A Progressive Multi-view Learning Approach for Multi-loss Optimization in 3D Object Recognition

Shitala Prasad, *Member, IEEE*, Yiqun Li, Dongyun Lin, Sheng Dong, Ma Tin Lay Nwe

***Abstract*—3D object recognition is a well studied 2D multi-view object classification task that achieves high accuracy if the object textures are distinctive. However, if objects are texture-less and are only differentiable by their shapes but at certain viewpoints. Thus, the problem is still very challenging. Furthermore, the existing methods are mostly based on supervised learning with lots of images per object which are difficult to collect and label them for training. In this letter, we introduced a multi-loss view invariant stochastic prototype embedding to minimize and improve the recognition accuracy of novel objects at different viewpoints by using a progressive multi-view learning approach. An extensive experimental results show that the proposed method outperforms the state-of-the-art methods on different types datasets and also on different backbones.**

***Index Terms*—3D Unseen Learning, DCNN, Progressive Multi-view Learning, Object Detection, Self-Supervised Learning**

I. Introduction

As we know, objects in real world are represented in a 3-dimensional space while the current visual intelligence is mostly tested on 2-dimensional images that are captured from different views of object [1], [2]. Thanks to deep convolutional neural network (DCNN), which rapidly encapsulated computer vision (CV) in supervised learning and achieve human-level performance to some tasks [3], [4]. However, despite of DCNN's great success in the field of vision [5], [6], [4], using 2D images to recognize 3D objects is still challenging due to viewpoint variation in shapes. In real vision system, as stated in [1], the viewer-centered representation plays a vital role in object level recognition and so are mostly followed by several psychophysical and computational researchers [7], [8], [9].

In digital CV, to obtain 3D object recognition there are several multi-view attempts for object recognition. In past, author try to aggregate the features obtained from multiple views either by using recurrent neural networks [10] or by integrating graph with other modalities [11], [12]. In practice, it's not always feasible to obtain a dense view of 3D object which covers all the visual aspects of it. There are researchers focusing on the missing viewpoints for recognition in inference phase while a complete view set is assumed for training [13]. These approaches works very well on the seen classes but suppressively performances for the unseen categories [11].

The industrial or mechanical parts are mostly texture-less and sometimes even the shapes are similar. For example, washers may only vary in their dimension. On the other hand, in many real world applications it's not always possible to re-train the network whenever a new product, part or design is released. Therefore, a prior knowledge of feature representation of similar objects will ease the recognition deployment process. This is usually resolved *via* embedded learning where the unseen classes are well patterned for classification. However, the images used for training are not evenly sampled from all the possible viewpoints of an object, thus the viewpoint difference between the training and testing sets will affect the existing system's performance. To code this, multi-view self-supervised learning (MVSSL), zero-shot learning (ZSL) or learning from side information are used by [14], [15].

To fathom above problems, Ho *et al.* has introduced a lightweight unsupervised multi-view object recognition embedding using MVSSL which they termed as view invariant stochastic prototype embedding (VISPE) [14]. They proposed a randomizer where *softmax* parameters are sampled stochastically from the embedding space of the object's viewpoint during the training, which suppose to simplify the embedding. The randomization in VISPE actually complemented the invariants and enable the classifier to remain stable whatever the viewpoint is. However, we found that the 3D object representations are still not robust for both seen and unseen classes. A special attention mechanism is required to deal with such multi-view object representation. In this letter, we introduce a classroom-based multi-loss function to optimize the multi-view object representation and improve the recognition performance.

Furthermore, we optimize the deep feature embedding which is suitable to represent unseen objects and thus we improve the embedding space by introducing a progressive multi-view learning approach. An extensive ablation studies on popular datasets prove that our methods are able to make the network converge faster and show a significant improvement in terms of recognition and computational cost. We further used the largest mechanical components benchmark (MCB) dataset [16] for novel 3D object recognition and found a consequential improvement. Lastly, we investigate our hypothesis over various state-of-the-art (SOTA) deep networks that are used for image classification.

In this letter, the further sections are arranged as below: Section 2 reviews the related work which is then followed by the proposed method and the serious of experiments that are performed on different datasets in Sections 3 and 4, respectively. The letter is then summarized with the key future works in Section 5.

