Lecture 12: New playground for Efficient AI: AR/VR

Notes: Final Presentation

- May 13 from 9am-3pm: 2MTC, 907.
- May 14 from 9am-3pm: RH 202.
- Will send out a signup spreadsheet.
- Presentation time:
  - <30 mins (25mins + 5mins QA)

Notes: Final Report

- Due on May 14 Midnight
- Four-six pages (Will send out the template)
  - Introduction
  - Problem Description
  - Related work
  - Method
  - Experiment results
  - Conclusion

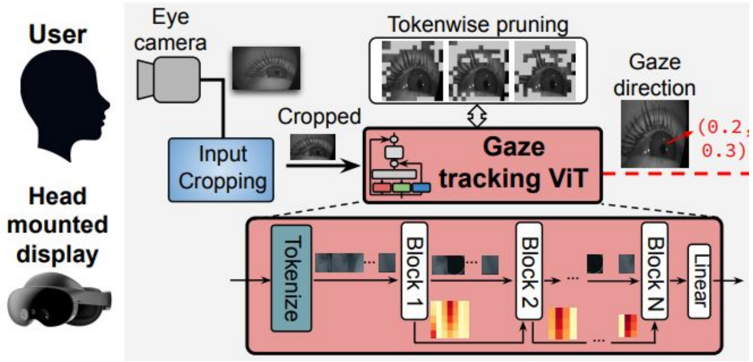
Agenda

- FovealNet: Advancing AI-Driven Gaze Tracking Solutions for Efficient Foveated Rendering in Virtual Reality
- FovealSeg: Efficient Gaze-driven Instance Segmentation for Augmented Reality

Liu, Wenxuan, et al. "Fovealnet: Advancing ai-driven gaze tracking solutions for efficient foveated rendering in virtual reality." *IEEE Transactions on Visualization and Computer Graphics* (2025).
Zeng, Hongyi, et al. "Foveated Instance Segmentation." in Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

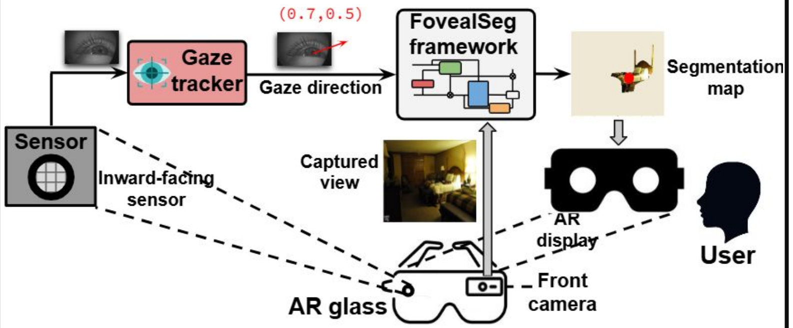
Agenda

FovealNet



AI for ARVR

FovealSeg



ARVR for AI

NYU SAI LAB

Liu, Wenxuan, et al. "Fovealnet: Advancing ai-driven gaze tracking solutions for efficient foveated rendering in virtual reality." *IEEE Transactions on Visualization and Computer Graphics* (2025).

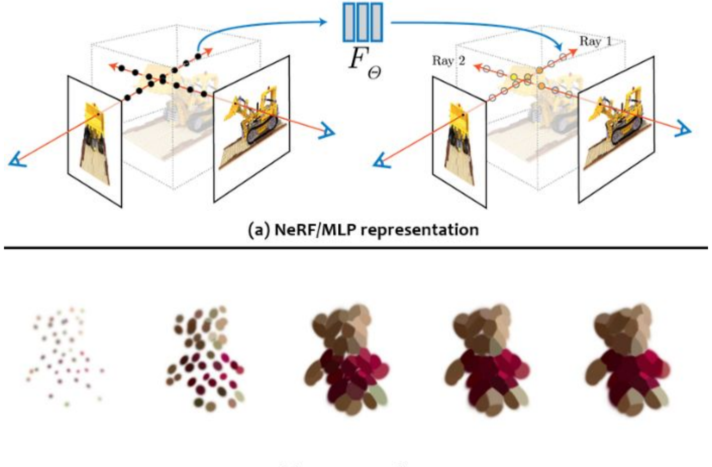
Image Rendering in Virtual Reality

Quest Pro

- Image rendering is one of the most important CV applications in AR/VR.
- Achieving real-time rendering that feels seamless and interactive requires sophisticated algorithms and powerful hardware.
- However, VR Platforms are usually have limited computational capability.

NYU SAI LAB

Image Rendering



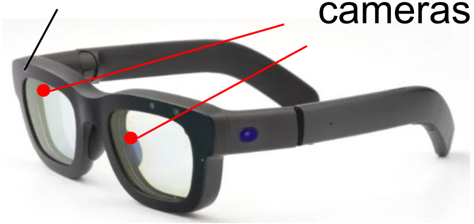
(a) NeRF/MLP representation

(b) Gaussian Splatting

- **Image rendering** is the process of generating a final visual image from a set of data, typically using computer algorithms.
- It is a key step in computer graphics, where scenes (made up of geometry, lighting, textures, and camera perspective) are converted into 2D images.

