2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 979-8-3503-1892-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/WACV57701.2024.00186

Few-shot Shape Recognition by Learning Deep Shape-aware Features

Wenlong Shi[⋆,▽], Changsheng Lu[⋆,◇], Ming Shao[†,§], Yinjie Zhang[▽], Siyu Xia[†,▽], Piotr Koniusz[†,◇,♣]
[▽]School of Automation, Southeast University [◇]The Australian National University
[§]University of Massachusetts, Dartmouth [♣]Data61/CSIRO
{wenlong.shi, xsy}@seu.edu.cn, {changsheng.lu, piotr.koniusz}@anu.edu.au, mshao@umassd.edu

Abstract

Traditional shape descriptors have been gradually replaced by convolutional neural networks due to their superior performance in feature extraction and classification. The state-of-the-art methods recognize object shapes via image reconstruction or pixel classification. However, these methods are biased toward texture information and overlook the essential shape descriptions, thus, they fail to generalize to unseen shapes. We are the first to propose a few-shot shape descriptor (FSSD) to recognize object shapes given only one or a few samples. We employ an embedding module for FSSD to extract transformation-invariant shape features. Secondly, we develop a dual attention mechanism to decompose and reconstruct the shape features via learnable shape primitives. In this way, any shape can be formed through a finite set basis, and the learned representation model is highly interpretable and extendable to unseen shapes. Thirdly, we propose a decoding module to include the supervision of shape masks and edges and align the original and reconstructed shape features, enforcing the learned features to be more shape-aware. Lastly, all the proposed modules are assembled into a few-shot shape recognition scheme. Experiments on five datasets show that our FSSD significantly improves the shape classification compared to the state-of-the-art under the few-shot setting.

1. Introduction

Shape recognition has been a critical area of research in computer vision, with applications in numerous fields, such as industrial automation [47, 81], botanics classification [68], and fine-grained shape recognition for medical organs [80]. In industrial automation, it is usually required to identify the textureless components that possess a unique shape, *e.g*. the workpiece classification. In botany science, diverse herbs collected in the wild need to be classified and made taxonomy. Moreover, the shape recognition is useful to identify lesions or pathological changes in medical diagnosis [80].

As demonstrated in the variety of recognition tasks, convolutional neural networks (CNN) allow high-dimensional image data to be presented as semantic features. Such representations capture both shape and texture information. While existing studies argue that shape information plays dominant roles in general recognition tasks [36, 40, 62, 83], many studies highlight that local textures provide adequate information for various vision tasks [4, 8, 17, 19]. There are ongoing efforts to combine texture and shape information [25, 34]. Nonetheless, the majority of works focus on utilizing *texture information* to capture object shapes in object segmentation [44, 63], arbitrary-shaped text detection [9, 58, 72, 87], salient object detection [55, 59, 74, 77, 86], circle/ellipse detection [48, 49, 71], and shape classification [1, 2, 37, 84]. However, these texture-based methods do not target extracting generic shape information suitable for shape recognition of common textureless objects or novel textureless objects.

Finding discriminative representations for object shapes is non-trivial, and no ultimate mathematical definition has ever been established so far. Traditional shape descriptors mathematically approximate the geometric information of shapes and demonstrate their efficacy empirically [66]. These methods enable fast computation while maintaining high accuracy. Moreover, explicit mathematical definitions produce interpretable descriptors, and therefore, are still widely used today.

Zeiler *et al*. [83] argue that CNN features carry limited shape information. To address this issue, the approaches proposed by [22, 32, 41, 61, 64] attempt to combine dedicated shape descriptors with well-established CNN features so that the learned representation can carry more shape information. These methods provide some degree of interpretable modeling and representations. Nonetheless, these methods model semantic and shape representations separately, *i.e*., the feature integration is applied right before

[⋆]Co-first author, [†]co-correspondence. [§]This material is based upon work supported by the National Science Foundation under Grant No. 2144772.

2642-9381/24/$31.00 ©2024 IEEE
DOI 10.1109/WACV57701.2024.00186

output layers without further coupling in earlier layers.

In this paper, we formulate a generic shape representation for few-shot shape recognition. In contrast to existing representations, our model couples the shape descriptions with deep learning for few-shot learning. Our FSSD model is based on a *matching network* for few-shot learning, but with three unique contributions. Firstly, we propose to use group equivariant convolutional neural networks (G-CNN) instead of regular convolutional networks to help the embedding module handle rotations of input. Secondly, to encourage networks to *learn* interpretable shape features comparable to conventional *hand-crafted* shape descriptions, we propose novel shape primitives learned through dual attention architecture. The shape primitives serve as the basis to describe various shapes and help "explain out" the formulation of each input shape through reconstruction and visualization. Thirdly, we incorporate supervision through pairwise decoders to align the original and reconstructed shape features. This ensures the fidelity of shape primitives and the efficacy of the learned shape representation.

The main contributions of our work are as follows:

i. To our best knowledge, we propose the first few-shot shape recognition model capable of recognizing seen and unseen geometric shapes.
ii. We propose a dual attention architecture to learn shape primitives to improve shape representation learning. We add supervision through shape masks and the edge of objects to guide shape-aware feature learning.
iii. We collect a dedicated shape recognition dataset and transform an existing image dataset into shapes. We demonstrate that the proposed approach outperforms the state-of-the-art on five datasets.

2. Related works

2.1. Shape Descriptors

Defining the shape of an object is a challenging task, and current methods for shape descriptions generally focus on the following factors [66]: i) input representation (*i.e*., the description of an object can be based on its boundary or on its whole region); ii) ability to reconstruct the object; iii) ability to recognize incomplete shapes, and iv) robustness to various transformations such as translation, rotation, and scale. Therefore, most shape description methods can be classified into two categories: i) boundary-based methods and ii) region-based methods. Boundary-based methods include Hough transform [3, 13, 23, 33, 39, 56, 57, 60], Fourier descriptors [20, 82], curvature scale-space descriptors [26, 53], shape context [6], segment boundary descriptors [35], wavelet transform descriptors [10] and more [18,29,38,54,73,75,79]. Region-based methods include angular radial transformation [7, 28, 78], image moment [16], and general Fourier shape descriptors [85].

Unlike existing works, this paper presents a deep learning pipeline formed from the perspective of shape description. Such a network can learn boundary-based generic shape features that are robust to translations and rotations, and can reconstruct object shapes from the basis set.

2.2. Few-shot Learning (FSL)

FSL can perform image classification, object detection, or segmentation given limited training data. Existing FSL works can be divided into two categories: i) metric-based methods (*e.g*., Prototypical Networks, Relation Networks); and ii) optimization-based methods such as MAML. Relation Networks [67] introduce a relation module to learn the similarity between the features of two input samples. Prototypical Networks [65] map a set of samples per class into a prototype vector. Then, classification is performed by measuring the cosine similarity between query samples and prototypes. MAML [15] trains models iteratively on multiple tasks.

