High-dimensional Convolutional Networks for Geometric Pattern Recognition

Christopher Choy  
Stanford University

Junha Lee  
POSTECH

René Ranftl  
Intel Labs

Jaesik Park  
POSTECH

Vladlen Koltun  
Intel Labs

Abstract

Many problems in science and engineering can be formulated in terms of geometric patterns in high-dimensional spaces. We present high-dimensional convolutional networks (ConvNets) for pattern recognition problems that arise in the context of geometric registration. We first study the effectiveness of convolutional networks in detecting linear subspaces in high-dimensional spaces with up to 32 dimensions: much higher dimensionality than prior applications of ConvNets. We then apply high-dimensional ConvNets to 3D registration under rigid motions and image correspondence estimation. Experiments indicate that our high-dimensional ConvNets outperform prior approaches that relied on deep networks based on global pooling operators.

1. Introduction

Finding structure in noisy data is a general problem that arises in many different disciplines. For example, robust linear regression requires finding a pattern (line, plane) in noisy data. 3D registration of point clouds requires the identification of veridical correspondences in the presence of spurious ones [6]. Structure from motion (SfM) pipelines use verification based on prescribed geometric models to filter spurious image matches [38]. A variety of such applications can benefit from improved methods for the detection of geometric structures in noisy data.

Such detection is challenging. Data points that belong to the sought-after structure often constitute only a small fraction, while the majority are outliers. Various algorithms have been proposed over the years to cope with noisy data [3, 9, 14, 19, 20, 32, 36, 39], but they are usually specific to a subset of problems.

Recent works have advocated for using deep networks [33, 46, 49] to learn robust models to classify geometric structures in the presence of outliers. Deep networks offer significant flexibility and the promise to replace hand-crafted algorithms and heuristics by models learned directly from data. However, due to the unstructured nature of the data in geometric problems, existing works have treated such data as unordered sets, and relied on network architectures based predominantly on global pooling operators and multi-layer perceptrons (MLPs) [31, 47]. Such network architectures lack the capacity to model local geometric structures and do not leverage the nature of the data, which is often embedded in a (high-dimensional) metric space and has a meaningful geometric structure.

In this work, we introduce a novel type of deep convolutional network that can operate in high dimensions. Our network takes a sparse tensor as input and employs high-dimensional convolutions as the fundamental operator. The distinguishing characteristic of our approach is that it is able to effectively leverage local neighborhood relations together with global context even for high-dimensional data. Our network is fully-convolutional, translation-invariant, and incorporates best practices from the development of ConvNets for two-dimensional image analysis [16, 21, 34].

To demonstrate the effectiveness and generality of our approach, we tackle various geometric pattern recognition problems. We begin with the diagnostic setting of linear subspace detection and show that our construction is effective in high-dimensional spaces and at low signal-to-noise ratios. We then apply the presented construction to geometric pattern recognition problems that arise in computer vision, including registration of three-dimensional point sets under rigid motion (a 6D problem) and correspondence estimation between images under epipolar constraints (a 4D problem). In both settings, the problem is made difficult by the presence of outliers.

Our experiments indicate that the presented construction can reliably detect geometric patterns in high-dimensional data that is heavily contaminated by noise. It can operate in regimes where existing algorithms break down. Our approach significantly improves 3D registration performance when combined with standard approaches for 3D point cloud alignment [6, 37, 51, 52]. The presented high-dimensional convolutional network also outperforms state-of-the-art methods for correspondence estimation between images [46, 49]. All networks and training scripts are available at https://github.com/chrischoy/HighDimConvNets.
2. Related Work

Robust model fitting. Fitting a geometric model to a set of observations that are contaminated by outliers is a fundamental problem that frequently arises in computer vision and related fields. The most widely used approach for robust geometric model fitting is RANdom SAmple Consensus (RANSAC) [13]. Due to its fundamental importance, many variants and improvements of RANSAC have been proposed over the years [3, 9, 23, 32, 36, 39, 42, 41].

Alternatively, algorithms for robust geometric model fitting are frequently derived using techniques from robust statistics [14, 19, 20, 30], where outlier rejection is performed by equipping an estimator with a cost function that is insensitive to gross outliers. While the resulting algorithms are computationally efficient, they require careful initialization and optimization procedures to avoid poor local optima [51].

