Hand Shadow Puppet Recognition

Team Members

Syed Rifat Raiyan — **180041205**

Zibran Zarif Amio — **180041209**

CSE 4836

Pattern Recognition Lab

Introduction

Hand Shadow Puppetry

Definition:
Hand Shadow Puppetry, also known as *shadowgraphy* or *ombromanie* is the art of performing a story or show using images made by hand shadows. It can be called *"**cinema in silhouette**"*.

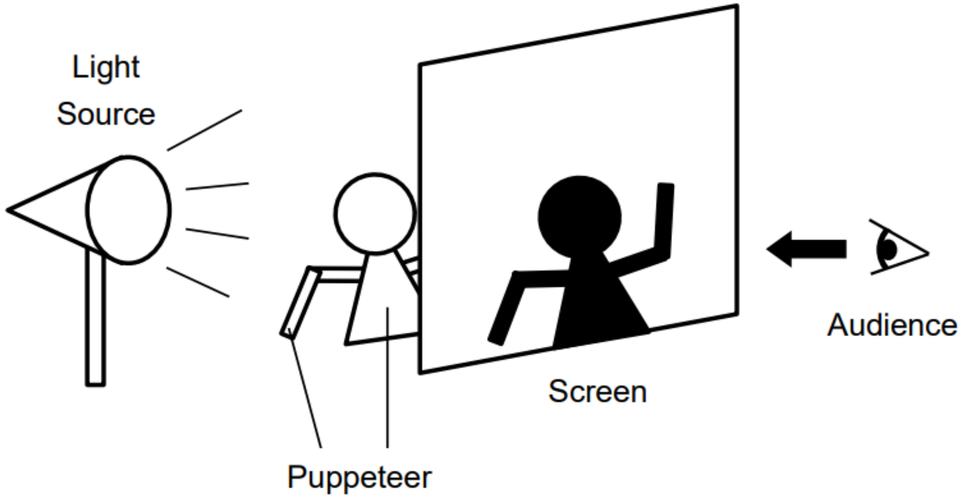

Rabbit


Bird


Dog

Goal:
Predicting the class labels of the animals from the silhouette shapes.

How does it work?
The puppeteer places his **hands between a light source and a translucent screen** to create shadows or silhouettes that resemble different **animals**.



Motivation

Our Inspiration to Pursue the Topic

- **Novelty factor** — To the best of our knowledge, **no** explicitly vision-related work or dataset exists on this topic of hand shadow puppet classification.
  - ➢ Some of the closely related works, however, will be mentioned in a bit…

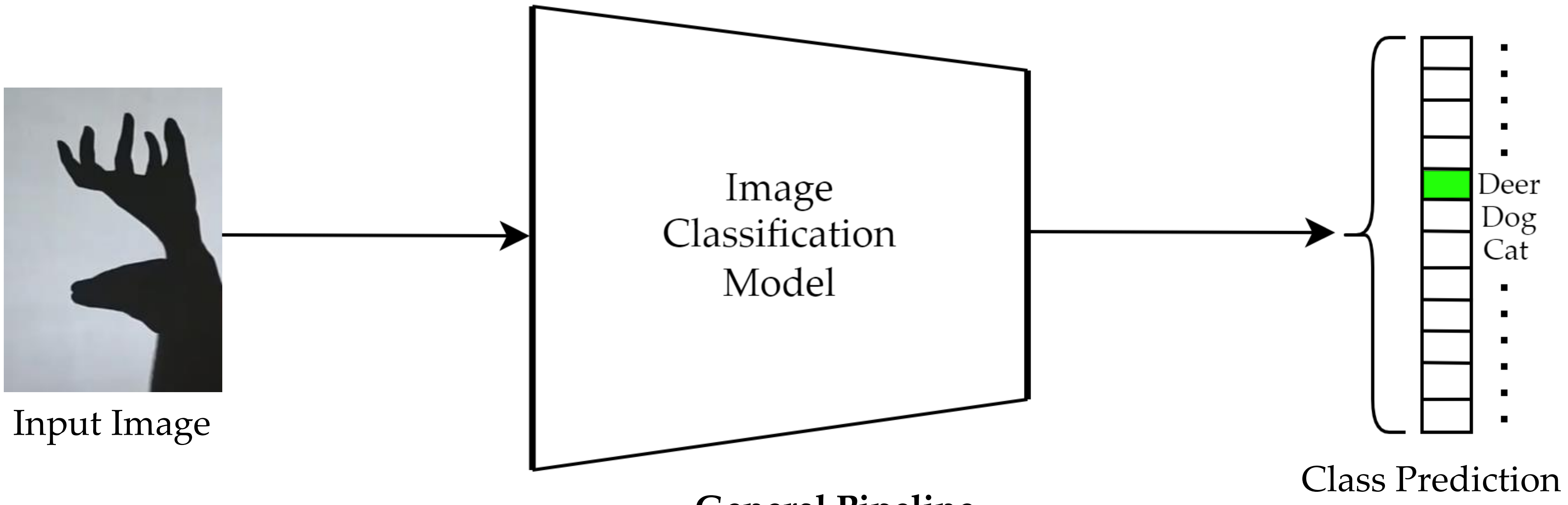
- **Utility**
  - ✓ **Tool for teaching** performance art
  - ✓ **Recreational app** for kids
  - ✓ Enabling the development of sophisticated algorithms for automatic **recognition**, **classification**, or even **generation** of ombromanie performances

