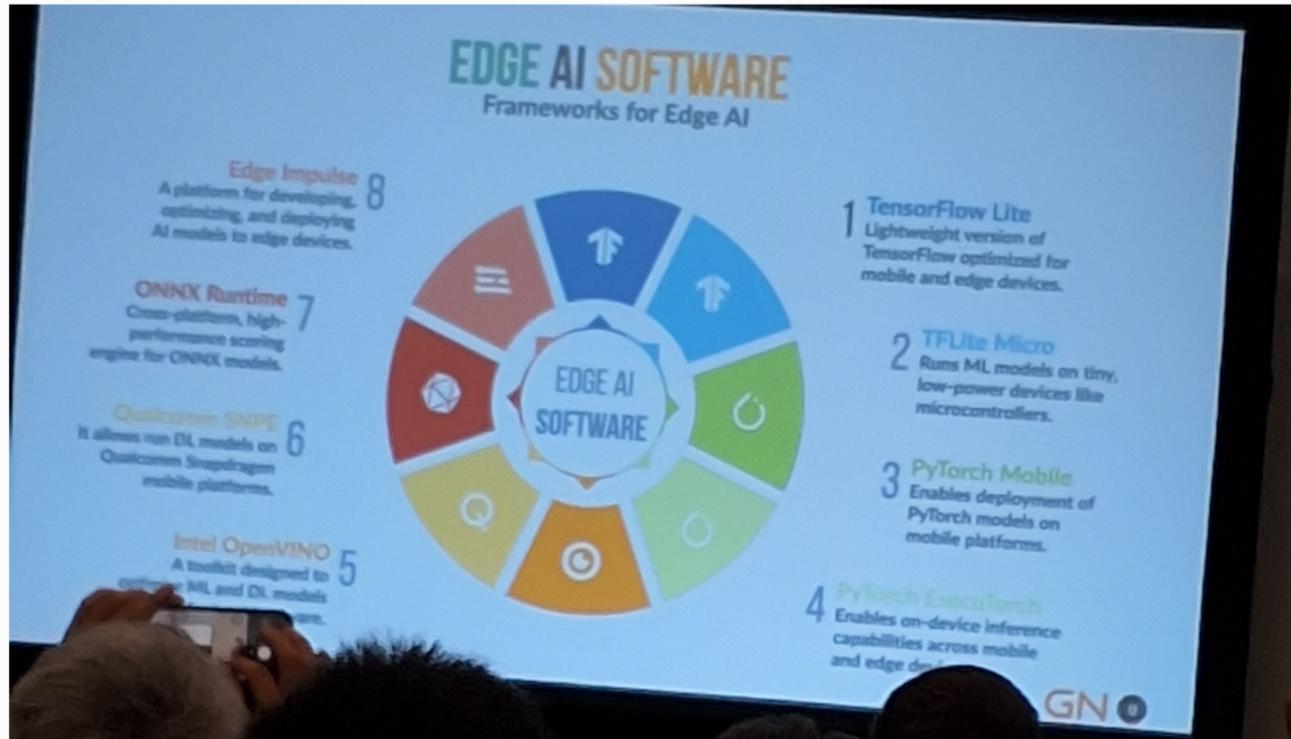
## [Edge AI in action]

\* I think I was wrong & thought model deployment is 1pm. Actually is at 3:45 pm. Other tutorial (vision foundation + robustness at inference) both

get long Q. Further, Vision got record, Robust got provide slides. So I just spot in Edge AI.



## Intro to Edge AI



## THE ROOFLINE MODEL

Operational Intensity (ops/byte)

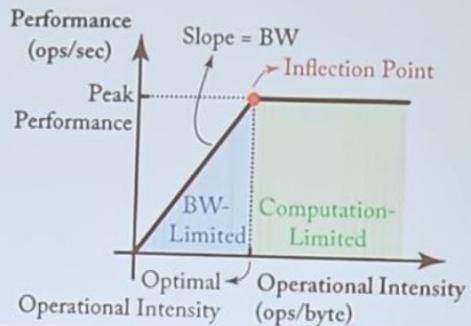
THE ROOFLINE MODEL IS A GRAPHICAL REPRESENTATION TO ILLUSTRATE AN ARCHITECTURE'S PERFORMANCE ACROSS DIFFERENT LEVELS OF OPERATIONAL INTENSITY.



