Lecture 13: New playground for Efficient AI: AR/VR

Notes: Final Presentation

- Final Presentation
  - 12/16/2025 whole day
  - 12/17/2025 whole day
  - Will be fully online (Only the team presenting needs to join during its assigned time slot.)
  - Signup spreadsheet can be access here.
  - Presentation time:
    - 25mins + 5mins QA, presentation must be less than <30 mins, a timer will be used.
    - The duration may be shorter (~20 mins) for projects involving a single student.
    - The presentation will include the following parts: Introduction, background, methodology, evaluation, conclusion.

Notes: Final Report

- Due on **Dec 18 11:59pm**
- NeurIPS format: https://www.overleaf.com/latex/templates/neurips-2024/tpsbbrdqcmsh
- Four-seven pages
    - Introduction
    - Individual contribution (if more than one student)
    - Problem Description
    - Related work
    - Method
    - Experiment results
    - Conclusion

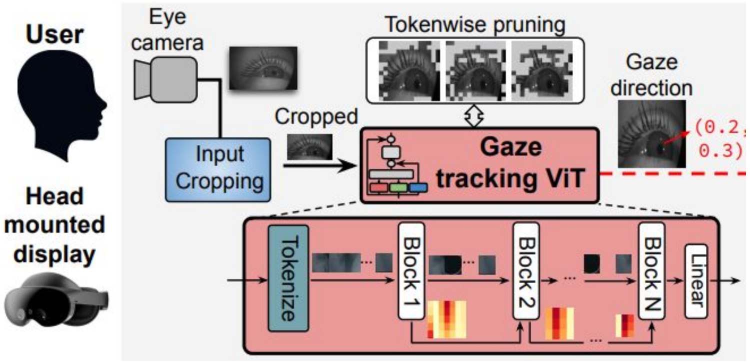
Notes: Course Evaluation

https://www.nyu.edu/students/student-information-and-resources/registration-records-and-graduation/final-exams-and-course-evaluations/course-evaluation.html?challenge=d06e90d7-4d8f-4b88-9d8c-10b73beb60f1

Friday, December 12, 2025 11:59 PM

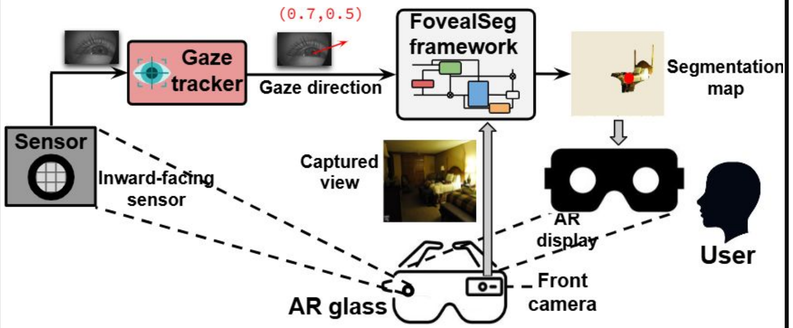
Topics

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).

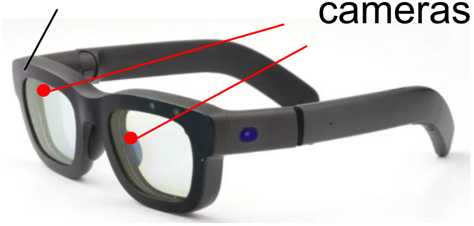
Image Rendering in Virtual Reality

Quest Pro

- Augmented and virtual reality (AR/VR) blend digital content with the physical world or create fully immersive virtual environments, enabling new forms of interaction, visualization, and computing.
- 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

AR/VR Device



Meta Orion AR Glass

Meta Quest Pro

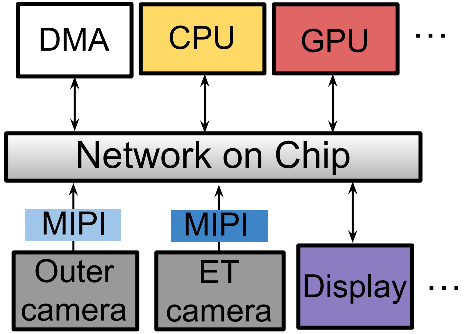
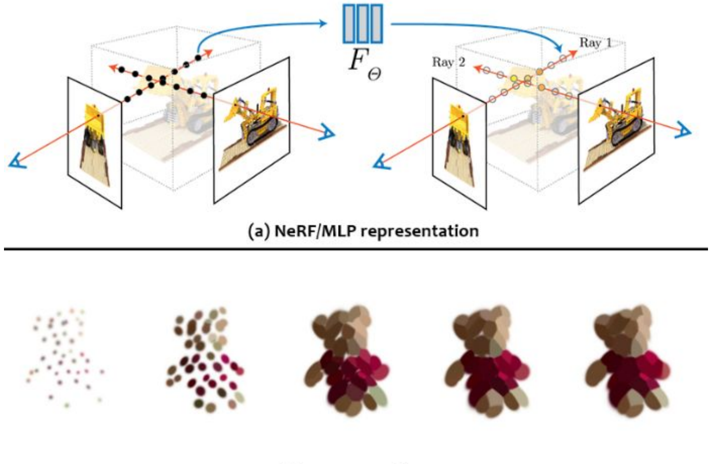


