3D Models Classification with Use of Convolution Neural Network

Karyna Khorolska, Bohdan Bebeshko, Alona Desiatko and Vitaliy Lazorenko

Kyiv National University of Trade and Economics, Kyiv, Ukraine

Abstract

Nowadays the most urgent challenge behind computer image recognition has become a problem of three dimensional reconstruction of the world or environment from the two dimensional representation like 2D images. Such a tendency is especially obvious in the example of architecture companies' requirements. In common one has no access to the ready-to-use 3D model. Therefore one has to somehow recognize and reconstruct a three dimensional model or object based on its two dimensional representation from different viewports. Novelty can be an approach that allows one not only to use 2D images for feature extraction and classification but also ready 3D models or objects of the same type. In such circumstances one can train neural network recognition models to utilize 3D models, especially their distinguishing features as voxel occupancy or surface curvature. It became obvious that most of the scientists and researchers that are researching the processes in this area commonly develop algorithms of 2D image recognition and extraction features of the same 2D images for further usage in image recognition systems with the aim to classify them and reconstruct the 3D model. Therefore one can build a classifier of three dimensional shapes using not only two dimensional images but also 3D models. Therefore in this paper we propose a new multi presentational 3D model classification framework. Precisely, in this work for the cross presentational information multiple two dimensional images of a three dimensional model as input was used, as well as the extraction of the high level cross presentational information using multiple 2D CNN in separated mode.

Keywords [1]

2D, 3D**,** image recognition**,** models classification**,** convolutional neural network, multi view convolutional neural network

1. Introduction

Nowadays the most urgent challenge behind computer image recognition has become a problem of three dimensional reconstruction of the world or environment from the two dimensional representation like 2D images. Such a tendency is especially obvious in the example of architecture companies' requirements. In common one has no access to the ready-to-use 3D model. Therefore one has to somehow recognize and reconstruct a three dimensional model or object based on its two dimensional representation from different viewports. Therefore model or object classification become a next critical problem in computer image recognition. Classified objects can be very helpful in future for tasks like 3D reconstruction, object detection or object tracking. Traditional approaches to classify objects or models is to extract features (e.g.: SURF, HOG) or to descript and then classify (e.g.: Bayes approach, SVM).

One can outline three main ways of input used to perform three dimensional model classification: 3D voxel, point cloud and multi presentation image.

Concluding above mentioned, it became obvious that most of the scientists and researchers that are researching the processes in this area commonly develop algorithms of 2D image recognition and

Information Technology and Implementation (IT&I-2021), December 01–03, 2021, Kyiv, Ukraine
EMAIL: k.khorolska@knute.edu.ua (A. 1); b.bebeshko@knute.edu.ua (A. 2); desyatko@gmail.com (A. 3); v.lazorenko@knute.edu.ua (A. 4)
ORCID: 0000-0003-3270-4494 (A. 1); 0000-0001-6599-0808 (A. 2); 0000-0002-2284-3418 (A. 3); 0000-0003-4492-3977 (A. 4)

extraction features of the same 2D images for further usage in image recognition systems with the aim to classify them and reconstruct the 3D model.

Novelty can be an approach that allows one not only to use 2D images for feature extraction and classification but also ready 3D models or objects of the same type. In such circumstances one can train neural network recognition models to utilize 3D models, especially their distinguishing features as voxel occupancy or surface curvature. Therefore one can build a classifier of three dimensional shapes using not only two dimensional images but also 3D models. Moreover such approaches become easy to achieve according to the emergency of a large variety of different 3D object repositories (e.g.: Shapeways, TurboSquid and others). Some authors [1] in their scientific studies overviews such an approach by presenting a 3D object classifier built on a DNN architecture that in turn was trained on the voxel models. Basically it was a classifier processing 3D objects to recognize and build up three dimensional shapes. Such an approach is totally consistent. Nevertheless, in this paper another way is proposed: to build a 3D object classifier by utilization of the two dimensional images that are renderings of the 3D model. Using such an approach one can greatly increase performance and outstand the approach with direct 3D representations use. Looking ahead, a convolutional neural network (CNN) that was trained on a N-size set of prerendered 3D model presentations with only a single presentation at a test iteration increases accuracy of the category recognition comparing to the model proposed by authors in their scientific paper [1] that was trained directly on the 3D objects. It should be also noted that by increasing the amount of presentations used at the test iteration one can increase performance, but it requires more computational resources. Although it became obvious that one can concatenate information of a 2D presentation range into a single descriptor using multi-view CNN. Such a descriptor in turn contains such an amount of the information for classification purposes as the full collection of view-based descriptors of the model. In fact it also amplifies efficiency of the retrieval while using both a similar 3D model or a simple picture regardless if it is digital or hand-drawn, without any resorting that tremendously slower methods that are based on pair comparisons of image.