- **Nostalgia** — incentivized by childhood memories during the load-shedding days.

Problem Formulation

The Precise Statement of the Problem

To classify 2D silhouette images of hand shadow puppets based on the animal being portrayed by the hands of the puppeteer.



General Pipeline

Related Work and Dataset

Research Literature on Shadow Puppetry

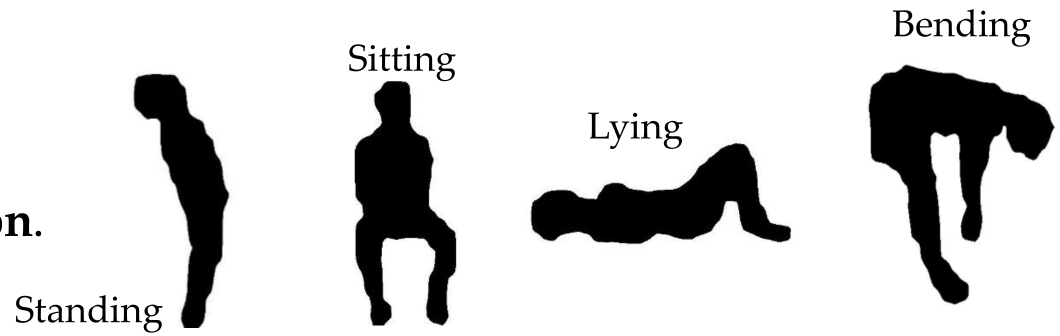
As mentioned before, we **haven't found any prominent work** on hand shadow puppet image classification.

Prominent Works: (closely related topics)

- In **Robotics,**
  - (Huang *et al.*)[1] — introduced a framework that enables **robotic arms** to **perform hand shadow puppetry** by matching shape correspondences of input image.
- In **Human-Computer Interaction,**
  - (Zhang *et al.*)[2] — worked on **emulating** the movements and body **gestures of a performer** on **Chinese shadow puppets** using Kinect sensor.
  - (Carr *et al.*)[3] — built a real-time **Indonesian shadow puppet** storytelling application using the Microsoft Kinect sensor capable of **mimicking full-body actions** of user.
- In **Computer Graphics,**
  - (Huang *et al.*)[4] — generated **3D models** of animals from shadow puppet images.

Dataset:

- Human Posture Silhouettes [5] — 4,800 binary images of silhouettes used for **human posture recognition**.



Our Work

What the project entails

- **Data Collection** — Gathered a total of **8,340 images** of hand shadow puppets.
  - ✓ From **45** professional hand shadow puppeteer clips and **22** amateur clips.
  - ✓ Across **11** classes.

