

3D Point Cloud Classification and Segmentation using 3D Modified Fisher Vector Representation for Convolutional Neural Networks

Yizhak Ben-Shabat
 Mechanical Eng. Dept.
 Techion IIT
 Haifa, Israel
 sitzikbs@gmail.com

Michael Lindenbaum
 Computer Science Dept.
 Techion IIT
 Haifa, Israel
 mic@cs.technion.ac.il

Anath Fischer
 Mechanical Eng. Dept.
 Techion IIT
 Haifa, Israel
 meranath@technion.ac.il

Abstract

The point cloud is gaining prominence as a method for representing 3D shapes, but its irregular format poses a challenge for deep learning methods. The common solution of transforming the data into a 3D voxel grid introduces its own challenges, mainly large memory size. In this paper we propose a novel 3D point cloud representation called 3D Modified Fisher Vectors (3DmFV). Our representation is hybrid as it combines the discrete structure of a grid with continuous generalization of Fisher vectors, in a compact and computationally efficient way. Using the grid enables us to design a new CNN architecture for point cloud classification and part segmentation. In a series of experiments we demonstrate competitive performance or even better than state-of-the-art on challenging benchmark datasets.

1 Introduction

Point clouds are commonly used for representing the sensory data associated with 3D objects and scenes. Recent sensing technologies make them more reliable and accurate and they are already in use in many applications such as modeling tools and autonomous systems.

We propose a new approach for analyzing a point cloud obtained by direct 3D sensors, using deep neural networks (DNNs), and especially convolutional neural networks (ConvNets). ConvNets have shown remarkable performance in image analysis. Adapting them to point clouds is not, however, straightforward. ConvNets are built for input data arranged in fixed size arrays, on which linear space invariant filters (convolutions) may be applied. Point clouds, unfortunately, are unstructured, unordered, and contain a varying number of points. Therefore, they do not fit naturally into a spatial array (grid).

Several methods for extending ConvNets to 3D point cloud analysis have been proposed [18, 33, 23]. A common approach is to rasterize the 3D point data into a 3D voxel

grid (array), on which ConvNets can be easily applied. This approach, however, suffers from a tradeoff between its computational cost and its approximation accuracy. We discuss this approach, as well as other common point cloud representations, in Section 2.

We take a different approach and use a new point cloud hybrid representation, the 3D Modified Fisher Vector (3DmFV). The representation describes points by their deviation from a Gaussian Mixture Model (GMM). It has some similarity to the Fisher Vector representation (FV) [19, 26] but modifies and generalizes it in two important ways: the proposed GMM is specified using a set of uniform Gaussians with centers on a 3D grid, and the components characterizing the set of points, that, for Fisher vectors, are averages over this set, are generalized to other functions of this set.

The representation is denoted hybrid because it combines the discrete structure of a grid with the continuous nature of the components. It has several advantages. First, because it keeps the continuous properties of the point cloud, it retains some of the point set's fine detail and, under certain conditions, is lossless and invertible (in principle), and is therefore equivalent to featureless input. Second, the grid-like structure makes it possible to use ConvNets, which yields excellent classification accuracy even with low resolutions (e.g. $8 \times 8 \times 8$). Finally, each component of the proposed representation represents a clear and meaningful property.

The main contributions of this work are:

- We introduce a new hybrid representation for 3D point clouds (3DmFV) which is structured and order independent.
- We design a new deep ConvNet architecture (3DmFV-Net) based on this representation and use it for point cloud classification, obtaining state of the art results.
- We conduct a thorough empirical analysis on the stability of our method.

- We extend the 3DmFV-Net to part segmentation of point clouds, and achieve state of the art results as well.

We first review related work on 3D classification, part segmentation, and the FV representation in Section 2. Then, in Section 3 we introduce and discuss the 3DmFV representation and 3DmFV-Net architecture. The classification and part segmentation results are presented in Section 4. Finally, we summarize in Section 5.

2 Related Work

2.1 Deep learning on 3D data

Point cloud features - Handcrafted features for point clouds have shown adequate performance for many tasks. They can be divided into two main groups: local descriptors [25, 13, 31, 9] and global descriptors [1, 32, 16]. The comprehensive performance evaluation in [8] suggests guidelines for feature selection. However, optimal feature selection remains non-trivial and highly data specific.

Deep learning on 3D representations - 3D data is commonly represented using one of the following representations: (a) Multi-view, (b) Volumetric grid, (c) Mesh, (d) Point clouds. Each representation requires a different approach for modifying the data to the form required by deep learning methods.

Rendering 2D images of a 3D object from multiple views, as in [23], transforms the learning domain from 3D to the well-researched 2D domain. Some information is lost in the projection process, but using multiple projections partially compensates.

The volumetric, voxelized, representation discretizes 3D space similarly to an image discretizing a camera projection plane. This enables a straightforward extension of learning using 3D CNNs [18, 33, 23]. A volumetric representation is associated with a quantization tradeoff: choosing a coarse grid leads to quantization artifacts and to substantial loss of information, whereas choosing a fine grid significantly increases the number of voxels, which are mostly empty but still induce a high computational cost. Usually a grid size of $32 \times 32 \times 32$ is chosen. A recent improvement applies an ensemble of very deep networks that extend the principles of Inception [29] and Resnet [10] to voxelized data [4], thus achieving the highest accuracy to date at the price of high computational cost and training time of weeks.

For the mesh representation, Spectral ConvNets [5, 17, 2] and anisotropic ConvNets [3] can be applied. These approaches utilize mesh topological structure, which is not always available.

The point cloud representation is challenging because it is both unstructured and point-wise unordered. To overcome these challenges, the PointNet approach [22, 24] applies a symmetric function that is insensitive to the order, on

a high-dimensional representation of the individual points. The Kd-Network [14] imposes a kd-tree structure on the points and uses it to learn shared weights for nodes in the tree. Here we propose the 3DmFV representation, which is directly linked to the point cloud representation but can be used as input to a CNN.

Part segmentation of 3D point clouds - The objective here is to assign a label, for each point, corresponding to a semantically meaningful part of the shape, e.g., chair back or chair seat. The ShapeNet dataset, introduced in [34], was used in [22] to extend the PointNet approach from point cloud classification to point-wise classification (part segmentation) by concatenating high-dimensional learned global and local features. The Kd-Network architecture was extended to part segmentation by mimicing an encoder-decoder architecture with skip-connections [14]. Here, we extend the 3DmFV classification network to perform part segmentation using locally and globally learned features.

2.2 Fisher vectors

Before the age of deep learning, the bag of visual words (BoV) [6] was a popular choice for image classification tasks. It extracted a set of local descriptors and assigned each of them to the closest entry in a codebook (visual vocabulary), leading to a histogram of occurrences. Perronnin and Dance [19], and Perronnin and Thomas [21] proposed an alternative descriptor aggregation method, called the Fisher Vector (FV), based on the Fisher Kernel (FK) principle of [12]. The FV characterizes data samples of varying sizes by their deviation from a generative model, in this case a Gaussian Mixture Model (GMM). It does so by computing the gradients of the sample's log-likelihood w.r.t. the model parameters (i.e., weight, mean and covariance). FV can be viewed as a generalization of the BoV as the histogram is closely related to the derivative w.r.t. the weight. Furthermore, it was shown in [12] that, when the label is included as a latent variable of the generative model, the FK is asymptotically as good as the maximum a posteriori (MAP) decision rule for this model. The FV representation optimality and independence of sample size make it a natural choice for representing point cloud data.

