Deep Fundamental Matrix Estimation without Correspondences

Omid Poursaeed[*1,2], Guandao Yang[*1], Aditya Prakash[*3], Qiuren Fang[1], Hanqing Jiang[1], Bharath Hariharan[1], and Serge Belongie[1,2]

[1] Cornell University
[2] Cornell Tech
[3] Indian Institute of Technology Roorkee

Abstract. Estimating fundamental matrices is a classic problem in computer vision. Traditional methods rely heavily on the correctness of estimated key-point correspondences, which can be noisy and unreliable. As a result, it is difficult for these methods to handle image pairs with large occlusion or significantly different camera poses. In this paper, we propose novel neural network architectures to estimate fundamental matrices in an end-to-end manner without relying on point correspondences. New modules and layers are introduced in order to preserve mathematical properties of the fundamental matrix as a homogeneous rank-2 matrix with seven degrees of freedom. We analyze performance of the proposed models using various metrics on the KITTI dataset, and show that they achieve competitive performance with traditional methods without the need for extracting correspondences.

Keywords: Fundamental Matrix · Epipolar Geometry · Deep Learning · Stereo.

The Fundamental matrix (F-matrix) contains rich information relating two stereo images. The ability to estimate fundamental matrices is essential for many computer vision applications such as camera calibration and localization, image rectification, depth estimation and 3D reconstruction. The current approach to this problem is based on detecting and matching local feature points, and using the obtained correspondences to compute the fundamental matrix by solving an optimization problem about the epipolar constraints [27, 16]. The performance of such methods is highly dependent on the accuracy of the local feature matches, which are based on algorithms such as SIFT [28]. However, these methods are not always reliable, especially when there is occlusion, large translation or rotation between images of the scene.

In this paper, we propose end-to-end trainable convolutional neural networks for F-matrix estimation that do not rely on key-point correspondences. The main challenge of directly regressing the entries of the F-matrix is to preserve its mathematical properties as a homogeneous rank-2 matrix with seven degrees of freedom. We propose a reconstruction module and a normalization layer (Sec. 2.2) to address this challenge. We demonstrate that by using these layers, we can accurately estimate the fundamental matrix, while a simple regression approach does not yield good results. Our detailed

* Indicates equal contribution

network architectures are presented in Sec. 2. Empirical experiments are performed on the KITTI dataset [13] in Sec. 3. The results indicate that we can achieve competitive results with traditional methods without relying on correspondences.

1 Background and Related Work

1.1 Fundamental Matrix and Epipolar Geometry

When two cameras view the same 3D scene from different viewpoints, geometric relations among the 3D points and their projections onto the 2D plane lead to constraints on the image points. This intrinsic projective geometry is referred to as the epipolar geometry, and is encapsulated by the fundamental matrix $\mathbf{F}$. This matrix only depends on the cameras' internal parameters and their relative pose, and can be computed as:

$$\mathbf{F} = \mathbf{K_2}^{-T}[\mathbf{t}]_{\times}\mathbf{R}\mathbf{K_1}^{-1} \tag{1}$$

where $\mathbf{K_1}$ and $\mathbf{K_2}$ represent camera intrinsics, and $\mathbf{R}$ and $[\mathbf{t}]_{\times}$ are the relative camera rotation and translation respectively [16]. More specifically:

$$\mathbf{K}_i = \begin{bmatrix} f_i^{-1} & 0 & c_x \\ 0 & f_i^{-1} & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

$$\mathbf{t}_{\times} = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \tag{3}$$

$$\mathbf{R} = \mathbf{R_x}(r_x)\mathbf{R_y}(r_y)\mathbf{R_z}(r_z) \tag{4}$$

in which $(c_x, c_y)^T$ is the principal point of the camera, f_i is the focal length of camera $i = 1, 2$, and t_x, t_y and t_z are the relative displacements along the x, y and z axes respectively. $\mathbf{R}$ is the rotation matrix which can be decomposed into rotations along x, y and z axes. We assume that the principal point is in the middle of the image plane.