Another line of work proposes to find globally optimal solutions to the consensus maximization problem [5, 24, 45]. However, these approaches are currently computationally too demanding for many practical applications.

3D registration. Finding reliable correspondences between a pair of surfaces is an essential step for 3D reconstruction [6, 10, 12, 29]. The problem has been conventionally framed as an energy minimization problem which can be solved using various techniques such as branch and bound [44], Riemannian optimization [35], mixed-integer programming [22], robust error minimization [51], semi-definite programming [18, 26], or random sampling [6].

Recent work has begun to leverage deep networks for geometric registration [2, 28, 30]. These works are predominantly based on PointNet and related architectures, which detect patterns via global pooling operations [31, 47]. In contrast, we develop a high-dimensional convolutional network that operates at multiple scales and can leverage not just global but also local geometric structure.


All of these works employ variants of PointNets to classify inliers in an unordered set of putative correspondences. These architectures, based on pointwise MLPs, lack the capacity to model local geometric structure. In contrast, we develop convolutional networks that directly leverage neighborhood relations in the high-dimensional space of correspondences.

3. High-Dimensional Convolutional Networks

In this section, we introduce the two main building blocks of our high-dimensional convolutional network construction: generalized sparse tensors and generalized convolutions.

3.1. Sparse Tensor and Convolution

A tensor is a multi-dimensional array that represents high-order data. A D-th order tensor \( T \) requires \( D \) indices to uniquely access its elements. We denote such indices or coordinates as \( x = [x^1, ..., x^D] \) and the element at the coordinate as \( T[x] \) similar to how we access components in a matrix. Likewise, a sparse tensor is a high-dimensional extension of a sparse matrix where the majority of the elements are 0. Concretely,

\[
T[x] = \begin{cases} 
    f_i & \text{if } x_i \in C \\
    0 & \text{otherwise,}
\end{cases}
\]

where \( C = \{x_i \mid x_i \in \mathbb{N}^D, T[x_i] \neq 0\}^{N}_{i=1} \) is the set of coordinates with non-zero values, \( N \) is the number of non-zero elements, and \( f_i \) is the non-zero value at the \( i \)-th coordinate.

A sparse tensor feature map is a \( (D+1) \)-th order tensor with \( f_i \in \mathbb{R}^{N^{D+1}} \) as we use the last dimension to denote the feature dimension. A sparse tensor has the constraint that \( x_i \in \mathbb{N}^D \). We extend the sparse tensor coordinates to integer indices \( x_i \in \mathbb{Z}^D \) and define \( T \in \mathbb{R}^{N_0 \times N_0^{D+1}} \) where \( N_0 \) denotes the cardinality of the integer space \( |\mathbb{Z}| \) to define a generalized sparse tensor.

A convolution on this generalized sparse tensor can then be defined as a simple extension of the generalized sparse convolution [7]:

\[
f^\text{out}_x = \sum_{i \in \mathbb{N}^D(x) \cap C^\text{in}} W_i f^\text{in}_{x+i} \quad \text{for } x \in C^\text{out},
\]

where \( C^\text{in} \) and \( C^\text{out} \) are the set of input and output locations which is predefined by the user, \( W_i \) is a weight matrix, and \( \mathbb{N}^D(x) \) defines a set of neighbors of \( x \) which is defined by the shape of the convolution kernel. For example, if the convolution kernel is a hypercube of size \( K \), \( \mathbb{N}^D(x) \cap C^\text{in} \) is a set of all the non-zero elements of the input sparse tensor centered at \( x \) within the \( \ell_{\infty} \)-ball of extent \( K \).

3.2. Convolutional Networks

We design a high-dimensional fully-convolutional neural network for sparse tensors (sparse tensor networks) based on generalized convolution [7, 8]. We use U-shaped networks [34] to capture large receptive fields while maintaining the original resolution at the final layer. The network has residual connections [16] within layers with the same resolution and across the network to speed up convergence.
and to recover the lost spatial resolution in the last layer. The network architecture is illustrated in Fig. 1.