Image Rendering

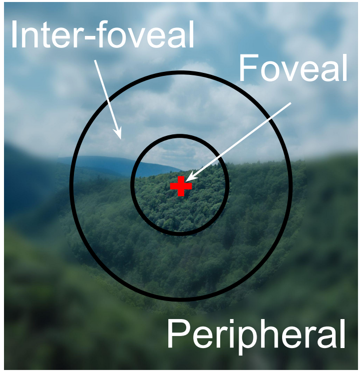


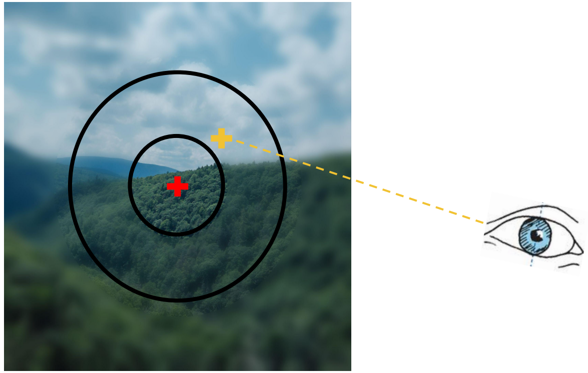
(a) NeRF/MLP representation

(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.

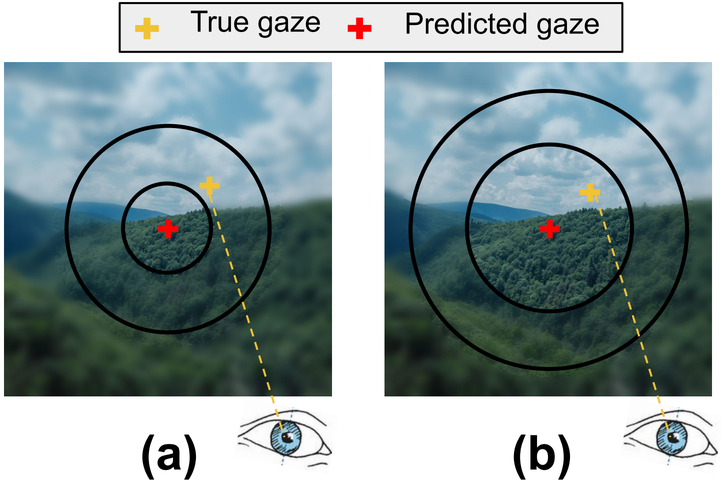
Foveated Rendering

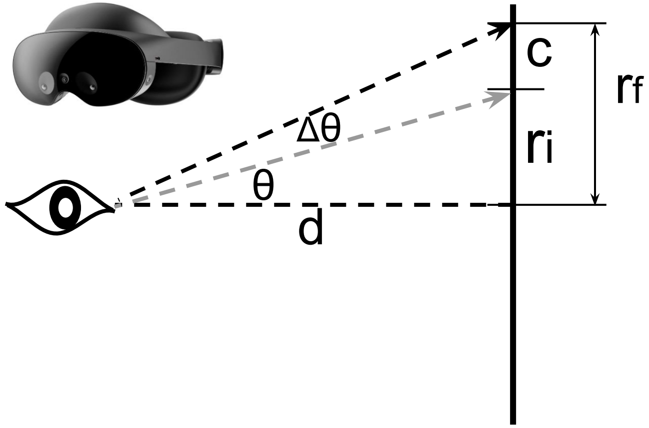




- Image rendering plays a pivotal role in the performance and user experience of VR systems.
- Foveated rendering emerges as an ideal solution, drastically reducing rendering latency without any noticeable degradation in visual quality.
- However, an accurate gaze tracking mechanism is required to make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

Foveated Rendering

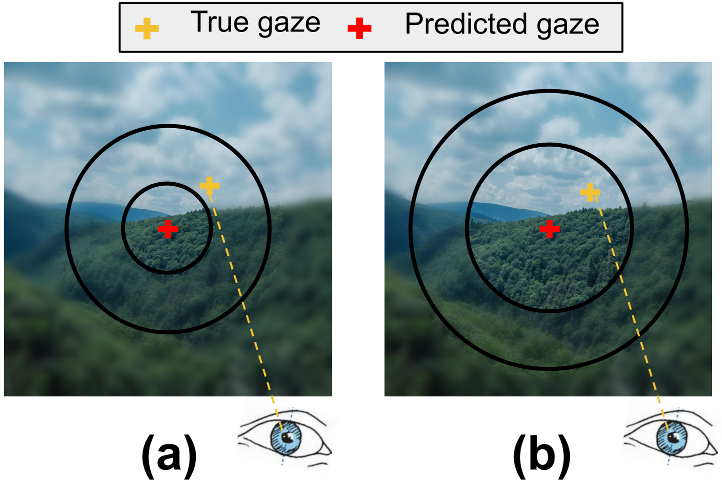


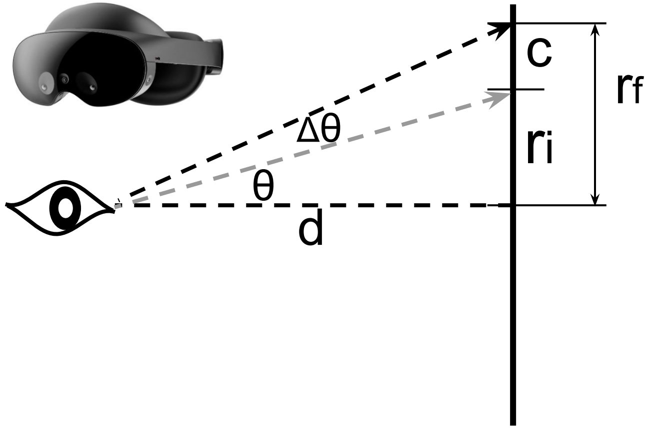


- Visual quality degradation due to tracking error, and then the foveal region is enlarged for better visual quality.

