Towards Improved Interest Point Detection and Description using Deformable Convolutions

Gunjan Sethi
Masters in Robotic Systems Development
Carnegie Mellon University
gunjans@andrew.cmu.edu

Sai Shruthi Balaji
Masters in Robotic Systems Development
Carnegie Mellon University
saishrub@andrew.cmu.edu

***Abstract*—Learning-based interest point detection methods like SuperPoint are extremely fast to execute and perform well real time compared to classical interest point detection methods like SIFT [1], however, they achieve poor repeatability in comparison [2]. Repeatability is a metric that evaluates the geometric stability of interest points under different transformations. We propose the use of deformable convolutions [3] in the detector head of learning based interest point detectors such as SuperPoint, since they are known for their ability to learn features that can adapt to geometric variations of objects. With this improvement implemented in the detector head of SuperPoint model, we demonstrate the improvement in a variety of metrics. We also evaluate the D-SuperPoint model on downstream tasks like 3D reconstruction using ColMap and analyze the results.**

I. Introduction

We aim to answer the question – *do deformable convolutions improve the performance of CNN-based interest point detectors?*

Several geometric vision tasks like Structure from Motion (SfM) and camera calibration require a robust set of interest points between two image frames. Interest points are pixels that represent unique geometric information about objects in the image - lines, corners, etc. Further, interest points must also be repeatable under varying lighting and views. Most algorithms for vision tasks assume a reliable set of extracted and matched interest points. However, it is challenging to extract high quality interest points from real-world images.

Classical interest point detectors such as Harris Corner Detector, SIFT, and so on, use a sliding window approach. In contrast, learning-based detectors rely on fully convolutional layers to extract features. Classical methods are efficient and produce interest points with high repeatability. However, they are not suitable for real-time applications due to high inference times. This is where learning-based detectors can significantly outweigh classical detectors with low inference times and with some architecture tweaks, improved reliability.

While regular convolutions, used by learning-based detectors, have achieved success for visual recognition tasks, their ability to model geometric transformations is limited. This is due to the static nature of the convolution layer that samples the input image only at fixed locations. To tackle this, the training data requires extensive data augmentation, or new ways of feature pooling may be required. Deformable Convolutional Networks [3] propose a new type of convolution layer called Deformable Convolution, proven to improve a model's ability to model geometric transformation.

We hypothesize that if interest points are focused on the spatial geometry of the objects in the image, they can be more repeatable and robust. This is because objects of interest are repeated between image pairs and often have more interesting features than the background. Therefore, observing maximum interest points on the object and its boundary can improve interest point detection repeatability. We propose the use of deformable convolution layers in current state-of-the-art learning-based interest point detectors. For the purpose of this project, we picked SuperPoint.

A. SuperPoint

SuperPoint is a self-supervised framework for training interest point detectors and descriptors. [2] Interest points are semantically ambiguous. What may seem like a good interest to one, may not be the same for another. Since most learning-based methods today rely on human-annotated ground truths, how can we overcome this problem of supervision? SuperPoint solves this problem by adopting a self-supervised approach. MagicPoint, SuperPoint's interest point detector, is first trained on millions of synthetic images. These synthetic images have simple geometrical figures like lines, quadrilaterals, ellipses, and so on, and have obvious interest points. To adapt this interest point detector to real-world images, Homographic Adaptation (HA) is used. Target, unlabeled, real-world images are warped into random homographic transformations and ground-truth interest points are generated. These generated ground truths are used to train a fully-convolutional network that extracts interest points and descriptors from test images.

MagicPoint combined with the descriptor head completes the SuperPoint architecture. Figure 1 shows the different stages of the SuperPoint training pipeline.

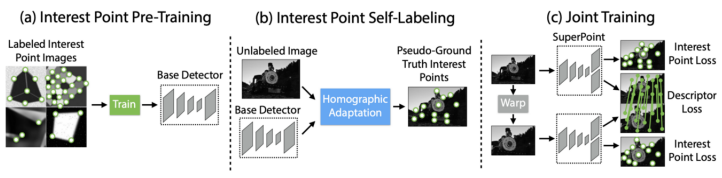

Fig. 1. SuperPoint training pipeline

The authors of SuperPoint conclude that the network has a competitive repeatability score against classical methods and outperforms when evaluated for homography estimation.

B. Deformable Convolutions

Deformable convolutions add 2D offsets to the regular kernel grid used in standard convolutions. This allows the convolution kernel to change its shape from a 2D $n \times n$ and "deform" into a free-form shape, slowly adapting to the geometric spatial extent of the object of interest. The aforementioned offsets are learned by the model. The inference time of CNNs is not affected by these additional learned parameters since the number of parameters is increased only linearly. Figure 2 shows a deformable convolution layer with learned offsets.

