This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Selective Multi-View Deep Model for 3D Object Classification

Mona Alzahrani[1,2] Muhammad Usman[1,3,4*] Saeed Anwar[1,3] Tarek Helmy[1,4]
[1]Department of Information & Computer Science, KFUPM, Dhahran, Saudi Arabia
[2]College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia
[3]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, KFUPM, Dhahran, Saudi Arabia
[4]Center for Intelligent Secure Systems, KFUPM, Dhahran, Saudi Arabia
{g201908310, muhammad.usman, saeed.anwar, helmy}@kfupm.edu.sa

Abstract

3D object classification has emerged as a practical technology with applications in various domains, such as medical image analysis, automated driving, intelligent robots, and crowd surveillance. Among the different approaches, multi-view representations for 3D object classification have shown the most promising results, achieving state-of-the-art performance. However, there are certain limitations in current view-based 3D object classification methods. One observation is that using all captured views for classification can confuse the classifier and lead to misleading results for certain classes. Additionally, some views may contain more discriminative information for object classification than others. These observations motivate the development of smarter and more efficient selective multi-view classification models. In this work, we propose a Selective Multi-View Deep Model that extracts multi-view images from 3D data representations and selects the most influential view by assigning importance scores using the cosine similarity method based on visual features detected by a pre-trained CNN. The proposed method is evaluated on the ModelNet40 dataset for the task of 3D classification. The results demonstrate that the proposed model achieves an overall accuracy of 88.13% using only a single view when employing a shading technique for rendering the views, pre-trained ResNet-152 as the backbone CNN for feature extraction, and a Fully Connected Network (FCN) as the classifier.

1. Introduction

3D object classification is vital in 3D computer vision and has significant applications in fields like medical imaging, autonomous driving, robotics, and various reality technologies [1, 9–11]. Classification methods hinge on how 3D objects are represented, leading to three primary approaches [11, 23]: voxel-based methods using 3D grids, point-based methods working with point clouds, and view-based methods relying on 2D projections of the objects.

Nonetheless, recent view-based 3D object classification methods [4, 5, 8, 16, 23] have certain limitations. They often utilize all captured views, including views that are not discriminative for classification, leading to potential confusion and noise to the classifier, as it may struggle to distinguish between relevant and irrelevant information. Consequently, the classification accuracy may be compromised, and the model's overall performance may suffer. Another limitation is the potential presence of redundant views that does not contribute significantly to the classification task. Including such redundant views adds unnecessary complexity to the classification process and may lead to overfitting or reduced generalization. Furthermore, processing many views requires significant computational resources and can result in increased computational costs and longer processing times. This becomes challenging when dealing with complex 3D models or large datasets.

These shortcomings have motivated the development of more intelligent and efficient selective multi-view classification models focusing on extracting informative view, thereby improving the classification performance and efficiency of the system and demonstrates the most encouraging results. We propose a Selective Multi-View Deep Model that extracts multi-view images from 3D data representations and selects discriminative views using importance scores. These scores are based on visual features detected by a pre-trained CNN. The key contributions of this study can be summarized as follows:

- Development of a new Selective Multi-View Deep Model for 3D object classification.
- Identification of the highly significant camera image (sin-

*Corresponding author.
Project code: https://github.com/Mona-Alzahrani/SelectiveMV

gle view) that has more influence on the prediction result of the model using the Cosine Similarity technique.
- Evaluation of different CNNs and classifiers to suggest the most effective components for the model pipeline.
- Analysis of correctly predicted classes using Grad-CAM technique to highlight significant regions on the views.

2. Related Works

View-based and selective view-based methods are currently the top-performing approach in 3D object classification.

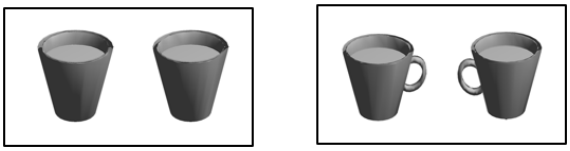
2.1. View-based 3D Object Classification

The pioneering MVCNN [16] introduced the concept of using multiple 2D views to represent and classify 3D objects but treated all views equally to extract the final shape descriptor, which was a limitation. To tackle this, the GVCNN considered the correlation between views to extract discriminative information. However, both methods assumed known viewing poses, an unrealistic scenario in real-world settings with occlusions. RotationNet [8] approached this by treating viewpoints as latent variables learned unsupervisedly, aiding in classification, though it relied on uniform view configurations. In contrast, view-GCN [23] represented multi-views as a graph, allowing for hierarchical feature extraction that considers inter-view relations, ultimately outperforming MVCNN [16], GVCNN [4], and RotationNet [8] in experiments.

