NYU

Instant Reality: Gaze-Contingent Perceptual Optimization for 3D Virtual Reality Streaming

Shaoyu Chen, Budmonde Duinkharjav, Xin Sun, Li-Yi Wei, Stefano Petrangeli, Jose Echevarria, Claudio Silva, Qi Sun

22.03.10

PART 01

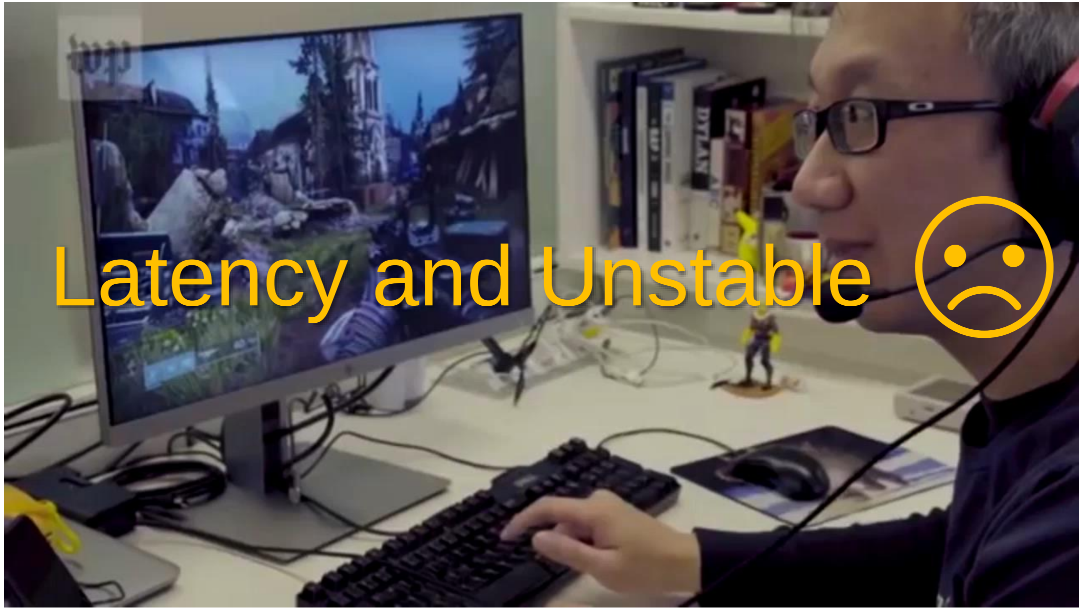
Introduction

Background

- Cloud-based streaming has widespread applications

Background

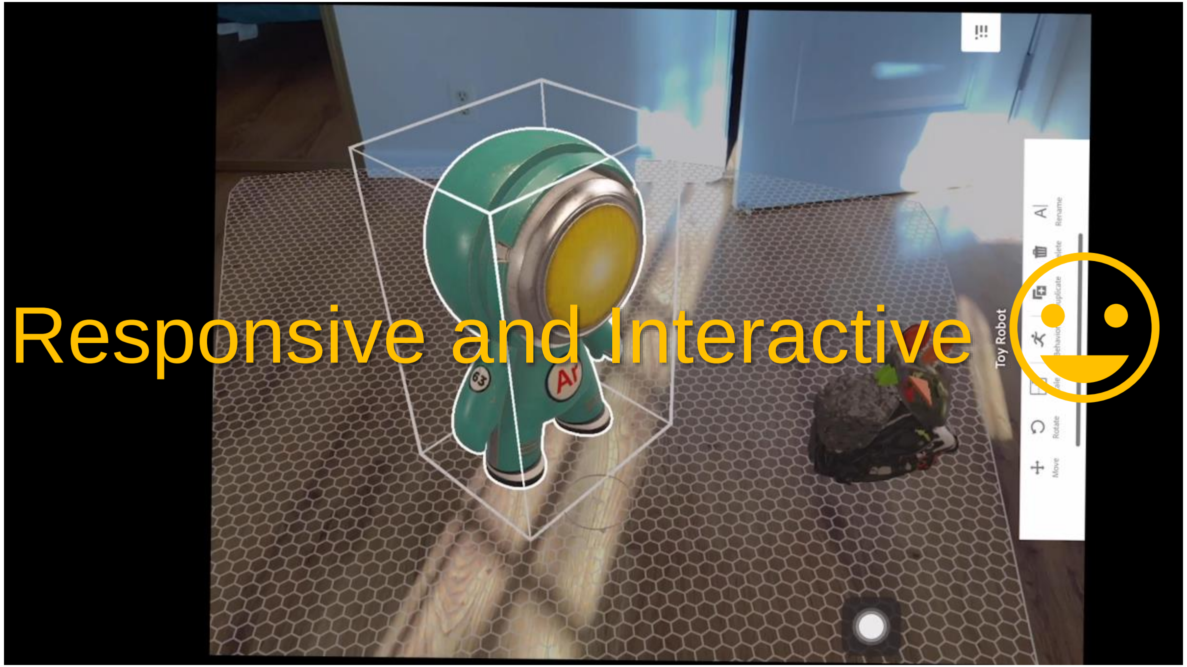
- The latency from traditional 2D video streaming may cause issues
- VR rendering needs to handle **7x** pixels/second than 2D screen



Latency and Unstable ☹

Background

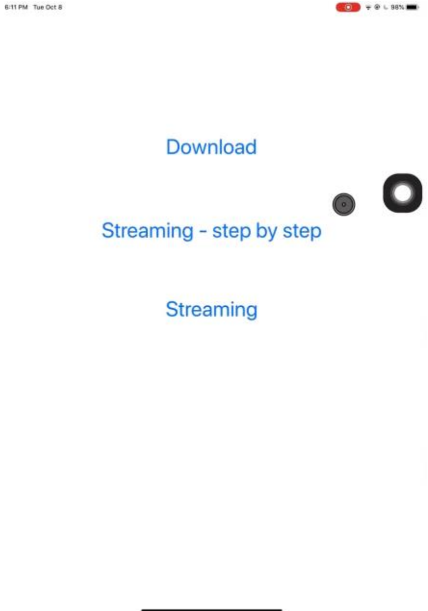
- In comparison, 3D assets can enable responsive interaction



Responsive and Interactive

Background

- 3D assets could yet be handled by existing network bandwidth
- GPU has **2.5x** FLOP while global internet bandwidth grows **26%**

