Learning Data-Adaptive Interest Points through Epipolar Adaptation

Guandao Yang[1]
[1]Cornell University

Tomasz Malisiewicz[2]
[2]Magic Leap

Serge Belongie[1,3]
[3]Cornell Tech

Abstract

Interest point detection and description have been cornerstones of many computer vision applications. Handcrafted methods like SIFT and ORB focus on generic interest points and do not lend themselves to data-driven adaptation. Recent deep learning models are generally either supervised using expensive 3D information or with synthetic 2D transformations such as homographies that lead to improper handling of nuisance features such as occlusion junctions. In this paper, we propose an alternative form of supervision that leverages the epipolar constraint associated with the fundamental matrix. This approach brings useful 3D information to bear without requiring full depth estimation of all points in the scene. Our proposed approach, Epipolar Adaptation, fine-tunes both the interest point detector and descriptor using a supervision signal provided by the epipolar constraint. We show that our method can improve upon the baseline in a target dataset annotated with epipolar constraints, and the epipolar adapted models learn to remove correspondence involving occlusion junctions correctly.

1. Introduction

Interest point detection and description are important building blocks for many computer vision tasks including SLAM [6] and tracking [7]. Classic methods such as SIFT and ORB are based on heuristics [13, 19, 1]. Thus, they lack the ability to adapt to new data or to address known pathological cases. When deploying an interest point detector or descriptor in real-world applications in robotics or mixed reality, inputs are environment-dependent. For example, being able to detect interest points that are well suited to animal tracking are unlikely to serve the purpose of a Mars rover. As a result, it is desirable to make the interest point detector and descriptor data-driven or learning-based.

Several recent works seek such adaptivity through use of deep neural networks [2, 25, 5, 16]. These approaches generally rely on 3D information, such as depth [16] or outputs from SfM [25], to provide ground truth point-to-point correspondences for supervision. Such 3D information, however, is usually difficult to obtain. Obtaining depths for a monocular scan, for example, requires the use of a special device such as Kinect sensor during the data collection process. Prior works have also generated synthetic datasets that are capable of providing such ground truth interest points and correspondences [6, 5], yet creating such synthetic dataset is laborious and requires sophisticated domain knowledge. SuperPoint circumvents this challenge by using self-supervised methods such as homographic warping to create synthetic ground truth [5]. Self-supervised methods relying only on 2D image transformation finds it hard to reject misleading points such as occlusion junction or cast-shadow features. For a purely rotating camera, occlusion junctions are valid features. Similarly, for a stereo rig, features formed by cast shadows are valid. In the case of monocular video, both types of features lead to incorrect or unreliable information about the scene. A data-adaptive interest point detector and descriptor, yet, should be able to cope with all abovementioned scenarios.

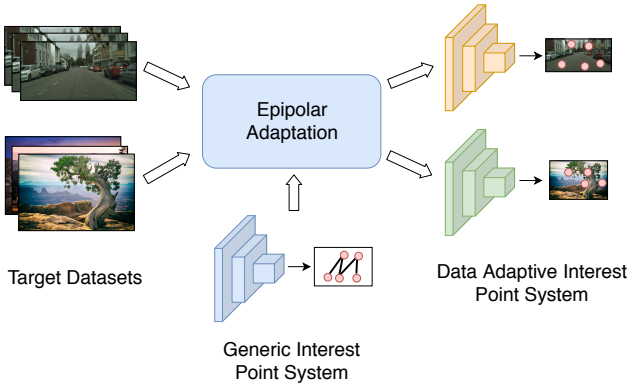

Figure 1. Epipolar Adaptation takes a generic interest point system and makes it adaptive to the target dataset of interest.

In this paper, we propose using the epipolar constraints as an alternative source of supervision to make a data-adaptive interest point detector and descriptor. Characterized by the fundamental matrix, the epipolar constraint of two views map a point from one view to the epipolar line of the other. This constraint provides a large number of negative correspondences since the corresponding point from the other view must lay on the epipolar line. Compared to

supervision from full 3D data, epipolar constraints are easy to obtain since they are properties of the camera configuration of two views. Only the camera intrinsic and the relative pose between two cameras are needed to compute the fundamental matrix between two views. Therefore, epipolar constraints depend solely on the camera configurations between two views, independent of 3D world geometry. It is unlikely that a problematic interest point might lay in the epipolar line throughout the whole data collection process. As a result, epipolar constraints can rule out occlusion junctions or time-dependent interest points such as corners induced by shadows, provided that there are enough variations from the dataset.