During the last decades Deep Convolutional Neural Networks (DCNN) became popular and took huge advances due to their ability for image classification. Dozens of images are classified using DCNN into thousands of possible categories. Opposed to the single presentation DCNN milti presentation CNN states for learning convolutional models in the parametric settings where a range of presentation data is available. In other worlds - it integrates different presentations' discriminative information, which in turn produce much more exhaustive representation for the sequential learning process. If one does not consider the above mentioned scientific paper [1] which proposes shape descriptors from the voxel-based view of a model through 3D convolutional neural networks, previous researches in field of 3D model descriptors were largely pioneer sketches according to a particular geometric property of the model surface or volume. As an example, model shapes can be interpreted as a histogram or set of model features it was constructed of like distance, angels, triangles and normals [12] that are in turn gathered in the predefined surface points [13] as well as properties of functions defined by volumetric grid [14], local model measured diameters relative to the surface points [15], kernel signatures plotted on the polygon meshes [16] or SIFT and SURF feature descriptor extensions for the voxel grids. Therefore any development of the classifiers or any other machine learning models with teachers on the basis of previously mentioned model descriptors defines a range of accompanying problems to be solved. Biggest issue is that the size of well organized repositories with labeled 3D objects is much more limited than image datasets available for research purposes. To make it clear - Model Net repository counts nearly 170 thousand models, opposed the Image Net database [17] stores millions of labeled images. Next issue is concerning 3D model descriptors themselves, 3D model descriptors are very multidimensional that results in overfitting. As an another example, which is mostly common in computer graphics setups, can be the Light Field descriptor [19]. Light Field descriptor extracts Fourier and geometric parameters set from silhouettes of the object rendered in different viewports. However, the object's silhouette itself can be decomposed into parts and then be represented in the form of the acyclic graph. Authors [20] defined resemblance parametric metrics based on the curves that are matching and therefore grouped similar presentations, that are called 3D model aspect praphs [21]. In study [22] authors attempted to compare human drawn sketches with line drawings of 3D models created from several different presentations based on local Gabor filters. In [23] the author proposed to use Fisher vectors on SIFT model features for representing shapes in human sketches. Nevertheless, mentioned descriptors are mostly narrowly designed and therefore do not fit across different domains.

Concerning the view-based methods, they translate 3D models into a range of 2D representations and then use the features extracted from the two dimensional classification CNN. As an example Multi View Convolutional Neural Network (MVCNN) [2] uses a range of two dimensional representations of the rendered 3D model for its input. Nevertheless view-based descriptors have a range of desirable characteristics: they are relatively low-dimensional in comparison to above mentioned, moreover they are efficient in the course of evaluation process, and solid in terms of 3D models representation artifacts, such as holes, flipped polygon mesh tessellations or uproarious surfaces. However the rendered models representations can be directly matched with any other two dimensional picture, image, sketch or even silhouette. Early research of the view-based model was demonstrated in the scientific paper [18]. Proposed model was able to recognize models by comparison of its appearance in parametric eigenspaces built from large sets of three dimensional models rendered to two dimensional images in various poses, angles and under different illuminations. According to scientific paper [3] proposed decomposition method outsmarts MVCNN but only due to increased computational resource consumption at the training stage. Such decomposition approaches utilize two CNN: one for presentation pair selection purposes and second for pair labeling. Each of the approaches CNN uses CNN(M) model [4], therefore they have to be trained separately. In addition to the MVCNN, RotationNet [5] explores multiple presentations from various angles, taking a part of the entire multi presentation image of a 3D model as an input, and defines the category of the model through the rotation process. Scientific paper [6] defines the multiple presentation group information and in turn proposes the Group View Convolutional Neural Network (GVCNN). GVCNN groups the presentation level features to generate so-called group level features. Further group level features are combined together in order to result in the model level feature. The recursive clustering and pooling layer introduced by the authors in the scientific research [7] was developed to concatenate the multi presentation features, which are in turn providing more exhaustive capabilities for 3D model classification. In the course of this paper one only retrieves information from the range of similar presentations in contrast to MVCNN.

Concerning the volume-based methods, they are simply applying three dimensional CNN on the voxelized shapes of the objects. As an example, authors in research [8] utilize 3D Shape Net and therefore propose the application of the Convolutional Deep Belief Network (cDBN) in order to interpret a three dimensional geometry as a probability distribution over a three dimensional voxel grid. In work [9] authors describe VoxNex as an extension upcasting 2D convolutional neural network kernel to 3D convolutional neural network kernel. In [10] researchers introduced VRN Ensemble presenting deep convolutional network models for modeling generative and discriminative voxel. In the same research authors explore issues of representations based on the voxel utilizing models. In research [11] authors present 3D A Nets developing an adversarial neural network for 3D purposes with the aim to efficiently solve problems concerning processing of the 3D volumetric data. However, every convolution centric model for 3D purposes has a huge disadvantage in terms of excessive technical complexity and even more exorbitant GPU resources requirements.

