THE UNIVERSITY OF TEXAS AT DALLAS

Create Video Clips using Frame Interpolation

Group 12

Narendra Kanayalal Gangwani
Akshay Kumar Jha
Karneeshwar Sendilkumar Vijaya
Madhumita Ramesh

Problem Overview

- Creating a 360-degree view of an object using few images
- One common technique used - 3d model reconstruction - involves masking of the background, multiple camera angles, more computation
- Our approach is to capture original state of the object - without involvement of masking its environment
- Gives a realistic view of the object actually being captured by a video
- Solves the application of generating a video from few real-time images of the same object

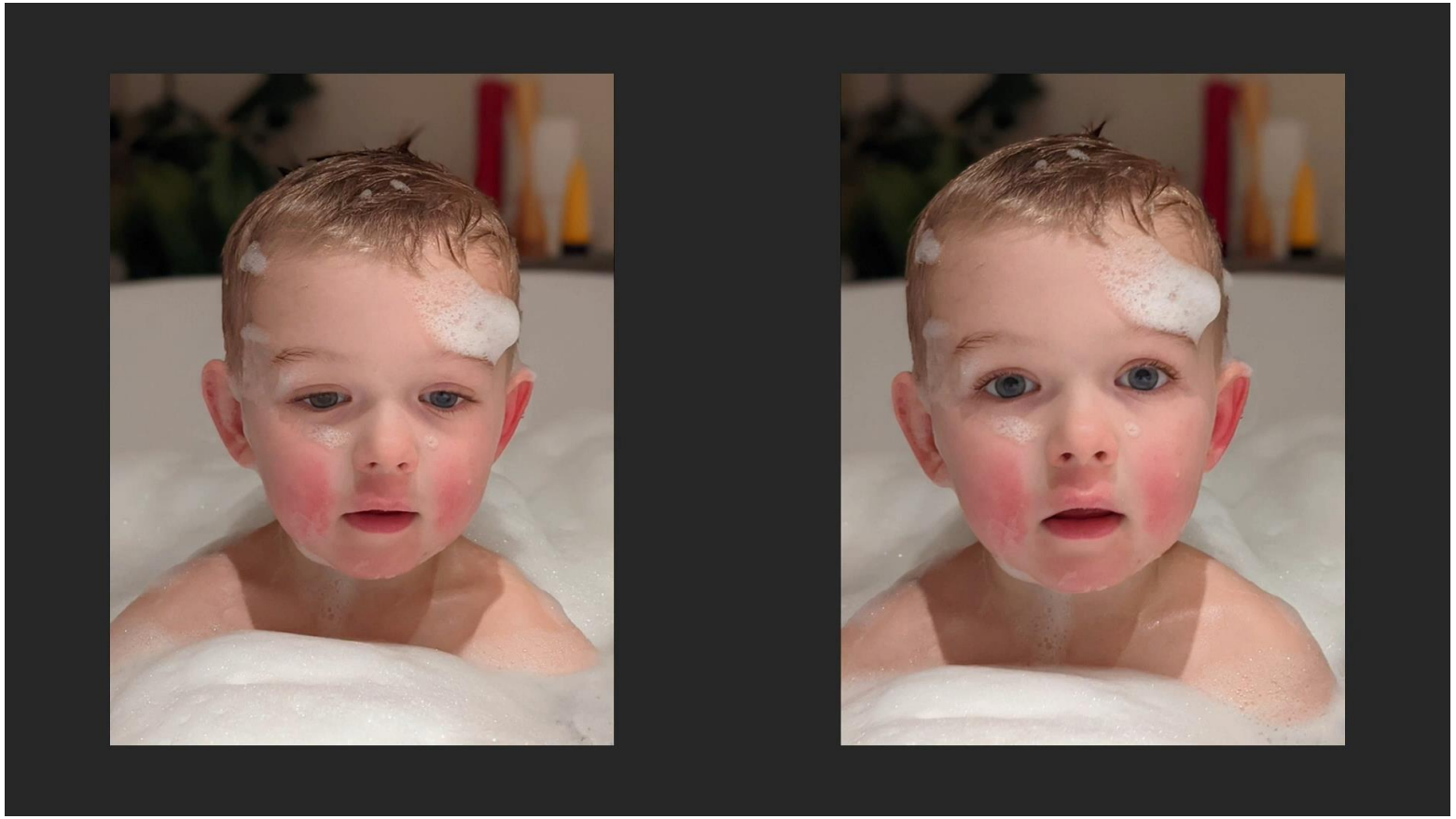
Applications

- Paparazzi clips
- Taj Mahal Monument capture
- Online Market Place

Algorithm used

- Using FILM - Large Motion Frame Interpolation algorithm over 3d model reconstruction
- Useful technique for improving quality of video content
- Useful for applications where background is required
- Can generate a realistic 360 view unlike other methods

Existing model and its working



Dataset

- The FILM is pretrained with Vimeo-90k dataset, a large-scale high-quality video dataset for lower level video processing.
- Contains over 90000 video URLs from Vimeo.
- From wide range of topics, including music, art, education, sports, and entertainment.
- It was created by researchers from the University of Amsterdam and Qualcomm AI Research using Vimeo's public API to collect the video URLs.
- The dataset is provided as a text file (.txt) with one URL per line.