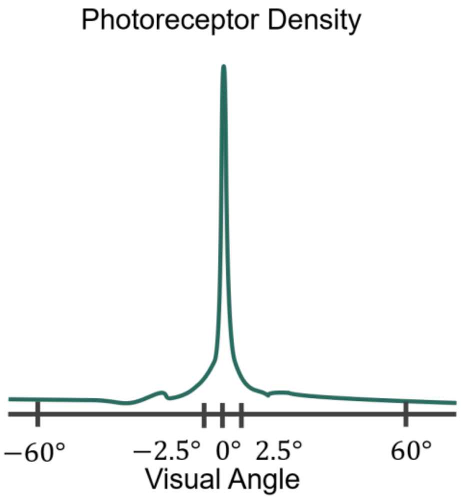


4X playback

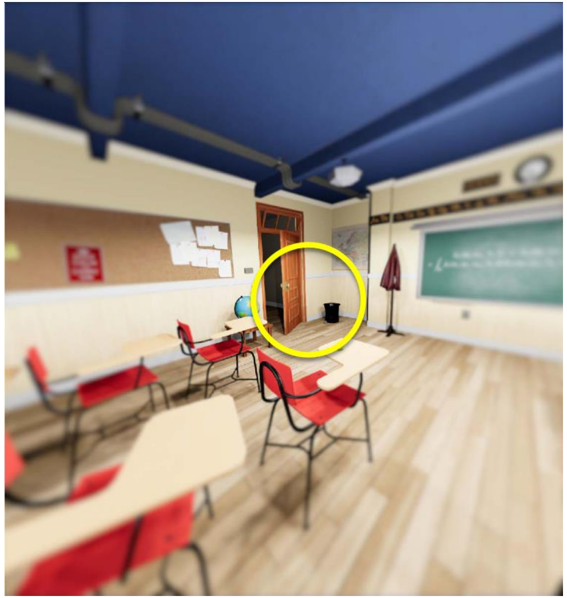
sec

Background

- Foveated rendering only works for rendering with **streamed** assets

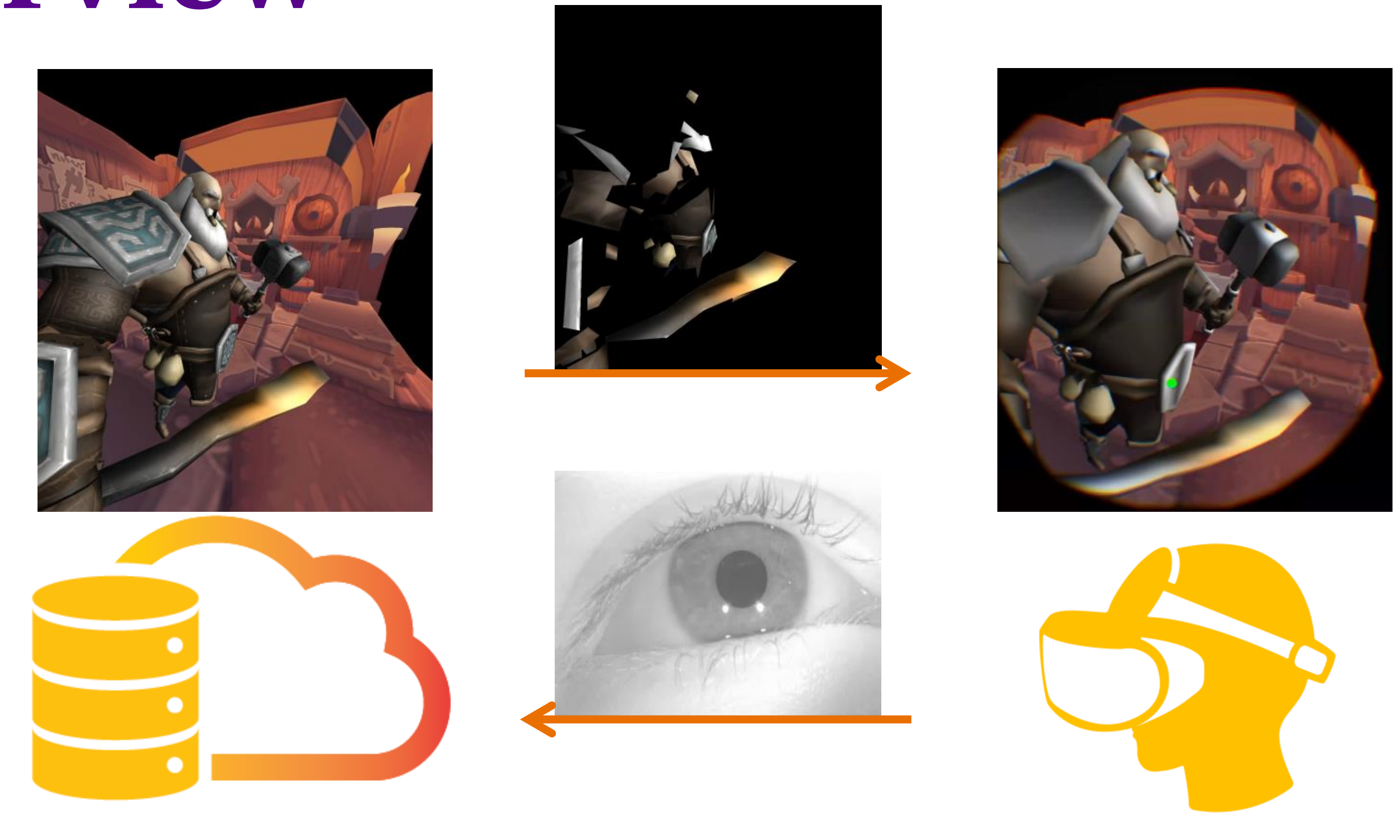


[Tursun et al., 2019]

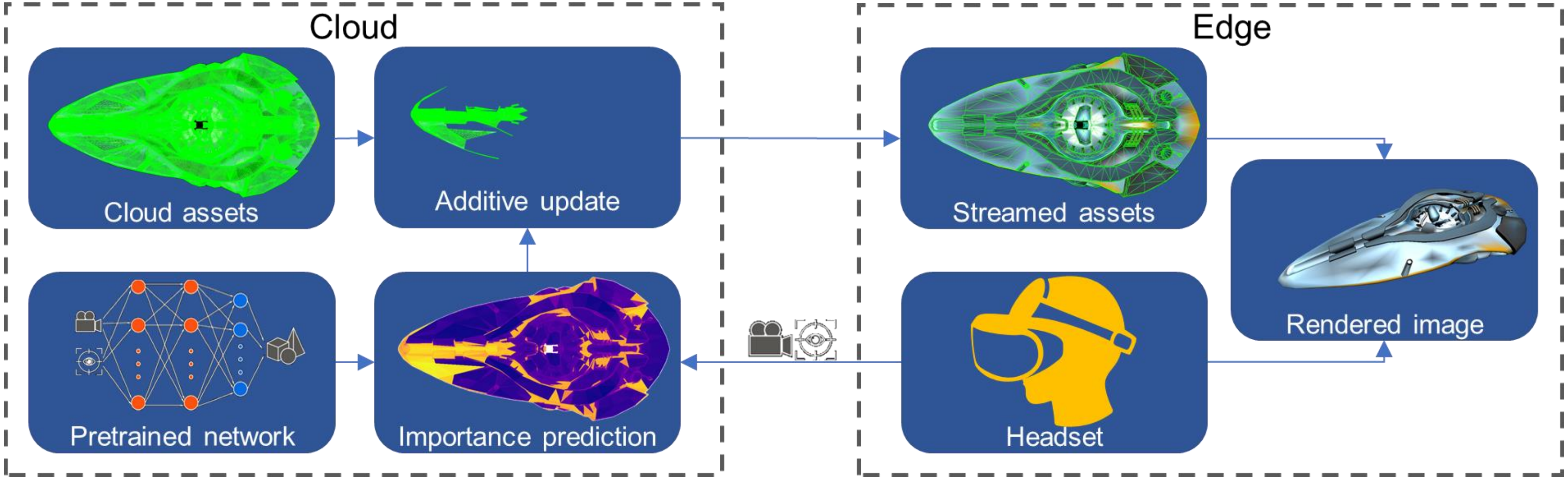


[Patney et al., 2016]

Overview

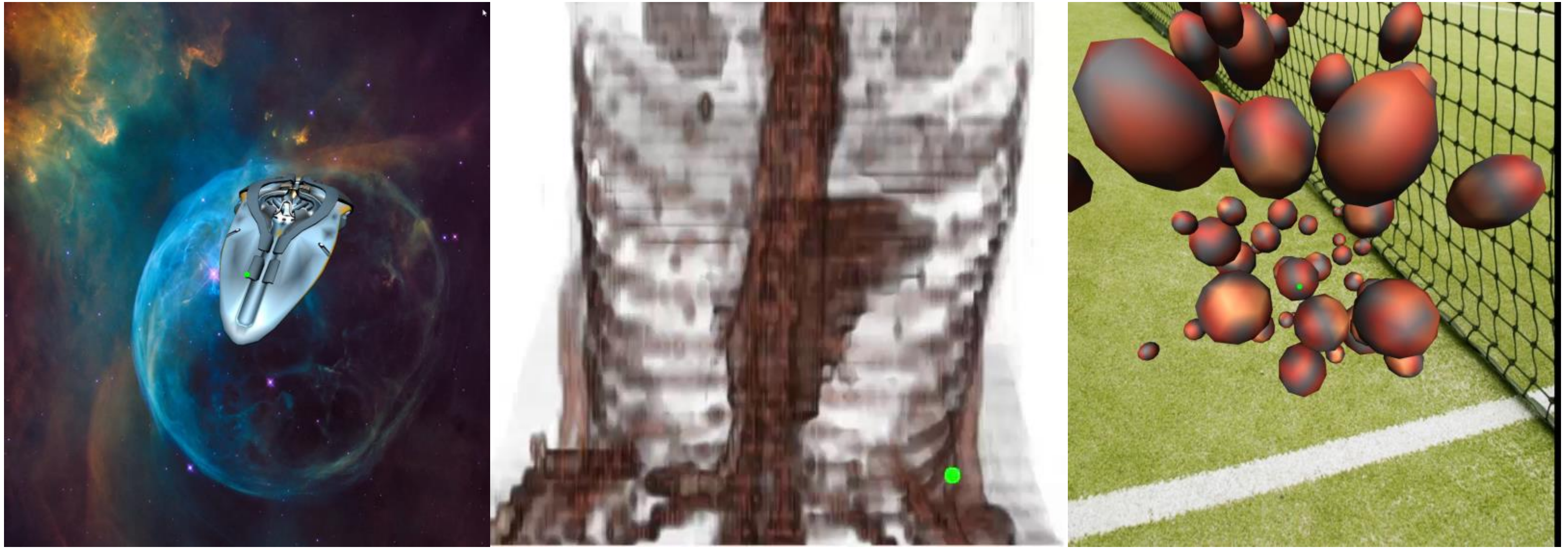


Overview


Cloud
Edge
Cloud assets
Additive update
Streamed assets
Rendered image
Pretrained network
Importance prediction
Headset