The most rapid and primitive way to obtain a solution to multi presentation 3D model classification with use of 2D CNN could be to merge all of the presentations of the model in form of features as a single presentation for the neural networks input. But there is one valuable disadvantage - such merged input will result in the reduced consistent interpretability of presentation information among different presentations. However, some presented models perform 2D convolutional neural networks on presentations separately and therefore concatenate them into a pool layer, nevertheless such pooling models commonly disregard the content relationships among different presentations. To solve such issues this paper proposes an idea for learning the discriminative cross presentational information simultaneously, preserving the content relationship among the range of presentations, cross-presentational information. Moreover, the proposed idea integrates the mentioned two sorts of information using milti presentation loss fusion method for end-to-end three dimensional model classification.

Therefore in this paper we propose a new multi presentational 3D model classification framework. Precisely, in this work for the cross presentational information multiple two dimensional images of a three dimensional model as input was used, as well as the extraction of the high level cross presentational information using multiple 2D CNN in separated mode. It is supportive to define the intrinsic attributed information for each presentation. For the cross presentational information the cross presentational information of the single presentation and that of all other various presentations are

utilized in this scientific paper to evaluate the outer products, which in turn obtain the correlation parametric matrices between different parameters of each presentation pair. Further the amplified correlation matrix was captured by the maximization operation at the corresponding locations of obtained correlation matrices in the direction of various presentation pairs. Afterall, one dimensional convolution and completely connected (CC) transformation over the amplified correlation matrix was applied in order to obtain high-level cross presentational information of each presentation. It is obliging to describe the content relationship among the presentations. When the above information was gained, it was merged and given as input parameter into the presentation specific CC layer, which in order obtains the presentation specific loss value as well as label prediction. For the cross presentational loss fusion method, a Z_0 constrained optimization problem was formulated with the regard to the weights of the various presentations and therefore obtained the optimal weight distribution. It was valuable to select different discriminative and informative presentations determined by the high weights and use their corresponding predictions to build a joint decision. The main goals of this research may be concluded as following:

- Present and propose a new multi presentation framework that stores the discriminative information with its relationships among presentations for different presentations and designs presentation set mechanism with use of the multi presentational fusion method in order to perform end-to-end 3D models classification.
- Propose the discriminative information with relationships by merging cross presentational and presentational information itself, where presentational information itself is generated as a result of one dimensional convolution as well as CC transformation application over the amplified correlation matrix which in turn was gained by the outer product and presentation pair pooling.
- Moreover, a multi-presentational loss fusion method was proposed by solving a Z_0 constraint optimization to build a joint decision for inferring the category.

2. Previous researches

For instance, in scientific research [2] authors describe MVCNN where multi presentation images were obtained by the 3D rotations passed through shared convolutional neural networks separately, therefore merging at the presentation pool layer and further used as input parameter for another convolutional neural network. Nevertheless, the disadvantage of MVCNN is that its pooling layer disregards the divergence between different presentations, where some of the presentations are distinctive while others have common information. In [6] authors proposed Group View Convolutional Neural Network introducing the presentation, the group, and the shape level descriptor, therefore providing a grouping scheme to divide the presentations in terms of the discrimination weights. However, the parameters of the thresholds of grouping weights inside the grouping module cannot be guided by more discriminative information.

Convolutional Neural Networks have shown promising results for 3D geometry prediction. They can make predictions from very little input data such as a single color image. A major limitation of such approaches is that they only predict a coarse resolution voxel grid, which does not capture the surface of the objects well. We propose a general framework, called hierarchical surface prediction (HSP), which facilitates prediction of high resolution voxel grids. The main insight is that it is sufficient to predict high resolution voxels around the predicted surfaces. [32]

As for the 2D image classification and data extraction where also several scientific works.