Most previous 3D object classification methods utilize all available views, which can burden the classifier with non-discriminative or misleading information. As evidenced by Fig. 1 showing different object views, some views provide distinctive details for classification, such as a cup shows the cup's handle in Fig. 1b, while others offer little to no classification help, such as a cup in Fig. 1a. Therefore, a selection mechanism is crucial to filter out non-useful views to avoid classifier confusion, reduce computational load, and improve performance. The best-performing models on the ModelNet40 dataset, like RotationNet [8] and View-GCN [23], implement selection mechanisms, yet they have their limitations, such as requiring a minimum number of views or lacking a complete active selection strategy. This indicates that determining the optimal number of views remains an open research problem.

2.2. Selective View-based 3D Object Classification

To address the inefficiency of using all views for 3D object classification, certain methods proposed selective mechanisms [21] to choose the most informative views for the task. Meanwhile, other methods [3, 7, 18, 27] experimented with varying the number of views to see its impact on classification performance. A summary of these selective view-based classification methods, including details on datasets, selection mechanisms, number of training views, and backbone networks can be found in Tab. 1.



(a) Neighboring views. (b) Symmetric views.

Figure 1. Discriminative Analysis of 3D Cup Views: (a) Non-informative neighboring views; (b) Informative symmetric views revealing cup's handle.

The OVPT [21] developed by Wang et al. enhances recognition by minimizing view redundancy. OVPT uses an entropy-based mechanism to select the most informative views from 20 spherical views of each 3D object. These optimal views are then processed by a pre-trained ResNet-34 for feature extraction and transformed into a sequence for the transformer. A pooling transformer module subsequently combines these features into global descriptors for classification. OVPT requires only six views for state-of-the-art classification results, outperforming other deep learning-based [8, 26] and transformer-based [20, 22] methods in efficiency and computational resource usage.

Other view-based methods [3, 7, 18] have explored the impact of classifying 3D objects using a single, randomly chosen view. The transformer-based MVT model [3] was trained on 12 views from the ModelNet10 dataset and then tested with one random view. Similarly, MVCNN [7] and DeepCCFV [7] used the ModelNet40 dataset, trained with 12 views, and tested with a single view, using VGG-11 and ResNet-50 networks, respectively. ViewFormer [18] also applied a random single view selection from 20 views of the ModelNet40 dataset, utilizing ResNet-18 networks. MVSG-DNN [27], trained like OVPT [21] and ViewFormer [18] with 20 views, employed an LSTM for adaptive view selection, achieving stable classification results with a single view and an AlexNet backbone.

3. Methodology

The overall architecture of the proposed model is shown in Fig. 2. Our model has five phases: A) multi-view extraction, B) feature extraction, C) vectorization, D) view selection, and E) object classification. In the multi-view extraction stage, we extract multiple views from a given 3D object from different viewpoints and angles. Then, in the feature extraction stage, each extracted view is fed to a pre-trained CNN to extract the corresponding feature stack of the detected visual features. After that, in the vectorization stage, the detected feature stacks corresponding to the same object are converted to feature vectors. Then, in the view selection stage, feature vectors are compared based on cosine

Selective Model	Year	Model Type	Selection Mechanism	ModelNet Dataset 10	40v1	40v2	Training Views	Feature Extractor
MVSG-DNN [27]	2019	DL	Saliency LSTM	✓			20 views	AlexNet
MVSG-DNN [27]	2019	DL	Saliency LSTM			✓	20 views	AlexNet
MVCNN [16]	2019	DL	Random selection		✓		12 views	VGG-11
MVCNN [16]	2019	DL	Random selection		✓		12 views	ResNet-50
DeepCCFV [7]	2019	DL	Random selection		✓		12 views	VGG11-BN
DeepCCFV [7]	2019	DL	Random selection		✓		12 views	ResNet-50
MVT [3]	2021	Transformer	Random selection	✓			12 views	-
OVPT [21]	2022	Transformer	Information entropy			✓	20 views	ResNet-34
ViewFormer [18]	2023	Transformer	Random selection			✓	20 views	ResNet-18

Table 1. Selective view-based 3D object classification methods experimented with a single view (DL: Deep Learning).

similarity, and a critical score is given to each feature vector. Later in this stage, the important scores for all the extracted views corresponding to the same object are normalized and compared to select only the discriminative view useful for classification and contribute to the correct class. The chosen feature vector is considered a global descriptor of the object. Then, in the object classification stage, the global descriptor feeds a classifier to predict the object's class.

3.1. Multi-view Extraction

The multi-view of a 3D object k is obtained by applying a function E that renders m views $V_1, V_2, \ldots, V_m$ from predefined angles ϱ onto its Computer-Aided Design (CAD) model O_k, as described in Eq. (1). Each view V_i is captured by a virtual camera at a specific angle ϱ.

$$V_1, V_2, \cdots, V_m = E(\varrho, O_k). \tag{1}$$

Our proposed work experiments with two camera configurations as illustrated in Fig. 3: a circular with 12 views and a spherical with 20 views, both of which have contributed to achieving state-of-the-art results in the literature [8, 16, 23].