Overview

- Our method can be applied to meshes, volume, and dynamic scene



PART 02

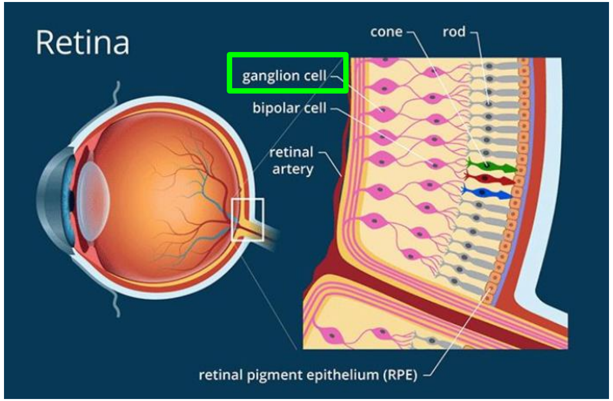
Method

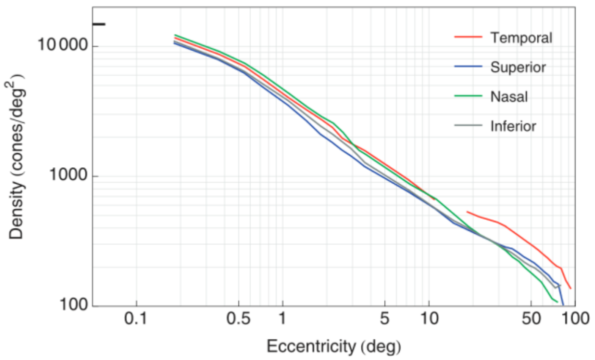
Method

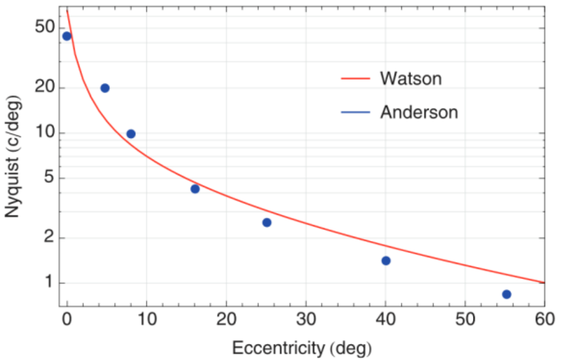
- Modeling spatio-temporal vision
    - Spatial visual acuity
    - Popping artifacts
    - Change blindness during saccade

Spatial visual acuity

- Distribution of retinal cells is not uniform
- As a result, spatial visual acuity is also non-uniform



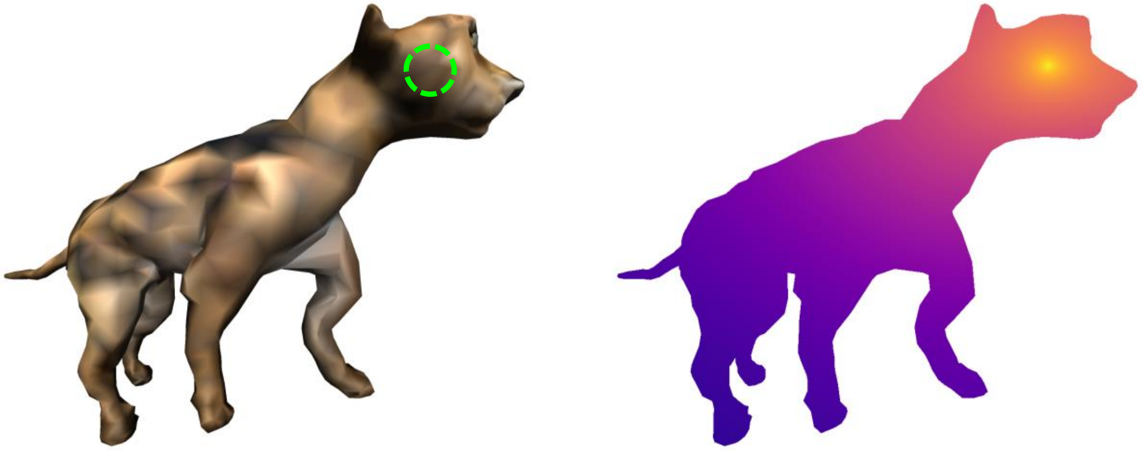




[Watson, 2014]

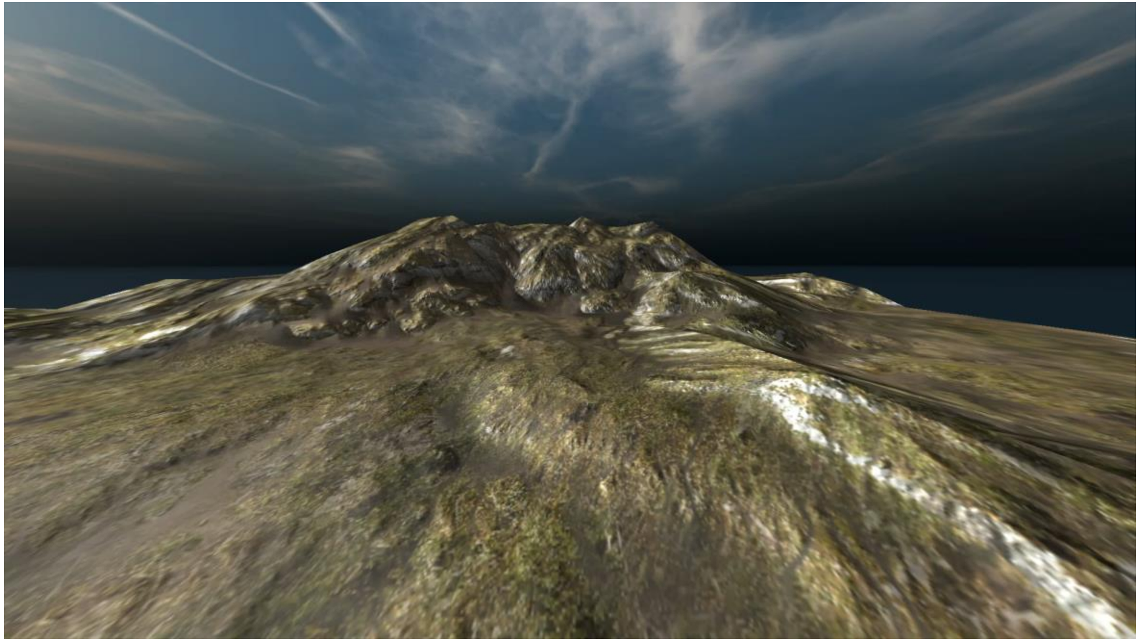
Eccentricity importance

- The importance is given by: $\hat{P}_{ec}(\mathbf{g},\mathbf{x}) = E(\mathbf{g}-\mathbf{x})$
  - ***E*** is the cell density function, ***x*** is pixel position and ***g*** is gaze position