This research/project is supported by A*STAR under its RIE2020 INDUSTRY ALIGNMENT FUND - INDUSTRY COLLABORATION PROJECTS (IAF-ICP) Grant No. I2001E0073.

The authors are currently with Visual Intelligence department at Institute for Infocomm Research (I^2R), A*STAR, Singapore. Their corresponding mailing address are: {shitalap, yqli, lin_dongyun, dong_sheng, tlnma}@i2r.a-star.edu.sg

II. Related Work

This segment briefly highlights few prior works related to multi-view object recognition that includes ZSL, context-aware and knowledge graph based recognition, multi-view recognition (MVR) and self-supervised learning (SSL).

A. Zero-Shot Learning

A wide group of researchers working on using external knowledge based ZSL to learn visual representation of unseen object classes [17], [18]. In addition to ZS recognition, ZS detection is also explored widely, aiming to localize objects that are never seen before [19]. Among various ZS approaches, some focus on proposal generation and some on loss optimization for novel categories. Cacheux *et al.* used ZSL for object recognition in deep feature space with various tricks [20]. Inspite of getting a satisfactory performance, ZS recognition and detection tasks are still limited in terms of context information in scenes.

B. Context-aware and Knowledge Graph Based Recognition

In traditional object detection, context plays major role before DCNN [21], [22], [23]. Whereas for weakly-supervised object detection, researcher uses common-sense based knowledge graph to optimize the detector [24]. These graph-based NNs often propagate information over the knowledge graph [25]. Such methods are mostly designed for fully-supervised settings and therefore cannot be directly applied to ZS environment.

C. Multi-view Recognition

As we know, multi-view recognition is a 2D image-based object recognition task which takes multiple views of an object and perform feature embedding using DCNN such as MVCNN [26] and MLVCNN [27]. These approaches are highly supervised and observed that the viewpoints are interrelated. Thus, recurrent NNs and graph-based DCNNs were proposed by [28] and [11] to resolve the above issues. But it is observed that such methods experience performance drop when partial views are included in inference which to some extent is decoded by considering views as an intermediate state or hierarchical embedding [14]. Therefore, in this letter we include partial viewpoints progressively for training which relax the above constraints, to some extent and enhances the recognition performance (see Section 4).

D. Self-Supervised Learning

In SSL, context-based, motion-based, sequence-based and view-based object recognition methods are now widely imported in deep learning [29]. In cluster-based SSL methods, data are grouped with some visual similarities into clusters which discriminates them [30]. In this approaches, the feature maps are updated once per epoch that might be a noisy generalization for 3D object recognition due to viewpoints. Therefore, multiple views are used to avoid this problem [14]. This letter optimization the multi-view embedding by introducing progressive multi-view learning. This proposed approach also helps in reducing the overall training cost.

III. Proposed Methodology

In this section, we detailed our proposed progressive multi-view embedded learning approach to optimize the overall performance of multi-view unseen object recognition by introducing a new classroom-based multi-loss concept. The overall flow-chart of the proposed concept is shown in Figure 1.

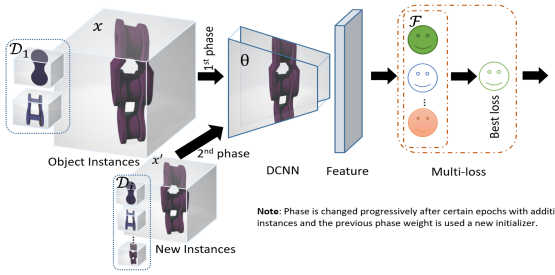


Fig. 1: The overview of our proposed approach.

