Final Project

CS 3570 Multimedia

Final Project Guideline

- Group Project: 3~4 people
    - Grouping Form
- Topic Candidate
    - Topic 1: Audio Super Resolution
    - Topic 2: Image Retrieval
    - Topic 3: Spatial and Temporal Video Super-resolution
    - Topic 4: Image Vectorization
    - Other (Topic Proposal)
- Fill out the group and topic forms: 5/6(Mon)
- One-page proposal: 5/13(Mon)
- Presentation: **6/7 (Fri)**
    - Details will be announced later

Topic 1: Audio Super Resolution

- overview:
  - audio super resolution is a task that reconstructs low resolution audio to high resolution audio.

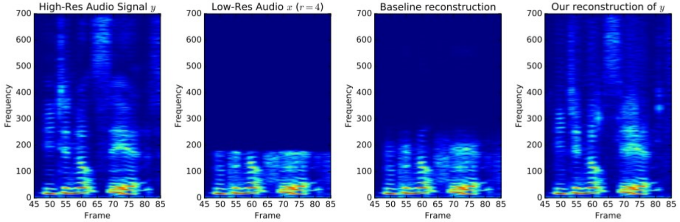


Figure 2: Audio super-resolution visualized using spectrograms. A high-quality speech signal (leftmost) is subsampled at $r = 4$, resulting in the loss of high frequencies (2nd from left). We recover the missing signal using a trained neural network (rightmost), greatly outperforming the cubic baseline (second from right).

- dataset: VCTK
  - 109 speaker who speaks 400 sentences.
  - synthetic low-resolution audio and high-resolution audio

- Goal:
  - Implement your own audio super-resolution algorithm or modify existence method
  - input : low-resolution audio
  - output : high-resolution audio

Topic 1: Audio Super Resolution

BaseLine Method:

1. Interpolation on temporal space
2. U-Net

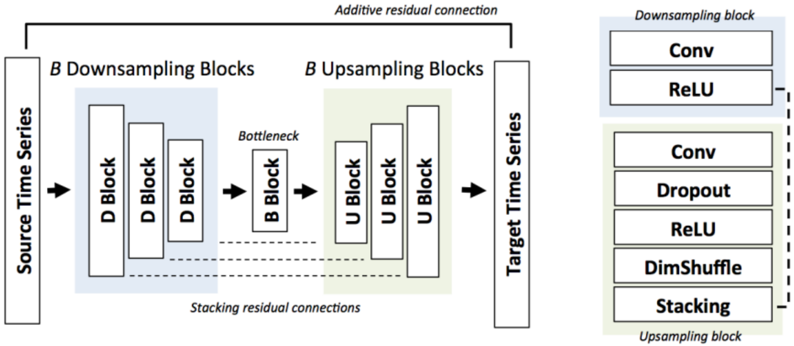


Figure 1: Deep residual network used for audio super-resolution. We extract features via B residual blocks; upscaling is done via stacked SubPixel layers.

Topic 1: Audio Super Resolution

$$D_{LS} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{P(\omega)}{\hat{P}(\omega)}\right]^2 d\omega},$$

- evaluate metric:
  - PSNR
  - LSD (**Log-Spectral Distance**)
- testing data:
  - VCTK dataset with downsample rate 4
- bonus:
  - VCTK dataset with different downsample rate and different filter
- rule :
  - you can’t use testing set to training
  - you can’t use pretrained model related to audio.
  - you can’t use extra dataset for training