Fei-Fei, L., Fergus, R., & Perona, P. used the Caltech 101 set that was among the first standardized datasets for multi-category image classification, it contains 101 object classes with generally 30 training sample images per single class. Later the set was reworked by Gregory and Holub Griffin and Alex and Perona Pietro's and became Caltech 256, which is the set with the increased number of object classes to 256 and added images with greater scale and background variability. However, since all of the mentioned data sets have not been manually verified, they contain many errors that are in a way making it unsuitable for precise algorithm evaluation activities.[28, 33]

Few publications [31, 32] states that it is possible to generate a 3D model, using a single input image and convolutional neural networks (CNN) (Figure 1). One can assume that it can be considered as simple enough for the third consumer-friendly requirement. The assumption that requires testing is the

limitation of the output resolution and swiftness of generation time, which has not been ever specified in any of the papers mentioned above. The key purpose of this work is to define the possible limitations of convolutional neural networks usage for 3D models generation, taking into account output resolution and generation swiftness. Moreover, this work also focuses on balancing, between an increase in the resolution quality and time consumption for output generation.

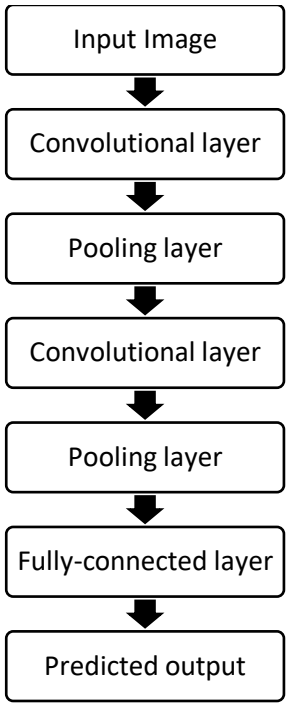

Figure 1: Visual Representation of the Convolutional Neural Network Algorithm

3. Formulation of the proposed method

As it was mentioned above, the goal in this research is to develop a presentation based descriptor for 3D models that are trainable, produce informative representations for recognition and evaluation problems, that won't be complex and will be efficient to compute. In this section of the article, the proposed method is illustrated in detail, which is a joint multi-presentation 2D CNN learning framework aimed to integrate the cross presentation and presentation information of the 3D models by the multi presentation convolutional representation with loss fusion. The input data of the proposed method is rendered by multiple 2D presentations of a 3D model, which belongs to the presentation based approach. 12 rendered presentations were created by placing 12 virtual cameras around the mesh every 30 degrees. The reason for rendering from such viewports is that it is unknown exactly which one can provide a good representative overview of the model. In this research one use of multiple 2D presentations to describe a 3D model and one 2D image per presentation. It was found that the multi presentation representation contains rich information of 3D models and can be applied to various practical problems. For the CNN features, it used the Res Net 18 [24] as the base architecture which consists of 17 convolutional layers followed by one CC layer, in order to capture the cross presentation information for each presentation. The Res Net-18 was pre-trained on the ImageNet repository image set consisting of 1000 categories and then was calibrated on all 2D presentations of 3D models in the training set. The CNN features can capture the high-level information for each presentation, which results in better performance on classification compared with some previously proposed descriptors [25-27, 29-30].

Based on the above subsection, given a 3D model, one first takes a set of two dimensional input images captured from different angles, and each image is passed through a 2D CNN to get the high level representation in the presentation level. Assumed that $x^v \in R^{H\times W\times D}$ and $x^v_{cross} = f_{cnn}(x^v) \in R^{D_{cross}}$ are the input image and the learned features before CC layer by CNN from the v-th presentation,

respectively, where H, W, and D determine the height, width, and channel. For the v-th presentation, a set S_v which contains different presentation pairs with respect to the v-th presentation was defined, resulting in,

$$S_v \left| M_{v=1} = \{(v, \bar{v})\}_{\bar{v}=\frac{\{1,\ldots,M\}}{v}} \right| M_{v=1,} \tag{1}$$

where M is the number of two dimension input images and $(v, \bar{v})$ = ($\bar{v}$, v). Therefore, the proposed presentation information for the v-th presentation x_{cross}^{v} across presentations can be calculated by using the outer product, presentation pair pooling, and one dimension convolution. Described as:

$$x_{en}^{v,\bar{v}} = x_{cross}^{v} \otimes x_{cross}^{\bar{v}} \tag{2}$$

$$x_{en}^{S_v} = \left\{x_{en}^{v,\bar{v}}\right\}_{\bar{v}=\frac{\{1,\ldots,M\}}{v}} \tag{3}$$

