GRAPH NEURAL NETWORKS FOR MULTI-IMAGE MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image feature matching is a fundamental part of many geometric computer vision applications, and using multiple images can improve performance. In this work, we formulate multi-image matching as a graph embedding problem then use a Graph Neural Network to learn an appropriate embedding function for aligning image features. We use cycle consistency to train our network in an unsupervised fashion, since ground truth correspondence is difficult or expensive to acquire. In addition, geometric consistency losses can be added at training time, even if the information is not available in the test set, unlike previous approaches which optimize cycle consistency directly. To the best of our knowledge, no other works have used learning for multi-image feature matching. Our experiments show that our method is competitive with other optimization based approaches.

1 INTRODUCTION

Feature matching is an essential part of Structure from Motion and many geometric computer vision applications. The goal in multi-image feature matching is to take 2D image locations from three or more images and find which ones correspond to the same point in the 3D scene. Methods such as SIFT feature matching (Lowe (2004)) followed by RANSAC (Fischler & Bolles (1981)) have been the standard for decades. However RANSAC-based approaches are limited to matching pairs of images, which can lead to global inconsistencies in the matching. Other works, such as Wang et al. Wang et al. (2017), have shown improvement in performance by optimizing cycle consistency, i.e. enforcing the pairwise feature matches to be globally consistent.

However, these multi-view consistency algorithms struggle in distributed and robust settings. Having image features suited for this task would help improve performance, and deep learning has revolutionized how image features are computed (Yi et al. (2016)). In this paper, we want to leverage the power of deep representations in order to compute feature descriptors that are robust across multiple views.

Unfortunately, there are obstacles to applying multi-view constraints directly to deep learning. To train networks, we need a large amount of labeled data. In the case of multi-image feature matching, one would need hand labeled point correspondences between images, which are difficult and expensive to obtain. Multi-view constraints are formulated in terms of sparse features, which traditional convolutional neural nets are not designed to handle. Additionally, in the case of multi-image feature matching, geometric constraints would be helpful in rejecting outlier matches. Geometric constraints, such as the epipolar constraint, can help disambiguate between visually similar features during training, and help the network learn better features robust to this kind of noise.

In this work, we propose to solve these problems using Graph Neural Networks (GNNs) to operate on the correspondence graph. The proposed method works directly on the correspondence graph, which is agnostic to how the correspondences were computed, thus allowing the algorithm to work in a broad class of environments. To the best of our knowledge this work is the first to apply deep learning to the multi-view feature matching problem. We use an unsupervised loss - the cycle consistency loss - to train the network, and thus avoiding the difficulty of expensive hand labeling. Geometric consistency losses can aid training, even if such information is not available at test time. Although our network is simple, it shows promising results compared to baselines which optimize for cycle-consistency without learned embeddings, using a matrix factorization loss (Zhou et al. (2015); Leonardos et al. (2016)). Furthermore, since inference requires only a single forward pass

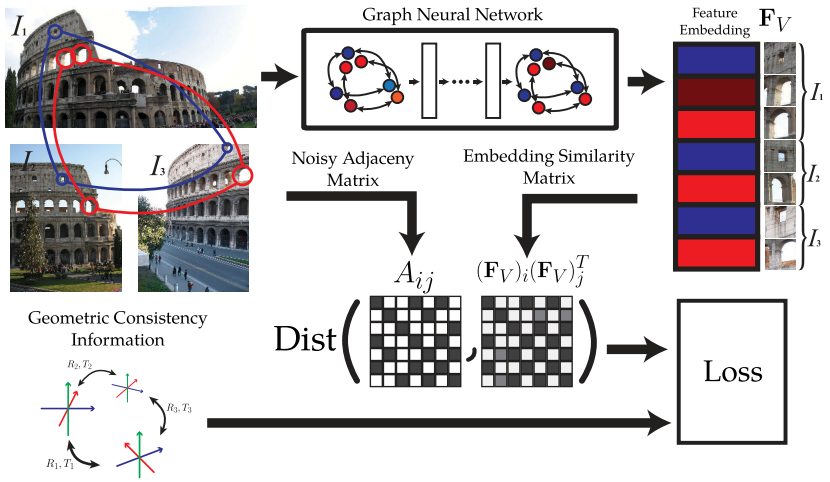


Figure 1: An illustration of the approach of this work. The Graph Neural Neural Network (GNN) (Battaglia et al. (2018)) takes as input the graph of matches and then outputs a low rank embedding of the adjacency matrix of the graph. The GNN operates on an embedding over the vertices of the graph. In the figure, the GNN vertex embeddings are represented by different colors. The final embedding is used to construct a pairwise similarity matrix, which we train to be a low dimensional cycle-consistent representation of the graph adjacency matrix, thus pruning the erroneous matches. We train the network using a reconstruction loss on the similarity matrix with the noisy adjacency matrix, and thus do not need ground truth matches. In addition, we can use geometric consistency information, such as epipolar constraints, to assist training the network.

over the neural network, our approach is faster (to achieve comparable accuracy) than methods which must solve an optimization problem every time. We perform experiments on the Rome16K dataset (Li et al. (2010)) to test the effectiveness of our method compared to optimization based methods. Our contributions in this work are:

- We use a novel architecture to address the multi-image feature matching problem using GNNs with graph embeddings.
- We introduce an unsupervised cycle consistency loss that does not require labeled correspondences to train.
- We demonstrate the effectiveness of geometric consistency losses in improving training.