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

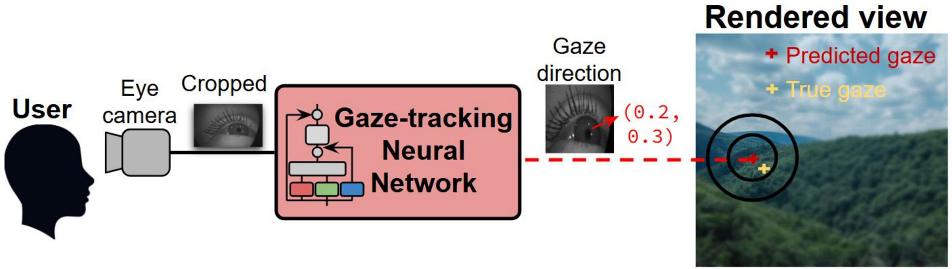
Foveated Rendering

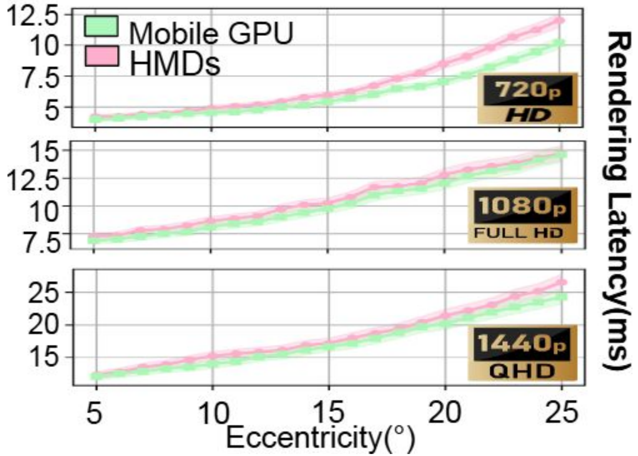




- C represents the changes due to the gaze tracking error.
- The smaller the tracking error is, the smaller the size of the foveal region is.
- A smaller foveal region will have a better system performance.

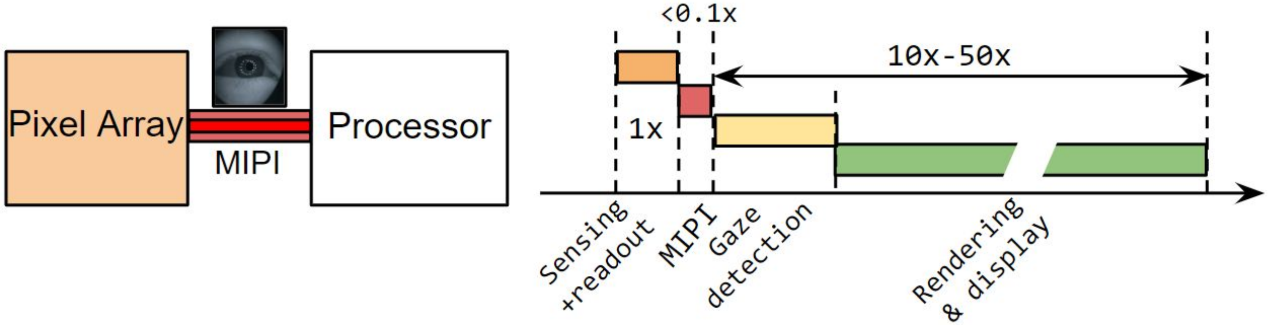
Efficient AI for Gaze-tracked Foveated Rendering





- 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.

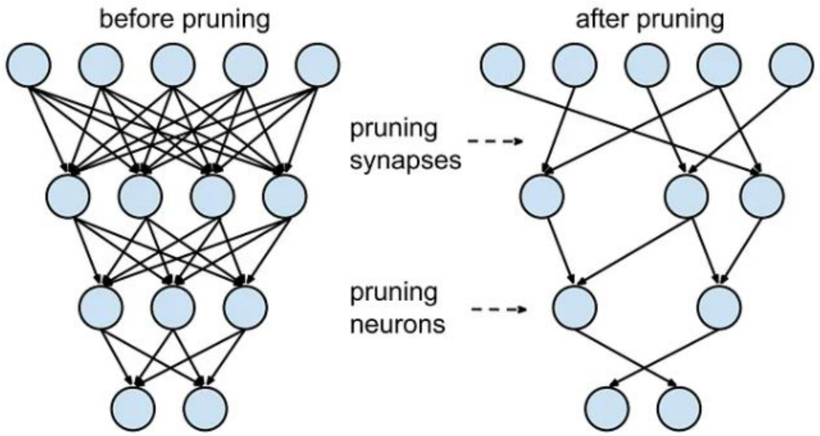
Efficient AI for Gaze-tracked Foveated Rendering

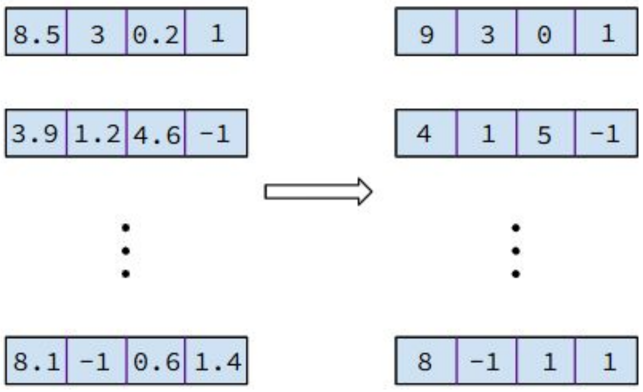


- Gaze detection with rendering and display will take majority of the processing time.
- It is critical to design an gaze tracking solution to minimize the rendering latency as well as the processing latency for gaze tracking neural networks.
- To reduce rendering latency, the gaze-tracking DNN needs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