In the context of image classification, the combination of FVs and DNNs was already considered [27, 20]. A network composed of Fisher layers was suggested in [27]. Each layer performs semi-local FV encoding (on dense hand-crafted features) followed by a dimensionality reduction, spatial stacking, and normalization. A network composed of unsupervised and supervised layers was proposed in [20]. The unsupervised layers calculate features, FVs, and reduce their dimension. They are followed by supervised fully connected layers, trained with back propagation.

The proposed 3DmFV shares some properties with the

representation types above. Like the volumetric approach, it is based on a grid, but not a grid of voxels. It thus maintains the grid structure, which makes it a convenient input to a ConvNet, but it suffers less from quantization. Like the PointNet approach, its features are symmetric functions, making them order and structure independent. The architecture we propose, like that of [20], combines unsupervised and supervised layers. However, it relies on the spatial properties of point clouds, which enables the use of ConvNets, substantially improving its performance.

3 The 3DmFV-Net

The proposed 3DmFV-Net classification architecture consists of two main modules. The first converts an input point cloud to the 3D modified Fisher vector (3DmFV) representation and the second processes it in a CNN architecture. These main modules are illustrated in Figure 1 and described below.

3.1 Describing point clouds with Fisher vectors

The proposed representation builds on the well-known Fisher vector representation. We start by formally describing the Fisher vectors (following the formulations and notation of [26]) in the context of 3D points, and then discuss some of their properties that lead to the proposed generalization and make them attractive as input to deep networks.

The Fisher vector is based on the likelihood of a set of vectors associated with a Gaussian Mixture model (GMM). Let $X = \{\mathbf{p}_t \in \mathbb{R}^3, t = 1, \dots, T\}$ be the set of 3D points of a given point cloud, where T denotes the number of points in the set. Let λ be the set of parameters of a K component GMM $\lambda = \{(w_k, \mu_k, \Sigma_k), k = 1, \dots, K\}$, where w_k, μ_k, Σ_k are the mixture weight, expected value, and covariance matrix of k -th Gaussian. The likelihood of a single 3D point (or vector) p associated with the k -th Gaussian density is

$$u_k(\mathbf{p}) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{p} - \mu_k)' \Sigma_k^{-1} (\mathbf{p} - \mu_k) \right\}. \quad (1)$$

The likelihood of a single point associated with the GMM density is therefore:

$$u_\lambda(\mathbf{p}) = \sum_{k=1}^K w_k u_k(\mathbf{p}). \quad (2)$$

Given a specific GMM, and under the common independence assumption, the Fisher vector, \mathcal{G}_λ^X , may be written as the sum of normalized gradient statistics, computed here

for each point \mathbf{p}_t :

$$\mathcal{G}_\lambda^X = \sum_{t=1}^T L_\lambda \nabla_\lambda \log u_\lambda(\mathbf{p}_t), \quad (3)$$

where L_λ is the square root of the inverse Fisher Information Matrix. The following change of variables, from $\{w_k\}$ to $\{\alpha_k\}$, ensures that $u_\lambda(x)$ is a valid distribution and simplifies the gradient calculation [15]:

$$w_k = \frac{\exp(\alpha_k)}{\sum_{j=1}^K \exp(\alpha_j)}. \quad (4)$$

The soft assignment of point \mathbf{p}_t to Gaussian k is given by:

$$\gamma_t(k) = \frac{w_k u_k(\mathbf{p}_t)}{\sum_{j=1}^K w_j u_j(\mathbf{p}_t)}. \quad (5)$$

The normalized gradients are:

$$\mathcal{G}_{\alpha_k}^X = \frac{1}{\sqrt{w_k}} \sum_{t=1}^T (\gamma_t(k) - w_k), \quad (6)$$

$$\mathcal{G}_{\mu_k}^X = \frac{1}{\sqrt{w_k}} \sum_{t=1}^T \gamma_t(k) \left(\frac{\mathbf{p}_t - \mu_k}{\sigma_k} \right), \quad (7)$$

$$\mathcal{G}_{\sigma_k}^X = \frac{1}{\sqrt{2w_k}} \sum_{t=1}^T \gamma_t(k) \left[\frac{(\mathbf{p}_t - \mu_k)^2}{\sigma_k^2} - 1 \right]. \quad (8)$$

These expressions include the normalization by L_λ under the assumption that $\gamma_t(k)$ (the point assignment distribution) is approximately sharply peaked [26]. For grid uniformity reasons, discussed in Section 3.4, we adopt the common practice and work with diagonal covariance matrices. The Fisher vector is formed by concatenating all of these components:

$$\mathcal{G}_{FV_\lambda}^X = \left(\mathcal{G}_{\alpha_1}^X, \dots, \mathcal{G}_{\alpha_K}^X, \mathcal{G}_{\mu_1}^{X'}, \dots, \mathcal{G}_{\mu_K}^{X'}, \mathcal{G}_{\sigma_1}^{X'}, \dots, \mathcal{G}_{\sigma_K}^{X'} \right). \quad (9)$$

To avoid the dependence on the number of points, the resulting FV is normalized by the sample size T [26]:

$$\mathcal{G}_{FV_\lambda}^X \leftarrow \frac{1}{T} \mathcal{G}_{FV_\lambda}^X. \quad (10)$$

See [26] for derivations, efficient implementation, and more details.

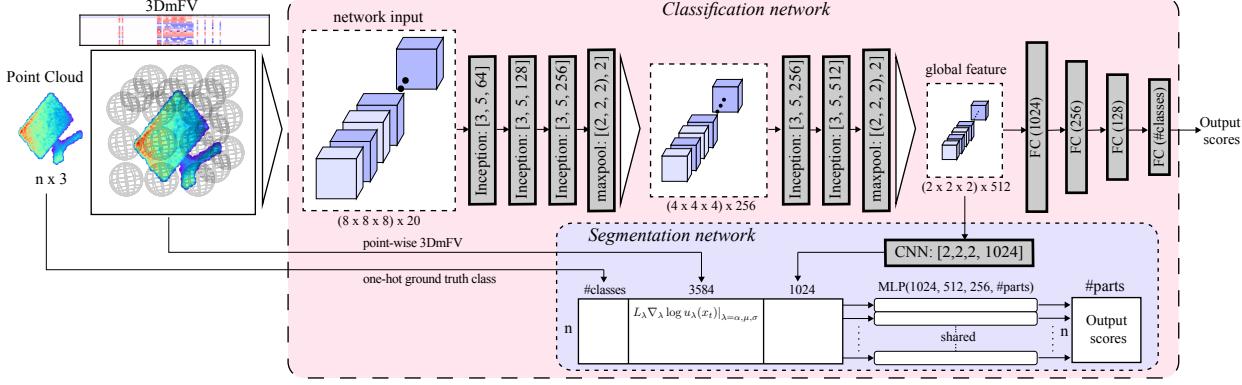


Figure 1: 3DmFV-Net architecture

3.2 Advantages of Fisher vectors as inputs to DNNs

A Fisher vector, representing a point set, may be used as an input to a DNN. It is a fixed size representation of a possibly variable number of points in the cloud. Its components are normalized sums of functions of the individual points. Therefore, FV representation of a point set is invariant to order, structure, and sample size.