While the fundamental matrix is independent of the scene structure, it can be computed from correspondences of projected scene points alone, without requiring knowledge of the cameras' internal parameters or relative pose. If p and q are matching points in two stereo images, the fundamental matrix $\mathbf{F}$ satisfies the equation:

$$q^T\mathbf{F}p = 0 \tag{5}$$

Writing $p = (x, y, 1)^T$ and $q = (x', y', 1)^T$ and $\mathbf{F} = [f_{ij}]$, equation 5 can be written as:

$$x'xf_{11} + x'yf_{12} + x'f_{13} + y'xf_{21} + y'yf_{22} + y'f_{23} + xf_{31} + yf_{32} + f_{33} = 0. \tag{6}$$

Let $\mathbf{f}$ represent the 9-vector made up of the entries of $\mathbf{F}$. Then equation 6 can be written as:

$$(x'x, x'y, x', y'x, y'y, y', x, y, 1)\mathbf{f} = 0 \tag{7}$$

A set of linear equations can be obtained from n point correspondences:

$$\mathbf{A}\mathbf{f} = \begin{bmatrix} x'_1x_1 & x'_1y_1 & x'_1 & y'_1x_1 & y'_1y_1 & y'_1 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_nx_n & x'_ny_n & x'_n & y'_nx_n & y'_ny_n & y'_n & x_n & y_n & 1 \end{bmatrix} \mathbf{f} = 0 \tag{8}$$

Various methods have been proposed for estimating fundamental matrices based on equation 8. The simplest method is the eight-point algorithm which was proposed by Longuet-Higgins [27]. Using (at least) 8 point correspondences, it computes a (least-squares) solution to equation 8. It enforces the rank-2 constraint using Singular Value Decomposition (SVD), and finds a matrix with the minimum Frobenius distance to the computed (rank-3) solution. Hartley [17] proposed a normalized version of the eight-point algorithm which achieves improved results and better stability. The algorithm involves translation and scaling of the points in the image before formulating the linear equation 8.

The Algebraic Minimization algorithm uses a different procedure for enforcing the rank-2 constraint. It tries to minimize the algebraic error $\mathbf{A}\|\mathbf{f}\|$ subject to $\|\mathbf{f}\| = 1$. It uses the fact that we can write the singular fundamental matrix as $\mathbf{F} = \mathbf{M}[e]_\times$ where $\mathbf{M}$ is a non-singular matrix and $[e]_\times$ is a skew-symmetric matrix with e corresponding to the epipole in the first image. This equation can be written as $\mathbf{f} = E\mathbf{m}$, where $\mathbf{f}$ and $\mathbf{m}$ are vectors comprised of entries of $\mathbf{F}$ and $\mathbf{M}$, and E is a 9×9 matrix comprised of elements of $[e]_\times$. Then the minimization problem becomes:

$$\text{minimize } \|\mathbf{A}E\mathbf{m}\| \text{ subject to } \|E\mathbf{m}\| = 1 \tag{9}$$

To solve this optimization problem, we can start from an initial estimate of $\mathbf{F}$ and set e as the generator of the right null space of $\mathbf{F}$. Then we can iteratively update e and $\mathbf{F}$ to minimize the algebraic error. More details are given in [16].

The Gold Standard geometric algorithm assumes that the noise in image point measurements obeys a Gaussian distribution. It tries to find the Maximum Likelihood estimate of the fundamental matrix which minimizes the geometric distance

$$\sum_i d(p_i, \hat{p}_i)^2 + d(q_i, \hat{q}_i)^2 \tag{10}$$

in which p_i and q_i are true correspondences satisfying equation 5, and $\hat{p}_i$ and $\hat{q}_i$ are the estimated correspondences.