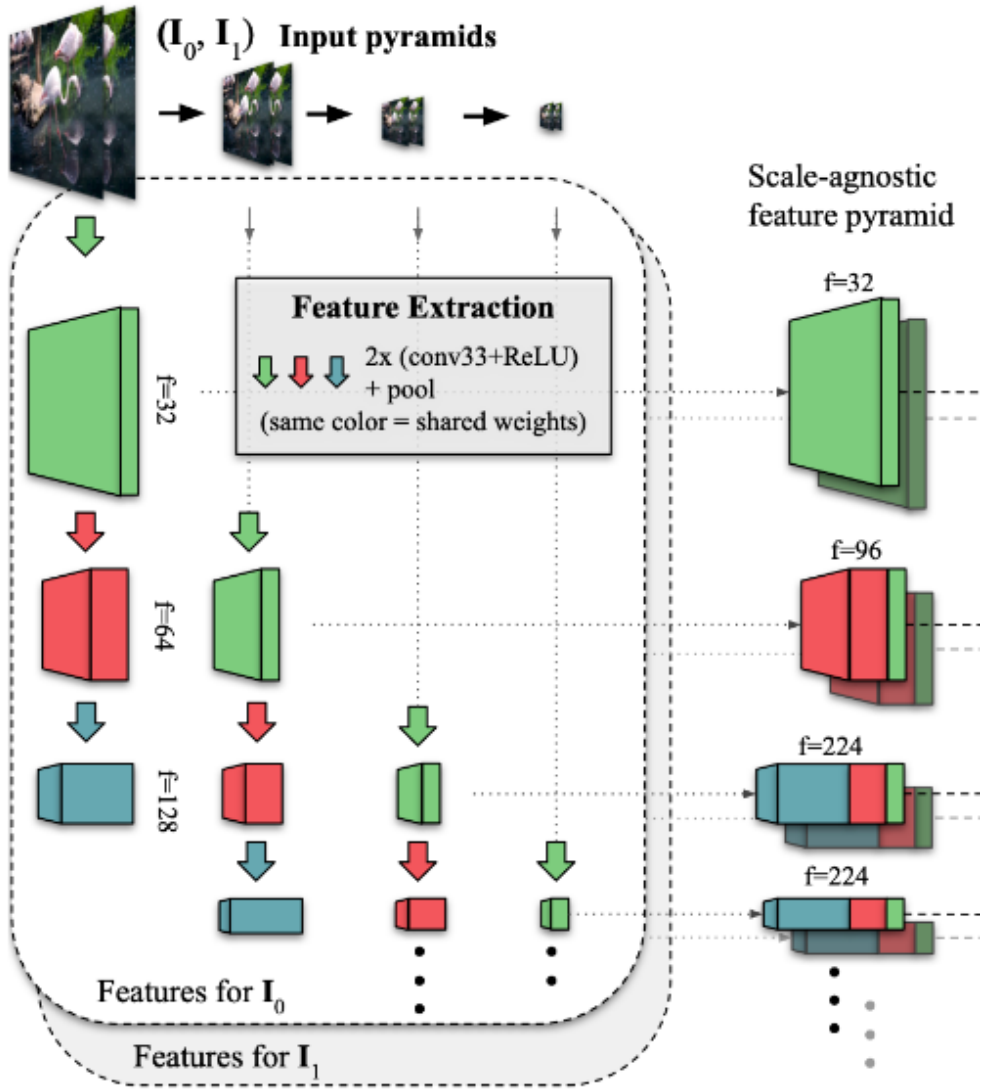
Architecture

- **SCALE AGNOSTIC FEATURE EXTRACTION**
- **BI-DIRECTIONAL FLOW ESTIMATION**
- **FUSION AND FRAME GENERATION**

SCALE AGNOSTIC FEATURE EXTRACTION