While existing FSL pipelines target various vision tasks, *e.g*., few-shot object detection [14, 30], few-shot segmentation [42, 43], and few-shot keypoint detection [46, 50], few-shot shape recognition and deep shape-aware features are still under-explored. We argue and show that good shape descriptors are essential for good performance compared to other factors, *e.g*., distance metric in similarity measurement.

2.3. Shape Recognition

Most shape recognition methods utilize simple CNNs to classify binary shape images [1,2,37,84]. Some other works encode and map shapes into high-dimensional spaces for classification [27, 52]. In shape detection, special detection frameworks are used instead of bounding boxes. For example, Kang *et al*. [31] proposed the use of bounding masks to regress the object edges, while ellipse detection [12, 71] utilizes elliptical bounding boxes to detect the most common ellipse shapes in the natural world. Tasks such as arbitrary-shaped text detection and salient object detection are closely related to shape reconstruction. In arbitrary-shaped text detection [9, 58, 72, 87], the aim is to capture regions around various text shapes rather than using bounding boxes. Salient object detection [55, 59, 74, 77, 86] focuses on reconstructing more accurate object boundaries.

This paper proposes a novel dual attention mechanism to learn a finite set of universal shape primitives. The learned set of primitives through known shapes can be extended to compose, represent, and interpret unseen shapes, enabling robust yet discriminant shape features for the classification of new shapes.

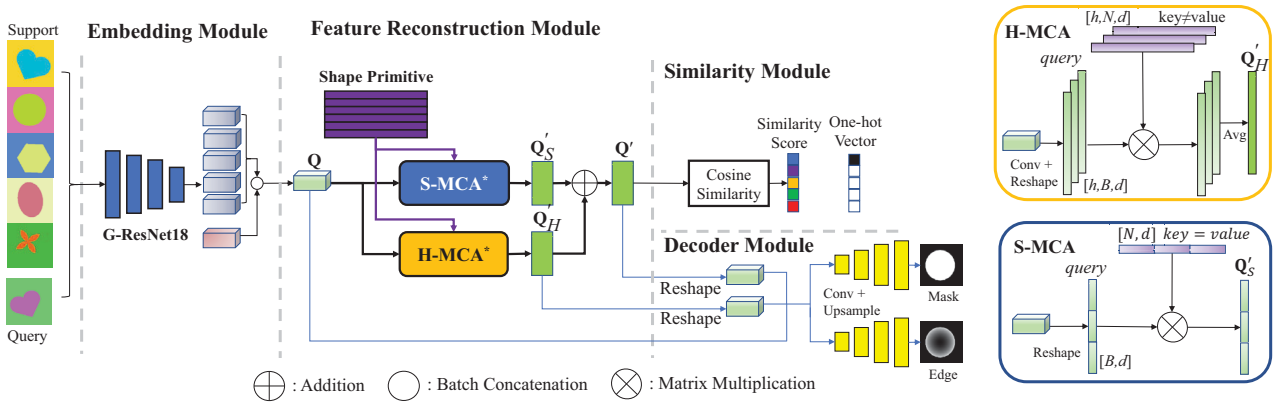

Figure 1. The overall architecture of the proposed FSSD consists of four modules (from left to right): embedding module, feature reconstruction module, similarity module, and decoder module. FSSD follows a few-shot learning pipeline, where feature reconstruction module learns a set of primitives for shapes and reconstructs the output features of embedding module using the input support set and query set. Feature reconstruction module consists of Holistic Multi-head Cross-Attention (H-MCA) and Standard Multi-head Cross-Attention (S-MCA). The pairwise decoders enables supervision through ground truth shape masks and edges to align original and reconstructed shape features. The similarity is calculated using the reconstructed support set features and query set features, enabling shape classification.

3. Proposed method

We develop a few-shot model which can focus on shape characteristics and maintain a high accuracy of shape classification. The entire model consists of four parts: (i) embedding module, (ii) feature reconstruction module, (iii) similarity module, and (iv) decoding module.

3.1. Problem Statement

This paper uses episode-based training [70] often used in few-shot learning. In each episode, the so-called c-way m_1-shot learning is applied. Specifically, c classes are randomly selected from the training set, with m_1 samples randomly selected from each class to form the support set $\mathcal{S} = (\mathbf{x}_i, y_i)_{i=1}^{m_1 \times c}$, and m_2 samples randomly selected from each class to form the query set $\mathcal{Q} = (\mathbf{x}_j, y_j)_{j=1}^{m_2 \times c}$.

A typical few-shot learning network consists of two parts: an embedding module and a similarity module. To accommodate the shape primitives and supervision, we introduce a novel feature reconstruction module and a decoding module. Firstly, the embedding module employs a group-equivariant convolutional network (G-CNN) that incorporates group transformations into the convolutional operations by rotation-equivariant operations. Secondly, a novel dual attention architecture is implemented to learn shape primitives for feature reconstruction. Thirdly, the similarity module uses the cosine similarity to calculate the similarity between the support set features and the query set features. Finally, the original features and the support/query reconstructed features are fed into the decoding module to recover the shape mask and edges. The overall architecture of our method is shown in Figure 1.

3.2. Embedding Module

Shape descriptors should be robust to geometric transformations such as translation, rotation, and scale changes. Recent works have shown that group-equivariant networks can capture objects better as they do not require learning separately each object pose. Given the transformation group g, group-equivariant networks [11] use the following operator:

$$[T_g \mathbf{f}] \otimes \Psi = T_g[\mathbf{f} \otimes \mathbf{\Psi}], \tag{1}$$

where $\mathbf{f}$ is the feature map or image, $\mathbf{\Psi}$ is the convolution filter, $\otimes$ is the convolution, and T_g indicates the proposed transformation groups in [11], $p_4 = \{r, r^2, r^3, e\}$ and $p_4^m = \{e, r, r^2, r^3, mr, r^3m, rm, mr^3, r^2m, mr^2, rmr, m\}$, where e represents an equivalent transformation, r represents a counterclockwise rotation of 90° and m represents a mirror transformation. Next, take the p_4 group as an example. Let $T_g \in p_4$, $\mathbf{f}$ be the image, and $\mathbf{\Psi}$ be a CNN filter. Then, the equivariance property means that applying T_g to $\mathbf{f}$ followed by convolution filter $\mathbf{\Psi}$ is identical to applying filter $\mathbf{\Psi}$ first to $\mathbf{f}$ followed by T_g in the feature domain. By using the equivariance property, backbone is "invariant" to transformations so it does not have to learn each object pose separately. By working in the equivariant feature domain rather than on the raw images, our similarity module then learns invariance on equivariant features by learning to match shapes in different poses. This process has been shown in Figure 1.