2 Related Work

2.1 Feature Matching

Matching has a rich history of research in computer vision. Much work has be done using hand-crafted feature descriptors such as SIFT (Lowe (2004)), SURF (Bay et al. (2006)), BRIEF (Calonder et al. (2012)), or ORB (Mur-Artal et al. (2015)). RANSAC Fischler & Bolles (1981) is the most widely used robust estimation technique to filter out outliers from the matches. The combination of RANSAC and hand-crafted feature descriptors has constituted the bulk of the matching literature for the last 40 years. Finally, graph matching (Suh et al. (2015); Hu et al. (2016)) can be used as a final step for more robust matches.

2.2 Multi-image Matching

Multi-image matching has traditionally been done using optimization based methods minimizing a cycle consistency based loss (see Section 3.3). Pachauri et al. (2013) and Arrigoni et al. (2017) use the eigenvectors of the matching matrix to obtain a low dimensional embedding. However, the low Gaussian noise assumption is not realistic. Zhou et al. (2015) and Wang et al. (2017) use most sophisticated optimization techniques on the matching matrix and thus produce more robust

solutions. Leonardos et al. (2016) implement a distributed optimization scheme to solve for cycle consistency. As an alternative to optimization based techniques, Tron et al. (2017) used density based clustering techniques to compute multi-image correspondence. To the best of our knowledge, we are the first to use neural networks for multi-image matching.

2.3 Deep Learning for Matching

Previous attempts to improve image matching techniques using machine learning have focused on learning the descriptors given ground truth correspondence from curated datasets such as Zagoruyko & Komodakis (2015); Yi et al. (2016); and Brachmann et al. (2017). This approach is limited if one do not have the ability to get the ground truth correspondences. There are other methods to build correspondences such as Choy et al. (2016), but they only handle two-view constraints and require dense correspondences. Most similar to our work, Yi et al. (2018) attempts to improve correspondences by learning match probabilities for RANSAC for greater robustness and speed. However, they only focus on two view matching and do not exploit the advantages of the correspondence structure. Note that while Zhu et al. (2017) use cycle consistency in their loss, their method is restricted to pairwise cycle consistency (i.e. enforcing consistency by going back and forth between two images). We use multi-image cycle consistency, which require consistency between 3 or more images.

2.4 Graph Neural Networks

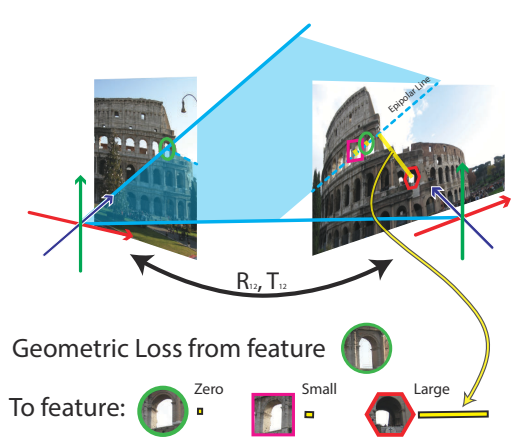
Graph neural networks have recieved more attention recently e.g. Bronstein et al. (2017); Bruna et al. (2013); Defferrard et al. (2016); Kipf & Welling (2017); Scarselli et al. (2009); Gama et al. (2018b;a); and Battaglia et al. (2018). Classical so-called Spectral methods used the eigenvectors of the graph Laplacian to compute convolutions as in Bruna et al. (2013), but requires an a-priori known graph structure. Non-spectral methods do not require a-priori knowledge, as seen in Bronstein et al. (2017); Kipf & Welling (2017); Scarselli et al. (2009); and Gama et al. (2018a). Most of these methods use polynomials of the graph Laplacian to compute neighborhood averages. Gama et al. (2018b;a) formalize this notion and generalize it beyond the use of the graph Laplacian. To improve performance, more sophisticated aggregation techniques and global information passing can be used as discussed in Battaglia et al. (2018).