Circular Configuration. The first camera configuration is the regular circle with cameras at a 30° elevation angle, targeting the object's center depicted in Fig. 3a, optimal for objects with a consistent upright orientation [4, 8, 11, 16, 23]. Cameras are spaced every 30° in azimuth, resulting in 12 views that mimic the output from 1D turntables [5, 8].

Spherical Configuration. The second setup uses a dodecahedron-based configuration with 20 virtual cameras distributed across its vertices as displayed in Fig. 3b, allowing for equal spacing without assuming an upright orientation for 3D objects [5, 8, 23]. This arrangement captures unaligned objects from diverse angles, leveraging the dodecahedron's many vertices to distribute a uniform viewpoint. Our study follows this approach as others did [8, 23].

Pre-Trained CNN	Size (MB)	No. of Layers	Feature Map Shape
VGG-16 [15]	56.4	19	$7 \times 7 \times 512$
VGG-19 [15]	76.7	22	$7 \times 7 \times 512$
ResNet-50 [6]	93.8	175	$7 \times 7 \times 2048$
ResNet-152 [6]	234	515	$7 \times 7 \times 2048$
GoogLeNet [19]	88.8	311	$5 \times 5 \times 2048$

Table 2. Details of pre-trained CNNs tested for feature extraction.

3.2. Feature Extraction

The pre-trained CNN architecture has excellent performance in 2D classification tasks. Hence, in this step, the role of the pre-trained CNN ω is to process each view V_i to produce the corresponding feature map fm_i at the beginning of the classification model as in Eq. (2).

$$fm_i = \omega(V_i), \quad \text{for } i = 1, 2, \cdots, m. \tag{2}$$

In this step, the tested CNNs architectures pre-trained on ImageNet [12] as feature extractors are: VGG-16 [15], VGG-19 [15], GoogLeNet (InceptionV3) [19], ResNet-50 [6], and ResNet-152 [6]. Tab. 2 summarizes the details of the tested feature extractor CNNs, including their size, the total number of layers, and the shape of each extracted feature map in the form of rows × columns × channels.

3.3. Vectorization

Each feature map fm_i is flattened ϱ in this phase to be treated as the feature vector fv_i as in Eq. (3). This phase enables the subsequent phase to compare the different feature vectors using cosine similarity.

$$fv_i = \varrho(fm_i), \quad \text{for } i = 1, 2, \cdots, m. \tag{3}$$

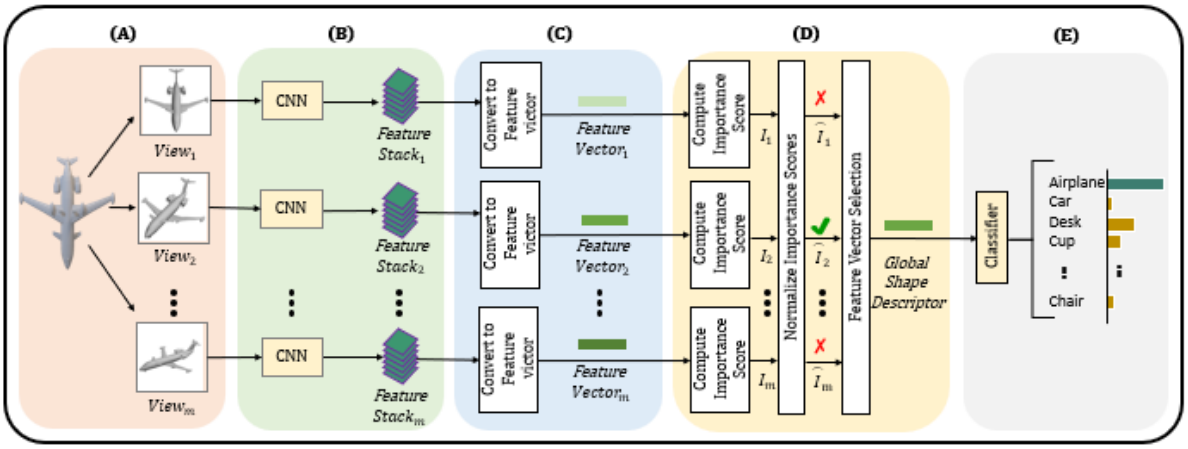
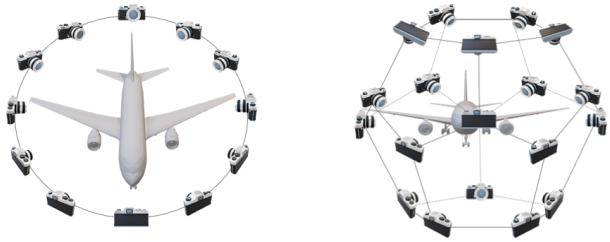


Figure 2. The architecture of the proposed selective multi-view deep model contains five phases. (A) Multi-view extraction: from a given 3D object, *m* multiple views are extracted from different viewpoints and angles. (B) Feature extraction: each extracted view is fed to a pre-trained CNN to extract the corresponding feature stack of the detected visual features. (C) Vectorization: the detected *m* feature stacks are converted to *m* feature vectors. (D) View selection: The feature vectors are compared based on their similarity using Cosine Similarity and give a vital score that is normalized later. The more discriminative view is selected as a global descriptor based on them. (D) Object classification: the global descriptor of the object feeds to a classifier to predict its class.