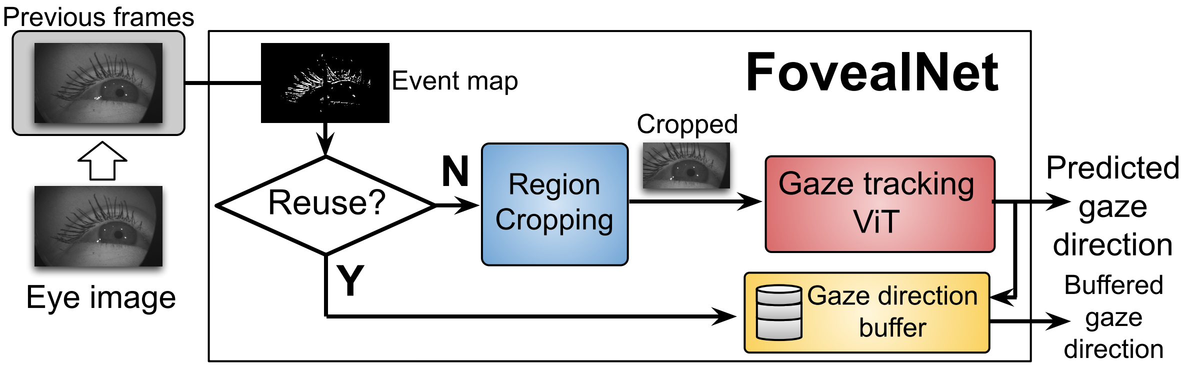




- Neural networks are highly redundant, meaning they often contain a large number of parameters and computations that contribute minimally to the final output.
- Pruning and quantization are two major approaches for neural network acceleration.

NYU SAI LAB

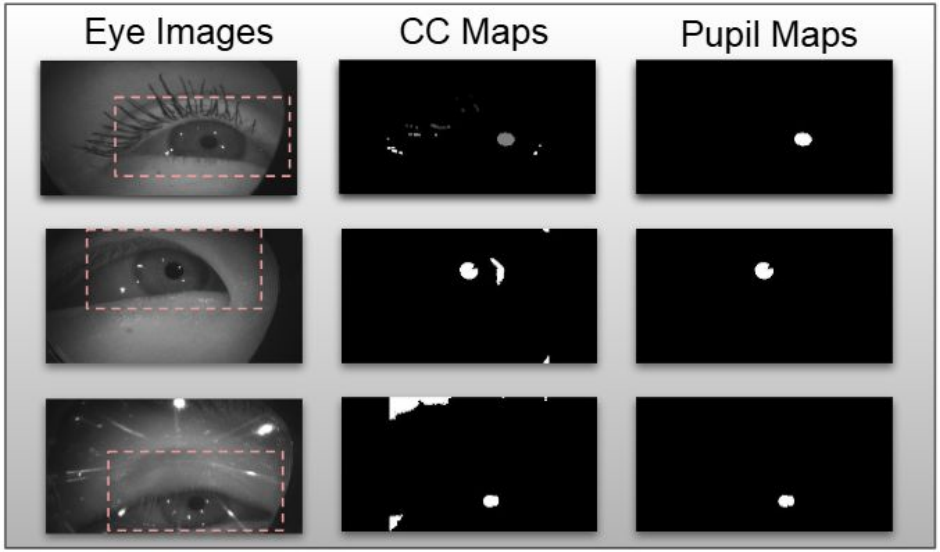
FovealNet: Overview



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

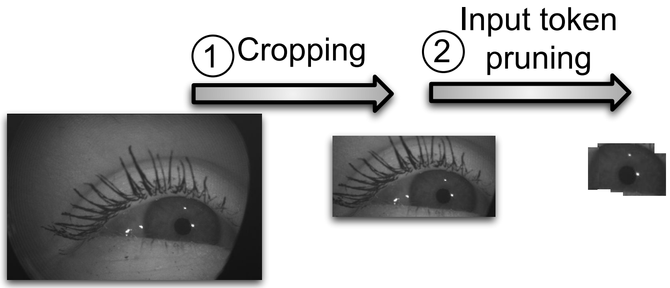
NYU SAI LAB

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.

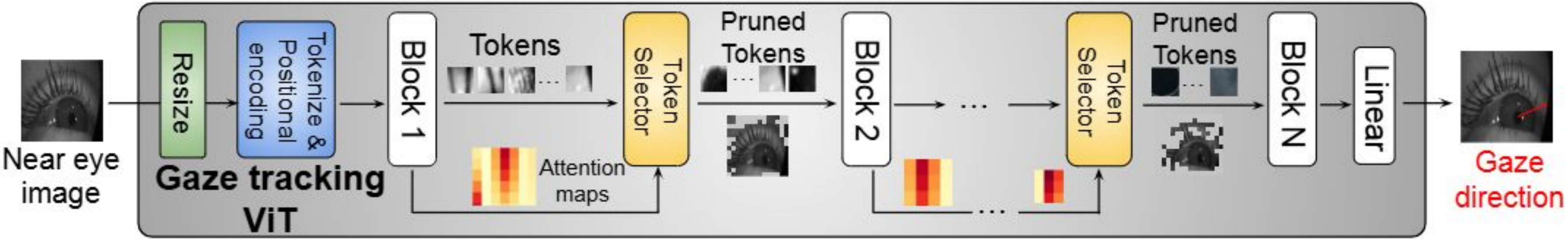
FovealNet: Gaze tracking Neural Network



- 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.

