# Understanding Scenery Quality: A Visual Attention Measure and Its Computational Model Supplementary

Yuen Peng Loh†, Song Tong, Xuefeng Liang, Takatsune Kumada, Chee Seng Chan†
IST, Graduate School of Informatics, Kyoto University, 606-8501 Kyoto, Japan
† Centre of Image and Signal Processing, University of Malaya, Kuala Lumpur, 50603 Malaysia
lohyuenpeng@siswa.um.edu.my, tong.song.53w@st.kyoto-u.ac.jp
xliang@i.kyoto-u.ac.jp, t.kumada@i.kyoto-u.ac.jp, cs.chan@um.edu.my

## Abstract

*This supplementary file shows additional details regarding the comparison results between our proposed computational model with the state-of-the-art method by Tong* et al. *[1], and tourist-taken photos we have collected.*

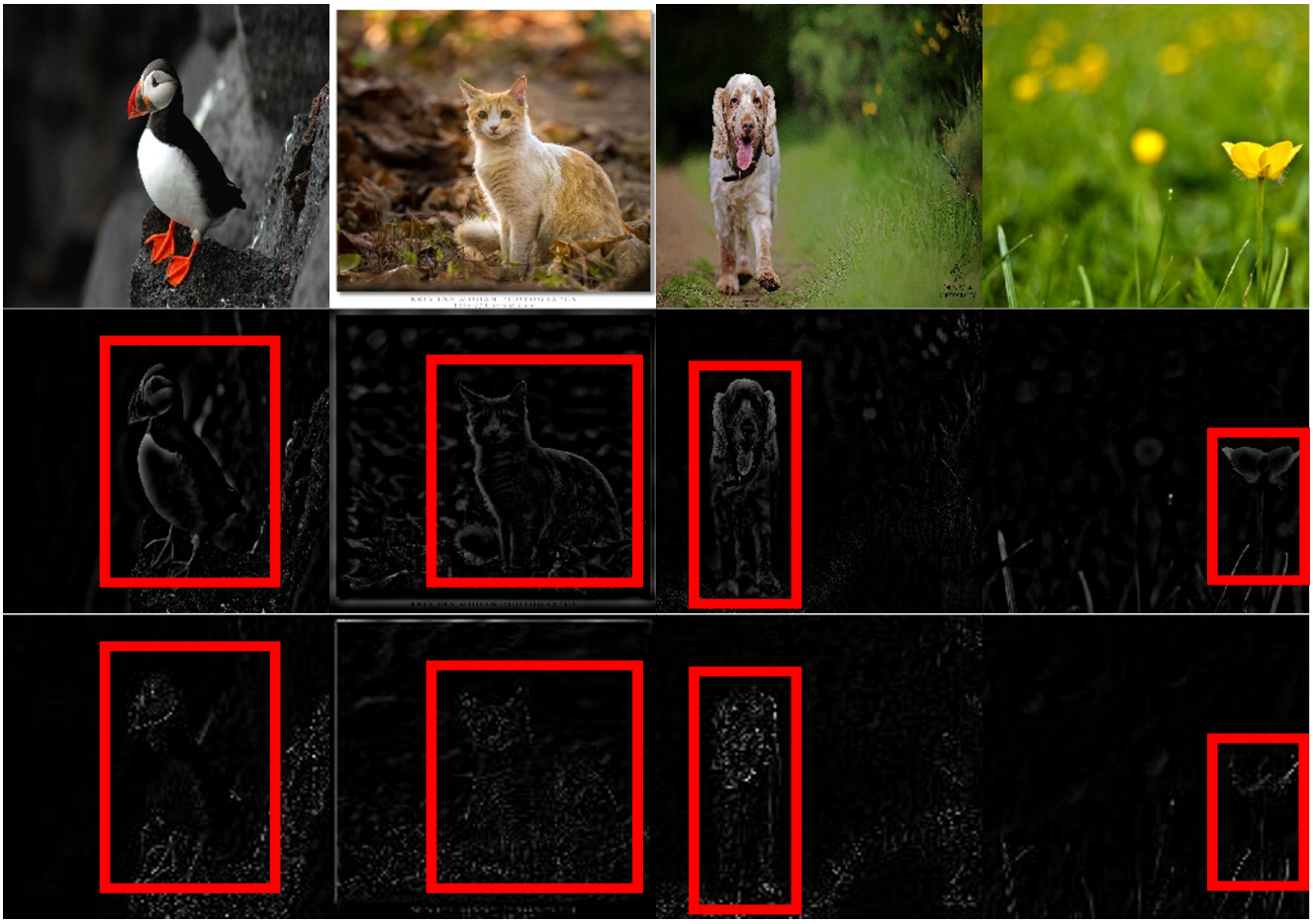
## 1. Comparison with State-Of-The-Art

As discussed in Section 4.5 of the main paper, our proposed framework using NSCT-3 + SURF(FV) and AdobeBING + CNN greatly