Reconstruct result

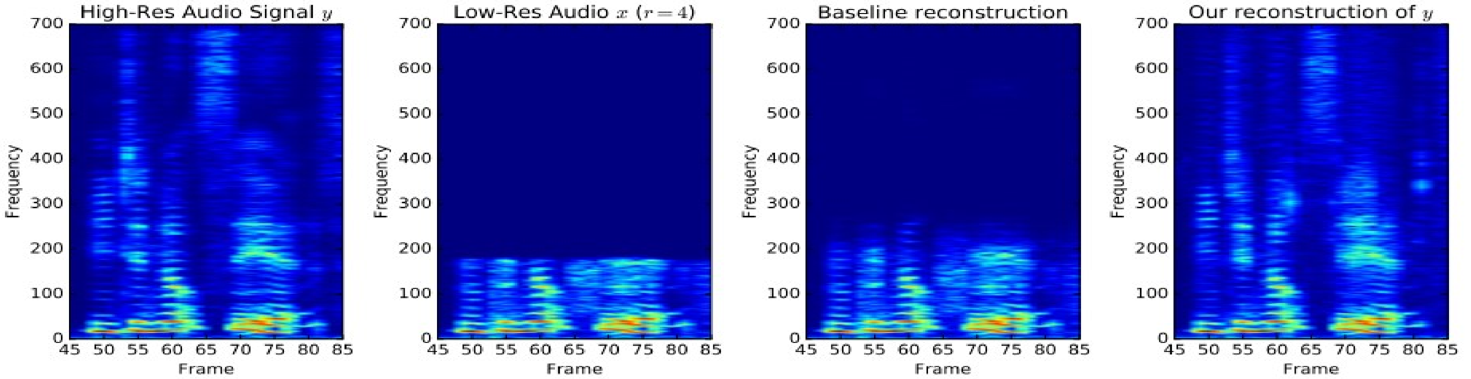
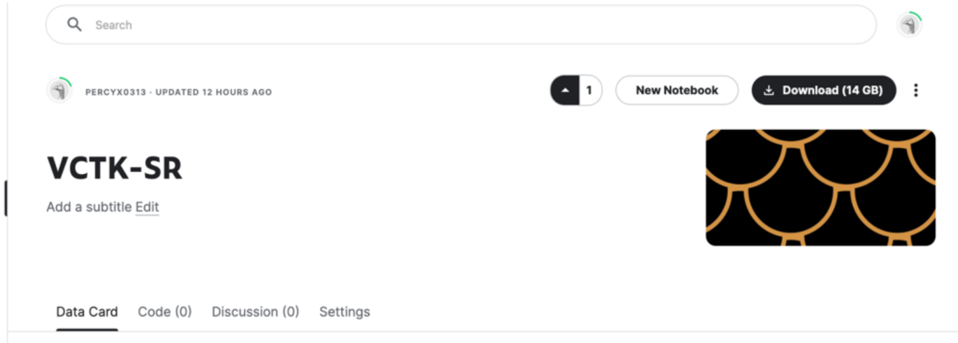


Figure 2: Audio super-resolution visualized using spectrograms. A high-quality speech signal (leftmost) is subsampled at $r = 4$, resulting in the loss of high frequencies (2nd from left). We recover the missing signal using a trained neural network (rightmost), greatly outperforming the cubic baseline (second from right).

