

NM-Net: Mining Reliable Neighbors for Robust Feature Correspondences

Chen Zhao¹, Zhiguo Cao¹, Chi Li¹, Xin Li², and Jiaqi Yang *¹

¹School of Artificial Intelligence and Automation, Huazhong University of Science and Technology

²Lane Department of Computer Science and Electrical Engineering, West Virginia University

¹{*hust_zhao, zgcao, li_chi, jqyang*} @hust.edu.cn

²Xin.Li @mail.wvu.edu

Abstract

Feature correspondence selection is pivotal to many feature-matching based tasks in computer vision. Searching for spatially k -nearest neighbors is a common strategy for extracting local information in many previous works. However, there is no guarantee that the spatially k -nearest neighbors of correspondences are consistent because the spatial distribution of false correspondences is often irregular. To address this issue, we present a compatibility-specific mining method to search for consistent neighbors. Moreover, in order to extract and aggregate more reliable features from neighbors, we propose a hierarchical network named NM-Net with a series of convolution layers taking the generated graph as input, which is insensitive to the order of correspondences. Our experimental results have shown the proposed method achieves the state-of-the-art performance on four datasets with various inlier ratios and varying numbers of feature consistencies.

1. Introduction

Searching for good feature correspondences (a.k.a. matches) is a fundamental step in computer vision tasks - e.g., structure from motion [32], simultaneous location and mapping [2], panoramic stitching [4], and stereo matching [14]. Finding consistent feature correspondences between two images relies on two key steps [22, 3, 23] - i.e., *feature matching* and *correspondence selection*. Specifically, initial correspondences can be obtained by matching local key-point features such as SIFT [22]. Due to various reasons (e.g., key-point localization errors, limited distinctiveness of local descriptors, and illumination/viewpoint changes), mismatches are often inevitable. To address this issue, correspondence selection can be employed as a postprocessing

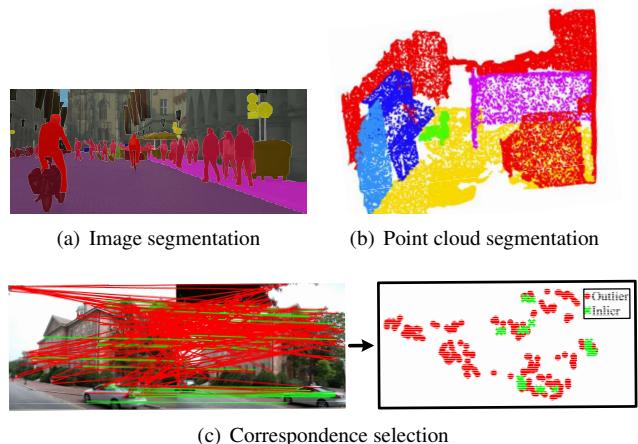


Figure 1. Comparison of (c) correspondence selection (can be viewed as a binary classification task for each correspondence) with standard segmentation problems including (a) image segmentation and (b) point cloud segmentation. Most spatially adjacent pixels and points in (a) and (b) are semantically consistent (belonging to the same class), yet the spatial distribution of mismatches in (c) is irregular, resulting in many outliers (red dots) contaminated in the local regions around inliers (green crosses). For visualization, image correspondences (4D) are projected to a 2D space via tsne [24].

step to ensure correct matches and improve the accuracy [3]. This paper focuses on a *learning-based* approach toward selecting correct matches from an initial set of feature correspondences [36].

Feature correspondence selection is challenging due to the scarcity of available information as well as the limitation of local feature representations. Spatial positions of matched features are discrete and irregular (note that RGB or texture information is not available any more). To effectively mine the consistency from raw positions, spatially local information is often employed in previous hand-crafted

*Corresponding author

algorithms [20, 3, 23]. Indeed, spatially local information has played an important role in image segmentation [21] and point cloud segmentation [26]. As shown in Fig. 1 (a) and (b), most feature points located in adjacent regions are semantically consistent (belonging to the same class). However, spatially local information is unreliable for correspondence selection due to irregular distribution of mismatches. As shown in Fig. 1 (c), around the vicinity of correct correspondences (marked by green crosses), a large number of mismatches (denoted by red dots) can be found. To overcome this difficulty, we present a compatibility-specific neighbor mining algorithm to search for top- k consistent neighbors of each correspondence. Correspondences are determined to be compatible if they meet the same underlying constraint [9, 1, 5]. When compared with spatially k -nearest neighbor (knn) search, the proposed neighbor mining approach is more reliable because potential inliers exhibit guaranteed consistency with each other [17].

Besides neighbor mining, another important issue is to find a proper representation for correspondence selection. Representations learned by various convolution neural networks (CNN) have become the standard in many computer vision tasks [16, 31, 13]. Correspondence selection can also be regarded as a binary classification problem for each match - *i.e.*, correct (inlier) vs. false (outlier). Nevertheless, it is often impractical to directly use a CNN to extract features from unordered and irregular correspondences. The first learning-based method for correspondence selection using multi-layer perceptron is proposed recently in [36], but unfortunately it has ignored useful local information such as those obtained by compatibility-specific neighbor mining (shown to be advantageous in our work). To fill this gap, we propose a hierarchical deep learning network called NM-Net (neighbors mining network), where features are successively extracted and aggregated. Compatibility-specific local information is utilized from two aspects: 1) a graph is generated for each correspondence where the nodes denote compatible neighbors found by our neighbor mining approach; 2) features are extracted and aggregated by a set of convolution layers that take the generated graph as input.

