HASPER: An Image Repository for Hand Shadow Puppet Recognition

Syed Rifat Raiyan, Zibran Zarif Amio, and Sabbir Ahmed, *Member, IEEE*

***Abstract*—Hand shadow puppetry, often referred to as *shadowgraphy* or *ombromanie*, is a form of theatrical art and storytelling where hand shadows are projected onto flat surfaces to create illusions of living creatures. The skilled performers create these silhouettes by hand positioning, finger movements, and dexterous gestures to resemble shadows of animals and objects. Due to the lack of practitioners and a seismic shift in people's standards of entertainment, this art form is on the verge of extinction. To facilitate its preservation and proliferate it to a wider audience by harnessing the power of machine learning, in this paper, we introduce HASPER, a dataset consisting of $8,340$ images of hand shadow puppets across 11 classes extracted from both professional and amateur hand shadow puppeteer clips. We provide a detailed statistical analysis of the dataset and employ a range of pretrained image classification models to establish baselines. For this task, our findings show a substantial performance superiority of traditional convolutional models over attention-based transformer architectures. Keeping the best-performing model INCEPTIONV3 under the limelight, we conduct comprehensive feature-spatial, explainability, and error analyses to figure out the lapses of the models. To the best of our knowledge, this is the first documented dataset and research endeavor to preserve the dying art of ombromanie for future generations with Computer Vision approaches. Our code and data are publicly available at https://github.com/Starscream-11813/HaSPeR.**

***Impact Statement*—The impact statement should not exceed 150 words. This section offers an example that is expanded to have only and just 150 words to demonstrate the point. Here is an example on how to write an appropriate impact statement: Chatbots are a popular technology in online interaction. They reduce the load on human support teams and offer continuous 24-7 support to customers. However, recent usability research has demonstrated that 30% of customers are unhappy with current chatbots due to their poor conversational capabilities and inability to emotionally engage customers. The natural language algorithms we introduce in this paper overcame these limitations. With a significant increase in user satisfaction to 92% after adopting our algorithms, the technology is ready to support users in a wide variety of applications including government front shops, automatic tellers, and the gaming industry.**

***Index Terms*—HaSPeR, Hand shadow puppetry, puppet, shadow, silhouette, Ombromanie, Shadowgraphy, Image classification, benchmark, dataset, Deep learning, Transfer learning, Computer vision, Mobile application**

I. INTRODUCTION

Ombromanie, the ancient art of hand shadow puppetry, is a form of art that involves the mesmerizing interplay of light and shadow through the construction and manipulation of shadow figures or silhouettes on a surface, typically a screen or a wall, using one's hands, body, or prop objects [1]. The alias *"cinema in silhouette"*[1] is sometimes used to refer to this proto-cinematic medium of entertainment. Its working principle is very simple—the puppeteer adeptly positions their hands between a radiant light source and a translucent screen, consequently conjuring shadows and silhouettes that emulate an array of diverse creatures [2] from the animal kingdom (as shown in Figure 1). Despite its rich history and captivating allure across a diverse range of cultures[2], there exists a notable dearth of resources specifically tailored to this artistic domain. With properly annotated and sourced data, researchers could delve into the intricacies of hand silhouette movements, shapes, and storytelling techniques, thereby enabling the development of sophisticated Artificial Intelligence (AI) systems for automatic recognition, classification, or even generation of ombromanie performances [3]. The generation aspect is particularly relevant given the demonstrable impotency of AI image generator models in accurately creating hands and fingers [4]. Apart from that, the development of applications that can facilitate the learning of ombromanie has the potential to breathe new life into this waning art form [5]. In 2011, UNESCO recognized shadow puppetry as an endangered artistic tradition by adding it to the Intangible Cultural Heritage list [6], which is why it necessitates more preservatory apparatus and research efforts.

In tandem with this motivation, this work introduces a seminal addition to the realm of data resources, HASPER (**Ha**nd **S**hadow **P**upp**e**t Image **R**epository), a methodically curated novel image dataset of ombromanie performances or hand shadow puppets. The dataset comprises an assemblage of $8,340$ samples, that we source and verify from 45 professional puppeteer clips and 22 amateur clips. We label and categorize

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

Syed Rifat Raiyan is affiliated with the Systems and Software Lab (SSL) of the Department of Computer Science and Engineering at the Islamic University of Technology, Boardbazar, Gazipur-1704, Dhaka, Bangladesh (e-mail: rifatraiyan@iut-dhaka.edu).

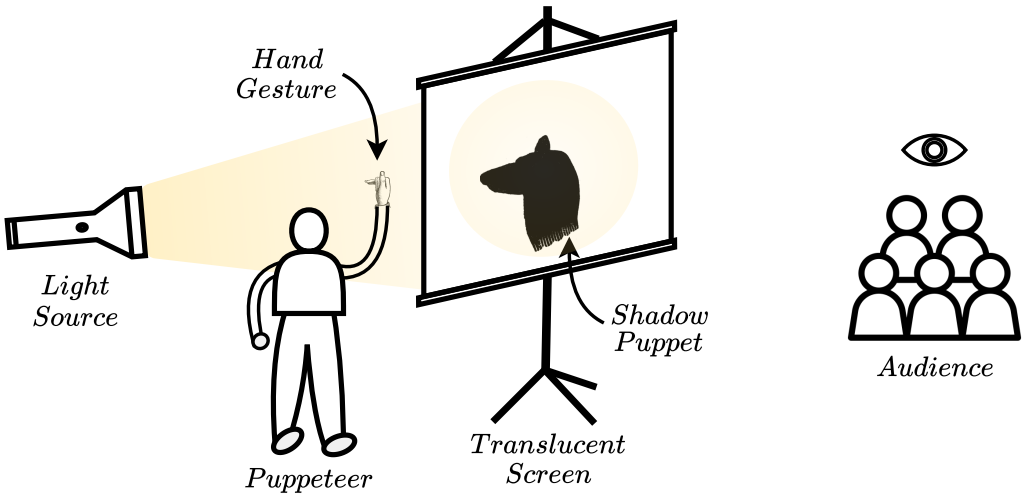
Zibran Zarif Amio, was affiliated with the Networking Research Group of the Department of Computer Science and Engineering at the Islamic University of Technology, Boardbazar, Gazipur-1704, Dhaka, Bangladesh. He is now with an AI-based software company, ReplyMind AI Ltd., Mirpur, Dhaka-1216, Bangladesh. (e-mail: zibranzarif@iut-dhaka.edu).

Sabbir Ahmed is affiliated with the Computer Vision Lab (CVLab) of the Department of Computer Science and Engineering at the Islamic University of Technology, Boardbazar, Gazipur-1704, Dhaka, Bangladesh (e-mail: sabbirahmed@iut-dhaka.edu).

This paragraph will include the Associate Editor who handled your paper.

[1] Shadowgraphy (performing art), Wikipedia — https://en.wikipedia.org/wiki/Shadowgraphy_(performing_art)

[2] Shadowgraphy, Magicpedia — https://www.geniimagazine.com/wiki/index.php/Shadowgraphy



(a) A generic hand shadow puppetry setup.

(b) Rabbit (c) Bird (d) Dog

Fig. 1: Ombromanie in a nutshell[3].

the images with utmost precision to ensure the elicitation of robustness in the image classification models that will undergo training with these images. The image samples portray the diversity of appearance since the source clips were recorded in a plethora of different poses, orientations, and background lighting conditions of the translucent screen. We posit that our dataset possesses the potential to offer a wealth of opportunities for exploration and analysis into the artistic domain of ombromanie. We conduct a detailed analysis of HASPER's statistical characteristics. We also employ a variety of state-of-the-art (SOTA) pretrained image classification models to establish a performance benchmark for validating the integrity of the dataset. Additionally, we conduct a thorough evaluation of several facets of the ace INCEPTIONV3 model, including its feature representations, interpretability, explainability, and classification errors that it encounters.

II. LITERATURE REVIEW

The recognition and classification of hand shadow puppet images are intriguing problemspaces, albeit relatively underexplored, since we could not find prior research works under the umbrella of this domain. After performing a rigorous analysis of existing literature on the topic, we could identify several somewhat related works.