A. Problem Definition

According to Ho *et al.* [14], the goal of light weight unsupervised multi-view object recognition is to learn the embedding in such a way that it can recognize new objects and new set of views of an objects $\mathcal{O}_k = \{o_i\}_k$, where o_i is the i-th object instance of k-th object. In DCNN-based embedding, the parameters θ are optimized by minimizing the risk factor $\mathcal{R}$ over N images in dataset $\mathcal{D}$:

$$\mathcal{R} = \sum_{i,j} -log \frac{\exp w_i^T f_\theta(x)}{\sum_{k=1}^{N} \exp w_k^T f_\theta(x)} \tag{1}$$

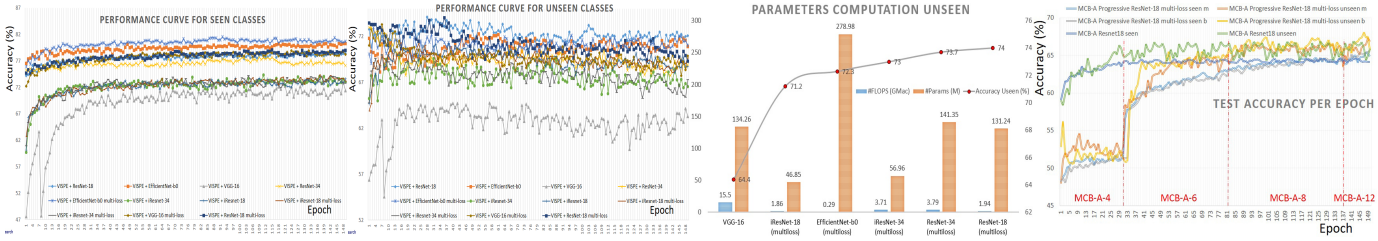
In *softmax*-based DCNN, the learning feature f_θ for image x is generally limited to the seen classes and therefore, metric learning based embeddings are more emphasized for unseen objects [31]. However, they agonize sample pairings in $\mathcal{D}$ and thus, requires a dedicated sampling methods which are hard to converge. Ho *et al.* used a randomization technique in multi-view datasets which actually updates the embedding along with the task, leading to a good embedding feature. This generalization is very useful for the unseen categories to learn instance classifier parameters w_i in Eq. 1.

The next challenge in 3D object recognition is that the embedding images of the same object are not very tightly clustered and so can be confused by other class(es). For this, a normalized embedding space is practised, *i.e.*, $(f_\theta(x) \rightarrow ||f_\theta(x)||_2 = 1)$ [14]. Thus, the weights in Eq. 1 can be replaced by randomly selected object's view embedding for the instance o_i. Hence, the *softmax* can be re-written as:

$$\mathcal{R} = \frac{\exp f_\theta(x_i^{v_i})^T f_\theta(x)/\tau}{\sum_{k=1}^{N} \exp f_\theta(x_i^{v_i})^T f_\theta(x)/\tau} \tag{2}$$

where τ controls the sharpness of posterior distribution and $v_i \epsilon \{1, 2, ...V_i\}$ is the view sampler per object instance i.

Further, the view invariant embedding is strengthened by minimizing the magnitude of variations in distribution which is the ultimate measure of view sensitivity. That is, if the distribution (Eq. 1) remains stable with the set of prototype, the embedding becomes more stronger. In [14], Kullback-Leibler



(a) (b) (c) (d)

Fig. 2: Comparison graphs for seen and unseen classes (a-b) and the improvement *vs* computation cost (c) for ModelNet40 test set is shown; and (d) shows with and without progressive multi-view learning for MCB-A. For best view, zoom 400%.

(KL) divergence is used to minimized this and so termed as VISPE.

Finally, the risk of classifying a training view $x^j_{\eta_i}$ of object instance label η_i of prototype set $s'_p \epsilon \mathbf{S} | s' = \{f^{\eta_i}_{\eta_i}\}^m_{k=1}$, is defined as:

$$\mathcal{L}_{s'_p}(i,j) = -log(P^{s'_p}_{Y|X}(\eta_i|x^j_{\eta_i})) \tag{3}$$

where $p = \{1, 2\}$. Using similar notations, the KL divergence is defined as:

$$\mathcal{L}_{KL} = \sum_{k=1}^{m} P^{s'_1}(k|x^j_{\eta_i}) log(\frac{P^{s'_1}(k|x^j_{\eta_i})}{P^{s'_2}(k|x^j_{\eta_i})}) \tag{4}$$