3 Method

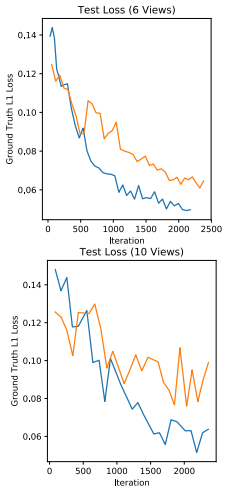
Our goal is to learn optimal features that capture multiple image views by filtering out noisy feature matches. An outline of our approach can be seen in figure 1. The input to our algorithm is a set of features and noisy correspondences, and the output is a new set of features where the pairwise similarities of these features correspond to the true matches. We do this by training the new set of feature embeddings to be cycle consistent. We formulate this problem in terms of the correspondence graph of the features. Graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ have a set of vertices $\mathcal{V}$ and of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ and implicitly in the subsequent discussions we assume the edges are directed. For a vertex $v \in \mathcal{V}$ we use $\mathcal{N}^h(v)$ to denote the h-hop neighbors of v, with the superscript left out for 1-hop neighbors. Similarly $\mathcal{E}(v)$ is used to denote the edges associated with v. To denote the vertices connected to an edge $e \in \mathcal{E}$ we write $e(v_1, v_2)$.

3.1 Correspondence Graph

We assume there is an initial set of feature matches represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with an associated adjacency matrix $\boldsymbol{A}$. The graph is constructed from putative correspondences of image features across images, typically constructed using feature descriptor distance (e.g. SIFT feature distance). While there are many interesting methods for computing these putative correspondences (Suh et al. (2015); Yi et al. (2018)), we do not explore them in this work. Typically putative correspondences are matched probabilistically, meaning a feature in one image matches to many features in another. The ambiguity in the matches could come from repeated structures in the scene, insufficiently informative low-level feature descriptors, or just an error in the matching algorithm. Filtering out these noisy matches is our primary learning goal.

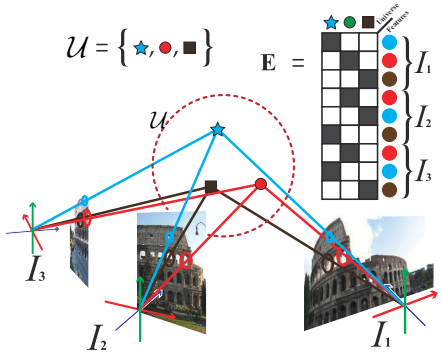


(a) Illustrated here is an example of how the geometric loss is computed for one feature.

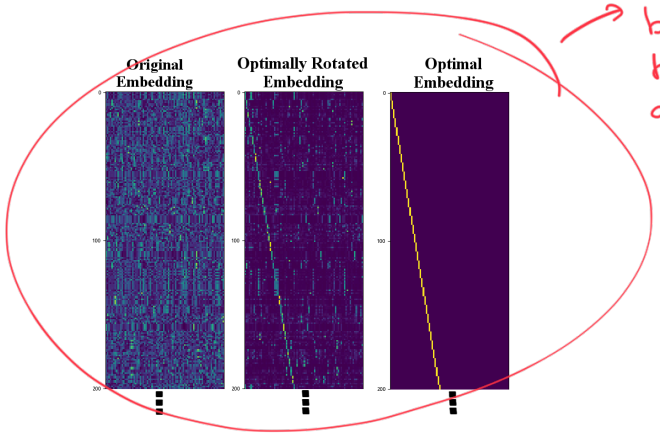


(b)

Figure 2: **(a)** Errors are computed via absolute distance from the epipolar line, as expressed by (6) via the epipolar constraint. The epipolar line is the line of projection of the feature on the first image, projected onto to the second. The distance to this line on the second image indicates how likely that point is to correspond geometrically to the original feature. There can be false positives along the projected line, as shown by the square feature in the figure, but other points will be eliminated, such as the hexagonal feature. Best viewed in color. **(b)** Training curves with and without Geometric Training loss, described in 8. The geometric training loss improves testing performance. While the training is higher due to the additional loss terms, the ground truth L1 error is substantially better with the Geometric Loss. Best viewed in color.



(a) Illustration of the universe of features.



(b)

Figure 3: **(a)** Each feature in each image corresponds to a 3D point in the scene. We can construct cycle consistent embeddings of the features by mapping each one to the one-hot vector of its corresponding 3D point. While there can be many features, there are fewer 3D points and thus this corresponds to a low rank factorization of the correspondence matrix. **(b)** Visualization of the learned embeddings. On the left we have the raw outputs, which are difficult to interpret. In the center, we rotated the features to best match the ground truth for a more interpretable visualization. On the right, we have the ground truth embeddings, given as indicator vectors for which feature in the world the points correspond to. Best viewed in color.