Another algorithm uses RANSAC [11] to compute the fundamental matrix. It computes interest points in each image, and finds correspondences based on proximity and similarity of their intensity neighborhood. In each iteration, it randomly samples 7 correspondences and computes the F-matrix based on them. It then calculates the re-projection error for each correspondence, and counts the number of inliers for which the error is less than a specified threshold. After sufficient number of iterations, it chooses the F-matrix with the largest number of inliers. A generalization of RANSAC is MLESAC [40], which adopts the same sampling strategy as RANSAC to generate putative solutions, but chooses the solution that maximizes the likelihood rather than

just the number of inliers. MAPSAC [39] (Maximum A Posteriori SAmple Consensus) improves MLESAC by being more robust against noise and outliers including Bayesian probabilities in minimization. A global search genetic algorithm combined with a local search hill climbing algorithm is proposed in [45] to optimize MAPSAC algorithm for estimating fundamental matrices. [42] proposes an algorithm to cope with the problem of fundamental matrix estimation for binocular vision system used in wild field. It first acquires the edge points using Canny edge detector, and then gets the pre-matched points by the GMM-based point set registration algorithm. It then computes the fundamental matrix using the RANSAC algorithm. [10] proposes to use adaptive penalty methods for valid estimation of Essential matrices as a product of translation and rotation matrices. A new technique for calculating the fundamental matrix combined with feature lines is introduced in [49]. The interested reader is referred to [1] for a survey of various methods for estimating the F-matrix.

1.2 Deep Learning for Multi-view Geometry

Deep neural networks have achieved state-of-the-art performance on tasks such as image recognition [24, 18, 38, 37], semantic segmentation [26, 3, 43, 47], object detection [14, 35, 34], scene understanding [23, 48, 32] and generative modeling [15, 33, 19, 44, 31] in the last few years. Recently, there has been a surge of interest in using deep learning for classic geometric problems in Computer Vision. A method for estimating relative camera pose using convolutional neural networks is presented in [29]. It uses a simple convolutional network with spatial pyramid pooling and fully connected layers to compute the relative rotation and translation of the camera. An approach for camera re-localization is presented in [25] which localizes a given query image by using a convolutional neural network for first retrieving similar database images and then predicting the relative pose between the query and the database images with known poses. The camera location for the query image is obtained via triangulation from two relative translation estimates using a RANSAC-based approach. [41] uses a deep convolutional neural network to directly estimate the focal length of the camera using only raw pixel intensities as input features. [2] proposes two strategies for differentiating the RANSAC algorithm: using a soft argmax operator, and probabilistic selection. [12] leverages deep neural networks for 6-DOF tracking of rigid objects.

[5] presents a deep convolutional neural network for estimating the relative homography between a pair of images. A more complicated algorithm is proposed in [8] which contains a hierarchy of twin convolutional regression networks to estimate the homography between a pair of images. [7] introduces two deep convolutional neural networks, MagicPoint and MagicWarp. MagicPoint extracts salient 2D points from a single image. MagicWarp operates on pairs of point images (outputs of MagicPoint), and estimates the homography that relates the inputs. [30] proposes an unsupervised learning algorithm that trains a deep convolutional neural network to estimate planar homographies. A self-supervised framework for training interest point detectors and descriptors is presented in [6]. A convolutional neural network architecture for geometric matching is proposed in [36]. It uses feature extraction networks with shared weights and a matching network which matches the descriptors. The output of the matching network is passed through a regression network which outputs the parameters of the

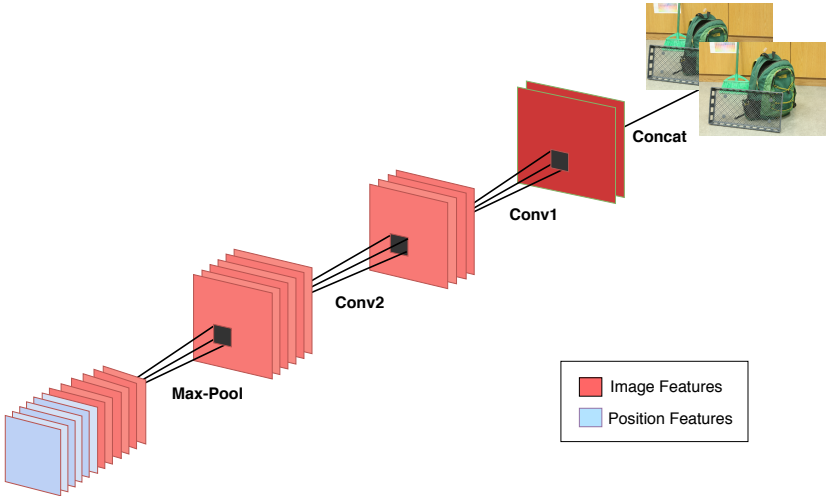

Fig. 1. Single-Stream Architecture. Stereo images are concatenated and passed to a convolutional neural network. Position features can be used to indicate where the final activations come from with respect to the full-size image.