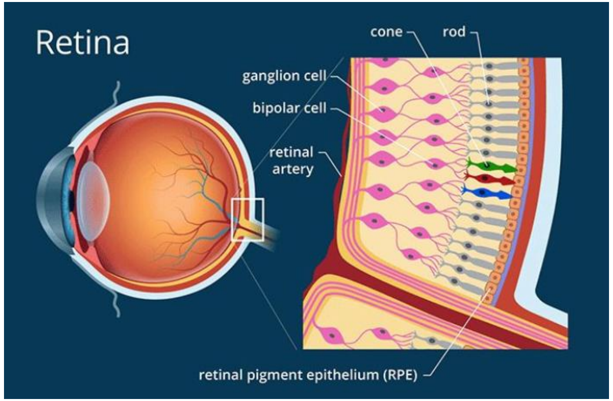
Popping artifacts

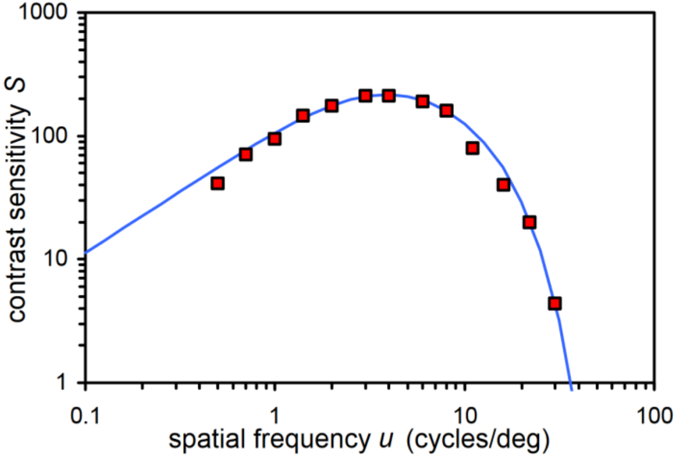
- A major problem of traditional LoD-based procedural rendering



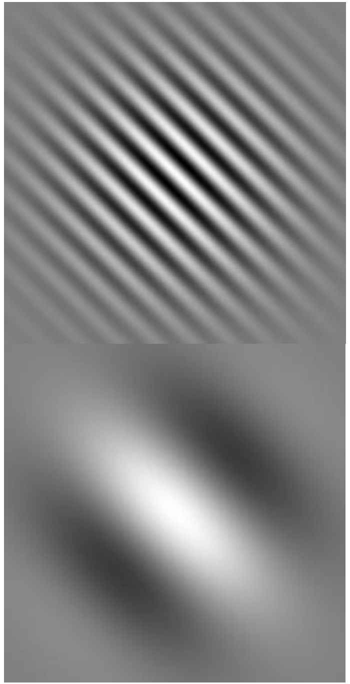
Model perception of images

- In order to minimize the perceive change
- We first model how human perceive static images





[Barten, 2004]



Model perception of images

- We model human perception on an image as:

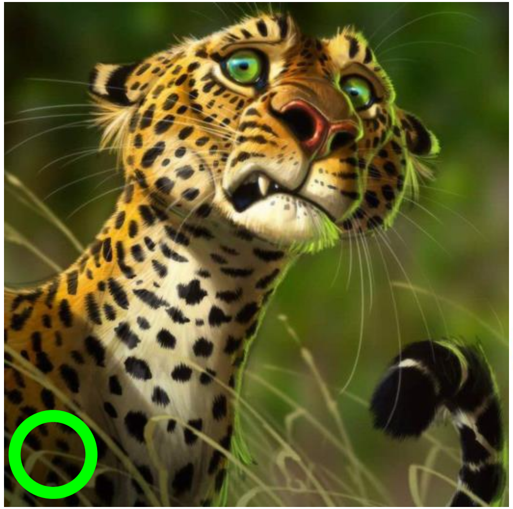
$$\hat{\Phi}(\mathbf{g},\mathbf{x},I) \approx \sum_{i=0}^{b-1} \underline{s\left(\mathbf{f}_i, L\right)}\, \underline{c\left(\mathbf{x},\mathbf{f}_i, I\right)}$$

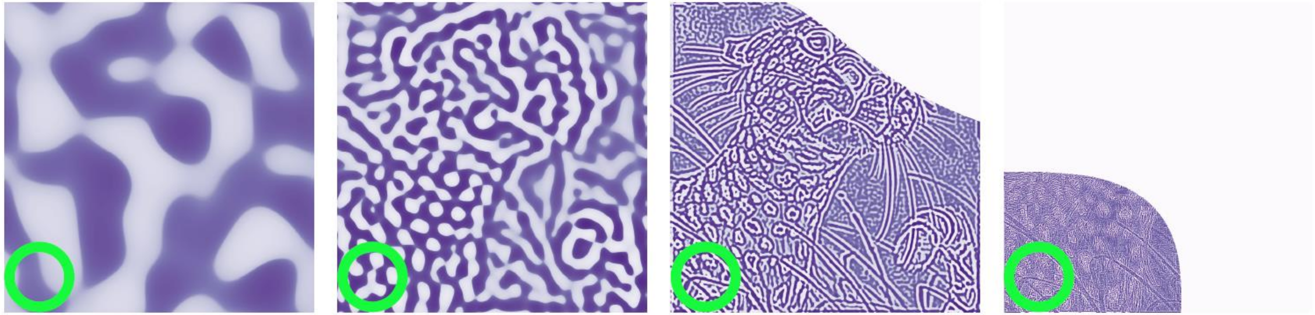
gaze pixel image

sensitivity function
given frequency band ***f***,
brightness of display ***L***

bandpass filtered
point contrast

Model perception of images





(a) 4 cycle / im (b) 16 cycle / im (c) 64 cycle / im (d) 256 cycle / im

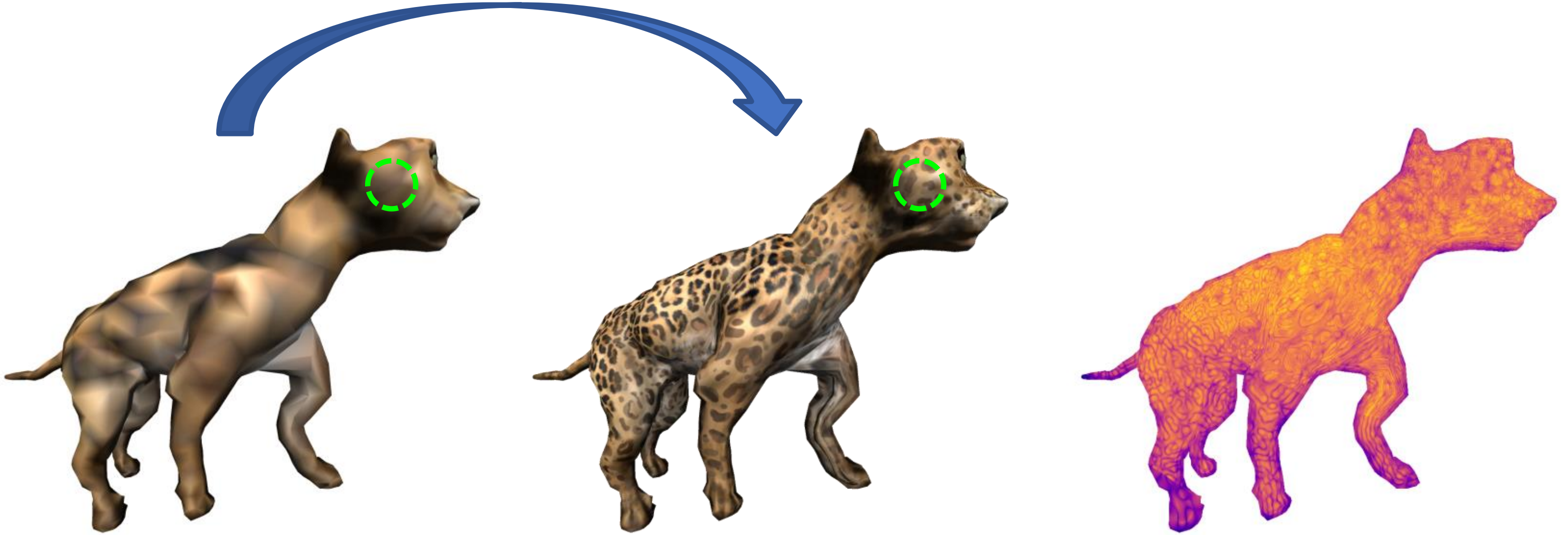
Decomposition visualization of bandpass filtered contrast
The periphery sensitivity was clamped by ***E***