Operational Intensity  
How computation-heavy an operation is relative to data movement.



Higher Operational Intensity  
More computations are performed for every byte fetched from memory.



THE ROOFLINE MODEL

## Model development for Edge AI

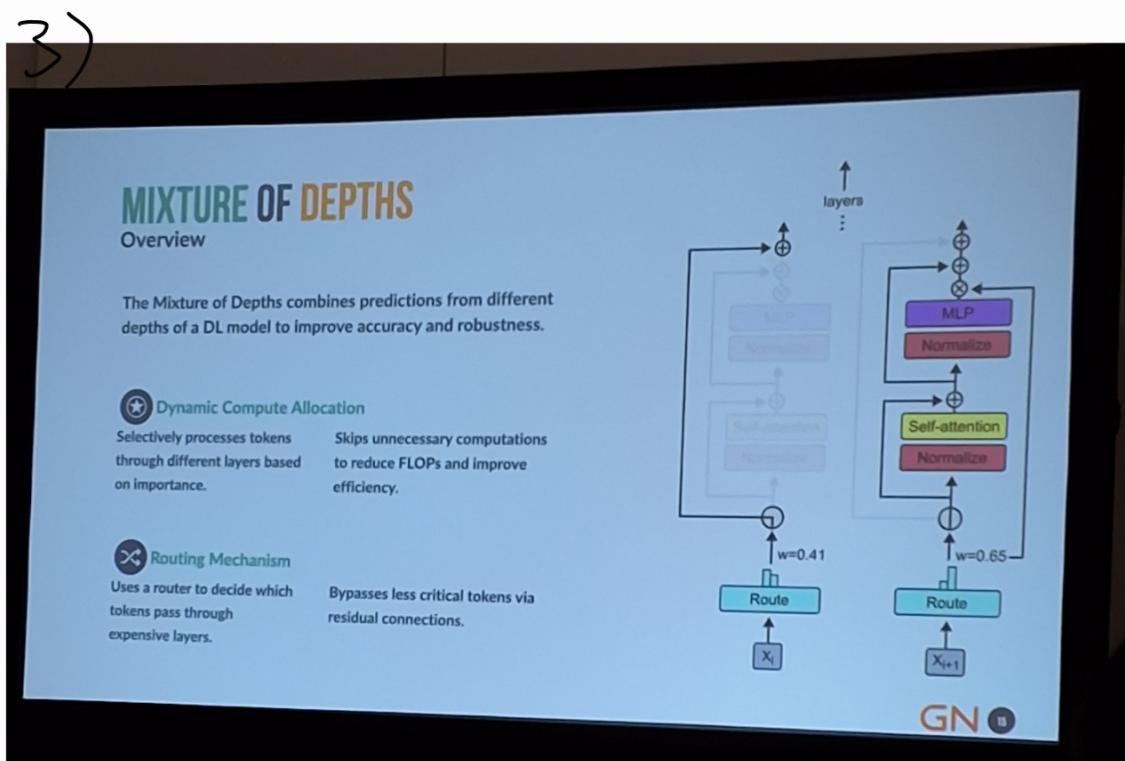
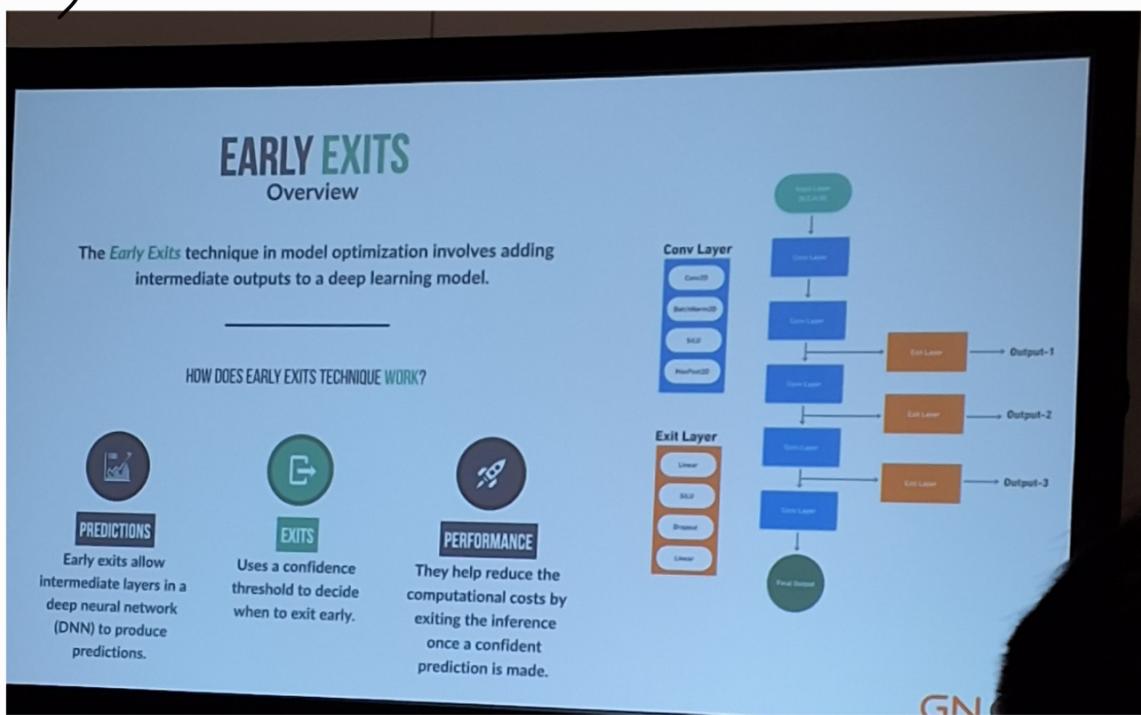
- Skip computation costly if feature size is large. Especially if is long, like from input to output
- Synthetic dataset create using blender
- 30 real, 70 synthetic gives them gd result
- Since their use case is video based, their dataset also is video based (like a video of person moving and instead of discrete images of different person at different location). This also mean that their architecture also have some temporal layers.
  - ↳ they have more video data vs static images
  - ↳ they freeze the GRU layers (temporal layer) when there is no video data in the training