We present an algorithm called Epipolar Adaptation that leverages the negative correspondences provided by the epipolar constraint to improve interest points detector and descriptors. The Epipolar Adaptation algorithm takes a generic interest point detector and descriptor, such as a pretrained SuperPoint [5], and fine-tunes such model on a new dataset with each image pair with the fundamental matrix between them. We apply Epipolar Adaptation on two types of datasets: stereo datasets and monocular datasets. The image pairs from the stereo dataset are collected by a fixed stereo camera rig, while the image pairs from the monocular datasets are two frames of the monocular video sequence. We demonstrate that Epipolar Adaptation improves upon the pretrained SuperPoint in synthetic stereo dataset SceneFlow [15] and in real-world monocular sequences from Freiburg RGB-D dataset [23].

2. Related Works

Interest Point Detection and Description. Most traditional interest-point detection and description methods are hand-crafted [13, 1, 19]. These methods focus on interest points in a generic setting and lack provision to adapt toward a specific dataset. Many recent works try to use deep learning to make a data adaptive interest-point detectors and descriptors [2, 25, 5, 22, 16, 6, 20]. Most of them require ground truth point-to-point correspondence between two views [2, 25, 16, 22]. Such ground truth label is difficult to obtain since it requires information about the 3D scene behind the view (e.g. dense depth map or output of SfM). SuperPoint [5], MagicPoint [6], and Quad-Networks [20] use synthetic 2D transformations (e.g., homography) to supervise the interest point detector and descriptor. Since these 2D transformations treat all pixels as textures in a plane, it is difficult to handle nuisance features such as occlusion junctions. This paper proposes an alternative form of supervision using the epipolar constraint. We show that such supervision is easy to obtain and is capable of handling dataset specific features.

Epipolar consistency as supervision. Epipolar geometry has been used as supervision signals in many tasks, such as fundamental matrices estimation [17, 18], depth estimation [8], ground plane estimation [14], semantic segmentation [24], and pose detection [27]. Yi et al. [26] try to find good correspondences by learning the weights for the weighted 8-points algorithm using a regression toward the essential matrix. Dang et al. [4] propose an alternative formulation for eigenvalue-based optimization objectives that allow more stable training for the objectives such as regressing toward the fundamental matrix or the essential matrix. These works leverage the epipolar geometry to supervise the process of selecting the correspondences, but they do not train the interest point detector or descriptor. Our work focuses on the interest point detection and description task. MONET [11] uses epipolar geometry between views to reduce the need for expensive annotations in the key-point detection task. MONET relies on a sophisticated data augmentation schema as well as a small number of ground truth labels to bootstrap the training, while Epipolar Adaptation is completely self-supervised.

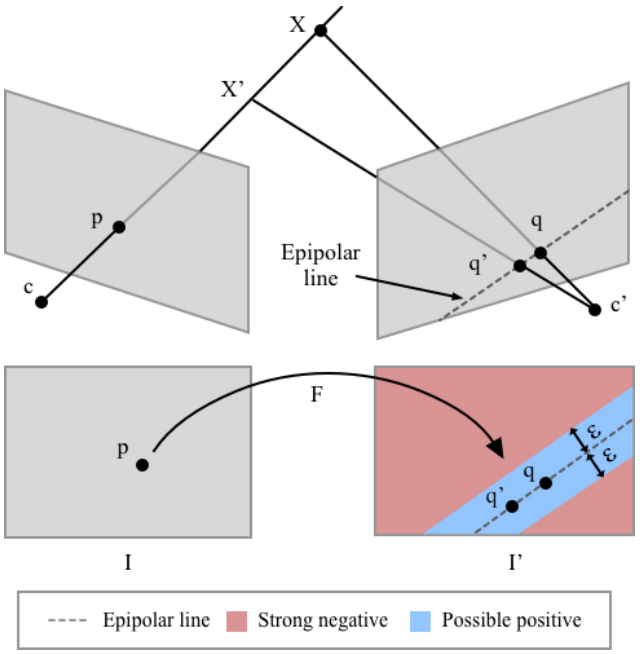

Figure 2. The epipolar geometry between two views maps a point from one view to the epipolar line of the other view. This point-to-line relationship provides a large amount of strong negative correspondences, but leaves the positive matches ambiguous.