NYU SAI LAB

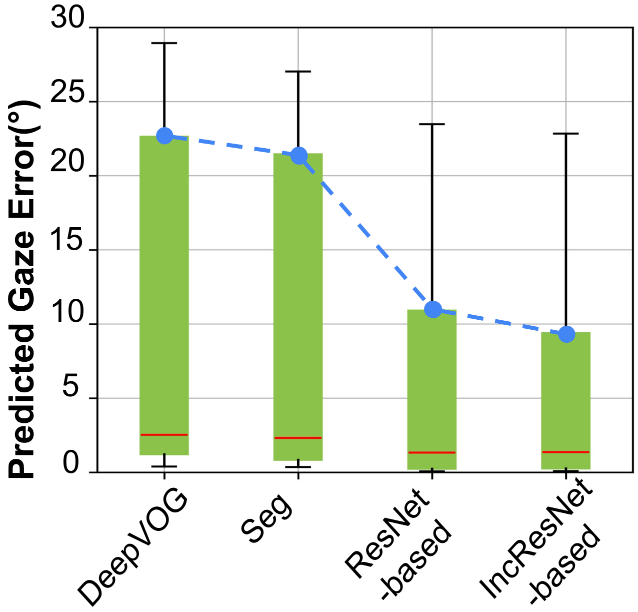
FovealNet: Gaze tracking Neural Network



- 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

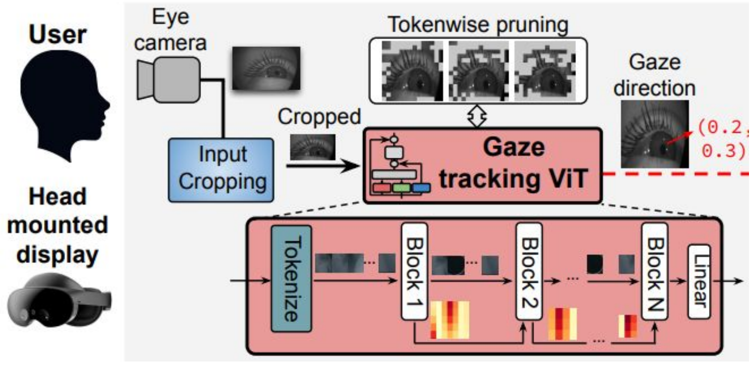
FovealNet: Evaluation Results



NYU SAI LAB

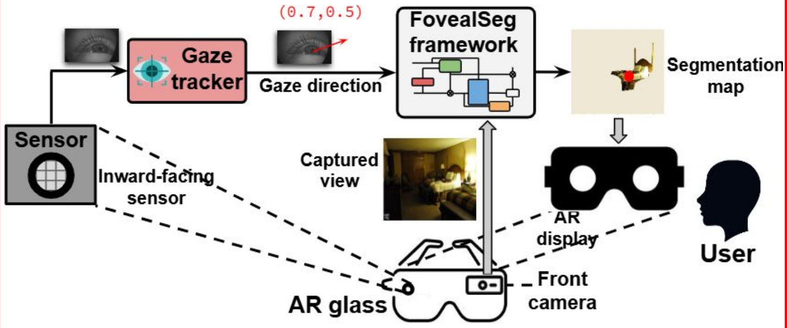
Topics

FovealNet



AI for ARVR

FovealSeg



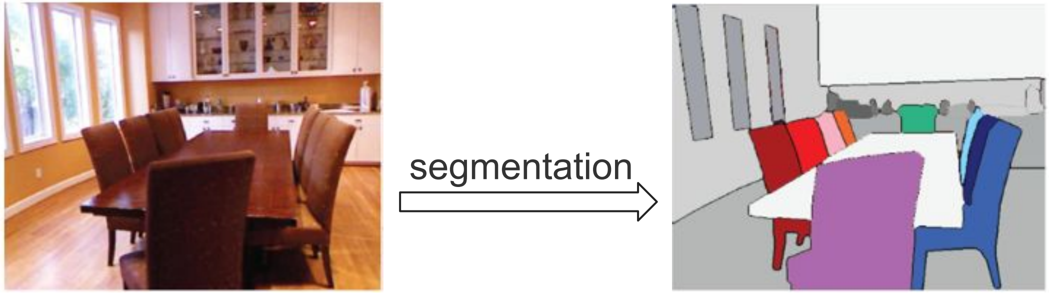
ARVR for AI

NYU SAI LAB

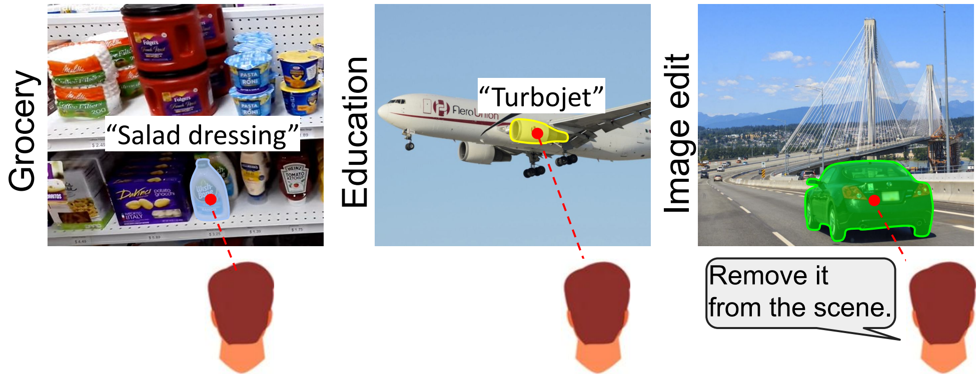
Zeng, Hongyi, et al. "Foveated Instance Segmentation." in Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