## Model deployment



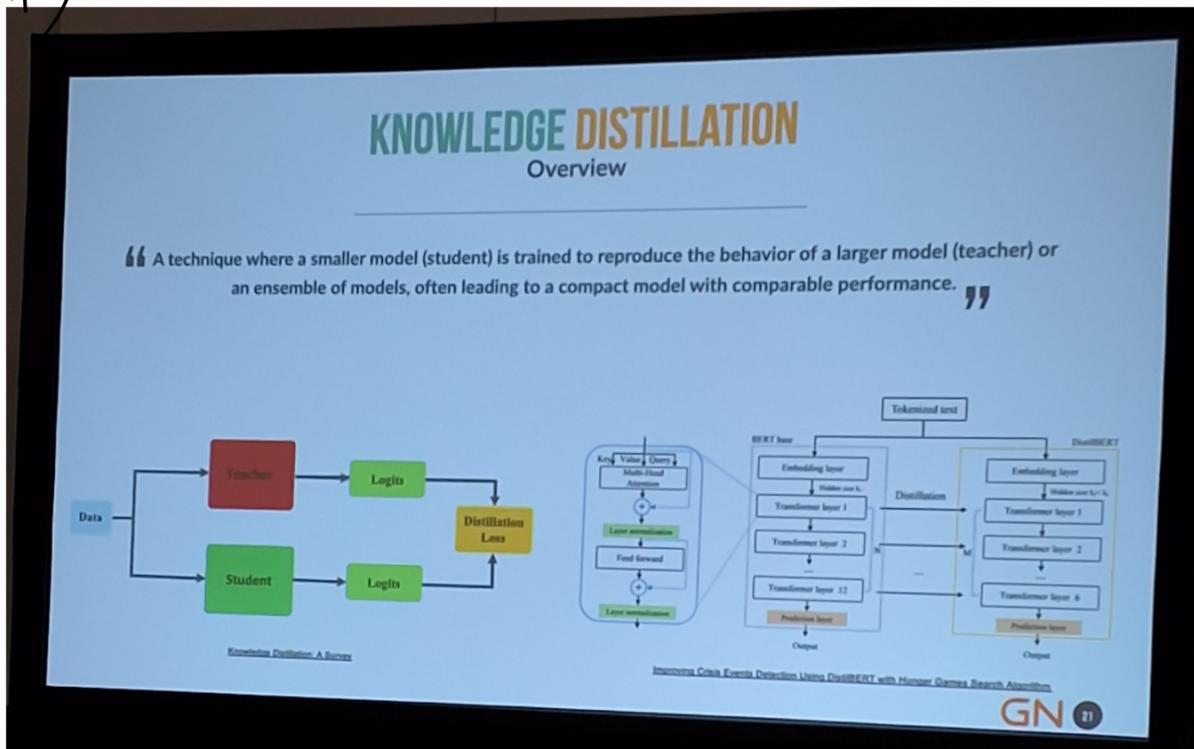
## Techniques

- 1) Neural architecture search to find the optimal architecture first
- 2)



