VITA: ViT Acceleration for Efficient 3D Human Mesh Recovery via Hardware-Algorithm Co-Design

Shilin Tian

Advisor: **Dr. Hao Zheng**

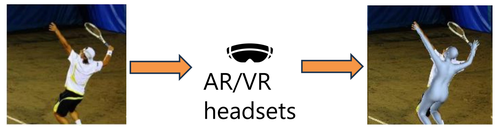
Department of Electrical and Computer Engineering, University of Central Florida

Email: Shilin.Tian@ucf.edu

UCF | Electrical and Computer Engineering

UNIVERSITY OF CENTRAL FLORIDA

Background and Motivation

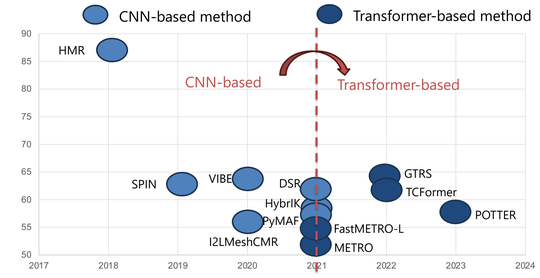


3D Human Mesh Recovery (HMR) is pivotal in enhancing AR/VR experiences by enabling realistic and dynamic human representations in various applications such as:

- **Avatar Creation**: Generating lifelike avatars for immersive experience.
- **Immersive Gaming**: Enhancing player interaction with realistic character animations.
- **Motion Capture**: Capturing and replicating human movements in virtual environments.
- **Virtual Try-On**: Allowing users to try on clothes and accessories virtually.

HMR task pipeline[1] is a complex 2D-to-3D lifting process encompasses:

- Objective Detection and Feature Extraction
- 3D Pose and Shape Recovery
- Post-Processing and Validation



Methods are lifting from CNN-based to ViT-based for:

- Strong ability to model global dependency

Challenges of ViT for HMR:

- Poor data locality
- Irregular memory access pattern
- High computation and memory footprint

Contributions

On the algorithm side:

- We adopt a pooling attention framework to replace conventional multi-head self-attention, significantly reducing the number of parameters and MAC operations.
- We further optimize attention framework to be data locality aware by proposing average pooling structure.

On the hardware side:

- We adopt reconfigurable interconnect that supports different dataflow, significantly reduce DRAM access.
- We propose unified PE architecture handling various ViT operations, efficiently optimizing spatial and temporal data locality.

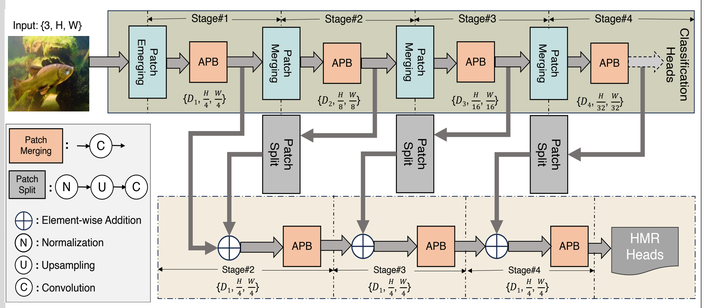
Proposed VITA Algorithm

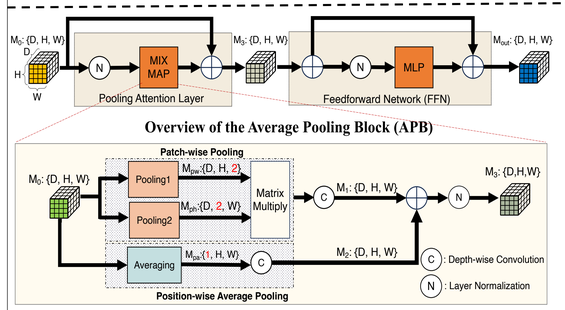
Design Goal:

- Fuse global feature captured at low resolution and local feature extraction at high resolution.
- Optimized for hardware efficiency

Design Features:

- Dual-Stream Structure
- Top steam hierarchically reduces patch size on each stage
- Bottom stream retains same input patch size

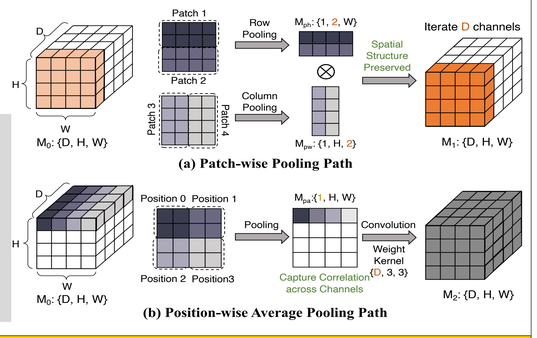




Architecture of mixed average pooling block (MIX MAP)

Average Pooling Block(APB) Feature:

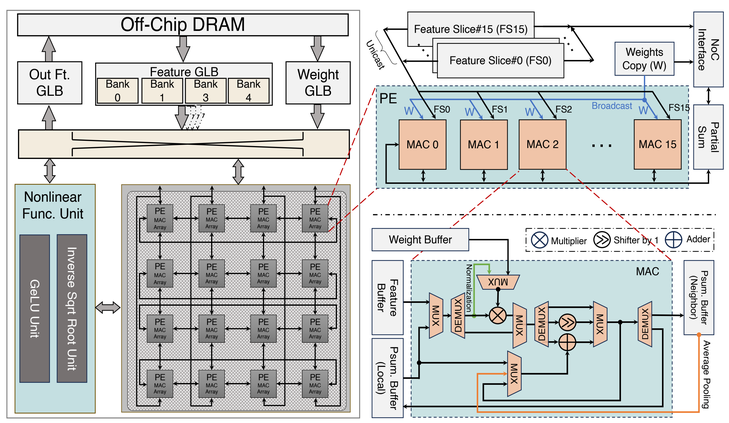
- Proposed APB mitigates irregular memory access patterns via two-path design
- Patch-wise pooling correlates features from patches within one channel
- Position-wise pooling correlates features from patches across different channels



Two Separate Pathways:

- Patch-wise Pooling:
  - Apply pooling within each dimension
  - Encapsulate inter-patch correlations
- Position-wise Pooling:
  - Delve into detailed features within each patch
  - Provide fine-grained control over patches

Proposed VITA Hardware Design



4 operation modes : Patch-wise pooling, position-wise pooling, matrix multiplication and depth-wise convolution.

Support Various Operations:

- Normalization, pooling, convolution etc. in ViT.

Accelerator Features:

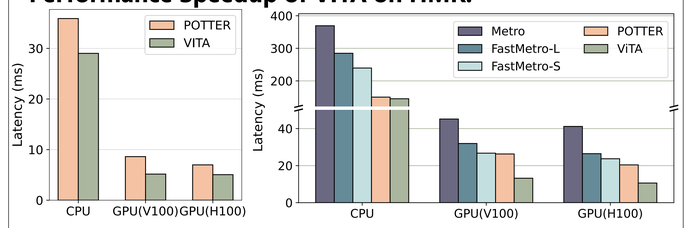
- Universal PE supports 4 operation modes (see below).
- Configurable interconnect supports various dataflows for these operation modes.
- Novel MAC unit supports square operation, division by 2, accumulation and multiplication, improving PE utilization and data locality.

Result and Evaluation

ViT Models			
Name	Params(M)	MACs(G)	Top-1 Acc(%)
ViT-L/16	307	190.7	76.5
RSB-ResNet-18	12	1.8	70.6
RSB-ResNet-34	22	3.7	75.5
POTTER_S12	12	1.8	77.2
VITA	**12**	**1.8**	**78.08**

- Datasets: ImageNet-1K and Human3.6M/3DPW (for HMR)
- Top-1 Accuracy: 78.08%

Performance Speedup of VITA on HMR:



On CPU(Xeon 6240), Nvidia Tesla V100, and Nvidia H100:

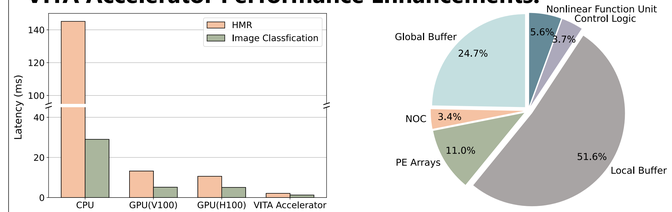
VITA vs. POTTER[2]:

- CPU: 1.23x
- Tesla V100:1.67x
- H100: 1.38x

VITA vs. Other Baselines:

- Metro: 2.95x
- POTTER: 1.65x
- FastMetro-L: 2.65x.
- FastMetro-S: 1.96x

VITA Accelerator Performance Enhancements:



Compared to CPU(Xeon 6240), Nvidia Tesla V100, and H100:

Image Classification Task:

- CPU: 23.05x
- Tesla V100: 4.01x
- H100: 4.10x

HMR Task:

- CPU: 69.12x
- Tesla V100: 6.29x
- H100: 5.05x

Conclusions

- VITA is the first ViT accelerator designed for HMR task that incorporates various hardware constraints into the algorithm design.
- VITA achieve 5.05× and 69.12× speedups on average over the state-of-the-art GPUs and CPUs for HMR task.

Future Work

End-to-End Implementation

[1] Goel, Shubham, et al. "Humans in 4d: Reconstructing and tracking humans with transformers." In *Proc. Of CVPR*, 2023.

[2] Zheng, Ce, et al. "Potter: Pooling attention transformer for efficient human mesh recovery." In *Proc.of CVPR*, 2023.