geometric transformation. [22] presents a model which takes a set of images and their corresponding camera parameters as input and directly infers the 3D model.

2 Network Architecture

We leverage deep neural networks for estimating the fundamental matrix directly from a pair of stereo images. Each network consists of a feature extractor to obtain features from the images and a regression network to compute the entries of the F-matrix from the features.

2.1 Feature Extraction

We consider two different architectures for feature extraction. In the first architecture, we concatenate the images across the channel dimension, and pass the result to a neural network to extract features. Figure 1 illustrates the network structure. We use two convolutional layers, each followed by ReLU and Batch Normalization [20]. We use 128 filters of size 3×3 in the first convolutional layer and 128 filters of size 1×1 in the second layer. We limit the number of pooling layers to one in order not to lose the spatial structure in the images.

Location Aware Pooling. As discussed in Sec. 1, the F-matrix is highly dependent on the relative location of corresponding points in the images. However, down-sampling layers such as Max Pooling discard the location information. In order to retain this information, we keep all the indices of where the activations come from in the max-pooling layers. At the end of the network, we append the position of final features with

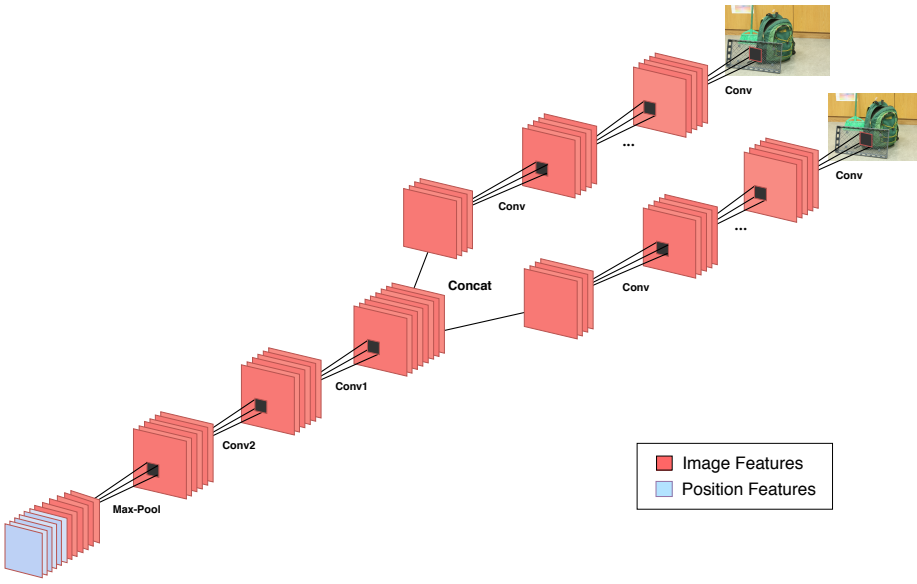

Fig. 2. Siamese Architecture. Images are first passed to two streams with shared weights. The resulting features are concatenated and passed to the single-stream network as in figure 1. Position features can be used with respect to the concatenated features.

respect to the full-size image. Each location is indexed with an integer in $[1, h \times w \times c]$ normalized to be within the range $[0, 1]$, in which h, w and c are the height, width and channel dimensions of the image respectively. In this way, each feature has a position index indicating from where it comes from. This helps the network to retain the location information and to provide more accurate estimates of the F-matrix.