Using a vector of non-learned features instead of the raw data goes against the common wisdom of deep neural network users. The reason is that features usually select some of the raw data properties and may lose some important characteristics. This is true especially when the feature extraction involves discretization, as is the case with voxel based representation. We argue, however, that the Fisher vector representation, being continuous on the point set, suffers less from this disadvantage. We shall give three arguments (not proofs) in favor of this claim.

a. Equation counting argument - Consider a K component GMM. A set of T points, characterized by $3T$ scalar coordinates, is represented using $7K$ components of the Fisher vector, every one of which is a continuous function of the $3T$ variables. Can we find another point set associated with the same Fisher components? We argue that for $T < 7K/3$, the set of equations specifying unknown points from known Fisher components is over-determined and that it is likely that the only solution, up to point permutation, is the original point set. While this equation counting argument is not a rigorous proof, such claims are common for sets of polynomial equations and for points in general position. If the solution is indeed unique, then the Fisher representation is lossless and using it is equivalent to using the raw point set itself.

b. Reconstructing the represented point structure in simplified, isolated cases

b.1 A single Gaussian representing a single point - Here, $T = 1$. By the sharply peaked $\gamma_t(k)$ assumption, there is

only one Gaussian for which $\gamma_t(k) = 1$. Inserting its FV components in Eq. 7 provides the point location:

$$\mathbf{p}_1 = \sigma_k \mathcal{G}_{\mu_k}^X + \mu_k. \quad (11)$$

b.2 A single Gaussian representing multiple points on one plane - We now show that, given a set of points sampled on a plane, it is possible to reconstruct a plane from the FV representing the points. The plane equation is $\hat{n}^T p = \rho$, where $\hat{n} = (a, b, c)^T$ is the unit normal to the plane and ρ is its distance from the origin. Using the assumption that $\gamma_t(k)$ is approximately sharply peaked [26], we consider the k -th Gaussian and the T points for which $\gamma_t(k) \approx 1$. For this Gaussian, eq. 7 is simplified to:

$$\mathcal{G}_{\mu}^X = \frac{1}{\sigma \sqrt{w}} \sum_{t=1}^T (\mathbf{p}_t - \mu). \quad (12)$$

Changing the coordinate system to $x' y' z'$, for which the origin is at the Gaussian center and an axis x' coincides with \hat{n} , leads to the following expression for the plane parameters (see Appendix 6.1 for a proof and illustrations) :

$$a = \frac{\mathcal{G}_{\mu_x}}{\|\mathcal{G}_{\mu}\|}, b = \frac{\mathcal{G}_{\mu_y}}{\|\mathcal{G}_{\mu}\|}, c = \frac{\mathcal{G}_{\mu_z}}{\|\mathcal{G}_{\mu}\|}, \rho = \sigma \|\mathcal{G}_{\mu}\| \sqrt{w} \quad (13)$$

Objects are often approximately polyhedral, with each facet having at least one close Gaussian for which only this facet is close. This implies that such models may be reconstructed from the FV even for very large T .

c. Point cloud reconstruction from FV representation using a deep decoder - A direct expression for reconstructing a point cloud from its FV representation is not available for $K > 1$. For illustration, we show now such a reconstruction obtained with a deep decoder, taking FV as an input and providing a point cloud. We consider a special FV,



Figure 2: Point cloud reconstruction from FV representation using a deep decoder. The original (left), and the reconstructed point cloud (right).

associated with a GMM with Gaussian centered on a grid; see Section 3.4 below. The decoder architecture is identical to the convolutional part of the classification network presented in Section 3.5 followed by two fully connected layers: $\text{FC}(T), \text{FC}(3T)$. The loss function between the original and the reconstructed point sets, S_1 and S_2 , should be invariant to point order. We use a loss function that is the sum of Chamfer distance and the Earth mover’s distance, which were used (separately) in [7]. Figure 2 shows a qualitative comparison between the original point cloud and the reconstructed point cloud. It shows that the decoder captured the overall shape while not positioning the points exactly in their original position.

3.3 Generalizing Fisher vectors to 3D modified Fisher vectors

We propose to generalize the Fisher vector along two directions:

Choice of the mixture model - Originally, the mixture model was defined as a maximum likelihood model. This makes the model optimally adapted to the training data and gives the Fisher representation of each Gaussian the nice property of being sensitive only to the deviation from the average training data. It is not the only valid option, however, and not a good choice if we prefer to maintain a grid structure. Therefore, in the proposed generalization we shall use other mixture models that rely on Gaussian grids.

Choice of the symmetric function - As apparent from eq. (3), the components of the Fisher vector are sums over all input points, regardless of their order and any structure they create. They are *symmetric* in the sense proposed in [22] and are therefore adequate for representing the orderless and structureless set of points. Note also that any other symmetric function, applied over the summands in eq. (3), would also induce a vector that can represent orderless sets. We will propose such functions and use them instead of or in addition to the Fisher vector sums.

3.4 The proposed 3DmFV generalization

Changing the mixture model - For the underlying density model, we use a mixture of Gaussians with Gaussian centers (μ_k) on a uniform 3D $m \times m \times m$ grid. Such Gaussians

induce a Fisher vector that preserves the point set structure: the presence of points in a specific 3D location would significantly influence only some, pre-known, Fisher components. The other GMM parameters, weight and covariance, are common to all Gaussians. The weights are selected as $w_k = \frac{1}{K}$ and the covariance matrix as $\Sigma_k = \sigma_k I$ with $\sigma_k = \frac{1}{m}$. Recall that all points are contained in the unit sphere. The uniformity is essential for shared weight (convolutional) filtering. The size of the mixture model is moderate and ranges from $m = 3$ to 9.

The proposed uniform model is not as effective as the maximum likelihood model for representing the distribution of point clouds. Recall, however, that the GMM does not represent a specific model or a specific class but rather the average model, which is much closer to uniform. The inaccuracy is more than compensated for by the power of the convolutional network, as we shall see in the comparison between the different models.

Changing/Adding other symmetric functions - For Fisher vectors, the sum is used as a symmetric function. While the sum is asymptotically optimal, it does not give full information about the input points for finite point sets. For small point sets the Fisher vectors may be invertible, as suggested above, implying that the FVs implicitly carry the full information about the set. For the practical case of large finite point sets, we propose to add information. To maintain the order independence, other summarizing features should be symmetric as well.