3. Epipolar supervision

In this section, we will provide a brief overview of epipolar geometry in Section 3.1. We will then show how to obtain epipolar supervision from stereo image pairs (Sec 3.2) and from pose-tagged monocular videos (Sec 3.3).

3.1. Epipolar geometry

The epipolar geometry defines the intrinsic projective geometry between two views and is independent of the scene structure [9]. Consider a pair of stereo images I and I' with camera center c, c' respectively, and p is a point on image I. All possible 3D points X corresponding to pixel x will lay on the ray connecting c and p. Any projection of $X' \in ray(c, p)$ onto image plane I' will be a potential correspondence pixel q'. The set of q' forms the *epipolar line* on the right image I' for point p. Knowing the epipolar geometry thus constraints the search of correspondences into a single line, thus significantly reducing the search space. Figure 2 illustrates this point to line relationship. Since a pixel on the left image could only map to pixels near its epipolar line, we can obtain a large number of negative correspondences.

The epipolar geometry between two views is uniquely characterized by the fundamental matrix F, which is a rank-2 matrix satisfying $q^T F p = 0$ for all correspondences $p \in I$ and $q \in I'$. The fundamental matrix is usually estimated by RANSAC and eigenvalue methods such as Normalized Eight Point Algorithm [10, 9]. In the following sections, we will introduce how to obtain the ground truth fundamental matrix in the stereo setting and the monocular setting.

3.2. Epipolar constraints in stereo pairs

The fundamental matrix can be retrieved from a pair of stereo images if we know the configuration of two cameras. We assume that we have access to a rigid stereo camera rig, and the stereo dataset $\mathcal{S} = \{(X_i, Y_i)\}_i$ is collected from this fixed camera rig. The intrinsic parameters and relative camera pose can be obtained by the standard camera calibration procedure. Once we obtain the intrinsic parameters for two cameras :K, K', and the relative rotation R and translation t, the fundamental matrix can be obtained using

$$F = (K')^{-T}[t]_\times R K^{-1},$$

where $[t]_x$ is the outer product matrix for vector t. Since the fundamental matrix only depends on the intrinsic camera parameters and the relative camera pose, it is independent of the 3D scene structure [9]. The camera calibration needs to be done only once, since as long as the relative position between two cameras remains unchanged, the fundamental matrix will remain unchanged for all pairs (X_i, Y_i).

3.3. Epipolar constraints in monocular sequence

Typical ways to obtain point-to-point correspondences from a monocular video sequence usually require reconstructing the 3D scene. For example, prior works [23, 3] scan the room with a Kinect sensor to obtain dense depth maps for all the frames. Such 3D information is difficult to obtain in high quality. However, establishing the epipolar constraints between arbitrary two frames in a monocular video is much easier. Since all frames are obtained from the same camera, all frames share the same intrinsic parameters K, which can be obtained from the standard camera calibration process. The key is to figure out the relative rotation R and the relative translation t between camera centers for two frames. Assume that the translation from the camera coordinate of frame i to the canonical coordinate is t_i and its relative rotation is R_i. A point x_i in the camera coordinate of frame i is transformed to the canonical coordinate by $x_c = R_i x_i + t_i$, and $x_c = R_j x_j + t_j$ for frame j. The transformation from i to j is then given by: $x_j = R_j^{-1} R_i x_i + R_j^{-1}(t_i - t_j)$. The relative rotation from i^{th} frame to the j^{th} one is $R_{i\to j} = R_j^{-1} R_i$, and the relative translation is $t_{i\to j} = R_j^{-1}(t_i - t_j)$. Finally, the fundamental matrix from frame i to frame j is

$$F_{i\to j} = K^{-T}[t_{i\to j}]_\times R_{i\to j} K^{-1}.$$

To obtain the fundamental matrices between all pairs of frames, we only need to annotate each frame with its camera pose information and calibrate the camera intrinsic. This process does not require a full 3D reconstruction provided by the public AR/VR libraries such as ARCore.