3.3. Feature Reconstruction Module

Discovering the basis or principal components has been popular practice in signal processing and representation

learning, *e.g.*, Fourier transform, principal component analysis (eigenfaces), low-rank matrix analysis, sparse coding. Inspired by such works, we let shape descriptors be approximated by a finite set of shape primitives. We propose to learn a finite set of shape primitives $\mathbf{\Phi}$ to ground the shape features of the images for both training and unseen data.

Attention-based Primitives. The support & query features in an episode are combined into a matrix $\mathbf{Q} \in \mathbb{R}^{B\times d}$, where B is the batch size (*i.e.*, the total number of support & query images) and d is the channel dimension.

Firstly, the feature vector per image, extracted from G-CNN, is regarded as a single query. Each shape primitive is corresponding to one key-value pair. Once we obtain the query-key-value, the attention can be established. Similar to the conventional attention mechanism [69], we have:

$$\begin{aligned} \mathbf{W} &= \text{softmax}\left(\frac{\mathbf{Q}\mathbf{\Phi}^T}{\sqrt{d_k}}\right), \\ \mathbf{Q}' &= \mathbf{W}\mathbf{\Phi}, \end{aligned} \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{B\times N}$ contains attention scores, d_k is a scaling factor, and $\mathbf{Q}' \in \mathbb{R}^{B\times d}$ contains reconstructed shape features. For any image with index j, the corresponding reconstructed shape feature vector $\mathbf{q}'_j$ can be written as:

$$\mathbf{q}'_j = \sum_{i=1}^{N} w_{ji}\boldsymbol{\phi}_i. \tag{3}$$

One can see that the reconstructed feature vector $\mathbf{q}'$ is obtained by a linear combination of shape primitives $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N]$. This also demonstrates that using attention to model feature vectors with shape primitives is reasonable. In Section 4.4, we empirically show that the linear addition of primitives results in continuously changing decoded images.

Similarity Module. The similarity score is computed for pairwise reconstructed support-query feature vectors. We use the cosine similarity as metric due to its simplicity. Given a pair of query and support samples, the similarity score s is obtained by the so-called weighted sum-kernel:

$$\begin{aligned} s &= \langle \mathbf{q}'_s,\ \mathbf{q}'_q \rangle \\ &= \left(\sum_{i=1}^{N} w_i^s \boldsymbol{\phi}_i\right)^{\top} \left(\sum_{j=1}^{N} w_j^q \boldsymbol{\phi}_j\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i^s w_j^q \langle \boldsymbol{\phi}_i,\ \boldsymbol{\phi}_j \rangle. \end{aligned} \tag{4}$$

If $i = j$, $\langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle$ becomes larger than the case of $i \neq j$ (assuming the unit ℓ_2 norm of vectors). Based on this fact and Eq. (4), if two samples are indeed similar, the similarity function encourages larger weights on the same primitives from both samples.

Dual Attention Architecture. In this section, we apply multi-head unit to extend single-head based shape primitives. Since we use $\mathbf{\Phi}$ as the "basis" to reconstruct inputs, a common $\mathbf{\Phi}$ is learned and used. Standard Multi-head Cross-Attention (S-MCA) [69] will learn different sub-primitive matrices for each head and thus cannot achieve our aim. Thus, we propose a Holistic Multi-head Cross-Attention (H-MCA) that encourages a common $\mathbf{\Phi}$ across different heads. By "holistic" we mean maintaining the integrity of primitives.

In the reminder of this paper, we use $\mathbf{Q}'_H$ to indicate the reconstruction via H-MCA. To that end, first, the dimensions of $\mathbf{Q}$ and $\mathbf{\Phi}$ (used as key) are mapped to $d' = d \times h$ dimension, while on the other hand $\mathbf{\Phi}$ (used as value) is duplicated h times (h is the number of heads). Thus, the output of each head is obtained from a linear combination of common primitives from set $\mathbf{\Phi}$, which preserves the structural integrity of the primitives, and helps visualize primitives. However, H-MCA itself lacks flexibility in reconstructing the features $\mathbf{Q}$, sometimes leading to poor $\mathbf{Q}'$ due to inadequate shape characteristics. To accommodate and supplement H-MCA, S-MCA is added into H-MCA to achieve enhanced reconstruction:

$$\mathbf{Q}' = \text{H-MCA}(\mathbf{Q}) + \text{S-MCA}(\mathbf{Q}) = \mathbf{Q}'_H + \mathbf{Q}'_S, \tag{5}$$

where $\mathbf{Q}'_S$ is the reconstruction by S-MCA. We term Eq. (5) as *dual attention architecture*. One can easily see that the enhanced $\mathbf{Q}'$ satisfies Eq. (3) while using supplementary term $\mathbf{Q}'_S$. This practice is also similar to residual structure in ResNet, where H-MCA basically reconstructs the identity feature $\mathbf{Q}'_H$, whereas S-MCA learns the subtle changes $\mathbf{Q}'_S$ to complete the reconstructed features $\mathbf{Q}'$.

Figure 2(a) shows an example of how S-MCA works. The attention tensor $\mathbf{W} \in \mathbb{R}^{2\times 1\times 2}$ assumes the number of primitives, heads, and batch size are 2, 2, and 1, respectively. Please note that S-MCA breaks the integrity of the primitives as the attention is performed on each head. Thus, if S-MCA is used alone, the reconstructed $\mathbf{Q}'$ is not obtained by the linear weighting of holistic primitives $\mathbf{\Phi}$, which contradicts Eq. (3).

3.4. Decoder Module

The reconstructed features $\mathbf{Q}'$ are used by the similarity module for few-shot learning. $\mathbf{Q}'$ also requires additional constraints to achieve faithful reconstruction akin to input features $\mathbf{Q}$. One might align $\mathbf{Q}'$ and $\mathbf{Q}$ via the Least Squares Error or KL-divergence. One might align both terms to enhance edges and the shape mask [74]. In such a way, the shape supervision can be injected directly on top of the shape features $\mathbf{Q}$.

In our case, the alignment between $\mathbf{Q}$ and $\mathbf{Q}'$ is achieved through two decoders with common supervision: one decoder focuses on the edges, and the other focuses on the

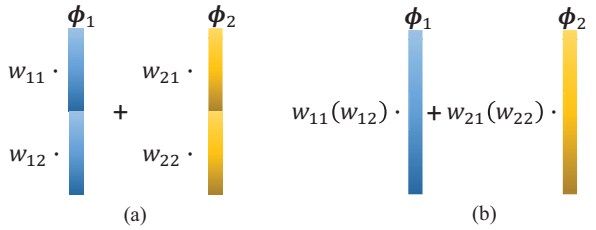

Figure 2. (a) Standard Multi-head Cross-Attention (S-MCA). Each primitive ϕ_1, ϕ_2 ($\mathbf{\Phi} = [\phi_1, \phi_2]$) is essentially divided into two pieces. (b) Holistic multi-head cross-attention (H-MCA). The primitives are complete regardless of the number of heads. In both cases, the number of primitives is 2, h=2, B=1, and attention tensor $\mathbf{W} \in \mathbb{R}^{2\times1\times2}$.