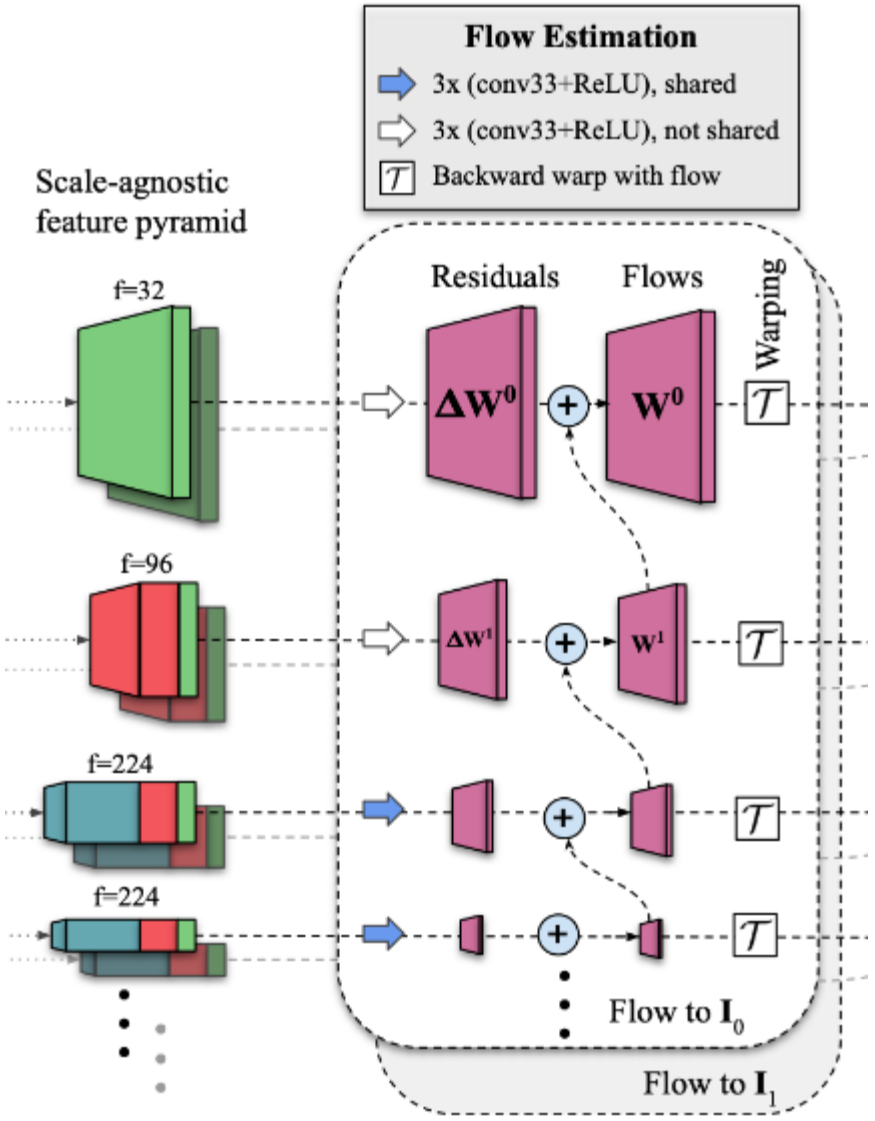
- Multi-resolution feature pyramids
- Shared - weights across scales
  - Large motion at shallow levels with small motion at deeper levels
  - Compact network with fewer weights