(a) Circular (12 views). (b) Spherical (20 views).

Figure 3. The two mostly experimented with camera configurations: (a) Circular and (b) Spherical (dodecahedral).

3.4. View Selection

In this phase λ, each feature vector fv_i is assigned an importance score I_i as in Eq. (4) by computing the cosine distance between fv_i and all other feature vectors as in Eq. (5), in a similar fashion as Yang and Wang [25]. The importance score reflects the comparative distinctiveness of each feature vector, with $\varepsilon = 10^{-5}$ ensuring numerical stability by preventing division by zero.

$$I_i = \lambda(fv_i), \quad \text{for } i = 1, 2, \cdots, m. \tag{4}$$

$$I_i = \sum_{j=1}^{m} \left(1 - \frac{fv_i \cdot fv_j}{\max(|fv_i||fv_j|, \varepsilon)}\right) \tag{5}$$

The views' importance scores are normalized to sum to one for each object. The normalization facilitates the comparison of views from the same object and assigns each view a normalized score $\hat{I}_i = Ii / \sum_{j=1}^{m} I_j$, where $i = 1, 2, \ldots, m$. The most informative view is then identified using a selection method ξ in Eq. (6) to create a global descriptor G_k for the object O_k.

$$G_k = \xi\left\{\hat{I}_1, \hat{I}_2, \cdots, \hat{I}_m\right\} \tag{6}$$

In our study, we evaluate two techniques for identifying the best discriminative view for an object. The first method selects the Most Similar View (MSV), which is presumed to be discriminative due to its high similarity and importance score across other views of the same object. Conversely, the second method opts for the Most Dissimilar View (MDV), valuing its distinct and non-redundant features, characterized by its lower similarity and importance score.

Fig. 4 shows examples from the ModelNet40v1 dataset where objects are rendered from 12 views and scored to identify their MSV or MDV for classification. MSVs are marked in green, and MDVs in brown. For objects with similar views, like "Bowl," where importance scores are almost identical, the model may recognize several MSVs but randomly choose one for classification.

3.5. Object Classification

Feature vectors of trained 3D objects train a deep network classifier δ for classification. In the testing phase, δ classifies the global descriptor G_k of each 3D object O_k to determine its class C_k. Two types of networks have been tested as classifier δ: a single-layer Fully Connected Layer (FCL) with softmax activation, and a more complex Fully Connected Network (FCN) as suggested by Seeland and Mäder [13], which has a 1024-neuron fully-connected layer

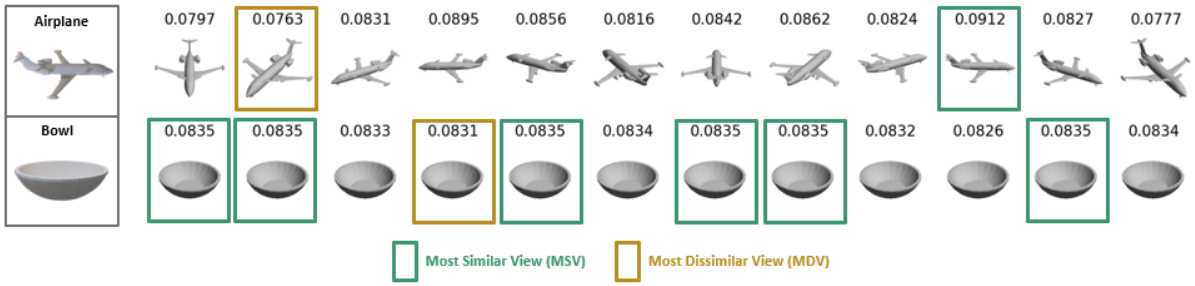


Figure 4. Display of 12 circular views from sample objects with corresponding importance scores: Most Similar Views (MSV) with the highest importance scores in green and Most Dissimilar Views (MDV) with the lowest importance scores in brown.

with ReLU activation and a 0.5 dropout rate for regularization, followed by a softmax layer.

4. Experimental Setup

This section covers 3D datasets, implementation specifics, and evaluation metrics for classification performance.

4.1. 3D Object Dataset

ModelNet40 [24] is a large-scale 3D dataset provided by Wu et al. from Princeton University's Computer Science Department. It contains manually cleaned 3D objects without color information that belong to 40 class categories. In all of our experiments, and for a fair comparison, we have experimented with two versions of that dataset based on the camera settings from the literature:

ModelNet40v1 (Balanced and aligned dataset): in this version, the same training and testing splits of ModelNet40 as in [7, 8, 16, 24] were experimented. Where for each category, they used the first 80 training objects (or all if there are less than 80) for training, and for balanced testing, they used the first 20 testing objects. They used the circular configuration to extract the 12 aligned views. So, they ended up with 3,983 objects consisting of 3,183 training objects (38,196 views) and 800 testing objects (9,600 views).