Figure 3. (a) shows a polar bear from the AwA2 dataset, and (d) shows an eight-petaled flower from the simple shape dataset. (a)(d) original image, (b)(e) mask image, and (c)(f) edge image.

shape mask. Moreover, we encourage the similarity between $\mathbf{Q}'_H$ and $\mathbf{Q}$ as high as possible.

By passing $\mathbf{Q}$, $\mathbf{Q}'_H$ and $\mathbf{Q}'$ through two decoders, the decoders enforce their alignment as well as learning of shape primitives. Both masks and edges (Figures 3(b) and (c)) are used as ground truth, allowing the network to focus on the shape information of objects. Additionally, the decoder module also accomplishes the task of image reconstruction.

3.5. Loss Function

Our loss function can be summarized as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{res}, \tag{6}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{res}$ are the loss functions for classification and image reconstruction, respectively. The loss function $\mathcal{L}_{res}$ includes six components:

$$\mathcal{L}_{res} = \mathcal{L}^{\mathbf{Q}}_{mask} + \mathcal{L}^{\mathbf{Q}}_{edge} + \mathcal{L}^{\mathbf{Q}'_H}_{mask} + \mathcal{L}^{\mathbf{Q}'_H}_{edge} + \mathcal{L}^{\mathbf{Q}'}_{mask} + \mathcal{L}^{\mathbf{Q}'}_{edge},$$

where $\mathcal{L}_{mask}$ and $\mathcal{L}_{edge}$ are the loss functions for mask and edge image reconstruction, respectively. $\mathcal{L}^{Q}$, $\mathcal{L}^{Q'_H}$ and $\mathcal{L}^{Q'}$ are the image reconstruction loss functions for the original features, holistically reconstructed features and dual reconstructed features, respectively. For the reconstruction loss function, MSE loss is used.

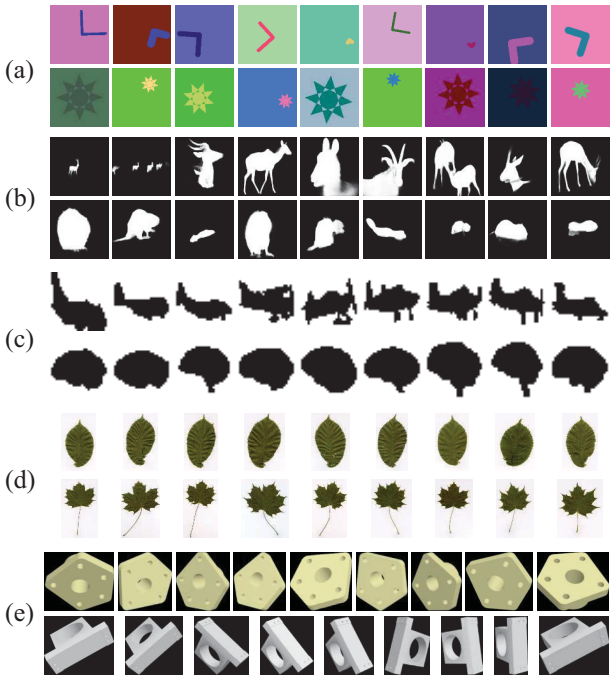

Figure 4. Example of shape images from *five* datasets. Each row showcases the data from one class. (a) Simple shape dataset, *e.g.*, *folding line*, *octagram*. (b) The shape-AwA2 dataset, including *antelope*, *beaver*. (c) Caltech 101 shapes with classes: *airplanes side 2* and *brain*. (d) Swedish leaf dataset. (e) Workpiece dataset.

4. Experiments

4.1. Datasets

We validate the effectiveness of the proposed FSSD on five datasets. Firstly, we create *simple shape dataset* and create *shape-AwA2 dataset* by transforming the popular animal dataset *AwA2* [5]. Secondly, the public *Caltech 101 shapes* dataset [51], *Workpiece* dataset [21, 45, 47, 76] and *Swedish leaf* dataset [68] are also used. Figure 4 shows some examples of the five datasets. Specifically, the simple shape dataset and shape-AwA2 are obtained as follows:

- The simple shape dataset is developped using basic geometric elements such as lines, arcs, angles, and circles. It is created using PIL (Python Imaging Library) by randomly generating shapes with varying sizes, positions, orientations, and foreground/background colors. There are a total of 25 classes with 250,000 images in this dataset.

- AwA2 dataset is a popular few-shot learning dataset. However, extracting shape features directly from color images is challenging. Therefore, in this paper, a state-of-the-art salient object detector [77] is employed to extract masks of animals from AwA2, creating the

Table 1. Comparison with traditional shape descriptors and classical few-shot learning methods across five datasets. The 5-way 1-shot and 5-way 5-shot results averaged over 5,000 test episodes are reported.

Algorithm	Backbone	Params	Simple-Shape (↑)		Shape-AwA2 (↑)		Caltech 101 (↑)		Workpiece shapes (↑)		Swedish leaf (↑)	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Shape context [6]	-	0M	34.25%	50.64%	22.69%	31.92%	52.80%	66.00%	30.80%	53.60%	60.32%	70.02%
Fourier descriptors [82]	-	0M	55.76%	65.60%	18.49%	23.77%	38.17%	48.91%	46.57%	62.53%	38.09%	54.96%
Hu moment [24]	-	0M	80.05%	87.49%	21.85%	22.72%	47.74%	50.63%	47.32%	50.16%	47.73%	64.44%
ProtoNet [65]	ResNet18	43M	83.90%	80.79%	23.66%	34.07%	66.28%	85.41%	67.84%	86.55%	47.56%	91.01%
RelationNet [67]	ResNet18	48M	80.84%	83.23%	22.74%	42.45%	59.53%	86.74%	66.76%	89.18%	43.56%	92.50%
Ours	G-ResNet18	45M	**91.02%**	**92.58%**	**43.35%**	**54.45%**	**80.19%**	**89.13%**	**97.29%**	**97.79%**	**88.58%**	**93.76%**

Table 2. Results for several backbones on the simple shape dataset.

Model	Backbone	Acc (↑)
Ours	4-layer conv	78.72%
	ResNet18	85.24%
	ResNet50	85.83%
	G-ResNet18	**91.02%**

Model	Backbone	Acc (↑)
ProtoNet	ResNet18	83.90%
	G-ResNet18	86.59%
RelationNet	ResNet18	80.84%
	G-ResNet18	84.71%

masks of AwA2 dataset specifically for shape classification research. This shape dataset consists of 50 classes with 37,322 images.