In a nutshell, the contributions of this paper are as follows:

- We suggest that compatibility-specific neighbors are more reliable (have stronger local consistency) for feature correspondences than spatial neighbors.
- We propose a deep classification network called NM-Net¹ that fully mines compatibility-specific locality for correspondence selection, with hierarchically extracted and aggregated local correspondences. Our

¹The code will be available at <https://github.com/sailor-z/NM-Net>

network is also insensitive to the order of correspondences.

- Our method achieves the state-of-the-art performance on comprehensive evaluation benchmarks including correspondences with various proportions of inliers and varying numbers of feature consistencies.

2. Related Work

Parametric methods. Generation-verification is arguably the most popular formulation of parametric methods such as RANSAC [9] and its variations (*e.g.*, PROSAC [6], LO-RANSAC [7], and USAC [27]). The consistency (compatibility) of correspondences is searched under a global constraint. Specifically, the generation and verification procedures are alternatively used to estimate a global transformation - *e.g.*, homography matrix or essential matrix. Correspondences consistent with the transformation are selected as inliers. Parametric methods have two fundamental weaknesses: 1) the accuracy of estimated global transformation severely degrades when initial inlier ratio is low [18] because the sampled correspondences may include no inlier; 2) the assumed global transformation is unsuitable for the case of multi-consistency matching [37] and non-rigid matching [23].

Non-parametric methods. Leveraging local information for correspondence selection is a popular strategy in non-parametric methods. For example, a locality preserving matching algorithm is presented in [23], which assumes that local geometric structures in the vicinity of inliers are invariant under rigid/non-rigid transformations, where the spatially knn search is utilized to represent variations of local structures. The spatially local information is exploited in a statistical manner in [3]. The similarity of local regions between two images is measured by the number of correspondences; all correspondences located in the regions are considered inliers if the number is larger than a predefined threshold. Additionally, local compatible information has been explored by other non-parametric approaches - *e.g.*, [1] measured the compatibility of each two correspondences as a payoff in a game-theoretic framework, where the probability of correct correspondences is iteratively calculated by ESS's algorithm [34]; [17] estimated an affinity matrix to represent the compatibility of correspondences and proposed a spectral technique to select inliers. Although these algorithms involve the compatibility information among correspondences, they do not sufficiently mine local information from compatible correspondences. By contrast, we use compatibility-specific neighbors to integrate local information to each correspondence via a data-driven manner.

Learning-based methods. Deep learning has achieved great success in recent years - *e.g.*, image classification [16,

31, 13], object detection [11, 10, 29], and image segmentation [21]. However, directly employing a standard CNN is infeasible to correspondence selection because correspondence representations are irregular and unordered. In [36], a deep learning framework based on multi-layer perceptron is employed to find inliers but without any local feature extraction or aggregation. Considering that point-cloud data has the similar characteristics with correspondences, PointNet [25] and PointNet++ [26] can be referred for correspondence selection, which have been developed for point cloud classification and segmentation recently. Nonetheless, each point is individually processed in PointNet without any local information involved; spatially nearest information is exploited by PointNet++ in a grouping layer even though such spatially local information could be unreliable for correspondences. Different from these learning-based methods for irregular data, our approach covers both locality selection and locality integration concerns via a compatibility metric and a hierarchical manner, respectively.

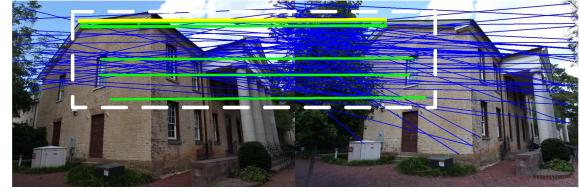
3. Motivation

Finding consistency (compatibility) among matches and selecting good correspondences is a chicken-and-egg problem: finding the consistency (for an assumed global transformation as an example) would require a set of inliers (*i.e.*, the knowledge about good correspondences); but meantime, the selection of inliers also relies on the results of finding reliable consistency. To get around this circular problem, we propose to utilize local information of correspondences as an surrogate representation for feature consistency.

Local information has been the cornerstone in many learning-based methods for image/point cloud classification and segmentation[31, 13, 21, 25], where local context features in convolution kernels are commonly extracted. Since correspondence selection can be considered as a binary classification problem for each correspondence (*i.e.*, inlier vs. outlier), it appears plausible to mine reliable local information for establishing good correspondences. As aforementioned, hand-crafted methods have employed spatially local information to select correct matches. However, unlike the cases of images and point clouds, directly using spatially local information for feature correspondence selection is not a good idea as shown in Fig. 2 (a) - the spatially selected k -nearest neighbors of an inlier are incompatible with two outliers. By contrast, the neighbors picked by a compatibility metric (which will be formally defined in the next section) are consistent as shown in Fig. 2 (b), but their positions are not necessarily to be spatially adjacent to the query correspondence. Since the matched keypoints of an inlier indicate the same 3D position from different viewpoints in the real world, the consistency among inliers is readily guaranteed. Such observation motivates us to develop a compatibility-specific method to mine consis-