kaggle dataset

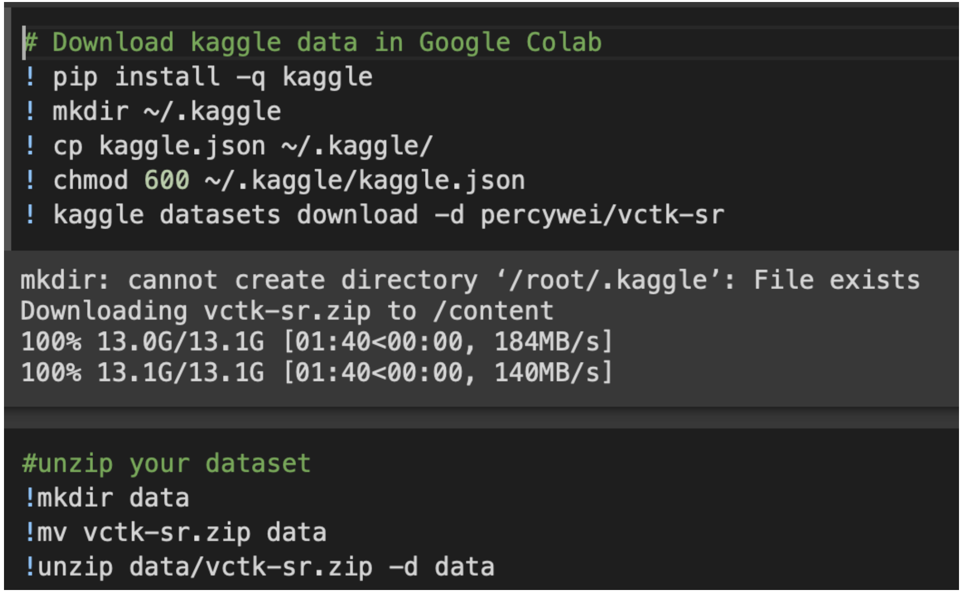
https://www.kaggle.com/datasets/percywei/vctk-sr



PERCYX0313 · UPDATED 12 HOURS AGO

VCTK-SR

Add a subtitle Edit



```
# Download kaggle data in Google Colab
! pip install -q kaggle
! mkdir ~/.kaggle
! cp kaggle.json ~/.kaggle/
! chmod 600 ~/.kaggle/kaggle.json
! kaggle datasets download -d percywei/vctk-sr
```

```
mkdir: cannot create directory ‘/root/.kaggle’: File exists
Downloading vctk-sr.zip to /content
100% 13.0G/13.1G [01:40<00:00, 184MB/s]
100% 13.1G/13.1G [01:40<00:00, 140MB/s]
```

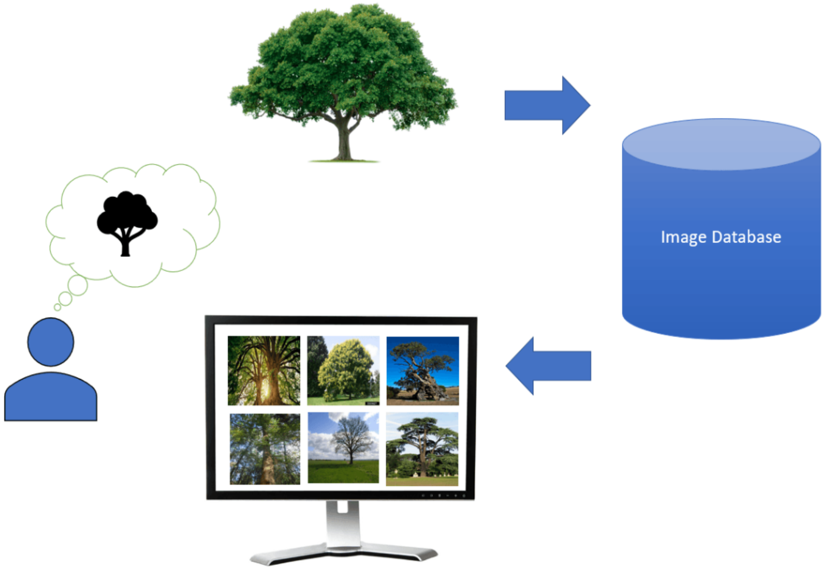
```
#unzip your dataset
!mkdir data
!mv vctk-sr.zip data
!unzip data/vctk-sr.zip -d data
```

Reference

1. audio interpolation
   https://www.alpha-ii.com/Info/AudioInt.html
2. U-Net:
   https://kuleshov.github.io/audio-super-res/
3. U-Net + AFiLM(attention)
   https://arxiv.org/pdf/2108.11637v1.pdf
4. you can design some filter to do data augmentation