A. Image Classification and Recognition

Among the pioneering endeavors in hand shadow image classification was that of Huang et al. [7], who created SHADOW VISION—a system to emulate an immersive virtual shadow puppet theater experience, employing a user's hand gestures over an overhead projector to control the creation and manipulation of objects within a 3D Open Inventor[4] environment. The chain of stages underlying the implementation of SHADOW VISION were acquisition, segmentation, feature extraction, and recognition of the shadow puppet images. These images were captured using an infrared camera during the acquisition stage, and subsequently thresholded to obtain the pixels representing the hands of the puppeteer in the segmentation stage. The authors extracted salient features from the curated images using a modified chain code algorithm [8] akin to the CG scan algorithm. The recognition stage entailed the application of the centralized contour moments modeling technique of Mertzios and Tsirikolias [9]. Huang et al. [7] adopted a 3-layer neural network for this purpose as well, using 13 features (7 moments of the object, length, angle, and the 4 endpoints of the axis of inertia). The data used for this study isn't publicly available, and the methodology can be deemed somewhat obsolete in the modern purview, given the emergence of deep learning models in the computer vision domain. Some recent works explore different convolutional models to assess their efficacy in classification and

[3]The shadowgraphy cliparts are adapted from ClipArt ETC, Florida Center for Instructional Technology, College of Education, University of South Florida. Link: https://etc.usf.edu/clipart/galleries/266-hand-shadow-puppetry

[4]Open Inventor™ toolkit — https://www.openinventor.com/

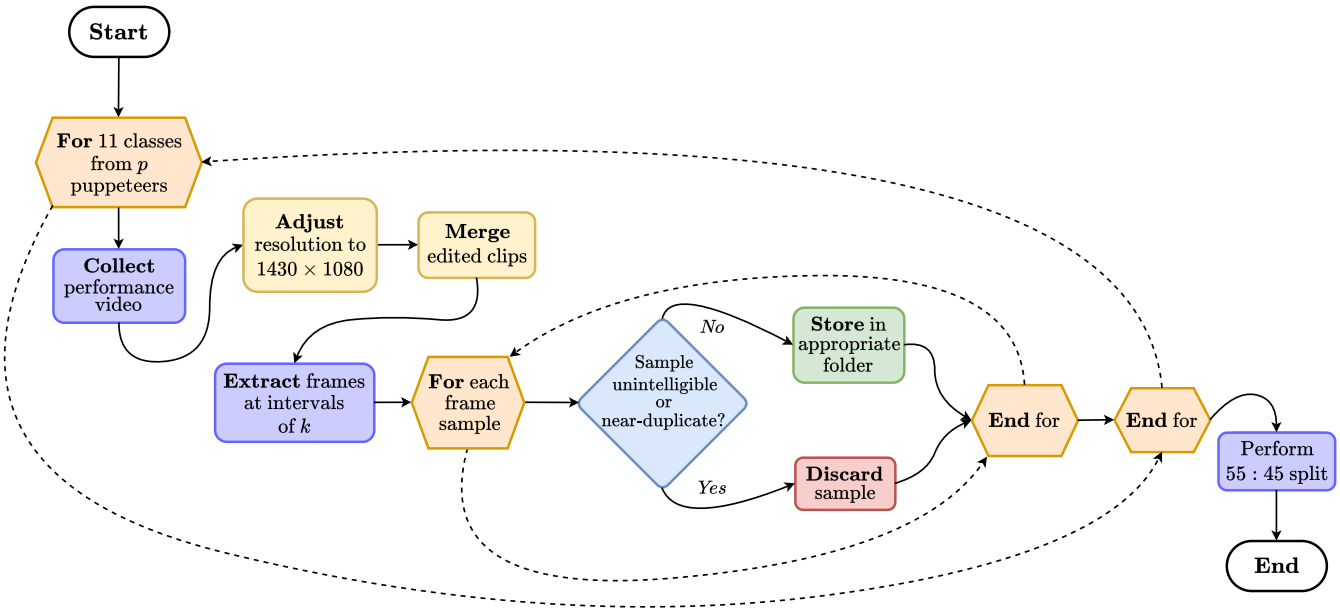


Fig. 2: A flowchart depicting the dataset construction process.

recognition tasks of Indonesian shadow puppets . Sudiatmika et al. [10], Sudiatmika and Dewi [11] used the deep CNN models ALEXNET [12] and VGG-16 [13] to perform the task of Indonesian shadow puppet classification. They constructed an augmented dataset of 2,530 images spanning across 6 classes of puppets (Arjuna puppet, Yudistira Puppet, Rahwana Puppet, Sahadewa Puppet, Dewi Sinta Puppet, and Gatot Kaca Puppet) collected from some puppet museums in Bali. For this image classification task, the authors also experimented with other convolutional models, such as MASK R-CNN [14] and MOBILENET [15], in two separate studies [16, 17].

In a similar spirit, our work is an endeavor towards establishing a performance benchmark of the recent SOTA feature extractor models for hand shadow puppet contour images, with a more large-scale and comprehensive approach.

B. 3D Modeling and Human Motion Capture

One of the earliest works involving silhouettes is a study that explored the mapping of monocular monochromatic 2D shadow image sequences of humans to animated 3D body poses using a configural and dynamical manifold [18]. The author characterized this manifold from data utilizing a hidden Markov model (HMM) endowed with special topological properties acquired via the process of entropy minimization without resorting to any articulatory body model. Several advances in vision-based human motion capture and analysis since then leveraged human silhouette templates [19, 20], more specifically hand and finger silhouettes [21, 22, 23].

C. Robotics

Huang et al. [24] introduced computer vision-aided shadow puppetry with robotics by matching shape correspondences of input images. They claimed that due to the physical limitations of human arms, it is often not feasible to construct complex shadow forms. Instead, they developed a framework that enabled them to produce shadow images with the mechanical arms of a robot. The authors built a library of shadow images, used these images to enable the robotic arms to orient themselves into a formation resembling the intended shadow puppet, and verified their acceptability.

D. Human-Computer Interaction

The authors of [25] proposed a framework for controlling two Chinese shadow puppets—a human model and an animal model, with the use of body gestures via a Microsoft Kinect sensor. Carr and Brown [26] conducted a similar work by building a real-time Indonesian shadow puppet storytelling application that is capable of mimicking the full-body actions of the user, using the Microsoft Kinect sensor. In order to leverage contactless gesture recognition (CGR) to teach traditional Chinese shadow puppetry to beginners, Tsai and Lee [27] developed a system using Leap Motion sensors. These studies on digitizing the art of shadow puppetry, or puppetry in general, were influenced to some extent by other similar works in the gesture recognition domain [28, 29, 6, 30, 31, 32]. Tang et al. [33] developed an intelligent shadow play system, called SHADOWTOUCH, which includes a multidimensional somatosensory interaction module coupled with an automatic choreography module, to facilitate natural interaction between the shadow play figures and the human users.

The motif of our work tessellates well with the core objectives of the aforementioned research works. The innovative utilization of digitized traditional arts serves as a means to preserve their inherent legacies and HASPER can be a potent contribution to the contemporary pool of resources to facilitate such innovative digitization for ombromanie.

TABLE I: Statistical summary of HASPER.

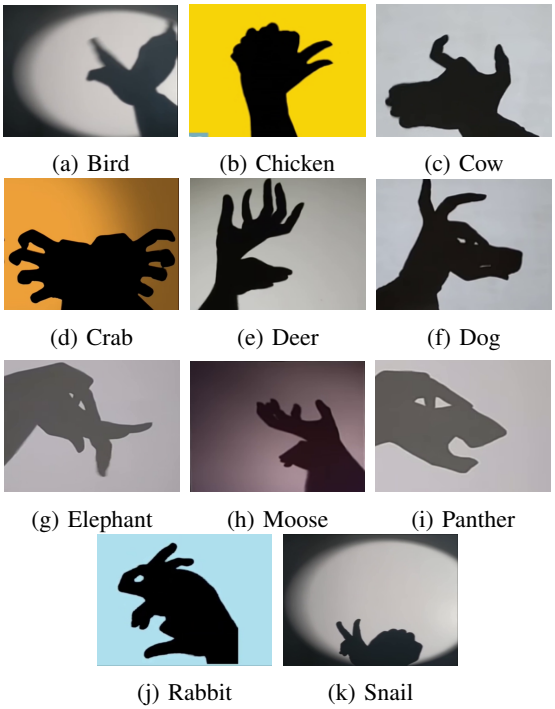
Silhouette Class	Clips		Sample Distribution		
	Professional	Amateur	# of Training Samples	# of Validation Samples	Total # of Samples
Bird	6	2	447	361	808
Chicken	2	2	431	389	820
Cow	2	2	412	223	635
Crab	4	2	367	328	695
Deer	6	2	423	379	802
Dog	7	2	466	365	831
Elephant	5	2	420	315	735
Moose	3	2	394	322	716
Panther	2	2	367	396	763
Rabbit	4	2	415	348	763
Snail	4	2	461	311	772
Total	45	22	4603	3737	8340
	67				

III. Dataset Construction

The series of steps involved in our data acquisition process is broadly divided into three tasks — procuring the performance clips, extraction of the frames, and categorization of each sample frame with a proper label. Figure 2 portrays this workflow behind our dataset preparation. We incorporate manual oversight at each step of the dataset creation in order to reconcile any exigencies pertaining to the quality of HASPER.

A. Collating Shadowgraphy Clips

In order to facilitate the creation of the dataset, we procure 45 different clips of 9 different professional ombromanie performers from YouTube[5], each performer varying in terms of their hand structure and stylistic choices. The video sources are licensed under fair use and a list consisting of the links to all of them is available in our GitHub[6] repository. We record the relevant portions of the performance videos using the open-source recording software OBS Studio[7]. Two novice volunteer shadowgraphists collectively produce 22 additional clips, with each contributing one clip for every class. As a consequence, the total number of source clips aggregates to $45 + (11 \times 2) = 67$.