outperforms the state-of-the-art by Tong *et al*. [1] using DWT + MWHFE and EdgeBoxes + CNN, with accuracy increment of 9.17%. We further present additional qualitative comparison to show the justification of our framework's superior performance in Section 1.1 and 1.2.

### 1.1. NSCT-3+SURF(FV) vs. DWT+MWHFE

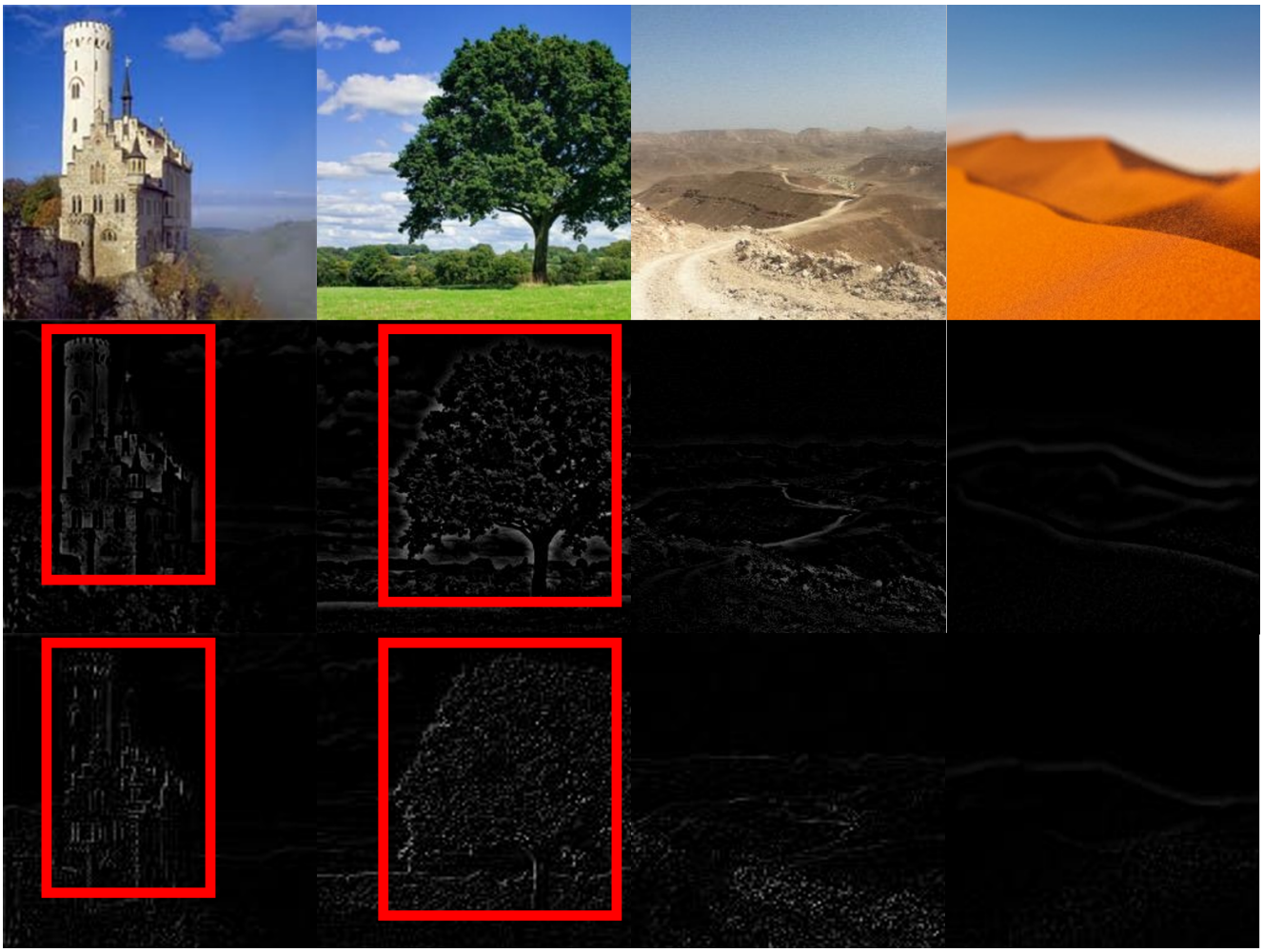
The first comparison is between the different approaches used in the focus modeling. Figure 1 shows examples of close-up views misclassified by the focus cue of [1], but is correctly classified by our proposal. The NSCT-3 used in our proposal produces more sophisticated high frequency signals that preserves object appearance, whereas the DWT loses much of the details, as shown in the second and last rows of Fig. 1 respectively. Particularly, the appearance of objects in the bounding boxes of NSCT-3 are well defined, while the DWT looks like scatter signals.