4. Epipolar adaptation

In this section, we will introduce the Epipolar Adaptation algorithm, which improves a pretrained interest point system on a dataset with epipolar supervision.

4.1. Notations

Let $\mathcal{S} = \{(X_i, Y_i, F_i)\}_i$ be a dataset with epipolar supervision available, and F_i be the fundamental matrix from image X_i to Y_i. We want to fine-tune a pretrained SuperPoint model on $\mathcal{S}$. Recall that SuperPoint will take an image $I \in \mathbb{R}^{H\times W\times 3}$ and predict a feature map with size $Z \in \mathbb{R}^{H_c\times W_c\times h}$. Each pixel in this feature map Z corresponds to an 8×8 pixels area in the original image. From Z, SuperPoint will predict another feature map $\mathcal{X}(I) \in \mathbb{R}^{H_c\times W_c\times 65}$, where each pixel $\mathcal{X}(I)_{i,j} \in \mathbb{R}^{65}$ in this feature map encodes the probability of how likely one of the 8×8 pixels inside the cell will be an interest point. The last dimension of $\mathcal{X}(I)_{i,j}$ represents that there is no interest point in the cell. SuperPoint will also predict a feature map $\mathcal{D}(I) \in \mathbb{R}^{H_c\times W_c\times D}$ that contains the D-dimensional feature descriptors for each cell. The feature descriptor of a pixel from the original image is computed by bilinear interpolating $\mathcal{D}(I)$. In the following sections, we will introduce how to use the outputs of the SuperPoint network and the fundamental matrices to generate epipolar consistent correspondences. These correspondences will be used to fine-tune the SuperPoint network.

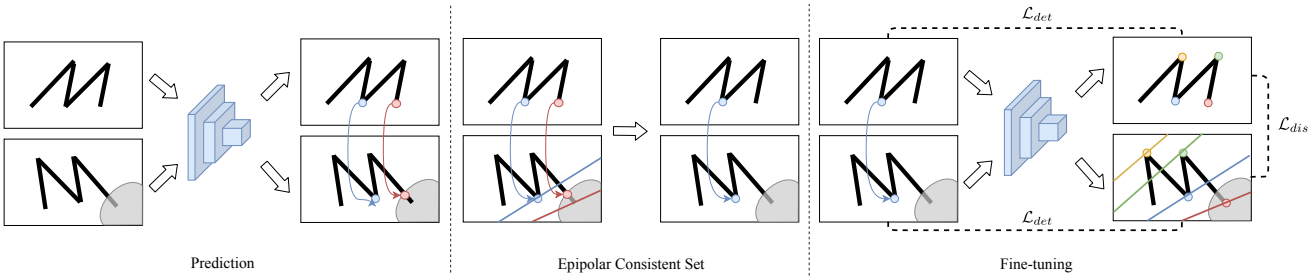

Figure 3. An illustration of three steps of the Epipolar Adaptation algorithm. First, we use a pretrained model to generate matches for all image pairs in the dataset. Second, for each pair of images, we filter out the matches that are not epipolar consistent to create an epipolar consistent set of matches. Finally, we use the epipolar consistent set of matches as ground truth to fine-tune the pre-trained models.

4.2. Epipolar consistent labels

To make the input generic model more data adaptive, we need to provide supervision signal about whether a pair of predicted correspondence is good or not in the target dataset. Note that a good match should fulfill the epipolar constraint, and the matches that do not fulfill the epipolar constraint must be bad matches. Therefore, we can test whether a match is epipolar consistent and can filter out those that are not. One way to test whether a correspondence agrees with the epipolar constraint is to compute the *symmetrical epipolar distance* (SED), which captures how far a point from the correspondence deviate from the epipolar line of the other point [9]. If the correspondence has large SED, then it must be an invalid match. Our network should then be updated to reject such match either by stopping proposing points in the correspondence as interest points or by pushing the descriptor of these two points apart so that they will not be matched again.