(a) Spatially k -nearest neighbors



(b) Compatibility-specific k -nearest neighbors

Figure 2. Visual illustration of (a) spatially k -nearest neighbors and (b) compatibility-specific k -nearest neighbors of feature correspondences. The blue lines represent initial feature correspondences between two images, the yellow line denotes a sampled inlier, and the green lines and red lines respectively imply the neighbors of the sampled inlier being inliers and outliers. Two outliers are included in (a) the spatially k -nearest neighbors, which are considered as inconsistent portions; (b) the compatibility-specific k -nearest neighbors are consistent without outliers, but their positions are not necessarily to be spatially close.

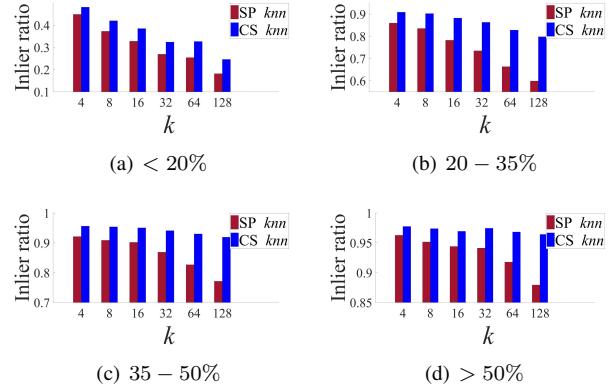


Figure 3. Comparison between the spatially k -nn (SP k -nn) and the compatibility-specific k -nn (CS k -nn). Hundreds of image pairs are sampled from our experimental datasets (Sect. 5) and divided into 4 parts according to inlier ratios of the initial correspondence sets, - *i.e.*, (a) lower than 20%, (b) from 20% to 35%, (c) from 35% to 50%, and (d) higher than 50%. The assessment metric is the average inlier ratio of the neighbors of correct correspondences.

tent neighbors.

The superiority of compatibility-specific neighbor mining can be further justified by the statistics collected from experimental datasets (Sect. 5) as shown in Fig. 3. The percentages of inliers in compatibility-specific nearest neighbors are remarkably higher than those results in spatially

nearest neighbors in all cases, with the gap being more dramatic as k increases. These empiric findings strongly suggest that neighbors chosen by the compatibility-specific method are more reliable.

4. Method

In addition to neighbor mining, correspondence selection requires an appropriate representation of the information provided by reliable neighbors. Built upon the success of deep learning in many visual recognition tasks [13, 29, 21], a learning-based method was developed for correspondence selection in [36]. A key strategy behind [36] is to use multi-layer perceptron that individually processes unordered correspondences. Unfortunately, this recent work fails to integrate local information for each correspondence; a key new insight of our work is to demonstrate the benefits of exploiting locality for correspondence selection by our proposed NM-Net. Our framework employs ResNet [13] as the backbone and compatibility-specific neighbor mining algorithm as the grouping module. The 4D raw correspondences are taken as input and the classification of each correspondence (*i.e.*, inlier or outlier) is output.

4.1. Problem Statement

Given a pair of images (I, I') , two sets of discrete keypoints $(\mathcal{K} = \{\mathbf{k}\}, \mathcal{K}' = \{\mathbf{k}'\})$ are detected and local patterns around those keypoints are described as $(\mathcal{P} = \{\mathbf{p}\}, \mathcal{P}' = \{\mathbf{p}'\})$. The initial correspondence set $\mathcal{C} = \{c\}$ is generated by brute-force matching between $(\mathcal{K}, \mathcal{K}')$ based on descriptor similarities. The correspondence selection boils down to a binary classification problem focusing on the classification of each c as an inlier $c_{inlier} \in \mathcal{C}_{inlier}$ or an outlier $c_{outlier} \in \mathcal{C}_{outlier}$.

4.2. Mining Neighbors

Compatibility-specific neighbor mining plays a crucial role in our network for the following two reasons. First, it explores the local space of each correspondence and extracts local information by our proposed compatibility metric. Second, it integrates unordered correspondences into a graph in which nodes correspond to the mined neighbors so that convolutions can be performed for further feature extraction and aggregation. As mentioned in Sect. 3, given a pair of correspondences (c_i, c_j) , quantifying the compatibility score denoted by $s(c_i, c_j)$ is non-trivial due to lacking of label information. One promising attack is that the variations of local structures around $(\mathbf{k}_i, \mathbf{k}'_i)$ and $(\mathbf{k}_j, \mathbf{k}'_j)$ are similar if c_i and c_j are compatible [1]. Based on this important observation, we propose to compute $s(c_i, c_j)$ by exploiting the variations as follows.