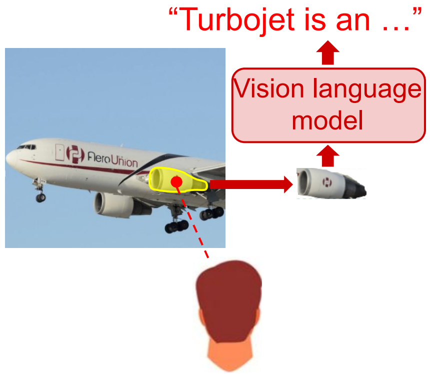
Why Segmentation is Necessary for AR?

- 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.



Instance Segmentation in AR





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

NYU SAI LAB

Segmentation is Expensive

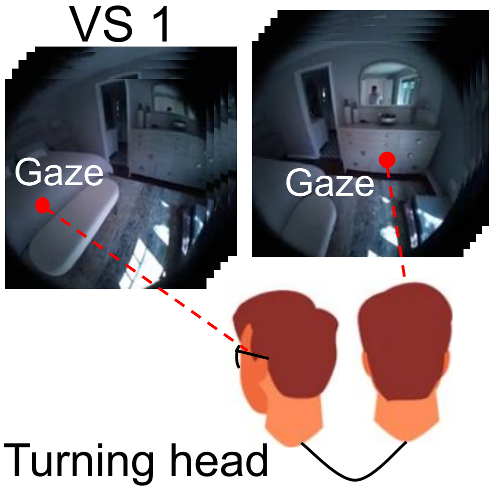
Platform	HRNet	Segformer	SAM-B	ESAM-S
Jetson Orin NX	779 ms	1419 ms	5462 ms	1307 ms
Qualcomm XR2	252 ms	880 ms	3471 ms	464 ms

1408x1408 input resolution

- Segmentation is computationally expensive.
- This latency breaks the real-time requirement essential for immersive AR experiences (70 ms).

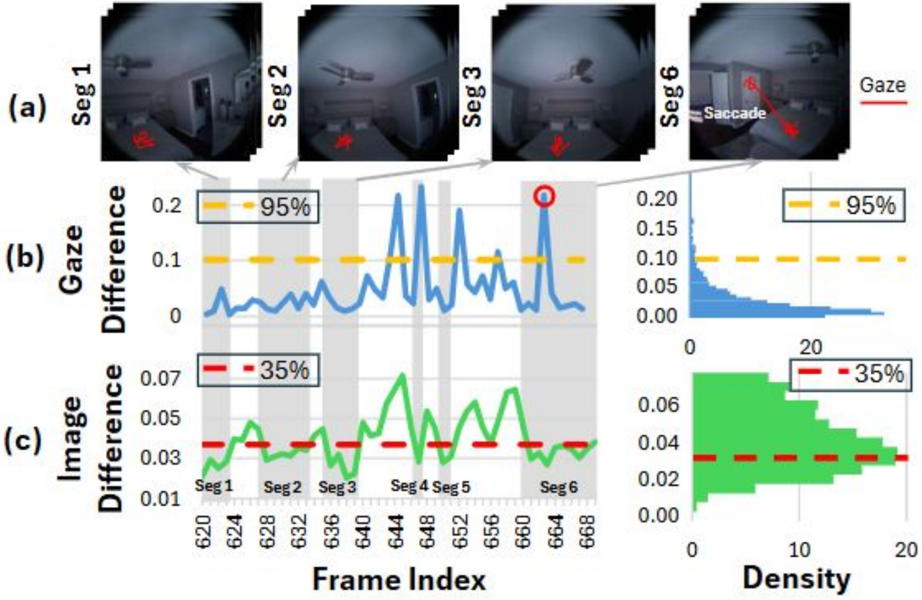
Tracked Foveated Instance Segmentation

- Human gaze alternates between **fixation** and **saccade**.
- Fixation: gaze remains still.
  - Reuse segmentation results
- Saccade: gaze moves rapid.
  - Skip segmentation