Each vertex of the graph $v \in \mathcal{V}$ is an image feature, corresponding to some ground truth 3D point $\boldsymbol{p}(v)$. Each edge $e = (v_1, v_2) \in \mathcal{E}$ is a potential correspondence. Associated with each vertex v is an embedding $\boldsymbol{f}_v \in \mathbb{R}^m$, which can include the visual feature descriptor, position, scale, orientation, etc. Similarly, each edge e has an associated feature $\boldsymbol{f}_e \in \mathbb{R}^p$ (in this work, initially just the weight of the feature association). We use these features as the initialization for our learning algorithm.

In the absence of noise or outliers, this graph would have a connected component for each visible point in the world, all mutually disjoint. Without noise, vertices v would only match with other vertices v' that correspond to the same 3D point in the scene. Since features in this case represent unique locations in the scene, no points in the same image would have edges e between them. Mathematically, this can be expressed as $e = (v_1, v_2) \in \mathcal{E} \implies \boldsymbol{P}(v_1) = \boldsymbol{P}(v_2)$. In the noisy case we expect this structure to be corrupted, i.e. there are some edges $e = (v_1, v_2) \in \mathcal{E}$ such that $\boldsymbol{P}(v_1) \neq \boldsymbol{P}(v_2)$. Thus we need to prune the erroneous edges.

However, standard CNNs cannot operate on this general graph structure. Thus we cannot use standard convolutional nets to learn features for this task. Instead we use graph networks to learn feature representations on this space, which we describe in the next section.

3.2 Graph Neural Networks

As input to our method we are given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the features described ~~last section~~ $\boldsymbol{f}_v \; \forall v \in \mathcal{V}$ and $\boldsymbol{f}_e \; \forall e \in \mathcal{E}$. ~~As with any neural network,~~ GNNs have layered outputs. We describe the output of layer k as $\boldsymbol{f}_v^{(k)} \in \mathbb{R}^{m_k} \; \forall v \in \mathcal{V}$ and $\boldsymbol{f}_e^{(k)} \in \mathbb{R}^{p_k} \; \forall e \in \mathcal{E}$, with the initial embeddings denoted $\boldsymbol{f}_v^{(0)} = \boldsymbol{f}_v$ and $\boldsymbol{f}_e^{(0)} = \boldsymbol{f}_e$. It will be useful to write these features as matrices in later sections, so we denote the vertex embedding matrix as $\boldsymbol{F}_V^{(k)}$ and the edge embedding matrix as $\boldsymbol{F}_E^{(k)}$. If a superscript is not specified then it refers to the final output of the network.

First we describe older methods of GNNs to give context, then we describe the method we use in this work. Many older methods assume we have the adjacency matrix $\boldsymbol{A}$ of the graph known a-priori Bruna et al. (2013), and can encode graph convolutions using the eigenvectors of $\boldsymbol{A}$. However, we do not have this luxury, as the correspondence structure changes from image set to image set, and thus we use non-spectral Graph Neural Networks. There are many variations on non-spectral methods, often which ultimately amount to message passing between vertices of the graph with learned non-linear transformations between various steps (Kipf & Welling (2017); Defferrard et al. (2016); Gama et al. (2018b;a)). Some works such as Gama et al. use pooling operations on the vertices to make the graph smaller and thus aid computation, but as we need labels on every vertex of the original graph, so we cannot use this. Most ~~of~~ GNNs used in these works can be expressed mathematically as:

$$\tilde{\boldsymbol{f}}_v^{(k+1)} = \sigma\left(b^{(k)} + \boldsymbol{W}_0^k \boldsymbol{f}_v^{(k)} + \sum_{h=0}^{H} \sum_{v' \in \mathcal{N}^h(v)} f_{e(v,v')} \boldsymbol{W}_h^k \boldsymbol{f}_{v'}^{(k)}\right)$$

The weights/biases $\boldsymbol{W}_h^k$, $b^{(k)}$ are all learned, with no learning done on the edge weights $f_{e(v,v')}$. Note that this is just sums or averages over h-hop neighborhoods, where the weights on the edges remain static through the computation. Given that we are trying to prune edges, it would be sensible to add features over edges to learn which ones to prune and which to keep such as in Scarselli et al. (2009).

~~Therefore,~~ in this work we use the method and implementation described in Battaglia et al. (2018) (more specifically repetitions of the architecture described in Battaglia et al. (2016)). This work is similar to previous works with one key difference: edges also have weights. Thus therefore there is intermediate processing on the edges before information is passed to the vertices.