Let F be the fundamental matrix between two views. Then, the *symmetrical epipolar distance* (SED) of a pair of point (p, q) is given by:

$$SED(p, q, F) = \frac{\bar{q}^T F \bar{p}}{\sqrt{(F\bar{p})_1^2 + (F\bar{p})_2^2}} + \frac{\bar{p}^T F^T \bar{q}}{\sqrt{(F^T\bar{q})_1^2 + (F^T\bar{q})_2^2}},$$

where $\bar{p}, \bar{q}$ represent the homogeneous coordinates of p, q respectively, and v_1, v_2 represent the first and second coordinate of vector v respectively. Geometrically, $SED(p, q, F)$ represents the sum of the distance from p to epipolar line of q and the distance of q from the epipolar line of p [9]. The unit of SED is pixel. Small SED indicates that the correspondence fulfills the epipolar constraint.

Consider an image pair $(X_i, Y_i) \in \mathcal{S}$ with fundamental matrix F_i. We assume that one can obtain a list of matches for them using the outputs of the pretrained SuperPoint network : $\mathcal{X}(X_i), \mathcal{X}(Y_i), \mathcal{D}(X_i), \mathcal{D}(Y_i)$. Note that our algorithm is independent of the the inference algorithm. Let this list of matches be $M_i = \{(p_j^i, q_j^i)\}_j$, where p_j^i is a point in left image X_i and q_j^i is p_j^i's corresponding point on right image Y_i. We would like to filter out correspondences that do not satisfy the epipolar constraints between X_i and Y_i:

$$\text{EpiAda}(M_i) = \{(p, q) \in M_i, SED(p, q, F_i) < \tau\}.$$

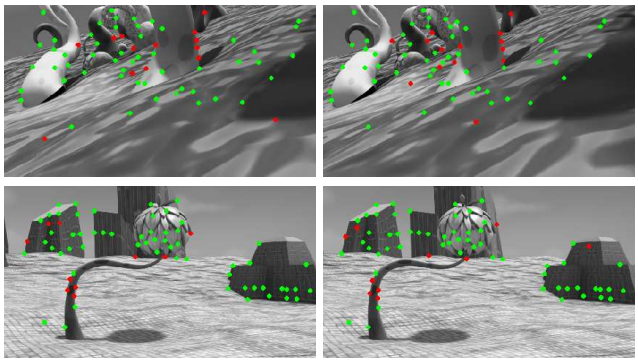

Figure 4. Example of epipolar adapted labels. Each row displays an image pair. All the interest points are predicted by the pretrained SuperPoint model. The red points will be filtered in the epipolar adaptation process. Note that many of the filtered out points are occlusion junctions.

Figure 4 provides examples of the epipolar consistent labels. We can see that some of the interest points that violate the epipolar constraints are occlusion junctions, which can be filtered out when obtaining epipolar consistent labels.

4.3. Training Objectives

A good interest point detector should only detect points that are matchable and correct in the target dataset. Points in the epipolar adapted correspondences fulfill both criterion with respect to the point-to-line mapping provided by the epipolar geometry. We will use interest points from the epipolar adapted correspondences as ground truth. Specifically, let $\mathcal{Y}(I) \in \mathbb{R}^{H_c \times W_c}$ be such ground truth for $\mathcal{X}(I)$. Our detector loss is SuperPoint's original loss applied to ground truth $\mathcal{Y}(I)$:

$$\mathcal{L}_{det}(I) = \frac{1}{H_c W_c} \sum_{h,w=1}^{H_c, W_c} -\log\left(\text{softmax}(\mathcal{X}(I)_{ij})_{\mathcal{Y}(I)_{ij}}\right).$$

A good interest point descriptor should also be able to reject correspondences that are invalid under the epipolar constraint. This provides us with a large amount of high quality negative pairs and few positive pairs, since the epipolar constraint allow a point to be matched to any point in the epipolar line on the target image. Learning only with point-to-line constraints, however, tends to create bad matches since the network can experience catastrophic forgetting of its prior knowledge about how to match one point to the other. To address such issue, we use the epipolar consistent interest points pairs in the epipolar adapted set $\mathrm{EpiAda}(\cdot)$ as the strong positive pairs. Point pairs that are epipolar consistent but not in $\mathrm{EpiAda}(\cdot)$ will be marked as neither positive nor negative.