The second architecture is shown in figure 2. We first process each of the input images in a separate stream using an architecture similar to the Universal Correspondence Network (UCN) [4]. Unlike the UCN architecture, we do not use Spatial Transformers [21] in these streams since they can remove part of the information needed for estimating relative camera rotation and translation. The resulting features from these streams are then concatenated, and passed to a single-stream network similar to figure 1. We can use position features in the single-stream network as discussed previously. These features capture the position of final features the with respect to the concatenated features at the end of the two streams. We refer to this architecture as 'Siamese'. As we show in Sec. 3, this network outperforms the Single-Stream one. We also consider using only the UCN without the single-stream network. The results, however, are not competitive with the Siamese architecture.

2.2 Regression

A simple approach for computing the fundamental matrix from the features is to pass them to fully-connected layers, and directly regress the nine entries of the F-Matrix. We

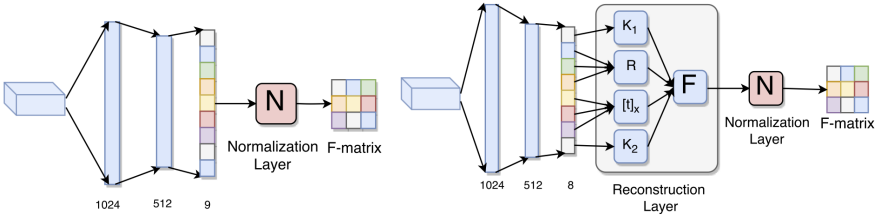

Fig. 3. Different regression methods for predicting F-matrix entries from the features. The architecture to directly regress the entries of the F-matrix is shown on the left. The network with the reconstruction and normalization layers is shown on the right, and is able to estimate homogeneous F-matrices with rank two and seven degrees of freedom.

can then normalize the result to achieve scale-invariance. This approach is shown in figure 3 (left). The main issue with this approach is that the predicted matrix might not satisfy all the mathematical properties required for a fundamental matrix as a rank-2 matrix with seven degrees of freedom. In order to address this issue, we introduce Reconstruction and Normalization layers in the following.

F-matrix Reconstruction Layer. We consider equation 1 to reconstruct the fundamental matrix:

$$\hat{\mathbf{F}} = \mathbf{K_2}^{-T}[\mathbf{t}]_{\times}\mathbf{R}\mathbf{K_1}^{-1} \tag{11}$$

we need to determine eight parameters $(f_1, f_2, t_x, t_y, t_z, r_x, r_y, r_z)$ as shown in equations (2–4). Note that the predicted $\hat{\mathbf{F}}$ is differentiable with respect to these parameters. Hence, we can construct a layer that takes these parameters as input, and outputs a fundamental matrix $\hat{\mathbf{F}}$. This approach guarantees that the reconstructed matrix has rank two. Figure 3 (right) illustrates the Reconstruction layer.

Normalization Layer. Considering that the F-matrix is scale-invariant, we also use a Normalization layer to remove another degree of freedom for scaling. In this way, the estimated F-matrix will have seven degrees of freedom and rank two as desired. The common practice for normalization is to divide the F-matrix by its last entry. We call this method **ETR-Norm**. However, since the last entry of the F-matrix could be close to zero, this can result in large entries, and training can become unstable. Therefore, we propose two alternative normalization methods.

FBN-Norm: We divide all entries of the F-matrix by its Frobenius norm, so that all the matrices live on a 9-sphere of unit norm. Let $\|\mathbf{F}\|_F$ denote the Frobenius norm of matrix $\mathbf{F}$. Then the normalized fundamental matrix is:

$$\mathcal{N}_{FBN}(\mathbf{F}) = \|\mathbf{F}\|_F^{-1}\mathbf{F} \tag{12}$$

ABS-Norm: We divide all entries of the F-matrix by its maximum absolute value, so that all entries are restricted within $[-1, 1]$ range:

$$\mathcal{N}_{ABS}(\mathbf{F}) = (\max_{i,j} |\mathbf{F}_{i,j}|)^{-1}\mathbf{F} \tag{13}$$

During training, the normalized F-matrices are compared with the ground-truth using both L_1 and L_2 losses. We provide empirical results to study how each of these normalization methods influences performance and stability of training in Sec. 3.

