ReFrame: Layer Caching for Accelerated Inference in Real-Time Rendering

Lufei Liu and Tor M. Aamodt
University of British Columbia

THE UNIVERSITY OF BRITISH COLUMBIA

ICML International Conference On Machine Learning

Project Website + Code

15 out of 16 pixels are AI-generated in real-time rendering

Accelerating neural network inferences is necessary for better graphics

- Real-time rendering is important for video games, AR/VR applications, scientific simulations, and 3D design.
- Neural network inferences are commonly used in the post-processing stage of real-time rendering to augment low-quality renderings achieved using rasterization or ray tracing.
- Several rendering stages are required to create every frame and strict latency requirements enforce desired frame rates.

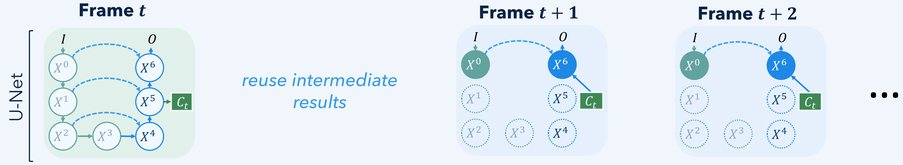
Diffusion Models vs. Rendering

- Neural networks for real-time rendering share similarities with diffusion models.
- Techniques designed for diffusion models can be adapted to support rendering workloads.

Diffusion Model	Rendering
Often applies a U-Net / Encoder-Decoder architecture.	
Relies on repeated forward passes to generate output.	
Exhibits high temporal redundancy between forward-pass inferences.	
Behavior of forward passes follows a predictable pattern.	Behavior of forward passes is dependent on real-time inputs.
Errors from one forward pass can be corrected before the final output.	Errors from each forward pass is directly visible and accumulates.
Inference time is best-effort but quality is important.	Image quality is best-effort but inference time is strict.

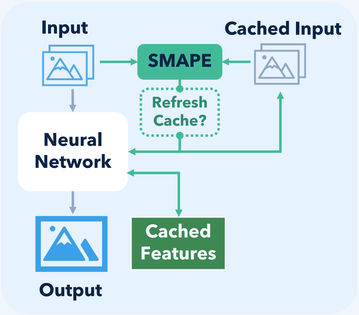
Overview of ReFrame

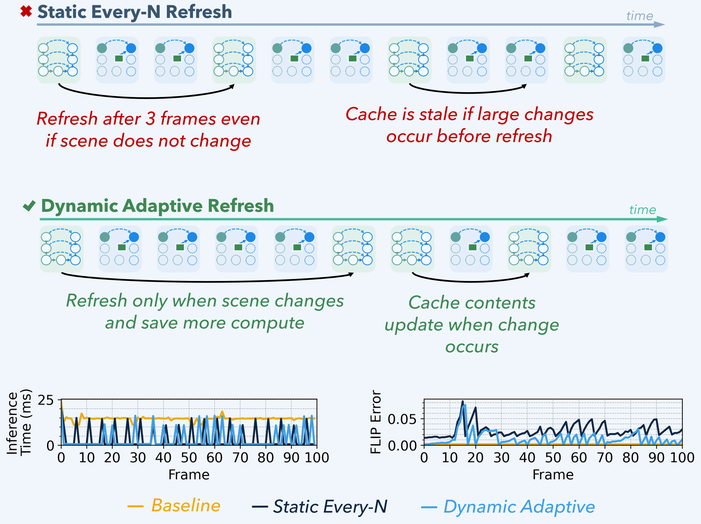
We extend the caching scheme introduced by **DeepCache** to rendering workloads:



Dynamic Refresh Policy

- There is no predetermined pattern in a real-time application
- Changes in the input is a good indicator for changes in the output
- Compute full inference and refresh cache when input changes significantly





Evaluation

We evaluate **ReFrame** on three real-time rendering workloads:

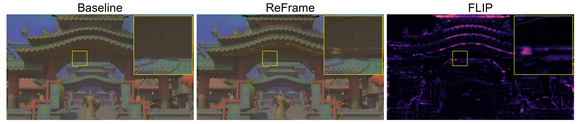
FE: Frame Extrapolation *ExtraNet*

SS: Supersampling *Fourier-Based Super Resolution*

IC: Image Composition *Implicit-Depth*

Results

13-50% of the frames in our workloads can take advantage of the cached features, which eliminates 6-29% of FLOPs in the encoder-decoder network, at a small cost to image quality.



Workload	Scene	Skipped Frames ↑	Eliminated Enc-Dec FLOPs ↑	Inference Speedup ↑	FLIP Image Quality Score ↓
FE	Sun Temple	50%	27%	1.42	0.0169
	Cyberpunk	30%	16%	1.10	0.0207
	Asian Village	35%	19%	1.24	0.0241
SS	Sun Temple	40%	29%	1.30	0.0490
IC	Garden Chair	13%	6%	1.05	0.0006

ReFrame achieves:

up to 1.05-1.85x inference speedup with negligible FLIP image error of 0.006-0.1 by extending existing caching techniques with dynamic policies