4.2. Experiment Setup

G-CNN with ResNet18 backbone is used for basic feature extraction. We use 5-way 1/5-shot setting with 15 query samples per class. We set the number of episodes to be 50,000 for training and 5,000 for testing. The Adam optimizer with learning rate 0.001 is used. The learning rate decays by 0.5 every 8,000 episodes. The classification accuracy is used to measure the performance of few-shot models, while PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) are used to measure the performance of image reconstruction.

4.3. Results of Few-shot Shape Recognition

In this section, we compare the proposed FSSD with traditional shape descriptors such as Shape Context [6], Fourier Descriptors [82], and Hu Moment [24], and classical few-shot learning methods such as ProtoNet [65] and RelationNet [67]. We compute 5-way 1/5-shot classification accuracy on five datasets by averaging 5,000 randomly generated episodes from the testing set. Table 1 shows the superior performance of our method compared to other approaches on five datasets where the shape plays a prominent role. Our method demonstrates superior performance and efficacy in capturing shape information and outperforms the compared methods. Unlike us, methods such as CNNs struggle to capture shape information in 5-way 1-shot setting on the Workpiece and Swedish leaf datasets. While CNNs begin to improve performance with more templates being used, the achieved performance is still inferior to our method. Moreover, we visualize shape matching using our method on the shape-AwA2 dataset, the simple shape

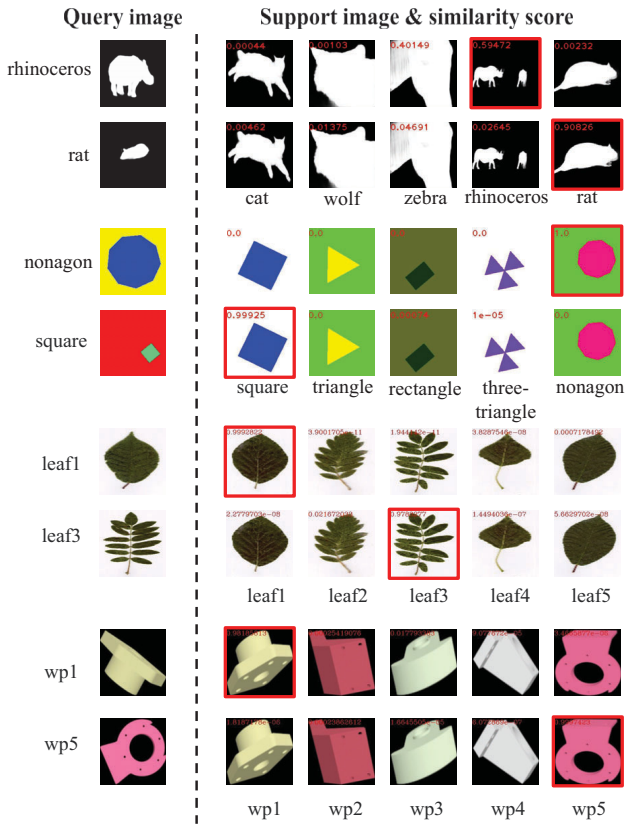

Figure 5. Visualizations of shape matching. Matched shapes are stressed by red frames. Similarity score is shown in red in the top-left corner. *Best viewed in zoom.*

dataset, the Swedish leaf dataset, and the Workpiece dataset in Figure 5, which shows that our model is able to successfully perform shape retrieval.

4.4. Ablation Studies

To our best knowledge, there is no existing research on shape-based neural networks for few-shot shape recognition. Therefore, we use state-of-the-art few-shot learning model [67] as the baseline and conduct ablation experiments on: (1) backbone types, (2) decoding methods, (3) similarity calculation methods, (4) attention types, and (5)

Table 3. Results of different decoders using G-CNN.

Dataset	Decoder	Acc(↑)	Mask		Edge	
			SSIM(↑)	PSNR(↑)	SSIM(↑)	PSNR(↑)
Simple Shape	-	89.01%	-	-	-	-
	Mask	89.09%	0.8636	40.01	-	-
	Edge	88.90%	-	-	0.8943	38.41
	Mask+Edge	**91.02**%	0.8926	40.01	0.9131	38.81

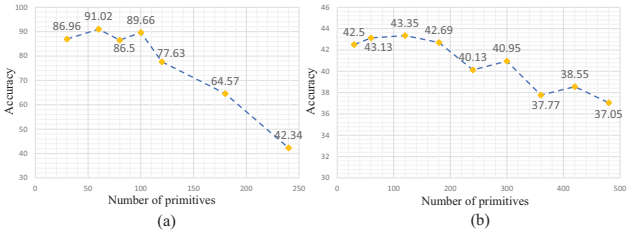

Figure 6. (a) Shape recognition performance using 30, 60, 80, 100, 120, 180, and 240 shape primitives on the simple shape dataset. (b) Same experiments tested on the shape-AwA2 dataset.

model hyper-parameters.

Types of backbone. To examine the importance of G-CNN, the experiments are performed by only changing the backbone while keeping other modules the same. The impact of different backbones is shown in Table 2. On the simple shape dataset, we observe that the performance increases along with the model size. However, the margin becomes smaller between ResNet18 and ResNet50. If replacing ResNet18 by G-ResNet18 (G-CNN version), our model leads to significant improvements in performance (5.8%). Moreover, using G-ResNet18 has very tiny overhead compared to the original CNN model (45.41M *vs*. 43M). Compared to ResNet50 (105.5M), G-ResNet18 (45.41M) enjoys much lower model capacity and higher performance, which shows the G-CNN extracts better shape representations.

Different decoder methods. To investigate the influence of shape mask/edge decoders, we only vary the decoders while the other modules are same. Table 3 shows the results of different decoding methods. One can see how both decoders contribute to the classification and image reconstruction measured by SSIM and PSNR. Using both decoders simultaneously provides superior performance compared to each single one. This also implies that the mask and edge groundtruth are complementary; the former provides the global information while the latter captures the edge information.

Calculation of similarity. This experiment compares cosine similarity *vs*. the relation module popular in few-shot learning [67]. The relation module learns a good "metric" in a data-driven way and enjoys better performance than conventional distances such as the Euclidean or cosine distance. Table 4 shows that using the cosine similarity in few-shot shape recognition is reasonable. We believe this is mainly due to the superior shape features learned through our model, thus, the performance depends less on the distance metric.

Attention architecture. This experiment uses G-ResNet18 as the backbone, a pair-wise decoder, and the cosine similarity while exploring the efficacy of different attention architectures. As shown in Table 5, a standalone S-MCA achieves higher classification accuracy and better image reconstruction quality than H-MCA. Nonetheless, S-MCA is not meant to discover the common primitives across different heads and leads to weak interpretations in visualization.

On the other hand, H-MCA provides comparable performance in accuracy but shows relatively weak performance in reconstruction, as indicated in the methodology section. The dual attention architecture enables the integration of both modules to complement each other. As a result, the combined model obtains the best classification results and better reconstruction than the single H-MCA.