ModelNet40v2 (Imbalanced and unaligned dataset): here, the whole ModelNet40 as in [5, 8, 23] were experimented. This version is not balanced where there is a diverse number of objects across diverse categories. It contains 12,311 3D objects split into 9,843 for training and 2,468 for testing. The literature used a spherical configuration to extract the 20 unaligned views from each object to end up with a total of 196,860 for training and 49,360 for testing.

4.2. Implementation Details

For each experiment, the classifiers were trained with all the features of the extracted views. In the testing phase, the 3D objects are classified using only the features of a selected view. The learning rate was initialized to 0.0001 and tested twice with 20 epochs as done by [21, 22] (its results in supplementary materials) and with 30 epochs as done by [18, 20, 23]. The network structure was optimized using Stochastic Gradient Descent (SGD) with 0.9 momentum and 0.001 weight decay. At the same time, the batch size is set to 400 and 384 images for the 20-view and the 12-view versions, respectively.

4.3. Evaluation Metrics

The proposed multi-view object classification model is assessed using two accuracy metrics: **Overall Accuracy (OA)**, which measures the proportion of correctly classified samples out of the total test samples, and **Average Accuracy (AA)**, which calculates the mean accuracy per class. While OA and AA are equivalent in balanced datasets like ModelNet40v1, they differ in imbalanced datasets like ModelNet40v2. OA and AA can be calculated as in [11] (see supplementary material for more information).

5. Results and Discussion

The classification accuracy of the proposed models using the ModelNet40v1/v2 datasets are summarized in Tab. 3. The proposed approach achieves the best results, an OA of 83.63% and AA of 83.63%, when only a single view is used for classifying 3D objects. This is observed when the pre-trained ResNet-152 model is employed for feature extraction, and the FCN is used as the classifier, trained with 12 views from ModelNet40v1 dataset (model M_{13} of Tab. 3). Additionally, when the same feature extractor is trained with 20 views from the ModelNet40v2 dataset, the proposed approach with the FCL classifier demonstrates competitive performance, achieving an OA of 83.7%, but with an AA of 80.39% (model M_{15} of Tab. 3).

5.1. Grad-CAM Visualization

Grad-CAM [14] is a technique that generates visual explanations for CNN predictions by highlighting important regions in images using gradient information from the last convolutional layer. This method produces high-resolution, class-discriminative visualizations called guided

Model #	Feature Extractor	Classifier FCN	Classifier FCL	Selected View MSV	Selected View MDV	ModelNet40v1 OA	ModelNet40v1 AA	ModelNet40v2 OA	ModelNet40v2 AA
M_1	VGG-16	✓		✓		78.00%	78.00%	63.25%	53.95%
M_2		✓			✓	69.00%	69.00%	52.87%	41.29%
M_3			✓	✓		80.87%	80.87%	**75.93%**	**70.83%**
M_4			✓		✓	73.30%	73.3%	70.54%	64.47%
M_5	VGG-19	✓		✓		79.50%	79.50%	64.22%	55.48%
M_6		✓			✓	70.50%	70.49%	54.38%	44.51%
M_7			✓	✓		**81.13%**	**81.13%**	75.41%	70.05%
M_8			✓		✓	73.88%	73.88%	70.14%	63.15%
M_9	ResNet-50	✓		✓		82.50%	82.50%	78.24%	71.47%
M_{10}		✓			✓	76.63%	76.63%	69.65%	60.96%
M_{11}			✓	✓		82.00%	82.00%	**83.31%**	**79.12%**
M_{12}			✓		✓	74.88%	74.88%	74.39%	67.64%
M_{13}	ResNet-152	✓		✓		**83.63%**	**83.63%**	80.99%	76.30%
M_{14}		✓			✓	75.50%	75.50%	66.20%	71.15%
M_{15}			✓	✓		82.75%	82.75%	**83.70%**	**80.39%**
M_{16}			✓		✓	75.25%	75.25%	72.53%	64.31%
M_{17}	GoogLeNet	✓		✓		10.25%	10.25%	04.05%	02.50%
M_{18}		✓			✓	10.63%	10.63%	04.25%	03.13%
M_{19}			✓	✓		**71.00%**	**71.00%**	**51.95%**	**45.92%**
M_{20}			✓		✓	66.88%	66.88%	50.00%	44.23%

Table 3. The classification accuracy of our proposed model on ModelNet40v1/v2 datasets is rendered as 12 views and 20 views for each object, respectively. Each model is trained for 30 epochs. The best results are shown in bold and underlined.