Deformable Convolutions v2 incorporates the deformable convolution layer from earlier but introduces a new modulation strategy where samples are also modulated based on learned feature amplitude.

Fig. 2. Example of detected features and resulting segmentation mask using Deformable Convolutions.

II. Prior Work

Classical Interest Point Detectors such as Harris Corner detectors [3] and SIFT [4] use sliding window or kernel-based techniques to detect the amount of change seen around a pixel. However, these are all hand-crafted, computationally heavy techniques (except for ORB which is an efficient rotation invariant version of the FAST detector) and do not perform well real time.

CNN-Based Interest Point Detectors such as LIFT [7], SuperPoint [2], SIPs [4] and KP2D [8] have the central idea of leveraging CNNs to learn per pixel interest scores using a series of VGG-like convolutional layers. These methods have lower repeatability scores than classical methods, and do not perform on par with classical techniques when there are geometric transformations involved.

There is no significant work that integrates deformable convolutions [3] with interest point detectors and descriptors. However, d-convs have been successfully used for feature extraction in several applications such as 3D object detection and segmentation [10], learning temporal pose estimation from sparsely-labeled videos, video super-resolution [9], and so on. These works have shown that deformable convolutions help in learning better geometric features and adapt well to geometric transformations.

III. Implementation

Our implementation majorly involves writing the deformable convolution layer in PyTorch, replacing the convolutions in SuperPoint's detector head with deformable convolutions and replicating the training and evaluation steps of SuperPoint.

A. The Deformable Convolution Layer

The Deformable Convolution layer was implemented with the help of PyTorch TorchVision's `deform_conv2d`. This implements version 2 of deformable convolutions that have learnable offsets as well as learnable modulation scalars.

For convolution kernels of dimension $n \times n$, offsets are of dimension $2 \times n \times n$ in order to learn feature offsets along the height and width of the image. Offsets spread the learned convolution kernel features across space, thus giving flexibility to learn a particular object's features even when it undergoes geometric transformations.

The modulation scalars are of the same dimension as the kernels $(n \times n)$. The modulation units are capable of modulating the input feature amplitudes, and can thus choose to modify the features in certain spatial bins of the image. This leads to more flexibility for learning focused representations of objects of interest. Figure 3 illustrates this approach.

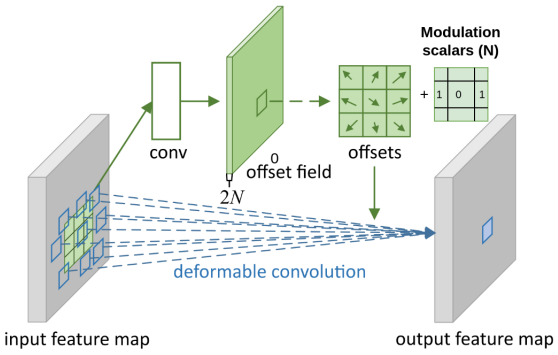

Fig. 3. Deformable Convolution V2

We believe that these two characteristic features of deformable convolutions will result in interest points that are repeatable across homographies, and are also focused on objects of interest.

B. MagicPoint Dataset Generation and Training

This step is adapted from SuperPoint's implementation, and involves simultaneously training the MagicPoint Architecture while generating the synthetic shapes data. Figure 4 illustrates the auto labeled synthetic shapes dataset which is trained on the base detector that has been replaced with deformable convolutions.

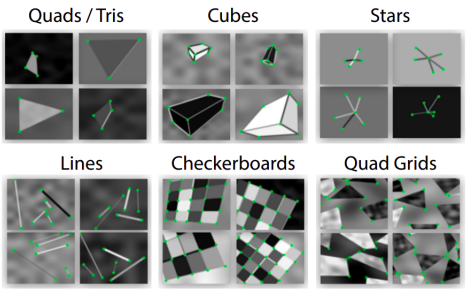

Fig. 4. Labeled Synthetic Shapes Dataset

C. SuperPoint Data Setup and Training

With the trained MagicPoint model, we generate interest point ground truth labels on the MS-COCO dataset. This step also performs the homography adaptation from SuperPoint where ground truth labels are generated for randomly homographized MS-COCO images as well.

Using the ground truth labels generated, joint training of the full SuperPoint Architecture is performed. By joint training, it means that the training optimizes two losses, namely the interest point loss and the descriptor loss.