(a) Bird (b) Chicken (c) Cow

(d) Crab (e) Deer (f) Dog

(g) Elephant (h) Moose (i) Panther

(j) Rabbit (k) Snail

Fig. 3: Samples from each class of the dataset.

B. Extracting Samples

To mitigate the presence of excessively similar and redundant image samples, we extract frames from these clips at reasonable intervals of k after downsampling the clips to a resolution of 1430×1080. The values of k are judiciously chosen for the clips of each class, and every kth frame is selected as a candidate image sample (*e.g.*, with $k \approx 180, 200, 220$ for a 60FPS clip). Table I encapsulates some essential statistical information germane to HASPER and provides a superficial overview of the dataset.

C. Labeling

After the extraction of the frames, the samples undergo manual scrutiny by two annotators, both pursuing undergraduate studies in Computer Science and Engineering. If a series of contiguous samples *prima facie* exhibit substantial similarity in terms of spatial properties, we only keep a single image from that set of samples. The rest are discarded to avoid redundancy and to instill diversity in the dataset. Another criterion that dictates the legitimacy of an image sample is its intelligibility. If the annotators agree on the unintelligibility of a sample, they discard it in unison. After performing this omission of unsuitable samples for each class, we end up with 11 different directories of images, each containing

[5]YouTube — https://www.youtube.com

[6]GitHub repository of HASPER — https://github.com/Starscream-11813/HaSPeR

[7]Open Broadcaster Software® — https://obsproject.com/

the curated samples of a particular class. The images in these folders are then further partitioned into training and validation sets, maintaining a $55 : 45$ split approximately. We also pragmatically incorporate a proper distribution of the samples sourced from amateur clips over both the training and validation sets, to avoid making the latter unfairly difficult for the classification models.

IV. Dataset Description

To provide a tangible exposition of the diverse samples constituted within HASPER, Figure 3 presents a compendium of representative images across the 11 classes. With a minimally astute perspicacity, we can observe that the samples vary in terms of the nature of the backgrounds, the anatomical structure of the puppeteers' hands, the photometric opacity/sharpness of the projected silhouettes, and a panoply of other aspects.

A. Background Variance

The hand shadow puppetry setup that a puppeteer's crew arranges prior to the performance greatly dictates the nature of the background against which the shadow puppets are displayed. If the location of the light source is very near to the wall or the translucent screen, then we can observe an elliptical shadow contour on the background as evident in Figures 3a and 3k. The angular directionality of the light also manifests a gradient effect on the background as can be seen in Figures 3d and 3h. The temperature and color of the light emanated by the light sources onto the screens also add to the diversity of the samples in the dataset.

(a) Sharp and high opacity (b) Diffuse and low opacity

Fig. 4: Samples with different silhouette properties.

B. Nature of the Silhouettes

The positioning of the light source with respect to the puppeteer's hands plays a role in shaping the shadows' quality. Proximity to the light source yields crisp, well-defined shadows (*e.g.*, Figure 4a), while increasing distance fosters softer, more diffuse shadows (*e.g.*, Figure 4b). The higher the contrast between the silhouettes and their respective backdrops, the more visible and well-contoured the shadow puppets are. The direction of light sources influences the orientation and shape of dark shadows. Shadows cast by overhead lighting sources may appear elongated or distorted, while shadows cast by low-angle lighting sources may exhibit softer edges and less pronounced contrast and sharpness. The shadows also differ in terms of the magnitude of their opacity, *i.e.*, they differ in the degree to which the puppeteers' hands prevent or significantly reduce the transmission of light being projected onto the screen.

Fig. 5: Samples of the 'Deer' class with different artistic representations.

C. Hand Anatomy and Stylistic Flair of the Puppeteers

The physiological properties of the puppeteers' hands can vary significantly due to a combination of genetic factors, environmental influences, and lifestyle choices. These nuanced anatomical variations pertaining to the wrists, palms, and digits of the puppeteers and the different stylistic choices they employ in their choreography contribute as yet another avenue of diversity of the image samples in HASPER. Figure 5 pristinely demonstrates the stylistic and morphological variations of hand shadow puppets of the 'Deer' class.

D. Comparative Analysis

1) Inter-class Similarity

Due to the conspicuous resemblance in the anatomical structures of certain animal species, the samples belonging to the classes corresponding to those animals exhibit a notable degree of similarity as well. These similarities make the image classification task on HASPER quite a challenging endeavor and culminate to being a reason behind a lot of misclassifications, as discussed in Section VI-C.

2) Intra-class Dissimilarity

Some classes include samples of multiple species of the same animal, and these samples are starkly different in appearance from one another. Given the presence of such quasi-disparate samples along with the individualistic flair of puppeteers that manifests through their stylistic choices, a particular class may show a lot of intra-class dissimilarity.

V. Statistical Analysis

Table I provides an overview of the statistical properties of the HASPER dataset. The histogram in Figure 6 illustrates the proportion of samples belonging to each of the 11 classes and their corresponding training-validation splits. Upon analyzing this histogram, it appears that the 'Dog' class has the highest number of images (831 samples $\approx 9.96\%$), while on the contrary, the 'Cow' class possesses the least number of images (635 samples $\approx 7.61\%$). This suggests that overall, hand shadow puppetry of dogs is the most popular—a fact that concurs well with our anecdotal observations while searching for hand shadow puppetry performances. The proportions of samples belonging to the 'Crab', 'Elephant', and 'Moose' classes (8.33%, 8.81%, and 8.59% respectively) are also

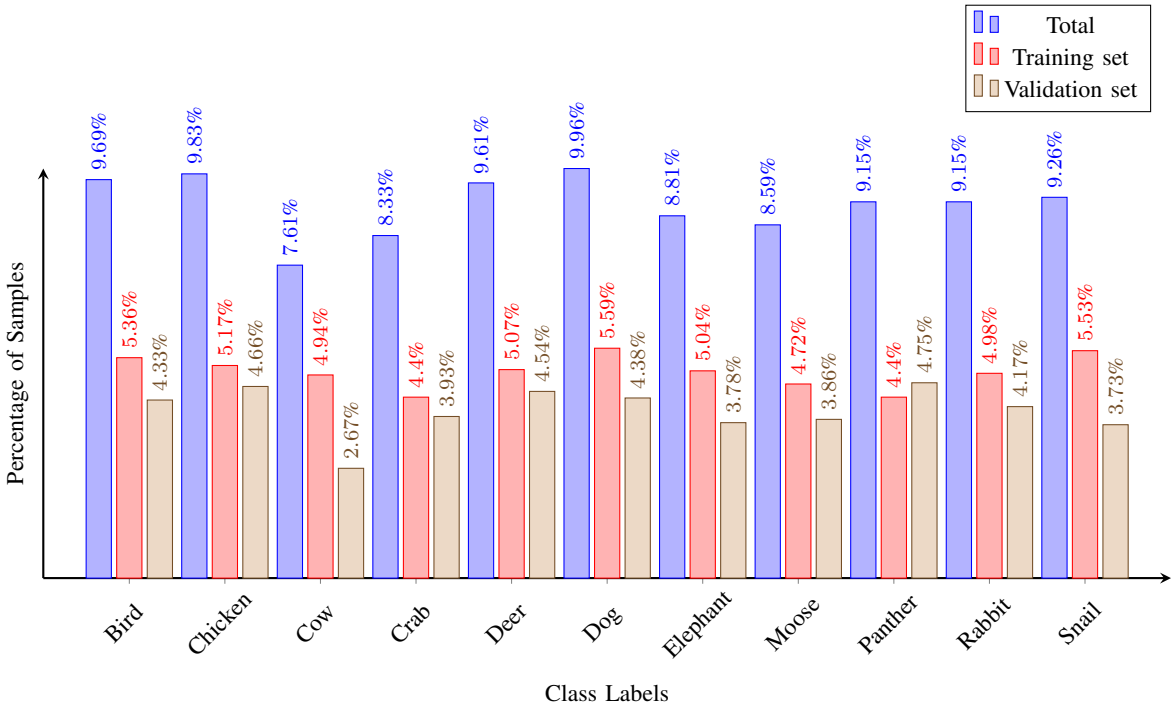


Fig. 6: Class Distribution of the Dataset.

slightly low due to a scarcity of performance clips starring hand shadow puppets of these classes. As evident in both Figure 6 and Table I, the image samples are quite evenly distributed across all 11 classes. Each class has $\approx 80\%$ samples sourced from clips of professional performers and the rest $\approx 20\%$ samples sourced from amateur clips. In totality, we end up with 4603 samples in the training set and 3737 samples in the validation set, thereby partitioning HASPER in accordance with a $55:45$ ratio.