The same observation is seen in the high frequency decompositions for distant view images shown in Fig. 2. In Fig. 2a - 2b, both NSCT-3 and DWT is able to maintain relatively understandable object features and is central to the image, hence could have confused the MWHFE extractor used by [1]. On the other hand, our proposed framework uses SURF which is a higher level feature representation that is able to partially differentiate the close-up and distant view based on the object texture. Even in Fig. 2c - 2d where the images are clearly distant views with distributed high frequency details, DWT + MWHFE misclassified them while NSCT-3 + SURF(FV) is able to maintain robustness.



(a) (b) (c) (d)

Figure 1: Examples of images misclassified by the focus cue of [1], and their respective frequency decompositions using NSCT-3 (second row), and DWT (last row).



(a) (b) (c) (d)

Figure 2: Examples of images misclassified by the focus cue of [1], and their respective frequency decompositions using NSCT-3 (second row), and DWT (last row).

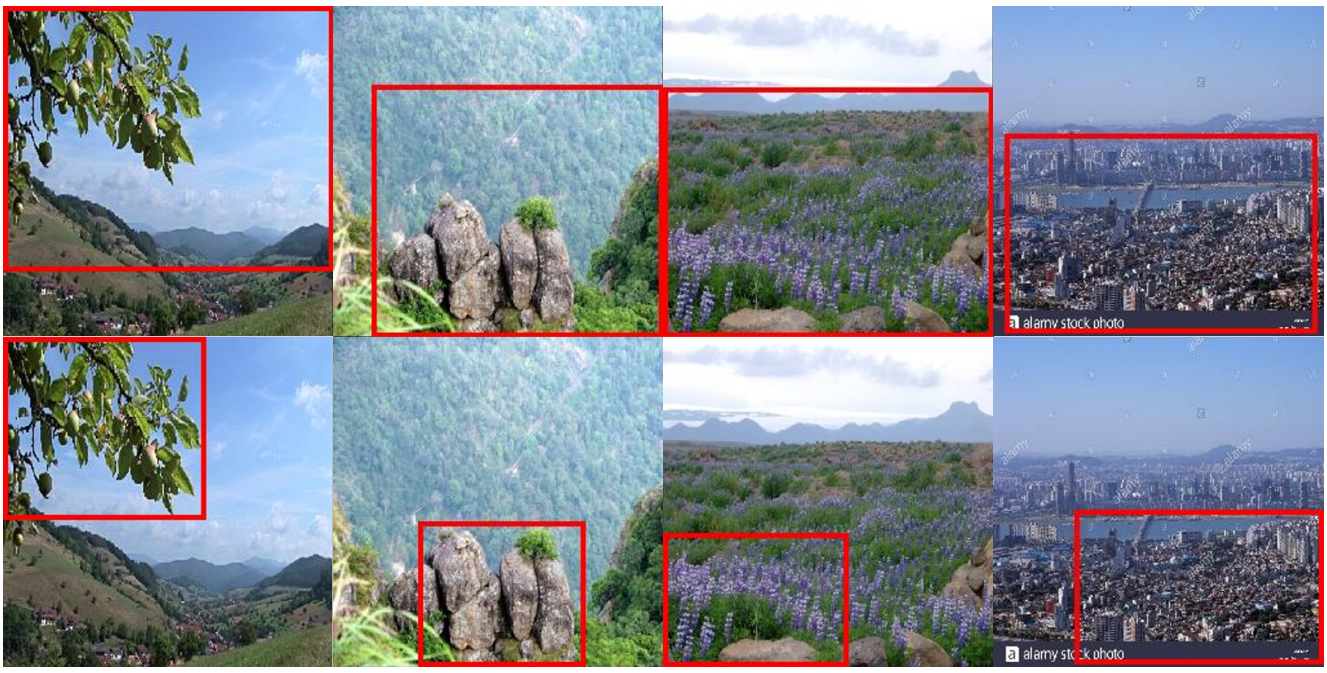
### 1.2. AdobeBING+CNN vs. EdgeBoxes+CNN