$$x_{cross}^{v} = f_{conv}(\Gamma(x_{en}^{S_v})) \tag{4}$$

where $x_{en}^{v,\bar{v}} \in R^{D_{cross} \times D_{cross}}$ defines the outer product of a presentation pair $(v, \bar{v})$, which captures correlations by multiplying each element of x_{cross}^{v} by each element of $x_{cross}^{\bar{v}}$. Extending to all the presentation pairs of the v-th presentation, $x_{en}^{S_v} \in R^{D_{cross} \times D_{cross} \times (M-1)}$ stores the correlation information of the v-th presentation with respect to other M−1 presentations. Moreover, $\Gamma(x_{en}^{S_v}) \in R^{D_{cross} \times D_{cross}}$ maximizes the correlations of M −1 presentation pairs in S_v along the direction of different presentation pairs for the v-th presentation, where Γ is the presentation pair pooling operation. Therefore, the high-level cross presentation information $x_{cross}^{v} \in R^{D_{cross}}$ is generated by applying f_{conv} over $\Gamma(x_{en}^{S_v})$, which consists of two steps that transforms each row of $\Gamma(x_{en}^{S_v})$ into a K-dimension vector by applying a one dimensional convolution (with kernel size=1) and merging D_{cross} K-dimension vectors to project into a D_{cross}-dimension vector (x_{cross}^{v}) through a CC layer.

Afterwards, x v intra and x v inter were combined by a merging operation and then used as input parameters for the CC layer in order to obtain the corresponding loss and label prediction of each presentation. Described as,

$$x_{con}^{v} = f_{cat}(x_{cross}^{v}, x^{v}) \tag{5}$$

$$z^{v} = f_{cc}(x_{con}^{v}) \tag{6}$$

where $x_{con}^{v} \in R^{D_{cross}+D}$ defines the comprehensive information of each presentation and $z^{v} \in R^{N_c}$ is produced by f_{cc} with input x_{con}^{v}, indicating the probability distribution over the possible classes for each presentation, and N_c is the number of categories. After, it was proposed a new adaptive-weighting loss fusion method with proper meager for multiple predictions $z^{v}|M_{v=1}$ to build a joint decision and implement the multi presentation 3D model classification, which can be shown as,

$$\alpha \top 1 = 1, \alpha \geq 0, \left|\left|\alpha\right|\right|_0 = s \sum_{v=1}^{M} \alpha_v^{\gamma} L^{v}(z^{v}, y) \tag{7}$$

$$L^{v}(z^{v}, y) = -log\left(\frac{exp(z_y^{v})}{\sum_{o=1}^{N_c} exp(z_o^{v})}\right) \tag{8}$$

where $\alpha \in R^{M}$ is a weight vector corresponding to multiple presentations, $y \in R$ defines the common label information of all the presentations for an object, and $L^{v}(z^{v}, y) \in R$ is the cross-entropy loss of the v-th presentation. $\gamma > 1$ is the power exponent parameter of the weight α_v, which adjusts the weight distribution of different views flexibly and avoids the trivial solution of α during the classification. $\left|\left|\alpha\right|\right|_0$= s is used to constrain the sparseness of the weight vector α, where $s \in N_+$ denotes

the number of nonzero elements in α. Crucially, the Z_0-norm constraint is able to capture the global relations among different views and is able to achieve presentation-wise sparseness such that only a few discriminative and informative views are selected during the optimization to make decisions

Afterwards, the proposed method on the Model Net 40 dataset was evaluated and comparisons with several state-of-the-art methods was done. As it is common - classification in 3D is mainly based on the Computer Aided Design (CAD) model. Only one widely used repository is Model Net [8] has more than 130 thousand 3D CAD models from over 600 categories. Model Net 40 [27] provided on the Princeton Model Net website is a subset of the Model Net and has around 12 thousand models from 40 common categories. Figure 1 selects 4 kinds of simple categories to intuitively show 6 2D presentations rendered from 3D models, where 6 presentations are generated from 360 degrees with an interval of 60 degrees (for the experiment itself one used 12 presentations of the same models with an interval of 30 degrees).