The interest point loss minimizes the distance of the detected interest points from the ground truth interest points using a simple convolutional cross entropy loss. Given a known homography (and hence known correspondences) between a pair of images, the descriptor loss maximizes the probability of repeatability of the interest point detections in the transformed image. Figure 5 shows an example image pair with detected interest points in the original image as well as the homographized image.

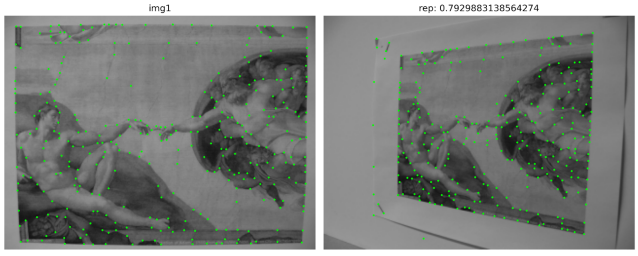

Fig. 5. Interest Point detections on original and transformed image

IV. Evaluation

A. Experiments

We conducted two major experiments. First, we trained DMagicPoint (MagicPoint with deformable convolutions). We generated the synthetic dataset and trained the DMagicPoint model for 100,000 iterations. Some results can be seen in Figure 4. We performed only qualitative analysis on DMagicPoint since it is merely trained on synthetic shapes. Qualitatively, DMagicPoint performs on par with MagicPoint. In most cases, the model is capable of detecting the interest points of the shape. Next, we used Homographic Adaptation to generate pseudo-ground-truth labels on the MS COCO dataset. We trained DSuperPoint (SuperPoint with deformable convolutions) supervised by these generated groundtruth labels, for 50,000 iterations. This model was then qualitatively and quantitively evaluated on the HPatches dataset for several metrics mentioned in the following section. The HPatches [11] is a dataset used for local patch descriptor evaluation. It comprises 116 image sequences of 6 images with known homography.

B. Metrics

1) Mean Localization Error (MLE): Localization error is defined as the average L1 distance error between the detected interest point and ground truth interest point, across all interest points. The value of this error is between 0 and ϵ, where a lower value signifies better interest points.

2) Repeatabilty: Given an interest point in an original image, repeatability is the probability of finding the same interest point in the warped image as well. A high repeatability score shows that the interest point detector is invariant to homography transformations.

3) NN mAP: This metric measures the differentiating ability of the descriptor. At different distance thresholds, nearest-neighbor matching is performed and the area under curve is computed and averaged.

4) M Score: This metric is a measure of the efficacy of the interest point detector and descriptor. It is a ratio of the number of ground truth correspondences recovered over the total number of features in the shared region of the image pair.

5) Homography Estimation: Given a known homography between a pair of images, we estimate the ability to calculate this homography from the detected and matched interest points.

C. Comparison to Baseline SuperPoint on HPatches Dataset

Qualitatively, we can observe that our hypothesis was correct: DSuperPoint is able to detect more interest points on the objects of interest rather than on the background. Some of these results are illustrated in Figure 6.

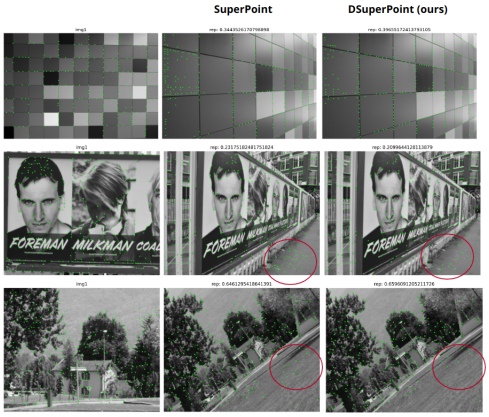

Fig. 6. SuperPoint vs DSuperPoint results. The red circles show minimum interest points detected in low-interest areas.

Quantitatively, our DSuperPoint model reaches loss convergence in 50,000 iterations as opposed to SuperPoint which is trained for 100,000 iterations. DSuperPoint outperforms SuperPoint in the detector metrics, namely repeatability and MLE as shown in Figure 7.

	Detector Metrics		Descriptor Metrics	
	Repeatability	MLE	NN mAP	M Score
DSuperPoint (Ours)	**.624**	**1.100**	.682	.396
SuperPoint	.581	1.158	**.821**	**.470**
LIFT	.449	1.102	.664	.315
SIFT	.495	**0.833**	.694	.313
ORB	**.641**	1.157	.753	.266

Fig. 7. Detector and Descriptor Metrics

DSuperPoint also achieves significantly higher scores in homography estimation, especially with smaller distance thresholds as furnished in Figure 8.