VI. Developing Benchmark for HASPER

A series of pretrained image classification models are used as feature extractors to develop a benchmark for the dataset. The models are pretrained on the IMAGENET [34, 35] dataset and finetuned on HASPER. We implement the training pipeline using the Pytorch[8] framework. An overview of the models, evaluation metrics, results, and analysis of the experimental results are delineated in this section.

A. Experimental Setup

1) Baseline Models

We use 31 models—VITB16 [36], VITL32 [36], ALEXNET [12], VGG16 [13], VGG19 [13], GOOGLENET [37], RESNET18 [38], RESNET34 [38], RESNET50 [38], RESNET101 [38], RESNET152 [38], SQUEEZENET1_1 [39], WIDERESNET101_2 [40], WIDERESNET50_2 [40], DENSENET121 [41], DENSENET201 [41], SHUFFLENETV2X10 [42], MOBILENETV2 [43], MOBILENETV3SMALL [44], MOBILENETV3LARGE [44], CONVNEXTLARGE [45], EFFICIENTNETB0 [46], EFFICIENTNETV2S [47], MNASNET13 [48], SWINB [49], SWINV2B [50], CONVNEXT [45], MAXVIT [51], REGNETX32GF [52], RESNEXT101_32X8D [53], and INCEPTIONV3 [54]—as baselines. Some of these models have a track record of good performance across various image classification tasks. We consider both traditional Convolutional Neural Networks (CNN) and CNNs with attention mechanisms. Some models have multiple variants in terms of size or number of parameters, and we compare the performance among those variants as well.

2) Performance Metrics

We use top-k validation accuracy values (with $k = 1, 2, 3$), Precision, Recall, and F1-score as evaluation metrics to perform comparative analyses among the aforementioned models. The latter three judgement criteria are used due to the slightly imbalanced nature of HASPER, as evident in Figure 6.

3) Classifier Network

We adopt two approaches to arrive at the final 11-dimensional layer since there are a total of 11 classes to predict from. The first approach is to directly append an 11-dimensional fully connected layer at the tail-end of the vanilla models. The second approach is to incorporate the classifier block portrayed in Figure 7.

4) Hyperparameters and Optimizer

We use Stochastic Gradient Descent (SGD) [55, 56] with a learning rate of $\alpha = 0.001$ and momentum of $\gamma = 0.9$ as the optimizing method and Cross Entropy Loss as the loss metric for all the models. To decay the learning rate, we use Step Scheduler, which decays the learning rate by 0.1 every 5 epochs. Each model undergoes training for 50 epochs to ensure equitable comparison, and we empirically ascertain that 50

[8]PyTorch models and pretrained weights — https://pytorch.org/vision/stable/models.html

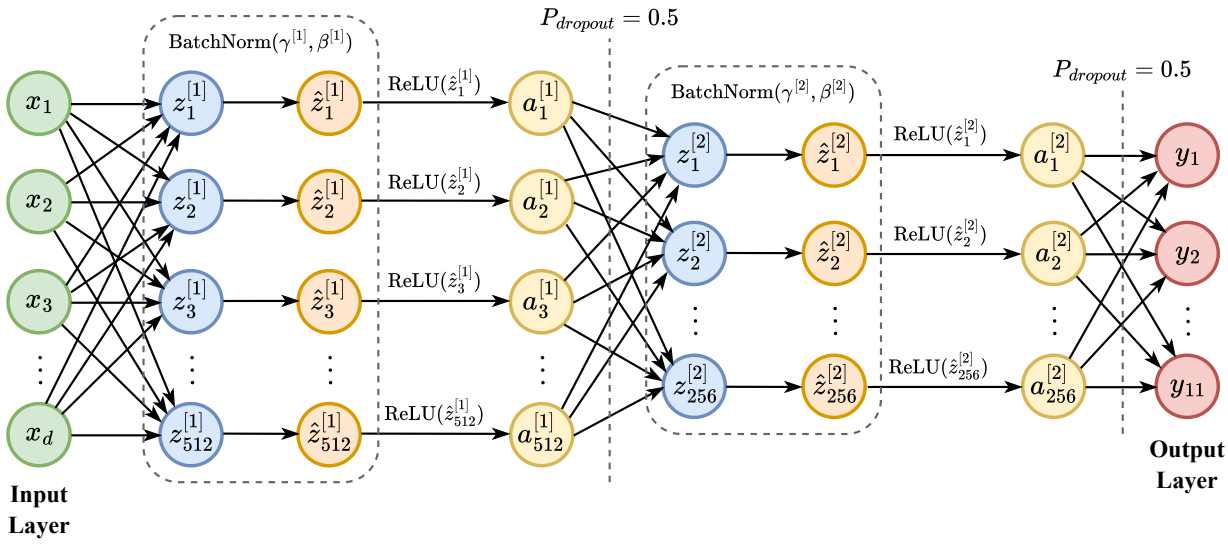


Fig. 7: Classifier block attached to the tail-end of the pretrained models. Here, d is the feature dimension of the anterior pretrained image classification model. The output features of layer l is $z^{[l]} = W^{[l]}a^{[l-1]}$, where $a^{[l-1]}$ denotes the activation values of the preceding $(l-1)$th layer. The batch normalized value of the ith output feature $z_i^{[l]}$ is $\hat{z}_i^{[l]} = \gamma^{[l]} z_{norm}^{[l](i)} + \beta^{[l]}$, where γ and β are learnable parameters. The activation values of layer l are denoted by $a^{[l]} = g(\hat{z}^{[l]})$ which is computed using the activation function $g = \text{ReLU}$.

epochs are sufficient for the majority of the models to achieve convergence.

5) Data Augmentation and Preprocessing

In order to generate a more diverse pool of training samples, we also incorporate data transformation techniques[9]—Random Resize, Random Perspective, Color Jitter, Random Horizontal Flip, Random Crop, Random Rotation, Gaussian Blur, and Random Affine with translation and shearing—while training the models. We choose these data augmentation techniques since the classes in HASPER are mostly rotationally asymmetric and incongruent. Consequently, the augmented samples aid in eliciting better generalization abilities and robustness for all the models. The input images that are fed to the models are appropriately resized a priori using the Bicubic Interpolation method.

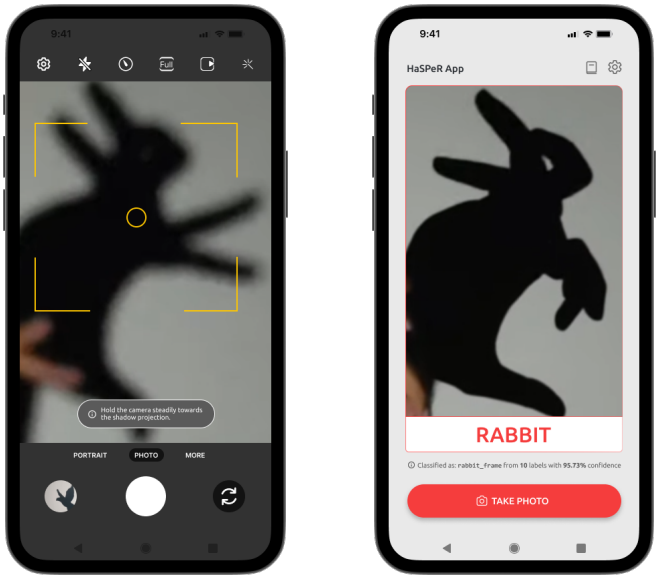
B. Results and Findings

Table II provides a summary of the experimental results for HASPER.

1) Performance Analysis

The INCEPTIONV3 model yielded the best performance with a top-1 accuracy of 88.97%. The vanilla version of the model also yields reasonably high Precision, Recall, and F1-score of 0.8979, 0.8943, and 0.8828 respectively. In terms of top-2 and top-3 accuracy values, we observe the competence of transformer architectures like SWINB and MAXVIT, with each achieving the highest top-2 and top-3 accuracy scores, respectively. Upon being equipped with the classifier block shown in Figure 7, the DENSENET201

[9]Torchvision transforming and augmenting images — https://pytorch.org/vision/stable/transforms.html