This section shows the qualitative comparison between our proposed AdobeBING+CNN for the scale modeling with EdgeBoxes+CNN [1]. As the scale modeling is divided into two components, we first looking into the spatial evaluation using the object proposal methods, AdobeBING and EdgeBoxes. As both methods implement different assumptions, they produce distinctly different bounding box proposals as seen in Fig. 3. The main objective of the spatial evaluation is to classify distant views based on small bounding box proposals. In this stage, it is crucial that small bounding boxes are not mis-proposed for close-up view images as they remain as a misclassification by the scale model. However, a big bounding box proposed for distant views will be analyzed by conceptual size evaluation, hence the classification can be rectified. As seen in Fig. 3, the bounding box proposed by AdobeBING is more precise, while EdgeBoxes proposed small boxes that classifies these images as distant view, subsequently degrading the overall framework's performance.

On the other hand, for the conceptual evaluation, both our method and [1] uses the same CNN architecture. Hence, the performance of this component relies on the area bounded by the object proposal methods. Not only does the EdgeBoxes mis-propose small bounding boxes in close-up view images (Fig. 3), the tight bounding boxes also localizes objects that confuses the conceptual classifier, such as the leaves in Fig. 4a and the rocks in 4b. However, our proposed framework does not suffer from this as the area proposed by AdobeBING has wider perspectives that assists the CNN for accurate classification.

(a) (b) (c) (d)

Figure 3: Examples of images misclassified by the scale cue of [1] and their bounding box proposals using AdobeBING (top row), and EdgeBoxes (bottom row).




(a) (b) (c) (d)

Figure 4: Examples of images misclassified by the scale cue of [1] and the bounding box proposed by AdobeBING (top row), and EdgeBoxes (bottom row) used for conceptual classification.

## 2. Tourist Spots Photos

In this section, we present additional information regarding the selected tourist spots for the analysis described in Section 2 of the main paper. Among the many tourist spots found throughout the world, the 12 spots were selected based on popularity, objectivity, and generality, where 10 are high rating spots (above 9.0), and 2 are lower rating ones, by *TripAdvisor.com*. The selected spots are listed below and shown in Fig. 5.

- **High rating spots.** Bryce Canyon (United States of America), Chichen Itza (Mexico), Colosseum (Italy), Eiffel Tower (France), Fushimi Inari Taisha (Japan), Golden Gate Bridge (United States of America), Opera House (Australia), Parthenon (Greece), Taj Mahal (India), and Tower Bridge (United Kingdom).
- **Lower rating spots.** Stonehenge (United Kingdom), and The Little Mermaid (Denmark).