For computational efficiency in high dimensions, we use the cross-shaped kernel \([7]\) for all convolutions. We denote the kernel size by \(K\). The cross-shaped kernel has non-zero weights only for the \(K - 1\) nearest neighbors along each axis, which results in one weight parameter for the center location and \(K - 1\) weight parameters for each axis. Note that a cross-shaped kernel is similar to separable convolution, where a full convolution is approximated by \(D\) one-dimensional convolutions of size \(K\). Both types of kernels are rank-1 approximations of the full hyper-cubic kernel \(K^D\), but separable convolution requires \(KD\) matrix multiplications, whereas the cross-shaped kernel requires only \((K - 1)D + 1\) matrix multiplications.

3.3. Implementation

We extend the implementation of Choy et al. \([7]\), which supports arbitrary kernel shapes, to high-dimensional convolutional networks. To implement the sparse tensor networks, we need an efficient data structure that can generate a new sparse tensor as well as find neighbors within the sparse tensor. Choy et al. \([7]\) use a hash table that is efficient for both insertion and search. We replaced the hash table with a faster and more efficient variant \([1]\). In addition, as the neighbor search can be run in parallel, we can create an iterator function that can run in parallel with OpenMP \([27]\) by dividing the table into smaller parallelization blocks.

Lastly, U-shaped networks generate hierarchical feature maps that expand the receptive field. Choy et al. \([7]\) use stride-\(K\) convolutions with kernel size \(\geq K\) to generate lower-resolution hierarchical feature maps. Still, such implementation requires iterating over at least \(K^D\) elements within a hypercubic kernel as the coordinates are stored in a hash table, which results in \(O(NK^D)\) complexity where \(N\) is the cardinality of input. Consequently, it becomes infeasible to store the weights on the GPU for high-dimensional spaces. Instead of strided convolutions, we propose an efficient implementation of stride-\(K\) sum pooling layers with kernel size \(K\). Instead of iterating over all possible neighbors, we iterate over all input coordinates and round them down to multiples of \(K\), which requires only \(O(N)\) complexity.

4. Geometric Pattern Recognition

Our approach uses convolutional networks to recognize geometric patterns in high-dimensional spaces. Specifically, we classify each point \(x_i\) in a high-dimensional dataset \(X = \{x_i\}_{i=1}^N\) as an inlier or an outlier. We start by validating our approach on synthetic datasets of varying dimensionality and then show results on 3D registration and essential matrix estimation.

![Figure 1: A generic U-shaped high-dimensional convolutional network architecture. The numbers next to each block indicate kernel size, stride, and the number of channels. The strided convolutions that reduce the resolution of activations are shifted upward to indicate different levels of resolution.](image-url)
strong correlation between F1 score and the mean squared error. The final metric we use is average precision (AP), which measures the area under the precision-recall curve. We report the results in Tab. 1 and provide qualitative examples in Fig. 2. Tab. 1 lists the inlier ratio to indicate the difficulty of each task.

In a second experiment, we create another synthetic dataset where the inlier pattern is sampled from a plane spanned by two vectors: \( \{c_1 v_1 + c_2 v_2 + c \mid c_1, c_2 \in \mathbb{R}\} \). The two basis vectors are sampled uniformly from the D-dimensional unit hypercube. We use the same training procedure for our network and the baselines, and report the results in Tab. 2.

We found that the convolutional network is more robust to noise in high-dimensional spaces than the PointNet variants. In addition, the convolutional network training converges quickly, as shown in Fig. 3, which is a further indication that the architecture can effectively leverage the structure of the data.

### 4.2. 3D Registration

A typical 3D registration pipeline consists of 1) feature extraction, 2) feature matching, 3) match filtering, and 4) global registration. In this section, we show that in the match filtering stage the correct (inlier) correspondences form a 6-dimensional geometric structure. We then extend our geometric pattern recognition networks to identify inlier correspondences in this 6-dimensional space.