Temporal consistency

- Similarly, we model the perceived change as temporally adapted **Weber's contrast** to individual frequency band

$$\hat{P}_{op}(\underbrace{\mathbf{g}}_{\text{gaze}}, \underbrace{I}_{\substack{\text{current}\\\text{image}}}, \underbrace{I'}_{\substack{\text{changed}\\\text{image}}}, \underbrace{\mathbf{x}}_{\text{pixel}}) = \sum_{i=0}^{b-1} \underbrace{s\left(\mathbf{f}_i, L\right)}_{\substack{\text{sensitivity function}\\\text{given frequency band } \boldsymbol{f},\\\text{brightness of display } \boldsymbol{L}}} \times \frac{\left|c(\mathbf{x}, \mathbf{f}_i, I) - c(\mathbf{x}, \mathbf{f}_i, I')\right|}{\underbrace{\left|c(\mathbf{x}, \mathbf{f}_i, I)\right|}_{\substack{\text{bandpass filtered}\\\text{point contrast}}} + \underbrace{\omega}_{\substack{\text{balancing parameter}\\\text{for low-intensity stimuli}}}}$$

Temporal consistency



LOD 1 image

LOD 0 image

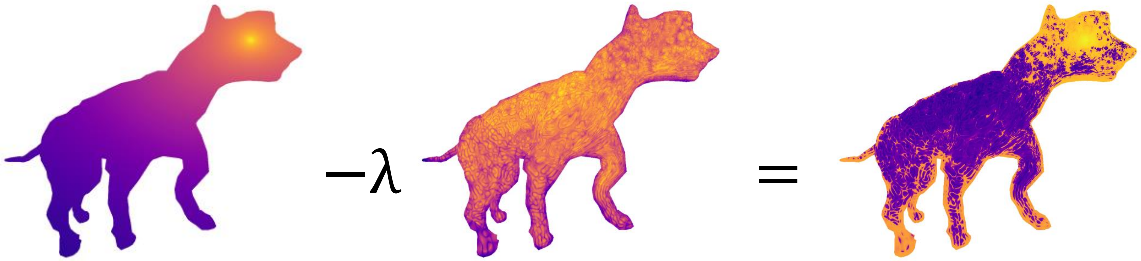
Temporal consistency heatmap

Saccade

- Fast eye movements with gaze speed > 180 deg/sec

Per-pixel importance

$$\hat{P}(\mathbf{g}, I, I', \mathbf{x}) = \begin{cases} \hat{P}_{ec}(\mathbf{g}, \mathbf{x}) - \lambda \hat{P}_{op}(\mathbf{g}, I, I', \mathbf{x}) & \text{during fixation} \\ \int_{\mathbf{g}' \in I'} \hat{P}_{op}(\mathbf{g}', I, I', \mathbf{x}) \mathrm{d}\mathbf{g}' & \text{during saccade} \end{cases}$$



$\hat{P}_{ec}(\mathbf{g}, \mathbf{x})$

Eccentricity importance

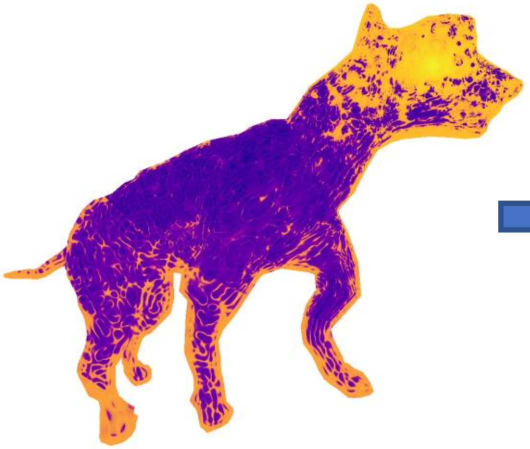
$\hat{P}_{op}(\mathbf{g}, I, I', \mathbf{x})$

Temporal consistency

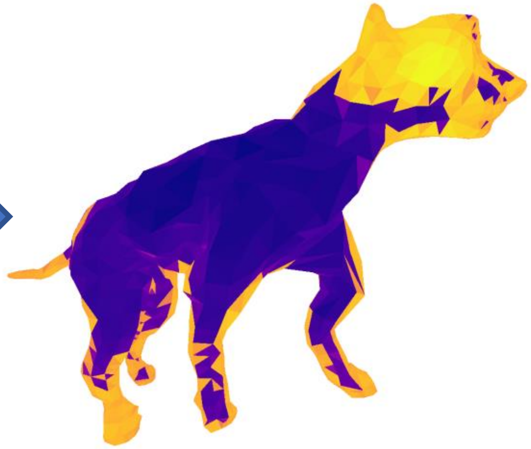
$\hat{P}(\mathbf{g}, I, I', \mathbf{x})$

Per-pixel importance

Mapping from 2D to 3D



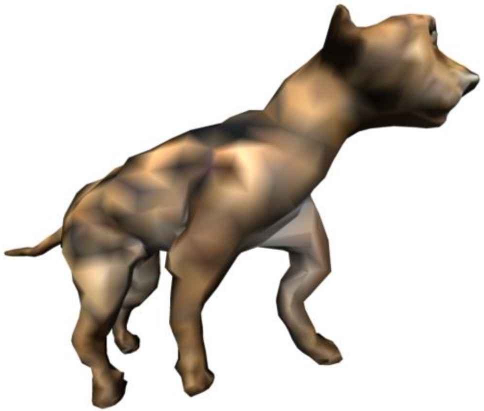
Per-pixel importance



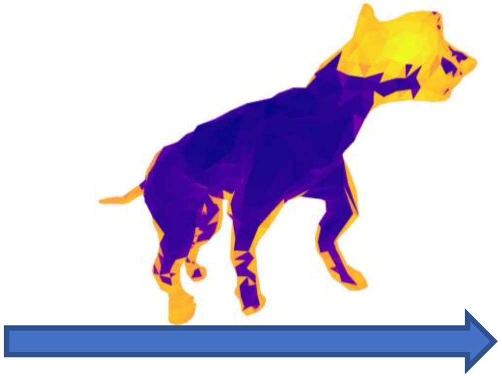
Per-3D-unit importance

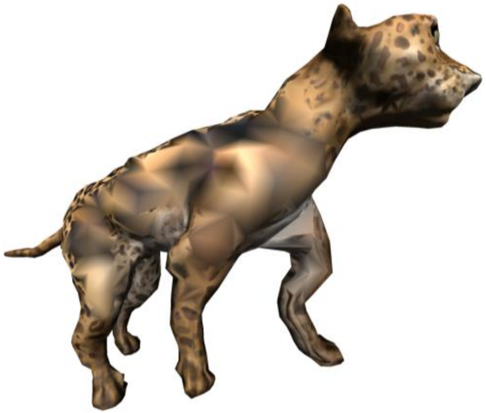
Streaming

- We use a greedy approach to fill the update to be streamed



Before update





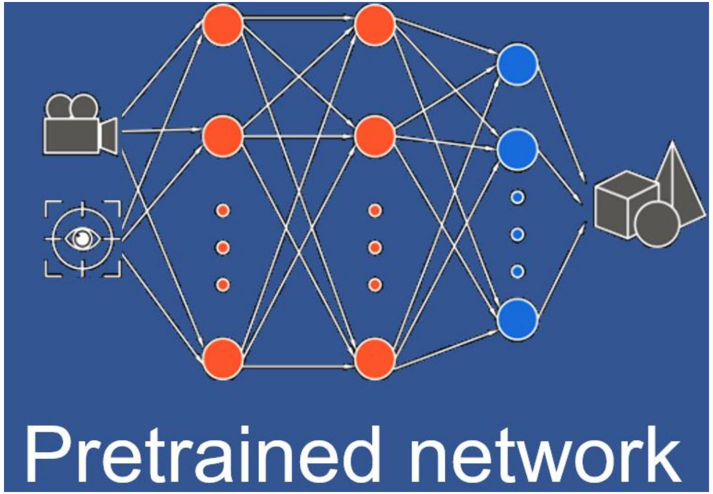
After update

Neural Acceleration

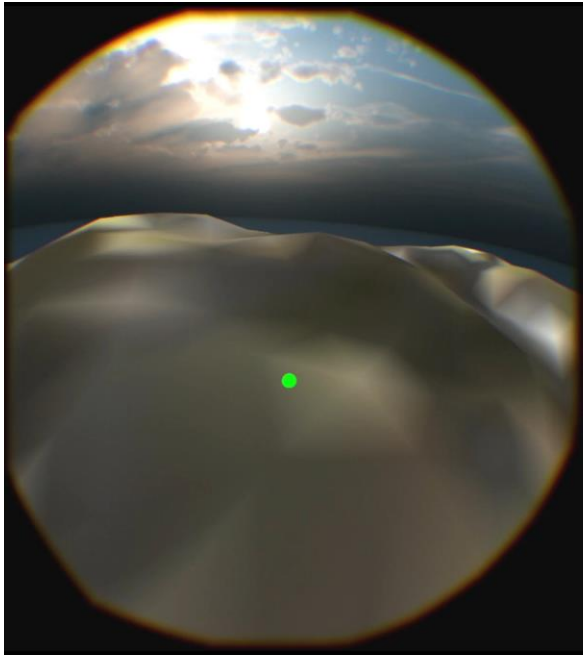
- Intolerable latency can be introduced during the heavy frequency domain decomposition for the temporal consistency calculation
- For **fast prediction** of the importance, a multilayer perceptron neural network is trained
- Cloud can **skip rendering** the actual image with neural acceleration