To reduce computational burden, we will only consider the maximum activated point in each cell. Let $\mathrm{loc}(I) \in \mathbb{R}^{H_c \times W_c}$ represent the maximum activated pixel for each cell of image I. Let F be the fundamental matrix between images X and Y, and let M be the predicted match of the pretrained network for this image pair. Let $\mathcal{C}(X, Y, F) \in \mathbb{R}^{H_c \times W_c \times H_c \times W_c}$ represent the label for all possible pairs of cells between X and Y. Then

$$\mathcal{C}(X,Y,F)_{iji'j'} = \begin{cases} -1 & \text{If } SED(\mathrm{loc}(X)_{ij}, \mathrm{loc}(Y)_{ij}, F) > \tau \\ 1 & \text{If } (\mathrm{loc}(X)_{ij}, \mathrm{loc}(Y)_{ij}) \in \mathrm{EpiAda}(M) \\ 0 & \text{Otherwise} \end{cases}.$$

For each pair of cells with index i, j, i', j', let $d = \mathcal{D}(X)_{ij}$, $d' = \mathcal{D}(Y)_{i'j'}$, and $c = \mathcal{C}(X, Y, F)_{iji'j'}$. Then, the loss for this pair of cells is:

$$\mathcal{L}_{cell}(d, d', c) = \lambda_{pos}\mathbb{I}(c = 1)\max(0, m_p - d^T d') + \lambda_{neg}\mathbb{I}(c = -1)\max(0, d^T d' - m_n),$$

where λ_{pos}, λ_{neg}, m_p and m_n are hyper-parameters, and $\mathbb{I}(\cdot)$ is the indicator function. Note that the descriptor loss will be zero for the cells that fulfills the epipolar constraint but are not predicted by the network previously. The final descriptor loss is:

$$\mathcal{L}_{dis}(X, Y, F) = \sum_{ijkl} \frac{\mathcal{L}_{cell}(\mathcal{D}(X)_{ij}, \mathcal{D}(X)_{kl}, \mathcal{C}(X,Y,F)_{ijkl})}{(H_c W_c)^2}$$

Our final training objective is a combination of both the detector loss and the descriptor loss:

$$\mathcal{L}(X, Y, F) = \mathcal{L}_{det}(X) + \mathcal{L}_{det}(Y) + \mathcal{L}_{des}(X, Y, F).$$

5. Experiment

In this section, we empirically evaluate the effectiveness of our method in both stereo and monocular datasets. [1]

[1] Codes will be available at https://github.com/stevenygd/SuperPointEA

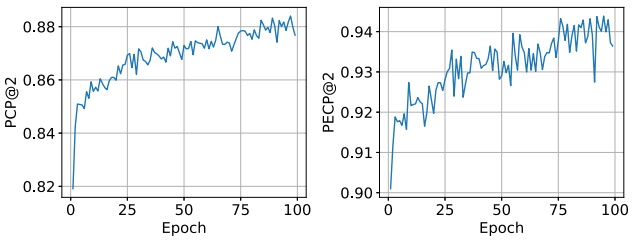

Figure 5. Validation performance. We can see that PECP@T and PCP@T align well during training, indicating that PECP@T can be used as a good approximation for PCP@T.

5.1. Dataset

We will evaluate the performance of our method in both stereo and monocular datasets. We use the SceneFlow dataset [15] for the stereo experiment. SceneFlow dataset contains rectified stereo pairs of synthetic images. Rectified stereo images provide the epipolar constraints where the epipolar line is horizontal. We re-size all images down to 464×256, a dimension that roughly matches that of pretrained dataset. For monocular experiments, we use the Teddy sequence of Freiburg RGB-D dataset [23]. We first filter out frames that are blurry. We then pre-compute the fundamental matrix using the provided intrinsic and camera pose for all pairs of frames. We further filter out frame pairs that do not have sufficient feature overlap. Finally, we obtain 1000 frame pairs from the monocular sequence.

5.2. Evaluation Metrics

Let I_l be the set of interest point predicted from the left image and I_r from the right image. For all $p \in I_l$, let $G(p)$ be p's correct corresponding point on I_r. We use the following metrics to evaluate an interest point system.

