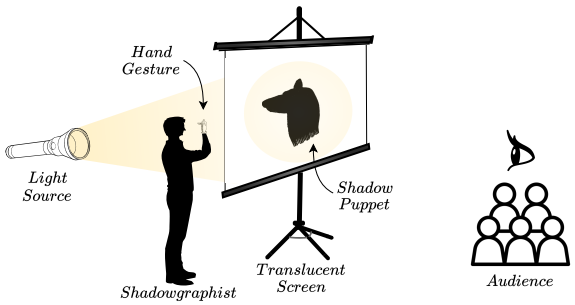
HASPER: An Image Repository for Hand Shadow Puppet Recognition

Syed Rifat Raiyan[1] Zibran Zarif Amio Sabbir Ahmed[2]
[1]Systems and Software Lab (SSL) [2]Computer Vision Lab (CVLab)
Islamic University of Technology (IUT), Boardbazar, Gazipur-1704, Dhaka, Bangladesh
{rifatraiyan,zibranzarif,sabbirahmed}@iut-dhaka.edu



(a) A generic hand shadow puppetry setup.

(b) Rabbit

(c) Bird

(d) Dog

Figure 1. Ombromanie in a nutshell.[1]

Abstract

Hand shadow puppetry, also known 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 entertainment standards, this art form is on the verge of extinction. To facilitate its preservation and proliferate it to a wider audience, we introduce HASPER*, a novel dataset consisting of 15,000 images of hand shadow puppets across 15 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. Our findings show a substantial performance superiority of skip-connected convolutional models over attention-based transformer architectures. We also find that lightweight models, such as* MOBILENETV2*, suited for mobile applications and embedded devices, perform comparatively well. We surmise that such low-latency architectures can be useful in developing ombromanie teaching tools, and we create a prototype application to explore this surmission. Keeping the best-performing model* RESNET34 *under the limelight, we conduct comprehensive feature-spatial, explainability, and error analyses to gain insights into its decision-making process and explore architectural improvements. To the best of our knowledge, this is the first documented dataset and research endeavor to preserve this dying art for future generations, with computer vision approaches. Our code and data are publicly available at https://github.com/Starscream-11813/HaSPeR.*

> "Will he not fancy that the shadows which he formerly saw are truer than the objects which are now shown to him?"
>
> Plato, *The Republic* (Book VII, Allegory of the Cave)

1. 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 props [1, 54]. The alias *"cinema in silhouette"*[2] is sometimes used to refer to this proto-cinematic medium of entertainment. Its working principle is very straightforward—the puppeteer adeptly positions their hands between a radiant light source and a translucent screen, consequently conjuring shadows and silhouettes that emulate different creatures, as shown in Fig. 1. Despite its rich history and captivating allure across many cultures,[3] there exists a notable dearth of resources

[1]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

[2]https://en.wikipedia.org/wiki/Shadowgraphy_(performing_art)

[3]https://www.geniimagazine.com/wiki/index.php/Shadowgraphy

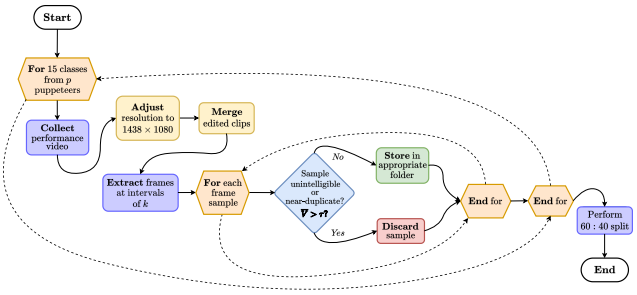


Figure 2. A flowchart depicting the dataset construction process.

specifically tailored to this artistic domain. With properly annotated and sourced data, researchers could study 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 [39]. The generation aspect is particularly relevant given the demonstrable impotency of AI image generator models in accurately creating hands and fingers [47]. 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 [49]. In 2011, UNESCO recognized shadow puppetry as an endangered artistic tradition by adding it to the Intangible Cultural Heritage list [36], 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**uppe**t** Image **R**epository), a methodically curated novel image dataset of hand shadow puppets. The dataset comprises an assemblage of 15,000 samples, that we painstakingly source and verify from 68 professional shadowgraphist clips and 90 amateur shadowgraphist clips. We label and categorize the images with utmost precision to elicit robustness in the image classification models that will undergo training with these images. The samples in HASPER are diverse in nature since the source clips are recorded in a plethora of different poses, orientations, and background lighting conditions of the translucent screen. We also inculcate silhouette motion diversity via optical flow estimation [17] in the frame extraction process. 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 RESNET34 model, including its feature representations, feature fusions, interpretability, explainability, and classification errors that it encounters. In an effort to assess the potential of digitized ombromanie teaching tools, we create a simple and lightweight prototype Android application using Flutter for classifying hand shadow puppet images from the phone's camera feed. We posit that our dataset possesses the potential to offer a wealth of opportunities for exploration and analysis into the artistic domain of hand shadow puppetry.

Silhouette Class	Clips		Sample Distribution		
	Pro.	Nov.	Training	Validation	Total
Bird	6	6	600	400	1000
Chicken	2	6	600	400	1000
Cow	2	6	600	400	1000
Crab	4	6	600	400	1000
Deer	6	6	600	400	1000
Dog	7	6	600	400	1000
Elephant	5	6	600	400	1000
Horse	8	6	600	400	1000
Llama	2	6	600	400	1000
Moose	3	6	600	400	1000
Panther	2	6	600	400	1000
Rabbit	4	6	600	400	1000
Snail	4	6	600	400	1000
Snake	3	6	600	400	1000
Swan	10	6	600	400	1000
Total	68	90	9000	6000	15000
	158				

Table 1. Statistical summary of HASPER.

2. Dataset Construction

The series of steps involved in our data acquisition process is broadly divided into three tasks—(a) procuring the performance clips, (b) extraction of the frames, and (c) categorization of each sample frame with a proper label. Fig. 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.