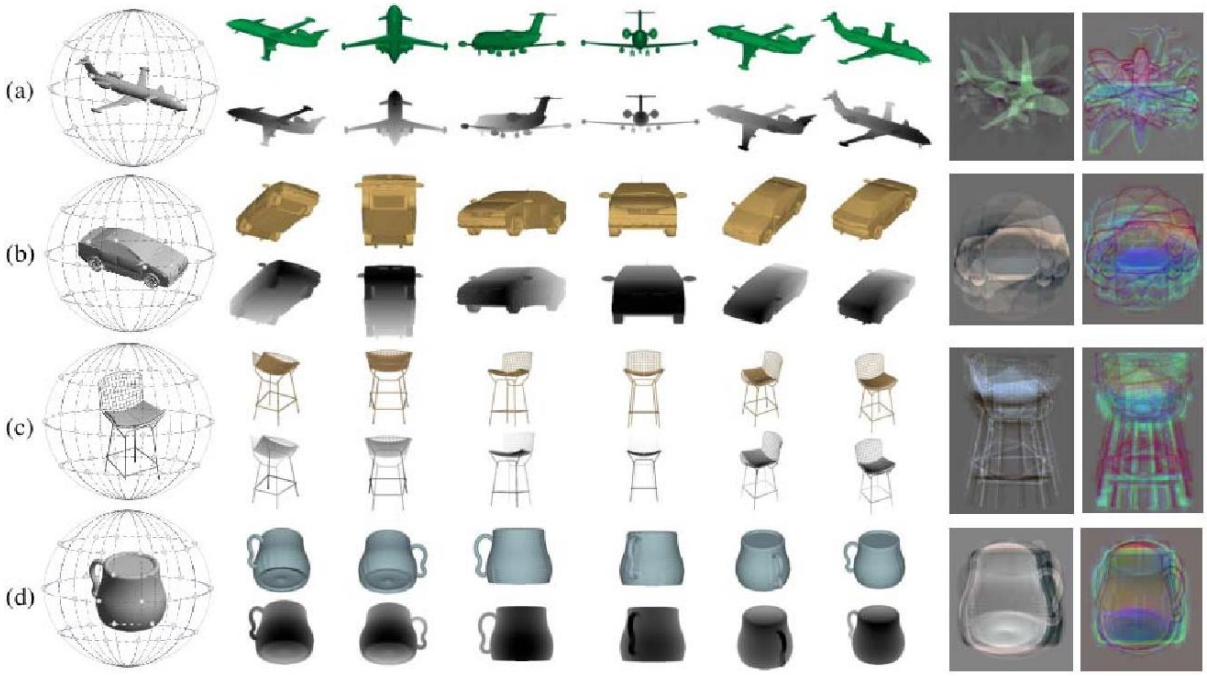

Figure 2: Example presentations generated by categories of 3D models in Model Net 40

Proposed method was compared with several state-of-theart methods for multi presentation 3D models classification, including both presentation and volume-based methods MVCNN Multi Res, Minto, common volume-based methods 3DShapeNets, common point-based methods PointNet , and common presentation based methods MVCNN.

With consideration of the FLOPs of the Deeper Convolutional Neural Network and a better trade-off between accuracy and memory consumption levels compared to other classical CNN Res Net 18 is a good choice but not limited to this convolutional neural network architecture. To evaluate the base architectures, the results of Multi View Convolutional Neural Network were compared with ResNet-18, whose results are shown in Table 1. Evidently, the usage of the Res Net 18 can result in the performance increase of Multi View Convolutional Neural Network. For example, Multi View Convolutional Neural Network (ResNet18) with 12 views achieves 3.3% improvements compared with standalone Multi View Convolutional Neural Network. For proposed method, the parameters of ResNet-18 were calibrated using the Model Net 40 dataset and use Adam with $learning\ rate = 5 * 10 - 6$, $\beta 1 = 0.7$, $\beta 2 = 0.933$, $weight\ decay = 0.0001$, $batch\ size = 8$, $epoch = 60$ for optimization. Furthermore, there are two parameters s and γ in the proposed method, where s represents the number of nonzero elements in α and γ is the power exponent of each element of α. s was calibrated in the range of [6, 12] with step 1 to select a few discriminative and informative presentations to build a joint decision during classification. γ was varied from 1.5 to 10 with a step of 1 to explore the influence on different values of γ on classification accuracy. Based on the proper parameters $s = 9$

and $\gamma = 2.5$, one can train an optimal model to improve the performance of classifying 3D models tremendously (see Table 1).

Table 1
Accuracy results using existing methods and proposed method

Method	Accuracy
MVCNN	89.90%
MVCNN (ResNet18)	92.20%
Proposed (ResNet18)	93.01%

The performance of different modules of the proposed method was evaluated and reported in Table 2. The demonstrated researches outline that the weight distribution α, the sparsity of multiple presentations, and the cross presentation information for any different presentations play different roles during classifying 3D models.

Table 2
Accuracy results using different approaches to the proposed method

Method	Accuracy
Proposed (merge)	88.02%
Proposed (α)	92.61%
Proposed (s)	92.23%
Proposed (α + s)	92.98%
Proposed (α + s+cross)	93.01%