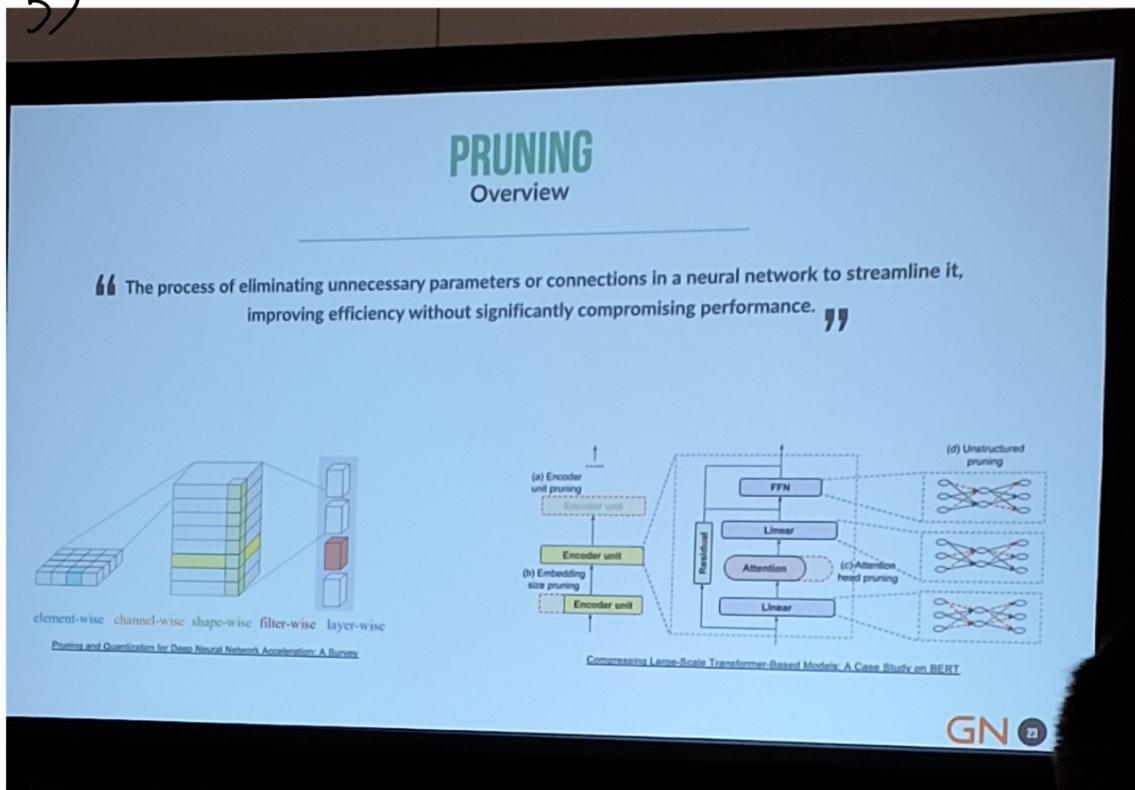
## - Fast ViT

4)



↳ Frozen teacher, Train student

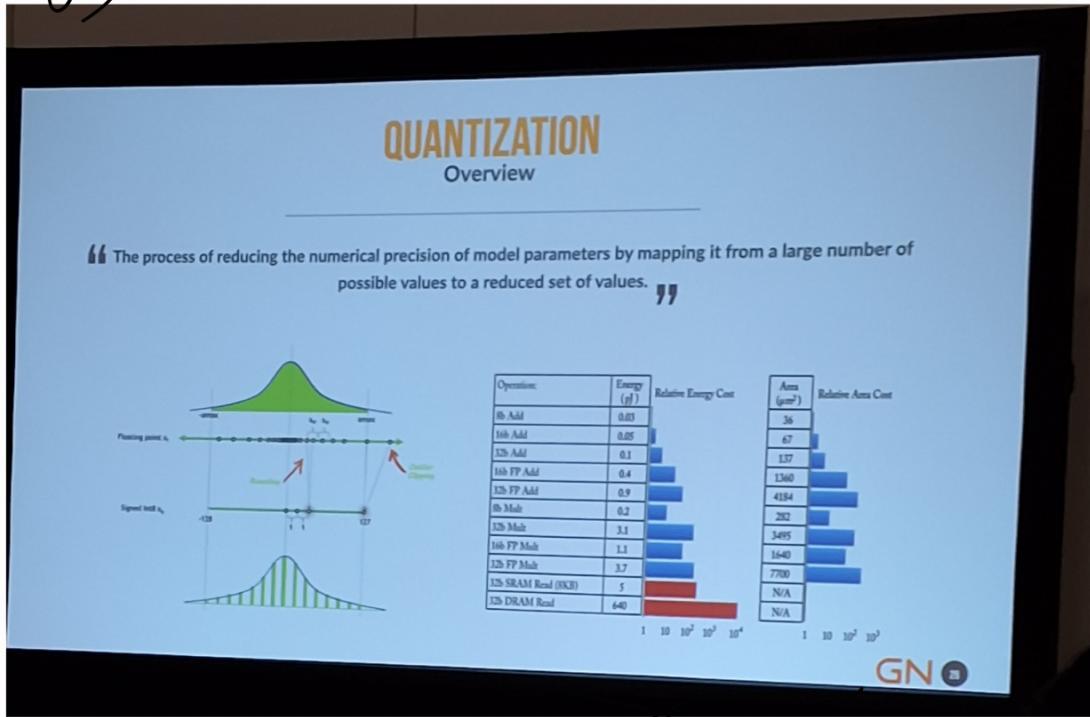
5)



↳ need be careful for Transformers

↳ only do if hardware supports it? like some new hardware already auto only compute non-zeros