First, the Hessian-affine detector [15] is used to detect keypoints; it provides the local affine information around

keypoints that is required by the introduced compatibility metric. Local affine information is critical for searching for consistent correspondences when images undergo viewpoint or scale changes [5]. Second, a transformation characterizing the variation between local structures around $(\mathbf{k}_i, \mathbf{k}'_i)$ of c_i can be calculated by

$$\mathbf{H}_i = \mathbf{T}'_i \mathbf{T}_i^{-1}, \quad (1)$$

where a pair of 3×3 matrices $(\mathbf{T}_i, \mathbf{T}'_i)$ describe the positions and local structures of the matched keypoints. Then we can calculate \mathbf{T}_i by (the same for \mathbf{T}'_i)

$$\mathbf{T}_i = \begin{bmatrix} \mathbf{A}_i & \mathbf{k}_i \\ \mathbf{0} & 1 \end{bmatrix}, \quad (2)$$

where \mathbf{A}_i is a 2×2 matrix representing the local affine information extracted by the Hessian-affine detector and \mathbf{k}_i is the position of the keypoint. Third, intuitively, c_i and c_j are compatible if the corresponding transformations \mathbf{H}_i and \mathbf{H}_j are consistent; in other words, local structure variations estimated by a consistent pair of transformations should be similar. Consequently, we have adopted reprojected errors that represent local structure variations to measure the dissimilarity of $(\mathbf{H}_i, \mathbf{H}_j)$ by

$$e_j(c_i) = \left| \rho(\mathbf{H}_j \cdot \begin{bmatrix} \mathbf{k}_i \\ 1 \end{bmatrix}) - \rho(\mathbf{H}_i \cdot \begin{bmatrix} \mathbf{k}_i \\ 1 \end{bmatrix}) \right|, \quad (3)$$

where $\rho([a \ b \ c]^T) = [a/c \ b/c]^T$. Note that the compatibility of (c_i, c_j) is negatively correlated with the sum of reprojected errors $(e_j(c_i) + e_i(c_j))$. As a strategy of normalizing s_{ij} to the range of $(0, 1]$, we propose to use a Gaussian kernel - *i.e.*,

$$s(c_i, c_j) = e^{-\lambda(e_j(c_i) + e_i(c_j))}, \quad (4)$$

where λ is a hyper-parameter. Note that λ will not affect the ranking of $s(c_i, c_j)$; so the search of compatible neighbors for each correspondence is insensitive to λ . For any given c_i , a graph \mathcal{G}_i is generated by first selecting the neighbors of c_i , - *i.e.*, $\{c_i^j\} (j = 1, 2, \dots, k)$, as those correspondences with top- k $s(c_i, c_i^j)$, and then sequentially linking c_i with all its neighbors.

4.3. Network Architecture

The network architecture as shown in Fig.4 includes two key modules - *i.e.*, *grouping* and *ResNet* blocks. Our network design is partially inspired by hierarchically extracted features to fully leverage correspondence-level local information (*e.g.*, PointNet++ [26]). The details of NM-Net are given as follows.

Feature extraction and aggregation. In NM-Net, features are extracted and aggregated along three lines. First, we

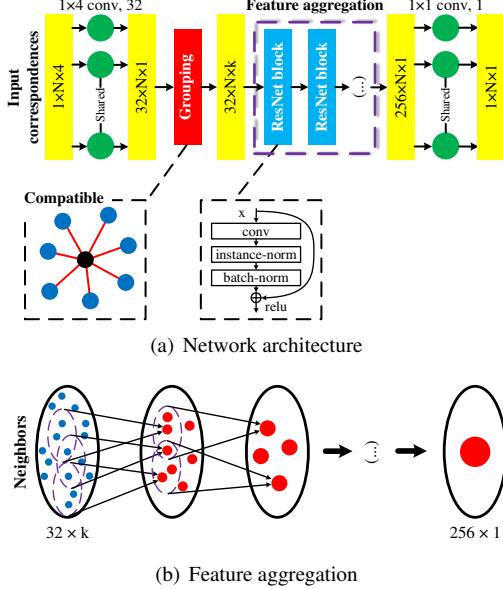


Figure 4. NM-Net architecture. NM-Net is (a) a classification network for feature correspondences. A grouping module is designed to first mine reliable neighbors via the compatibility metric and then convert them into a graph for each correspondence to achieve ordered representations. In the bottom-left dashed box of (a), the black dot denotes a query correspondence and those blue dots are its compatibility-specific neighbors. (b) illustrates the hierarchical aggregation of the features extracted from neighbors by a series of ResNet blocks.

use a grouping module to extract local information for each correspondence, where compatibility-specific knn search is adopted. The unordered raw correspondences are converted to N graphs $\mathcal{G}_i (i = 1, 2, \dots, N)$, where the nodes are sorted by $s(c_i, c_j^l)$, resulting in a set of regular organizations that are invariant to the order of correspondences. Second, features are hierarchically extracted and aggregated by a set of convolutions as

$$\mathbf{f}_j^n = \sum_{j=1}^d \omega_j \mathbf{f}_j^{n-1}, \quad (5)$$

where \mathbf{f}_j^n is the feature of the j -th node in \mathcal{G}_i at the n -th layer, d is the size of convolution kernel, and ω_j denotes the learned weight. The $N \times k$ feature map is successively aggregated into $N \times 1$, with the feature dimension being increased from 32 to 256 as shown in Fig. 4 (b). In contrast to [36], our convolutions take regular graphs as input instead of isolated correspondences, with reliably captured local features. Third, for global feature extraction, Instance Normalization [33] is used to normalize the feature map in each ResNet block, which has been proven more effective than average-pooling and max-pooling in [36].