Let \( \mathcal{X} \) be a set of points sampled from a 3D surface, \( \mathcal{X} = \{x_i \mid x_i \in \mathbb{R}^3\}_{i=1}^N \), and let \( \mathcal{X}' \) be a subset of \( \mathcal{X} \) that went through a rigid transformation \( T \): \( \mathcal{X}' = \{T(x) \mid x \in S, S \subseteq \mathcal{X}\} \). For example, \( \mathcal{X}' \) could be a 3D scan from a different perspective that has an overlap with \( \mathcal{X} \). We denote a correspondence between points \( x_i \in \mathcal{X} \) and \( x'_j \in \mathcal{X}' \) as \( x_i \leftrightarrow x'_j \). When we form an ordered pair \( (x_i, x'_j) \in \mathbb{R}^6 \), the ground truth correspondences satisfy \( T(x) = x' \) along the

### Table 1: Line detection in high-dimensional spaces in the presence of extreme noise. All networks are trained with the cross-entropy loss for 40 epochs. Inlier ratios are listed on the left. In the 32-dimensional setting, only 7 out of 10,000 data points are inliers. The table reports Mean Squared Error (MSE), F1 score, and Average Precision (AP) of our approach (a high-dimensional ConvNet) versus baselines (PointNet variants). MSE: lower is better. F1 and AP: higher is better.

<table>
<thead>
<tr>
<th>Dim.</th>
<th>Inlier Ratio</th>
<th>Qi et al. [31]</th>
<th>Zaheer et al. [47] + BN + IN</th>
<th>Yi et al. [46]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MSE</td>
<td>F1 (AP (AUC))</td>
<td>F1 (AP (AUC))</td>
<td>F1 (AP (AUC))</td>
</tr>
<tr>
<td>4</td>
<td>15.59%</td>
<td>1.337 0.025 0.164</td>
<td>6.33E-4 0.867 0.936</td>
<td>9.11E-5 0.981 0.996</td>
<td>2.33E-5 0.998 0.999</td>
</tr>
<tr>
<td>8</td>
<td>5.54%</td>
<td>2.369 0.022 0.065</td>
<td>0.001 0.891 0.955</td>
<td>2.45E-4 0.946 0.989</td>
<td>1.64E-5 0.999 0.999</td>
</tr>
<tr>
<td>16</td>
<td>2.75%</td>
<td>3.854 0.012 0.034</td>
<td>4.86E-4 0.970 0.992</td>
<td>0.002 0.962 0.986</td>
<td>3.39E-5 0.999 0.999</td>
</tr>
<tr>
<td>24</td>
<td>0.40%</td>
<td>5.372 0.021 0.011</td>
<td>0.676 0.634 0.691</td>
<td>0.775 0.610 0.674</td>
<td>5.34E-5 0.994 0.996</td>
</tr>
<tr>
<td>32</td>
<td>0.07%</td>
<td>6.715 0.012 0.003</td>
<td>6.71E-5 0.100 0.100</td>
<td>0.010 0.050 0.050</td>
<td>0.010 0.669 0.689</td>
</tr>
</tbody>
</table>

### Table 2: Plane detection in high-dimensional spaces in the presence of extreme noise. Inlier ratios are listed on the left. In the 32-dimensional setting, fewer than 5 out of 100,000 data points are inliers. The table reports the F1 score and Average Precision (AP) of our approach (a high-dimensional ConvNet) versus baselines (PointNet variants). Higher is better.

<table>
<thead>
<tr>
<th>Dim.</th>
<th>Inlier Ratio</th>
<th>Qi et al. [31]</th>
<th>Zaheer et al. [47] + BN + IN</th>
<th>Yi et al. [46]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F1 (AP (AUC))</td>
<td>F1 (AP (AUC))</td>
<td>F1 (AP (AUC))</td>
<td>F1 (AP (AUC))</td>
</tr>
<tr>
<td>4</td>
<td>29.96%</td>
<td>0.0 0.315 0.980</td>
<td>0.996</td>
<td>0.993 0.999</td>
<td>0.991 0.998</td>
</tr>
<tr>
<td>8</td>
<td>8.07%</td>
<td>0.0 0.088 0.985</td>
<td>0.999</td>
<td>0.990 0.999</td>
<td>0.998 0.999</td>
</tr>
<tr>
<td>16</td>
<td>0.34%</td>
<td>0.0 0.004 0.155</td>
<td>0.299</td>
<td>0.182 0.359</td>
<td>0.951 0.961</td>
</tr>
<tr>
<td>24</td>
<td>0.01%</td>
<td>0.0 1.61E-4 0.032</td>
<td>0.133</td>
<td>0.0 0.081</td>
<td>0.304 0.346</td>
</tr>
<tr>
<td>32</td>
<td>4.64E-3%</td>
<td>0.0 5.56E-5 0.0</td>
<td>0.221</td>
<td>0.0 0.023</td>
<td>0.138 0.240</td>
</tr>
</tbody>
</table>
common 3D geometry \( S \) whereas an incorrect correspondence implies \( T(x) \neq x' \). For example, in Fig. 4, we visualize the ordered pairs from 1-dimensional sets. Note that the inliers follow the geometry of the inputs and form a line segment. Similarly, the geometry \( (x, x') \in R^6 \) or \( (x, T(x)) \) for \( x \in S \) forms a surface in 6-dimensional space.