Neural Acceleration

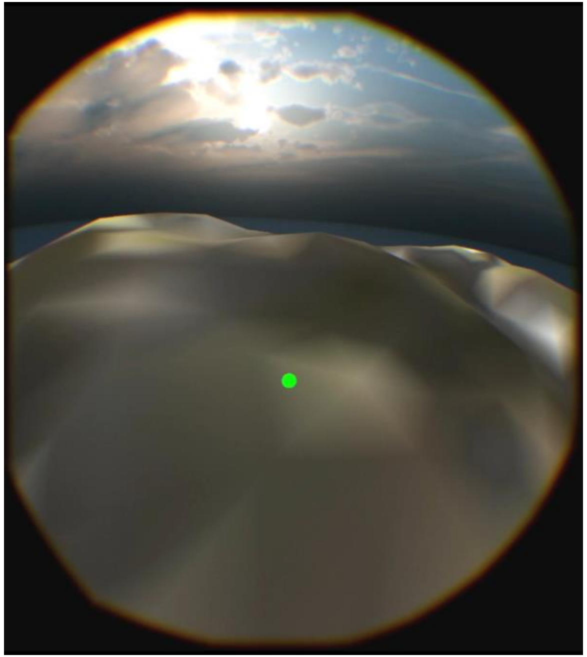
- Trained for a specific scene
- Input: camera position, camera direction and gaze position
- Output: predicted importance of each 3D asset in the scene



Neural Acceleration



w/o acceleration



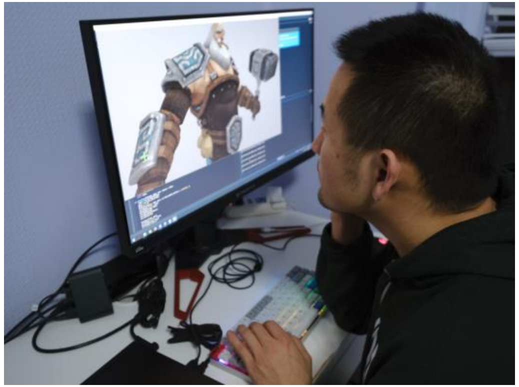
with acceleration

PART 03

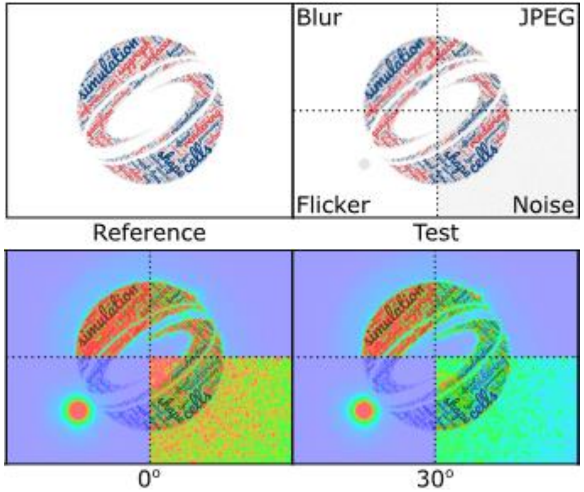
User Study

Evaluation

Eye-tracked study

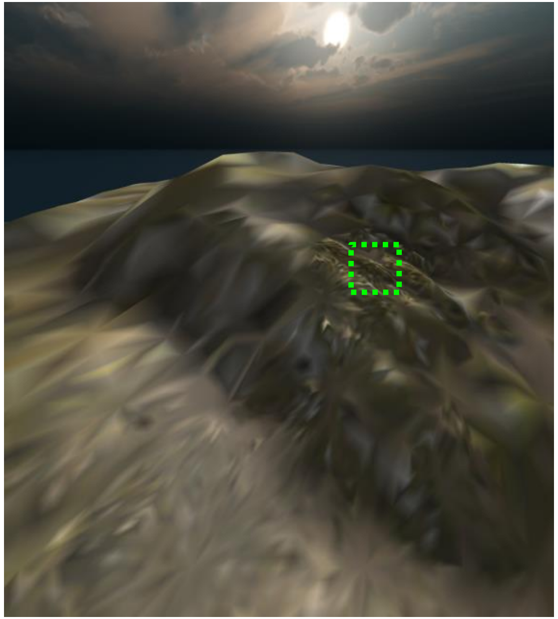


Screen-based study

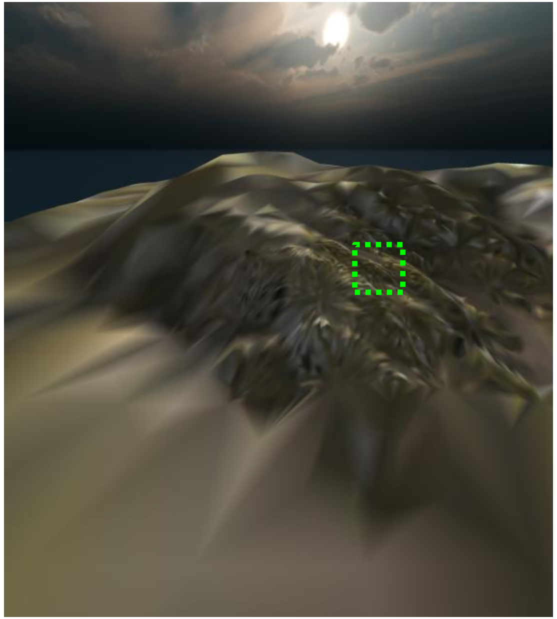


Objective analysis

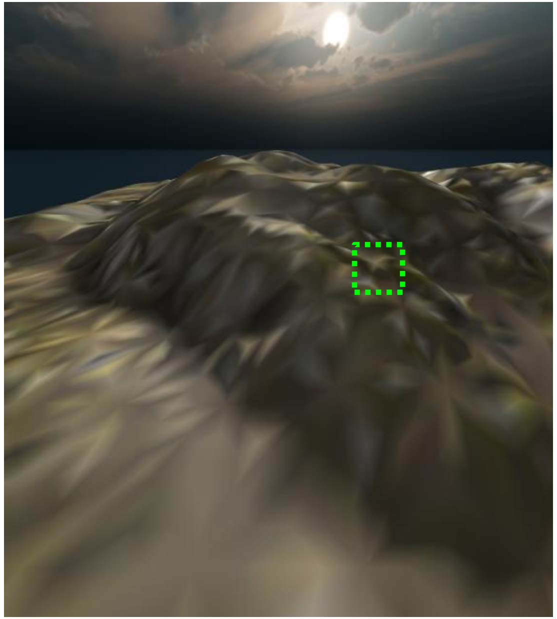
Evaluation



OURS



ECC

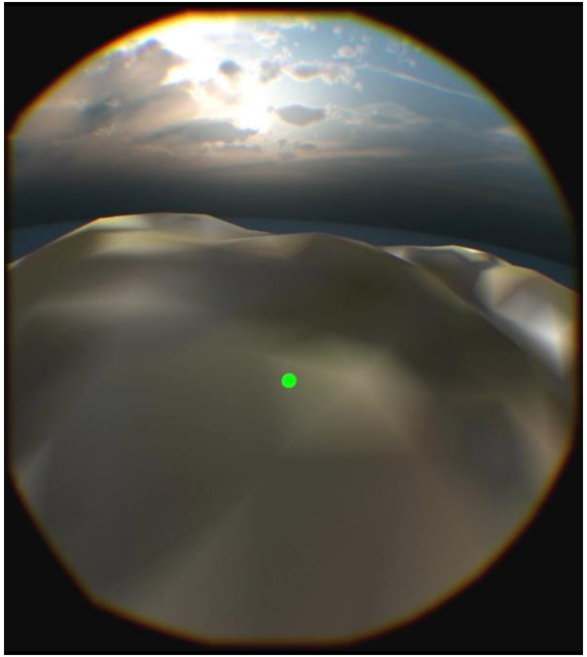


UNI

$$\hat{P}_{ec}(\mathbf{g},\mathbf{x}) \cancel{-\lambda \hat{P}_{op}(\mathbf{g},I,I',\mathbf{x})}$$