Loss function. A simple yet effective cross-entropy loss function is used to calculate the deviation between the out-

puts and corresponding labels - *i.e.*,

$$L(\omega, \mathcal{C}) = \frac{1}{N} \sum_{i=1}^N \alpha_i H(y_i, S(g(c_i, \omega))), \quad (6)$$

where g is the output of NM-Net, S indicates the logistic function, y_i denotes the ground-truth label of c_i , H represents a binary cross entropy function, and α_i is a self-adaptive weight to balance positive and negative samples. The regression loss in [36] is not used because the ground-truth global transformation may be unavailable in some applications such as multi-consistency matching and non-rigid matching. As we will show next, our simpler form of loss function can achieve even better performance.

5. Experiments

This section includes extensive experimental evaluations on four standard datasets covering a variety of contexts - *i.e.*, narrow and wide baseline matching, matching for reconstruction (*i.e.*, structure from motion), and matching with multiple consistencies. We also present comprehensive comparisons with several state-of-the-art methods including both hand-crafted approaches (*i.e.*, RANSAC [9], GTM [1], and LPM [23]) and learning-based approaches (*i.e.*, PointNet [25], PointNet++ [26], and LGC-Net [36]).

5.1. Experimental Setup

Optimization and architecture details. The configuration of NM-Net (Fig. 4 (a)) is C(32, 1, 4)-GP-R(32, 1, 3)-R(32, 1, 3)-R(64, 1, 3)-R(64, 1, 3)-R(128, 1, 3)-R(128, 1, 3)-R(256, 1, 3)-R(256, 1, 3)-C(256, 1, 1)-C(1, 1, 1), where C(n, h, w) denotes a convolution layer with n output channels and a $h \times w$ convolution kernel, GP indicates the grouping module, and R(n, h, w) represents a ResNet block that includes two convolution layers C(n, h, w). Every convolution layer is followed by Instance Normalization, Batch Normalization, and ReLU activation, except for the last one. NM-Net is trained by Adam [8] with a learning rate being 10^{-3} and batch size being 16. For LGC-Net, we use the code released by the authors to train the model. For PointNet and PointNet++, we adopt the ResNet backbone following the accomplishment in [36]. To verify the effectiveness of compatibility-specific neighbor mining method, another version of NM-Net is also implemented (named NM-Net-sp), where compatibility-specific knn search is replaced by spatially knn search. Parameters including the number of neighbors k and λ (Eq. 4) are set to 8 and 10^{-3} , respectively.

Benchmark datasets. Four datasets are employed in our experiments - *i.e.*, NARROW, WIDE, COLMAP [30], and MULTI [37] (Table 1). The first two datasets are collected by us using a drone in four scenes, and we respectively keep a sample interval of 10 and 20 frames to attain narrow and wide baseline matching data. For NARROW,

Dataset	# Image pairs	# Training	# Validation	# Testing	Inlier ratio (%)	Challenges
NARROW	24070	16849	3610	3610	40.827	VP changes
WIDE	11426	7998	1713	1713	32.771	VP changes
COLMAP [30]	18850	13195	2827	2827	7.496	VP changes, rotation
MULTI [37]	45	-	-	-	40.828	Dynamic scenarios

Table 1. **Properties of the experimental datasets.** VP means viewpoint and the inlier ratio indicates the average proportion of inliers in initial correspondence sets computed over a whole dataset.

Method	Precision (%)	Recall (%)	F-measure (%)	MSE	MAE	Median	Max	Min (10^{-2})
RANSAC [9]	86.923	60.397	69.194	2.017	2.622	2.809	4.978	0.755
GTM [1]	88.707	52.949	65.653	2.042	2.728	2.886	4.968	2.467
LPM [23]	72.667	68.504	70.173	2.087	2.879	3.107	4.869	25.453
PointNet [25]	79.003	86.163	82.102	2.293	2.787	3.503	4.728	1.180
PointNet++ [26]	83.677	85.045	84.112	2.248	2.773	3.328	5.128	3.899
LGC-Net [36]	95.238	98.405	96.611	2.096	2.255	3.021	5.006	0.558
NM-Net-sp	96.946	97.659	97.283	2.482	2.664	3.687	5.038	0.245
NM-Net	97.169	97.870	97.501	2.436	2.608	3.630	5.021	0.390

Method	Precision (%)	Recall (%)	F-measure (%)	MSE	MAE	Median	Max	Min (10^{-2})
RANSAC [9]	80.740	51.198	60.350	2.052	2.711	3.125	5.040	1.347
GTM [1]	79.989	47.711	58.881	2.040	2.784	3.068	5.041	1.046
LPM [23]	62.940	64.487	62.828	2.038	2.921	3.080	4.926	19.782
PointNet [25]	64.730	77.287	70.068	2.282	2.863	3.458	4.905	5.528
PointNet++ [26]	73.926	81.856	77.245	2.180	2.771	3.255	5.013	3.020
LGC-Net [36]	88.139	97.138	91.264	2.059	2.230	2.995	5.061	1.226
NM-Net-sp	91.742	94.039	92.749	2.513	2.731	3.751	5.110	0.650
NM-Net	92.332	94.251	93.145	2.488	2.718	3.781	5.113	0.553