NYU SAI LAB

AR/VR Device

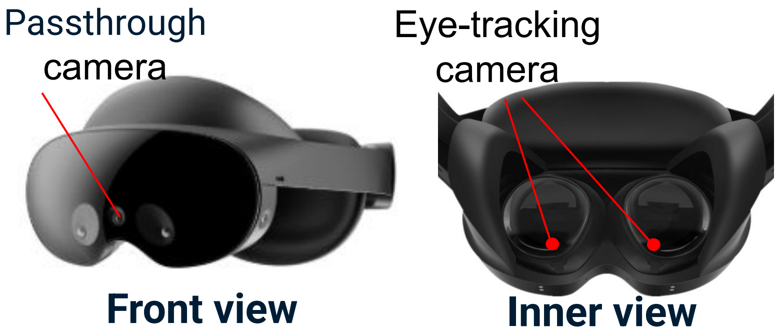


Meta Orion AR Glass

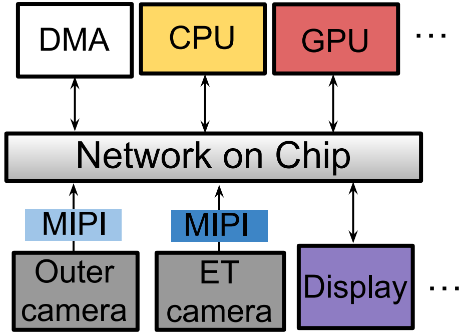
Front view **Inner view**

Meta Quest Pro

Hardware Architecture of AR/VR Device

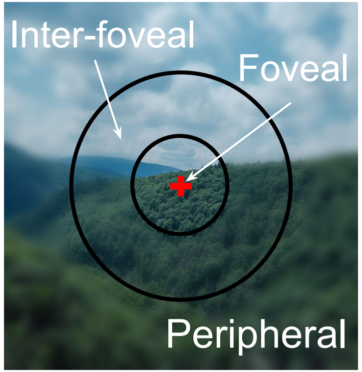


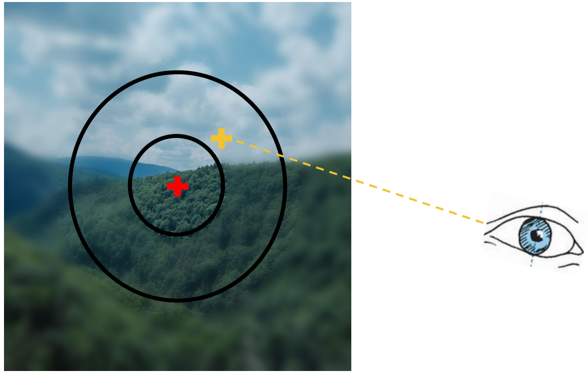
Meta Quest Pro



NYU SAI LAB

Foveated Rendering

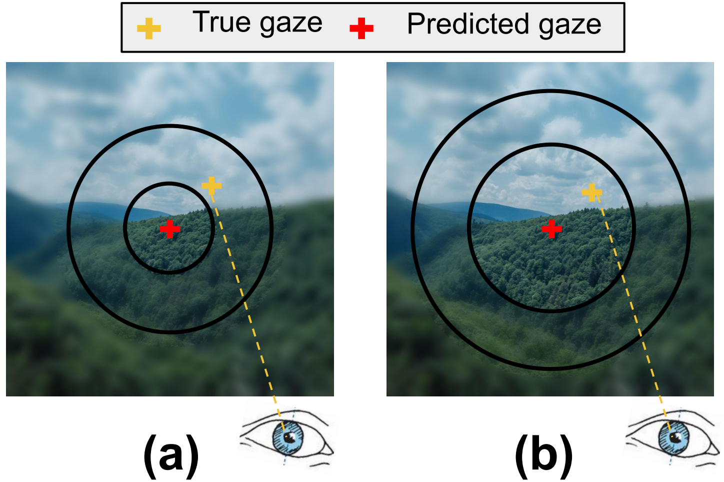


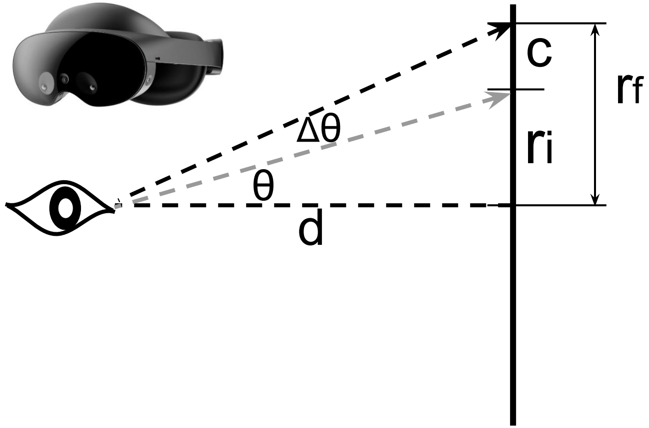


- Image rendering plays a pivotal role in the performance and user experience of VR systems.
- Foveated rendering emerges as an ideal solution, drastically reducing rendering latency without any noticeable degradation in visual quality.
- However, an accurate gaze tracking mechanism is required to make foveated rendering works well without impacting use experience.

NYU SAI LAB

Foveated Rendering



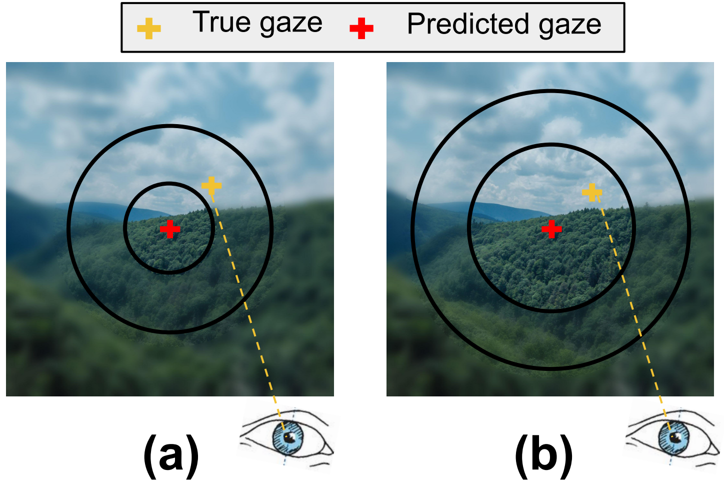


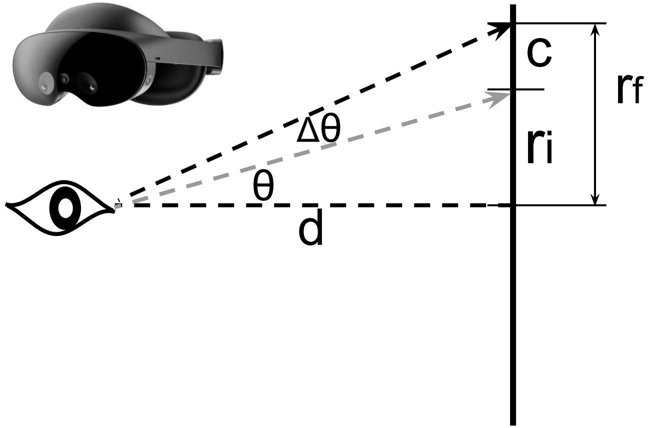
- Visual quality degradation due to tracking error, and then the foveal region is enlarged for better visual quality.

$$r_f = r_i + c = d \cdot \tan(\theta_i + \Delta\theta) = d\tan(\theta_f)$$

NYU SAI LAB

Foveated Rendering

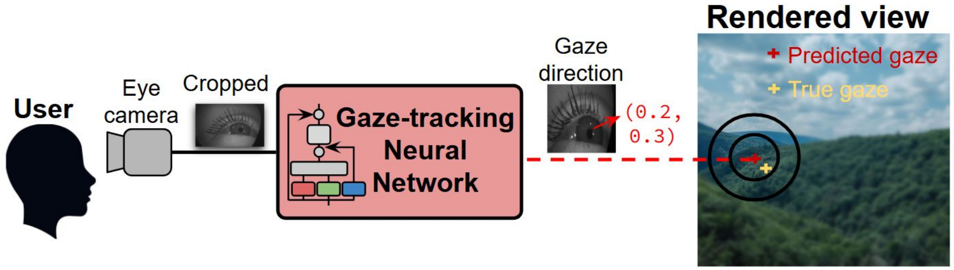


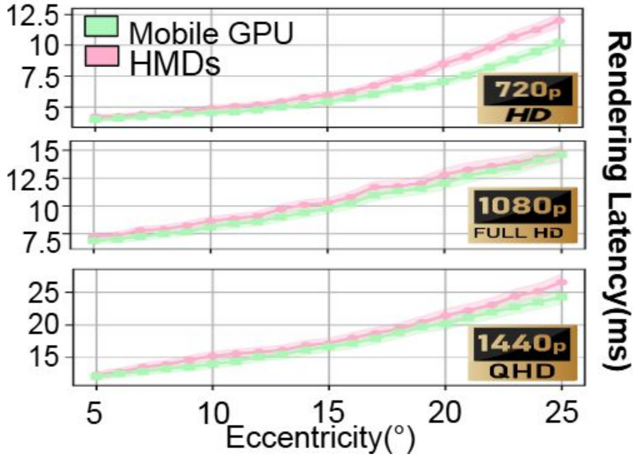


- C represents the changes due to the gaze tracking error.
- The smaller the tracking error is, the smaller the size of the foveal region is.
- A smaller foveal region will have a better system performance.

NYU SAI LAB

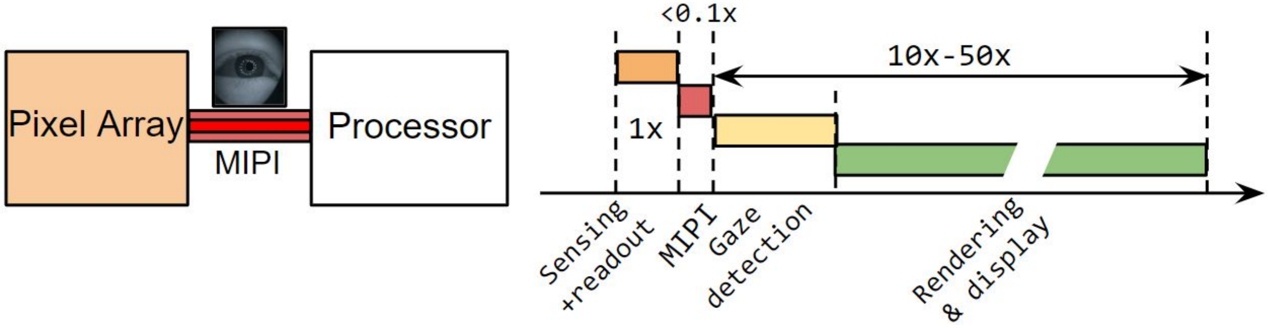
Efficient AI for Gaze-tracked Foveated Rendering





- In gaze-tracked foveated rendering (TFR), an accurate gaze-tracking solution needs to be developed with high tracking accuracy.
- The gaze tracking is usually performed using deep neural networks.