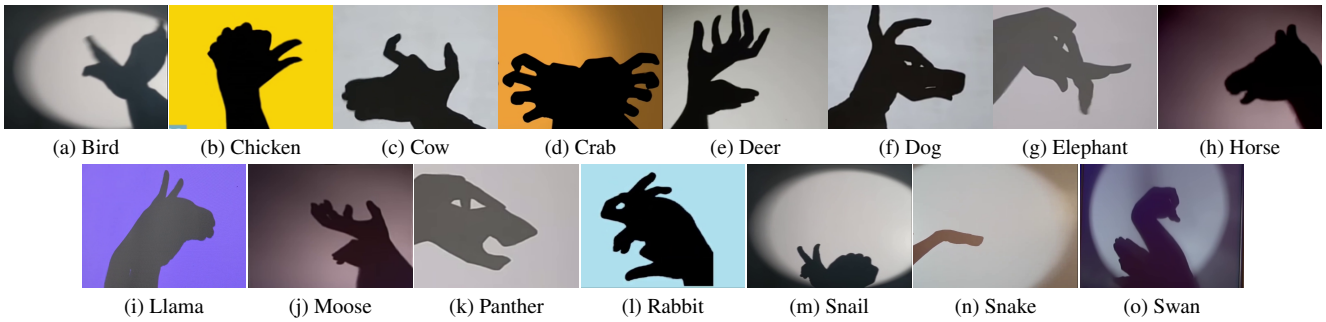
2.1. Collating Shadowgraphy Clips

At the outset of the process, we procure 68 different clips of 14 different professional shadowgraphists from YouTube.[4] The video sources are licensed under fair use and a list consisting of the links to all of them is available in our GitHub[5] repository. We record the relevant portions of the performance videos using the open-source recording software OBS Studio.[6] Six novice volunteer shadowgraphists collectively produce 90 additional clips, with each contributing one clip for every class. As a consequence, the total number of source clips aggregates to $68 + (15 \times 6) = 158$.

[4] https://www.youtube.com

[5] GitHub repository—https://github.com/Starscream-11813/HaSPeR

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



(a) Bird (b) Chicken (c) Cow (d) Crab (e) Deer (f) Dog (g) Elephant (h) Horse

(i) Llama (j) Moose (k) Panther (l) Rabbit (m) Snail (n) Snake (o) Swan

Figure 3. Samples from each class of the dataset.

2.2. 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 1438×1080. 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 60 FPS clip). From this sequence of extracted candidate frames, we prioritize those exhibiting significant motion. To this end, we estimate the motion vector field by calculating the optical flow between the consecutive tth frame and the $(t+k)$th frame. The magnitude of this motion is quantified by the mean L2 norm of the resulting flow field. We retain the frame pairs with an average flow magnitude surpassing a certain requisite threshold τ, thereby ensuring the inclusion of dynamically distinct frames. We synergistically amalgamate two optical flow estimation methods: the Lucas–Kanade (LK) method [37] and the Total Variation L1 Regularization (TV-L^1) method [72]. The undergirding assumption beneath the LK method is brightness constancy and spatial coherence of the flow in a local neighbourhood of the pixel (say, the patch W) under consideration. It employs a multi-scale gradient descent optimization approach for the constraint equation shown in Eq. (1).

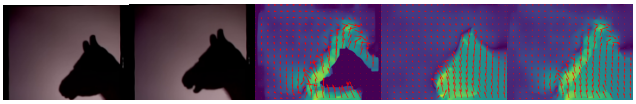
$$I_x \cdot u + I_y \cdot v + I_t = 0 \tag{1}$$

where, $I_x = \frac{\partial I}{\partial x}$ and $I_y = \frac{\partial I}{\partial y}$ are the spatial gradients of the image intensity I, and $I_t = \frac{\partial I}{\partial t}$ is the temporal gradient. u and v are the horizontal and vertical components of the optical flow vector, respectively. The motion is then estimated by iteratively minimizing the cost function in Eq. (2) at increasingly granular image resolutions, from coarse to fine.

$$\begin{bmatrix} u \\ v \end{bmatrix}_{\text{LK}} = \arg\min_{u,v} \sum_{\{i,j\} \in W} \left[I_x(i,j) \cdot u_{ij} + I_y(i,j) \cdot v_{ij} + I_t(i,j) \right]^2 \tag{2}$$

which can be then solved with the closed-form solution in Eq. (3).

$$\begin{bmatrix} u \\ v \end{bmatrix}_{\text{LK}} = \begin{bmatrix} \sum_{\{i,j\} \in W} I_x^2 & \sum_{\{i,j\} \in W} I_x I_y \\ \sum_{\{i,j\} \in W} I_x I_y & \sum_{\{i,j\} \in W} I_y^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_{\{i,j\} \in W} I_x I_t \\ -\sum_{\{i,j\} \in W} I_y I_t \end{bmatrix} \tag{3}$$



(a) tth frame (b) $(t+k)$th frame (c) LK method (d) TV-L^1 method (e) $\max(\text{LK}, \text{TV-}L^1)$

Figure 4. Optical flow estimation of contiguous candidate frames from the 'Horse' class.

As evident from Fig. 4c, the LK method can track the edge movement of the homogeneous silhouette patches, given the slight texture offered by the penumbral region of the shadow. However, as it is limited by local window constraints, it fails to capture the global motion of the shadow puppet. To ameliorate this issue, we resort to the TV-L^1 method, which is a variational method that minimizes the total variation of the flow field, subject to the L1 norm of the data fidelity term, which together form the energy function E in Eq. (4).

$$\begin{bmatrix} u \\ v \end{bmatrix}_{\text{TV-}L^1} = \arg\min_{u,v} E(u,v) \tag{4}$$

$$= \arg\min_{u,v} \int_{\Omega} (\lambda \underbrace{\|\nabla I \cdot \vec{w} + I_t\|_1}_{\text{Data term}} + \underbrace{\|\nabla u\|_1 + \|\nabla v\|_1}_{\text{L1 Regularization term}}) \, \mathrm{d}x \, \mathrm{d}y \tag{5}$$

where $\vec{w} = \langle u, v \rangle$ is the optical flow vector, ∇u and ∇v are the spatial gradients of the flow, λ is the parameter for balancing data fidelity and regularization, and $\Omega \subseteq \mathbb{R}^2$ represents the spatial domain of the entire image. The TV-L^1 method is more adept at capturing the global motion of the homogeneous and spatially consistent inner portion of the shadow puppet, as depicted in Fig. 4d. We then take the element-wise maximum of the LK and TV-L^1 flow fields' L2 norms to obtain a more hoilistic optical flow field, as portrayed in Fig. 4e.