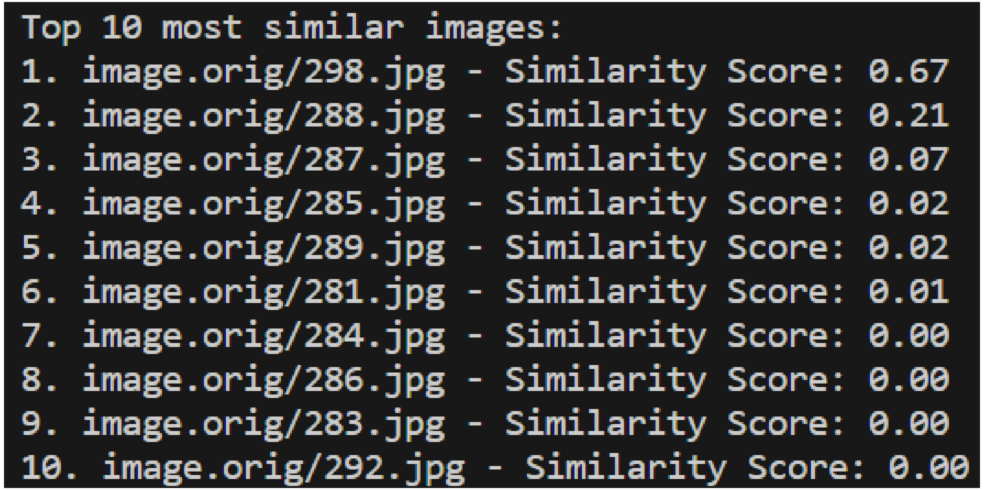
Topic 2: Image Retrieval

- Description: Image-to-image retrieval is a task that aims to find images in a large database that are similar to a given query image

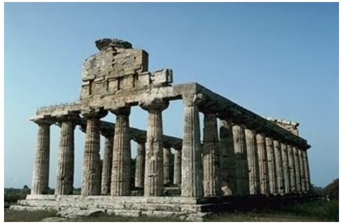


Topic 2: Image Retrieval

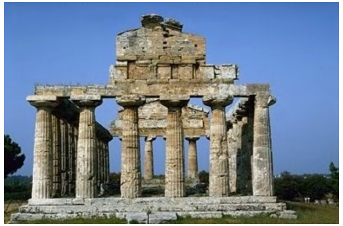
- CLIP



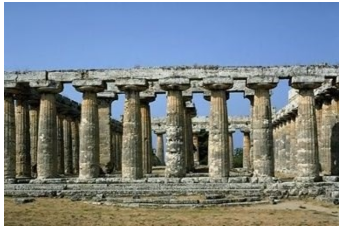
```
Top 10 most similar images:
1. image.orig/298.jpg - Similarity Score: 0.67
2. image.orig/288.jpg - Similarity Score: 0.21
3. image.orig/287.jpg - Similarity Score: 0.07
4. image.orig/285.jpg - Similarity Score: 0.02
5. image.orig/289.jpg - Similarity Score: 0.02
6. image.orig/281.jpg - Similarity Score: 0.01
7. image.orig/284.jpg - Similarity Score: 0.00
8. image.orig/286.jpg - Similarity Score: 0.00
9. image.orig/283.jpg - Similarity Score: 0.00
10. image.orig/292.jpg - Similarity Score: 0.00
```