Mathematically, this is expressed as:

$$\tilde{\boldsymbol{f}}_{e(v_1,v_2)}^{(k+1)} = \sigma\left(a^{(k)} + \boldsymbol{U}_1^{(k)} \boldsymbol{f}_e^{(k)} + \boldsymbol{U}_1^{(k)} \boldsymbol{f}_{v_1}^{(k)} + \boldsymbol{U}_2^{(k)} \boldsymbol{f}_{v_2}^{(k)}\right) \tag{1}$$

$$\tilde{\boldsymbol{f}}_v^{(k+1)} = \sigma\left(b^{(k)} + \boldsymbol{W}_0^{(k)} \boldsymbol{f}_v^{(k)} + \sum_{e \in \mathcal{E}(v)} \boldsymbol{W}_1^{(k)} \boldsymbol{f}_e^{(k+1)}\right) \tag{2}$$

Here the learned weights are denoted $\boldsymbol{W}$ and $\boldsymbol{U}$, and the biases $a^{(k)}$ and $b^{(k)}$. In Battaglia et al. (2018), they allow for more sophisticated aggregation functions, but in this work we simply use the mean function. In practice, we use MLPs in message passes between vertices and edges for better expressiveness

Method	Same Point Similarities	Different Point Similarities
Ideal	1.00e+0 ± 0.00e+0	0.00e+0 ± 0.00e+0
Initialization Baseline	5.11e-1 ± 1.68e-2	2.56e-1 ± 2.06e-1
3 Views, Noiseless	9.96e-1 ± 7.70e-3	1.16e-1 ± 1.32e-1
5 Views, Noiseless	1.00e+0 ± 4.15e-4	1.22e-1 ± 1.67e-1
3 Views, Added Noise	9.96e-1 ± 7.70e-3	1.16e-1 ± 1.32e-1
5 Views, Added Noise	9.89e-1 ± 2.47e-2	7.67e-2 ± 1.56e-1
6 Views, Added Noise	9.84e-1 ± 3.16e-2	7.46e-2 ± 1.57e-1
3 Views, 5% Outliers	9.29e-1 ± 1.79e-1	1.41e-1 ± 1.48e-1
3 Views, 10% Outliers	9.27e-1 ± 1.79e-1	1.40e-1 ± 1.51e-1

Table 1: Results on Synthetic correspondence graphs. The 'Same Point Similarities' column is the mean and standard deviation of similarities for true corresponding points, while the 'Different Point Similarities' is the same for points that do not correspond. For the 'Same Point Similarities' column higher is better, and for 'Different Point Similarities' lower is better. Losses tested against ground truth correspondence graph adjacency matrices. Our method was not trained on ground truth correspondences but using unsupervised methods.

3.3 Cycle Consistency

Let $\boldsymbol{M}$ be the noiseless set of matches between our features, with $\boldsymbol{M}_{ij}$ being the matches between image i and image j. If the pairwise matches are globally consistent, then for all i, j, k:

$$\boldsymbol{M}_{ij} = \boldsymbol{M}_{ik}\boldsymbol{M}_{kj} \tag{3}$$

In other words, the matches between two images stay the same no matter what path is taken to get there. This constraint is known as *cycle consistency*, and has been used in a number of works to optimize for global consistency (Zhou et al. (2015); Wang et al. (2017); Leonardos et al. (2016)). Stated in this form, there are $O(n^3)$ cycle consistency constraints to check. A more elegant way to represent cycle consistency is to first create a 'universe' of features that all images match to (see figure 3a). Then, one can match the i^{th} set of features to the universe using a ground-truth matching matrix $\boldsymbol{X}_i$. Then the cycle consistency constraint becomes:

$$\boldsymbol{M}_{ij} = \boldsymbol{X}_i\boldsymbol{X}_j^\top \tag{4}$$

This reduces the number of our constraints from $O(n^3)$ to $O(n^2)$. We try to learn vertex embeddings $\boldsymbol{F}_V$ to approximate $\boldsymbol{X}$ - in other words the final embedding should be an encoding of the universe of features. As we do not have the ground truth matches $\boldsymbol{M}$, we approximate it using the noisy adjacency matrix $\boldsymbol{A}$ of our correspondence graph. Thus our loss would be

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{F}_V) = \mathcal{D}(\boldsymbol{A}, \boldsymbol{F}_V\boldsymbol{F}_V^\top) \tag{5}$$

Here $\mathcal{D}$ could be an L_2 loss, L_1 loss, or many others. In this work, we use the L_1 loss. Note that because of this formulation, we can determine our embeddings only up to a rotation, as $\boldsymbol{E}R(\boldsymbol{E}R)^\top = \boldsymbol{E}RR^\top\boldsymbol{E}^\top = \boldsymbol{E}\boldsymbol{E}^\top$ Thus when visualizing embeddings, we rotate them to make them more interpretable (see figure 3b).