$$\begin{bmatrix} u \\ v \end{bmatrix}^* = \arg\max_{u,v} \left(\left\| \begin{bmatrix} u \\ v \end{bmatrix}_{\text{LK}} \right\|_2, \left\| \begin{bmatrix} u \\ v \end{bmatrix}_{\text{TV-}L^1} \right\|_2 \right) \tag{6}$$

Then from this maximum combination $M \times N$ flow field, we compute the mean L2 norm $\overline{V}$ using Eq. (7).

$$\overline{V} = \frac{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \sqrt{{u_{ij}^*}^2 + {v_{ij}^*}^2}}{MN} \tag{7}$$

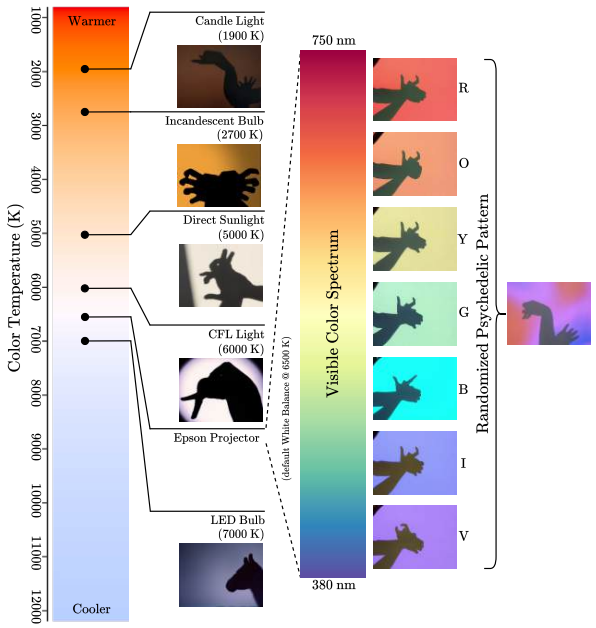


Figure 5. Light sources for background diversity in HASPER.

If $\overline{V} > \tau$, we retain the corresponding frame pair as candidate samples, otherwise we continue the process with the $(t + k)$th and the $(t + 2k)$th frames. Tab. 1 encapsulates some essential statistical information related to HASPER and provides a superficial overview of the dataset.

2.3. Labeling

After the extraction of the frames, the samples undergo manual scrutiny by 3 annotators who are pursuing undergraduate studies in Computer Science and Engineering (CSE). If a series of contiguous samples *prima facie* exhibit substantial similarity, we only keep a single image from that set of samples. The rest are discarded to avoid redundancy and to instill diversity. Another criterion that dictates the legitimacy of an image sample is its intelligibility. If the majority of 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 15 different directories of images, each containing the curated samples of a particular class. The images in these folders are then further partitioned into training and validation sets, maintaining a 60:40 split. 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.

3. Dataset Description

To provide a tangible exposition of the diverse samples in the dataset, Fig. 3 presents a collection of representative images across all 15 classes. With 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 and sharpness of the projected silhouettes, and a panoply of other aspects.

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

Figure 6. Samples with different silhouette properties.

3.1. Background Variance

The hand shadow puppetry setup that a puppeteer's crew arranges before the performance greatly dictates the nature of the background on 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 Figs. 3a and 3m. The angular directionality of the light also manifests a gradient effect on the background as can be seen in Figs. 3d and 3h. The temperature and color of the light emanated by the light sources onto the screens also add to the diversity. To achieve this, we use six different light sources—candlelight, an incandescent bulb, sunlight, a CFL light, an Epson EB-972 XGA projector, and an LED bulb—each with different color temperatures. Historically, many other light sources were used by shadowgraphists such as marrow-fat lamps, flame torches, halogen lamps, lime lights, etc. Fig. 5 depicts an overarching illustration of the monocular polychromatic background lighting diversity that we maintain in HASPER. We avail the overhead projector to emit light from across the visible spectral range (380–750 nm). Additionally, we use a random-patterned combination of these colors (colloquially referred to as the *psychedelic pattern*) as the backdrop for a subset of the organically created samples.

3.2. 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. As per the natural laws of optics, proximity to the light source yields crisp, well-defined shadows (*e.g.*, Fig. 6a), while increasing the distance fosters softer, more diffuse shadows (*e.g.*, Fig. 6b) with a central umbra and peripheral penumbra. The higher the contrast between the silhouettes and their respective backdrops, the more visible and well-contoured the shadow puppets are. The direction of the light source influences the orientation and shape of the shadows. Shadows cast by overhead lighting sources may appear elongated, while shadows cast by low-angle lighting sources may exhibit softer edges and less pronounced con-

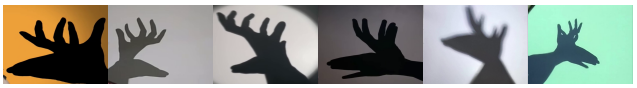


Figure 7. 'Deer' samples with different artistic representations.

trast and sharpness. Similarly, due to varied values of the lateral incident angle at which the light sources are kept relative to the screen's normal, we see horizontally elongated and compressed shadows. The shadows also differ in terms of the magnitude of their opacity, *i.e.*, the degree to which the hands prevent the transmission of light being projected onto the screen.

3.3. Puppeteers' Hand Anatomy and Stylistic Flair

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 of the wrists, palms, and digits of the puppeteers, along with the different stylistic choices they employ in their choreography, contribute as yet another avenue of diversity of the image samples in HASPER. Human beings, by nature, exhibit morphometric variations in finger length, palm width, and forearm thickness based on age and gender. As such, the cohort of novice shadowgraphists ($n = 6$) that we employ for the creation of amateur samples comprises a balanced gender representation of 3 males and 3 females, spanning an age range from 9 to 25 years. Among the adults, the hand anthropometric measurements are of 18.75 ± 1.55 cm in length and 8.66 ± 0.77 cm in width. For the minors, who obviously have proportionally smaller hand dimensions, these measurements are 14.23 ± 1.16 cm and 6.73 ± 0.82 cm respectively. The gender representation among the 14 professional shadow puppeteers is however imbalanced, with 12 male and 2 female shadowgraphists. Fig. 7 pristinely demonstrates the morphological variations of hand shadow puppets belonging to the 'Deer' class due to anatomical and stylistic diversity.