query img: 299.jpg



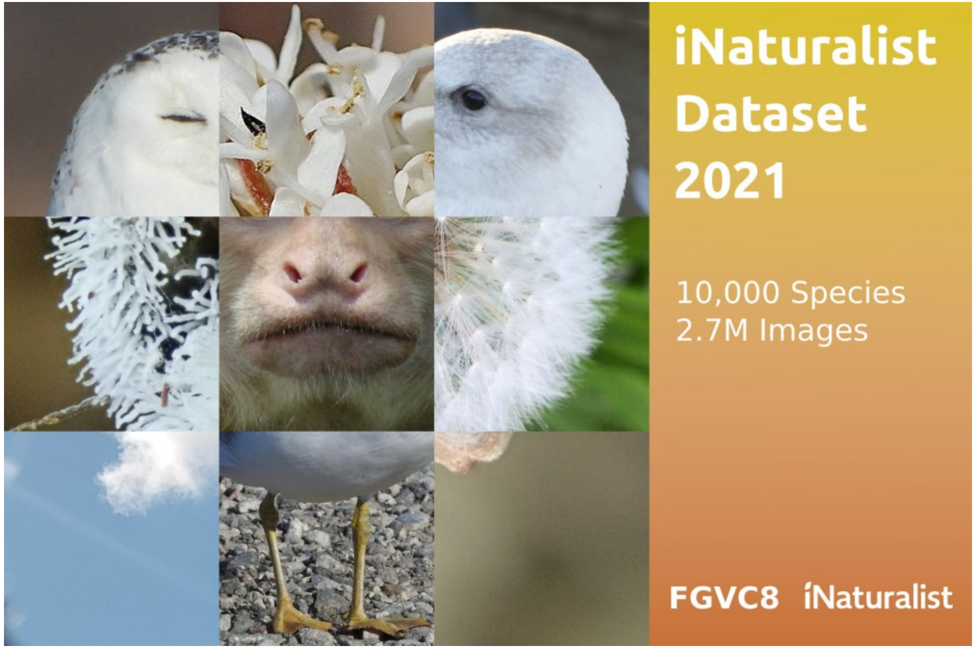
298.jpg



288.jpg

Topic 2: Image Retrieval

- iNaturalist 2021 Dataset
  - https://github.com/visipedia/inat_comp/tree/master/2021
  - The full training dataset contains nearly 2.7M images in 13 classes



Topic 2: Image Retrieval

- Metrics
  - Precision and Recall
    - Precision = A / B
      - A: # of relevant retrieved images
      - B: # of total retrieved images
    - Recall = A / C
      - C: # of relevant images in database
  - Speed
    - Measure the execution time
    - For fairness, TA will test your execution time on the same computer, but you are still encouraged to test it by yourselves

Topic 2: Image Retrieval

- Reference
    - A Comprehensive Analysis on Deep Learning based Image Retrieval
      https://ieeexplore.ieee.org/document/10200622
    - Learning Transferable Visual Models From Natural Language Supervision
      https://arxiv.org/abs/2103.00020
    - Awesome and classical image retrieval papers
      https://github.com/willard-yuan/awesome-cbir-papers
    - Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models
      https://arxiv.org/abs/2108.04024

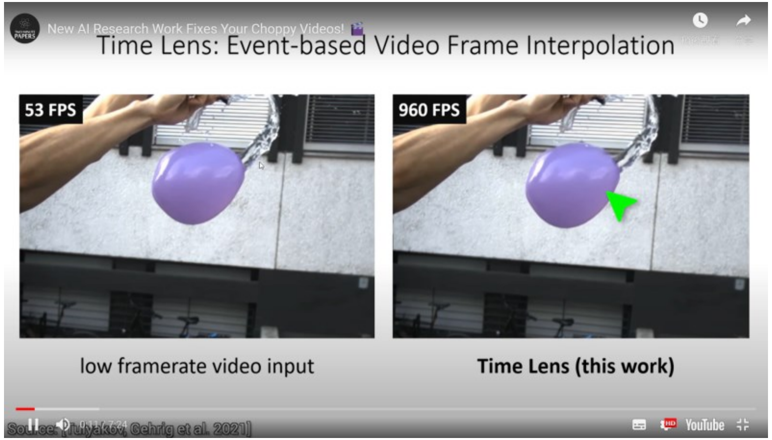
Topic 3: Spatial and Temporal Video Super-resolution

Video Interpolation (Temporal)

- Generate intermediate frames between two or more existing frames in a video sequence.
- Create smoother and more fluid video playback.