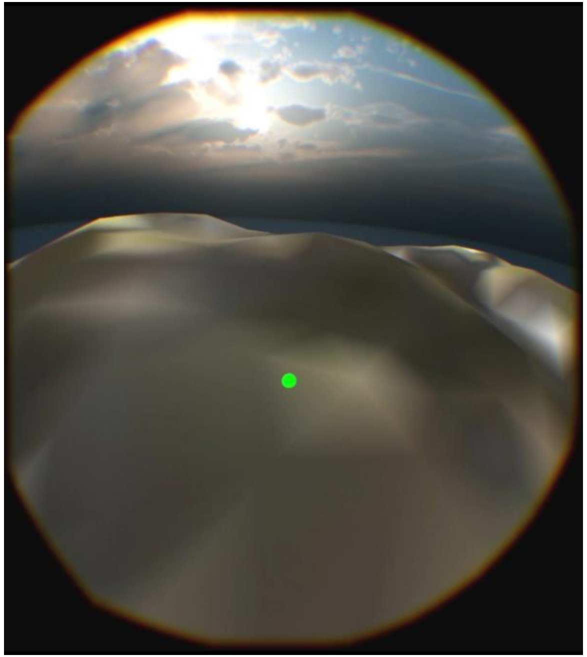
Eye-tracked study

- Task - two-alternative-forced-choice (2AFC) experiment
  - Each trial consists of a pair of conditions among the UNI/ECC/OURS
  - Participants select which condition appeared more smoothly and comfortably updated with fewer artifacts over the entire duration



Eye-tracked study

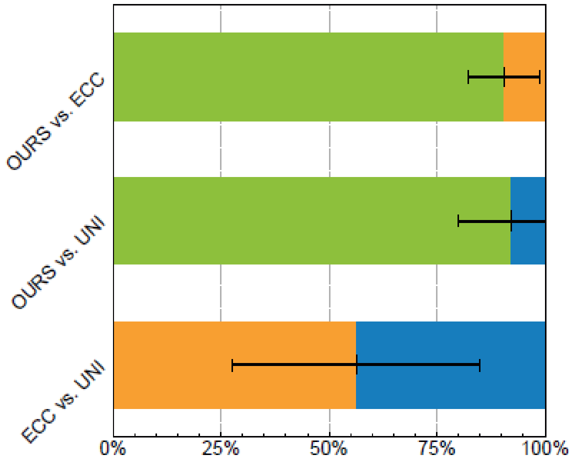
- Why didn’t task focus on visual quality?
  - Participants cannot focus on two different aspects
  - There exists objective metrics for visual quality like FovVideoVDP
  - Limited human visual perception during natural, active viewing conditions



NYU

Eye-tracked study

- Each pair of comparison contains:
  - 8 participants * 8 trials/participant
  - = 64 trials in total
- Consistency: $OURS > ECC \approx UNI$

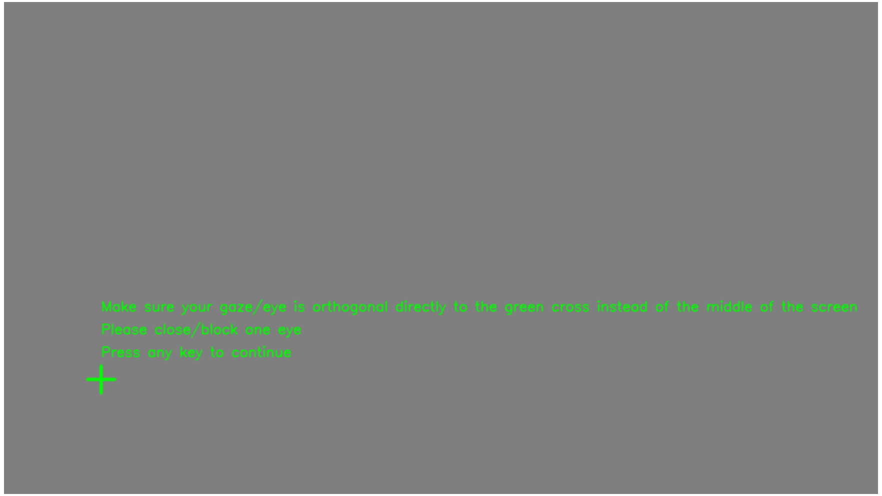


VR eye-tracked temporal consistency

NYU

Screen-based study

- Visual stimuli rendered with 1920×1080 resolution and 60 degree of vertical FoV
- Our protocol automatically compute and inform participants of the correct eye-display distance



Make sure your gaze/eye is orthogonal directly to the green cross instead of the middle of the screen

Please close/block one eye

Press any key to continue

Screen-based study

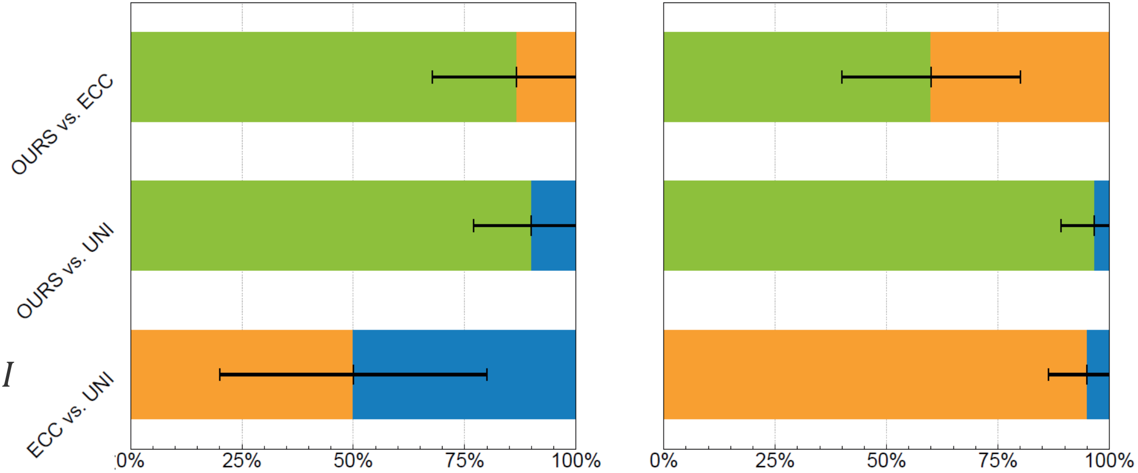
- Task 1 – temporal consistency
  - Similar to eye-tracked study
  - Except that user gaze is fixed so that there is no saccade

Screen-based study

- Task 2 – visual quality
  - First observe full-quality rendering
  - Then, 2 static images of different conditions are sequentially displayed
  - The 2 images are sampled from the sequences in task 1 at the same timestamp

Screen-based study

- Each pair of comparison contains:
  - 12 participants * 5 trials/participant
  - = 60 trials in total
- Consistency: $OURS > ECC \approx UNI$
- Quality: $OURS \approx ECC > UNI$



Video-based temporal consistency

Video-based quality

PART 04

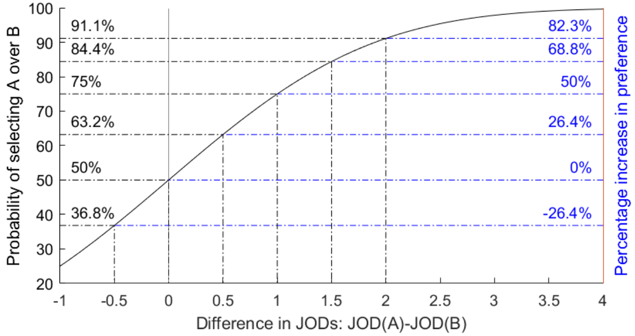
Objective Analysis

FovVideoVDP

- Full-reference visual quality metric predicts perceptual difference
- Report quality in the JOD (Just-Objectionable-Difference) units