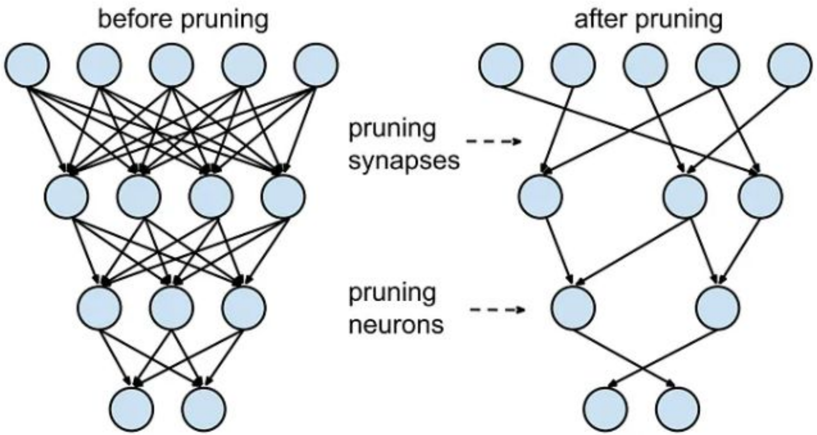
Efficient AI for Gaze-tracked Foveated Rendering

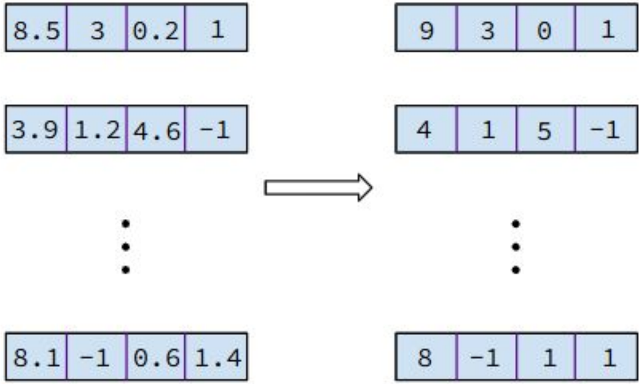


- Gaze detection with rendering and display will take majority of the processing time.
- It is critical to design an gaze tracking solution to minimize the rendering latency as well as the processing latency for gaze tracking neural networks.
- To reduce rendering latency, the gaze-tracking DNN needs to achieve high accuracy.
- To minimize the latency in gaze tracking, we will implement efficient DNN algorithms.

NYU SAI LAB

Neural Network is Highly Redundant

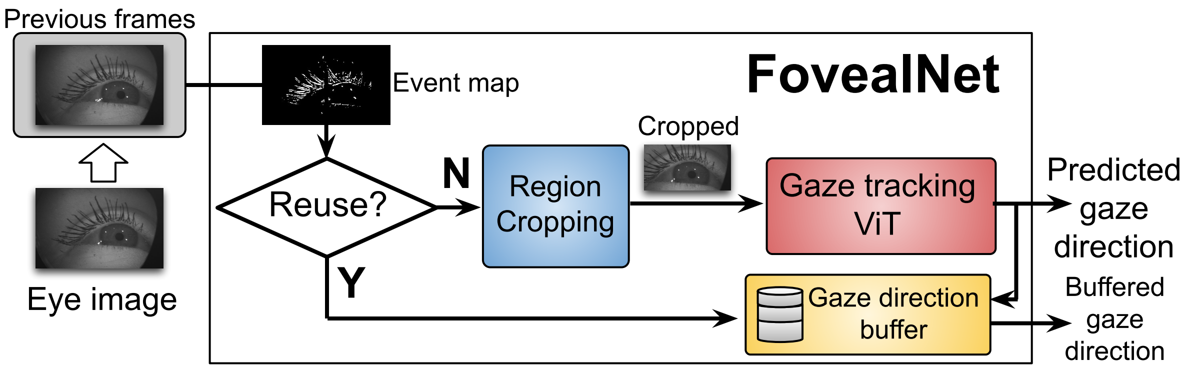




- Neural networks are highly redundant, meaning they often contain a large number of parameters and computations that contribute minimally to the final output.
- Pruning and quantization are two major approaches for neural network acceleration.

NYU SAI LAB

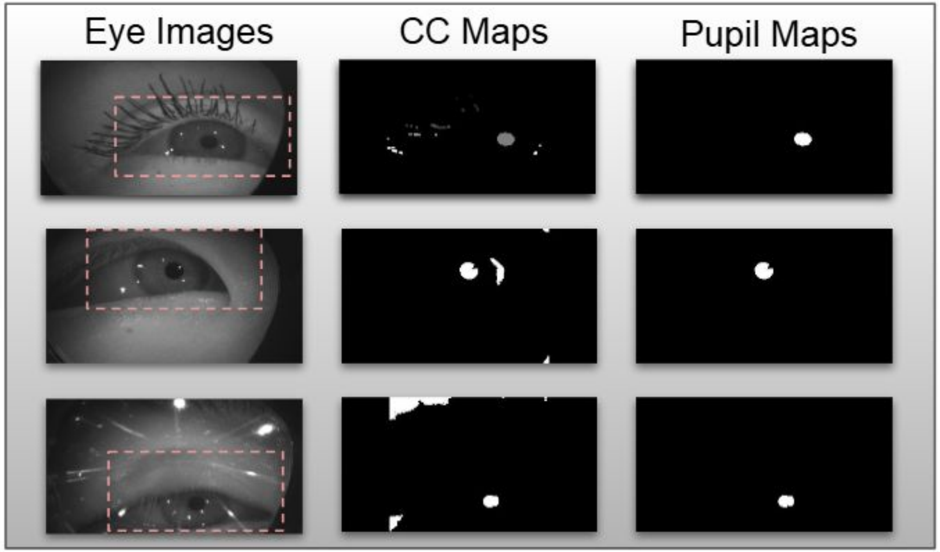
FovealNet: Overview



- We design FovealNet, an efficient gaze tracking solution for consecutive frames.

NYU SAI LAB

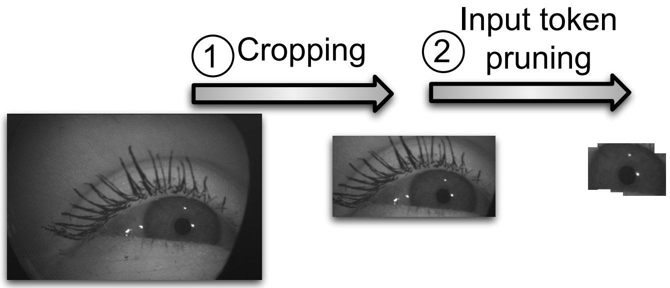
FovealNet: Input Cropping Algorithm



- Given the input eye image captured by the eye camera, we first apply an analytical solution to predict the pupil location.
- Given the gaze direction, the eye image can then be cropped using a bounding box of predefined size.

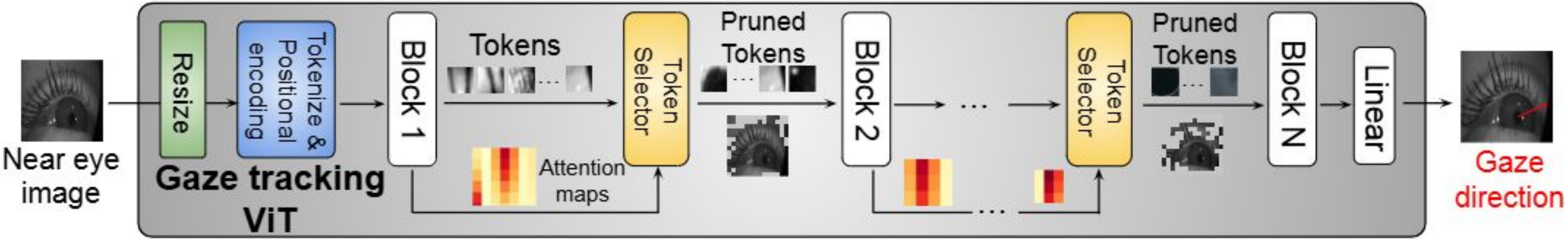
NYU SAI LAB

FovealNet: Gaze tracking Neural Network



- A key advantage of ViT over CNN is its ability to fine-grain prune input tokens, enabling the removal of image tokens with unimportant content.
- The attention score reflects the importance of each token in relation to the gaze prediction result.
- Using these scores, we employ a top-k selector to remove unimportant tokens, which further reduces the computational cost of subsequent ViT blocks.