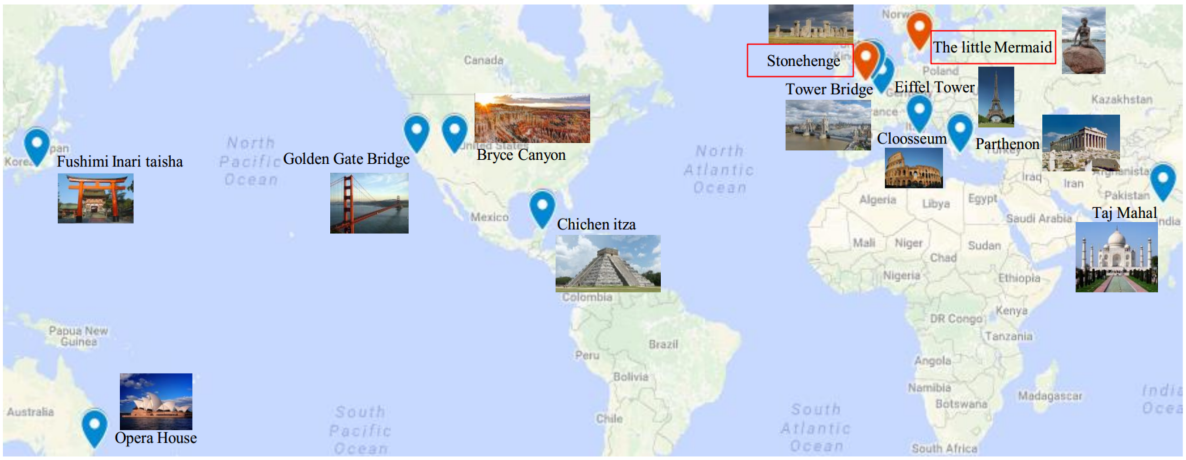



Figure 5: The 12 selected tourist spots for analysis.

After identifying these locations, 2000 geo-tagged photos of each spot were obtained from *Flickr.com* for a total of 24,000 photos, as pointed out in the main paper's Section 2 (line 264-266). In order to ensure a reliable photo set of the specified locations, they were downloaded based on their geo-locations that are in the radii of the respective spots, as seen in Fig. 6. The radius of each spot is determined at a distance where the density of tourist-taken photos beyond that region becomes much sparser. This is affected by the structures found in the locations, for example, the radii of the Golden Gate Bridge and Bryce Canyon are set at 1.60 km and 1.50 km respectively, more than double the radii of the other locations. The Golden Gate Bridge has the length of approximately 2.70 km, therefore, many visitors would capture photos further away from the structure to be able to capture the whole bridge in the photo. Whereas, the Bryce Canyon is a national park covering a large area, hence visitors would capture photos around a wide space. On the other hand, the other locations concentrates on comparatively smaller structures, such as The Little Mermaid statue, Opera House building, and Colosseum building to name a few, that would gather tourists to take photos within a smaller radius.

Figure 7-9 show the distribution maps of the photos obtained from these spots, with the distant/panoramic view and close-up/local view indicated with red and black dots respectively, and some example photos of each spot as well. Based on these distribution maps, a quick impression can be derived from the clear red gathering dots in Fig. 8-9 that high rating spots provide more distant view photos. Whereas, the maps of the two lower rating spots in Fig. 7 show almost equal distribution of distant view (red) and close-up view (black) photos. Subsequently, this enabled us to deduce the relationship between the ratio of view photos and the quality of the tourist spot that motivated our hypothesis that is justified in the main paper. Additionally, the example images of each spot also show variations of close-up images with and without focus and fringe attributes, and distant views with variety of objects with different spatial sizes and conceptual scales.

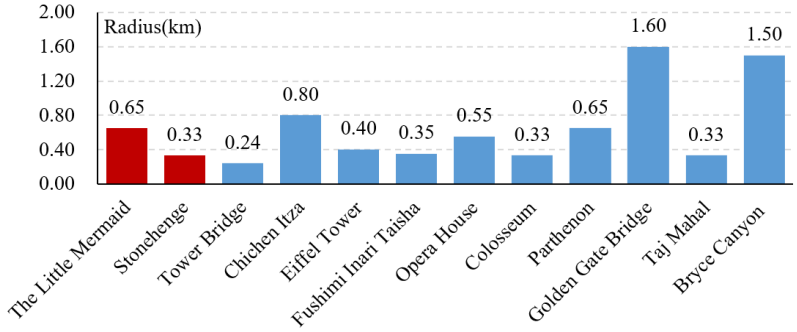



Figure 6: The radii of the 12 selected tourist spots.


Figure 7: Distribution maps of distant and close-up view photos (left) and their examples (right) (Lower rating tourist spots: The Little Mermaid and Stonehenge) .