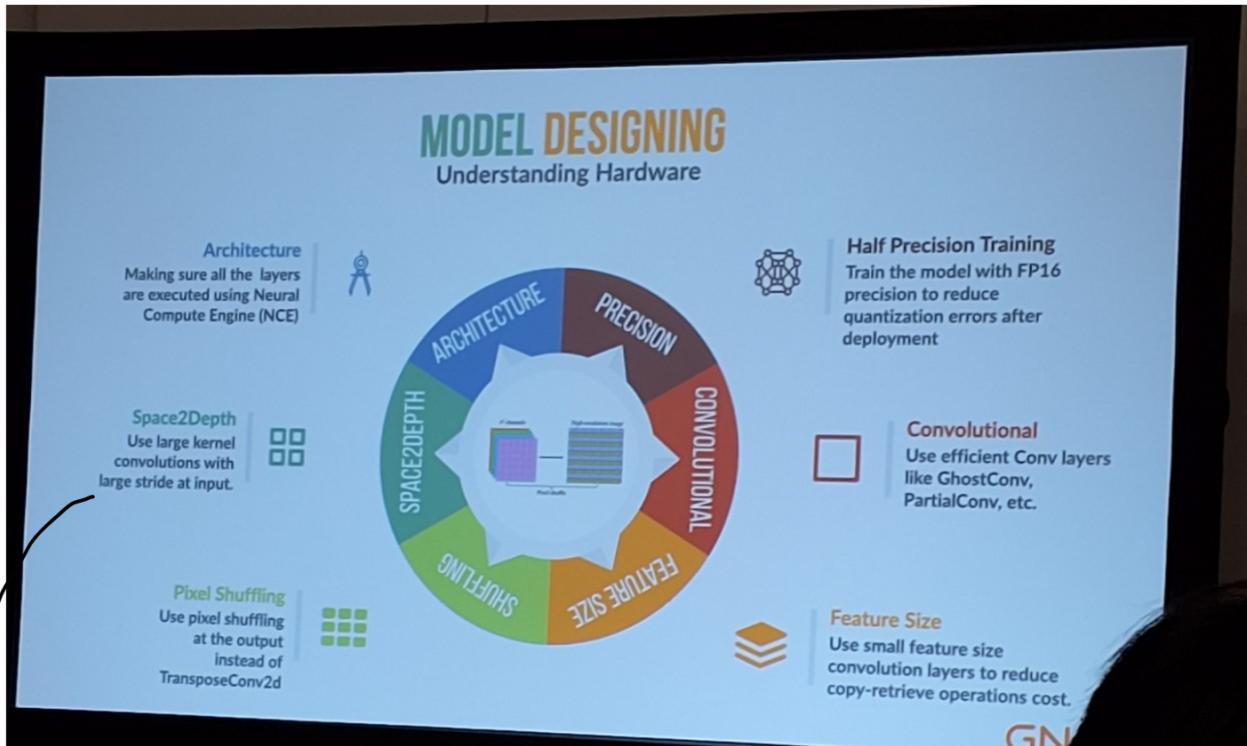
6)



need to see hardware supports what

↳ Post Training or Quantize aware training

↳ if want use, back prop shd still be higher precision  
else gonna  $\rightarrow 0$

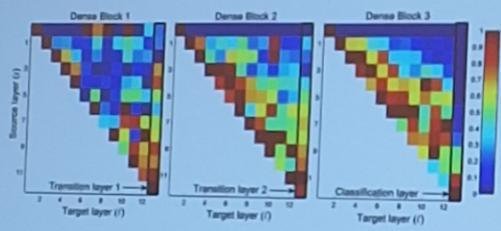


→ Pushing features into depth instead of dealing with larger resolution

# OPTIMIZING DOWN SAMPLE CONVOLUTIONS

## Model Optimization

- 01 Dense Connections, promotes feature reuse across layers, saving on parameters and computations.



- 02 Unique Concatenation, combines features from prior layers, enhances feature richness, avoids duplication, and conserves memory bandwidth.

- 03 Diverse Learning, dense links foster varied feature learning due to added supervision from loss.

- 04 Enhanced Propagation, ensures improved feature spread and minimizes overfitting.

- 05 Efficiency in Bandwidth, reduced parameters and redundancy lead to less memory usage, conserving memory bandwidth.

GN

- Emailed cam vision in the wild & robustness of inference  
↳ don't have ↳ email back (Mohit).  
got WACV recording  
on youtube