Image Interpolation (Spatial)

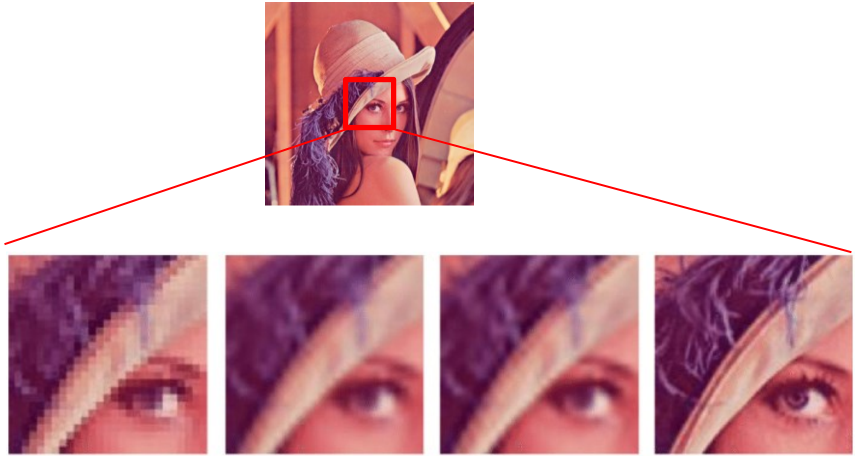
- Generate intermediate pixels between existing pixels in an image.
- Enhance the resolution and detail of images for clearer and more refined visual presentation.



New AI Research Work Fixes Your Choppy Videos!

Time Lens: Event-based Video Frame Interpolation

Source: [Tulyakov, Gehrig et al. 2021]



Topic 3: Spatial and Temporal Video Super-resolution

- You will be provided with half-sized frame 0 and half-sized frame 2
- Need to interpolate Frame 1 both temporally and spatially.
  The predicted frame 1 will be compared with the ground truth frame 1 using metrics such as PSNR and SSIM.

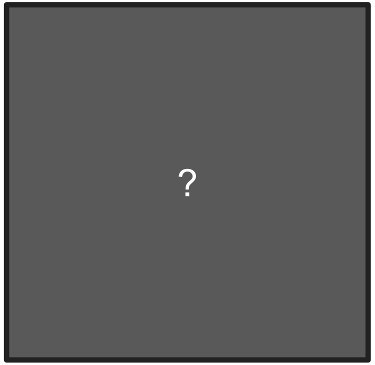
Frame 0

Frame 2

Frame 1 Ground truth

SSIM、PSNR

Your Prediction



Topic 3: Spatial and Temporal Video Super-resolution

Dataset:

The test dataset used for this project consists of two sets – public and private
public set：200 image pairs with ground truth data
private set：200 image pairs w/o ground truth data.

Rule:

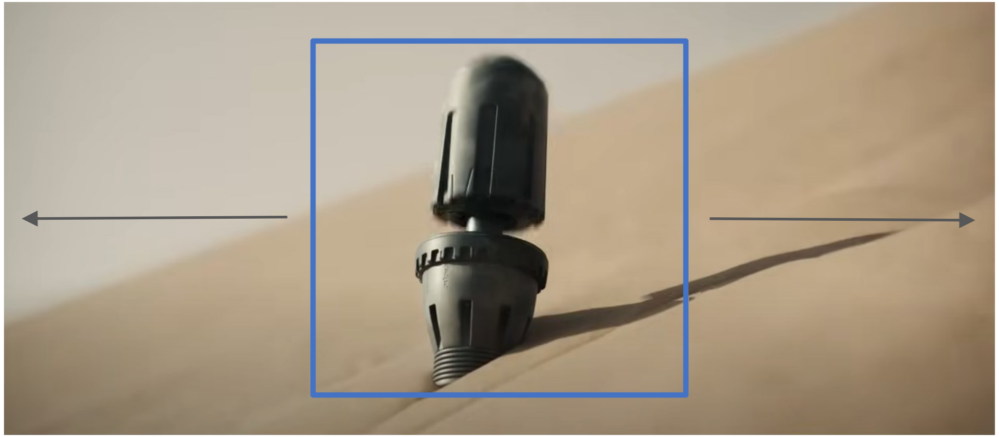
We can't use the pre-trained weights of the same task.
We would later specify a file structure for the evaluation of test sets.
You should not only pursue performance, but also try to provide some novelty and ingenuity.