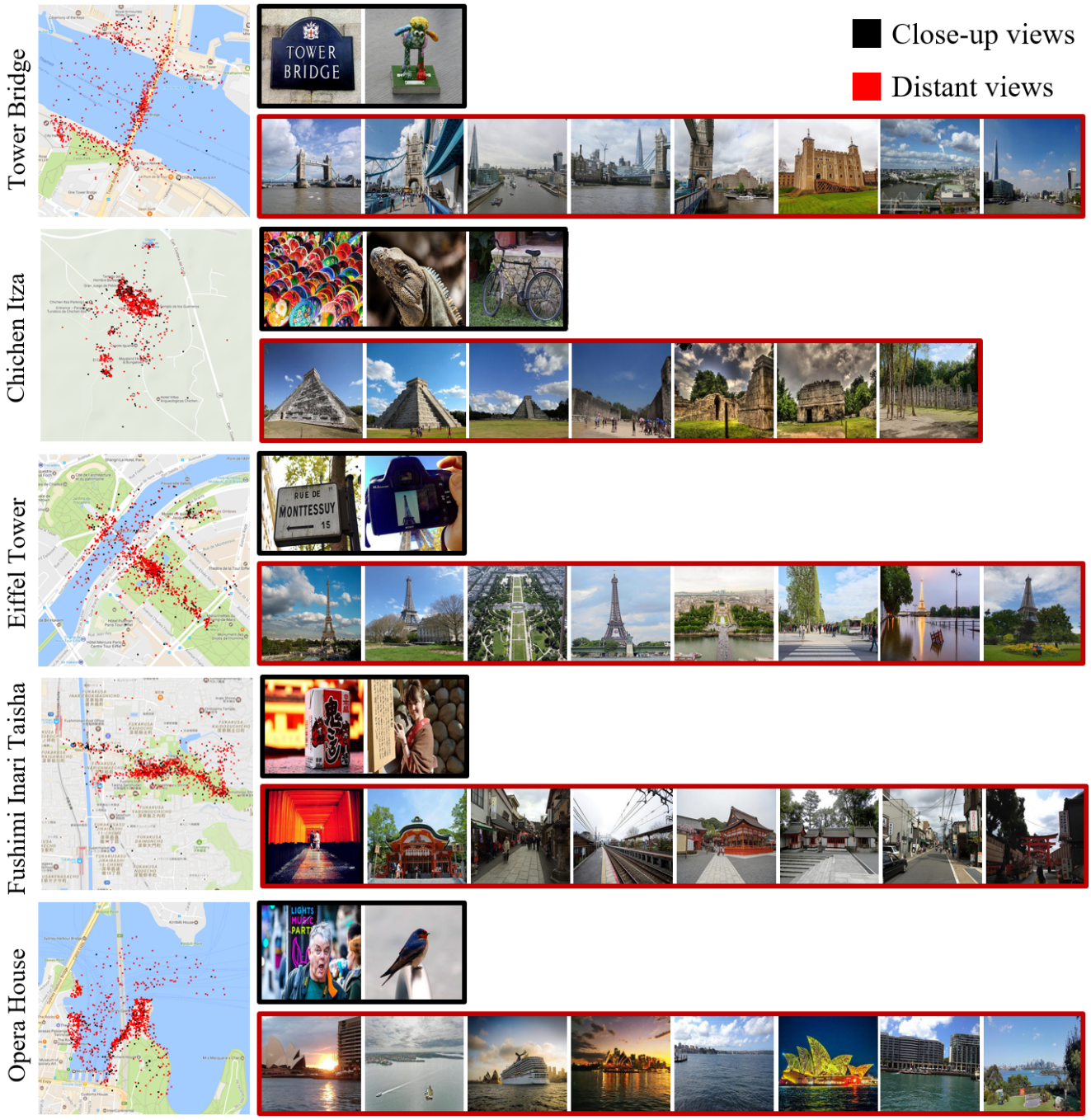



Figure 8: Distribution maps of distant and close-up view photos (left) and their examples (right) (High rating tourist spots: Tower Bridge, Chichen Itza, Eiffle Tower, Fushimi Inari Taisha, and Opera House) .