FovealNet: Gaze tracking Neural Network

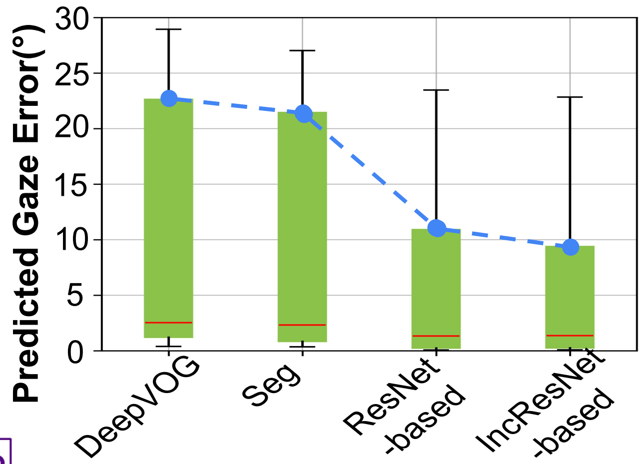


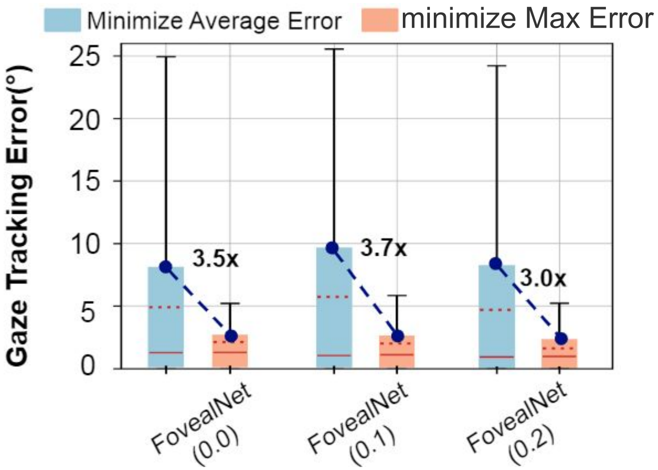
- The cropped eye images containing informative content are first resized to a smaller square (224×224) and then processed by the gaze tracking DNN to predict gaze direction.
- The ViT contains 8 transformer block, each block consists of 6 heads with an embedding dimension of 128.

NYU SAI LAB

FovealNet: Loss Function Design

$$\min \sum_{d \in D_{train}} (||\theta_d - \theta_d^g||^2) \Rightarrow \min \max_{d \in D_{train}} (||\theta_d - \theta_d^g||^2)$$





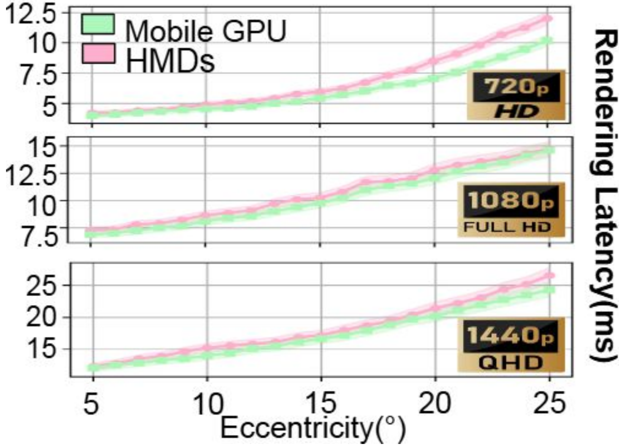
NYU SAI LAB

FovealNet: Loss Function Design

$$\min \sum_{d\in D_{train}} (||\theta_d - \theta_d^g||^2) \Rightarrow \min \max_{d\in D_{train}} (||\theta_d - \theta_d^g||^2)$$

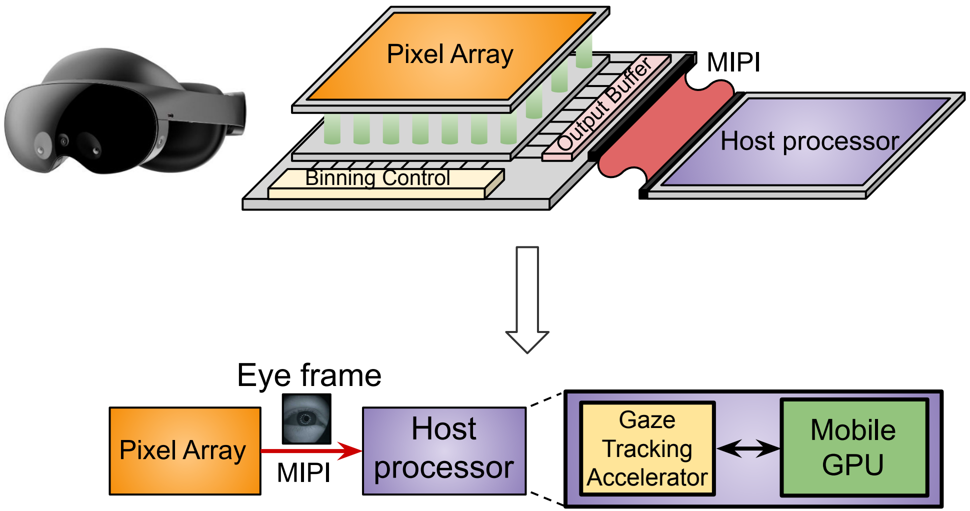
$$\Downarrow$$

$$\sum_{b\in B} U\left(\frac{1}{N}\ln\left(\sum_{d\in D^b_{train}} e^{N||\theta_d - \theta_d^g||^2}\right)\right) \Leftarrow \sum_{b\in B} \frac{1}{N}\ln\left(\sum_{d\in D^b_{train}} e^{N||\theta_d - \theta_d^g||^2}\right)$$



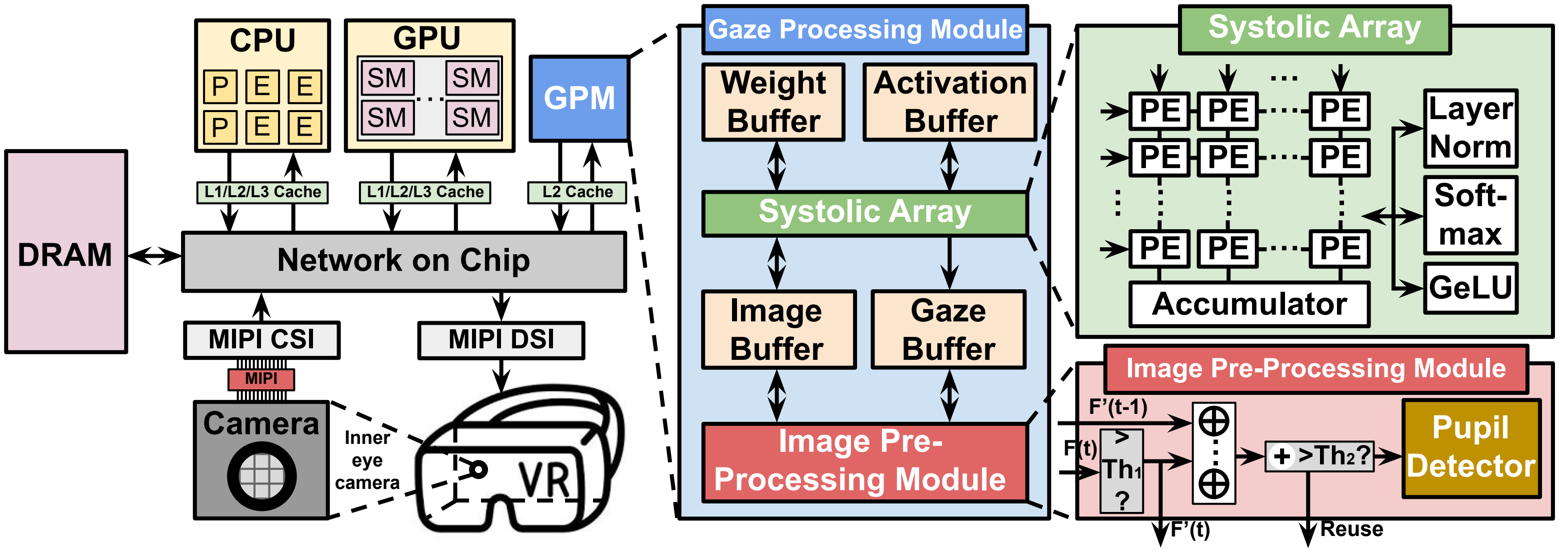
- To fully utilize the training dataset, we find it more effective to optimize an approximate version by replacing the max operation with an alternative approach.
- Finally, we can directly relate the gaze error to the TFR latency, enabling us to optimize the rendering latency directly.