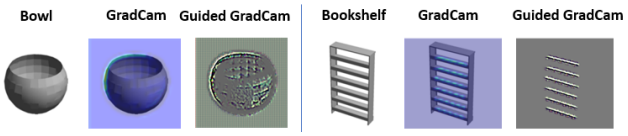


Figure 5. Feature map samples with Grad-CAM highlights indicating regions responsible for correct classification.

Grad-CAM. It can pinpoint relevant areas in an image even when multiple pieces of evidence are present. We apply Grad-CAM to visualize and understand which regions in the views lead to a model's classification decision (Fig. 5). The feature maps show how the proposed model selects the views that contain distinguishing features, such as shelves in bookshelves and circular edges in bowls.

5.2. Predicted Classes Analysis

The confusion matrices of model M_{13} (the best result from the ModelNet40v1 dataset) and model M_{15} (the best result from the ModelNet40v2 dataset), were constructed (provided in supplementary material), which provides a detailed breakdown of the model's predictions across different classes. It has been found that top confusions happened when: i) "flower pot" predicted as "plant", ii) "dressers" predicted as "night stand", and iii) "plant" predicted as "flower pot". As shown in Fig. 6, even for human observers, distinguishing between these specific pairs of classes can be challenging due to the ambiguity present.

5.3. The Effect of the Number of Training Views

We observed that the classification performance is influenced by the number of training views, specifically when using different feature extractors such as VGG-16, VGG-19, GoogLeNet, and ResNet architectures. When the feature extractors VGG-16, VGG-19, or GoogLeNet were utilized, increasing the number of training views resulted in a significant decrease in classification accuracy, ranging from 5.07% to 19.05% in terms of OA. However, when employing ResNet architectures as feature extractors, we noticed a slight increase in classification accuracy as the number of training views increased, albeit by a small margin. The improvement ranged from 0.07% to 0.43% in terms of OA.

5.4. The Effect of the Selected Testing Views

Tab. 3 demonstrate that classification accuracy improves when using the single view MSV as global descriptors for categorizing 3D objects, regardless of any changes in the feature extractor or classifier. This suggests that the MSV (most similar view) is more effective in distinguishing objects than the MDV (most dissimilar view) because it captures the common and shared features found in most extracted views of the same object. So, the final proposed model selects the view with the highest score (MSV), and uses its feature vector to classify the object.

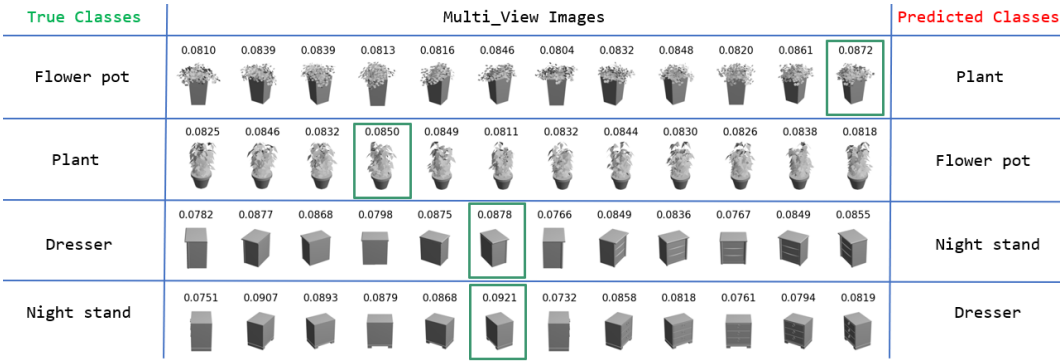


Figure 6. Multi-view samples from ModelNet40v1 dataset of the most wrongly classified objects by the proposed model.

Selective Model	ModelNet Dataset 40v1	40v2	Shaded40	Training Views	Selection Mechanism	Feature Extractor	Accuracy OA	AA
MVCNN [16]	✓			12 views	Random selection	VGG-11	64.28%	-
MVCNN [16]	✓			12 views	Random selection	ResNet-50	48.11%	-
DeepCCFV [7]	✓			12 views	Random selection	VGG11	82.11%	-
DeepCCFV [7]	✓			12 views	Random selection	ResNet-50	70.39%	-
Ours	✓			12 views	Cosine Similarity (MSV)	ResNet-50	82.88%	82.88%
Ours	✓			12 views		ResNet-152	**83.63%**	**83.63%**
Ours		✓		20 views		ResNet-50	83.31%	79.12%
Ours		✓		20 views		ResNet-152	**83.70%**	**80.39%**
Ours			✓	12 views		ResNet-152	**88.13%**	**85.28%**

Table 4. Comparison with the selective view-based 3D object classification methods experimented with a single view. OA is overall accuracy, and AA is average accuracy. The best results are shown in bold and underlined.