Epipolar Parametrization Given that the F-matrix has a rank of two, an alternative parametrization is specifying the first two columns $\mathbf{f}_1$ and $\mathbf{f}_2$ and the coefficients α and β such that $\mathbf{f}_3 = \alpha\mathbf{f}_1 + \beta\mathbf{f}_2$. Normalization layer can still be used to achieve scale-invariance. The coordinates of the epipole occur explicitly in this parametrization: $(\alpha, \beta, 1)^T$ is the right epipole for the F-matrix [16]. The corresponding regression architecture is similar to figure 3, but we interpret the final eight values differently: the first six elements represent the first two columns and the last two represent the coefficient for combining the columns. The main disadvantage of this method is that it does not work when the first two columns of $\mathbf{F}$ are linearly dependent. In this case, it is not possible to write the third column in terms of the first two columns.

3 Experiments

To evaluate whether our models can successfully learn F-matrices, we train models with various configurations and compare their performance based on the metrics defined in Sec. 3.1. The baseline model (**Base**) uses neither position features nor the reconstruction module. The **POS** model utilizes the position features on top of the **Base** model. Epipolar parametrization (Sec. 2.2) is used for the **EPI** model. **EPI+POS** uses the position features with epipolar parametrization. The **REC** model is the same as **Base** but uses the reconstruction module. Finally, the **REC+POS** model uses both the position features and the reconstruction module.

We use the KITTI dataset for training our models. The dataset has been recorded from a moving platform while driving in and around Karlsruhe, Germany. We use 2000 images from the raw stereo data in the 'City' category, and split them into 1600 train, 200 validation and 200 test images. Ground truth F-matrices are obtained using the ground-truth camera parameters. The same normalization methods are used for both the estimated and the ground truth F-matrices. The feature extractor and the regression network are trained jointly in an end-to-end manner.

3.1 Evaluation Metrics

We use the following metrics to measure how well the F-matrix satisfies the epipolar constraint (equation 5) according to the held out correspondences:

EPI-ABS (Epipolar Constraint with Absolute Value):

$$\mathcal{M}_{EPI-ABS}(\mathbf{F}, p, q) = \sum_i |q_i^T \mathbf{F} p_i| \tag{14}$$

EPI-SQR (Epipolar Constraint with Squared Value):

$$\mathcal{M}_{EPI-SQR}(\mathbf{F}, p, q) = \sum_i (q_i^T \mathbf{F} p_i)^2 \tag{15}$$

	Siamese Network			Single-stream Network		
Normalization	Models	EPI-ABS	EPI-SQR	Models	EPI-ABS	EPI-SQR
ETR-Norm	Base	3.77	27.16	Base	4.43	34.34
	POS	4.05	21.90	POS	2.47	9.79
	EPI	0.52	0.28	EPI	1.00	0.99
	EPI + POS	0.88	1.02	EPI + POS	1.00	1.00
	REC	0.56	0.45	REC	0.99	0.99
	REC + POS	0.97	0.98	REC + POS	1.00	0.99
	7-point	1.91	152.83	7-point	1.91	152.83
	LeMedS	1.09	25.50	LeMedS	1.09	25.50
	RANSAC	0.60	3.85	RANSAC	0.60	3.85
	Ground-truth	0.05	0.004	Ground-truth	0.05	0.004
FBN-Norm	Base	1.44	2.58	Base	2.45	9.99
	POS	1.97	5.66	POS	2.78	8.55
	EPI	0.07	0.01	EPI	0.91	0.91
	EPI + POS	0.06	0.005	EPI + POS	0.67	0.58
	REC	0.92	1.11	REC	0.78	1.24
	REC + POS	0.43	0.44	REC + POS	0.87	0.81
	7-point	1.06	11.7	7-point	1.06	11.7
	LeMedS	0.39	0.68	LeMedS	0.39	0.68
	RANSAC	0.27	0.21	RANSAC	0.27	0.21
	Ground-truth	0.05	0.004	Ground-truth	0.05	0.004
ABS-Norm	Base	4.76	30.63	Base	3.55	18.04
	POS	3.74	22.59	POS	2.87	10.4
	EPI	0.18	0.06	EPI	0.92	1.94
	EPI + POS	0.12	0.03	EPI + POS	0.82	0.77
	REC	0.22	0.06	REC	0.77	0.99
	REC + POS	0.28	0.10	REC + POS	0.87	0.81
	7-point	1.17	15.4	7-point	1.17	15.4
	LeMedS	0.72	3.88	LeMedS	0.72	3.88
	RANSAC	0.33	0.39	RANSAC	0.33	0.39
	Ground-truth	0.05	0.004	Ground-truth	0.05	0.004