We experimented with several options for additional symmetric functions, and eventually chose the maximum and minimum functions. Note that the maximum was also used in [22] as a single summarizing feature for each of the learned features. Thus, the components of the proposed generalized vector, denoted 3DmFV, are given in eq. 14 and obtained as follows: each component is either a sum, max, or min function, evaluated on the set of one gradient component. In our experiments we found that partial, more compact, representations (especially those focusing on the minimum and maximum function) may lead to improved accuracy, and we describe them as well. However, we also found that the minimal weight derivative is always a constant and omitted this specific function. The minimal value associated with a specific Gaussian corresponds to the farthest point and its $\gamma_t(k)$ value, which is 0 in practice.

$$\mathcal{G}_{3DmFV_\lambda}^X = \begin{bmatrix} \sum_{t=1}^T L_\lambda \nabla_\lambda \log u_\lambda(\mathbf{p}_t) \Big|_{\lambda=\alpha,\mu,\sigma} \\ \max_t (L_\lambda \nabla_\lambda \log u_\lambda(\mathbf{p}_t)) \Big|_{\lambda=\alpha,\mu,\sigma} \\ \min_t (L_\lambda \nabla_\lambda \log u_\lambda(\mathbf{p}_t)) \Big|_{\lambda=\mu,\sigma} \end{bmatrix} \quad (14)$$

For $K = m^3$ Gaussians, there are therefore $20 \times K$ components in the 3DmFV representation ($20 = 3(3+3)+2$). The 3DmFV is best visualized as a $20 \times K$ matrix. Figure 3

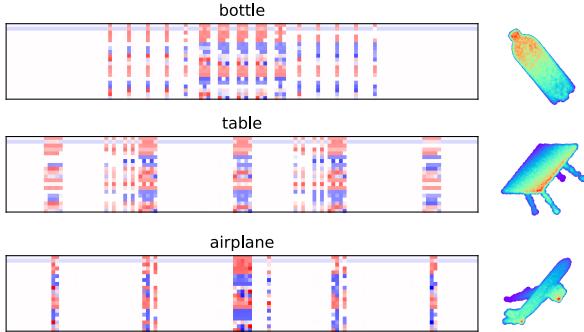


Figure 3: 3DmFV representation (left) and the corresponding point cloud (right)

depicts a point cloud (right) and its 3DmFV representation ($m = 5$) as a color coded image (left). Each column of the image represents a single Gaussian in a $5 \times 5 \times 5$ Gaussian grid. Zero values are white whereas positive and negative values correspond respectively to the red and blue gradients. Note that the representation lends itself to intuitive interpretation. For example, many columns are white, except for the first two top entries. These correspond to Gaussians that do not have model points near them; see eq. 6.

Normalization Following [21] (Sec. 2.3), we applied two consecutive normalizations on the 3DmFV representation: First, we applied an element-wise signed square root normalization, and then an L2 normalization over all 20 vectors corresponding to all Gaussians and a single feature λ_i . This last normalization equalizes the derivatives with respect to different parameters.

3.5 3DmFV-Net classification architecture

The proposed network receives a point cloud and converts it to a 3DmFV representation on a grid. The main parts of the network consist of an Inception module [30], visualized in Figure 4, maxpooling layers, and finally four fully connected layers. The network output consists of classification scores; see Figure 1. The network is trained using back propagation and standard softmax cross-entropy loss with batch normalization after every layer and dropout after each fully connected layer. The network has approximately 4.6M trained parameters, the majority of which are between the last maxpooling layer and the first fully connected layer.

3.6 3DmFV-Net part segmentation architecture

The 3DmFV-Net architecture is extended to perform part segmentation. We solve this problem using a per-point classification approach. The proposed architecture combines local geometrical information of each point with the global context of the entire point cloud. The local information is encoded in the per-point 3DmFV representation (i.e., the

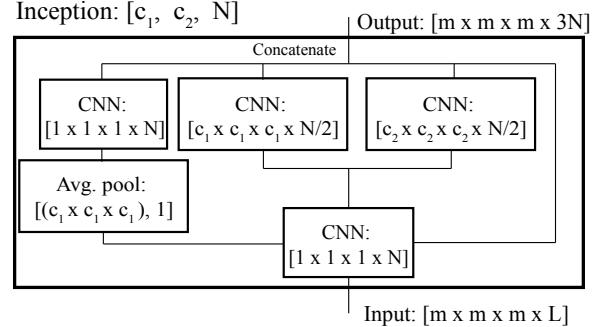


Figure 4: Inception module used in 3DmFV-Net

derivative value of each GMM Gaussian with respect to λ at that specific point), and the global information is added by concatenating the output of the convolutional part of the classification architecture (before the FC classifiers) to each point’s 3DmFV representation. Additionally, to fairly compare to other methods that assume a given label, we concatenate a one-hot representation of the ground truth label; however, this is not strictly necessary since omitting the label only slightly reduces performance (0.2%). These features are then aggregated into a combined representation using a multilayer perceptron with shared weights across points, similarly to [22]. The details of the 3DmFV-Net segmentation architecture are illustrated in Figure 1.

4 Experiments

We first evaluate the classification performance of our proposed 3DmFV-Net and compare it to previous approaches. We then evaluate several variants of the proposed representation. Next, we analyze our method’s robustness to noise. Finally, we evaluate the part segmentation results obtained by our method and compare them to the state of the art.

4.1 Implementation

Datasets - We evaluate all classification algorithms on the ModelNet40 dataset [33]. It consists of 12311 CAD models from 40 object categories, represented as triangle mesh. The data is split into a training set (9843 models) and a test set (2468). To generate point clouds, the mesh is sampled as described in [22]. We also experimented with the ModelNet10 dataset, which contains 4899 CAD models from 10 object classes split into 3991 for training and 908 for testing.

Training details: Unless otherwise specified, the 3DmFV-Net (Figure 1) was trained using an Adam optimizer with a learning rate of 0.001 with a decay of 0.7 every 20 epochs. The point cloud (of 2K points) is centered around the origin and scaled to fit a cube of edge length 2. Data is augmented using random anisotropic scaling (range:

Method	ModelNet10	Modelnet40
MVCNN [28]	-	90.1
3DShapeNets [33]	83.5	77.32
VoxNet [18]	92.0	83.0
VRN (One-View) [4]	-	88.98
VRN [4]	93.6	91.33
VRN ensemble [4]	97.14	95.54
FusionNet [11]	93.1	90.8
PointNet [22]	-	89.2 ^a
PointNet++ [24]	-	90.7 ^a
Kd-network [14]	94.0 ^b / 93.3 ^a	91.8 ^b / 90.6 ^a
3DmFV+VoxNet	94.3	88.5 ^c
Our 3DmFV-Net	95.2 ^{ac}	91.6 ^c /91.4 ^a

Table 1: Classification accuracy on ModelNet40 and ModelNet10 datasets. The point based methods use ^a1024, ^b32768, ^c2048 points.

[0.66, 1.5]) and random translation (range: [−0.2, 0.2]) in each axis, similarly to [14]. We used Tensorflow on a NVIDIA Titan Xp GPU. Training took ∼ 7 hours.