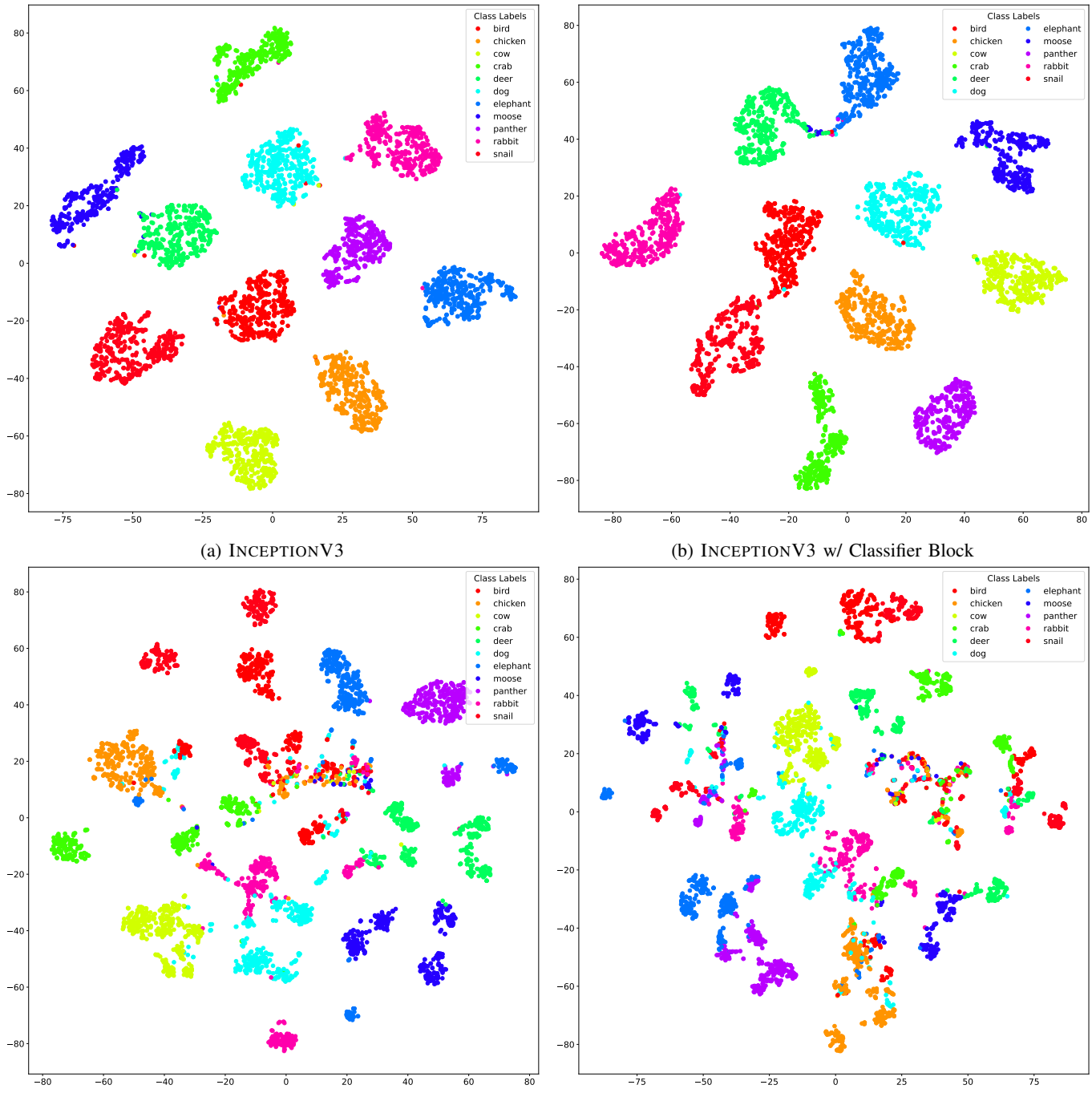
(a) Camera feed (b) Classification verdict

Fig. 8: Android application prototype for hand shadow puppet recognition.

model yields the highest top-1 accuracy and F1-score. In contrast, the RESNEXT101_32X8D model demonstrates the best performance across all the other evaluation metrics. In this recess of the performance analysis, we consider the top-1 accuracy metric to be the most significant metric. As evident in Table II, the RESNEXT101_32X8D model lags behind the INCEPTIONV3 model when it comes to the top-1 accuracy value in both cases, which is why we adjudicate that the latter is the best-performing model. It is noteworthy to

TABLE II: Performance comparison of the vanilla and modified versions of the image classification models.

Models	Performance Metrics											
	Vanilla						w/ Classifier Block					
	Top-k Accuracy (%)			Precision	Recall	F1-score	Top-k Accuracy (%)			Precision	Recall	F1-score
	Top-1	Top-2	Top-3				Top-1	Top-2	Top-3			
ShuffleNetV2X10 [42]	46.29	60.71	72.97	0.6006	0.4576	0.4225	81.64	87.26	92.23	0.8405	0.823	0.7989
ViTB16 [36]	65.9	75.06	79.82	0.7214	0.6649	0.6462	59	68.15	73.72	0.6576	0.5965	0.5792
ViTL32 [36]	75.94	82.82	89.08	0.7998	0.7716	0.7443	77.89	85.14	90.5	0.8158	0.7891	0.7614
AlexNet [12]	79.04	84.05	88.19	0.826	0.8015	0.7718	81.85	88.41	92.05	0.8448	0.8261	0.8022
SqueezeNet1_1 [39]	80.67	85.57	88.94	0.83	0.8127	0.7912	80.35	85.68	89.77	0.8296	0.8135	0.7902
MobileNetV3Small [44]	81.08	86.13	92.18	0.8529	0.8221	0.7993	81.72	87.1	90.98	0.8513	0.8258	0.8012
ConvNextLarge [45]	81.58	87.53	91.78	0.8441	0.8208	0.7997	84.93	88.86	92.9	0.8783	0.8522	0.8384
EfficientNetB0 [46]	81.91	85.38	90.42	0.8625	0.829	0.8012	83.54	88.22	92.96	0.8611	0.8418	0.8231
WideResNet50_2 [40]	83.32	88.76	95.53	0.8596	0.8416	0.8158	85.65	91.06	94.32	0.8751	0.861	0.8444
MNasNet13 [48]	83.43	86.35	88.94	0.8495	0.8388	0.8173	82.57	88.28	92.58	0.8445	0.8315	0.8094
VGG16 [13]	83.48	87.87	91.94	0.8462	0.8376	0.8187	84.85	90.95	94.4	0.857	0.8516	0.8364
VGG19 [13]	83.97	90.07	93.57	0.8636	0.8449	0.8307	84.02	90.09	93.25	0.8623	0.8401	0.8274
MobileNetV3Large [44]	84.23	90.71	94.86	0.8727	0.8468	0.83	83.32	87.79	93.57	0.8682	0.842	0.8218
EfficientNetV2S [47]	84.72	89.24	92.15	0.8667	0.8523	0.8329	85.97	90.01	92.26	0.876	0.8628	0.8482
MobileNetV2 [43]	84.8	90.28	94.88	0.8742	0.8563	0.8373	87.23	92.32	94.48	0.8837	0.8774	0.8668
GoogLeNet [37]	84.93	90.12	93.84	0.8615	0.8541	0.8348	83.97	88.97	92.4	0.8554	0.8448	0.8244
SwinV2B [50]	84.98	89.96	94.11	0.8747	0.8551	0.8439	81.34	86.88	93.22	0.853	0.8227	0.8021
ConvNeXt [45]	85.17	89.8	95.34	0.8751	0.856	0.8431	86.75	91.03	94.46	0.8803	0.8728	0.8608
ResNet18 [38]	85.36	92.56	94.4	0.8807	0.863	0.8411	85.84	92.61	95.21	0.8793	0.8665	0.8494
SwinB [49]	85.84	92.72	**97.35**	0.8789	0.8643	0.8535	85.12	91.67	95.23	0.8772	0.8564	0.843
ResNet101 [38]	86.37	93.17	95.85	0.8837	0.8706	0.8508	86.54	93.97	97	0.8764	0.8719	0.8556
MaxViT [51]	86.54	**93.84**	97.05	0.8903	0.8691	0.8642	85.2	91.54	95.18	0.8722	0.857	0.847
RegNetX32GF [52]	86.67	91.14	95.63	0.8879	0.8735	0.857	86.99	92.13	95.21	0.8823	0.8742	0.8608
DenseNet201 [41]	86.83	90.23	92.74	0.8853	0.8701	0.8548	**88.89**	93.49	95.42	0.9042	0.8953	**0.8844**
ResNet152 [38]	86.86	90.42	94.4	0.8929	0.877	0.8585	86.64	91.57	94.51	0.8844	0.8715	0.8576
ResNet50 [38]	86.91	92.74	95.47	0.8953	0.8771	0.8607	84.69	90.12	94.08	0.8717	0.8538	0.8357
ResNet34 [38]	87.55	92.93	95.55	0.8943	0.8809	0.8699	85.79	91.73	93.79	0.8767	0.8634	0.8461
WideResNet101_2 [40]	87.77	92.37	95.1	0.8955	0.8833	0.8684	86.27	92.45	96.14	0.8783	0.869	0.8533
DenseNet121 [41]	87.9	92.4	94.75	0.8947	0.8844	0.8719	86.08	90.95	94.64	0.8771	0.866	0.8498
ResNext101_32X8D [53]	88.62	93.17	95.85	**0.9045**	0.8927	0.8803	86.54	**93.97**	**97**	**0.9073**	**0.896**	0.8839
InceptionV3 [54]	**88.97**	92.77	94.19	0.8979	**0.8943**	**0.8828**	88.3	92.29	95.34	0.8905	0.8849	0.8777