- Shared - UNet Convolutional Encoder
- Horizontally concatenating features with same spatial dimensions



BI-DIRECTIONAL FLOW ESTIMATION

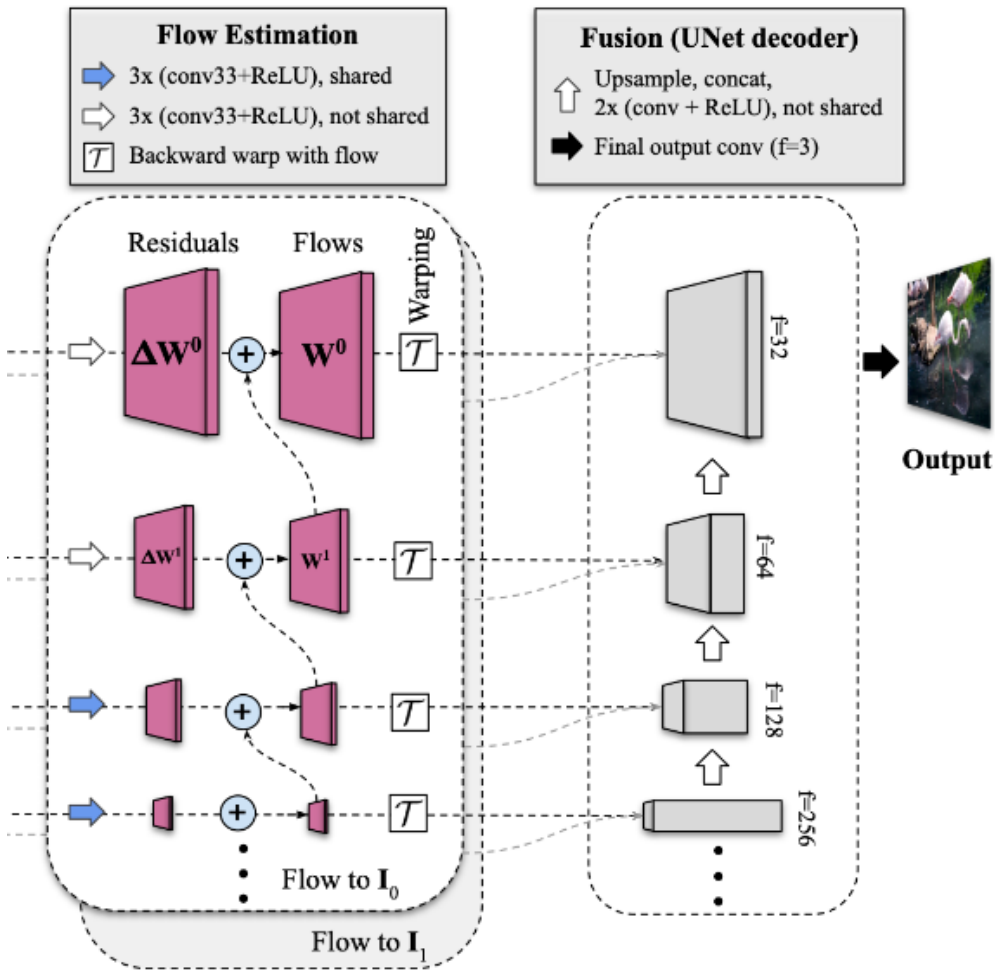
- Pyramid based residual flow estimation to compute flows from middle image
- Flow estimate is done once for each input starting at coarsest level to finer levels
- Adding residual correction to the upsampled estimate from the next deeper level