Gaze-tracking Foveated Rendering System Design



- We propose a plug-in module to the host processor of modern VR device.
- The plug-in module will accelerate the execution of gaze analyzer.
- The mobile GPU will take the output from the gaze analyzer and adaptively changes the rendering resolution.
- We simulate its PPA using EDA tools. Together with some user study to ensure the visual experience.

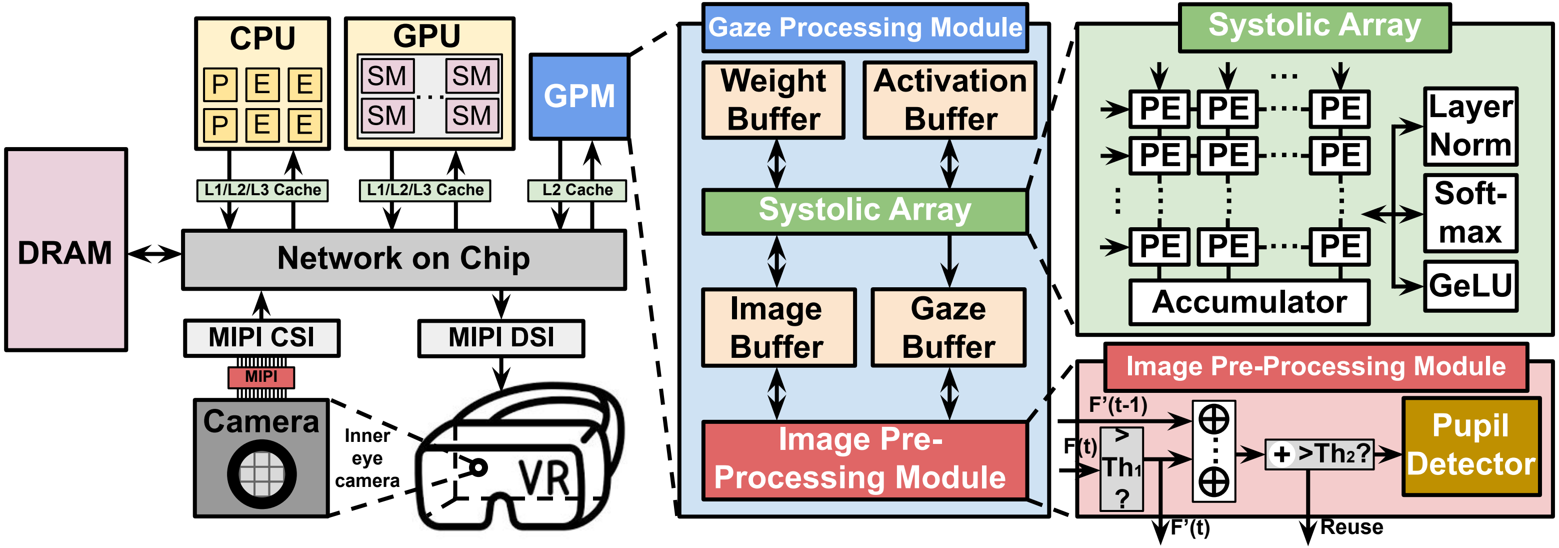
Hardware Accelerator Design



- We design a gaze processing module that is integrated with the modern VR device.

NYU SAI LAB

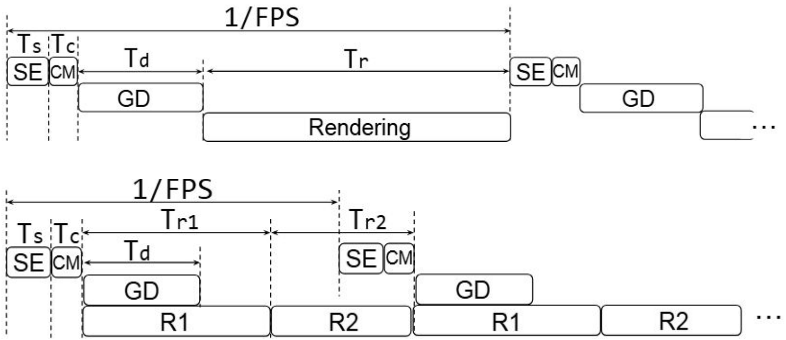
Hardware Accelerator Design

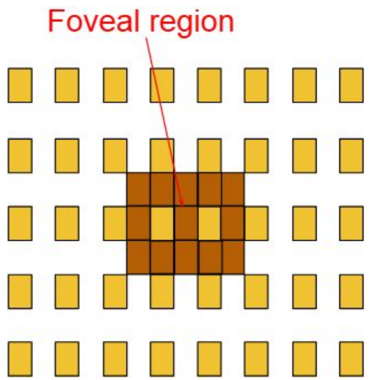


- The Input frame will first be sent to the image Pre-Processing Module which returns the cropped image.
- The resultant image will then send to the systolic array for gaze prediction.

NYU SAI LAB

Hardware Accelerator Design

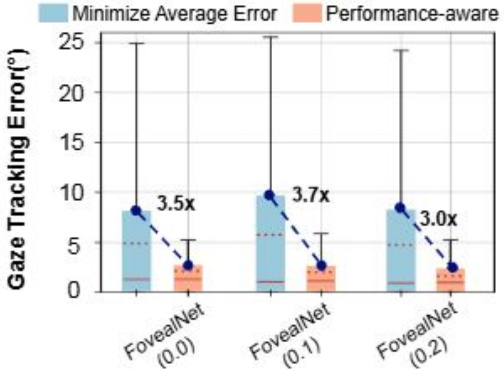




- The accelerator is integrated with other SoC components via the Network-on-Chip (NoC), enabling efficient communication with the CPU, GPU, DMA, and additional components
- The gaze tracking and background rendering process can be overlapped to save the processing latency.

Tracking Performance Evaluation

Network	Mean	P90	P95	Min	Max	FLOPS (billions)
NVGaze [31]	6.81	13.07	18.62	0.94	42.30	0.021
DeepVoG [10]	3.47	17.76	23.77	0.55	29.06	36.5
Seg [11]	3.25	18.29	22.80	0.52	28.42	2.6
ResNet-based [29]	1.52	5.96	13.15	0.07	26.46	3.6
IncResNet-based [28]	1.72	6.23	12.4	0.12	25.47	13.12
FovealNet (0.2)	1.27	4.92	**8.09**	**0**	24.92	2.08
FovealNet (0.1)	1.05	5.75	9.63	**0**	25.54	2.42
FovealNet (0.0)	**0.93**	**4.71**	8.21	**0**	**24.2**	2.80

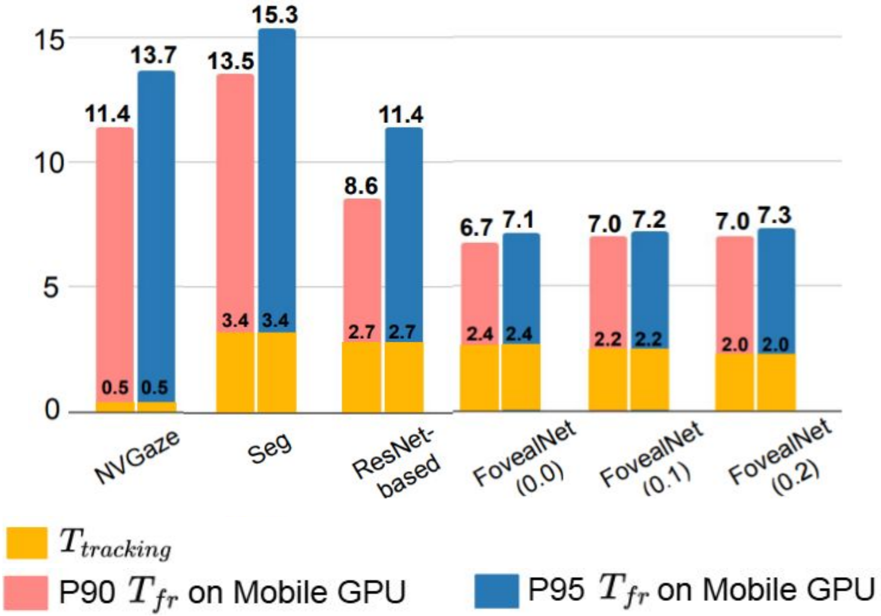


- We change the tokenwise pruning ratio of FovealNet over three ratios: 0.0, 0.1, 0.2.
- We evaluate the performance in terms of mean, P90, and P95 tracking error.
- FovealNet achieves the lowest gaze tracking error compared with other baselines, while maintaining the lowest FLOPs.