4.2 Classification performance

Table 1 compares the 3DmFV-Net with previous approaches on the ModelNet40 and ModelNet10 datasets. Clearly, the proposed method wins over most methods. It is comparable to the Kd-network, which requires a much larger input(∼ 32K points) and is not very robust to rotations and noise [14]. It is less accurate only than the more complex VRN ensemble method [4], which operates on voxelized input and averages 6 models, each trained for 6 days. However, it is slightly better than a single VRN model. Note also that direct comparison is somewhat unfair because, unlike the voxelized description, point based methods (like ours) do not have direct access to the mesh.

We also tested a combination of the 3DmFV representation ($m = 8$) with the simpler convolutional network that mimics the one used in VoxNet [18]. Although this combination (denoted 3DmFV+VoxNet) uses a lower resolution grid than VoxNet (32^3 voxels), it is more accurate. Thus, the performance boost of the 3DmFV-Net may be attributed to both the representation and the architecture.

4.3 Testing variations of the 3DmFV representation

We now test partial, more compact variants of the proposed representations. The first variant is the Fisher vectors. The 3DmFV generalization of FV uses different symmetric functions (and not only the sum, as in FV), and different GMMs. We considered the combinations of several symmetric functions, with GMMs obtained either from the

Rep.	ML + LinCls	ML + NonLinCls	Grid + NonLinCls
FV	58.4	82.8	84.5
3DmFV-ss	58.8	85.0	84.4
3DmFV-min	67.7	87.7	86.1
3DmFV-max	68.6	87.4	85.3
3DmFV	76.8	88.0	87.7

Table 2: Classification accuracy of the proposed 3DmFV and FV representation computed on EM-learned Gaussians and Gaussians positioned on a grid using a linear and non-linear classifier. Note that convolution layers are not used in this experiment.

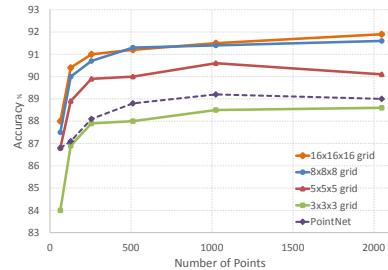


Figure 5: Effects of grid resolution and number of input points on the evaluation accuracy.

maximum likelihood (ML) optimization obtained by the expectation maximization (EM) algorithm or from Gaussians on a 3D $5 \times 5 \times 5$ grid. The tested symmetric functions include maximum (3DmFV-max), minimum (3DmFV-min), and sum of squares (3DmFV-ss). We tested these combinations with a nonlinear classifier (4 fully connected layers of sizes (1024, 256, 128, 40) with ReLU activation). For reference to the original FVs, we tested the ML GMM with a linear classifier as well. All models were represented by 1024 points. Table 2 reveals that the 3DmFV representation always wins over the FV representation, and that using grid GMM is comparable to the optimal ML GMM. Using maximum or minimum as a symmetric function yields comparable results with fewer parameters. Note that with the convolutional network, possible only with grid GMM, the accuracy is much higher (Table 1).

Resolution and standard deviation We found that accuracy increases with both grid size (m) and the number of points representing the model; see Figure 5. Note that performance saturates in both parameters. PointNet [22] seems to be more sensitive to the number of points. This is probably because their descriptors are learned as well. The architecture is slightly different for each grid size; see appendix 6.3.

We also found that the model is insensitive to the selection of standard deviation as long as it is not too small. In

method	mean	aero	bag	cap	car	chair	ear	guitar	knife	lamp	laptop	motor	mug	pistol	rocket	skate	board
Yi [34]	81.4	81.0	78.4	77.7	75.7	87.6	61.9	92.0	85.4	82.5	95.7	70.6	91.9	85.9	53.1	69.8	75.3
3DCNN [22]	79.4	75.1	72.8	73.3	70.0	87.2	63.5	88.4	79.6	74.4	93.9	58.7	91.8	76.4	51.2	65.3	77.1
PointNet [22]	83.7	83.4	78.7	82.5	74.9	89.6	73.0	91.5	85.9	80.8	95.3	65.2	93	81.2	57.9	72.8	80.6
Kd-Net [14]	77.2	79.9	71.2	80.9	68.8	88.0	72.4	88.9	86.4	79.8	94.9	55.8	86.5	79.3	50.4	71.1	80.2
Ours	84.3	82.0	84.3	86.0	76.9	89.9	73.9	90.8	85.7	82.6	95.2	66.0	94.0	82.6	51.5	73.5	81.8

Table 3: 3DmFV-Net part segmentation performance compared to state of the art. The metric is mean IoU.

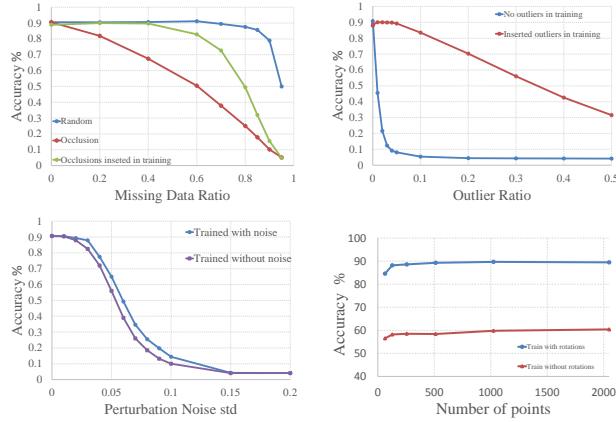


Figure 6: 3DmFV-Net robustness to data corruptions. Classification accuracy results for missing data (top-left), outlier insertion (top-right), perturbation noise (bottom-left), and rotations (bottom-right).

the case of very small σ , most points do not contribute to any Gaussian, rendering the FV representation empty; See appendix 6.4.

4.4 Robustness evaluation

To simulate real-world point cloud classification challenges, we tested 3DmFV-Net’s robustness under several types of noise:

Uniform point deletion - randomly deleting points is equivalent to classifying clouds that are smaller than those used for training.

Focused region point deletion - selecting a random point and deleting its closest points (the number of points is defined by a given ratio), simulating occlusions.

Outlier points - adding uniformly distributed points.

Perturbation noise - adding small translations, with a bounded Gaussian magnitude, independently to all points, simulating measurement inaccuracy.

Random rotation - randomly rotating the point cloud w.r.t. the global reference frame, simulating the unknown orientation of a scanned object.

The results (Figure 6) demonstrate that the proposed ap-

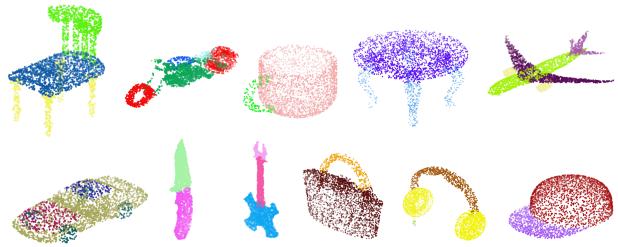


Figure 7: 3DmFV-Net part segmentation qualitative results

proach is inherently robust to perturbation noise and uniform point deletions. For the other types of data corruptions, training the classifiers with any of these types of noise made it robust to them.