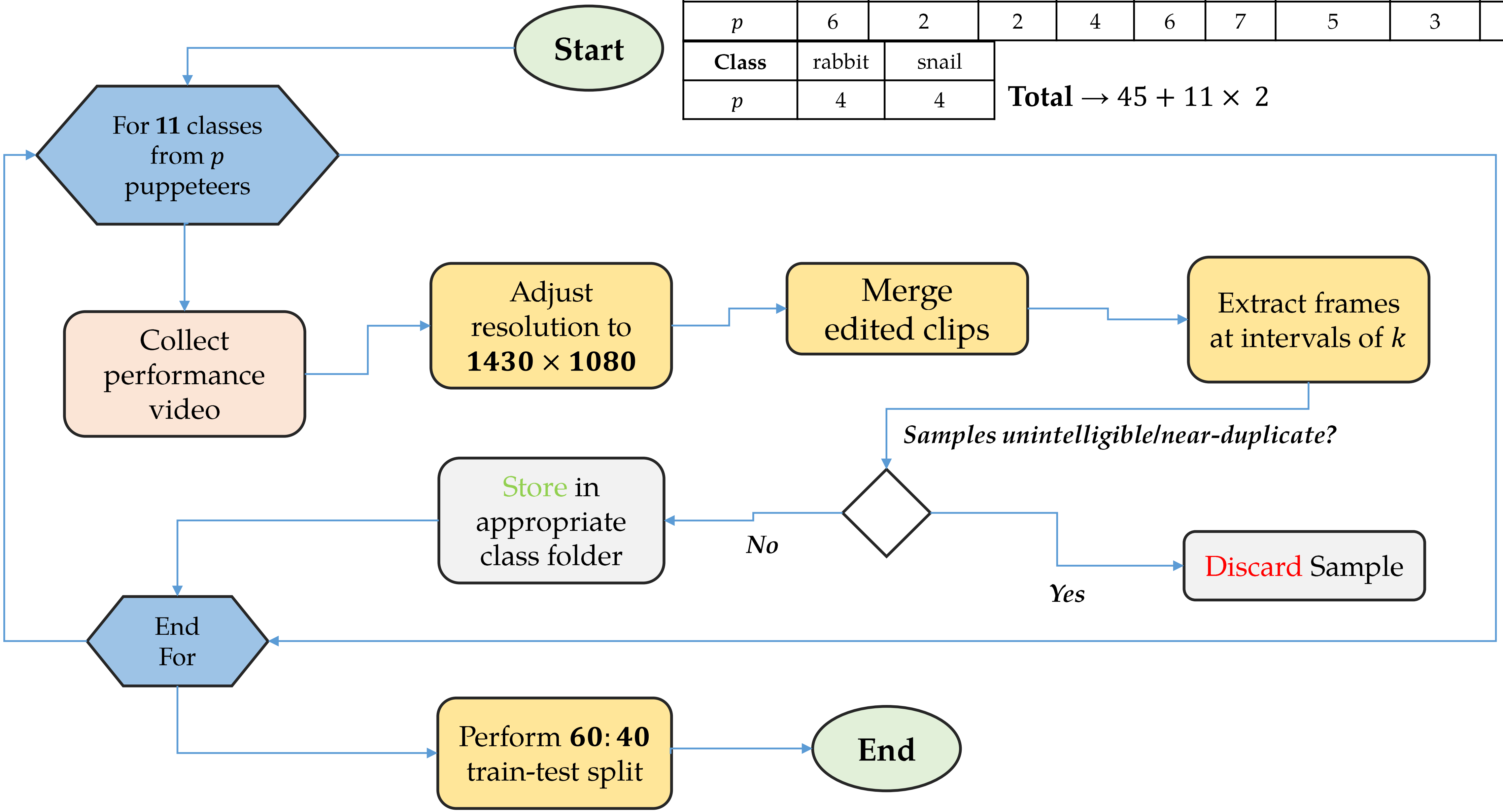
- **Benchmarking** — Established benchmarks for the dataset.
  - ✓ With **12** baseline pre-trained Pytorch image classification architectures.
  - ✓ Found **efficacy of convolutional models over transformer-based models** in silhouette classification.