Table 1. Results for Siamese and Single-stream networks on the KITTI dataset. Traditional methods such as 8-point, LeMedS and RANSAC are compared with different variants of our proposed model. Various normalization methods and evaluation metrics are considered.

The first metric is equivalent to the Algebraic Distance mentioned in [9]. We evaluate the metrics based on high-confidence key-point correspondences: we select the key-points for which the Symmetric Epipolar Distance based on the ground-truth F-matrix is less than 2 [16]. This ensures that the point is no more than one pixel away from the corresponding epipolar line.

4 Results and Discussion

Results are shown in Table 1. We compare our method with 8-point, LeMedS and RANSAC algorithms [46]. On average, 60 pairs of keypoints are used per image. As we can observe, the reconstruction module is highly effective, and without it the network is unable to recover accurate fundamental matrices. The position features are also helpful in decreasing the error. The Siamese network outperforms the Single-Stream architecture, and can achieve errors comparable to the ground truth. This shows that the two streams used to process each of the input images are indeed useful. Note that the networks are trained end-to-end without the need for extracting point correspondences between the images, yet they are able to achieve competitive results with classic algorithms. The epipolar parametrization generally outperforms the other methods. During the inference time, we just need to pass the images to the feature extraction and regression networks to estimate the fundamental matrices.

5 Conclusion and Future Work

We present novel deep neural networks for estimating fundamental matrices from a pair of stereo images. Our networks can be trained end-to-end without the need for extracting point correspondences. We consider two different network architectures for computing features from the images, and show that the best result is obtained when we first process images in two streams, and then concatenate the features and pass the result to a single-stream network. We show that the simple approach of directly regressing the nine entries of the fundamental matrix does not yield good results. Therefore, a reconstruction module is introduced as a differentiable layer to estimate the parameters of the fundamental matrix. Two different parametrizations of the F-matrix are considered: one based on the camera parameters, and the other based on the epipolar parametrization. We also demonstrate that position features can be used to further improve the estimation. This is due to the sensitivity of fundamental matrices to the location of points in the input images. In the future, we plan to extend the results to other datasets, and explore other parametrizations of the fundamental matrix.

References

1. Armangué, X., Salvi, J.: Overall view regarding fundamental matrix estimation. Image and vision computing **21**(2), 205–220 (2003)
2. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: Dsac-differentiable ransac for camera localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 3 (2017)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2018)
4. Choy, C.B., Gwak, J., Savarese, S., Chandraker, M.: Universal correspondence network. In: Advances in Neural Information Processing Systems. pp. 2414–2422 (2016)
5. DeTone, D., Malisiewicz, T., Rabinovich, A.: Deep image homography estimation. arXiv preprint arXiv:1606.03798 (2016)

6. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. arXiv preprint arXiv:1712.07629 (2017)
7. DeTone, D., Malisiewicz, T., Rabinovich, A.: Toward geometric deep slam. arXiv preprint arXiv:1707.07410 (2017)
8. Erlik Nowruzi, F., Laganiere, R., Japkowicz, N.: Homography estimation from image pairs with hierarchical convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 913–920 (2017)
9. Fathy, M.E., Hussein, A.S., Tolba, M.F.: Fundamental matrix estimation: A study of error criteria. Pattern Recognition Letters **32**(2), 383–391 (2011)
10. Fathy, M.E., Rotkowitz, M.C.: Essential matrix estimation using adaptive penalty formulations. J. Comput. Vision **74**(2), 117–136 (2007)
11. Fischler, M.A., Bolles, R.C.: 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)
12. Garon, M., Lalonde, J.F.: Deep 6-dof tracking. IEEE transactions on visualization and computer graphics **23**(11), 2410–2418 (2017)
13. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
16. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
17. Hartley, R.I.: In defense of the eight-point algorithm. IEEE Transactions on pattern analysis and machine intelligence **19**(6), 580–593 (1997)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., Belongie, S.: Stacked generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2, p. 4 (2017)
20. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
21. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)
22. Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L.: Surfacenet: An end-to-end 3d neural network for multiview stereopsis. arXiv preprint arXiv:1708.01749 (2017)
23. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680 (2015)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks pp. 1097–1105 (2012)
25. Laskar, Z., Melekhov, I., Kalia, S., Kannala, J.: Camera relocalization by computing pairwise relative poses using convolutional neural network. arXiv preprint arXiv:1707.09733 (2017)
26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

27. Longuet-Higgins, H.C.: A computer algorithm for reconstructing a scene from two projections. Nature **293**(5828), 133–135 (1981)
28. Lowe, D.G.: Object recognition from local scale-invariant features. In: Computer vision, 1999. The proceedings of the seventh IEEE international conference on. vol. 2, pp. 1150–1157. Ieee (1999)
29. Melekhov, I., Ylioinas, J., Kannala, J., Rahtu, E.: Relative camera pose estimation using convolutional neural networks. In: International Conference on Advanced Concepts for Intelligent Vision Systems. pp. 675–687. Springer (2017)
30. Nguyen, T., Chen, S.W., Skandan, S., Taylor, C.J., Kumar, V.: Unsupervised deep homography: A fast and robust homography estimation model. IEEE Robotics and Automation Letters (2018)
31. Poursaeed, O., Katsman, I., Gao, B., Belongie, S.: Generative adversarial perturbations. arXiv preprint arXiv:1712.02328 (2017)
32. Poursaeed, O., Matera, T., Belongie, S.: Vision-based real estate price estimation. arXiv preprint arXiv:1707.05489 (2017)
33. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
34. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
35. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
36. Rocco, I., Arandjelovic, R., Sivic, J.: Convolutional neural network architecture for geometric matching. In: Proc. CVPR. vol. 2 (2017)
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014)
38. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
39. Torr, P.H.S.: Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. International Journal of Computer Vision **50**(1), 35–61 (2002)
40. Torr, P.H., Zisserman, A.: Mlesac: A new robust estimator with application to estimating image geometry. Computer vision and image understanding **78**(1), 138–156 (2000)
41. Workman, S., Greenwell, C., Zhai, M., Baltenberger, R., Jacobs, N.: Deepfocal: a method for direct focal length estimation. In: Image Processing (ICIP), 2015 IEEE International Conference on. pp. 1369–1373. IEEE (2015)
42. Yan, N., Wang, X., Liu, F.: Fundamental matrix estimation for binocular vision measuring system used in wild field. In: International Symposium on Optoelectronic Technology and Application 2014: Image Processing and Pattern Recognition. vol. 9301, p. 93010S. International Society for Optics and Photonics (2014)
43. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
44. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: IEEE Int. Conf. Comput. Vision (ICCV). pp. 5907–5915 (2017)
45. Zhang, Y., Zhang, L., Sun, C., Zhang, G.: Fundamental matrix estimation based on improved genetic algorithm. In: Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2016 8th International Conference on. vol. 1, pp. 326–329. IEEE (2016)
46. Zhang, Z.: Determining the epipolar geometry and its uncertainty: A review. International journal of computer vision **27**(2), 161–195 (1998)

47. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2881–2890 (2017)
48. Zhou, B., Khosla, A., Lapedriza, A., Torralba, A., Oliva, A.: Places: An image database for deep scene understanding. arXiv preprint arXiv:1610.02055 (2016)
49. Zhou, F., Zhong, C., Zheng, Q.: Method for fundamental matrix estimation combined with feature lines. Neurocomputing **160**, 300–307 (2015)