NYU SAI LAB

Evaluation with Performance-aware Training Loss

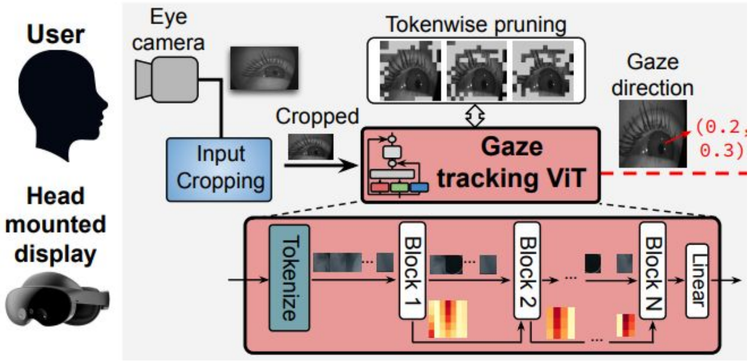
- We profile the processing latency $T_{tracking}$ of FovealNet on a Quadro RTX 3000 Mobile GPU.
- FovealNet (0.0) achieves the lowest gaze tracking latency of 6.7ms and 7.1ms when setting $\Delta\theta$ to P95 or P90 of the gaze error distribution.



NYU SAI LAB

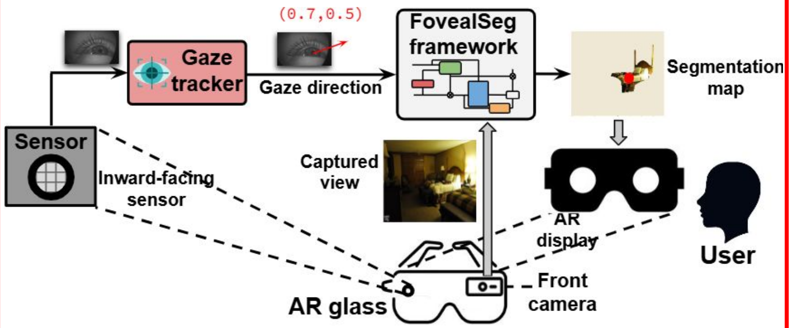
Agenda

FovealNet



AI for ARVR

FovealSeg

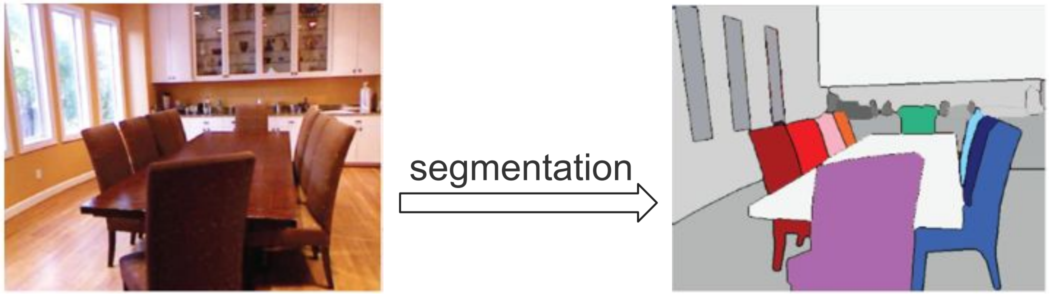


ARVR for AI

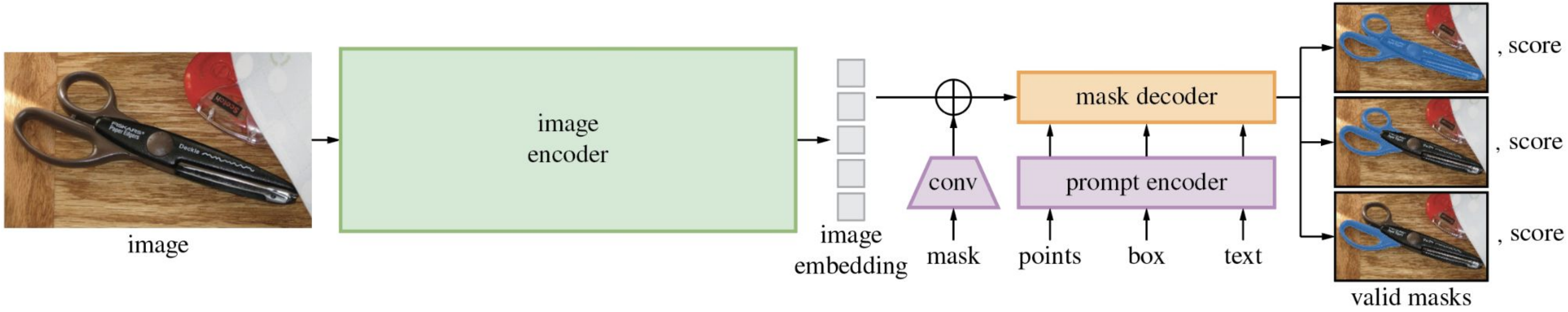
Zeng, Hongyi, et al. "Foveated Instance Segmentation." in Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

Why Segmentation is Necessary for AR?

- Enables the user to identify and isolate objects, allowing accurate overlay of virtual content.
- Helps AR systems understand spatial relationships for correct depth perception and perspective adjustments.
- Can be used as VLM input.



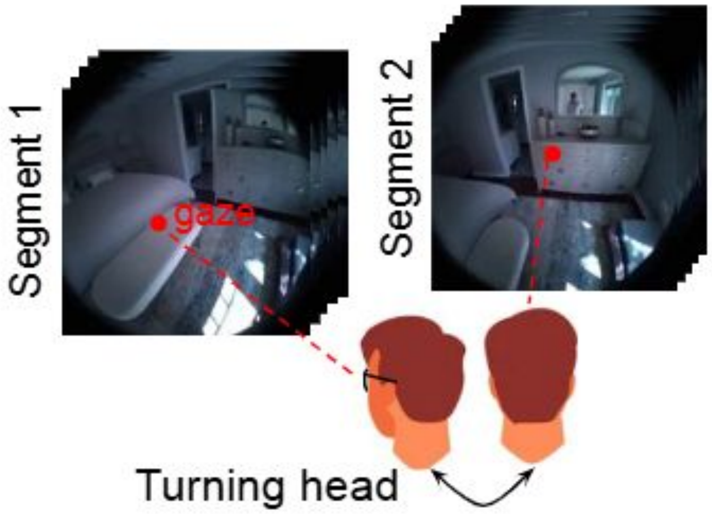
Segmentation is Expensive



Models	LVIS (640x640): GFLOPs
ViT-base	2.774
Efficient SAM	37.1
SAM	831

NYU SAI LAB

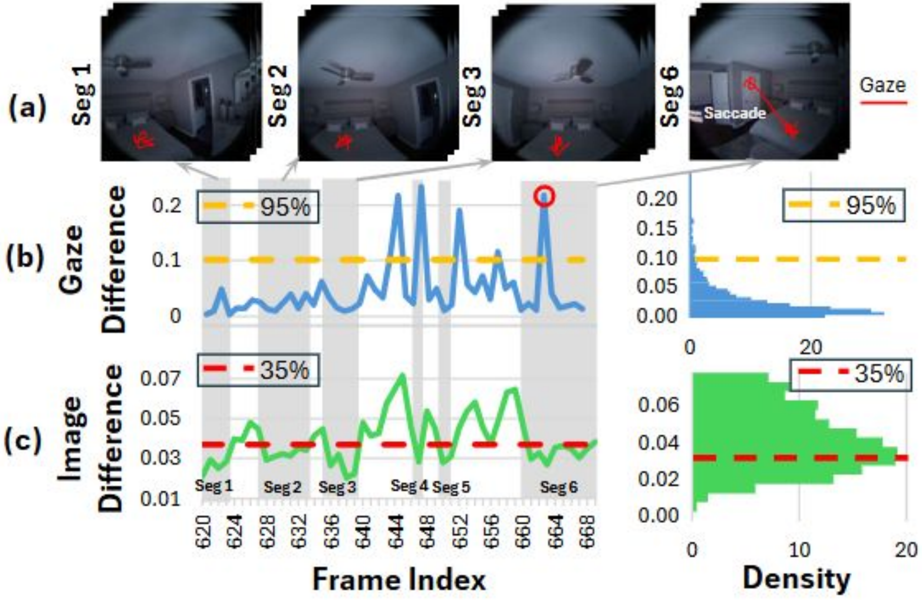
Tracked Foveated Instance Segmentation



- AR users typical have such behavior:
    - Focus on a single scene for a period of time.
    - Within each scene, observe only a small number of objects.
- This enables significant room for enhance computational efficiency for the instance segmentation tasks.