VS: video segment

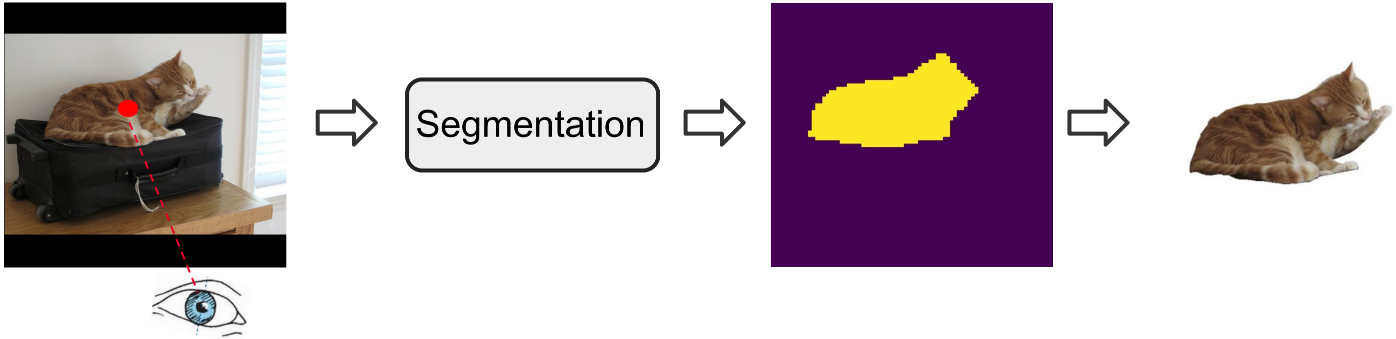
Tracked Foveated Instance Segmentation



- 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.

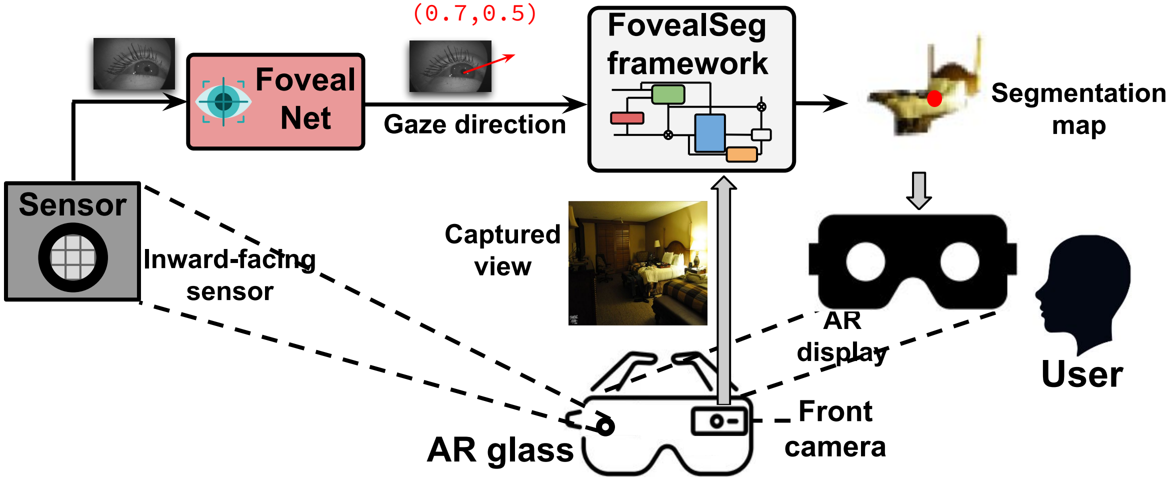
NYU SAI LAB

Instance Segmentation in AR



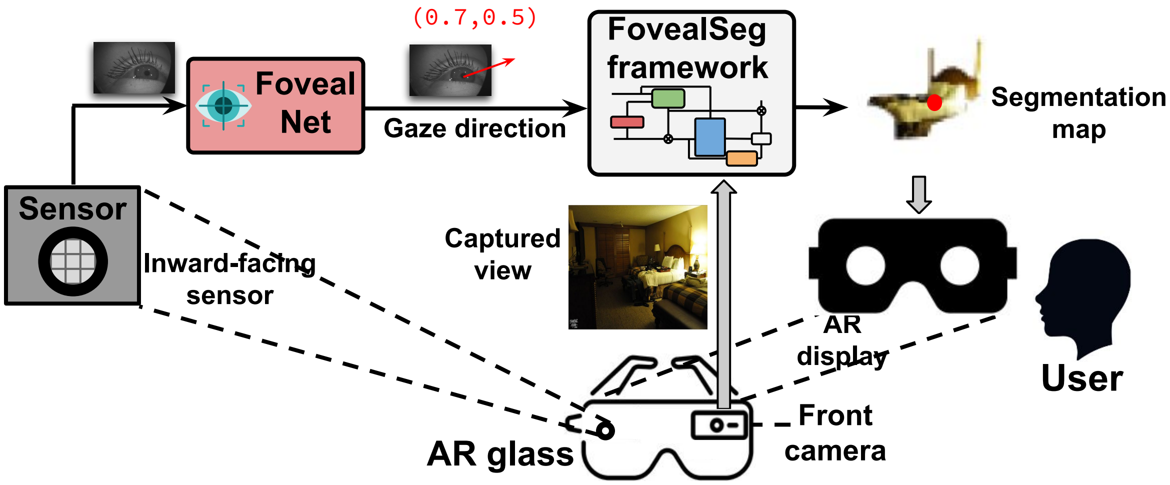
- While processing the entire image and then extracting the mask is possible, this approach would incur a significant computational cost.
- In AR, the user typically only needs to compute the segmentation masks for the instance of interest (IOI).

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).