Therefore, the cumulative training loss $\mathcal{L}$ for $(x^j_{\eta_i}, \eta_i)$, $i \epsilon \{1, 2, ...m\}$, $j \epsilon \{1, 2, ...V_{\eta_i}\}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{s_p} + \alpha \mathcal{L}_{KL} \tag{5}$$

B. Multi-loss Optimization

In 3D object recognition, it is observed that the learning curve highly depends upon the random selection of viewpoints per object instance o_i, which is the advantage of VISPE but can fail if wrongly selected. Therefore, a classroom-based approach is introduced in this letter, where randomly selected students of different learning capabilities for same task are trying to fetch identical information to compute a global loss. This multi-loss will enable DCNN to converge much faster compared to VISPE by every time selecting the best minimum error per epoch. Hence, the stochastic gradient descent (SGD) loss (Eq. 5) is transformed as:

$$\mathcal{L} = \mathcal{F}(\mathcal{L}^{\gamma}_{s_p} + \alpha \mathcal{L}^{\gamma}_{KL}) \tag{6}$$

where $\gamma \epsilon \{1, 2, ...r\}$ is the number of students in a classroom and $\mathcal{F}$ is the selection criteria to determine the best student for set s_p with random multi-views from $\mathcal{D}$. Note, since the same shared feature maps are used by all students, there is no additional computation cost involved. In our case, we set the selection function as $\mathcal{F} = min(.)$, *i.e.*, the best student with the minimum error is selected to update w. The best feature f_θ obtained from Eq. 6, reduces the overall training cost by converging faster. The approach also assist network to recognize the unseen object instances and boosts up the baseline performance (see Table I). In our classroom, we have cross-entropy, negative log likelihood and KL divergence loss functions, *i.e.*, $\gamma = 3$. But it is scalable and can further grow (see Figure 1).

C. Progressive Multi-view Learning

We observed that increasing viewpoints of an object highly influence the learning of deep feature embedding. In conventional training, after certain epochs the learning curve gets saturated. Thus, in this letter we introduce a progressive multi-view learning approach to minimize gradient vanishing problem and further optimize the object representation. For this, we split MCB-A dataset in four different subsets of viewpoints: 4-views $\mathcal{D}_1 \subset$ 6-views $\mathcal{D}_2 \subset$ 8-views $\mathcal{D}_3 \subset$ 12-views $\mathcal{D}$. We fixed the training subset $\mathcal{D}_i$ as the initial data and then progressively increase to $\mathcal{D}_{i+1}$ whenever the learning curve gets soaked. Here, i is the number of view subsets, which in our case is four. This progressive multi-view learning actually saves the training cost. That is, instead of training f_θ on $\mathcal{D}$ for e epochs, we split it's training set in such as way that the previous subset's trained weights act as the initializer for the successive subsets.

Implementation All experiments were conducted in Pytorch[1] and used a standard SGD with learning rate $\lambda = 0.001$, batch size $m = 32$, $\tau = 0.05$ and $\alpha = 5$. The network was trained over for max epoch $e = 150$, which is half of [14].

IV. Experiments and Results

In this section, we evaluate different SSL algorithms on two different types of 3D object recognition datasets and perform rigorous experiments to validate our hypothesis.

A. Datasets

Two different types of datasets are used in all further experiments: ModelNet40 [32] and MCB-A [16].

ModelNet40 is a 3D CAD dataset of 40 objects, in total 3,183 models. We followed VISPE, where 10 classes are subjected as unseen and rest other settings are kept the same, as in [14]. **ModelNet-c20** is a subset of ModelNet40 with additional 10 unseen classes: bowl, cone, cup, laptop, plant, radio, sink, stool, vase, and xbox. This dataset is created to test robustness of our hypothesis when novel objects are increased.

Mechanical Components Benchmark (MCB-A)[2] is the newest and largest annotated 3D CAD MCB dataset for classification and retrieval tasks. It has a total of 58,696 mechanical components under 68 classes which are aggregated

[1]PyTorch: https://pytorch.org/

[2]MCB-A dataset: https://mechanical-components.herokuapp.com/

TABLE I: Classification comparison with SOTA. *Results are carried from [14] and '–' means result unknown.