3.4. Comparative Analysis

3.4.1. 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. Figs. 3e and 3j are prime examples of such structural similitude that can be observed between the 'Deer' and 'Moose' classes. These similarities make the image classification task on HASPER quite challenging and culminate to being the reason behind a lot of misclassifications, as discussed in Sec. 4.3.

3.4.2. 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

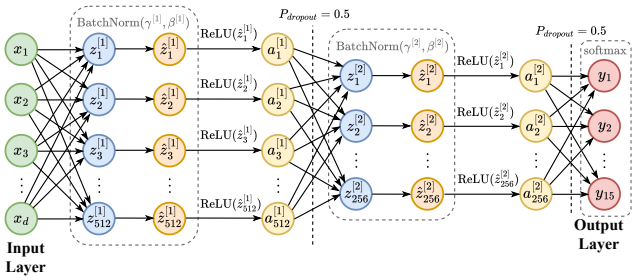


Figure 8. Classifier block attached to the tail-end of the pre-trained models. Here, d is the feature dimension of the anterior model. The output features of layer l is $z^{[l]} = W^{[l]} a^{[l-1]}$, where $a^{[l-1]}$ denotes the activation values of the preceding $(l-1)$th layer. The batch normalized 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 $g = \mathrm{ReLU}$. The predicted probabilities are determined by using the $\mathrm{softmax}$ function on the logits, *i.e.*, $P(\hat{y}_i = 1|x) = \mathrm{softmax}(y)_i = \frac{e^{y_i}}{\sum_{j=1}^{15} e^{y_j}}$

quasi-disparate samples, along with the individualistic flair that manifests through the puppeteers' stylistic choices, a particular class may show a lot of intra-class dissimilarity. As aforementioned, Fig. 7 portrays the heterogeneity of this nature among the samples from the 'Deer' class.

3.5. Statistical Analysis

Tab. 1 presents the statistical properties of the HASPER dataset. It tabulates the proportion of samples belonging to each of the 15 classes and their corresponding training-validation splits. While searching for hand shadow puppetry performances, we anecdotally observe that certain classes of puppets are more popular than others. Irrespective of this fact, we make sure all 15 classes of puppets are equitably represented in our dataset, with each class having 1,000 image samples ($\approx 6.\dot{6}\%$). As evident in Tab. 1, the image samples are evenly distributed across all 15 classes. The proportions of professionally sourced samples belonging to the 'Llama' and 'Snake' classes (14% and 27.3% respectively) are slightly low due to a scarcity of performance clips starring hand shadow puppets of these classes. For cases such as these, we supplement the classes with samples organically created by our novice shadowgraphist cohort. Each class has $\approx 47.827 \pm 1.414\%$ samples from clips of professional performers, and the rest $\approx 52.172 \pm 1.414\%$ samples are sourced from amateur clips. Of the 15,000 samples, approximately 76.53% feature male hands, while 23.64% represent female hands. In tandem, considering the target demographic of ombromanie, around 28.33% of the samples in HASPER consist of children's hands. In totality, we end up with 9,000 samples in the training set and 6,000 samples in the validation set, thereby partitioning HASPER by maintaining a 60:40 ratio.

Models	Params.	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 [38]	2.3M	61.73	78.41	86.10	0.6559	0.6173	0.5970	88.73	93.98	96.10	0.8995	0.8873	0.8853
ViTB16 [9]	86.6M	69.71	77.60	83.28	0.7276	0.6972	0.6969	68.88	76.65	81.36	0.7192	0.6868	0.6851
ViTL32 [9]	306.5M	85.10	91.56	94.48	0.8720	0.8510	0.8509	84.71	91.80	94.08	0.8632	0.8472	0.8465
AlexNet [27]	61.1M	87.01	93.61	95.46	0.8840	0.8702	0.8708	88.18	92.58	94.80	0.8887	0.8818	0.8809
SqueezeNet1_1 [24]	1.2M	87.56	92.45	94.15	0.8880	0.8757	0.8744	86.21	92.48	94.65	0.8754	0.8622	0.8637
MobileNetV3Small [18]	2.5M	89.48	94.31	95.76	0.9038	0.8948	0.8942	89.85	94.35	96.48	0.9082	0.8985	0.8976
SwinB [32]	87.8M	90.50	95.38	97.40	0.9128	0.9050	0.9042	90.20	95.40	97.08	0.9097	0.9020	0.9006
GoogLeNet [58]	6.6M	90.73	94.65	95.70	0.9105	0.9073	0.9059	92.18	95.65	96.58	0.9283	0.9218	0.9206
ResNet18 [15]	11.7M	90.91	95.28	96.60	0.9176	0.9092	0.9069	91.25	95.43	97.05	0.9229	0.9125	0.9119
MobileNetV3Large [18]	5.5M	91.20	94.48	95.98	0.9185	0.9120	0.9110	90.40	94.53	95.26	0.9147	0.9040	0.9024
ConvNeXt [34]	88.6M	91.46	96.33	98.05	0.9220	0.9147	0.9140	92.55	96.36	97.96	0.9306	0.9255	0.9246
SwinV2B [33]	87.9M	91.58	96.25	97.61	0.9210	0.9158	0.9151	91.48	96.00	97.55	0.9209	0.9148	0.9144
VGG16 [52]	138.4M	91.61	95.08	96.65	0.9248	0.9162	0.9168	91.00	95.21	96.45	0.9235	0.9100	0.9119
MNasNet13 [62]	6.3M	91.66	95.65	97.01	0.9240	0.9167	0.9149	91.45	95.86	97.26	0.9231	0.9145	0.9133
ConvNextLarge [34]	197.8M	91.88	95.90	97.70	0.9254	0.9188	0.9181	88.00	94.70	96.56	0.8942	0.8800	0.8782
EfficientNetB0 [60]	5.3M	91.93	95.26	96.71	0.9257	0.9193	0.9178	90.40	93.75	95.10	0.9131	0.9040	0.9022
MaxViT [68]	30.9M	92.01	96.50	97.81	0.9268	0.9202	0.9214	92.08	95.98	97.36	0.9320	0.9208	0.9237
EfficientNetV2S [61]	21.5M	92.31	95.75	96.76	0.9375	0.9232	0.9245	**94.45**	**97.35**	**98.30**	0.9498	**0.9445**	0.9438
VGG19 [52]	143.7M	92.36	95.13	96.10	0.9354	0.9237	0.9242	91.80	95.06	96.15	0.9296	0.9180	0.9187
MobileNetV2 [48]	3.5M	92.38	94.98	96.05	0.9303	0.9238	0.9233	92.31	95.38	96.91	0.9311	0.9232	0.9225
WideResNet50_2 [73]	68.9M	92.46	96.28	97.28	0.9331	0.9247	0.9235	93.35	95.73	97.15	0.9421	0.9335	0.9330
ResNet50 [15]	25.6M	92.58	95.56	96.75	0.9332	0.9258	0.9252	93.08	96.48	97.20	0.9363	0.9308	0.9299
RegNetX32GF [45]	107.8M	92.86	95.71	96.93	0.9348	0.9287	0.9269	92.91	95.71	96.95	0.9366	0.9292	0.9282
DenseNet121 [21]	8.0M	92.93	95.75	96.88	0.9367	0.9293	0.9282	92.95	95.51	96.56	0.9360	0.9295	0.9285
ResNext101_32X8D [70]	88.8M	93.00	96.41	97.23	0.9364	0.9310	0.9303	94.20	96.61	97.58	**0.9520**	0.9420	0.9423
WideResNet101_2 [73]	126.9M	93.36	95.81	96.90	0.9423	0.9337	0.9332	92.73	96.35	97.63	0.9337	0.9273	0.9267
InceptionV3 [59]	27.2M	93.50	96.48	97.35	0.9401	0.9350	0.9338	93.71	96.36	97.06	0.9446	0.9372	0.9371
DenseNet201 [21]	20.0M	93.56	95.78	96.73	0.9450	0.9357	0.9353	94.43	97.00	97.61	0.9492	0.9443	**0.9442**
ResNet101 [15]	44.5M	93.81	96.23	97.71	0.9432	0.9382	0.9406	93.23	96.93	98.13	0.9386	0.9323	0.9321
ResNet152 [15]	60.2M	94.06	97.06	98.05	0.9447	0.9407	0.9394	93.05	96.73	97.48	0.9374	0.9305	0.9297
ResNet34 [15]	21.8M	**94.97**	**97.23**	**98.23**	**0.9516**	**0.9497**	**0.9491**	91.98	95.95	97.20	0.9266	0.9198	0.9189
ResNet34 w/ Silhouette Polygonization	21.8M	92.72	96.41	97.51	0.9328	0.9272	0.9257	92.95 (+1.05%)	95.75	96.61	0.9352 (+0.93%)	0.9295 (+1.05%)	0.9283 (+1.02%)
ResNet34 w/ Topological Features	21.8M	93.72	96.43	97.78	0.9432	0.9372	0.9359	94.05 (+2.25%)	96.45 (+0.52%)	97.53 (+0.34%)	0.9476 (+2.27%)	0.9405 (+2.25%)	0.9401 (+2.31%)