Data Collection and Preprocessing

The workflow of our dataset preparation

Class	bird	chicken	cow	crab	deer	dog	elephant	moose	panther	rabbit	snail
p	6	2	2	4	6	7	5	3	2	4	4

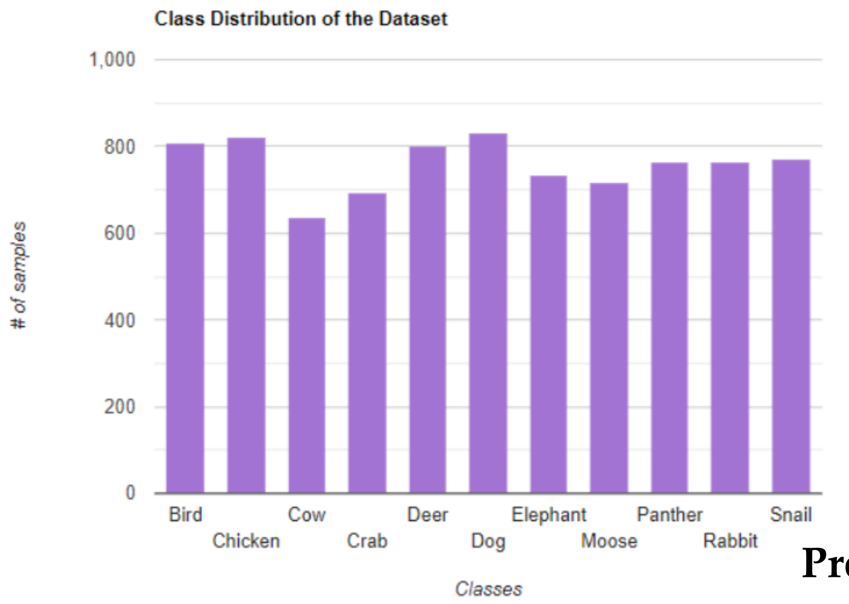
Total $\rightarrow 45 + 11 \times 2$



Dataset Statistics

General statistical analysis of the dataset

Class Label	# of samples
Bird	808
Chicken	820
Cow	635
Crab	695
Deer	802
Dog	831
Elephant	735
Moose	716
Panther	763
Rabbit	763
Snail	772



Examples of some sample images

- **Variations**
  - ✓ Different species of animals. (**Low intra-class similarity**)
  - ✓ Different orientations.
  - ✓ Different background colors.
  - ✓ **Inter-class similarity** between some classes.
  - ✓ Different hand structures (more **diversity**).