4.5 Failure cases

Analyzing misclassifications, we found that most of them occur for similar class pairs: table-desk, dresser-night stand, and plant-flower pot, containing objects which are difficult to discriminate even for humans. See appendix 6.2 for some typical failures and for the confusion matrix (computed on the test set).

4.6 Part segmentation

We evaluate the performance of our part segmentation architecture on the ShapeNet part dataset from [34]. It contains 16881 point clouds with 50 annotated parts from 16 categories. Learning this dataset is challenging because it is highly imbalanced. The evaluation metric is mean intersection over union (IoU). It is calculated first for each part of each point cloud and is then averaged over each point cloud, and over all point clouds in that category, yielding the category IoU. The total mean IoU is computed as a weighted average of all category IoUs, where the weights are the respective number of point clouds in that category. We report in Table 3 the total mean IoU and the category IoUs and compare to other state of the art methods. We achieve best performance in 9/16 categories while no other method wins in more than 4/16. Additionally, we achieve best mean IoU.

Qualitative part-segmentation results are presented in Figure 7, and additional qualitative results are presented in the appendix 6.5.

5 Conclusion

In this work, we propose a new unsupervised 3D point cloud representation, the 3D modified Fisher vector. It preserves the raw point cloud data while using a grid for structure. This allows the use of the proposed CNN architecture (3DmFV-Net).

Representing data by non-learned features goes against the deep network principle that the best performance is obtained only by learning all components of the classifier using an end-to-end optimization. The proposed representation achieves state of the art results relative to all methods that use point cloud input, and therefore provides evidence that end-to-end learning is not always essential.

References

- [1] Aitor Aldoma, Federico Tombari, Radu Bogdan Rusu, and Markus Vincze. Our-cvfh-oriented, unique and repeatable clustered viewpoint feature histogram for object recognition and 6dof pose estimation. In *Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium*, pages 113–122. Springer, 2012.
- [2] Davide Boscaini, Jonathan Masci, Simone Melzi, Michael M Bronstein, Umberto Castellani, and Pierre Vandergheynst. Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks. In *Computer Graphics Forum*, volume 34, pages 13–23. Wiley Online Library, 2015.
- [3] Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 3189–3197, 2016.
- [4] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Generative and discriminative voxel modeling with convolutional neural networks. *arXiv preprint arXiv:1608.04236*, 2016.
- [5] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.
- [6] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cdric Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.
- [7] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image. *arXiv preprint arXiv:1612.00603*, 2016.
- [8] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, Jianwei Wan, and Ngai Ming Kwok. A comprehensive performance evaluation of 3d local feature descriptors. *International Journal of Computer Vision*, 116(1):66–89, 2016.
- [9] Yulan Guo, Ferdous Sohel, Mohammed Bennamoun, Min Lu, and Jianwei Wan. Rotational projection statistics for 3d local surface description and object recognition. *International Journal of Computer Vision*, 105(1):63–86, 2013.
- [10] 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.
- [11] Vishakh Hegde and Reza Zadeh. Fusionnet: 3d object classification using multiple data representations. *arXiv preprint arXiv:1607.05695*, 2016.
- [12] Tommi Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*, pages 487–493, 1999.
- [13] Andrew E. Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999.
- [14] Roman Klokov and Victor Lempitsky. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [15] Josip Krapac, Jakob Verbeek, and Frédéric Jurie. Modeling spatial layout with fisher vectors for image categorization. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1487–1494. IEEE, 2011.
- [16] Zoltan-Csaba Marton, Dejan Pangercic, Nico Blodow, and Michael Beetz. Combined 2d–3d categorization and classification for multimodal perception systems. *The International Journal of Robotics Research*, 30(11):1378–1402, 2011.

- [17] Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 37–45, 2015.
- [18] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *The IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928. IEEE, 2015.
- [19] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007.
- [20] Florent Perronnin and Diane Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3743–3752, 2015.
- [21] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. *Computer Vision–ECCV 2010*, pages 143–156, 2010.
- [22] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [23] Charles R Qi, Hao Su, Matthias Niessner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5648–5656, 2016.
- [24] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.
- [25] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *The IEEE International Conference on Robotics and Automation(ICRA)*, pages 3212–3217. IEEE, 2009.
- [26] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.
- [27] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep fisher networks for large-scale image classification. In *Advances in Neural Information Processing Systems*, pages 163–171, 2013.
- [28] 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 (CVPR)*, pages 945–953, 2015.
- [29] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017.
- [30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.
- [31] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *European Conference on Computer Vision*, pages 356–369. Springer, 2010.
- [32] Walter Wohlkinger and Markus Vincze. Ensemble of shape functions for 3d object classification. In *The IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 2987–2992. IEEE, 2011.
- [33] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaou 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.
- [34] Li Yi, Vladimir G Kim, Duygu Ceylan, I Shen, Mengyan Yan, Hao Su, ARCEwu Lu, Qixing Huang, Alla Sheffer, Leonidas Guibas, et al. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (TOG)*, 35(6):210, 2016.

6 Appendix

6.1 Point reconstruction from 3DmFV representation

In Section 3.2 we provide an expression for a plane’s parameters, given the FV representation of points sampled on it. This result holds in the asymptotic sense and under the assumption of uniform sampling. It holds for a single Gaussian. Here we prove this result, generalize it to grid GMM and verify it experimentally.

The plane equation is given by $\hat{n}^T \mathbf{p} = \rho$, where $\hat{n} = (a, b, c)^T$ is a unit normal to the plane and ρ is its distance from the origin. Using the assumption that $\gamma_t(k)$ is approximately sharply peaked [26], we consider the k -th Gaussian and the T_k points for which $\gamma_t(k) \approx 1$. From eq. 6 we get an expression for T_k :

$$T_k \approx \sum_{t=1}^T \gamma_t(k) = T\sqrt{w}(\mathcal{G}_{\alpha_k}^X + \sqrt{w}) \quad (15)$$

For this Gaussian, eq. 7 is simplified to:

$$\mathcal{G}_{\mu}^X = \frac{T_k}{\sigma\sqrt{w}} \sum_{t=1}^T (\mathbf{p}_t - \mu). \quad (16)$$

The \mathcal{G}_{μ}^X expression becomes clearer when we change the coordinate system to $x'y'z'$ (see Figure 8 top-right), for which the origin is at the Gaussian center, the axis x' coincides with \hat{n} and the other axes are chosen to constitute an orthogonal system. We select:

$$\hat{x}' = \hat{n}, \hat{y}' = \frac{\hat{n} \times \hat{y}}{|\hat{n} \times \hat{y}|}, \hat{z}' = \frac{\hat{y}' \times \hat{n}}{|\hat{y}' \times \hat{n}|}. \quad (17)$$

This yields the following transformation:

$$\mathbf{p} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mu + \begin{pmatrix} a & -\frac{c}{\sqrt{a^2+c^2}} & -\frac{ab}{\sqrt{a^2+c^2}} \\ b & 0 & \sqrt{a^2+c^2} \\ c & \frac{a}{\sqrt{a^2+c^2}} & -\frac{bc}{\sqrt{a^2+c^2}} \end{pmatrix} \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix}. \quad (18)$$

