Reviewer #1:

We thank R1 for confirming our contributions. As requested, we further clarify SAFA in Sec.3.2 as follows:

In SAFA, multiple embedding maps M are generated by the spatial-aware position embedding (SPE) modules. For each SPE module, the purpose of max-pooling along channels is to obtain the most dominant features across spatial locations, and the two fully-connected layers are used to select features among the prominent ones as well as encode the spatial information and responses of the chosen features. By doing so, the SPE is able to highlight salient features while mitigating the feature distortions between the cross-view images. Moreover, SPE modules are adopted in both the ground and aerial branches, and our objective forces them to encode corresponding features between these two branches. We employ multiple SPE modules and aggregate encoded features to increase the robustness and discriminativeness of our model. In doing so, our SPE modules are initialized with different weights and thus able to generate different attention maps. Figure 2 in the supplementary material provides the visualization of the generated different embedding maps. $\langle ., . \rangle$ denotes the Frobenius inner product of the two inputs.

Reviewer #2:

We thank R2 for the thorough and comprehensive comments. We are glad that R2 thinks the contribution is interesting.

1. Polar coordinate. The polar transform takes the center of each aerial image as the origin without using any ad hoc pre-centering process. In general, the north direction of a satellite map is often available, and thus we use it as angle 0° in the polar transform. During testing, we do not assume the ground-to-aerial pair is perfectly aligned. In fact, small offsets on the polar origin do not affect the appearance of polar-transformed aerial images obviously, and the small appearance changes will be reduced by our SPE modules. On the contrary, when a large offset occurs, the aerial image should be regarded as a negative sample and the polar-transformed aerial image will be significantly different from the ground-truth one. In this manner, our polar transform effectively increases the discriminativeness of our model.

2. Actual retrieval. In Sec. 4.2, our experiment on CVACT_test is a real retrieval case. The aerial images in the database densely cover the city of interest. However, there is no guarantee that the location of the query ground-level image corresponds to the center of an aerial image. We still take the center of each aerial image as the polar origin, and Fig. 6(c) demonstrates the robustness and effectiveness of our method to unknown center offsets.

3. Multiple embeddings in SPE. Only one polar transform is applied to aerial images and the multiple embeddings are obtained by different SPE modules. Please refer to the response to R1 for the details of SPE.

4. Robust to changes. The aerial and the ground images in our experiments are obtained at different times [22, 12]. Therefore, our model is able to tolerate appearance changes. Once an updated satellite map is available, we can directly apply the trained model to re-extract the features of the database images and then perform the retrieval.

5. Other competitors and datasets. Cross-view image localization is a newly emerging problem and the two competitive algorithms included in our submission are the most advanced ones: the CVM-NET is published in CVPR2018 and the work of Liu *et al.*is published in CVPR2019. CVACT dataset is released recently in CVPR2019.

6. Settings in [18], limitations to panoramas. We test our algorithm on the dataset of [18] although it is not the case for our work. We achieve 71.5% on recall@1%, which is much higher than that of the original paper (59.9%). It also demonstrates that our proposed algorithm is not restricted to the panorama case.

Reviewer #3:

Thanks for R3's supportive comments and confirming our contributions. We response R3's specific concerns below:

1. Structure of SPE. Our SPE modules are trainable and we will clarify this in the revised version. Please refer to our response to R1 for more details.

2. Overall framework. SAFA takes the output of the last convolutional layer of VGG16 as input and the two branches do not share weights. We will release the source code soon.

3. Orientation misalignments. We convert rotational misalignments between ground and aerial images into translational shifts by the polar transform, thus facilitating CNNs to extract features. Moreover, since our SAFA embeds the relative spatial relationship between features instead of absolute locations, our method is insensitive to the misalignments. Our algorithm achieves 85% on CVUSA and 79% on CVACT at r@1 within ±20° orientation perturbations.

4. CVACT_val and CVACT_test. Both CVACT_val and CVACT_test are test sets and their names are inherited from [12]. As CVACT_val is evaluated by the same metric as CVUSA, we list them in the same Tables. In CVACT_test, when an estimated location from aerial images is less than 5 meters to the ground-truth GPS, it will be considered as successful geo-localization. This is different from the evaluation metric employed in CVACT_val and CVUSA. Therefore, we report the results on CVACT_test separately.