	Homography Estimation		
	ε = 1	ε = 3	ε = 5
DSuperPoint (Ours)	**.425**	**.703**	.768
SuperPoint	.310	.684	**.829**
LIFT	.284	.598	.717
SIFT	.424	.676	.759
ORB	.150	.395	.538

Fig. 8. Homography Estimation Metrics

D. Downstream Task evaluation: 3D reconstruction

Other than evaluation on just 2D images, the significance of interest point detection is more in downstream tasks that involve 3D such as SLAM and 3D Reconstruction. Therefore, we perform a comparative analysis of the reconstruction of a 3D scene using SuperPoint and DSuperPoint as the interest point detectors. For this analysis, we use `supercolmap` [12], an open source library which replaces the classical SIFT interest point detector in colmap with SuperPoint.

Figure 9 shows the qualitative results of the ColMap reconstruction. Though not obvious, on closer observation, we can notice sharper edges and better reconstruction towards the top of the building. Some areas that appeared sparse in the SuperPoint reconstruction appeared denser in the DSuperPoint Reconstruction. This proved that more interest points were produced on the object of interest rather than on the background. Quantitatively, DSuperPoint results in lesser reprojection error (0.61) in comparison with SuperPoint (0.62).

This evaluation was performed by only generating 5000 interest points. We believe that better results would be produced when more interest points are generated. Therefore, deformable convolutions produce promising quantitative results even on downstream 3D tasks.

V. Conclusion

In the work, we showed that deformable convolutions have the potential for improving the repeatability and minimizing

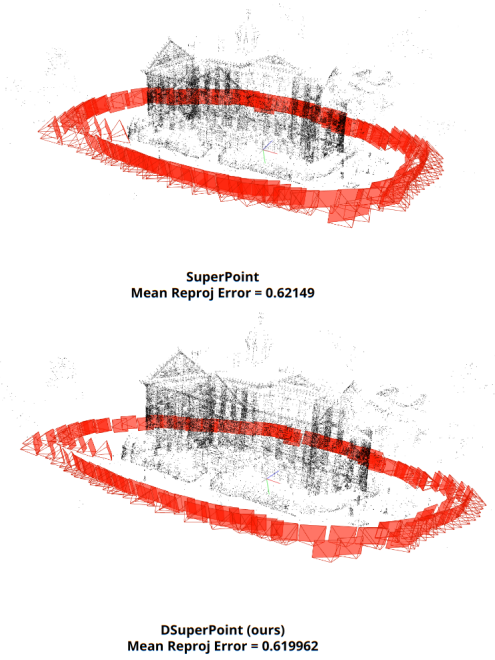

Fig. 9. 3D reconstruction using interest points from SuperPoint and Deformable Superpoint

localization error of learning-based interest point detectors. Furthermore, when used for downstream tasks like homography estimation or SfM, we proved that these new interest points perform better. While homography estimation gives significantly better results than learning-based and classical methods, SfM shows minor improvement.

References

[1] "On the Comparison of Classic and Deep Keypoint Detector and Descriptor Methods," arXiv:2007.10000v2 [cs.CV] 29 Jul 2020

[2] "SuperPoint: Self-Supervised Interest Point Detection and Description," https://arxiv.org/pdf/1712.07629.pdf

[3] "Deformable ConvNets v2: More Deformable, Better Results," https://arxiv.org/pdf/1811.11168.pdf

[4] "Succinct Interest Points from Unsupervised Inlierness Probability Learning" SIPs, https://rpg.ifi.uzh.ch/docs/3DV19_Cieslewski.pdf.

[5] "A Combined Corner and Edge Detector", http://www.bmva.org/bmvc/1988/avc-88-023.pdf

[6] "Distinctive Image Features from Scale Invariant Keypoints", https://www.cs.ubc.ca/ lowe/papers/ijcv04.pdf

[7] "LIFT: Learned Invariant Feature Points", https://arxiv.org/abs/1603.09114

[8] "Neural Outlier Rejection for Self-Supervised Keypoint Learning", https://openreview.net/pdf?id=Skx82ySYPH

[9] "Deformable 3D Convolution for Video Super-Resolution", https://arxiv.org/abs/2004.02803.

[10] "Flexible and Deformable Convolution for Point Clouds" KPConv, https://arxiv.org/abs/1904.08889.

[11] "A benchmark and evaluation of handcrafted and learned local descriptors" HPatches, https://hpatches.github.io/.

[12] "SuperColMap: SuperPoint replaces SIFT in the colmap framework", https://github.com/Xbbei/super-colmap.