For all points on the plane, $x' = \rho_0$. By symmetry, for random uniform sampling on the plane, the expected value of both y' and z' is zero. Therefore, for a large number of samples, inserting eq. 18 in eq. 16 yields

$$\mathcal{G}_{\mu}^X \approx \frac{T_k}{\sigma\sqrt{w}} \begin{pmatrix} a \\ b \\ c \end{pmatrix} x'. \quad (19)$$

In addition, recall that to remove the dependency on the total number of points T , we divide by T (eq. 10). A simple

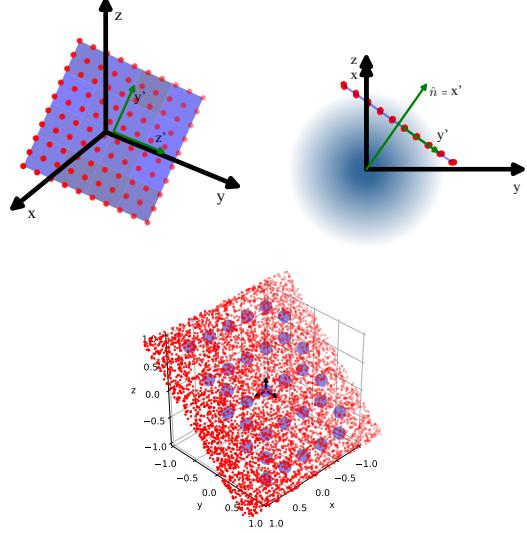


Figure 8: Points on a plane with their reconstructed plane. A view from the \hat{n} direction (top-left), a view from the y' direction, centered in one of the Gaussians (top-right), and an isometric view Gaussians that have a non-negligible $\sum_{t=1}^T \gamma_t(k)$ value with points on the plane (bottom).

inversion of eq. 19 yields:

$$a = \mathcal{G}_{\mu_x}^X \frac{\sigma\sqrt{w}T}{T_k\rho_0}, b = \mathcal{G}_{\mu_y}^X \frac{\sigma\sqrt{w}T}{T_k\rho_0}, c = \mathcal{G}_{\mu_z}^X \frac{\sigma\sqrt{w}T}{T_k\rho_0}. \quad (20)$$

Using the constraint that \hat{n} is a unit vector,

$$a^2 + b^2 + c^2 = 1, \quad (21)$$

we derive the expressions for the plane parameters as a function of (some of) the FV components:

$$a = \frac{\mathcal{G}_{\mu_x}}{\|\mathcal{G}_{\mu}\|}, b = \frac{\mathcal{G}_{\mu_y}}{\|\mathcal{G}_{\mu}\|}, c = \frac{\mathcal{G}_{\mu_z}}{\|\mathcal{G}_{\mu}\|}, \rho_0 = \frac{\sigma \|\mathcal{G}_{\mu}\|}{\mathcal{G}_{\alpha_k}^X + \sqrt{w}}. \quad (22)$$

Note that ρ_0 is the distance of the plane along \hat{n} in the local coordinate system. Therefore, to compute ρ in the global coordinate system we use:

$$\rho = \rho_0 + \mu^T \hat{n} \quad (23)$$

To validate these expressions in a more realistic case, where $\gamma_t(k)$ is not binary, we sampled points uniformly on a plane specified by $\frac{1}{\sqrt{2}}(0, 1, 1)\mathbf{p} = 0.05$. We then computed the point cloud’s FV representation using a $5 \times 5 \times 5$ Gaussian grid ($\sigma = \frac{1}{10}$, $w = \frac{1}{125}$) using eq. 6 to 8. Next, using eq. 22 and 23 we estimated the plane parameters from the FV representation of each Gaussian that has a non-negligible $\sum_{t=1}^T \gamma_t(k)$ value, see 8 (bottom). Figure 8 (top-left) shows a part of the original sampled points in red and the reconstructed surface in purple. We found that all the

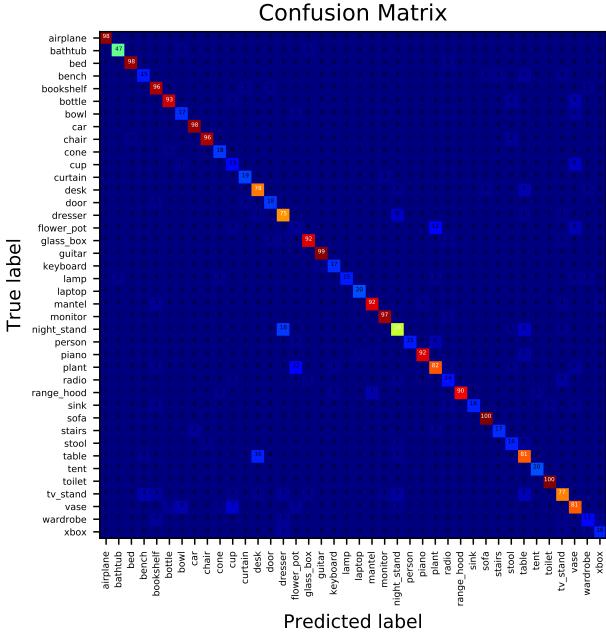


Figure 9: 3DmFV-Net classification confusion matrix

reconstructed planes are similar to the original one, which demonstrate that the FV components hold the information about the plane parameters.

6.2 Failure cases

In order to better understand the 3DmFV representation and the 3DmFV-Net, we explored failure cases. The confusion matrix in Figure 9 shows that the majority of misclassified point clouds belong to a few class pairs, specifically table-desk, plant-flowerpot and dresser-night stand. Further inspection using visualization of the failure cases presented in Figure 10 provides some insight. First, some classes are inherently hard to distinguish even for humans (table-desk) and second, the training data imposes a challenge since there is some overlap between categories (plant-flower pot).

6.3 3DmFV architecture for different grid sizes

We tested several grid resolutions for the 3DmFV representation (see Section 4, Figure 5). Each grid size imposes a slightly different network architecture. The architectures are detailed below. See Figure 4 for a description of the parametric inception module.