3.4 Geometric Consistency Loss

One of the main advantages of this approach over more traditional optimization based approaches is the ability to add geometric consistency information into the loss at training time, even if it is not available at test time. The simplest way to add geometric consistency losses, and the approach we use here, is to use the epipolar constraint. The epipolar constraint describes how the positions of features in different images corresponding to the same point should be related. An illustration of this is provided in figure 2a, showing how this loss can help reject erroneous points. Given a relative pose (R_{ij}, T_{ij}) between two cameras i and j (transforms j to i) the epipolar on corresponding feature locations X_i and X_j: $X_i^\top[T_{ij}]_\times R_{ij}X_j = 0$. In this work we use the two pose epipolar constraint (Tron & Daniilidis (2014)):

$$X_i^\top R_i^\top [T_j - T_i]_\times R_j X_j = 0 \tag{6}$$

Method (6 Views)	L_1	L_2	Area under ROC	Time (sec)
MatchALS 15 Iterations	0.101 ± 0.008	0.022 ± 0.004	0.918 ± 0.073	0.074 ± 0.008
MatchALS 35 Iterations	0.046 ± 0.016	0.010 ± 0.005	0.910 ± 0.072	0.139 ± 0.041
MatchALS 50 Iterations	0.029 ± 0.017	0.008 ± 0.005	0.905 ± 0.068	0.260 ± 0.048
PGDDS0 15 Iterations	0.017 ± 0.002	0.007 ± 0.001	0.918 ± 0.087	0.796 ± 0.147
PGDDS0 25 Iterations	0.016 ± 0.002	0.007 ± 0.002	0.919 ± 0.087	1.670 ± 0.328
PGDDS0 50 Iterations	0.015 ± 0.002	0.006 ± 0.002	0.920 ± 0.087	3.363 ± 0.528
Spectral	0.073 ± 0.006	0.027 ± 0.003	0.921 ± 0.083	0.036 ± 0.005
GNN (ours)	0.044 ± 0.005	0.031 ± 0.005	0.872 ± 0.081	0.765 ± 0.046
Method (10 Views)	L_1	L_2	Area under ROC	Time (sec)
MatchALS 25 Iterations	0.092 ± 0.008	0.019 ± 0.003	0.912 ± 0.055	0.228 ± 0.014
MatchALS 35 Iterations	0.065 ± 0.009	0.013 ± 0.003	0.907 ± 0.053	0.355 ± 0.073
MatchALS 50 Iterations	0.045 ± 0.012	0.011 ± 0.004	0.914 ± 0.051	0.455 ± 0.022
PGDDS 15 Iterations	0.017 ± 0.001	0.008 ± 0.001	0.903 ± 0.061	1.225 ± 0.159
PGDDS 25 Iterations	0.016 ± 0.001	0.007 ± 0.001	0.904 ± 0.061	2.637 ± 0.357
PGDDS 50 Iterations	0.016 ± 0.001	0.007 ± 0.001	0.905 ± 0.061	6.116 ± 1.009
Spectral	0.073 ± 0.005	0.029 ± 0.002	0.912 ± 0.057	0.081 ± 0.021
GNN (ours)	0.053 ± 0.006	0.035 ± 0.005	0.872 ± 0.061	2.438 ± 0.070

Table 2: Results on Rome16K Correspondence graphs, showing the mean and standard deviation of the L_1 and L_2. Our method was not trained on ground truth correspondences but using unsupervised methods and geometric side losses. Thus we test against ground truth correspondence graph adjacency matrices computed from the bundle adjustment output. Our method performs better than 25 iteration of the MatchALS (Zhou et al. (2015)) method, but does not perform as well as 50 iterations. We do not perform as well as the Projected Gradient Descent Doubly Stochastic (PGDDS) (Leonardos et al. (2016)) but we perform significantly faster than them. We perform better than a simple eigenvalue based method (Pachauri et al. (2013)). Note that we perform much better in L_1 performance rather than L_2, as we optimized the network weights using an L_1 loss.

The constraint assumes that the X_k are calibrated (i.e. the camera intrinsics are known). Given our vertex embeddings matrix $\boldsymbol{F}_V$, we can formulate a loss between all cameras i and j:

$$\mathcal{L}_{ij,geom}(\boldsymbol{F}_V) = \sum_{v \in \boldsymbol{V}(i), u \in \mathcal{V}(j)} (\boldsymbol{f}_v \cdot \boldsymbol{f}_u) \left| X_v^\top R_i^\top [T_j - T_i]_\times R_{c_j} X_u \right| \tag{7}$$

With $\mathcal{V}(i)$ being the vertices associated with camera i. For our purposes, since we use low rank embeddings $\boldsymbol{E}_i$, $\boldsymbol{E}_j$, the loss would read (where $c(k)$ is the appropriate camera for point index k):

$$\mathcal{L}_{geom}(\boldsymbol{F}_V) = \text{tr}(\boldsymbol{G}^\top \boldsymbol{F}_V \boldsymbol{F}_V^\top) = \sum_{k,l} (\boldsymbol{F}_V)_k \cdot (\boldsymbol{F}_V)_l (\boldsymbol{G})_{kl} \tag{8}$$

$$(\boldsymbol{G})_{kl} = \left| X_k^\top R_{c(k)}^\top [T_{c(l)} - T_{c(k)}]_\times R_{c(l)} X_l \right|$$