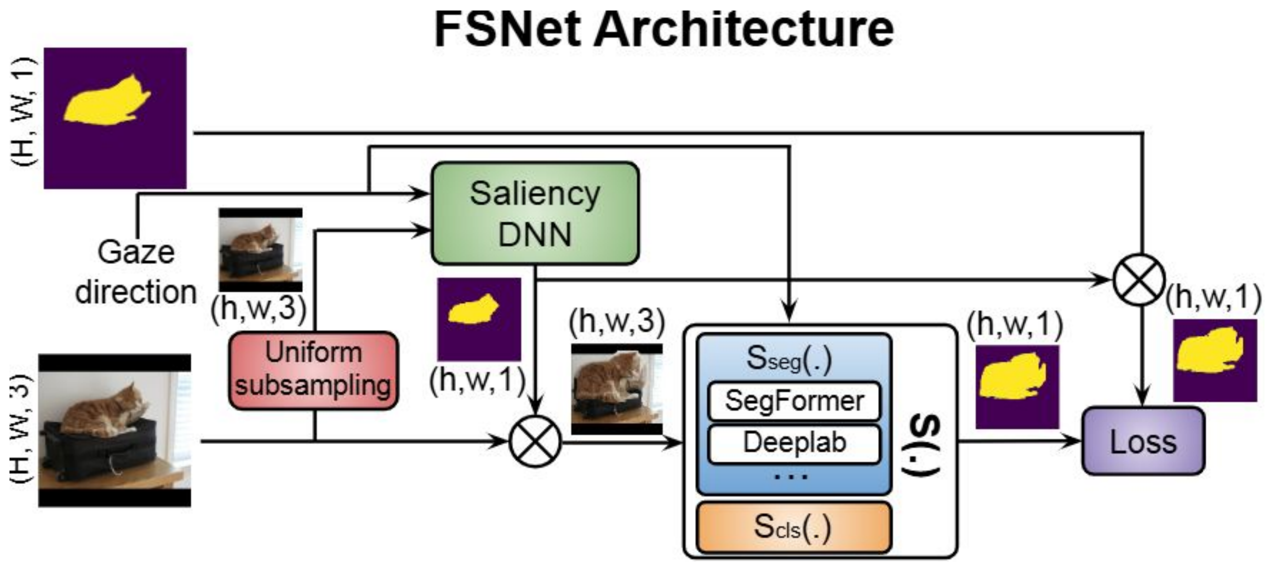
Foveated Instance Segmentation



- 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.

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 $\sum_{ij} |F^t_{ij} - F^{init}_{ij}| > \beta$ **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

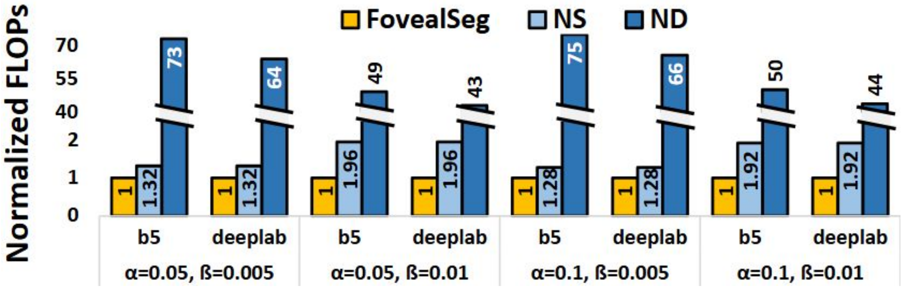
        Run FSNet with F^t and g_t, get M^t;

        $g_{last} \leftarrow g_t, M_{last} \leftarrow M_t$;

        return M_t

Evaluation Results

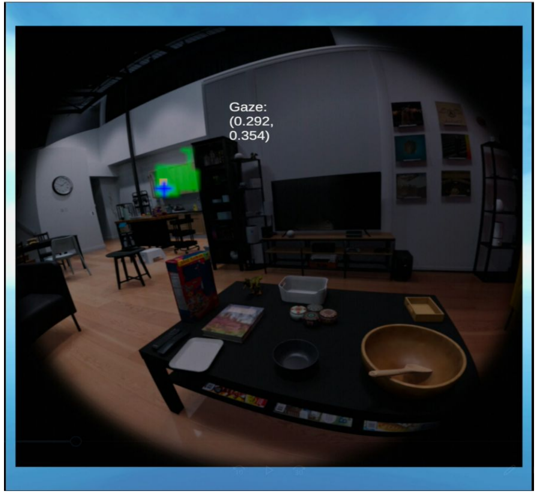
Method	Parameters(M) ↓	CityScapes (64 × 128)	
		IoU↑	IoU'↑
Avg+DeepLab	42.01	0.26	0.27
Avg+PSPNet	24.3	0.27	0.28
Avg+HRNet	67.12	0.20	0.21
Avg+SegFormer-B4	64.1	0.25	0.27
Avg+SegFormer-B5	84.6	0.27	0.29
LTD [18]	76.22	0.37	0.38
FSNet+DeepLab	42.26	**0.52**	**0.53**
FSNet+PSPNet	24.55	0.49	0.50
FSNet+HRNet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



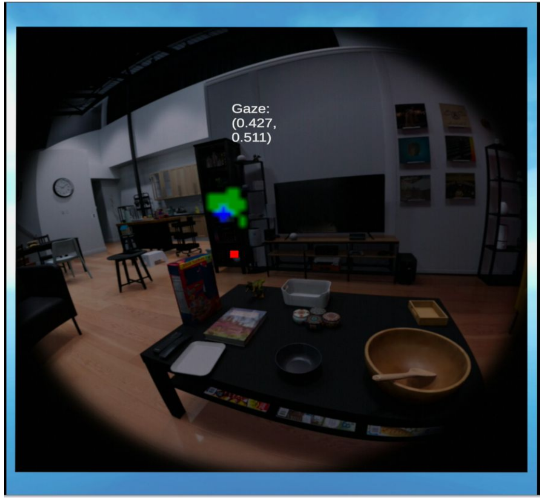
- FovealSeg (FSNet) achieves superior performance with much reduced computational cost.

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