1. Grid $3 \times 3 \times 3$: 3DmFV-inception(2,3,64) - inception(2,3,128) - inception(2,3,256) - inception(2,3,256) - inception(2,3,512) - FC(1024) - FC(256) - FC(128) - FC(#classes)

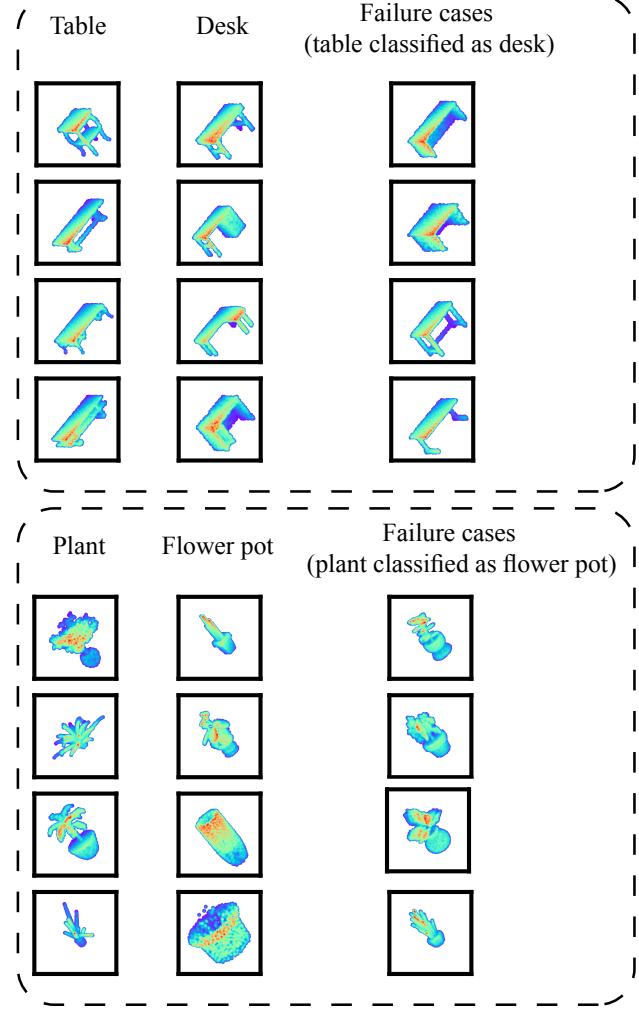


Figure 10: 3DmFV-Net classification failure cases. Table point clouds classified as desks (top), and plants classified as flower pots (bottom).

2. Grid $5 \times 5 \times 5$: 3DmFV - inception(3,5,64) - inception(3,5,128) - inception(3,5,256) - maxpool([3,3,3],2) - inception(2,3,256) - inception(2,3,512) - FC(1024) - FC(256) - FC(128) - FC(#classes)
3. Grid $8 \times 8 \times 8$ (3DmFV-Net): 3DmFV - inception(3,5,64) - inception(3,5,128) - inception(3,5,256) - maxpool([2,2,2],2) - inception(3,5,256) - inception(3,5,512) - maxpool([2,2,2],2) - FC(1024) - FC(256) - FC(128) - FC(#classes)
4. Grid $16 \times 16 \times 16$: 3DmFV - inception(4,8,64) - inception(4,8,128) - inception(4,8,256) - maxpool([2,2,2],2) - inception(3,5,256) - inception(3,5,512) - maxpool([2,2,2],2) - inception(2,3,512) - inception(2,3,512) - maxpool([2,2,2],2) - FC(1024) - FC(256) - FC(128) - FC(#classes)

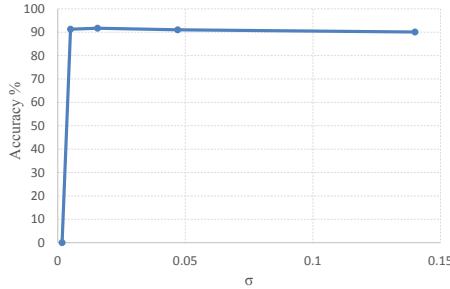


Figure 11: Effects of Gaussian standard deviation (σ) selection on the evaluation accuracy

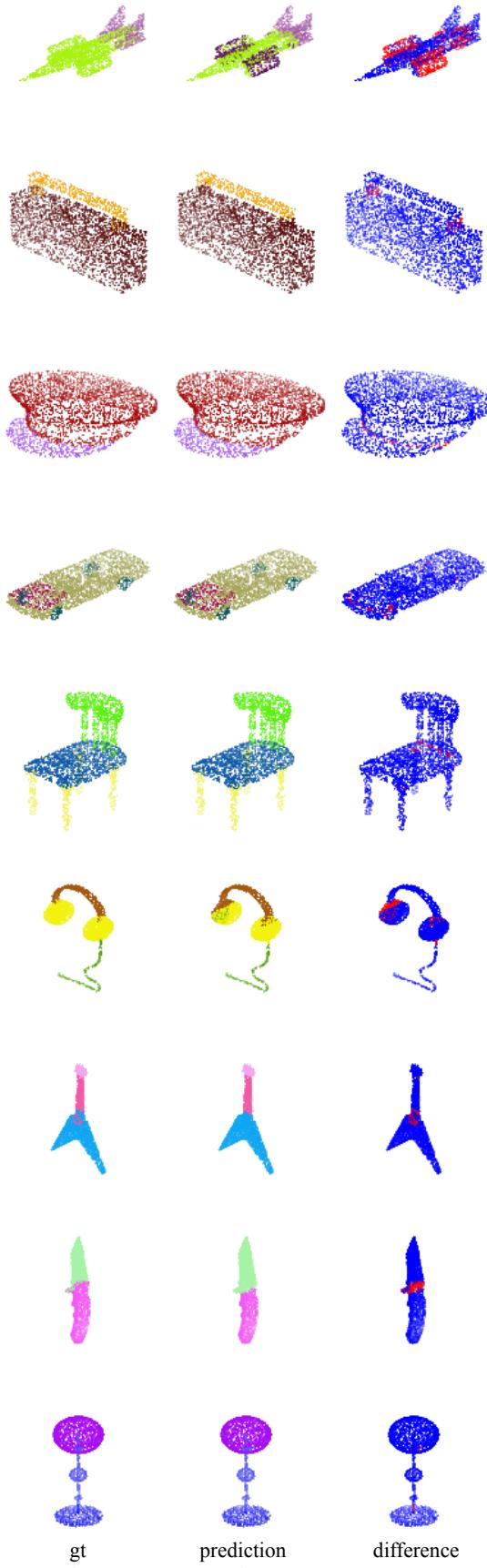
FC(256) - FC(128) - FC(#classes)

6.4 Additional robustness tests

One of the parameters of the 3DmFV representation is the Gaussian standard deviation σ . The value of σ qualitatively specifies spherical sub-volume whose points contribute to the 3DmFV component associated with a specific Gaussian. Very small σ values create Gaussians with very few or no contributing points and very large σ s create Gaussians that may be affected by many or all points in the cloud. We tested the robustness of 3DmFV-Net to the σ parameter selection. Figure 11 shows that the network is robust to the σ selection except for very small σ s, for which the network performs poorly (since the representation essentially fails to capture the point cloud).

6.5 Part segmentation

In Section 3.6 we presented qualitative results for part segmentation. Additional part segmentation results are presented here with a comparative visualization. In Figure 12, the left column shows the ground truth point labels, the middle column shows the labels predicted by the 3DmFV-Net segmentation network, and the right column shows a color coded comparison between the two, where correctly labeled points are shown in blue and mislabeled in red. It can be seen that mislabeled points sometimes appear in transitional locations between labels (e.g., red points are visible where the chair back meets the chair seat).



13 Figure 12: Part segmentation qualitative comparative results. Ground truth labels (left), predicted labels (middle), and color coded difference (right).