Numbers of shape primitives. Figures 6(a) and (b) show the results with different numbers of shape primitives on the simple shape dataset and shape-AwA2 dataset, respectively. While more shape primitives seem to provide the enriched basis for reconstruction, they may also contain redundant information, leading to the overfitting issue, which can be observed in reconstruction-based feature learning. One may observe in both figures that the performance of the network did not improve with increasing numbers of primitives. When the number of primitives exceeds a specific point, the classification performance drops quickly. Therefore, maintaining a moderate number of shape primitives is critical. According to Figures 6(a) and (b), for the optimal performance, we set the number of primitives to be 60 for the simple shape dataset and 120 for the shape-AwA2 dataset *by default* throughout all experiments.

Visualization of shape primitives. To investigate the visualization of shape primitives, we selected two primitives ϕ_1, ϕ_2, and performed interpolation on them as follows:

$$\begin{aligned} \phi_i^{inter} &= \alpha \times \phi_1 + (1 - \alpha) \times \phi_2 \\ \alpha &= i \times 0.1, \quad i = 0, \ldots, 10. \end{aligned} \tag{7}$$

Through the interpolation of ϕ_1 and ϕ_2, 11 interpolated feature vectors were obtained. These feature vectors were then fed into the decoder to decode them into corresponding masks, as shown in Figure 7. The figure shows that the reconstructed images corresponding to the interpolated features are continuously changing. This demonstrates that the decoded images corresponding to the linearly combined primitives are obtained by continuous variation in the reconstruction space. Figure 8 shows the visualization results of some primitives of the simple shape dataset when the number of primitives is set to 60. From the results, it is

Table 4. Comparison of different distances and similarity metrics.

Backbone	Decoder	Method	Acc(↑)	Mask		Edge	
				SSIM(↑)	PSNR(↑)	SSIM(↑)	PSNR(↑)
G-CNN	Mask+Edge	Relation module	66.23%	0.8815	42.77	0.9455	39.67
		Cosine similarity	**91.02**%	0.8926	40.01	0.9131	38.81

Table 5. Comparison of H-MCA and S-MCA.

Backbone	Attention	Acc(↑)	Mask		Edge	
			SSIM(↑)	PSNR(↑)	SSIM(↑)	PSNR(↑)
G-CNN	S-MCA	90.13%	0.9180	42.79	0.9395	39.39
	H-MCA	89.21%	0.8286	37.32	0.8854	37.12
	Dual attention	**91.02**%	0.8926	40.01	0.9131	38.81

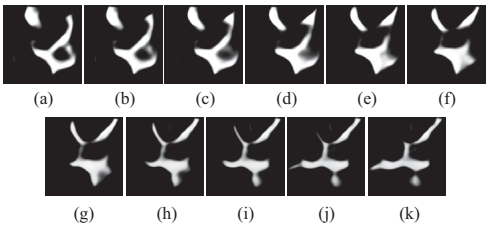

Figure 7. Images (a) and (k) visualize primitives ϕ_1 and ϕ_2, and (b)-(j) visualize the feature vectors interpolated by ϕ_1 and ϕ_2.

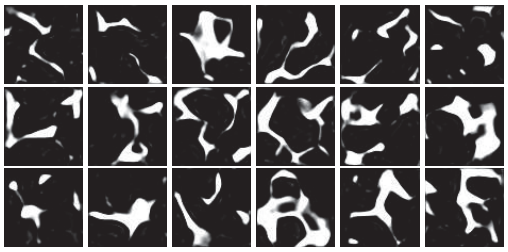

Figure 8. Partial visualization of primitives on the simple shape dataset. The number of shape primitives is set to 60.

not straightforward to link the visualization of primitives to a real yet complete shape. However, both the theoretical outcomes in Eq. 3 and the decoding of the reconstructed interpolated features in Figure 7 demonstrate that these primitives can approximate the original shape well.

Visualization of reconstructions. Figure 9 shows the visualization results of $\mathbf{Q}, \mathbf{Q}'_H$ and $\mathbf{Q}'$ on the test dataset through the decoder using the mask branch. One can see that for shapes not included in the training dataset, our model can still leverage primitives to reconstruct them in a precise manner. These results demonstrate that dual attention is needed for faithful and interpretable reconstruction. For instance, the pentagon in Figures 9(a), (c) and (d) represent ground truth, results by H-MCA and H-MCA+S-MCA, respectively. One may see that H-MCA is not sensitive to

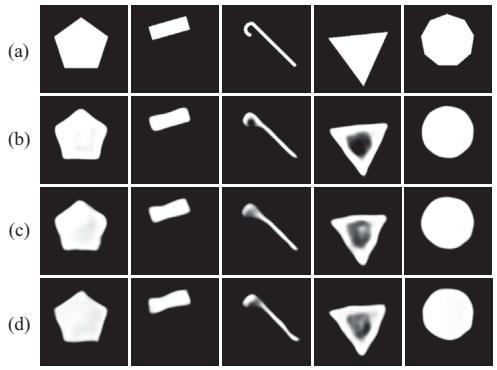

Figure 9. Visualization of reconstructed images: (a) ground truth; (b), (c), (d) are the decoding results of $\mathbf{Q}$, $\mathbf{Q}'_H$ and $\mathbf{Q}'$.

obtuse angles, and the decoder tends to reconstruct obtuse angles as arcs. However, the integration with S-MCA can do a better job of reconstructing the obtuse angles.

5. Conclusions

We have proposed to incorporate shape properties in the network architecture so that the network can perform efficient few-shot shape learning. Our proposed FSSD replaces the CNN with a G-CNN to enable the network obtain robustness to geometric variations of object poses. An image reconstruction module is used and a pair-wise decoder architecture to ensure the output image focuses on the edge information of the target. A dual attention architecture is used to implement shape primitives learning, and the shape primitives are used to reconstruct new features to approximate the original features. Our FSSD is robust to shape transformations and has the ability to reconstruct shapes due to improved shape modeling. The extensive experiments on few-shot shape recognition show the effectiveness of the proposed approach. Despite the learned geometric primitives differ from human-made geometric primitives (*e.g.*, lines, small arcs), we hope our work may inspire the future research in this field.