FUSION AND FRAME GENERATION

- We obtain a concatenated feature pyramid by stacking the 2 aligned feature maps, bi-directional flow and the input images.

- U-Net decoder synthesizes the interpolated output image.



LOSS FUNCTIONS

L1 Reconstruction: $$\mathcal{L}_1 = \|\hat{\mathbf{I}}_t - \mathbf{I}_t\|_1.$$

Perceptual Loss: $$\mathcal{L}_{\text{VGG}} = \frac{1}{L}\sum_{l=1}^{L} \alpha_l \left\| \Psi_l(\hat{\mathbf{I}}_t) - \Psi_l(\mathbf{I}_t) \right\|_1, \qquad \Psi_l(\mathbf{I}_i) \in \mathbb{R}^{\text{H}\times\text{W}\times\text{C}}$$

Style Loss: $$\mathcal{L}_{\text{Gram}} = \frac{1}{L}\sum_{l=1}^{L} \alpha_l \left\| \text{M}_l(\hat{\mathbf{I}}_t) - \text{M}_l(\mathbf{I}_t) \right\|_2, \quad \text{M}_l(\hat{\mathbf{I}}_t) = \left(\Psi_l(\hat{\mathbf{I}}_t)\right)^{\intercal}\left(\Psi_l(\hat{\mathbf{I}}_t)\right),$$

$$\mathcal{L}_S = w_l\mathcal{L}_1 + w_{\text{VGG}}\mathcal{L}_{\text{VGG}} + w_{\text{Gram}}\mathcal{L}_{\text{Gram}},$$

Experiments Conducted

- Experiments were conducted based on varying the following parameters.
  - Number of input images (15 or 30)
  - Interpolation parameter (3 or 5)

- Collected images of several test objects from different sides.

  - Type 1: 30 input images roughly at every 12 degs of the object
  - Type 2: 15 input images roughly at every 24 degs of the object

- Each type was tested with two interpolation parameters 3 and 5

- We ran these experiments on the free tier servers on Kaggle so we had to limit our experiments to a max of 5 interpolations between any two frames.

Testing Data Collection

This is a sample test image to show how we captured multiple images of an object along its sides to use as the input for the model to generate 360 degree video.