(a) INCEPTIONV3 (b) INCEPTIONV3 w/ Classifier Block

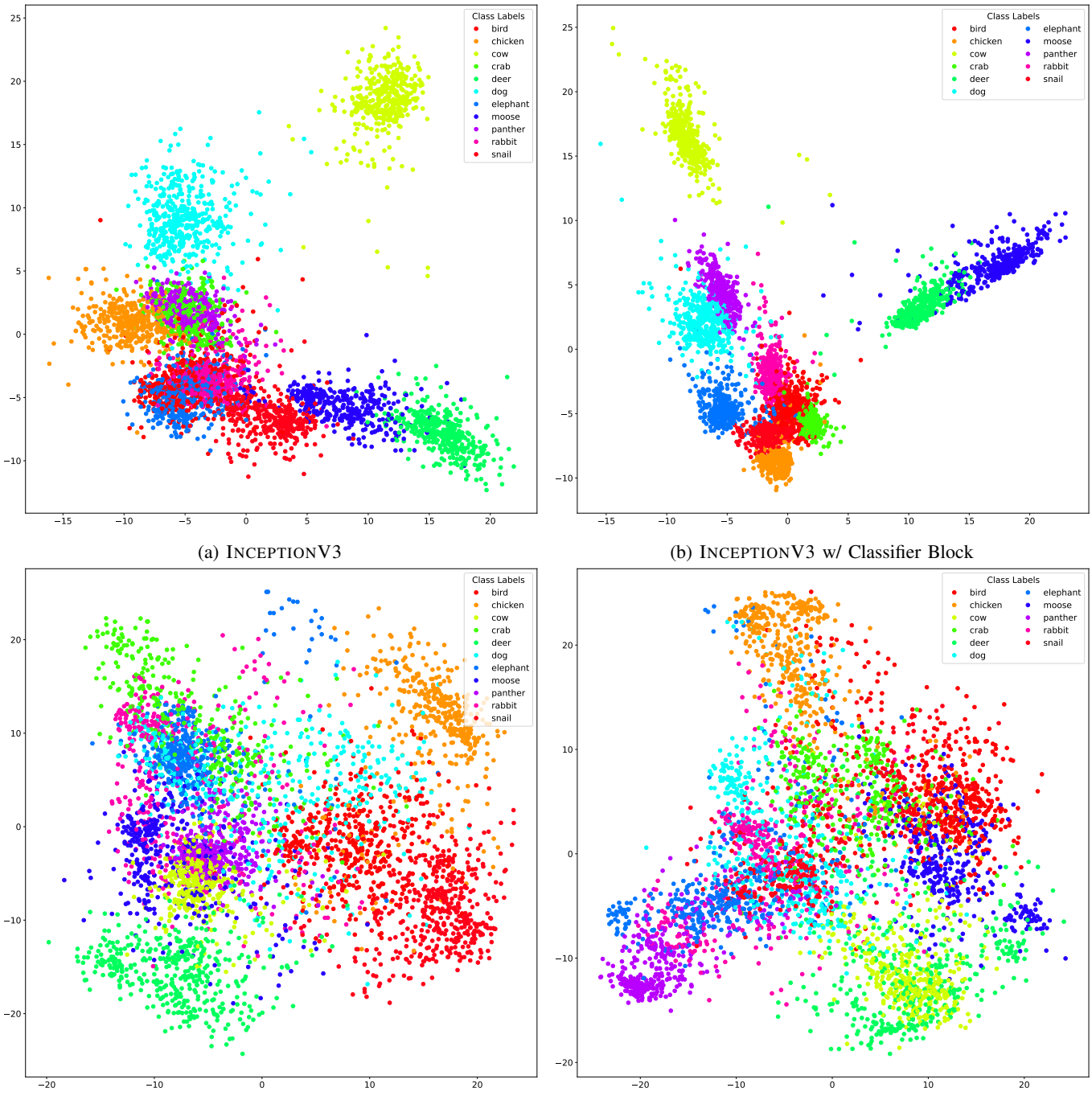
(c) VITB16 (d) VITB16 w/ Classifier Block

Fig. 9: t-Distributed Stochastic Neighbor Embedding (t-SNE) feature representations of INCEPTIONV3 and VITB16.

point out that MOBILENETV3LARGE, with only 5.4 million parameters, managed to surpass most of the other models in terms of performance. This indicates the suitability of this image classification task for lighter, low-latency models that can be used in mobile applications and embedded devices. We create a simple prototype Android application using Flutter to test the efficacy of MOBILENETV3LARGE in classifying hand shadow puppet images from the phone's camera feed. Figure 8 portrays the snapshots of the prototype application.

2) Feature Space Visualization and Analysis

In order to assess the reasons behind the stark performance disparity between the better and worse-performing models, we resort to two data visualization techniques via dimensionality reduction—Principal Component Analysis (PCA) [57, 58] and t-Distributed Stochastic Neighbor Embedding [59, 60]. Linear techniques, such as PCA, prioritize maintaining significant separation among low-dimensional representations of dissimilar datapoints by using Singular Value Decomposition (SVD) of the data distribution to project it down to a lower

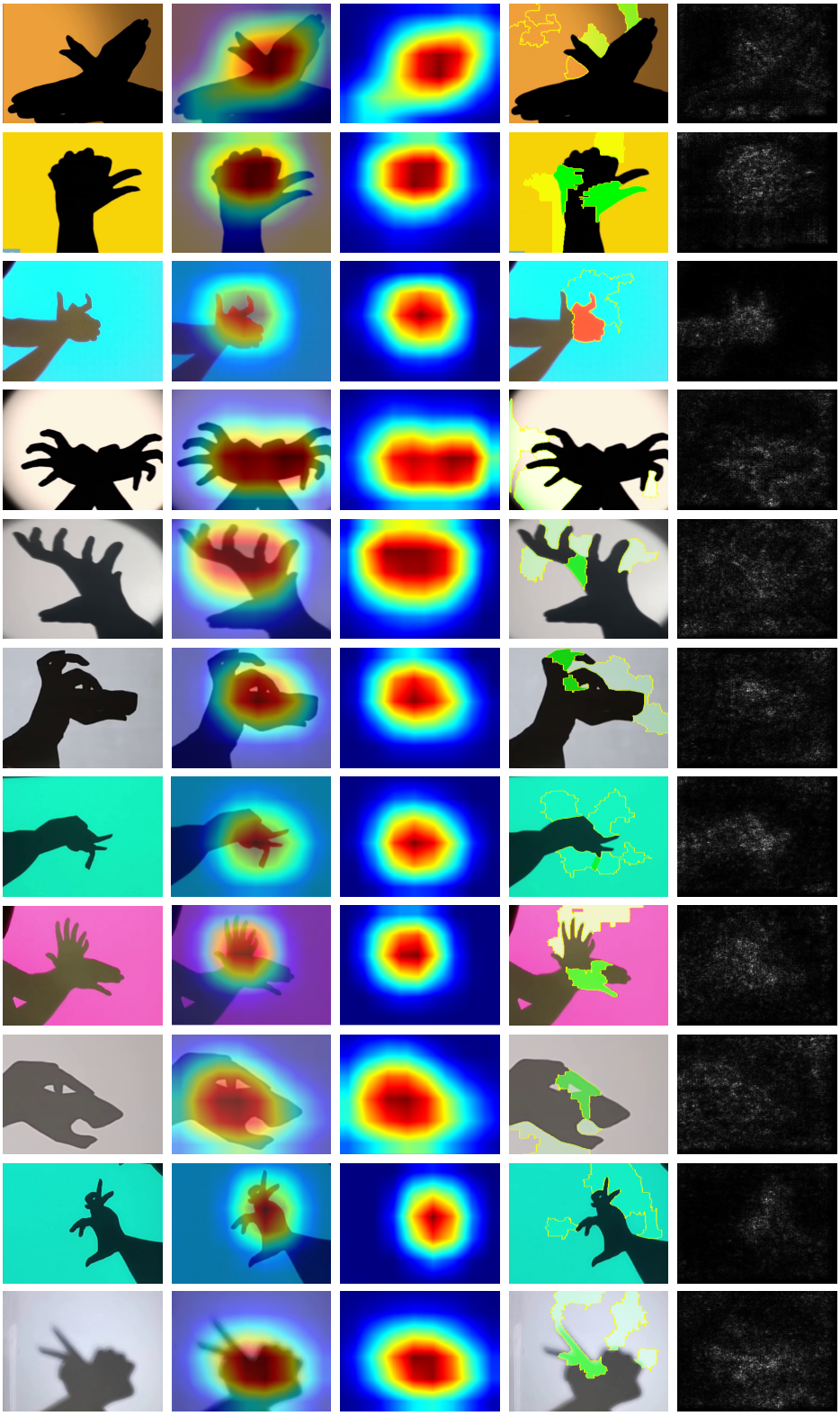


(a) INCEPTIONV3 (b) INCEPTIONV3 w/ Classifier Block

(c) VITB16 (d) VITB16 w/ Classifier Block

Fig. 10: Principal Component Analysis (PCA) feature representations of INCEPTIONV3 and VITB16.

dimensional vector space. However, for high-dimensional data residing on or proximate to a low-dimensional, non-linear manifold, it becomes imperative to preserve the proximity of the collapsed low-dimensional representations for closely resembling datapoints. Achieving such proximity preservation is often unattainable through linear mappings, which is why we opt for the t-SNE dimensionality reduction approach as well. Figure 9 and Figure 10 portray the t-SNE plots and PCA plots of the feature representations of the comparatively better and worse-performing models. They are all generated using the `scikit-learn` package [61]. It is pragmatic to put more importance on interpreting the t-SNE plots than the PCA plots because they are more intuitive and understandable. We can infer from the 2D-collapsed visualization of the high-dimensional feature representations in Figure 9a that the classes are nicely clustered and congealed with minimal overlaps and outliers. This enables the INCEPTIONV3 model to easily determine the decision surface in the high-dimensional feature space and perform very well on the classification task. Observing the t-SNE plot in Figure 9b and PCA plot in Figure



(a) Raw Image Sample (b) GradCAM Heatmap (c) GradCAM Score (d) LIME Visualization (e) Saliency Map

Fig. 11: The juxtaposition of original image samples from the HASPER dataset with their corresponding GradCAM Heatmaps, GradCAM Scores, LIME Visualizations, and Saliency Maps (for the best-performing INCEPTIONV3 model).