References

[1] Habibollah Agh Atabay. Binary shape classification using convolutional neural networks. *IIOAB J*, 7(5):332–336, 2016. 1, 2

[2] Habibollah Agh Atabay. A convolutional neural network with a new architecture applied on leaf classification. *IIOAB J*, 7(5):226–331, 2016. 1, 2

[3] TJ Atherton and DJ Kerbyson. Using phase to represent radius in the coherent circle hough transform. In *IEE colloquium on Hough transforms*, pages 5–1. IET, 1993. 2

[4] Pedro Ballester and Ricardo Araujo. On the performance of googlenet and alexnet applied to sketches. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 1

[5] Prianka Banik, Lin Li, and Xishuang Dong. A novel dataset for keypoint detection of quadruped animals from images. *arXiv preprint arXiv:2108.13958*, 2021. 5

[6] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, 24(4):509–522, 2002. 2, 6

[7] Miroslaw Bober. Mpeg-7 visual shape descriptors. *IEEE Transactions on circuits and systems for video technology*, 11(6):716–719, 2001. 2

[8] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019. 1

[9] Ying Cai, Yuliang Liu, Chunhua Shen, Lianwen Jin, Yidong Li, and Daji Ergu. Arbitrarily shaped scene text detection with dynamic convolution. *Pattern Recognition*, 127:108608, 2022. 1, 2

[10] GC-H Chuang and C-CJ Kuo. Wavelet descriptor of planar curves: Theory and applications. *IEEE Transactions on Image Processing*, 5(1):56–70, 1996. 2

[11] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 3

[12] Wenbo Dong, Pravakar Roy, Cheng Peng, and Volkan Isler. Ellipse r-cnn: Learning to infer elliptical object from clustering and occlusion. *IEEE Transactions on Image Processing*, 30:2193–2206, 2021. 2

[13] Richard O Duda and Peter E Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972. 2

[14] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020. 2

[15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2

[16] Jan Flusser and Tomas Suk. Pattern recognition by affine moment invariants. *Pattern recognition*, 26(1):167–174, 1993. 2

[17] Christina M Funke, Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Synthesising dynamic textures using convolutional neural networks. *arXiv preprint arXiv:1702.07006*, 2017. 1

[18] Davi Geiger and Zvi M Kedem. Quantum interference and shape detection. In *Energy Minimization Methods in Computer Vision and Pattern Recognition: 11th International Conference, EMMCVPR 2017, Venice, Italy, October 30–November 1, 2017, Revised Selected Papers 11*, pages 18–33. Springer, 2018. 2

[19] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1

[20] Chetan S Gode and Atish S Khobragade. Object detection using color clue and shape feature. In *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 464–468. IEEE, 2016. 2

[21] Changjian Gu, Chaochen Gu, Kaijie Wu, Liangjun Zhang, and Xinping Guan. Cad-based viewpoint estimation of texture-less object for purposive perception using domain adaptation. *International Journal of Robotics and Automation*, 34(6), 2019. 5

[22] You Hao, Qi Li, Hanlin Mo, He Zhang, and Hua Li. Ami-net: convolution neural networks with affine moment invariants. *IEEE Signal Processing Letters*, 25(7):1064–1068, 2018. 1

[23] Paul VC Hough. Machine analysis of bubble chamber pictures. In *International Conference on High Energy Accelerators and Instrumentation, CERN, 1959*, pages 554–556, 1959. 2

[24] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8(2):179–187, 1962. 6

[25] Kyu-hong Hwang, Ho-rim Park, and Young-guk Ha. Stfnet: Image classification model based on balanced texture and shape features. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 268–274. IEEE, 2021. 1

[26] Andrei C Jalba, Michael HF Wilkinson, and Jos BTM Roerdink. Shape representation and recognition through morphological curvature scale spaces. *IEEE Transactions on Image Processing*, 15(2):331–341, 2006. 2

[27] Saumya Jetley, Michael Sapienza, Stuart Golodetz, and Philip HS Torr. Straight to shapes: real-time detection of encoded shapes. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6550–6559, 2017. 2

[28] WooYeol Jun, JeongMok Ha, Byeongchan Jeon, JoonHo Lee, and Hong Jeong. Led traffic sign detection with the fast radial symmetric transform and symmetric shape detection. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 310–315. IEEE, 2015. 2

[29] Kenichi Kanatani and Naoya Ohta. Automatic detection of circular objects by ellipse growing. *International Journal of Image and Graphics*, 4(01):35–50, 2004. 2

[30] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019. 2

[31] Ba Rom Kang, Hyunku Lee, Keunju Park, Hyunsurk Ryu, and Ha Young Kim. Bshapenet: Object detection and instance segmentation with bounding shape masks. *Pattern Recognition Letters*, 131:449–455, 2020. 2

[32] Hoel Kervadec, Houda Bahig, Laurent Letourneau-Guillon, Jose Dolz, and Ismail Ben Ayed. Beyond pixel-wise supervision for segmentation: A few global shape descriptors might be surprisingly good! In *Medical Imaging with Deep Learning*, pages 354–368. PMLR, 2021. 1

[33] Akio Kimura and Takashi Watanabe. An extension of the generalized hough transform to realize affine-invariant two-dimensional (2d) shape detection. In *2002 International Conference on Pattern Recognition*, volume 1, pages 65–69. IEEE, 2002. 2

[34] Lingwei Kong, Jianzong Wang, Zhangcheng Huang, and Jing Xiao. A competition of shape and texture bias by multi-view image representation. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part IV 4*, pages 140–151. Springer, 2021. 1

[35] Piotr Koniusz and Krystian Mikolajczyk. On a quest for image descriptors based on unsupervised segmentation maps. In *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*, pages 762–765. IEEE Computer Society, 2010. 2

[36] Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016. 1

[37] Laksono Kurnianggoro, Alexander Filonenko, and Kang-Hyun Jo. Shape classification using combined features. In *Computational Collective Intelligence: 9th International Conference, ICCCI 2017, Nicosia, Cyprus, September 27-29, 2017, Proceedings, Part II 9*, pages 549–557. Springer, 2017. 1, 2

[38] Iago Landesa-Vázquez, Francisco Parada-Loira, and José L Alba-Castro. Fast real-time multiclass traffic sign detection based on novel shape and texture descriptors. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 1388–1395. IEEE, 2010. 2

[39] Tam T Le, Son T Tran, Seichii Mita, and Thuc D Nguyen. Real time traffic sign detection using color and shape-based features. In *ACIIDS (2)*, pages 268–278, 2010. 2

[40] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[41] Dilong Li, Xin Shen, Yongtao Yu, Haiyan Guan, Hanyun Wang, and Deren Li. Ggm-net: Graph geometric moments convolution neural network for point cloud shape classification. *IEEE Access*, 8:124989–124998, 2020. 1

[42] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8334–8343, 2021. 2

[43] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2869–2878, 2020. 2

[44] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[45] Changsheng Lu, Chaochen Gu, Kaijie Wu, Siyu Xia, Haotian Wang, and Xinping Guan. Deep transfer neural network using hybrid representations of domain discrepancy. *Neurocomputing*, 409:60–73, 2020. 5

[46] Changsheng Lu and Piotr Koniusz. Few-shot keypoint detection with uncertainty learning for unseen species. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19416–19426, June 2022. 2

[47] Changsheng Lu, Haotian Wang, Chaochen Gu, Kaijie Wu, and Xinping Guan. Viewpoint estimation for workpieces with deep transfer learning from cold to hot. In *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13-16, 2018, Proceedings, Part I 25*, pages 21–32. Springer, 2018. 1, 5