NYU SAI LAB

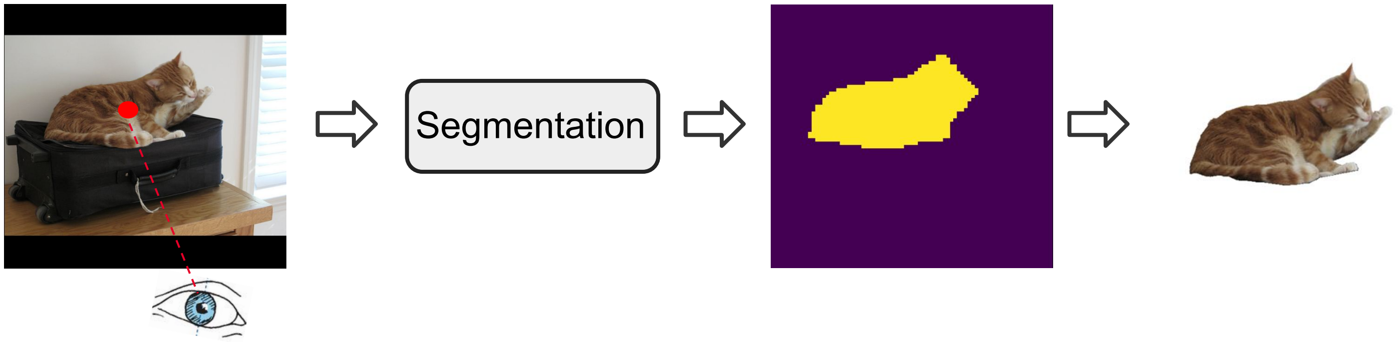
Tracked Foveated Instance Segmentation



- AR users typical have such behavior:
  - Focus on a single scene for a period of time.
  - Within each scene, observe only a small number of objects.
- This enables significantly room for enhance computational efficiency for the instance segmentation tasks.

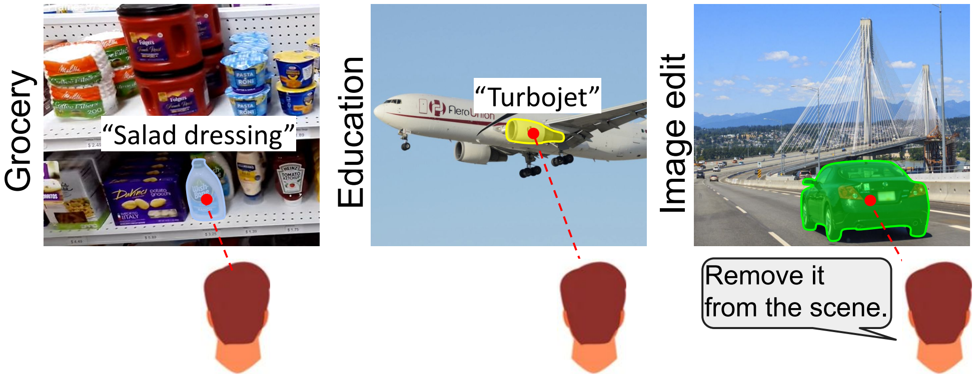
NYU SAI LAB

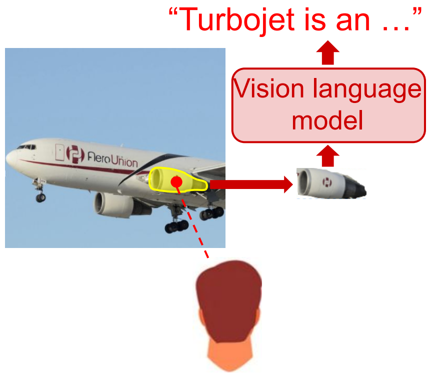
Instance Segmentation in AR



- While processing the entire image and then extracting the mask is possible, this approach would incur a significant computational cost.
- In AR, the user typically only needs to compute the segmentation masks for the instance of interest (IOI).

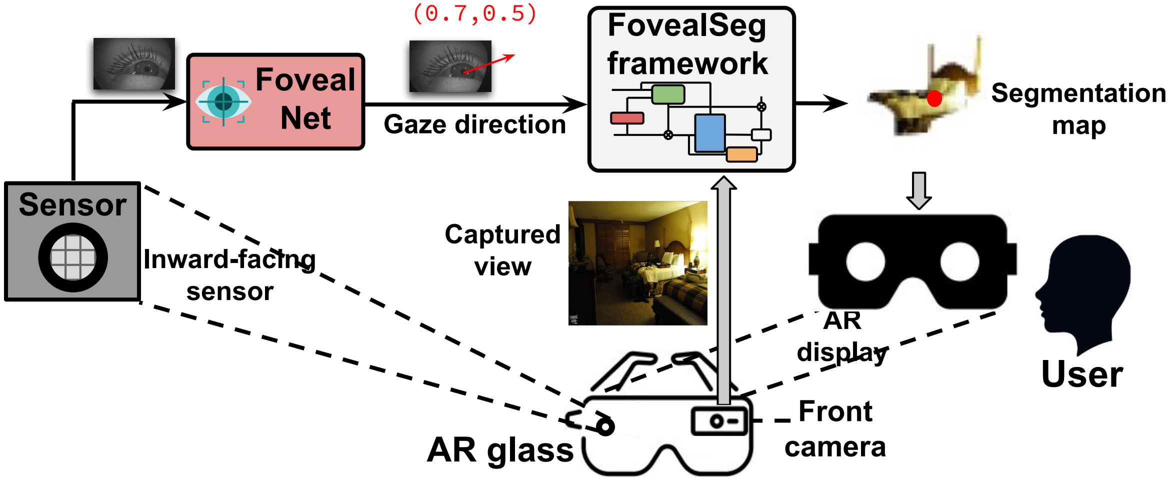
Instance Segmentation in AR





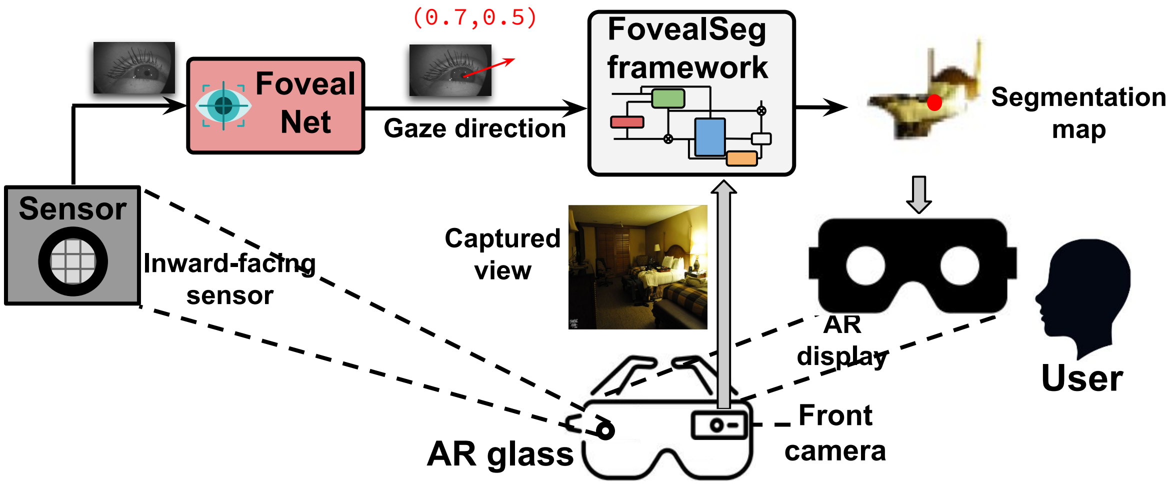
- Segmentation is the fundamental building block for a lot of AR applications.