Methods/Dataset	ModelNet40		Backbone
	seen	*unseen*	
Pretrained*, 2014	62.7	52.7	VGG-16
Autoencoder*, 2016	31.8	37.2	–
Egomotion*, 2015	32.4	34.7	BCNN
Puzzle*, 2016	34.4	41.5	CFN
ShapeCode*, 2017	39.4	46.5	VGG-16
MVCNN*, 2015	39.6	48.1	VGG-VD
UEL*, 2019	47.9	46.5	ResNet-18
Instance classification*	57.7	58.9	VGG-16
Triplet*, 2015	70.1	62.4	–
PE, 2020 [14]	69.7	61.7	VGG-16
MVSPE, 2020 [14]	70.3	63.2	VGG-16
VISPE, 2020 [14]	71.2	64.4	VGG-16
improved VISPE (multi-loss)	79.0	73.3	VGG-16
improved VISPE (backbone)	78.7	74.0	ResNet-18
improved VISPE (multi-loss)	79.1	**74.0**	ResNet-18
improved VISPE (backbone)	74.0	71.2	iResNet-18 [33]
improved VISPE (multi-loss)	74.3	71.2	iResNet-18 [33]
improved VISPE (backbone)	80.3	72.3	EfficientNet-b0
improved VISPE (multi-loss)	**81.5**	73.0	EfficientNet-b0

Methods/Datasets	ModelNet-c20		Backbone
	seen	*unseen*	
VISPE [14]	80.2	66.4	VGG-16
improved VISPE (backbone)	82.9	68.6	ResNet-18
improved VISPE (multi-loss)	**83.0**	**70.2**	ResNet-18

Methods/Datasets	MCB-A 12-views		Backbone
	seen	*unseen*	
VISPE, 2020 [14]	60.6	60.2	VGG-16
improved VISPE (multi-loss)	62.6	63.2	VGG-16
improved VISPE (backbone)	64.5	**66.7**	ResNet-18
improved VISPE (multi-loss)	**65.4**	65.7	ResNet-18
improved VISPE (backbone)	62.2	63.7	EfficientNet-b0
improved VISPE (multi-loss)	64.7	65.4	EfficientNet-b0

Note: Word *improved* means either new backbone or backbone + multi-loss.

from TraceParts, 3DWarehouse and GrabCAD. This dataset is highly imbalanced and therefore, are more challenging for seen/unseen object recognition task, compared to ModelNet40 dataset. Since the dataset is used for classification, we manually split out 12 unseen categories. The training and testing set involves 38 and 9 instances per objects. The unseen classes are: Castor, Clamps, Fan, Hinge, Knob, Nozzle, Pulleys, Roll pins, Springs, Studs, Toothed and Wheel.

B. Recognition and Comparison

We followed the same training and testing distribution as of [14]. For inference, we use *k*-nearest neighbors (NN) classification where *k* is set to 960 for ModelNet40 while for MCB-A and ModelNet-c20 datasets, we set it to 456 and 500, respectively. The results are averaged over three runs.

SOTA Comparison. Table I shows a detailed *k*-NN classification comparison for ModelNet40, ModelNet-c20 and MCB-A datasets. It is observed that the proposed method outperforms the SOTA methods for all considered datasets. Especially, our result shows an improvement for unseen classes. Noticeably, the complex EfficientNet-b0 and ResNet-18 architectures along with classroom-based multi-loss function respectively boosts the performance of ModelNet40 compared to simple VGG-16 by 10.3% and 7.9% for seen categories while for unseen classes, it's 8.6% and 9.6% for EfficientNet-b0 and ResNet-18, respectively. Secondly, the proposed approach is trained for 150 epochs which can significantly up-lift the performance compared to VISPE (*baseline*), which was trained for 300 epochs. Thirdly, for 3D object recognition ResNet-18 performs better compared to iResNet-18 [33].

Similarly for ModelNet-c20 and MCB-A datasets, our proposed method improves the performance significantly for both seen and unseen categories. The result on ModelNet-c20 proves that multi-loss with complex architecture learns a better representation to distinguish unseen classes, even when they are increased. For MCB-A dataset, our proposed approach achieves 65.4% for 56 seen classes and 66.7% for 12 unseen categories (Table I). That is, the higher the γ is, the better the convergence speed will be, depending up on the type of loss functions selected. Since original MCB-A was used for supervised tasks, here we use it for our ablation studies.