Tab. 5 highlights the relationship between importance score and accuracy. MSV consistently provides high accuracy across both datasets, despite occasional outperformance by a few lower-ranked views, which appear to be statistical outliers rather than indicative of a trend. The MDV invariably results in the poorest classification outcomes, reaffirming the importance of view selection in the 3D object classification process. This pattern highlights the critical role that the MSV plays in maintaining robust accuracy levels, suggesting that it holds the most discriminative features necessary for effective classification. In contrast to ViewFormer [18], which employs multi-view attention maps, our approach diverges by utilizing scoring to identify and classify based on the most informative single view. Unlike ViewFormer's technique of weighting and using all views, our approach streamlines the process by selecting only one view with the highest score, thereby improving efficiency and potentially lowering computational demands.

5.5. The Impact of the Pre-trained CNNs

The choice of pre-trained CNN for feature extraction is a key hyperparameter in our model. We assessed various CNN architectures in Tab. 2 with the ModelNet40v1/v2 datasets, and supplementary material includes plots of their best results. ResNet-150 and ResNet-50 yielded the highest performance on ModelNet40v1/v2, respectively. Specifically, with ResNet-150, the model reached 83.63% OA/AA on the balanced ModelNet40v1, and with ResNet-50, it achieved slightly lower scores of 82.88% OA/AA. On the unaligned and more complex ModelNet40v2, ResNet-150 achieved 83.7% OA and 80.39% AA, while ResNet-50 showed 83.31% OA and 79.12% AA. The increase in OA can be attributed to the larger number of samples, as we extracted more views from each object. However, the decrease in AA can be attributed to the imbalanced distribution of samples in the ModelNet40v2 dataset. In contrast, the performance on both datasets significantly dropped when using GoogLeNet. Furthermore, it can be observed that the performance improves as the number of layers increases. This is because using more layers has the potential to capture finer details and features of 3D objects from the rendered 2D views. Conversely, using fewer layers may miss essential features and details, underutilizing the feature extractor's potential for improvement.

5.6. The Role of the Classifiers

The FCL and FCN classifiers have been experimented with as hyper-parameters in the proposed model. The train-

Rank:	1st (MSV)	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th (MDV)
ModelNet40v1	**83.63%**	83.38%	83.13%	83.63%	83.25%	**83.75%**	**84.75%**	82.75%	82.00%	79.13%	78.25%	75.50%
Shaded ModelNet40	**88.13%**	**88.33%**	**88.45%**	87.60%	87.40%	87.03%	86.99%	86.63%	85.25%	85.33%	84.48%	80.67%

Table 5. Classification results for views ranked by importance from 1st (most significant view, MSV) to 12st (least significant view, MDV).

Figure 7. Different shape representations in the multi-view images: a) Original, and b) Shaded multi-view images.

Selective Model	Selected View	Classifier	Shaded ModelNet40 OA	Shaded ModelNet40 AA
Our model	MSV	FCN	**88.13%**	**85.28%**
Our model	MSV	FCL	**88.00%**	**85.95%**
Our model	MDV	FCN	80.67%	76.99%
Our model	MDV	FCL	82.10%	79.25%

Table 6. Results of the proposed model with shaded views.

ing accuracy and loss curves for FCN and FCL from the best-performing experiments are presented in the supplementary material. In testing, the majority of conducted experiments demonstrated that FCL consistently outperformed FCN (highlighted in bold in Tab. 3). Even in cases where FCN showed better performance, the proposed model achieved comparable results when the classifier was replaced with FCL, as observed in models M_{13} and M_{15}.

5.7. The Effect of Shape Representation

Here we investigated the effect of shape representation on the classification of a single view for rendering 3D objects. We utilized the ModelNet40v2 dataset for this experiment, with 12 views per 3D object. However, each 3D object was rendered using the Phong shading technique [2]. Shading techniques have been demonstrated to improve performance by more than 2% in models such as MVCNN [17] where it achieves OA of 95% using all the 12 views. The rendered views were grayscale images with dimensions of 224×224 pixels and black backgrounds, as depicted in Fig. 7. The camera's field of view was adjusted so that the image canvas tightly encapsulated the 3D object.

Tab. 6 displays the results of the proposed model when applied to the shaded ModelNet40 dataset with 12 views, utilizing ResNet-152 as the feature extractor. A comparison with the results presented in Tab. 3 reveals a significant performance improvement, with a margin ranging from 4.3% to 9.57% OA. Specifically, the proposed model classification performance increases from 83.7% to 88.13% OA when the shaded version of the dataset is employed. This demonstrates that enhancing the shape representation through shading can improve the model's performance, even with only a single view for 3D object classification. To ensure a fair comparison of single-view 3D object classification, we evaluated the results obtained by our approach alongside the MVCNN [16] and DeepCCFV [7] models, as reported in [7]. We selected these models because they are deep learning-based (not transformer-based) and were tested in the same settings we explored. When the ModelNet40v1 dataset was used with 12 views, our model outperformed the MVCNN [16] and DeepCCFV [7] models, even without a shading technique (see Tab. 4). It is worth noting that despite all models using ResNet-50 as the feature extractor, our model achieved significantly higher accuracy, with a margin ranging from 13.24% to 35.52% OA. This notable improvement can be attributed to the selection mechanism employed in the proposed model, which utilizes the most similar view (MSV).