4 Experiments

4.1 Synthetic Graph Dataset

We first test our method on synthetically generated data as a simple proof of concept. As these were simpler datasets, we the simpler edge-feature free model of (Kipf & Welling (2017)). To generate the data, we generate p points, each with its own randomly generated descriptor. To create the graph, we generate random permutation matrices, with a noise applied to it after it is generated. We initialize the input descriptors using the true descriptor, plus some added Gaussian noise. No geometric losses were added during training for these experiments. However, the method was robust in testing with different noise functions and parameters. The normalized noisy input descriptors are our baseline - they correlate with the true values but do not preserve the structure well. However, the GNN recovered the true structure very well, as shown in Table 1. All experiments were run with a 12 layer GNN with the ReLU nonlinearity and skip connections. The feature vector lengths were

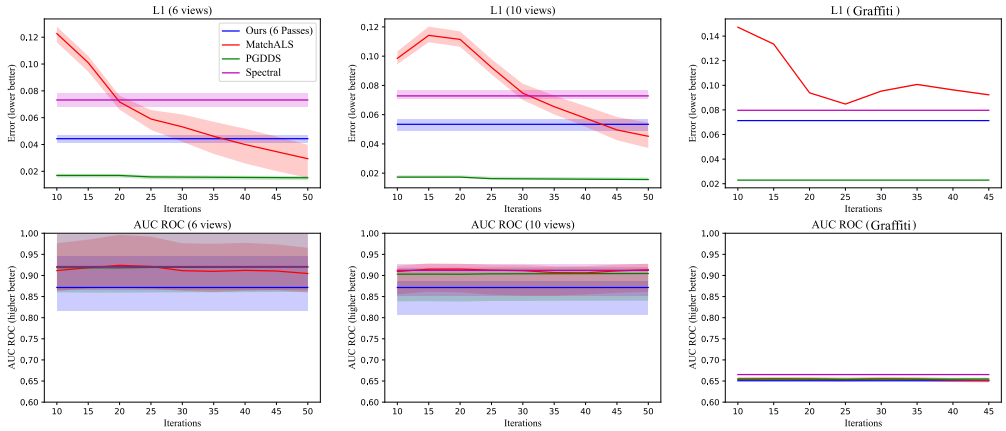


Figure 4: Plot of the losses of the baselines at different iteration numbers. The line shows the mean of the graph while the translucent coloring shows the 25^{th} to 75^{th} percentiles. The ROC AUC curves remain fairly consistent while the L1 loss goes noticibly down after more iterations. Our method compares to 35-45 iterations of MatchALS, while only having 16 layers and 8 message passes. PGDDS performs better than us in L_1 but we perform similarly in the ROC AUC metric.

32, 64, 128, 256, 512, 512, 512, 512, 512, 512, 1024, 1024, with skip connections between layers 1 and 6, 6 and 12, and 1 and 12. All were trained with the Adam optimizer (Kingma & Ba (2014)) and a learning rate 10^{-4} The network was implemented in Tensorflow (Abadi et al. (2015)), version 1.11. With this simple test on synthetic data passed, we now move to more challenging datasets.

4.2 Rome 16K Graph Dataset

We use the Rome16K dataset (Li et al. (2010)) to test our algorithm in real world settings. Rome16K consists of 16 thousand images of various historical sites in Rome extracted from Flickr, along with the 3D structure of the sites provided by bundle adjustment. While not a standard dataset to test cycle consistency, other datasets had insufficient data to train a network on. Rome16K is typically used to test bundle adjustment methods. Therefore, to use our method, we extract image triplets and quadruplets with overlap of 80 points or more to test our algorithm, with the points established as corresponding in the given bundle adjustment output. For the initial embedding we use the original 128 dimensional SIFT descriptors, normalized to have unit L_2 norm, the calibrated x-y position, the orientation, and log scale of the SIFT feature. To construct the graph, we take each feature as a vertex and create edges to the 5 nearest SIFT descriptors for the other images.

For these experiments we train with the L_1 norm and geometric consistency losses. We evaluate on a test set using the ground truth adjacency matrix, which we compute from the bundle adjustment given by the Rome16K dataset. However, we do not train with the ground truth adjacency matrix, only with the noisy adjacency matrix from the graph. We also add the geometric loss (8) which helps improve testing performance (see figure 2b) We use the L_1 and ROC AUC metrics to measure performance. For this method to work, we need the dimension of the embedding to be at least the number of unique points in the scene. Picking the correct number is difficult a-priori, and is a problem with all cycle consistency based methods. Here we use the ground truth dimension of the embedding to test both our method and the baselines.