Table 2. Performance comparison of the vanilla and modified versions of the image classification models.

4. Methodology for Benchmarking HASPER

A series of pretrained models are used as feature extractors to develop a benchmark for the dataset. The models are pretrained on the ImageNet [8] dataset and fine-tuned on HASPER. We implement the training pipeline using the Pytorch[7] framework. This section presents an overview of the models, evaluation metrics, and experimental results.

4.1. Experimental Setup

4.1.1. Baseline Models

For this classification task, we use 31 feature extractor models as baselines, which are listed in Tab. 2. Some of these models have a track record of good performance across various other image classification tasks [41]. We examine both conventional Convolutional Neural Networks (CNNs) and CNNs augmented 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. We fuse silhouette-specific features obtained via topological descriptors [3] and polygonization [10] with the extracted features from the best-performing vanilla model (see Appendix B for more information).

4.1.2. Classifier Network

We adopt two approaches to arrive at the final 15-dimensional layer since there are a total of 15 classes to predict from. The first approach is to directly append a 15-dimensional fully connected layer at the tail-end of the vanilla models. The second approach incorporates the classifier block portrayed in Fig. 8.

[7] https://pytorch.org/vision/stable/models.html

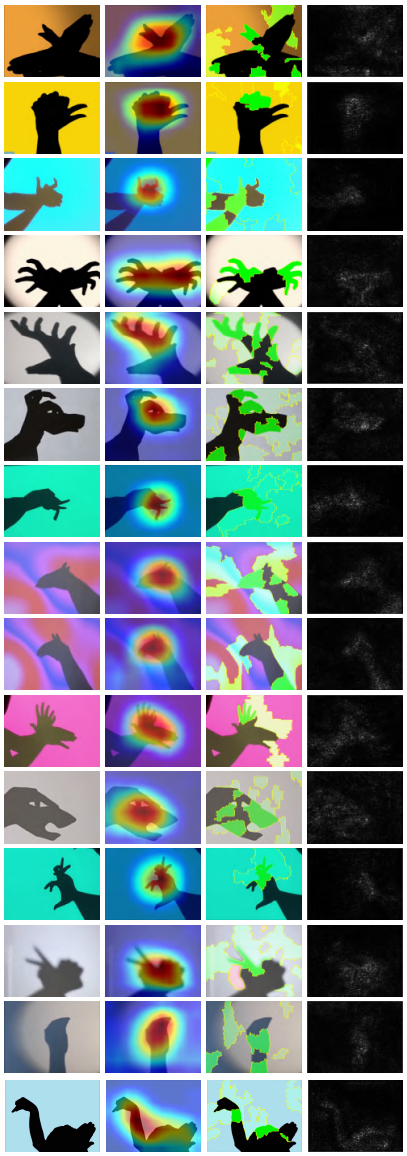
4.2. Results and Findings

4.2.1. Performance Analysis

The ResNet34 model yields the best performance with a top-1 accuracy of 94.97%. The vanilla version of the model also yields the highest top-2 accuracy, top-3 accuracy, Precision, Recall, and F1-scores of 97.23%, 98.23%, 0.9516, 0.9497, and 0.9491 respectively. Upon being equipped with the classifier block shown in Fig. 8, the EfficientNetV2S model yields the highest top-k accuracies and Recall. In contrast, the ResNext101_32X8D and DenseNet201 models demonstrate the best performance across Precision and F1-score metrics respectively. At this recess of the performance analysis, we consider the top-1 accuracy metric to be the most statistically significant metric. As evident in Tab. 2, the vanilla models listed in the upper part's penultimate row and above lag behind the ResNet34 model when it comes to the top-1 accuracy value (as well as the other metrics), which is why we adjudicate that ResNet34 is the best-performing model. We hypothesize that residual connections in ResNets help preserve low-level edge and contour information through identity mappings, ensuring that crucial silhouette boundaries aren't lost as the network deepens, making them better at capturing subtle variations.