Moose

Deer

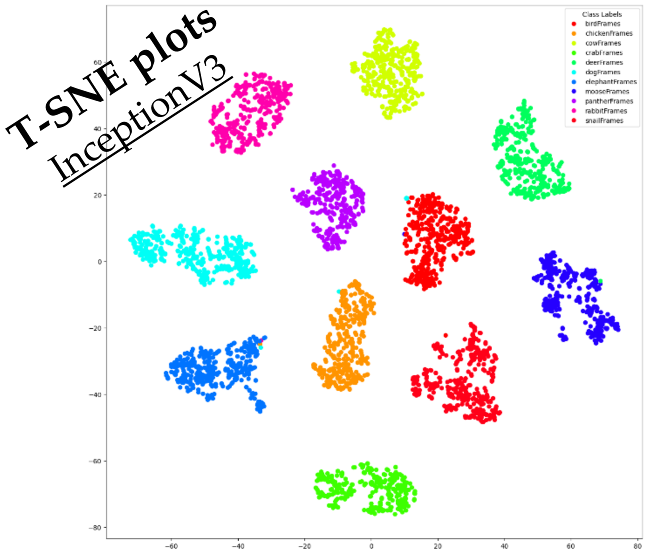
Pro:Novice ratio ≈ 80:20

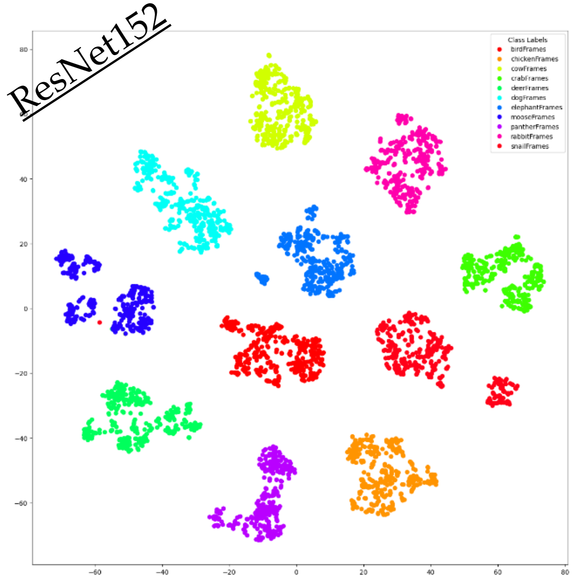
Results

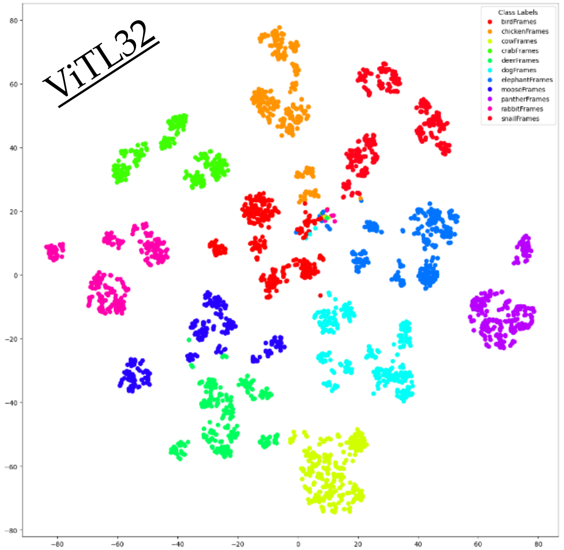
Tentative Benchmarking Results

Models	Val. Accuracy (%)		
	Top-1	Top-2	Top-3
ViTL32 (307M)	73.93	81.42	86.83
SwinV2B	83.06	88.25	92.85
ConvNeXt Large	83.35	88.11	92.10
GoogLeNet (23M)	83.38	88.54	93.04
MNasNet13	83.54	87.77	90.79
VGG16 (138M)	83.54	87.95	91.70
MaxViT	84.37	89.72	93.76
DenseNet201 (20.2M)	85.79	90.09	92.21
MobileNetV3Large (5.4M)	**86.11**	**90.25**	**93.52**
RegNetX32GF	86.45	91.46	93.34
ResNet152 (60.4M)	86.80	91.46	94.13
InceptionV3 (23.9M)	**88.27**	**93.65**	**95.42**

- **Key Takeaways**
  - ✓ **Inception V3 outperforms** every other model by ~2%.
  - ✓ Light-weight model **MobileNetV3Large** perform exceptionally **well**, despite having the lowest parameters (**mobile app!**).







- **InceptionV3** – tight, compact clustering
- **ResNet152** – comparatively less compact (2^{nd} best)
- **ViTL32** – too many overlaps and outliers (worst performance!)
- See our report for more…

Results

Attention Heatmaps

- **DenseNet201**'s dense block 4 (last layer) attention weights (using **GradCam**) –