Repeatability at radius T (REP@T) measures how likely an interest point can be repeated in the target image [21]. A point is considered repeated in the target image if there exists an interest point from the target image that is close to the correct corresponding point. Formally, for radius T, REP@T can be computed as:

$$REP_T(I_l, I_r) = \frac{|\{p | \exists q \in I_r, |G(p) - q| < T\}|}{\min |I_l|, |I_r|}.$$

Percentage of Correct Points at radius T (PCP@T) evaluates how well the system can match the interest point to the correct corresponding point on the target image. This metric corresponds to the PCK@T used in [11]. Let $C = \{(p, q) | p \in I_l, q \in I_r\}$ be the predicted correspondences. For radius T, PCP@T is given by:

$$PCP_T(C) = \frac{|\{(p, q) | (p, q) \in C, |q - G(p)| < T\}|}{|C|}.$$

Note that both REP@T and PCP@T requires knowing the dense ground truth correspondences between two im-

Table 1. Performance in the SceneFlow dataset. SuperPointEA is the model after Epipolar Adaptation. EA Labels are the epipolar consistent labels (Sec 4.2).

Metric	Model	Test Performance (%)
PCP@2	SuperPoint	73.25
	SuperPointEA	**87.67**
	EA Labels	86.65
REP@2	SuperPoint	58.27
	SuperPointEA	**60.65**
	EA Labels	88.69
PECP@2	SuperPoint	81.88
	SuperPointEA	**93.64**
	EA Labels	100

ages. Such ground truth correspondences are difficult to obtain. For the synthetic dataset SceneFlow, we compute such ground truth from the dense disparity map. For the real world monocular dataset, however, ground truth correspondences obtained from 3D information are usually very sparse. As a result, REP@T and PCP@T cannot be used as a useful metric in such scenarios. Note that one can obtain high quality ground truth fundamental matrices even in a real-world dataset. We thus relax PCP@T to test instead whether a correspondence fulfills the point-to-line relationship from the epipolar geometry in the following metric:

Percentage of Epipolar Correct Points at radius T (PECP@T) measures the percentage of interest point matches that satisfy the epipolar constraints. Let F be the fundamental matrix for the image pair of interest and T be the tolerance margin. PECP@T is computed as follow:

$$PECP_T(C) = \frac{|\{(p,q)|(p,q) \in C, SED(p,q,F) < T\}|}{|C|}.$$

Figure 5 shows that PECP@T is a good approximation for PCP@T since the align well during training.

5.3. Stereo experiment

We run Epipolar Adaptation on pretrained SuperPoint network with threshold $\tau = 2$. We use Adam [12] with learning rate $1e-5$ to fine-tune the network for 100 epochs. We set $\lambda_{pos} = 300$ and $\lambda_{neg} = 1$ to balance the number of positive and negative examples. Other hyper-parameters are adapted from SuperPoint [5]. Since we have ground truth correspondences between views, we report all three metrics for this experiment in Table 1 with the threshold set to be 2. Note that the model after Epipolar Adaptation outperforms the original SuperPoint model in all metrics. The performance of SuperPointEA in PCP@2 can even match the performance of the epipolar consistent labels used to fine-tune

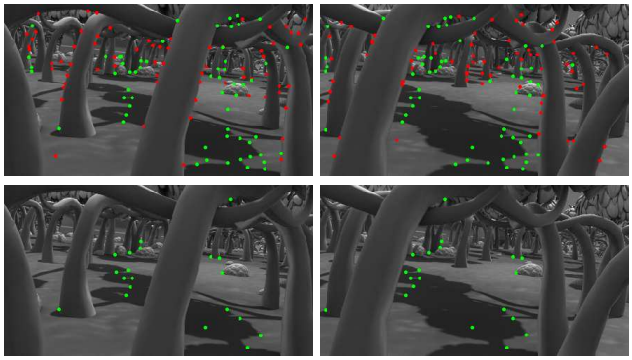

Figure 6. Interest points predictions of stereo image pairs from SuperPoint in the first row and SuperPointEA in the second row. Green points are epipolar consistent, while red points are not. We can see that SuperPointEA latches to points that are more epipolar consistent and SuperPointEA predicts fewer occlusion junctions.