We can thus use our high-dimensional convolutional network construction to segment the 6-dimensional set of correspondences into inliers and outliers by estimating the inlier likelihood for each correspondence.

**Network.** We use a 6-dimensional instantiation of the U-shaped convolutional network presented in Sec. 3. As the dimensionality is manageable, we use hypercubic kernels. The network takes an order-6 sparse tensor whose coordinates are correspondences \((x_i, x'_i) \in R^6\). We discretize the coordinates with the voxel size used to extract features. Our baseline is Yi et al. [46], which takes dimensionless mean-centered correspondences without discretization. We train the networks to predict the inlier probability of each correspondence with the balanced cross-entropy loss.

**Dataset.** We use the 3DMatch dataset for this experiment [48]. The 3DMatch dataset is a composition of various 3D scan datasets [15, 43, 48] and thus covers a wide range of scenes and different types of 3D cameras. We integrate RGB-D images to form fragments of the scenes following [48]. During training, we randomly rotate each scene on the fly to augment the dataset. We use a popular hand-designed feature descriptor, FPFH [37], to compute correspondences. Note, however, that our pipeline is agnostic to the choice of feature and can also be used with learned features [8].

We follow the standard procedures in the 3D registration literature to generate candidate correspondences. First, since 3D scans often exhibit irregular densities, we resample the input point clouds using a voxel grid to produce a regular point cloud. We use voxel sizes of 2.5cm and 5cm for our experiments. Next, we compute FPFH features and find the nearest neighbor for each point in feature space to form correspondences. The correspondences obtained from this procedure often exhibit a very low inlier ratio, as little as 0.87% with a 2.5cm voxel size. Among these correspondences, we regard \( x \leftrightarrow x' \) as an inlier if it satisfies \( \|T(x) - x'\|_2 < \tau \) and all others as outliers. We set \( \tau \) to be two times the voxel size.

Finally, we use a registration method to convert the filtered correspondences into a final registration result. We show results with two different registration methods. The first is Fast Global Registration [51], which directly minimizes a robust error metric. The second is a variant of RANSAC [13] that is specialized to 3D registration [52].

**Evaluation.** We use three standard metrics to evaluate registration performance: rotation error, translation error, and success rate. The rotation error measures the absolute angular deviation from the ground truth rotation \( R \), \( \arccos \frac{\text{Tr}(R^T R) - 1}{2} \). Similarly, the translation error measures the deviation of the translation \( ||t - t'||_2 \). When we report these metrics, we exclude alignments that exceed a threshold following [8] since the results of the registration methods [51, 13] can be arbitrarily bad when registration fails. Finally, the success rate is the ratio of registrations that were successful; registration is considered successful if both rotation and translation errors are within the respective thresholds. For all experiments, we use a rotation error of 15 degrees and a translation error of 30cm as the thresholds.
Figure 5: The success rate of baseline methods and ours on the 3DMatch benchmark [48] with 5cm voxel size. FGR denotes registration with FPFH [37] and Zhou et al. [51], Yi et al. + X denotes FPFH filtering with Yi et al. [46] and registration with X, and Ours + X denotes our method for filtering followed by registration with X.

Table 3: Pairwise registration on 3DMatch test scenes with 2.5cm downsampling. Translation Error (TE), Rotation Error (RE), success rate. Registration is considered successful if TE < 30cm and RE < 15°.