6. Conclusion

This study introduces a method for 3D object classification using a single testing view. The proposed approach involves extracting multi-view images from the 3D objects and selecting the most discriminative views using the cosine similarity method. The proposed method was evaluated on the ModelNet40 dataset, considering two camera configurations for multi-view extraction. Additionally, experiments were conducted to investigate the effect of various hyper-parameters on the classification performance of the proposed model. These hyper-parameters included the number of training views, similarity selection mechanisms, pre-trained CNNs, and classifiers. The results demonstrate the effectiveness of the proposed model for 3D object classification using only a single testing view. Future directions involve enhancing its performance by exploring and selecting different numbers of testing views.

7. Acknowledgment

The authors would like to acknowledge the support received from Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM) under SDAIA-KFUPM Joint Research Center for Artificial Intelligence Grant no. JRC-AI-RFP-19.

References

[1] Eman Ahmed, Alexandre Saint, Abd El Rahman Shabayek, Kseniya Cherenkova, Rig Das, Gleb Gusev, Djamila Aouada, and Bjorn Ottersten. A survey on deep learning advances on different 3d data representations. *arXiv preprint arXiv:1808.01462*, 2018. 1

[2] Phong Bui-Tuong. Illumination for computer generated pictures. *CACM*, 1975. 8

[3] Shuo Chen, Tan Yu, and Ping Li. Mvt: Multi-view vision transformer for 3d object recognition. *arXiv preprint arXiv:2110.13083*, 2021. 2, 3

[4] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2018. 1, 2, 3

[5] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2021. 1, 3, 5

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[7] Zhengyue Huang, Zhehui Zhao, Hengguang Zhou, Xibin Zhao, and Yue Gao. Deepccfv: Camera constraint-free multi-view convolutional neural network for 3d object retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8505–8512, 2019. 2, 3, 5, 7, 8

[8] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5010–5019, 2018. 1, 2, 3, 5

[9] Chih-Chia Li and I-Chen Lin. Unpaired translation of 3d point clouds with multi-part shape representation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(1):1–20, 2023. 1

[10] Yeray Mezquita, Alfonso González-Briones, Patricia Wolf, and Javier Prieto. Computer vision: A review on 3d object recognition. pages 117–125, 2023.

[11] Shaohua Qi, Xin Ning, Guowei Yang, Liping Zhang, Peng Long, Weiwei Cai, and Weijun Li. Review of multi-view 3d object recognition methods based on deep learning. *Displays*, page 102053, 2021. 1, 3, 5

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[13] Marco Seeland and Patrick Mäder. Multi-view classification with convolutional neural networks. *Plos one*, 16(1): e0245230, 2021. 4

[14] 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 *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 5

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[16] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 1, 2, 3, 5, 7, 8

[17] Jong-Chyi Su, Matheus Gadelha, Rui Wang, and Subhransu Maji. A deeper look at 3d shape classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 8

[18] Hongyu Sun, Yongcai Wang, Peng Wang, Xudong Cai, and Deying Li. Viewformer: View set attention for multi-view 3d shape understanding. *arXiv preprint arXiv:2305.00161*, 2023. 2, 3, 5, 7

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[20] Wenju Wang, Yu Cai, and Tao Wang. Multi-view dual attention network for 3d object recognition. *Neural Computing and Applications*, 34(4):3201–3212, 2022. 2, 5

[21] Wenju Wang, Gang Chen, Haoran Zhou, and Xiaolin Wang. Ovpt: Optimal viewset pooling transformer for 3d object recognition. In *Proceedings of the Asian Conference on Computer Vision*, pages 4444–4461, 2022. 2, 3, 5

[22] Wenju Wang, Xiaolin Wang, Gang Chen, and Haoran Zhou. Multi-view softpool attention convolutional networks for 3d model classification. *Frontiers in Neurorobotics*, page 255, 2022. 2, 5

[23] Xin Wei, Ruixuan Yu, and Jian Sun. View-gcn: View-based graph convolutional network for 3d shape analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1850–1859, 2020. 1, 2, 3, 5

[24] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 5

[25] Ze Yang and Liwei Wang. Learning relationships for multi-view 3d object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7505–7514, 2019. 4

[26] Zizhao Zhang, Haojie Lin, Xibin Zhao, Rongrong Ji, and Yue Gao. Inductive multi-hypergraph learning and its application on view-based 3d object classification. *IEEE Transactions on Image Processing*, 27(12):5957–5968, 2018. 2

[27] He-Yu Zhou, An-An Liu, Wei-Zhi Nie, and Jie Nie. Multi-view saliency guided deep neural network for 3-d object retrieval and classification. *IEEE Transactions on Multimedia*, 22(6):1496–1506, 2019. 2, 3