We thank the reviewers for their constructive feedback. Our answers to all the questions are presented below.

Reviewer 1

Edge effects. We use zero-padding when rotating an image. Downsampling doesn't need padding.

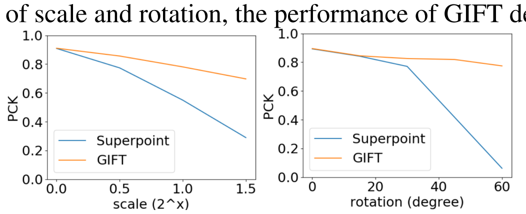
Compact group. Since the scale group is unbounded, in implementation we empirically select a reasonable range of scales and set feature values outside the range to zero. Thus, the permutation property doesn't rigorously hold near the boundary of the selected range. But empirical results show that this boundary effect will not obviously affect the final matching performance if the scale change is not too large.

Limitations of regular grid. We agree that regular grids and regular convolutions are only applicable to groups on which unit transformations can be defined. On groups without properly-defined unit transformations, we may resort to other techniques to compute group convolutions, e.g. G-FFT proposed in [9]. Provided well-defined group convolution, we can still exploit structures of group features to construct GIFTs.

Reviewer 2

Importance of affine-invariant descriptors. If the observed object is smooth, the perspective transformation of a local region can be well approximated by an affine transformation [38]. Most of the existing works about local descriptors [3, 7, 11, 13, 21, 32, 35, 38, 40, 51, 56] focus on affine transformations or the subset of rotation and scaling. To the best of our knowledge, GIFT is the first CNN-based descriptor with provable robustness to rotation and scaling.

Evaluation on datasets with systematic increase of scaling and rotation. The following figure shows the performance of GIFT on the HPatches dataset with systematic increase of scaling and rotation. It shows that, with the increase of scale and rotation, the performance of GIFT degrades much more slowly than the performance of Superpoint.



	error(°)	inlier
SIFT	35.88	104.97
Superpoint	20.83	60.48
GIFT-SP	17.02	65.66
GIFT-Dense	**15.14**	**927.31**

Table 1: Relative pose estimation.

	PCK	Std
GIFT	67.15	14.59
Superpoint	53.66	21.88

Table 2: Standard deviation on View-HP.

Experiments on more challenging datasets. We have evaluated GIFT on a non-planar indoor dataset SUN3D using the PCK as the metric in Section 4.3. In order to further demonstrate the potential of GIFT for real computer vision tasks, here we provide additional results for relative pose estimation on the SUN3D dataset. The mean error of estimated poses and the average number of inlier correspondences are listed in Table 1. GIFT-SP uses SuperPoint as the detector while GIFT-Dense uses a dense grid as keypoints.

Standard deviation. The standard deviations on the View-HP dataset are given in Table 2. We will add the standard deviations for all other experiments in the revised manuscript.

Clarity. We thank the reviewer for the suggestions and will definitely improve the clarity in the revision.
(1) In Table 4 of Section 4.3, the first row lists names of the detectors. More explanations will be added in the text.
(2) We will add symbols to Figure 1 in the revised manuscript, as suggested by the reviewer.
(3) About Equation 2, H contains 9 elements in the implementation. On every element $h \in H$, $W_i(h)$ is a vector with n_{l-1} elements where n_{l-1} is the number of input channels and i is the index of output channels. The group convolution defined in Equation 2 was originally proposed in [8] as mentioned in Line 89-91. Due to the space limit, we only give a brief introduction and refer the readers to [8] for more details.

Reviewer 3

Motivation of using bilinear pooling. As shown by Lemma 1 and Lemma 2, the transformation of an image results in a permutation of its group features. Thus, instead of using a permutation-sensitive FC layer, we adopt a permutation-invariant pooling operator to gain the invariance to transformations. We choose bilinear pooling rather than max or average pooling for two reasons. First, bilinear pooling collects expressive second-order statistics, retaining the distinctiveness of the resulting descriptors. Second, bilinear pooling makes GIFT a generalized model of former descriptors [21, 51, 56], as proved in the supplementary material. The ablation study in Table 2 shows that the bilinear pooling gives a better performance than max or average pooling.

Comparison to the model with FC layers and Group Equivariant CNNs. In Table 1 of the manuscript, we have compared GIFT-1 to GFC which uses FC layers instead of group convolutions. Here, we additionally provide the results of GFC with one more group convolution layer (GFC+GC) in Table 3. The original Group Equivariant CNN is not designed for this task and is not directly comparable. Instead, we have implemented a baseline model similar to Group Equivariant CNNs, which is described in Line 252-255, and the results are reported in Table 2 of the manuscript (max pooling).

	ER-HP	ES-HP
GIFT-1	39.68	21.74
GFC+GC	30.25	17.23

Table 3: PCK on ER- and ES-HP.

Advanced backbones. Advanced backbones with more training data may help. But our objective is to improve invariance with properly designed geometric components which can be integrated in any CNN backbones .