<table>
<thead>
<tr>
<th>Scene</th>
<th>Inlier Ratio</th>
<th>FPFH + FGR</th>
<th>FPFH + Ours + FGR</th>
<th>FPFH + RANSAC</th>
<th>FPFH + Ours + RANSAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kitchen</td>
<td>1.62%</td>
<td>10.98 4.99</td>
<td>37.15</td>
<td>6.25 2.17</td>
<td>5.90 1.98</td>
</tr>
<tr>
<td>Home 1</td>
<td>2.71%</td>
<td>11.12 4.40</td>
<td>45.51</td>
<td>7.07 2.19</td>
<td>7.69 2.44</td>
</tr>
<tr>
<td>Home 2</td>
<td>2.83%</td>
<td>9.61 3.83</td>
<td>36.54</td>
<td>5.91 2.46</td>
<td>5.91 2.46</td>
</tr>
<tr>
<td>Hotel 1</td>
<td>1.35%</td>
<td>12.31 5.09</td>
<td>33.19</td>
<td>7.95 2.65</td>
<td>7.38 2.38</td>
</tr>
<tr>
<td>Hotel 2</td>
<td>1.54%</td>
<td>12.27 5.22</td>
<td>62.23</td>
<td>7.86 2.56</td>
<td>7.86 2.56</td>
</tr>
<tr>
<td>Hotel 3</td>
<td>1.59%</td>
<td>13.52 7.04</td>
<td>27.78</td>
<td>5.91 2.46</td>
<td>5.91 2.46</td>
</tr>
<tr>
<td>Study</td>
<td>0.87%</td>
<td>16.10 6.01</td>
<td>16.78</td>
<td>10.05 3.01</td>
<td>10.05 3.01</td>
</tr>
<tr>
<td>Lab</td>
<td>1.59%</td>
<td>10.48 4.80</td>
<td>42.86</td>
<td>8.01 2.31</td>
<td>8.01 2.31</td>
</tr>
</tbody>
</table>

Average: 12.05 5.17 33.10 7.23 2.39 67.85 7.60 2.56 48.37 6.82 2.22 72.02

Tab. 3 shows the 3D registration pipelines with and without our network to filter the outliers. Note that for FGR [51], we observe a considerable improvement with our network since FGR assumes more accurate correspondences as inputs. The improvement is smaller with RANSAC, since it is more robust to high outlier rates. Although the inlier ratio is as low as 1% for many 3D scene pairs, our network generates very accurate predictions. Similar to the linear regression experiments in Fig. 3, we find that the network converges very quickly. We compare to the model of Yi et al. [46] in Fig. 5 with 5cm voxel size to study the robustness of the 6-dimensional convolutional network to voxel size and find that the convolutional network improves the registration success rate significantly even for the higher inlier ratio seen with a 5cm discretization resolution. Qualitatively, our network accurately filters out outliers even in the presence of extreme noise (Fig. 6).

4.3. Filtering Image Correspondences

In this section, we apply the high-dimensional convolutional network to image correspondence inlier detection. In the projective space \( \mathbb{P}^2 \), an inlier correspondence \( u \leftrightarrow u' \) must satisfy \( u'^\top E u = 0 \), where \( E \) is an essential matrix, \( u \) denotes a normalized homogeneous coordinate \( u = K^{-1} x \), \( x \) is the corresponding homogeneous image coordinate, and \( K \) is the camera intrinsic matrix. When we expand \( u'^\top E u = 0 \), we get \( u'^\top A u + u'^\top B u + u'^\top C u^2 + u'^\top D u^2 + E u'^\top F u^2 + G u'^\top H u^2 + I = 0 \), which is a quadratic equation. If there is a real-valued solution, there are infinitely many solutions that form either an ellipse (sphere), a parabola, or a hyperbola. These are known as conic sections. Thus a set of ground-truth image correspondences will form a hyper conic section in 4-dimensional space. We use a convolutional network to predict the likelihood that a correspondence is an inlier.

Dataset: YFCC100M. We use the large-scale phototourism dataset YFCC100M [40] for the experiment. The dataset contains 100M Flicker images of tourist hot-spots with metadata, which is curated into 72 locations with camera extrinsics estimated using SfM [17]. We follow Zhang et al. [49] to generate a dataset and use 68 locations for training and the others for testing. We filtered any image pairs that have fewer than 100 overlapping 3D points from SfM to guarantee non-zero overlap between images.

We use SIFT features [25] to create correspondences and label a correspondence to be a ground-truth inlier if the symmetric epipolar distance of the correspondence is below a certain threshold using the provided camera parameters, i.e.,

\[
\left( \frac{r^2}{l_1^2 + l_2^2} + \frac{r^2}{l_1^2 + l_2^2} \right) < \tau, \tag{3}
\]
where \( l = u^T E = (l_1, l_2, l_3) \) is a homogeneous line, \( l' = Eu \), and \( r = u^T Eu \).