[48] Changsheng Lu, Siyu Xia, Wanming Huang, Ming Shao, and Yun Fu. Circle detection by arc-support line segments. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 76–80. IEEE, 2017. 1

[49] Changsheng Lu, Siyu Xia, Ming Shao, and Yun Fu. Arc-support line segments revisited: An efficient high-quality ellipse detection. *IEEE Transactions on Image Processing*, 29:768–781, 2019. 1

[50] Changsheng Lu, Hao Zhu, and Piotr Koniusz. From saliency to dino: Saliency-guided vision transformer for few-shot keypoint detection, 2023. 2

[51] Benjamin Marlin, Kevin Swersky, Bo Chen, and Nando Freitas. Inductive principles for restricted boltzmann machine learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 509–516. JMLR Workshop and Conference Proceedings, 2010. 5

[52] Laurynas Miksys, Saumya Jetley, Michael Sapienza, Stuart Golodetz, and Philip HS Torr. Straight to shapes++: real-time instance segmentation made more accurate. *arXiv preprint arXiv:1905.11358*, 2019. 2

[53] Farzin Mokhtarian. Robust and efficient shape indexing through curvature scale space. *Proceedings of BMVC'96, Edinbourgh 9-13 Sept.*, 1996. 2

[54] Hankyu Moon, Rama Chellappa, and Azriel Rosenfeld. Optimal edge-based shape detection. *IEEE transactions on Image Processing*, 11(11):1209–1227, 2002. 2

[55] Prerana Mukherjee, Brejesh Lall, and Sarvaswa Tandon. Salprop: Salient object proposals via aggregated edge cues. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2423–2429. IEEE, 2017. 1, 2

[56] Eduardo A Murillo-Bracamontes, Miguel E Martinez-Rosas, Manuel M Miranda-Velasco, Horacio L Martinez-Reyes, Jesus R Martinez-Sandoval, and Humberto Cervantes-de Avila. Implementation of hough transform for fruit image segmentation. *Procedia Engineering*, 35:230–239, 2012. 2

[57] Ah-Reum Oh and Mark S Nixon. Extending the image ray transform for shape detection and extraction. *Multimedia Tools and Applications*, 74:8597–8612, 2015. 2

[58] Zhuo Qi, Wenyi Chen, Xiaofei Sun, Wangqian Sun, and Hui Yang. Ktext: Arbitrary shape text detection using modified k-means. *IET Computer Vision*, 16(1):38–49, 2022. 1, 2

[59] Min Qiao, Gang Zhou, Qiu Ling Liu, and Li Zhang. Salient object detection: An accurate and efficient method for complex shape objects. *IEEE Access*, 9:169220–169230, 2021. 1, 2

[60] Aaron Rasheed Rababaah. Angle histogram of hough transform as shape signature for visual object classification–(ahoc). *International Journal of Computational Vision and Robotics*, 10(4):312–336, 2020. 2

[61] Yongmei Ren, Jie Yang, Qingnian Zhang, and Zhiqiang Guo. Ship recognition based on hu invariant moments and convolutional neural network for video surveillance. *Multimedia Tools and Applications*, 80:1343–1373, 2021. 1

[62] Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning*, pages 2940–2949. PMLR, 2017. 1

[63] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[64] Jaspreet Singh and Chandan Singh. Learning invariant representations for equivariant neural networks using orthogonal moments. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022. 1

[65] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2, 6

[66] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2014. 1, 2

[67] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 2, 6, 7

[68] Oskar J. O. Söderkvist. Swedish leaf dataset. https://www.cvl.isy.liu.se/en/research/datasets/swedish-leaf/. Accessed: 2016-03-08. 1, 5

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[70] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 3

[71] Tianhao Wang, Changsheng Lu, Ming Shao, Xiaohui Yuan, and Siyu Xia. Eldet: An anchor-free general ellipse object detector. In *Proceedings of the Asian Conference on Computer Vision*, pages 2580–2595, 2022. 1, 2

[72] Wenhai Wang, Enze Xie, Xiang Li, Xuebo Liu, Ding Liang, Zhibo Yang, Tong Lu, and Chunhua Shen. Pan++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5349–5367, 2021. 1, 2

[73] Zhiqian Wang and Jezekiel Ben-Arie. Detection and segmentation of generic shapes based on affine modeling of energy in eigenspace. *IEEE transactions on image processing*, 10(11):1621–1629, 2001. 2

[74] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13025–13034, 2020. 1, 2, 4

[75] Gang Wu, Weijie Liu, Xiaohui Xie, and Qiang Wei. A shape detection method based on the radial symmetry nature and direction-discriminated voting. In *2007 IEEE International Conference on Image Processing*, volume 6, pages VI–169. IEEE, 2007. 2

[76] Xunjin Wu, Changsheng Lu, Chaochen Gu, Kaijie Wu, and Shanying Zhu. Domain adaptation for viewpoint estimation with image generation. In *2021 International Conference on control, automation and information sciences (ICCAIS)*, pages 341–346. IEEE, 2021. 5

[77] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7264–7273, 2019. 1, 2, 5

[78] Xianghua Xu, Jiancheng Jin, Shanqing Zhang, Lingjun Zhang, Shiliang Pu, and Zongmao Chen. Smart data driven traffic sign detection method based on adaptive color threshold and shape symmetry. *Future generation computer systems*, 94:381–391, 2019. 2

[79] Chengzhuan Yang, Lincong Fang, and Hui Wei. Learning contour-based mid-level representation for shape classification. *IEEE Access*, 8:157587–157601, 2020. 2

[80] Haichun Yang, Ruining Deng, Yuzhe Lu, Zheyu Zhu, Ye Chen, Joseph T Roland, Le Lu, Bennett A Landman, Agnes B Fogo, and Yuankai Huo. Circlenet: Anchor-free glomerulus detection with circle representation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 35–44. Springer, 2020. 1

[81] Fenggen Yu, Zhiqin Chen, Manyi Li, Aditya Sanghi, Hooman Shayani, Ali Mahdavi-Amiri, and Hao Zhang. Capri-net: learning compact cad shapes with adaptive primitive assembly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11768–11778, 2022. 1

[82] Charles T Zahn and Ralph Z Roskies. Fourier descriptors for plane closed curves. *IEEE Transactions on computers*, 100(3):269–281, 1972. 2, 6

[83] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014. 1

[84] Chaoyan Zhang, Yan Zheng, Baolong Guo, Cheng Li, and Nannan Liao. Scn: a novel shape classification algorithm based on convolutional neural network. *Symmetry*, 13(3):499, 2021. 1, 2

[85] Dengsheng Zhang and Guojun Lu. Shape-based image retrieval using generic fourier descriptor. *Signal Processing: Image Communication*, 17(10):825–848, 2002. 2

[86] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8779–8788, 2019. 1, 2

[87] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3123–3131, 2021. 1, 2