The network was implemented using the code provided by Battaglia et al. (2018) using Tensorflow 1.11 (Abadi et al. (2015)). Our network has 16 layers, with 8 message passing operations placed every other layer. All layers were simple Multi-layer Perceptrons, with no batch norm. The network was trained with the Adam optimizer (Kingma & Ba (2014)) with a learning rate of 10^{-4}, with an exponentially decaying learning rate. We incorporate skip connections between the input, 6^{th}, and 12^{th} layers (all possible pairs).

We compare our method to spectral and optimization based baselines with different maximum iteration cutoffs. Figure **??** illustrates this by plotting the means of various metrics and their 25^{th} and 75^{th} percentiles, with table 2 giving the exact numbers. Our network, though only using 8 message

How do the highlighted terms relate?

passes, has comparable accuracy to MatchALS (Zhou et al. (2015)) run 35 to 45 iterations, with an equivalent message passing step at each phase. Although our method does not outperform the Projected Gradient Descent - Doubly Stochastic (PGDDS) (Leonardos et al. (2016)) method, we perform comparably to them in the ROC AUC metric.

5 Conclusion

This sentence is actually just gibrish to me

We have shown a novel method for training feature matching using GNNs, using an unsupervised cycle consistency loss and geometric consistency losses. We have demonstrated has comparable the traditional optimization based baselines on a simple GNN, while still allowing for end-to-end training integration in deep learning pipelines. For future work, we will investigate robust losses for better outlier rejection, and using higher order geometric constraints, such as the tri-focal tensor, as additional loss terms. With this new architecture, we have the capability of training multi-image matching pipelines end to end, thus allowing us to train for image features explicitly for this task. We can also extend this to distributed settings where we can train for matching images from multiple distributed agents.

References

was thinking "holy crap, that's a lot of authors - "

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Federica Arrigoni, Eleonora Maset, and Andrea Fusiello. Synchronization in the symmetric inverse semigroup. In *International Conference on Image Analysis and Processing*, pp. 70–81. Springer, 2017.

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pp. 4502–4510, 2016.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pp. 404–417. Springer, 2006.

Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

Michael Calonder, Vincent Lepetit, Mustafa Ozuysal, Tomasz Trzcinski, Christoph Strecha, and Pascal Fua. Brief: Computing a local binary descriptor very fast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1281–1298, 2012.

Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. In *Advances in Neural Information Processing Systems*, pp. 2414–2422, 2016.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pp. 3844–3852, 2016.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Fernando Gama, Antonio G Marques, Geert Leus, and Alejandro Ribeiro. Convolutional neural network architectures for signals supported on graphs. *IEEE Transactions on Signal Processing*, 67(4):1034–1049.

Fernando Gama, Geert Leus, Antonio G Marques, and Alejandro Ribeiro. Convolutional neural networks via node-varying graph filters. In *2018 IEEE Data Science Workshop (DSW)*, pp. 1–5. IEEE, 2018a.

Fernando Gama, Antonio G Marques, Alejandro Ribeiro, and Geert Leus. Mimo graph filters for convolutional neural networks. *arXiv preprint arXiv:1803.02247*, 2018b.

Nan Hu, Boris Thibert, and Leonidas Guibas. Distributable consistent multi-graph matching. *arXiv preprint arXiv:1611.07191*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Spyridon Leonardos, Xiaowei Zhou, and Kostas Daniilidis. Distributed consistent data association. *arXiv preprint arXiv:1609.07015*, 2016.

Yunpeng Li, Noah Snavely, and Daniel P Huttenlocher. Location recognition using prioritized feature matching. In *European conference on computer vision*, pp. 791–804. Springer, 2010.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

Deepti Pachauri, Risi Kondor, and Vikas Singh. Solving the multi-way matching problem by permutation synchronization. In *Advances in neural information processing systems*, pp. 1860–1868, 2013.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

Yumin Suh, Kamil Adamczewski, and Kyoung Mu Lee. Subgraph matching using compactness prior for robust feature correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5070–5078, 2015.

Roberto Tron and Kostas Daniilidis. On the quotient representation for the essential manifold. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1574–1581, 2014.

Roberto Tron, Xiaowei Zhou, Carlos Esteves, and Kostas Daniilidis. Fast multi-image matching via density-based clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4057–4066, 2017.

Qianqian Wang, Xiaowei Zhou, and Kostas Daniilidis. Multi-image semantic matching by mining consistent features. *arXiv preprint arXiv:1711.07641*, 2017.

Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*, pp. 467–483. Springer, 2016.

Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, number CONF, 2018.

Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4353–4361, 2015.

Xiaowei Zhou, Menglong Zhu, and Kostas Daniilidis. Multi-image matching via fast alternating minimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4032–4040, 2015.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.