10b, we can deduce the reason for the slight degradation of performance for the INCEPTIONV3 model upon being coupled with the classifier block. It is clearly evident from those plots that there is a marginally higher overlap between the clusters representing the 'Moose' and 'Deer' classes, and the clusters are overall slightly less compact compared to the plots of the vanilla version of the model.

On the contrary, if we consider one of the comparatively worse-performing models, VITB16, we discern from Figures 9c, 9d, 10c, and 10d that its feature representation is quite sparse and inept. The clusters are quasi-compact and overlap more than desired, which may be the cause of numerous misclassifications. One probable reason for this underwhelming performance of the transformer-based models may be an inadequate number of training samples in the dataset, given the huge number of parameters of the model (307M).

(a) Actual Label: Crab, Predicted Class: Rabbit

(b) Similar sample from the 'Rabbit' class

(c) Actual Label: Moose, Predicted Class: Bird

(d) Similar sample from the 'Bird' class

Fig. 12: Misclassified samples with visually similar samples of the predicted class.

3) Qualitative Analysis and Explainability

The models put more emphasis on certain distinguishing aspects of each class. We adopt a plethora of Explainable AI (XAI) techniques on the best-performing INCEPTIONV3 model to visualize these aspects in Figure 11. While viewing the GradCAM (Gradient-weighted Class Activation Mapping) [62] attention heatmaps, it becomes apparent that the models generally put more gravitas on the common-sense distinguishing traits. For example, in Figure 11b, we observe the regions of the image samples predominantly influencing their respective classification scores—the wingspan and beak of a bird, the gallinaceous comb of a chicken, the horns and concave head of a cow, the appendages of a crab, the horns of a deer, the long-slanted head of a dog, the tusks of an elephant, the upright horns of a moose, the big eyes and small ears of a panther, the petite hands and head of a rabbit, as well as the shell and antennae of a snail. As human beings, we elicit these same distinguishing characteristics when we try to classify the images using our own reasoning faculties. As exemplified in Figure 11d, for local interpretation, we use the model-agnostic technique called LIME (Local Interpretable Model-agnostic Explanations) [63]. We also demonstrate the spatial support of the top-1 predicted classes by generating the saliency maps [64] in Figure 11e. These maps are rendered using a solitary back-propagation pass through the INCEPTIONV3 model, and they accentuate the salient areas of the given image, characterized by their discriminative attributes with respect to the given class.

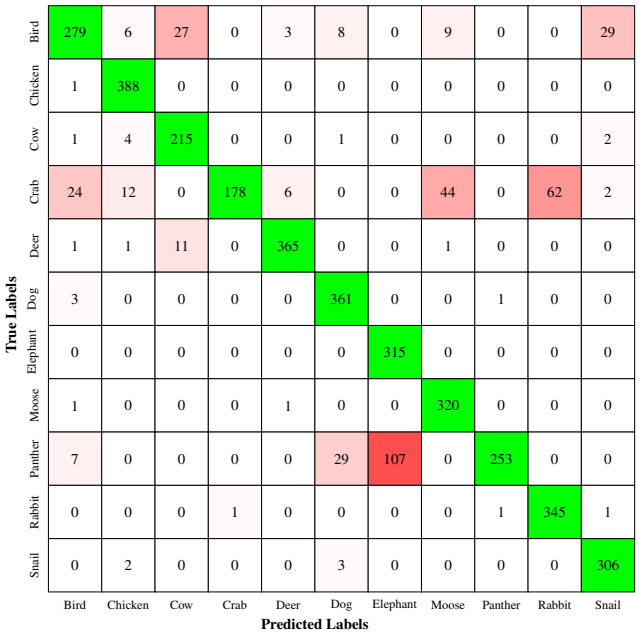


Fig. 13: Confusion Matrix of INCEPTIONV3.

C. Error Analysis

The confusion matrix for the INCEPTIONV3 model on our dataset, presented in Figure 13, reveals that the 'Panther' class exhibits the highest count of misclassifications. One obvious reason for this is the somewhat significant inter-class similarity among the 'Dog', 'Elephant', and 'Panther' classes. Most of the misclassified samples are from visually similar classes. We can posit that navigating the intricacies of visually similar classes poses a significant challenge in image classification tasks, as evident from the other pale-red entries of the confusion matrix in Figure 13. Even to the keen human eye, distinguishing between these classes may be perplexing, as they share common visual features, shapes, or color patterns that result in a high degree of resemblance. We examine various aspects, such as the distinctive features or characteristics that might have led to confusion and the degree of similarity between the misclassified classes. Figure 12a and 12b show the confusion between a 'Crab' sample and a 'Rabbit' sample which look visually quite similar. The same holds for the sole 'Moose' sample that is misclassified as a 'Bird' sample by the INCEPTIONV3 model, as depicted in Figure 12c and 12d. We observe that misclassifications of this type occur when images belonging to distinct but visually akin categories are erroneously assigned to the wrong class.

There are numerous other examples of such miscategorized validation samples, but we leave them out of this paper for brevity. The green entries along the diagonal of the confusion matrix in Figure 13 indicate the reasonably good classwise prediction performance of the INCEPTIONV3 model, even though there exists a slight class imbalance in HASPER as evident from Figure 6.

VII. LIMITATIONS

Despite our utmost efforts, we acknowledge that our work still harbors certain limitations. Although these limitations do not undermine the validity and significance of the findings, they provide insight into areas that warrant further investigation. Due to the physical constraints and novice skill level of the employed individuals constructing the shadows, we could not generate a varied range of shadow images that accurately represent the palm and wrist structures of a diverse array of individuals across different classes. There's also scope for augmenting our dataset with numerous permutations of arm positioning and finger movements to create diverse silhouettes. We posit that the dataset in a more supplemented state will be suitable enough to finetune models with a very large number of parameters. Using some posture/gesture detection technology such as MediaPipe[10] or Microsoft Kinect[11], we can leverage depth coordinates of hand landmarks to assess their efficacy in classifying hand shadow puppets, as is done in many recently published works on sign language recognition. Some salient digital image pre-processing techniques may be adopted prior to feeding the images as input to the models for yielding better performance. We explore this intuition to some extent by using rudimentary morphological image processing techniques such as Dilation, Erosion, Opening, and Closing as preprocessing steps; but in vain, as they yield mediocre results.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we extensively explore an intriguing subject matter, Hand Shadow Puppetry, in the realm of computer vision. We introduce HASPER, a dataset with a sizable collection of 8,340 hand shadow puppet images distributed across 11 classes, taken from both expert and amateur puppetry performance clips. We finetune 31 pretrained image classification models on HASPER to establish a benchmark for the dataset. To explain and analyze the performance of the most erudite model, INCEPTIONV3, in comparison with other baseline models, we visually manifest their feature spaces using dimensionality reduction techniques. We also perform thorough qualitative and error analyses for the INCEPTIONV3 model. We envisage the possibility of developing applications for imparting the art of shadowgraphy via mobile and embedded devices. We claim that this work is novel and significant since it is the first dataset and study on image classification benchmarking that focuses only on ombromanie. We wish to build upon this work in the future by addressing the limitations we have outlined in the penultimate section, Section VII. We hope our work will be deemed a worthy and meaningful contribution to this domain of research.

[10]MediaPipe — https://developers.google.com/mediapipe

[11]Kinect for Windows — https://learn.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows

IX. ACKNOWLEDGMENTS

We convey our heartfelt gratitude, in advance, to the anonymous reviewers for their constructive criticisms and insightful feedback which will surely be conducive to the improvement of the research work outlined in this paper. We also appreciate the Systems and Software Lab (SSL) of the Islamic University of Technology (IUT) for the generous provision of computing resources during the course of this project. Syed Rifat Raiyan, in particular, wants to thank his parents, Syed Sirajul Islam and Kazi Shahana Begum, for everything.