4. Conclusions

First, all of the multi presentation methods outperform the single presentation method, which verifies the advantages of multi presentation representations. Second, the classification accuracy of proposed $(\alpha + s)$ is better than that of proposed (α) and proposed (s), respectively. It is obvious that considering the weight distribution and the sparsity of multiple presentations simultaneously is reasonable and effective. Finally, proposed $(\alpha + s + cross)$ obtains better performance than any other method, which shows that cross presentation information across presentations also plays an important role. The experimental results of different methods and their comparisons are reported in Table 3. The proposed method $(\alpha + s + cross)$ with 12 presentations achieves the best classification accuracy. Firstly, compared with the 'view,volume'- based methods, for example MVCNN-MultiRes, the proposed method gains 0.81% improvements. It is obvious that the inputs of these methods contain both 2D and 3D information, however making them work well with each other needs to be improved. Secondly, compared with the volume-based methods, the proposed method obtains 7.28% improvements. It is found that these volume-based methods also cannot address 3D volumetric data processing effectively. Thirdly, making comparisons between the points-based methods and proposed methods, the performance of classifying 3D models can be achieved by 4.67% improvements. However, the problem of effectively modeling point clouds still needs to be solved. This verifies the superiority of the proposed method at merging the cross presentation information and a selective and adaptive weighting strategy into a unified multi presentation framework. Weights of different views on Model Net 40 dataset were evaluated. The higher weight indicates that the presentation provides more valuable information and makes more contributions during the multi presentation 3D model classification. In this paper, a new 2D CNN based multi presentation framework for 3D object classification was proposed. It takes the multiple 2D images rendered from the 3D CAD model as the inputs. It not only contains the discriminative information with relationships among presentations but also provides a novel presentation merging mechanism for fusing multiple presentations to build a joint decision for classifying 3D models. The experimental results verify the superiority and effectiveness of the proposed method in 3D modes classification.

Table 3
Accuracy comparison using different methods

Method	Input	Amount of presentations	Accuracy
MVCNN(MultiRes)	view,volume	-	92.20%
3D Shape Net	volume	1	85.73%
Point Net	points	1	88.34%
Proposed (α + s+cross)	view	12	93.01%

5. References

[1] H. Zeng, T. Zhao, R. Cheng, F. Wang and J. Liu, "Hierarchical Graph Attention Based Multi-View Convolutional Neural Network for 3D Object Recognition," in IEEE Access, vol. 9, pp. 33323-33335, 2021, doi: 10.1109/ACCESS.2021.3059853.

[2] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In ICCV, pages 945–953, 2015. https://arxiv.org/abs/1505.00880

[3] Johns, E., Leutenegger, S., & Davison, A.J. (2016). Pairwise Decomposition of Image Sequences for Active Multi-view Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3813-3822. DOI:10.1109/CVPR.2016.414

[4] Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the Devil in the Details: Delving Deep into Convolutional Nets. *ArXiv, abs/1405.3531*. DOI:10.5244/C.28.6

[5] Kanezaki, A., Matsushita, Y., & Nishida, Y. (2018). RotationNet: Joint Object Categorization and Pose Estimation Using Multiviews from Unsupervised Viewpoints. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5010-5019. DOI:10.1109/CVPR.2018.00526

[6] Feng, Y., Zhang, Z., Zhao, X., Ji, R., & Gao, Y. (2018). GVCNN: Group-View Convolutional Neural Networks for 3D Shape Recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 264-272. DOI:10.1109/CVPR.2018.00035

[7] Wang, C., Pelillo, M., & Siddiqi, K. (2017). Dominant Set Clustering and Pooling for Multi-View 3D Object Recognition. ArXiv, abs/1906.01592. DOI:10.5244/C.31.64

[8] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., & Xiao, J. (2015). 3D ShapeNets: A deep representation for volumetric shapes. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1912-1920. DOI:10.1109/CVPR.2015.7298801

[9] D. Maturana and S. Scherer, "VoxNet: A 3D Convolutional Neural Network for real-time object recognition," *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 922-928, doi: 10.1109/IROS.2015.7353481.

[10] Brock, A., Lim, T., Ritchie, J.M., & Weston, N. (2016). Generative and Discriminative Voxel Modeling with Convolutional Neural Networks. *ArXiv, abs/1608.04236*.

[11] Mengwei Ren, Liang Niu, and Yi Fang. 3d-a-nets: 3d deep dense descriptor for volumetric shapes with adversarial networks. arXiv preprint arXiv:1711.10108, 2017.

[12] B. K. P. Horn, "Extended Gaussian images," in *Proceedings of the IEEE*, vol. 72, no. 12, pp. 1671-1686, Dec. 1984, doi: 10.1109/PROC.1984.13073.

[13] Ma, Lingfei & Li, Ying & Li, Jonathan & Tan, Weikai & Yu, Yongtao & Chapman, Michael. (2019). Multi-Scale Point-Wise Convolutional Neural Networks for 3D Object Segmentation From LiDAR Point Clouds in Large-Scale Environments. IEEE Transactions on Intelligent Transportation Systems. PP. 1-16. doi: 10.1109/TITS.2019.2961060. .

[14] Wang, Wenju & Cai, Yu & Wang, Tao. (2021). Multi-view dual attention network for 3D object recognition. Neural Computing and Applications. 1-12. doi: 10.1007/s00521-021-06588-1.