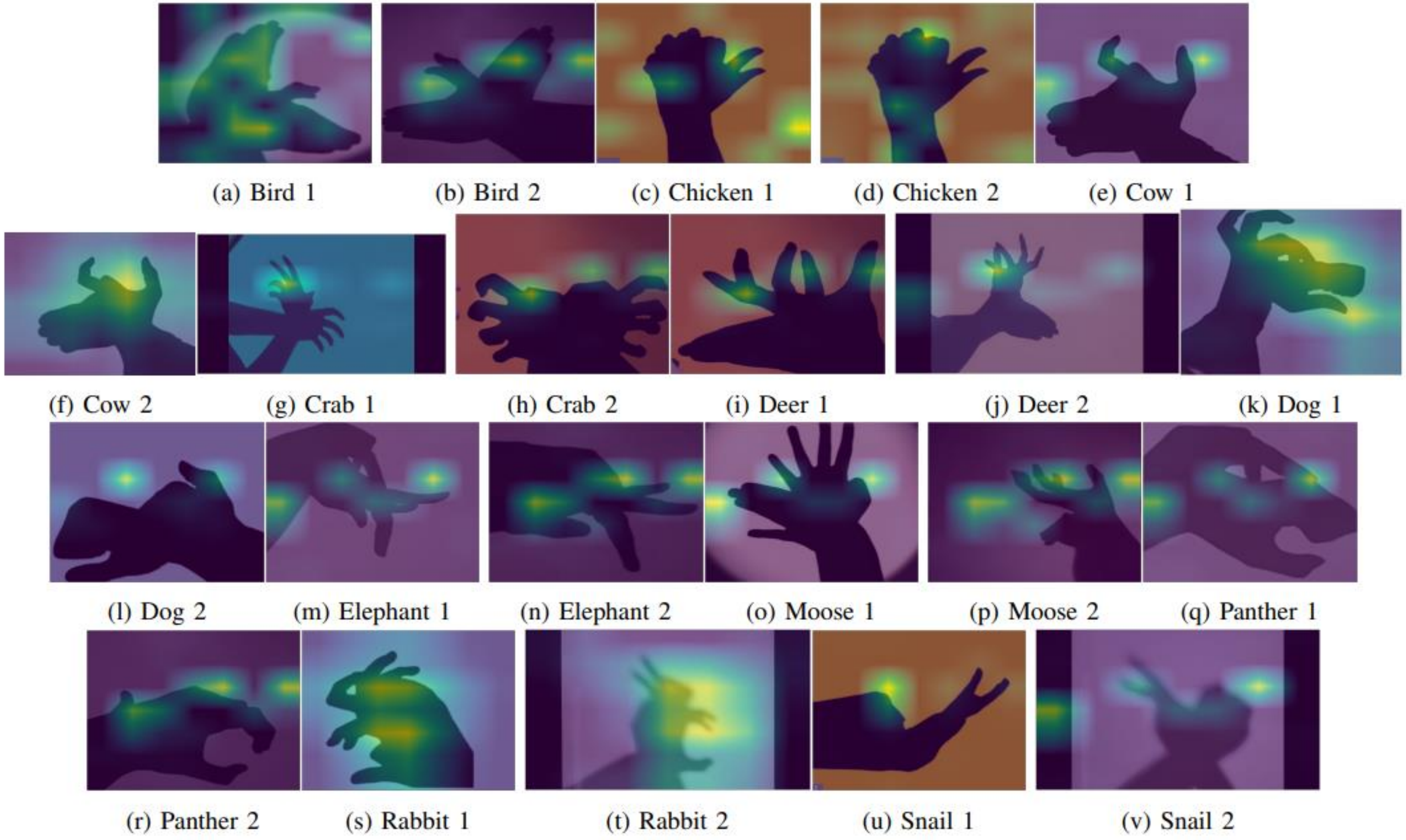


Fig. 6: Attention Heatmaps of 2 samples from each class.