Topic 3: Bonus: Image Extension

- Restore or reconstruct missing areas in images.
- Create complete and visually coherent images by intelligently filling outside the boundaries based on the surrounding pixel information
- No ground truth, provide visualize results and present your proposed methods in the final presentation.

Your Prediction

Image Extension

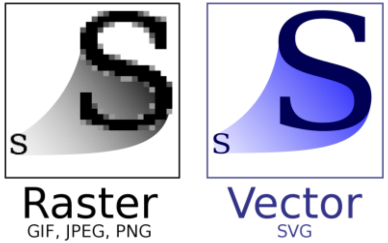


Topic 3: Spatial and Temporal Video Super-resolution

Reference:

- Provide an overview of the methods and techniques used in video interpolation in recent years. https://dl.acm.org/doi/pdf/10.1145/3556544
- Implement video interpolation using optical flow. https://learnopencv.com/optical-flow-in-opencv/
- Explore learning-based approaches. https://paperswithcode.com/task/video-frame-interpolation
- Image Interpolation https://www.researchgate.net/profile/Shreyas-Fadnavis/publication/301889708_Image_Interpolation_Techniques_in_Digital_Image_Processing_An_Overview/links/5abcee20a6fdcccda656f974/Image-Interpolation-Techniques-in-Digital-Image-Processing-An-Overview.pdf
- Image Super Resolution: A Comparison between Interpolation & Deep Learning-based Techniques to Improve Clarity of Low-Resolution Images https://medium.com/htx-s-s-coe/image-super-resolution-a-comparison-between-interpolation-deep-learning-based-techniques-to-25e7531ab207

Topic 4: Image Vectorization



- Overview

  Image vectorization is the process of converting a raster image consisting of pixels into a vector image consisting of lines, curves, and other geometric shapes. Such an image can be enlarged or reduced without loss of quality.

- Objective

  Implement your own image vectorization algorithm or modify the existing methods.

  - Input: raster image
  - Output: vector image
- Testing data:
  1. noto-emoji’s .png files
  - Containing 2458 images(.png)
  - 128*128 resolution

Topic 4: Image Vectorization

- Testing data:

  2. In the wild image: visualization data

  - 10 ~15 images for visaulization
- Evaluation metrics:
  - Quantitative evaluation: Calculate the average MSE loss between input raster image and rendered vector graphics of noto-emoji testing data
  - Visualization results: Show the results of vectorized image in visualization data
- Limitation
  - Do not train on the given testing data
  - Images can’t be vectorized manually

Reference Paper

- **Traditional algorithm**
  - https://wordsandbuttons.online/simple_image_vectorization.html
- **SAMVG (**ICASSP 2024**)**
  - https://arxiv.org/abs/2311.05276
- **Towards Layer-wise Image Vectorization (**CVPR 2022**)**
  - https://arxiv.org/abs/2206.04655
- **Differentiable Vector Graphics Rasterization for Editing and Learning (SIGRAPH 2020)**
  - https://people.csail.mit.edu/tzumao/diffvg/

Final Project Proposal

- One-page proposal includes:
  - Topic & Introduction
  - What is expected to be completed, e.g. implement what algorithm/model, system(application) design
  - References
- Due: **11:59pm, 5/13 (Mon)**
  - TA will discuss with you afterweard