Method	Precision (%)	Recall (%)	F-measure (%)	MSE	MAE	Median	Max	Min (10^{-2})
RANSAC [9]	25.156	14.477	17.464	1.984	2.985	3.169	5.044	1.383
GTM [1]	22.931	19.913	19.075	2.004	3.073	3.245	5.102	4.804
LPM [23]	15.879	34.293	19.595	2.019	3.037	3.113	4.693	17.298
PointNet [25]	13.596	41.765	19.710	2.051	2.864	3.193	4.878	14.138
PointNet++ [26]	18.659	41.953	24.301	2.060	2.902	3.200	4.877	4.150
LGC-Net [36]	26.383	71.132	33.949	1.981	2.554	3.071	4.717	1.265
NM-Net-sp	29.296	59.710	37.503	1.983	2.446	3.047	5.125	2.250
NM-Net	31.003	58.499	38.887	1.953	2.402	3.027	4.989	1.514

Table 2. **Evaluation results.** The three tables from top to bottom are results on NARROW, WIDE, and COLMAP datasets, respectively. Precision, recall, and F-measure are colored for highlighting because they explicitly measure the performance of correspondence selection. NM-Net-sp indicates a variant of our NM-Net with spatial neighbors. The best result in each column is rendered in bold.

WIDE, and COLMAP, the ground-truth camera parameters are obtained by VisualSfM [35] and the ground-truth labels of correspondences are calculated by comparing the corresponding epipolar distances [12] with a threshold (10^{-4}). MULTI is a tiny dataset consisting of 45 image pairs with available ground-truth labels for each correspondence. We use this dataset to test the generalization in cases of multi-consistency matching.

Evaluation criteria. To measure the correspondence selection performance, we employ precision (P), recall (R),

and F-measure (F) as in [19, 3, 23]. Moreover, considering that accurate estimation of the global transformation is required in some image alignment and 3D reconstruction tasks [32, 2, 4], the deviation between the essential matrix \mathbf{E} estimated by selected correspondences and ground-truth \mathbf{E}_{gt} is measured by MSE, MAE, median, max, and min as in [28]. Since P, R, and F explicitly reflect the performance of correspondence selection, we will focus on analyzing results from the perspective of these metrics.

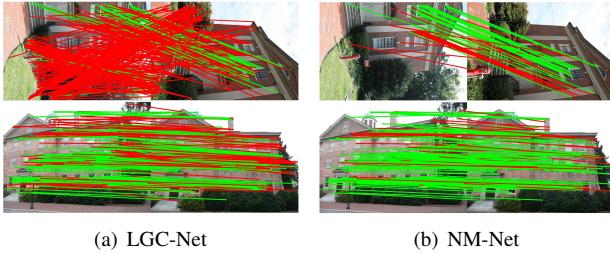


Figure 5. **Visual results on COLMAP dataset.** Green and red lines represent inliers and outliers in the selected correspondence set by LGC-Net and NM-Net, respectively.

5.2. Single Consistency

Finding a single consistency corresponds to a global transformation (*e.g.*, essential matrix) in static scenes is a popular application [32, 2] for feature correspondences. Our experimental results on NARROW, WIDE, and COLMAP datasets which contain a single consistency in each image pair are presented in Table 2.

As reported in Table 2 (a), (b), and (c), NM-Net significantly outperforms hand-crafted algorithms and other learning-based approaches in terms of F-measure. First, when compared with hand-crafted algorithms such as RANSAC, GTM (using the same binary item in Eq. 4), and LPM (employing spatially k -nearest information), NM-Net outperforms them by about 20 percentages on all three datasets. Second, NM-Net remarkably surpasses a representative set of learning-based approaches. For PointNet and LGC-Net, global features are extracted by average pooling and Context Normalization, respectively. For PointNet++, local information is added by spatially knn search for each correspondence. NM-Net also extracts both global features and local features but mines neighbors relying on the proposed compatibility metric of Eq. (4). The superiority of our framework can be easily verified. Third, NM-Net performs better than NM-Net-sp on all datasets; the gap becomes more dramatic on COLMAP which is a more challenging dataset with an extremely low initial inlier ratio (7.49%). It implies that our compatibility-specific knn search is more robust to high outlier ratios than standard spatially knn search. Some representative visual comparison results are presented in Fig. 5, in which NM-Net is compared against current state-of-the-art deep learning framework LGC-Net. More visual results can be found in the supplementary material.

5.3. Multiple Consistencies

Multi-consistency feature matching in the situation of dynamic scenarios remains an open research problem as mentioned in [37]. In contrast to a global transformation for static scenes, several local transformations corresponding to multiple consistencies are included into the initial

Method	P (%)	R (%)	F (%)
PointNet [25]	48.223	5.829	8.717
PointNet++ [26]	64.661	7.871	13.327
LGC-Net [36]	61.736	36.849	41.849
NM-Net	51.898	33.653	35.605

Table 3. **Generalization on MULTI dataset.** Tested models are pretrained on NARROW dataset. Metrics are precision (P), recall (R), and F-measure (F).