[15] Chaudhuri, Siddhartha & Koltun, Vladlen. (2010). Data-Driven Suggestions for Creativity Support in 3D Modeling. ACM Transactions on Graphics. 29. doi: 10.1145/1866158.1866205..

[16] Kokkinos, I., Bronstein, M.M., Litman, R., & Bronstein, A.M. (2012). Intrinsic shape context descriptors for deformable shapes. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 159-166. DOI:10.1109/CVPR.2012.6247671
[17] Mogalapalli, Harshit & Abburi, Mahesh & Balan, Nithya & Bandreddi, Surya. (2022). Classical–Quantum Transfer Learning for Image Classification. SN Computer Science. 3. doi: 10.1007/s42979-021-00888-y.
[18] Vyas, Shantanu & Chen, Ting-Ju & Mohanty, Ronak & Jiang, Peng & Krishnamurthy, Vinayak. (2021). Latent Embedded Graphs for Image and Shape Interpolation. Computer-Aided Design. 140. 103091. doi: 10.1016/j.cad.2021.103091.
[19] Wang, Wenju & Cai, Yu & Wang, Tao. (2021). Multi-view dual attention network for 3D object recognition. Neural Computing and Applications. 1-12. doi: 10.1007/s00521-021-06588-1.
[20] Yin, Junjie & Huang, Ningning & Tang, Jing & Fang, Mei-e. (2020). Recognition of 3D Shapes Based on 3V-DepthPano CNN. Mathematical Problems in Engineering. 2020. 1-11. doi: 10.1155/2020/7584576.
[21] Koenderink, J.J., van Doorn, A.J. The singularities of the visual mapping. Biol. Cybernetics 24, 51–59 (1976). doi: 10.1007/BF00365595
[22] Eitz, Mathias & Hildebrand, Kristian & Boubekeur, Tamy & Alexa, Marc. (2010). Sketch-Based Shape Retrieval. ACM Transactions on Graphics - TOG. 31. doi: 10.1145/2185520.2185527..
[23] Schneider, Rosalia & Tuytelaars, Tinne. (2014). Sketch Classification and Classification-driven Analysis Using Fisher Vectors. ACM Trans. Graph.. 33. 174:1-174:9. doi: 10.1145/2661229.2661231.
[24] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. DOI:10.1109/cvpr.2016.90
[25] Zhao, Yue & Wu, Yuwei & Chen, Caihua & Lim, Andrew. (2020). On Isometry Robustness of Deep 3D Point Cloud Models under Adversarial Attacks. CVPR2020 https://arxiv.org/abs/2002.12222
[26] Bebeshko, B., Khorolska, K., Kotenko, N., Kharchenko, O., & Zhyrova, T. (2021). Use of neural networks for predicting cyberattacks. Paper presented at the CEUR Workshop Proceedings, 2923 213-223. http://ceur-ws.org/Vol-2923/paper23.pdf
[27] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In CVPR, pages 1912–1920, 2015. https://arxiv.org/pdf/1406.5670.pdf
[28] Bebeshko, B., Khorolska, K., Kotenko, N., Desiatko, A., Sauanova, K., Sagyndykova, S., & Tyshchenko, D. (2021). 3D modelling by means of artificial intelligence. Journal of Theoretical and Applied Information Technology, 99(6), 1296-1308. http://www.jatit.org/volumes/Vol99No6/5Vol99No6.pdf
[29] Kumar, Amarjeet & Hampson, Gary & Rayment, Tom. (2021). Adaptive subtraction using a convolutional neural network. First Break. 39. 35-45. 10.3997/1365-2397.fb2021066.
[30] Khorolska K., Lazorenko V., Bebeshko B., Desiatko A., Kharchenko O., Yaremych V. (2022) Usage of Clustering in Decision Support System. In: Raj J.S., Palanisamy R., Perikos I., Shi Y. (eds) Intelligent Sustainable Systems. Lecture Notes in Networks and Systems, vol 213. Springer, Singapore. https://doi.org/10.1007/978-981-16-2422-3_49
[31] Wu J., Zhang C., Xue T., Freeman W. T., Tenenbaum J. B. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling, In NIPS, pages 82–90, 2016.
[32] Häne, C., Tulsiani, S., & Malik, J. (2017). Hierarchical Surface Prediction for 3D Object Reconstruction. *2017 International Conference on 3D Vision (3DV)*, 412-420. DOI:10.1109/3DV.2017.00054
[33] Igor Smirnov, Alexey Kutyrev, Nikolay Kiktev.Neural network for identifying apple fruits on the crown of a tree. E3S Web Conf. 270 01021 (2021) DOI: 10.1051/e3sconf/202127001021