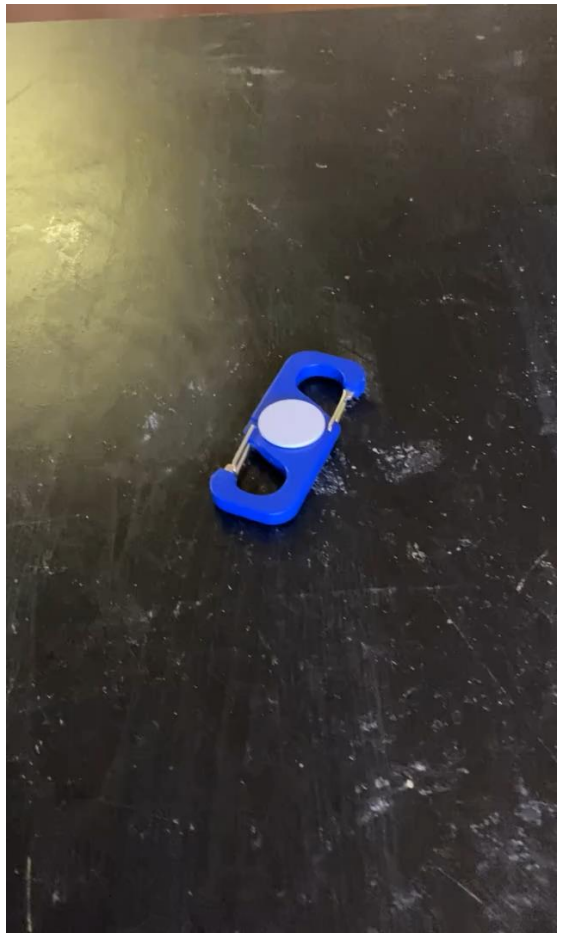


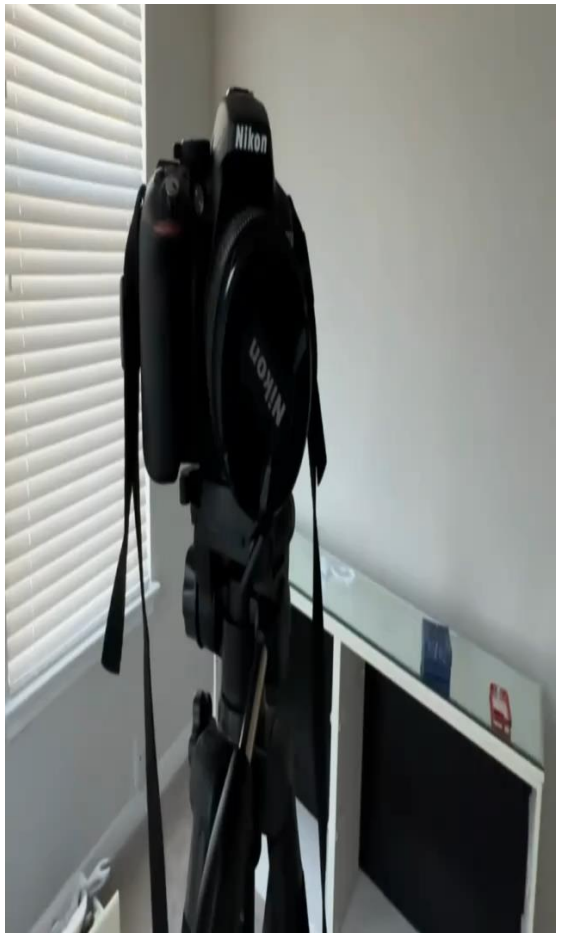

Object 1

Interpolation = 3

Interpolation = 5

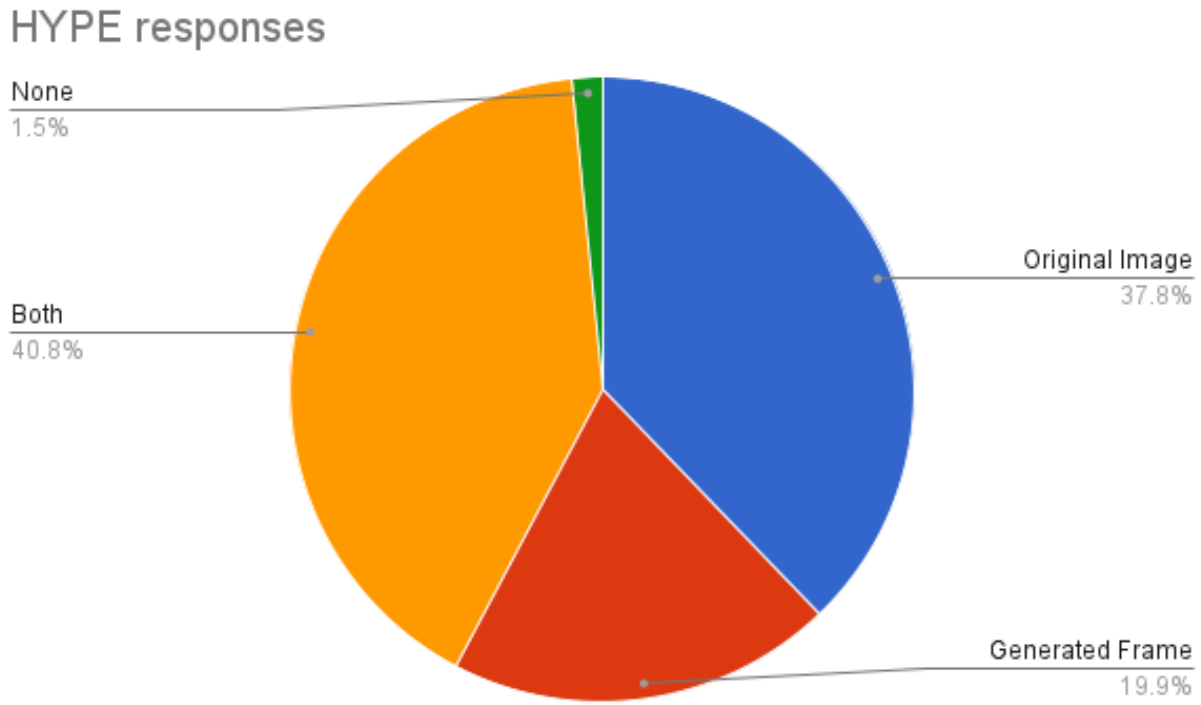
More Objects with Interpolation = 5



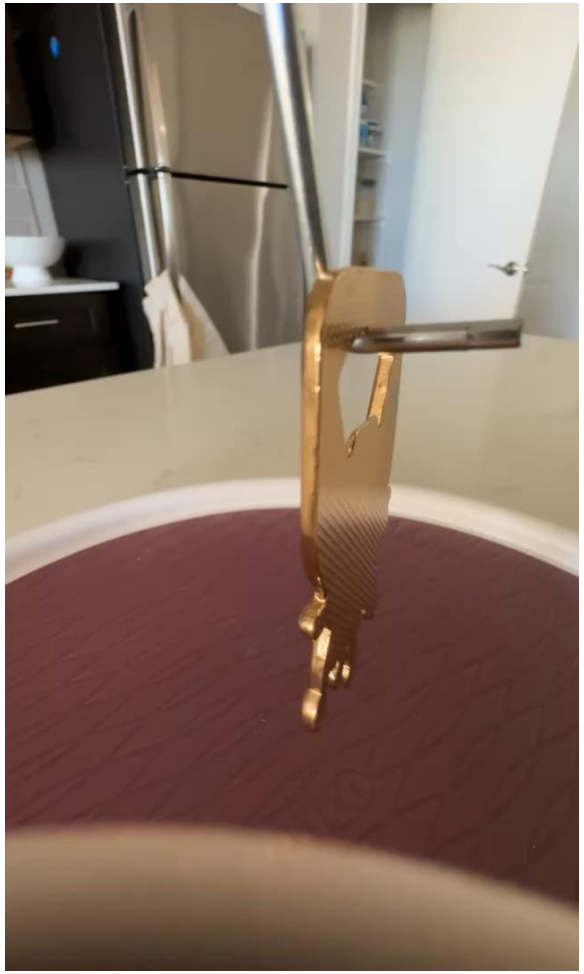


Validation

- HYPE (Human eYe Perceptual Evaluation), involves presenting human evaluators with pairs of images and asking them to choose the image that they believe is generated by a machine or realistic.



Testing challenges with complex scenarios



- **Lighting issues -** With 360 degree view one of the major issues was change in the light exposure between images because of which we were only able to get good results with higher interpolation
- **Computation power -** The computation power was a major challenge. Even with GPUs, we were not able to go over more than 2^5 interpolation images

Dataset Challenges

We used some unique innovative ways to create dataset. These are some of them.

Thank you