4.2.2. Qualitative Analysis and Explainability

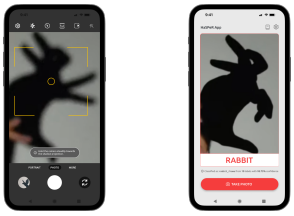
As depicted in Fig. 9, we adopt a plethora of explainable AI (XAI) techniques for the best-performing ResNet34



(a) Raw Image (b) GradCAM Heatmap (c) LIME (d) Saliency Map

Figure 9. The juxtaposition of original image samples from the HASPER dataset with their corresponding GradCAM Heatmaps, LIME Visualizations, and Saliency Maps (for the best-performing vanilla RESNET34 model).

model to understand its decision-making. While viewing the GradCAM (Gradient-weighted Class Activation Mapping) [50] attention heatmaps, it becomes apparent that the model puts more gravitas on the common-sense distinguishing traits. For example, in Fig. 9b, 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



(a) Camera feed (b) Prediction

Figure 10. Android application for shadow puppet recognition.

elephant, the long maxilla-mandibular jaw of a horse, the long-eared and tapered head of a llama, the upright horns of a moose, the big eyes and small ears of a panther, the petite hands and head of a rabbit, the shell and antennae of a snail, the lateral hood expansion of a snake, as well as the slender neck and wing feathers of a swan. As human beings, we evoke these same distinguishing characteristics while classifying the images using our own visual reasoning faculties. As exemplified in Fig. 9c, for local interpretation, we use the model-agnostic technique called LIME (Local Interpretable Model-agnostic Explanations) [46]. The green highlights indicate regions of the image that contribute positively to the probability of the assigned label, while the red highlights signify areas that reduce this probability. We also demonstrate the spatial support of the top-1 predicted classes by generating the saliency maps [53] in Fig. 9d. These maps are rendered using a solitary back-propagation pass through the RESNET34 model, and they accentuate the salient areas of the given image, characterized by their discriminative attributes with respect to the given class.

4.2.3. Practicality Analysis as a Teaching Tool

It is noteworthy to point out that MOBILENETV2, with only 3.5 million parameters, managed to surpass many 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 MOBILENETV2 in classifying hand shadow puppet images from the phone's camera feed. In order to make the prototype work as seamlessly as intended, we make sure the vicinity is well-lit, and the camera accurately captures a sharply focused silhouette. We find that the model has a memory footprint of 29 MB and achieves an average inference time of 880 μs on the Snapdragon 8 Gen 2 mobile chipset featured in the Samsung Galaxy S23 smartphone. Fig. 10 portrays the snapshots of the prototype application. There are several other practical implementation challenges involved in this endeavor that we can identify for an educational mobile application to comport well with the target demographics [2] and real-world settings.

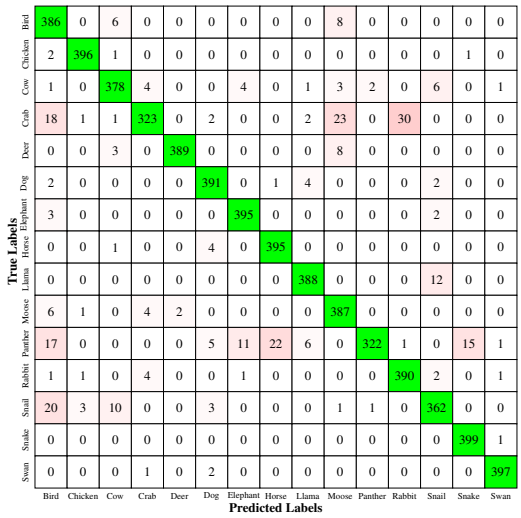
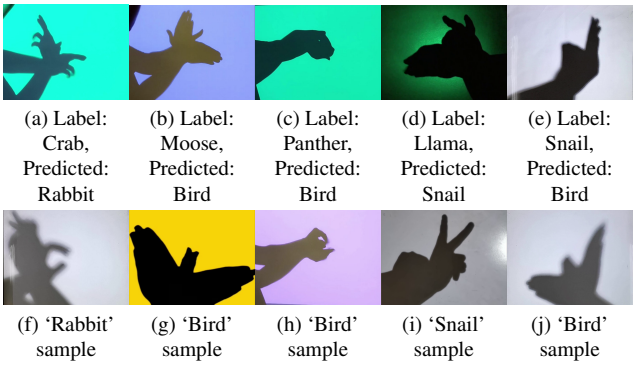


Figure 11. Confusion Matrix of vanilla RESNET34.

4.3. Error Analysis

The confusion matrix for the RESNET34 model on our dataset, presented in Fig. 11, reveals that the 'Crab' class exhibits the highest count of misclassifications. One obvious reason for this is the somewhat significant inter-class similarity among the 'Bird', 'Moose', 'Rabbit', and 'Crab' 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 this image classification task, as evident from the other pale-red entries of the confusion matrix in Fig. 11. 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. Figs. 12a and 12f show the confusion between a 'Crab' sample and a 'Rabbit' sample which look visually quite similar. The same holds for one of the six 'Moose' samples that are misclassified as 'Bird' samples by the RESNET34 model, as depicted in Figs. 12b and 12g. We observe that misclassifications of this type occur when images belonging to different, but visually akin categories, are erroneously assigned to the wrong class. Another reason for misclassifications is the ambiguity of shape present in mid-action frames. For instance, the 'Bird' sample in Fig. 12h is a transition frame between two successive wing flaps. However, due to the presence of this ambiguous sample in the training set of the 'Bird' class, the RESNET34 ends up misclassifying the 'Panther' sample in Fig. 12c as a 'Bird' sample. The misclassification portrayed by the pair of Figs. 12d and 12i is due to the combination of poor lighting and ineptitude of the amateur child puppeteer in creating 'Llama' shadows. The model confuses the ear protrusions of the 'Llama' sample to be the tentacular eyes of a 'Snail' sample. For Fig. 12e, we can pontificate the misclassification reasons to be the overlapping of the fingers and the distorted angle at which the sample was captured. The 'Snail' sample thereby gets wrongly classified as a 'Bird' sample due to the existence of the analogous sample portrayed in Fig. 12j. The green entries along the diagonal of the confusion matrix in Fig. 11 indicate the reasonably good classwise prediction performance of the RESNET34 model, which is, to some extent, due to the perfectly balanced sample distribution in HASPER.