Architecture Comparison. Next, Figure 2a-b shows a comparison of seen and unseen classes per epochs on ModelNet40 for different backbone networks. Figure 2c shows improvement comparison of different network's performance over their computational cost, in terms of FLOPS and parameters. From these graphs, it is clear that the proposed approach of multi-loss based object recognition is promising as it improves the overall training performance for almost all SOTA networks.

Progressive Multi-view. Lastly, we trained the proposed model using our progressive multi-view learning approach with $m = 32$ and $b = m/2$, as defined previously. The accuracy per epoch on MCB-A is shown in Figure 2d, where we can see that after few epoch of learning the view angles are increased for all objects in $\mathcal{D}$. This results in a better embedding without vanishing gradient issue. For seen categories the accuracy reaches to **65.6%** and for unseen its **68.1%** which is 0.2% and 2.4% high compared to multi-loss approach with ResNet-18. Not only this, the proposed learning approach also reduces the overall training cost by a factor of ≈42.8%[3].

***k*-shot Object Recognition.** The proposed concept is inherited for *k*-shot object recognition to test the generalizability of different embeddings. The classification accuracy for unseen classes with *k* images per object for ModelNet40 is tested and the proposed multi-loss approach achieves 45.8%@1 and 59.1%@3 compared to 43.1%@1 and 52.5%@3 from VISPE [14]. Thus, the proposed multi-loss significantly improves the performance and can be further explored in ZSL.

C. Discussion

Based on the above experiments, we analysed that for multi-view unseen object recognition a complex architecture is the best choice compared to a simple liner structure. Secondly, we examined that classroom-based multi-loss training strategy converges the network much faster and gives a significant boost in the performance by 6-8% in case of VGG-16 while 2-4% for complex architecture. Lastly, the proposed progressive learning is generic and can further reduce the training cost of various SOTA networks.

V. Conclusion and Future Works

In this letter, we proposed a multi-loss based progressive multi-view learning approach for 3D object recognition. The proposed method boosts the network learning capability and converges must faster (≈42%) with a significant performance improvement. The extensive experiment shows significant enhancements on all types of considered datasets and DCNN architectures. In future, we would like to further optimize the learning curve with minimal number of learning parameters.

[3]Training cost ($epoch \times \mathcal{D}$): 1. Conversional approach $\approx e \times \mathcal{D} = 150 \times \mathcal{D}$ 2. Our $\approx (e_i \times \mathcal{D}_1 + e_2 \times \mathcal{D}_2 + e_3 \times \mathcal{D}_3 + e_4 \times \mathcal{D})$ — *e.g.*, in our case $e_i = \frac{\mathcal{D}_i}{\mathcal{D}}$ then $\approx (9.9 + 25 + 38.86 + 12) \times \mathcal{D} = 85.76 \times \mathcal{D}$, (Figure 2d).

REFERENCES

[1] S. Liu, V. Nguyen, I. Rehg, and Z. Tu, "Recognizing objects from any view with object and viewer-centered representations," in *CVPR*, 2020, pp. 11 784–11 793.

[2] R. Luo, N. Zhang, B. Han, and L. Yang, "Context-aware zero-shot recognition." in *AAAI*, 2020, pp. 11 709–11 716.

[3] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "Foveabox: Beyound anchor-based object detection," *TIP*, vol. 29, pp. 7389–7398, 2020.

[4] F. Liang, L. Duan, W. Ma, Y. Qiao, J. Miao, and Q. Ye, "Context-aware network for rgb-d salient object detection," *PR*, vol. 111, p. 107630, 2021.

[5] W. Lee, J. Na, and G. Kim, "Multi-task self-supervised object detection via recycling of bounding box annotations," in *CVPR*, 2019, pp. 4984–4993.

[6] P. Tang, C. Ramaiah, Y. Wang, R. Xu, and C. Xiong, "Proposal learning for semi-supervised object detection," in *WACV*, 2020, pp. 2291–2301.