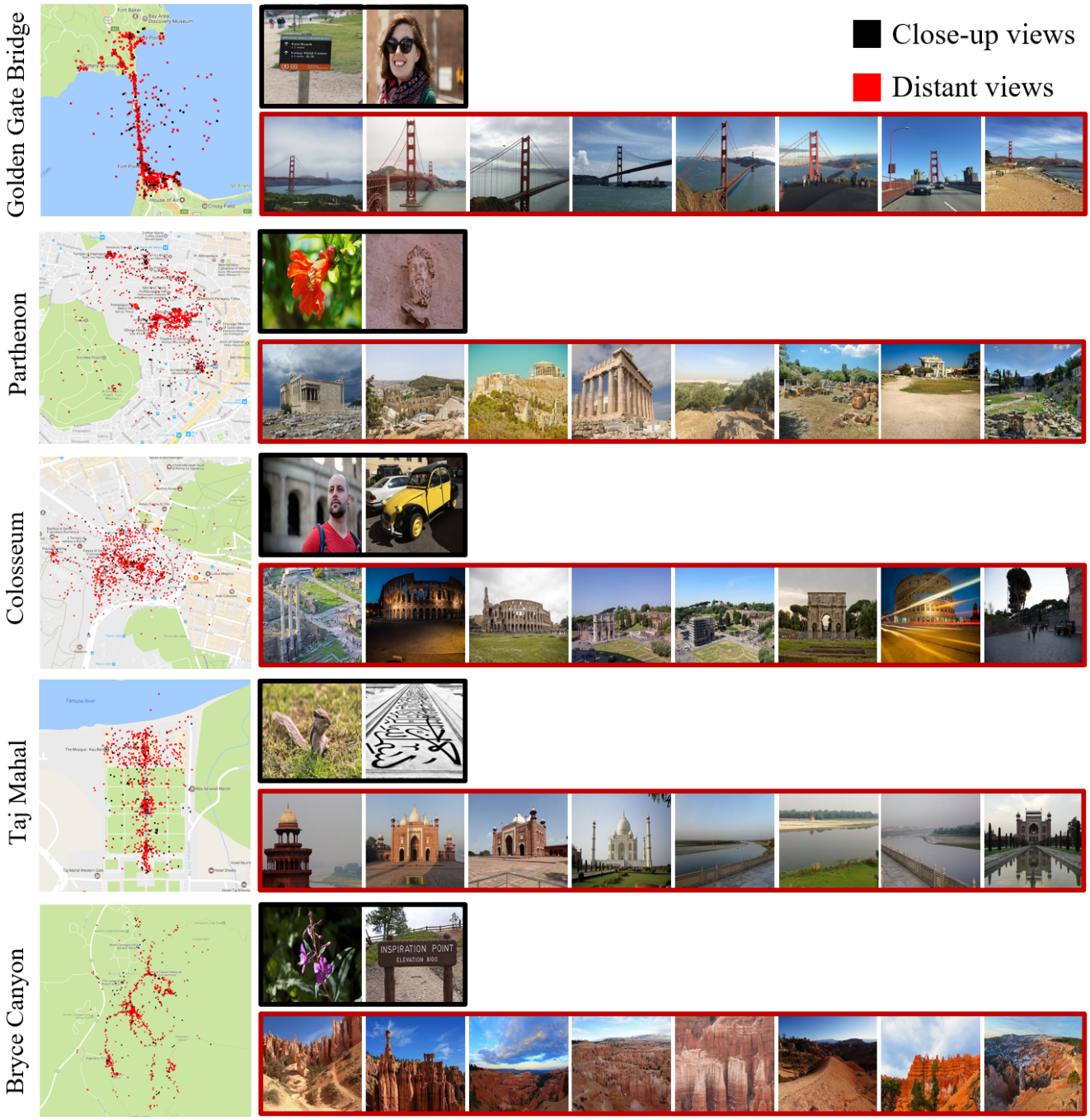



Figure 9: Distribution maps of distant and close-up view photos (left) and their examples (right) (High rating tourist spots: Colosseum, Parthenon, Golden Gate Bridge, Taj Mahal, and Bryce Canyon).

## References

[1] S. Tong, Y. P. Loh, X. Liang, and T. Kumada. Visual attention inspired distant view and close-up view classification. In *ICIP*, 2016. 1, 2, 3, 4, 5