(a) Label: Crab, Predicted: Rabbit (b) Label: Moose, Predicted: Bird (c) Label: Panther, Predicted: Bird (d) Label: Llama, Predicted: Snail (e) Label: Snail, Predicted: Bird

(f) 'Rabbit' sample (g) 'Bird' sample (h) 'Bird' sample (i) 'Snail' sample (j) 'Bird' sample

Figure 12. Misclassified samples with visually similar samples of the predicted class.

5. Conclusion and Future Work

This paper introduces HASPER, a 15,000-image dataset for hand shadow puppet recognition, curated from expert and amateur performances via optical flow-based frame extraction. We establish a benchmark by fine-tuning 31 pretrained image classification models on the dataset. We analyze the performance of our top-performing model, RESNET34, by visualizing its feature space using t-SNE and conducting comprehensive qualitative and error analyses. 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 publicly available dataset and study on image classification benchmarking that focuses only on ombromanie. There are many avenues in our work that warrant further investigation. We hope to reconcile those desiderata by enriching our dataset with numerous permutations of arm positions and finger movements, preferably by employing more skilled individuals with varying palm and wrist structures, thereby creating more diverse silhouettes. We also plan to experiment with a gesture detection technology such as MediaPipe[8] or Microsoft Kinect[9] for leveraging depth coordinates of hand landmarks [40], and assess their efficacy in classifying hand shadow puppets.

[8] MediaPipe — developers.google.com/mediapipe

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

6. Acknowledgments

We convey our heartfelt gratitude to the anonymous reviewers for their constructive criticisms and insightful feedback, which were surely 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, and the Department of Research, Extension, Advisory Services, and Publications (REASP) for funding our travel expenses. Additionally, we wish to acknowledge Shahriar Ivan, Department of Computer Science and Engineering, IUT, for his assistance in proofreading and offering a preliminary review of this manuscript. We further thank Mohammad Ishrak Abedin and Reaz Hassan Joader, from the same department, for their help in polishing this paper's illustrations and diagrams. Syed Rifat Raiyan, in particular, wants to thank his parents, Syed Sirajul Islam and Kazi Shahana Begum, for everything.

References

[1] Albert Almoznino and Y Pinas. The art of hand shadows, 2002. 1

[2] Najwa Altwaijry and Isra Al-Turaiki. Arabic handwriting recognition system using convolutional neural network. *Neural Computing and Applications*, 33(7):2249–2261, 2020. 7, 14

[3] Harry Blum. A Transformation for Extracting New Descriptors of Shape. In *Models for the Perception of Speech and Visual Form*, pages 362–380. The MIT Press, Cambridge, MA, 1967. 6, 13

[4] Alexander Dylan Bodner, Antonio Santiago Tepsich, Jack Natan Spolski, and Santiago Pourteau. Convolutional kolmogorov-arnold networks. *arXiv preprint arXiv:2406.13155*, 2024. 15

[5] M. Brand. Shadow puppetry. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, page 1237–1244, Corfu, Greece, 1999. IEEE. 12

[6] Benjamin M Carr and Guy J Brown. Shadow puppetry using the kinect. *The University of Sheffield*, 2014. 12

[7] R.T. Collins, R. Gross, and Jianbo Shi. Silhouette-based human identification from body shape and gait. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, page 366–371, Washington D.C., USA, 2002. IEEE. 12

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, Florida, USA, 2009. IEEE. 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6

[10] David H. Douglas and Thomas K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10(2):112–122, 1973. 6, 12, 13

[11] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI, 2008. 13

[12] Shang-zhen Gao. On the digital development of the chinese shadow play art. In *2011 International Conference on Internet Technology and Applications*, page 1–4, Wuhan, China, 2011. IEEE. 12

[13] Ugur Güdükbay, Fatih Erol, and Nezih Erdogan. Beyond tradition and modernity: Digital shadow theater. *Leonardo*, 33(4):264–265, 2000. 12

[14] Oğul Göçmen and Murat Emin Akata. Polygonized silhouettes and polygon coding based feature representation for human action recognition. *IEEE Access*, 11:57021–57036, 2023. 15

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016. IEEE. 6

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, 2020. 12

[17] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1–3):185–203, 1981. 2

[18] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le. Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 1314–1324, Seoul, Korea, 2019. IEEE. 6

[19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv 1704.04861*, 2017. 12

[20] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8(2):179–187, 1962. 13

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017. IEEE. 6

[22] Mark Huang, Shishir Mehrotra, and Flavia Sparacino. Shadow vision. 1999. 12

[23] Zhe Huang, Vamshi Krishna Madaram, Saad Albadrani, and Tam V. Nguyen. Shadow puppetry with robotic arms. In *Proceedings of the 25th ACM international conference on Multimedia*, page 1251–1252, Mountain View, California, USA, 2017. ACM. 12

[24] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *ArXiv*, abs/1602.07360, 2016. 6

[25] Itseez. Open source computer vision library. https://github.com/itseez/opencv, 2015. Accessed: 2025-6-13. 13

[26] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952. 14

[27] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014. 6, 12

[28] Louisa Lam, Seong-Whan Lee, and Ching Y. Suen. Thinning Methodologies – A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(9): 869–885, 1992. 13

[29] Hui Liang, Jian Chang, Shujie Deng, Can Chen, Ruofeng Tong, and Jianjun 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*, page 63–72, Kobe, Japan, 2015. ACM. 12

[30] Hui Liang, Jian Chang, Ismail K. Kazmi, Jian J. Zhang, and Peifeng Jiao. Hand gesture-based interactive puppetry system to assist storytelling for children. *The Visual Computer*, 33(4):517–531, 2016. 12