Foveated Instance Segmentation



- The inward-facing sensor in the AR glasses first captures the eye image, which is then processed using FovealNet.
- The predicted gaze direction will then be sent to the FovealSeg framework to generate segmentation maps on the instance of interest (IOI).

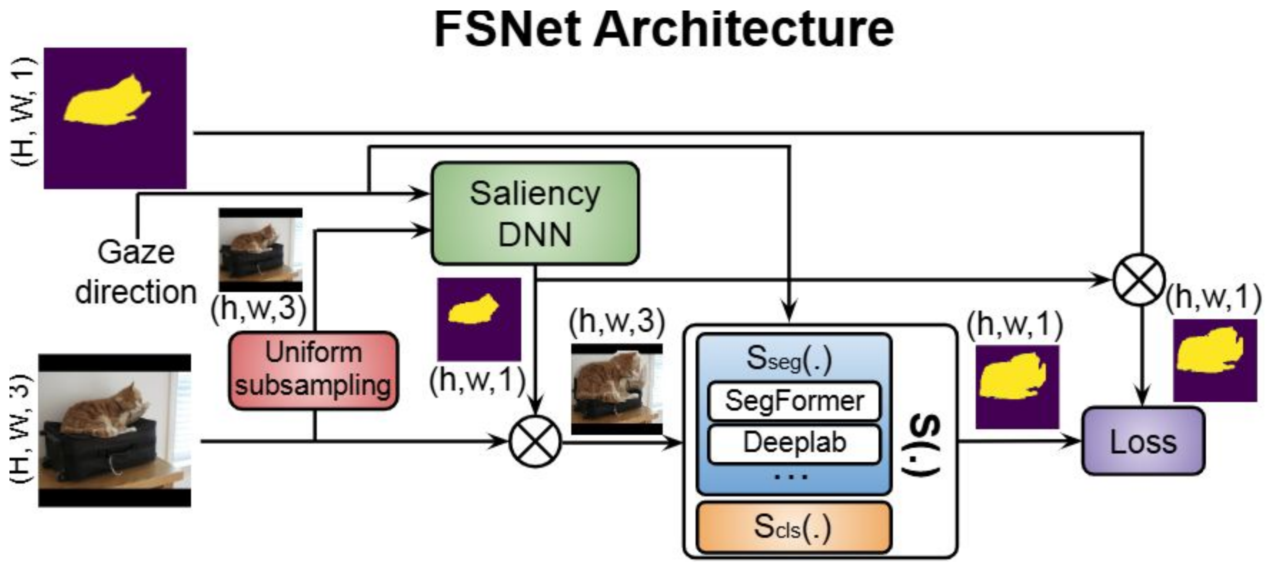
Foveated Instance Segmentation



- FovealSeg applies a learnable pooling layer to selectively remove the redundant information and only process the IOI with high resolution.

NYU SAI LAB

FSNet



- The saliency DNN is trained to generate the saliency score, which guides the subsampling process of the full-resolution input frame.
- The segmentation DNNs are fine-tuned to handle instance segmentation tasks.

NYU SAI LAB

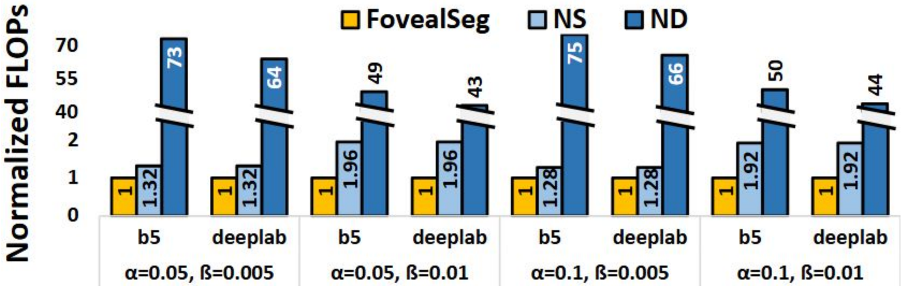
FovealSeg

- The FSNet is executed when:
  - No saccade is detected **and**
  - Input image has changed **or**
  - User gaze direction has moved

Initiation
  $F^{init} = \varnothing, g_{last} = \varnothing, M_{last} = \varnothing$
  for $1 \leq t \leq T$ **do**
    if $|g_t - g_{last}|^2 > \alpha$ **then**
      $g_{last} \leftarrow g_t$;
      Saccade detect, halt rest operations.
    else
      if $\sum_{ij} |F_{ij}^t - F_{ij}^{init}| > \beta$ **then**
        Run FSNet with F^t and g_t, get M^t;
        $F^{init} \leftarrow F^t, g_{last} \leftarrow g_t, M_{last} \leftarrow M_t$;
        return M_t
      else
        if g_t *is within IOI regions of* M_{last} **then**
          return M_{last}
        else
          Run FSNet with F^t and g_t, get M^t;
          $g_{last} \leftarrow g_t, M_{last} \leftarrow M_t$;
          return M_t

Evaluation Results

Method	Parameters(M) ↓	CityScapes (64 × 128)	
		IoU↑	IoU'↑
Avg+DeepLab	42.01	0.26	0.27
Avg+PSPNet	24.3	0.27	0.28
Avg+HRNet	67.12	0.20	0.21
Avg+SegFormer-B4	64.1	0.25	0.27
Avg+SegFormer-B5	84.6	0.27	0.29
LTD [18]	76.22	0.37	0.38
FSNet+DeepLab	42.26	**0.52**	**0.53**
FSNet+PSPNet	24.55	0.49	0.50
FSNet+HRNet	67.38	0.47	0.49
FSNet+SegFormer-B4	64.26	0.46	0.48
FSNet+SegFormer-B5	84.87	0.51	0.52

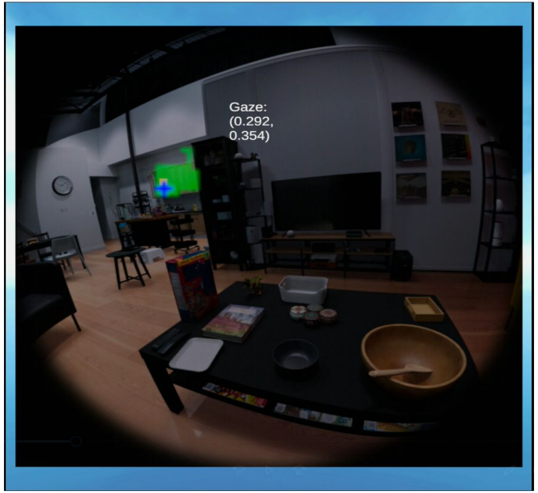


- FovealSeg (FSNet) achieves superior performance with much reduced computational cost.

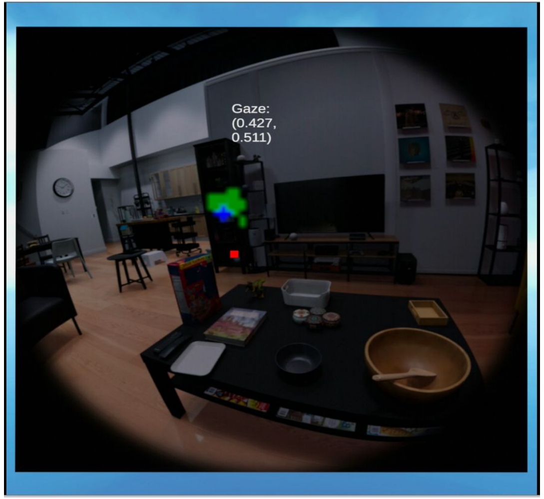
NYU SAI LAB

Implementation

FovealSeg



Conventional



User Study

- Green mask: segmentation mask
- Blue marker: gaze position of current segmentation mask
- Red square: real-time gaze position

NYU SAI LAB

Presentation

- Fusion-3D: Integrated Acceleration for Instant 3D Reconstruction and Real-Time Rendering (Franklyn and Josh)

- Exploiting Human Color Discrimination for Memory-and Energy-Efficient Image Encoding in Virtual Reality (Sancho and Archie)