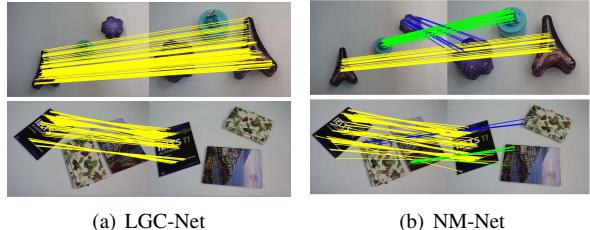


Figure 6. **Visual results on MULTI dataset.** Different colors represent different feature consistencies.

correspondence set. Because MULTI only contains 45 image pairs, models pretrained on NARROW that includes a similar inlier ratio (Table 1), are adopted to test the generalization from single consistency to multiple consistencies.

Table 3 and Fig. 6 show quantitative and qualitative results on MULTI dataset, respectively. Although LGC-Net achieves higher F-measure than NM-Net (still the second best), NM-Net is able to pick out all kinds of consistencies; while LGC-Net finds out only one kind of consistency. Note that LGC-Net is trained under the supervision of a hybrid loss that includes a regression loss corresponding to a global transformation, which explains why LGC-Net is less effective for multi-consistency matching with several local transformations. By contrast, NM-Net is trained with a classification loss that is insensitive to multiple consistencies.

5.4. Method Analysis

Parameter k . The number of neighbors k for each correspondence is a core parameter in NM-Net, which determines the receptive field of local features in each graph \mathcal{G} . Consequently, several versions of NM-Net with different numbers of neighbors are studied on NARROW, WIDE, and COLMAP. As shown in Fig. 7, NM-Net experiences performance degradation when k is either too small (4) or too large (32). When $k = 4$, the search space for each correspondence is too small to extract enough available local features; when $k = 32$, the consistency of neighbors in \mathcal{G}_{inlier} for an inlier c_{inlier} will decline and some outliers will be included undesirably as nuisance. Based on above analysis, we have used $k = 8$ as the default size of neighbors in NM-Net.

Learning effectiveness validation. Since the neighbors

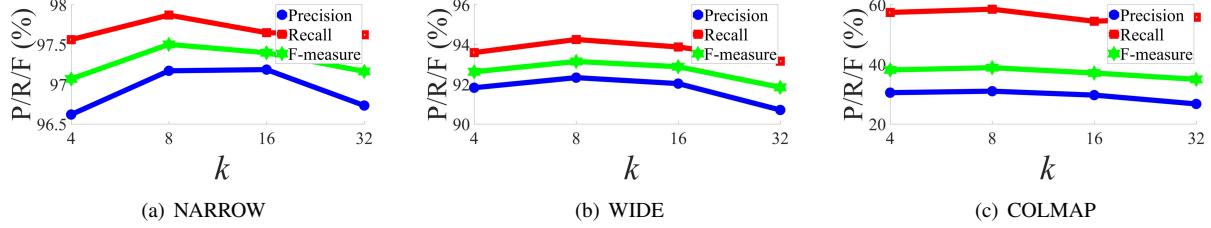


Figure 7. **Analysis of parameter k .** We train our NM-Net with different values of neighborhood size k ($\{4, 8, 16, 32\}$) while keep other settings identical, and examine the performance variation on NARROW, WIDE, and COLMAP datasets.

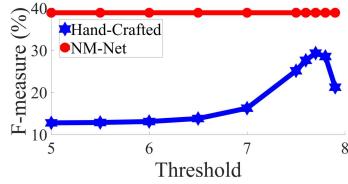


Figure 8. **Learning effectiveness validation.** The hand-crafted method judges the correctness of a correspondence by comparing the sum of compatibility scores of the neighbors with a threshold. This threshold is varied from 5 to 7.9 to alleviate the effect of threshold setting.

searched by the compatibility metric are more consistent for inliers than for outliers, compatibility scores of correspondences in \mathcal{G}_{inlier} should be higher than those in $\mathcal{G}_{outlier}$ in theory. A potential issue arises: can correspondences be directly classified based on the scores? To address this question, a hand-crafted approach is designed to calculate the sum of scores in each graph \mathcal{G}_i as s_i^{sum} . Then c_i will be determined as an inlier if s_i^{sum} is higher than a predefined threshold. A comparison between the hand-crafted approach and NM-Net on COLMAP dataset is included in Fig. 8. Clearly, NM-Net achieves a far higher F-measure than the hand-crafted one with all thresholds. Due to uncertainties such as viewpoint changes and camera rotation, the distribution of correspondences is distinctly different among various image pairs. Utilizing raw scores to distinguish inliers and outliers is therefore unreliable. In contrast to the hand-crafted approach, the scores are leveraged as indexes to search for compatible neighbors in our proposed method. The local features hidden in these neighbors can be fully explored by a powerful deep learning network.

Compatibility metric analysis. Since compatibility-specific knn search is a key procedure in NM-Net to generate the initial graph \mathcal{G} , the consistency of elements in \mathcal{G} directly affects the effectiveness of raw information. To shed more light to this important matter, an analysis of the compatibility metric (Eq. 4) is conducted as follows.