**Network.** We convert a set of candidate image correspondences into an order-5 sparse tensor with four spatial dimensions and vectorized features. The coordinates are defined as normalized image coordinates \( u \). We define integer coordinates by discretizing the normalized image coordinates with quantization resolution 0.01 and additionally use the normalized coordinates as features. We use state-of-the-art baselines on the YFCC dataset and two variants of convolutional networks for this task. The first variant is a U-shaped convolutional network (Ours); the second network is a ResNet-like network with spatial correlation modules (Ours + SC). Spatial correlation modules [49] are blocks of shared MLPs that encode global context from a set of correspondences. We use the balanced cross-entropy loss [46] for all experiments.

**Evaluation.** We use precision, recall, and F1 score to evaluate correspondence classification accuracy. We use \( \tau = 10^{-4} \) for the distance threshold to define ground truth correspondences. Our network predicts a correspondence to be accurate if its inlier probability prediction is above 0.5. We provide quantitative results in Tab. 4 and qualitative results in Fig. 7. Our approach (with the spatial correlation modules) outperforms the PointNet variants [46, 49] as measured by the F1 score, but not by a large margin. We attribute this to the sparse nature of SIFT keypoints. Unlike FPFH, where we densely sample keypoints, correspondences are only defined on sparse keypoints. Such sparsity leads to fewer neighbors in the high-dimensional space, which can result in the degeneration of the convolution to a multi-layer perceptron. Our convolutional networks remain effective in this setting, but their distinctive ability to leverage local geometric structure is not strongly utilized.

**5. Conclusion**

Many interesting problems in computer vision involve geometric patterns in high-dimensional spaces. We proposed high-dimensional convolutional networks for geometric pattern recognition. We presented a fully-convolutional network architecture that is efficient and able to find patterns in high-dimensional data even in the presence of severe noise. We validated the efficacy of our approach on tasks such as line and plane detection, 3D regis-
Table 4: Classification scores on the YFCC100M test set. We consider a correspondence a true positive if the symmetric epipolar distance is below $10^{-4}$ and the prediction confidence is above 0.5.

<table>
<thead>
<tr>
<th></th>
<th>Yi et al. [46]</th>
<th>Zhang et al. [49]</th>
<th>Ours</th>
<th>Ours + SC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec.</td>
<td>Recall</td>
<td>F1</td>
<td>Prec.</td>
</tr>
<tr>
<td>BUCKINGHAM</td>
<td>0.497</td>
<td>0.772</td>
<td>0.605</td>
<td>0.496</td>
</tr>
<tr>
<td>NOTRE DAME</td>
<td>0.581</td>
<td>0.894</td>
<td>0.705</td>
<td>0.629</td>
</tr>
<tr>
<td>REICHTAG</td>
<td>0.747</td>
<td>0.877</td>
<td>0.807</td>
<td>0.734</td>
</tr>
<tr>
<td>SACRE COEUR</td>
<td>0.658</td>
<td>0.871</td>
<td>0.750</td>
<td>0.662</td>
</tr>
<tr>
<td>Average</td>
<td>0.621</td>
<td>0.854</td>
<td>0.717</td>
<td>0.628</td>
</tr>
</tbody>
</table>

Figure 7: Matching results using Yi et al. [46] (first row), Zhang et al. [49] (second row), and Ours + SC (last row). We visualize correspondences with inlier probability above 0.5. We color a correspondence green if it is a true positive (symmetric epipolar distance smaller than $10^{-4}$) and red otherwise.

Acknowledgements

This work was partially supported by a National Research Foundation of Korea (NRF) grant and funded by the Korean government (MSIT) (No. 2020R1C1C1015260).
References


[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015. 1, 2


[37] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In ICRA, 2009. 1, 5, 6

[38] Schönberger, Johannes Lutz, and Jan-Michael Frahm. Structure-from-motion revisited. In CVPR, 2016. 1

[40] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 2016. 6

[41] Philip HS Torr and Andrew Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 2000. 2


[46] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *CVPR*, 2018. 1, 2, 3, 4, 5, 6, 7, 8