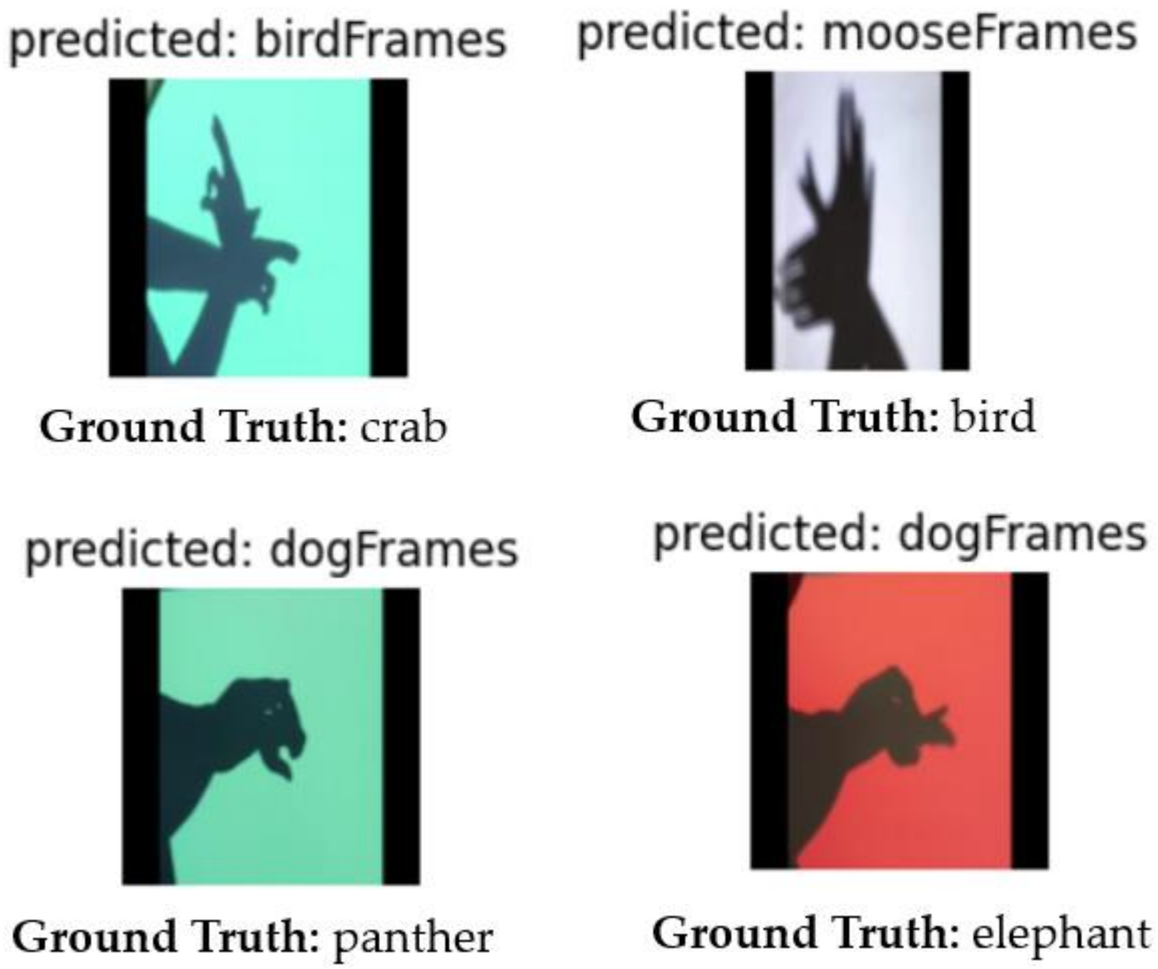
Common-sense distinguishing features

- ✓ **Bird** wingspan, beak
- ✓ **Chicken** comb
- ✓ **Cow** horn, concave head
- ✓ **Crab** appendages
- ✓ **Deer** horns
- ✓ **Dog** slanted head, ears
- ✓ **Elephant** tusks
- ✓ **Moose** upright horns
- ✓ **Panther** eyes and ears
- ✓ **Rabbit** small hands and mouth
- ✓ **Snail** shell, antennae

Results

Error Analysis

- Most of the misclassifications are among visually similar classes.