Table 2. Performance for monocular dataset. SuperPointEA refers to the model after Epipolar Adaptation.

Model	PECP@4 (%)
SuperPoint	33.73
SuperPointEA	**34.73**

SuperPointEA network. This result suggests that the Epipolar Adaptation algorithm is able to make the baseline model adaptive to the target stereo dataset. Figure 6 contains visualization for both SuperPoint and SuperPointEA, predicts fewer occlusion junctions.

5.4. Monocular experiment

The monocular experiment is conducted in the Teddy sequence from Freiburg RGB-D dataset. Since the resolution of this dataset is about 4 times as large as the stereo dataset, we set the threshold for Epipolar Adaptation to be $\tau = 4$. During training, we set $\lambda_{pos} = 1$ and $\lambda_{neg} = 1$. Other hyper-parameters are the same as in the stereo experiment. Since the monocular dataset has precise fundamental matrices between frames but no dense point-to-point correspondence, we only report PECP@T in this experiment. The results are presented in Table 2. SuperPointEA out-performs the baseline SuperPoint in this task, which suggests that Epipolar Adaptation is able to make SuperPoint adaptive to the target monocular dataset.

6. Conclusion and future works

This paper proposes and evaluates Epipolar Adaptation, an algorithm that uses epipolar supervision to make a generic interest point system adaptive to a dataset. Interesting future works include generalizing Epipolar Adaptation to more generic interest point systems.

7. Acknowledgement

This work was supported in part by a research gift from Magic Leap.

References

[1] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 1, 2

[2] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker. Universal correspondence network. In *Advances in Neural Information Processing Systems*, pages 2414–2422, 2016. 1, 2

[3] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3

[4] Z. Dang, K. M. Yi, Y. Hu, F. Wang, P. Fua, and M. Salzmann. Eigendecomposition-free training of deep networks with zero eigenvalue-based losses. *arXiv preprint arXiv:1803.08071*, 2018. 2

[5] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. *arXiv preprint arXiv:1712.07629*, 2017. 1, 2, 6

[6] D. DeTone, T. Malisiewicz, and A. Rabinovich. Toward geometric deep slam. *CoRR*, abs/1707.07410, 2017. 1, 2

[7] D. DeTone, T. Malisiewicz, and A. Rabinovich. Self-improving visual odometry. *CoRR*, abs/1812.03245, 2018. 1

[8] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. 2

[9] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3, 4

[10] R. I. Hartley. In defense of the eight-point algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19:580–593, 1997. 3

[11] Y. Jafarian, Y. Yao, and H. S. Park. Monet: Multiview semi-supervised keypoint via epipolar divergence. *arXiv preprint arXiv:1806.00104*, 2018. 2, 5

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 6

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1, 2

[14] Y. Man, X. Weng, and K. M. Kitani. Groundnet: Segmentation-aware monocular ground plane estimation with geometric consistency. *CoRR*, abs/1811.07222, 2018. 2

[15] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134. 2, 5

[16] Y. Ono, E. Trulls, P. Fua, and K. M. Yi. Lf-net: learning local features from images. In *Advances in Neural Information Processing Systems*, pages 6237–6247, 2018. 1, 2

[17] O. Poursaeed, G. Yang, A. Prakash, Q. Fang, H. Jiang, B. Hariharan, and S. Belongie. Deep fundamental matrix estimation without correspondences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2

[18] R. Ranftl and V. Koltun. Deep fundamental matrix estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–299, 2018. 2

[19] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE international conference on*, pages 2564–2571. IEEE, 2011. 1, 2

[20] N. Savinov, A. Seki, L. Ladicky, T. Sattler, and M. Pollefeys. Quad-networks: unsupervised learning to rank for interest point detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[21] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37:151–172, 2000. 5

[22] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015. 2

[23] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012. 2, 3, 5

[24] Y. Yao and H. S. Park. Multiview cross-supervision for semantic segmentation. *CoRR*, abs/1812.01738, 2018. 2

[25] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483. Springer, 2016. 1, 2

[26] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua. Learning to find good correspondences. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, number CONF, 2018. 2

[27] Y. Zhang and H. S. Park. Multiview supervision by registration. *CoRR*, abs/1811.11251, 2018. 2