[7] V. A. Diwadkar and T. P. McNamara, "Viewpoint dependence in scene recognition," *Psychological science*, vol. 8, no. 4, pp. 302–307, 1997.

[8] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman, "3d interpreter networks for viewer-centered wireframe modeling," *IJCV*, vol. 126, no. 9, pp. 1009–1026, 2018.

[9] H. Cao, R. Zhan, Y. Ma, C. Ma, and J. Zhang, "Lfnet: Local rotation invariant coordinate frame for robust point cloud analysis," *IEEE Signal Processing Letters*, vol. 28, pp. 209–213, 2020.

[10] C. Ma, Y. Guo, J. Yang, and W. An, "Learning multi-view representation with lstm for 3-d shape recognition and retrieval," *TMM*, vol. 21, no. 5, pp. 1169–1182, 2018.

[11] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *AAAI*, vol. 33, 2019, pp. 3558–3565.

[12] H. You, Y. Feng, R. Ji, and Y. Gao, "Pvnet: A joint convolutional network of point cloud and multi-view for 3d shape recognition," in *ACM MM*, 2018, pp. 1310–1318.

[13] C.-H. Ho, P. Morgado, A. Persekian, and N. Vasconcelos, "Pies: Pose invariant embeddings," in *CVPR*, 2019, pp. 12 377–12 386.

[14] C.-H. Ho, B. Liu, T.-Y. Wu, and N. Vasconcelos, "Exploit clues from views: Self-supervised and regularized learning for multiview object recognition," in *CVPR*, 2020, pp. 9090–9100.

[15] Z. Han, Z. Fu, and J. Yang, "Learning the redundancy-free features for generalized zero-shot object recognition," in *CVPR*, 2020, pp. 12 865–12 874.

[16] S. Kim, H.-g. Chi, X. Hu, Q. Huang, and K. Ramani, "A large-scale annotated mechanical components benchmark for classification and retrieval tasks with deep neural networks," in *ECCV*, 2020.

[17] K. Wei, C. Deng, and X. Yang, "Lifelong zero-shot learning," in *IJCAI*, 2020, pp. 551–557.

[18] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2371–2381.

[19] S. Rahman, S. H. Khan, and F. Porikli, "Zero-shot object detection: Joint recognition and localization of novel concepts," *IJCV*, vol. 128, no. 12, pp. 2979–2999, 2020.

[20] Y. L. Cacheux, H. L. Borgne, and M. Crucianu, "Zero-shot learning with deep neural networks for object recognition," *arXiv preprint arXiv:2102.03137*, 2021.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *TPAMI*, vol. 39, no. 6, pp. 1137–1149, 2016.

[22] S. Prasad and A. Wai Kin Kong, "Using object information for spotting text," in *ECCV*, 2018, pp. 540–557.

[23] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[24] K. Kumar Singh, S. Divvala, A. Farhadi, and Y. Jae Lee, "Dock: Detecting objects by transferring common-sense knowledge," in *ECCV*, 2018, pp. 492–508.

[25] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *CVPR*, 2018, pp. 6857–6866.

[26] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *ICCV*, 2015, pp. 945–953.

[27] J. Jiang, D. Bao, Z. Chen, X. Zhao, and Y. Gao, "Mlvcnn: Multi-loop-view convolutional neural network for 3d shape retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8513–8520.

[28] Z. Han, H. Lu, Z. Liu, C.-M. Vong, Y.-S. Liu, M. Zwicker, J. Han, and C. P. Chen, "3d2seqviews: Aggregating sequential views for 3d global feature learning by cnn with hierarchical attention aggregation," *TIP*, vol. 28, no. 8, pp. 3986–3999, 2019.

[29] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3d human pose using multi-view geometry," in *CVPR*, 2019, pp. 1077–1086.

[30] J. Zhang, C.-G. Li, C. You, X. Qi, H. Zhang, J. Guo, and Z. Lin, "Self-supervised convolutional subspace clustering network," in *CVPR*, 2019, pp. 5473–5482.

[31] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[32] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *CVPR*, 2015, pp. 1912–1920.

[33] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Improved residual networks for image and video recognition," *arXiv preprint arXiv:2004.04989*, 2020.