(a) Label: Crab, Predicted: Bird

(b) Label: Bird

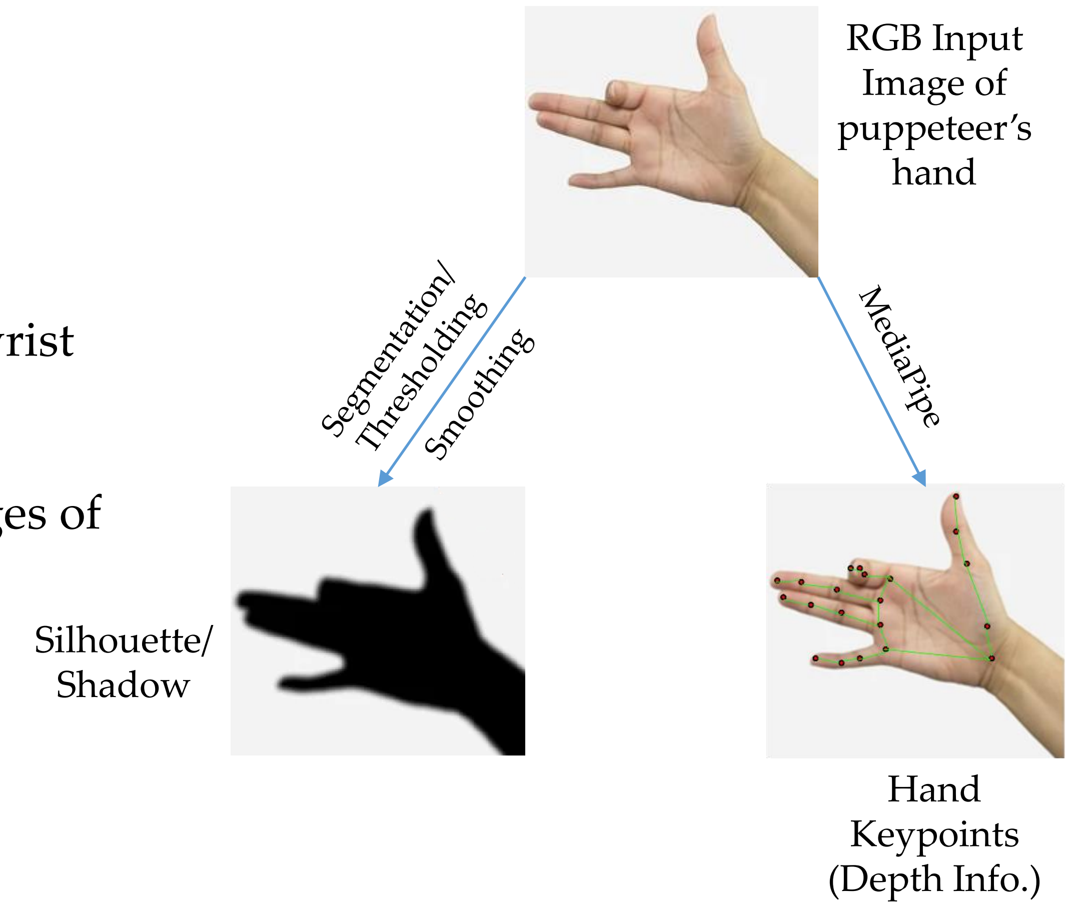
(c) Label: Deer, Predicted: Moose

(d) Label: Moose

Limitations

Scopes of improvement

- Our work still has some ground to build upon,
  - ✓ Introducing samples with **more diversified** hand/palm/wrist structures.
  - ✓ Exploring **two different approaches** to classify RGB images of the hand
    - ➢ Feature Extraction after **RGB to grayscale silhouette conversion** (using pre-processing DIP techniques).
    - ➢ Utilizing **depth information** and coordinates of **hand landmarks** as features (using MediaPipe).
      - ✓ Yields **high accuracy in Sign Language Recognition** tasks, as per recently published research works.



Conclusion and Future Work

Summary of our contributions

- We introduce the **first documented dataset** on Ombromanie/Hand Shadow Puppet images for image classification.
- We provide a **detailed statistical analysis** of our dataset.
- We employ a range of **12 image classification models** as baselines to perform **benchmarking** for the dataset.
- We provide **visualizations** of the models' **feature space** and **attention** tendencies.
- Our **findings** are,
  - ✓ Convolutional models outperform transformer-based models in silhouette classification.
  - ✓ May be due to insufficient training samples.

References

Works Cited in This Presentation

[1] Z. Huang, V. K. Madaram, S. Albadrani, and T. V. Nguyen, "Shadow puppetry with robotic arms," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1251–1252, 2017.
[2] H. Zhang, Y. Song, Z. Chen, J. Cai, and K. Lu, "Chinese shadow puppetry with an interactive interface using the kinect sensor," in *Computer Vision–ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pp. 352–361, Springer, 2012.
[3] B. M. Carr and G. J. Brown, "Shadow puppetry using the kinect," 2014.
[4] M. Huang, S. Mehrotra, and F. Sparacino, "Shadow vision," 1999.
[5] https://www.kaggle.com/datasets/deepshah16/silhouettes-of-human-posture

THANK YOU FOR LISTENING.