The inlier ratios of the neighbors searched by Eq. 4 on NARROW, WIDE, and COLMAP datasets are shown in Fig. 9. The compatibility metric is deemed to be reason-

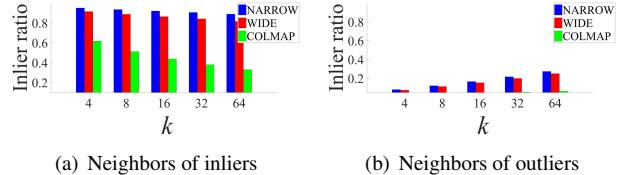


Figure 9. **Compatibility metric analysis.** The inlier ratios of neighbors recognized via our compatibility metric (Eq. 4) of (a) inliers and (b) outliers are calculated on NARROW, WIDE, and COLMAP datasets to examine if this metric can provide distinguishable local information.

able for the following reasons. First, the neighbors of inliers are significantly more consistent than the ones of outliers, with the inlier ratios being 3 times higher. Second, our approach achieves approximate 100% inlier ratios on both NARROW and WIDE datasets as shown in Fig. 9 (a), where almost all searched neighbors of inliers are consistent (*i.e.*, inliers). However, the inlier ratios of the neighbors of correct matches drop considerably on the more challenging COLMAP dataset. This result suggests that there is still a large room for further improvement from the perspective of robustness of the compatibility metric. We leave this to our future study.

6. Conclusion

We have presented a hierarchical classification network named NM-Net to select correct matches from initial correspondences, which fully mines compatibility-specific locality for each correspondence. Experiments demonstrate that NM-Net behaves favorably to the state-of-the-art (both hand-crafted and learning-based) approaches. The current shortcoming of our approach is the demand for key-point detectors with local affine information to compute the compatibility score. We expect developing a more advanced compatibility metric without such constraint in our future works.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 61876211 and by the 111 Project on Computational Intelligence and Intelligent Control under Grant B18024.

References

- [1] Andrea Albarelli, Emanuele Rodolà, and Andrea Torsello. Robust game-theoretic inlier selection for bundle adjustment. In *International Symposium on 3D Data Processing, Visualization and Transmission*, 2010.
- [2] Selim Benhimane and Ezio Malis. Real-time image-based tracking of planes using efficient second-order minimization. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, pages 943–948, 2004.
- [3] JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan Dat Nguyen, and Ming-Ming Cheng. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [4] Matthew Brown and David G Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007.
- [5] H. Y. Chen, Y. Y. Lin, and B. Y. Chen. Co-segmentation guided hough transform for robust feature matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2388, 2015.
- [6] Ondřej Chum and Jiří Matas. Matching with prosac—progressive sample consensus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 220–226, 2005.
- [7] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized ransac. *Pattern Recognition*, pages 236–243, 2003.
- [8] D.P.Kingma and J.Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014, 2014.
- [9] 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.
- [10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.
- [11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.
- [12] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.
- [13] 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.
- [14] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008.
- [15] Mikolajczyk K and Schmid C. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.
- [16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.
- [17] Marius Leordeanu and Martial Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1482–1489, 2005.
- [18] Xiangru Li and Zhanyi Hu. Rejecting mismatches by correspondence function. *International Journal of Computer Vision*, 89(1):1–17, 2010.
- [19] Wen-Yan Daniel Lin, Ming-Ming Cheng, Jiangbo Lu, Hongsheng Yang, Minh N Do, and Philip Torr. Bilateral functions for global motion modeling. In *Proceedings of the European Conference on Computer Vision*, pages 341–356. Springer, 2014.
- [20] Hairong Liu and Shuicheng Yan. Common visual pattern discovery via spatially coherent correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1609–1616, 2010.
- [21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.
- [22] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [23] Jiayi Ma, Ji Zhao, Hanqi Guo, Junjun Jiang, Huabing Zhou, and Yuan Gao. Locality preserving matching. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4492–4498, 2017.
- [24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [25] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 4, 2017.
- [26] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *International Conference on Neural Information Processing Systems*, pages 5099–5108, 2017.
- [27] R Raguram, O Chum, M Pollefeys, J Matas, and J. M. Frahm. Usac: A universal framework for random sample consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):2022–2038, 2013.
- [28] René Ranftl and Vladlen Koltun. Deep fundamental matrix estimation. In *Proceedings of the European Conference on Computer Vision*, pages 284–299, 2018.
- [29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *International Conference on Neural Information Processing Systems*, pages 91–99, 2015.
- [30] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, 2016.

- [31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [32] Noah Snavely, Steven M Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, 2008.
- [33] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4105–4113, 2017.
- [34] Jörgen W Weibull. *Evolutionary game theory*. MIT press, 1997.
- [35] Changchang Wu. Towards linear-time incremental structure from motion. In *International Conference on 3D Vision*, pages 127–134, 2013.
- [36] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [37] Chen Zhao, Jiaqi Yang, Yang Xiao, and Zhiguo Cao. Scalable multi-consistency feature matching with non-cooperative games. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1258–1262. IEEE, 2018.