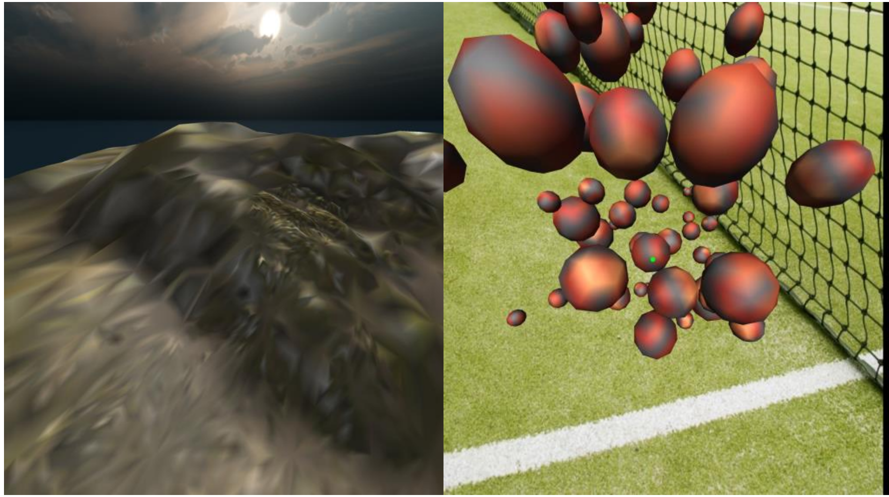
JOD 7.4506

JOD 6.4633



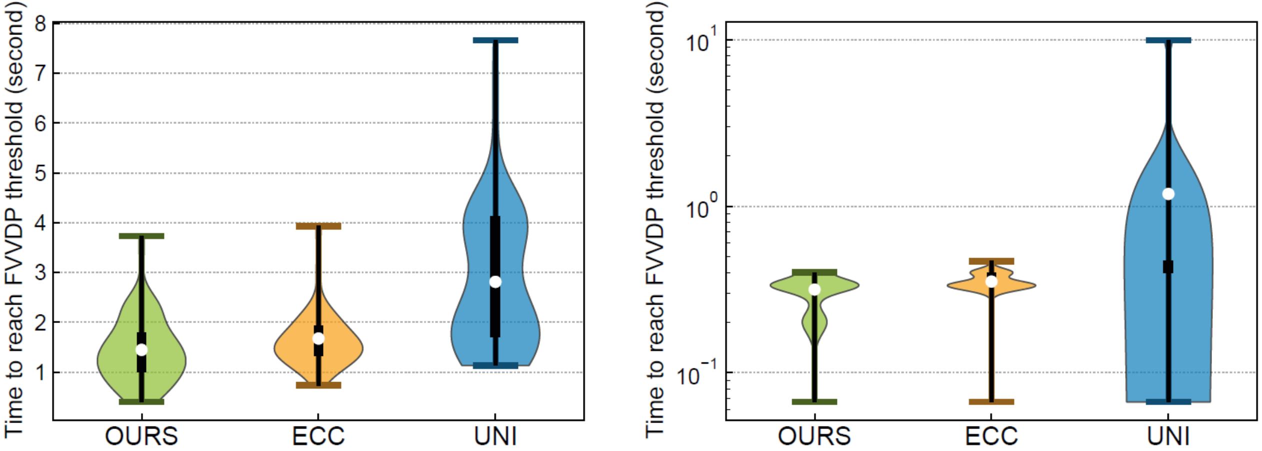
Visual quality

- Use FovVideoVDP as the metric
- Sample 10-second gaze sequences from eye-tracked user study
- Measure the timing when FovVideoVDP reaches a shared threshold



Visual quality

- $OURS \approx ECC > UNI$ in both static and dynamic scenes

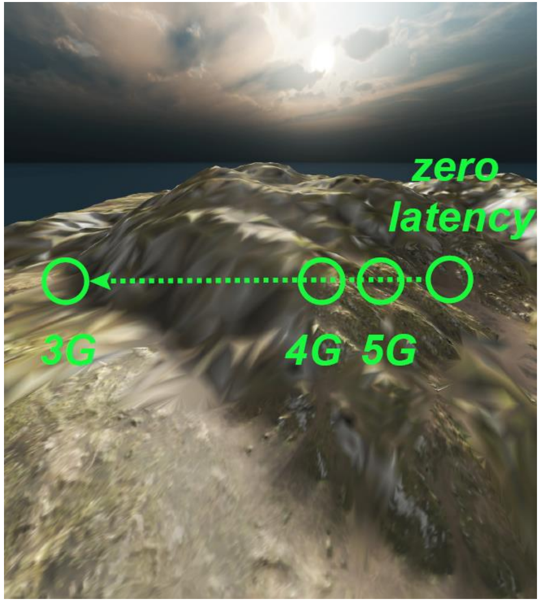


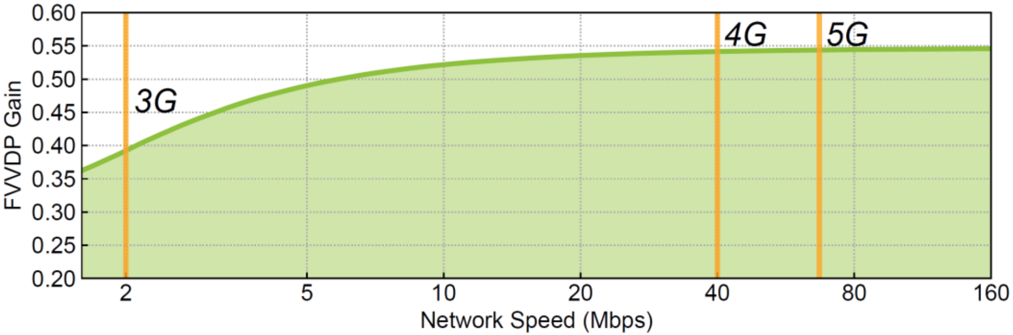
(a) static scene

(b) dynamic scene

Network

- We measure the FovVideoVDP for OURS and UNI under same network condition, and use the difference as the gain of OURS

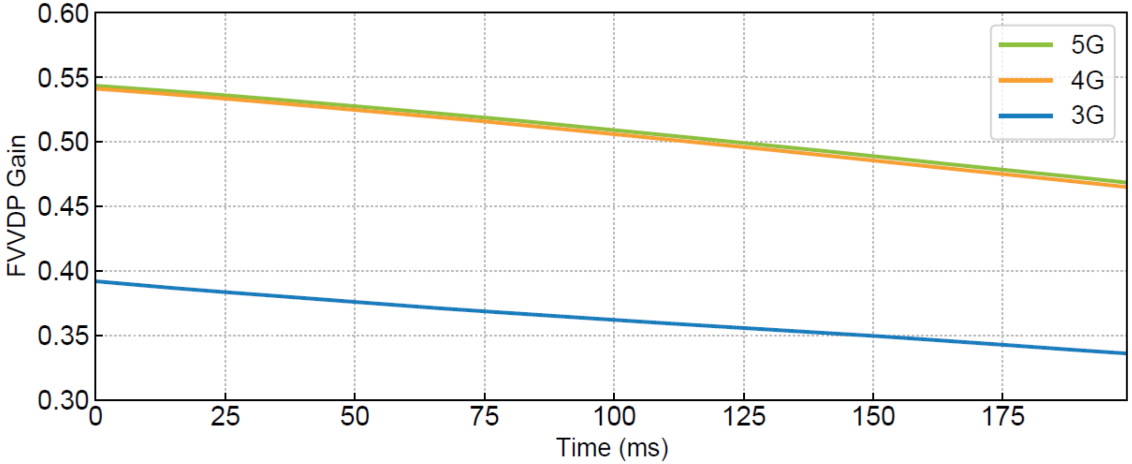




quality gain w.r.t. bandwidth

Network

- We also measure the gain under different latencies at 3G/4G/5G speed



quality gain w.r.t. artificially introduced network latencies

PART 05

Conclusion

Summary

- Compared with 2D frame-based streaming, our 3D streaming method enables low-latency interaction, low cloud overload
- Our system delivers a statistically significant reduction of temporal artifacts without compromising the visual quality
- Our system can work well under different network conditions

Limitation and future work

- Only **foveation** and **saccade** are used as the main perceptual mechanisms
- Neural network only trained in **static** scene
- Our framework only **mitigates** the perceived flickering
- Gaze motion **prediction** can be used in the future

NYU

Thank you!