REFERENCES

[1] A. Almoznino and Y. Pinas, *The art of hand shadows*. Courier Corporation, 2002.

[2] E. T. Clearinghouse, "Hand shadow puppetry gallery," https://etc.usf.edu/clipart/galleries/266-hand-shadow-puppetry, Florida Center for Instructional Technology, College of Education, University of South Florida, 2004, accessed: 2024.

[3] A.-S. Maerten and D. Soydaner, "From paintbrush to pixel: A review of deep neural networks in ai-generated art," *arXiv preprint arXiv:2302.10913*, 2023.

[4] D. Samuel, R. Ben-Ari, S. Raviv, N. Darshan, and G. Chechik, "Generating images of rare concepts using pre-trained diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4695–4703.

[5] R. Saritha, "An artist nurturing a dying art and his quest for its conservation," *YourStory*, 2017. [Online]. Available: https://yourstory.com/2017/05/artist-nurturing-dying-art-quest-conservation

[6] F. Lu, F. Tian, Y. Jiang, X. Cao, W. Luo, G. Li, X. Zhang, G. Dai, and H. Wang, "Shadowstory: creative and collaborative digital storytelling inspired by cultural heritage," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 1919–1928.

[7] M. Huang, S. Mehrotra, and F. Sparacino, "Shadow vision," 1999.

[8] D. H. Ballard and C. M. Brown, *Computer vision*. Prentice Hall Professional Technical Reference, 1982.

[9] B. G. Mertzios and K. D. Tsirikolias, "Fast shape discrimination using one-dimensional moments," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*. IEEE Computer Society, 1991, pp. 2473–2474.

[10] I. B. K. Sudiatmika *et al.*, "Indonesian traditional shadow puppet image classification: A deep learning approach," in *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*. IEEE, 2018, pp. 130–135.

[11] I. B. K. Sudiatmika and I. G. A. A. S. Dewi, "Indonesian shadow puppet recognition using vgg-16 and cosine similarity," *The IJICS (International Journal of Informatics and Computer Science)*, vol. 5, no. 1, pp. 1–6, 2021.

[12] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *arXiv preprint arXiv:1404.5997*, 2014.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[16] I. B. K. Sudiatmika, M. Artana, N. W. Utami, M. A. P. Putra, and E. G. A. Dewi, "Mask r-cnn for indonesian shadow puppet recognition and classification," in *Journal of Physics: Conference Series*, vol. 1783, no. 1. IOP Publishing, 2021, p. 012032.

[17] D. P. Prabowo, M. K. A. Nugraha, D. I. Ihya'Ulumuddin, R. A. Pramunendar, and S. Santosa, "Indonesian traditional shadow puppet classification using convolutional neural network," in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE, 2021, pp. 1–5.

[18] M. Brand, "Shadow puppetry," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2. IEEE, 1999, pp. 1237–1244.

[19] R. T. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," in *Proceedings of fifth IEEE international conference on automatic face gesture recognition*. IEEE, 2002, pp. 366–371.

[20] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[21] A. Tsuji, K. Ushida, and Q. Chen, "Real time animation of 3d models with finger plays and hand shadow," in *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*, ser. ISS '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 441–444. [Online]. Available: https://doi.org/10.1145/3279778.3279918

[22] A. Tsuji, K. Ushida, S. Yamaguchi, and Q. Chen, "Real-time collaborative animation of 3d models with finger play and hand shadow," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 1195–1196.

[23] A. Tsuji and K. Ushida, "Telecommunication using 3dcg avatars manipulated with finger plays and hand shadow," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2021, pp. 39–40.

[24] 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*, 2017, pp. 1251–1252.

[25] 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*. Springer, 2012, pp. 352–361.

[26] B. M. Carr and G. J. Brown, "Shadow puppetry using the kinect," 2014.

[27] T.-H. Tsai and L.-C. Lee, "A study of using contactless gesture recognition on shadow puppet manipulation [j]," *ICIC express letters. Part B, Applications: an international journal of research and surveys*, vol. 7, no. 11, pp. 2317–2322, 2016.

[28] U. Güdükbay, F. Erol, and N. Erdogan, "Beyond tradition and modernity: digital shadow theater," *Leonardo*, vol. 33, no. 4, pp. 264–265, 2000.

[29] S.-z. Gao, "On the digital development of the chinese shadow play art," in *2011 International Conference on Internet Technology and Applications*. IEEE, 2011, pp. 1–4.

[30] H. Liang, J. Chang, S. Deng, C. Chen, R. Tong, and J. Zhang, "Exploitation of novel multiplayer gesture-based interaction and virtual puppetry for digital storytelling to develop children's narrative skills," in *Proceedings of the 14th ACM SIGGRAPH International Conference on Virtual Reality Continuum and its Applications in Industry*, 2015, pp. 63–72.

[31] Z. Yan, Z. Jia, Y. Chen, and H. Ding, "The interactive narration of chinese shadow play," in *2016 International Conference on Virtual Reality and Visualization (ICVRV)*. IEEE, 2016, pp. 341–345.

[32] H. Liang, J. Chang, I. K. Kazmi, J. J. Zhang, and P. Jiao, "Hand gesture-based interactive puppetry system to assist storytelling for children," *The Visual Computer*, vol. 33, pp. 517–531, 2017.

[33] Z. Tang, Y. Hu, W. Weng, L. Zhang, L. Zhang, and J. Ying, "An intelligent shadow play system with multi-dimensional interactive perception," *International Journal of Human–Computer Interaction*, vol. 39, no. 6, pp. 1314–1326, 2023.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡1mb model size," *ArXiv*, vol. abs/1602.07360, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:14136028

[40] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[42] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[43] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[44] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[45] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[46] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[47] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[48] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2820–2828.

[49] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[50] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.

[51] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*. Springer, 2022, pp. 459–479.

[52] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 428–10 436.

[53] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[54] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[55] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, pp. 462–466, 1952.

[56] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[57] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[58] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[59] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[60] K. Bunte, S. Haase, M. Biehl, and T. Villmann, "Stochastic neighbor embedding (sne) for dimension reduction and visualization using arbitrary divergences," *Neurocomputing*, vol. 90, pp. 23–45, 2012.

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[62] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[63] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[64] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: visualising image classification models and saliency maps," in *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR, 2014.

Syed Rifat Raiyan was born in Agrabad, Chattogram, Bangladesh in 2001. He alma mater includes PlaySchool, Milestone College, RAJUK Uttara Model College, and Notre Dame College. He attained the Bachelor of Science (B.Sc.) degree with Honors from the Department of Computer Science and Engineering (CSE) at the Islamic University of Technology (IUT), Board Bazar, Gazipur-1704, Dhaka, in 2023.

He worked as an Industrial Trainee at Battery Low Interactive Ltd. in 2021. Since the 16th of August 2022, he has been working as a Lecturer of the Department of Computer Science and Engineering (CSE) at the Islamic University of Technology (IUT). He is currently affiliated with the Systems and Software Lab (SSL) research group of IUT. His research interests lie broadly in natural language processing, computer vision, and deep learning. His works have been published in the Findings of the Association for Computational Linguistics: ACL 2023 and in the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop). His current endeavors revolve around projects concerning mathematical reasoning in language models, in-context learning for text classification, Bangla sentiment analysis, and image classification.

Mr. Raiyan is a member of the Association for Computational Linguistics (ACL) since 2023.

Zibran Zarif Amio was born in Dulahazara, Cox's Bazar, Bangladesh in 2000. His alma mater includes BIAM Laboratory School, Cox's Bazar Govt. High School, and Dhaka Residential Model College. He attained the Bachelor of Science (B.Sc.) degree with First Class from the Department of Computer Science and Engineering (CSE) at the Islamic University of Technology (IUT), Board Bazar, Gazipur-1704, Dhaka, in 2023.

He worked as an Android Developer at BYDO Academy in 2020. He was an Industrial Trainee at Battery Low Interactive Ltd. in 2021. Since the 1st of June 2023, he has been working as the Chief Technology Officer at ReplyMind AI Ltd., Grameen Telecom Bhaban, Mirpur-1216, Dhaka. His work involves leading and developing Generative AI (GPT4, Claude) enabled software solutions. His research interests lie broadly in natural language processing, computer vision, and deep learning. His current endeavors include Meta Business Automation with Large Language Models.

Sabbir Ahmed was born in Dhaka, Bangladesh, in 1996. He received the B.Sc. Engg. degree (Hons.) in computer science (CS) from the Islamic University of Technology (IUT), Gazipur, Bangladesh, in 2017, where he is pursuing the M.Sc. degree in CS.

From 2018 to 2022, he was a Lecturer of the Department of Computer Science and Engineering, IUT, and since 2022 he has been working as an Assistant Professor. His research interests include pattern recognition, deep learning in computer vision, and intelligent agriculture. He received the IUT Gold Medal for his B.Sc. Engg. degree.