[31] Mariana Dória Prata Lima, Gilson Antonio Giraldi, and Gastão Florêncio Miranda Junior. Image classification using combination of topological features and neural networks. *arXiv preprint arXiv:2311.06375*, 2023. 15

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 9992–10002, Montreal, QC, Canada, 2021. IEEE. 6

[33] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 11999–12009, New Orleans, LA, USA, 2022. IEEE. 6

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022. IEEE. 6

[35] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024. 15

[36] Fei Lu, Feng Tian, Yingying Jiang, Xiang Cao, Wencan Luo, Guang Li, Xiaolong Zhang, Guozhong Dai, and Hongan Wang. Shadowstory: creative and collaborative digital storytelling inspired by cultural heritage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 1919–1928, Vancouver BC, Canada, 2011. ACM. 2, 12

[37] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc. 3

[38] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Computer Vision – ECCV 2018*, pages 122–138, Cham, 2018. Springer International Publishing. 6

[39] Anne-Sofie Maerten and Derya Soydaner. From paintbrush to pixel: A review of deep neural networks in ai-generated art. *arXiv preprint arXiv:2302.10913*, 2023. 2

[40] Hasan Mahmud, Mashrur M. Morshed, and Md. Kamrul Hasan. Quantized depth image and skeleton-based multimodal dynamic hand gesture recognition. *The Visual Computer*, 40(1):11–25, 2023. 8

[41] José Maurício, Inês Domingues, and Jorge Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9):5521, 2023. 6

[42] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2–3):90–126, 2006. 12

[43] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. 13

[44] Dwi Puji Prabowo, M. Kuncoro Aji Nugraha, Dimas Irawan Ihya' Ulumuddin, Ricardus Anggi Pramunendar, and Stefanus Santosa. Indonesian traditional shadow puppet classification using convolutional neural network. In *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, page 1–5, Semarang, Indonesia, 2021. IEEE. 12

[45] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollar. Designing network design spaces. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020. IEEE. 6

[46] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, 2016. Association for Computational Linguistics. 7

[47] Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4695–4703, 2024. 2

[48] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 4510–4520, Salt Lake City, UT, USA, 2018. IEEE. 6

[49] R Saritha. An artist nurturing a dying art and his quest for its conservation. *YourStory*, 2017. 2

[50] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks

via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, page 618–626, Venice, Italy, 2017. IEEE. 7

[51] Jean Serra. *Image Analysis and Mathematical Morphology*. Academic Press, London, 1982. 13

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6, 12

[53] K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proceedings of the ICLR*, 2014. 7

[54] Shafic A. Sraj. The hand in art: Shadowgraphy—a display of hand shadows as a performing art. *The Journal of Hand Surgery*, 37(4):817, 2012. 1

[55] Ida Bagus Kresna Sudiatmika and I Gusti Ayu Agung Sari Dewi. Indonesian shadow puppet recognition using vgg-16 and cosine similarity. *The IJICS (International Journal of Informatics and Computer Science)*, 5(1):1–6, 2021. 12

[56] Ida Bagus Kresna Sudiatmika, Pranowo, and Suyoto. Indonesian traditional shadow puppet image classification: A deep learning approach. In *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, page 130–135, Kuta, Bali, Indonesia, 2018. IEEE. 12

[57] Ida Bagus Kresna Sudiatmika, Made Artana, Nengah Widya Utami, Made Adi Paramartha Putra, and Eka Grana Aristyana Dewi. Mask r-cnn for indonesian shadow puppet recognition and classification. *Journal of Physics: Conference Series*, 1783(1):012032, 2021. 12

[58] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1–9, Boston, MA, USA, 2015. IEEE. 6

[59] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2818–2826, Las Vegas, NV, USA, 2016. IEEE. 6

[60] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114, Long Beach, California, USA, 2019. PMLR. 6

[61] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10096–10106, Virtual, 2021. PMLR. 6

[62] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2815–2823, Long Beach, CA, USA, 2019. IEEE. 6

[63] Zhichuan Tang, Yidan Hu, Weining Weng, Lekai Zhang, Lingtao Zhang, and Jichen Ying. An intelligent shadow play system with multi-dimensional interactive perception. *International Journal of Human–Computer Interaction*, 39(6): 1314–1326, 2022. 12

[64] Tsun-Hung Tsai and Lai-Chung Lee. A study of using contactless gesture recognition on shadow puppet manipulation. *ICIC express letters: an international journal of research and surveys*, 7(11):2317–2322, 2016. 12

[65] Amato Tsuji and Keita Ushida. Telecommunication using 3dcg avatars manipulated with finger plays and hand shadow. In *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, page 39–40, Kyoto, Japan, 2021. IEEE. 12

[66] Amato Tsuji, Keita Ushida, and Qiu 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*, page 441–444, Tokyo, Japan, 2018. ACM.

[67] Amato Tsuji, Keita Ushida, Saneyasu Yamaguchi, and Qiu 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)*, page 1195–1196, Osaka, Japan, 2019. IEEE. 12

[68] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *Computer Vision – ECCV 2022*, pages 459–479, Cham, 2022. Springer Nature Switzerland. 6

[69] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. http://jmlr.org/papers/v9/vandermaaten08a.html. 14

[70] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017. IEEE. 6

[71] Zifei Yan, Ziyuan Jia, Yuehua Chen, and Haolun Ding. The interactive narration of chinese shadow play. In *2016 International Conference on Virtual Reality and Visualization (ICVRV)*, page 341–345, Los Alamitos, California, USA, 2016. IEEE. 12

[72] C. Zach, T. Pock, and H. Bischof. *A Duality Based Approach for Realtime TV-L 1 Optical Flow*, page 214–223. Springer Berlin Heidelberg, 2007. 3

[73] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016*, pages 87.1–87.12, York, UK, 2016. British Machine Vision Association. 6

[74] David Zhang and Guangming Lu. Review of Shape Representation and Description Techniques. *Pattern Recognition*, 37(1):1–19, 2004. 12

[75] Hui Zhang, Yuhao Song, Zhuo Chen, Ji Cai, and Ke Lu. Chinese shadow puppetry with an interactive interface using the kinect sensor. In *Computer Vision – ECCV 2012. Workshops and Demonstrations*, pages 352–361, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 12