# Supplemental Material: Learning a Dynamic Map of Visual Appearance

Tawfiq Salem
Miami University

Scott Workman
DZYNE Technologies

Nathan Jacobs
University of Kentucky

## 1. Dynamic Visual Attribute Maps

We show additional dynamic attribute maps rendered from our model. See Figure 1 for examples of the *sunny* attribute and Figure 2 for examples of the *stressful* attribute. For both attributes, we show our approach (*sat+time+loc*) and a baseline that does not incorporate location as an input (*sat+time*). For each, we specified the time of day as 4pm, and show three different months. In both models, we observe trends that match our expectations. For example, there tends to be more sunshine at 4pm in July than in January. However, the *sat+time+loc* model does a better job of capturing large-scale spatial trends, such as the difference between the *sunny* attribute in the north and south during January and April.

## 2. Application: Image Localization

We evaluated the accuracy of our approach for the task of image geolocalization (Table 2 in the main paper). To summarize our method, we extracted the visual attributes of a query image and compared them against the visual attributes of an overhead image reference database, computed using the timestamp of the query image. To support this experiment, we created a new evaluation dataset that includes timestamps. The results show that our model, *sat+time+loc*, performs the best using all scoring strategies.

In Figure 3 we show qualitative localization results generated by our approach. For this experiment, we used $488\,224$ overhead images from CVUSA as our reference database. The heatmap represents the likelihood that an image was captured at a specific location, where red (blue) is more (less) likely. Additionally, we compare the different scoring strategies on each row. Similar to our quantitative results, using the *Combine* score produces heatmaps that more closely match the true location of the ground-level image.

## 3. Application: Metadata Verification

For time verification accuracy, Table 3 in the main paper demonstrates that our approach, *sat+time+loc*, outperforms all baseline methods. In Figure 4 and Figure 5, we show additional qualitative results for this task. The heatmaps reflect the distance between the visual attribute extracted from the ground-level image and the predicted attributes from the overhead image (varying the input time). This results in a distance for each possible time. The true capture time is indicated by the red $X$. As observed, our approach more accurately estimates the capture time of the ground-level image.

## 4. Discussion

Our model combines overhead imagery, time, and geographic location to predict visual attributes. We have demonstrated the superiority of this combination, but we think there are several questions that naturally arise when considering our model. Here we provide answers, which we believe are supported by the evaluation.

**Why do we need overhead imagery when it just depends on the location?** If our model was only dependent on geographic location, then we would need to learn a mapping between geographic location and the visual attribute. Consider something as simple as, "does this geographic location contain a road?". This would be a very complicated function to approximate using a neural network and we have seen that it does not work well. In contrast, it is relatively easy to estimate this type of information from the overhead imagery.

**Why do we need to include geographic location if we have overhead imagery?** We think it makes it easier to learn larger scale trends, especially those that relate to time. For example, the relationship between day length and latitude. If we didn't include latitude we would have to estimate it from the overhead imagery, which would likely be highly uncertain.

**Why don't we need an overhead image for each time?** The overhead image provides information about the type of

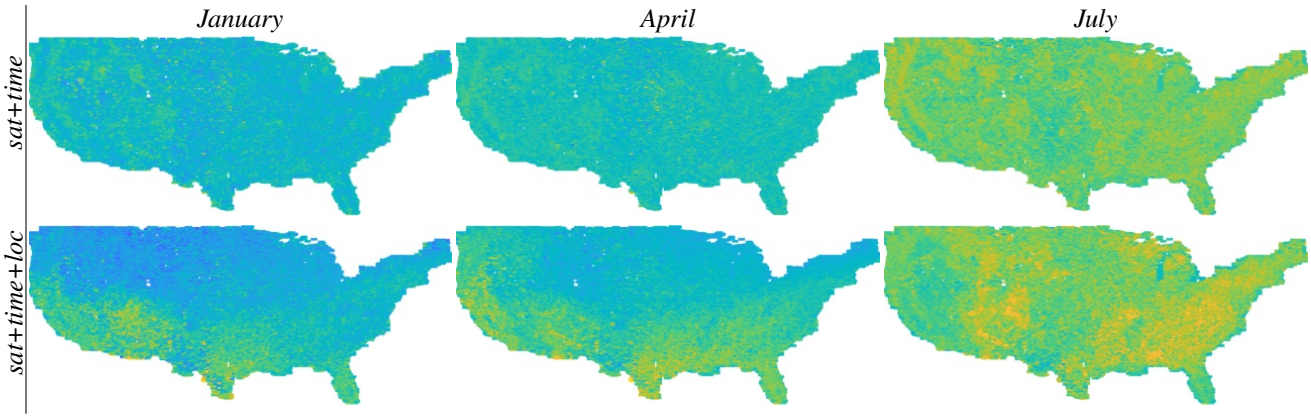


Figure 1: Dynamic visual attribute maps over time for the transient attribute *sunny*. In each, yellow (blue) corresponds to a higher (lower) value for the corresponding attribute.

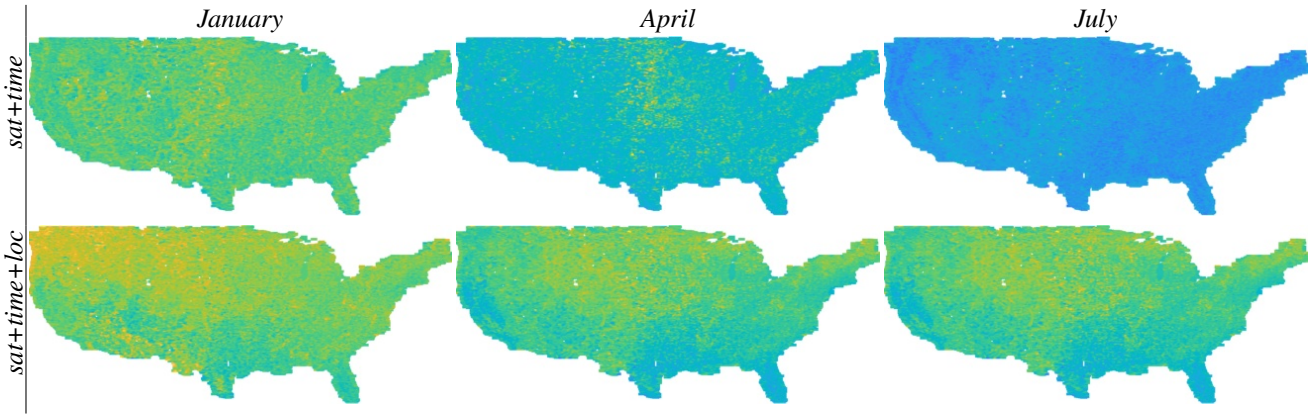


Figure 2: Dynamic visual attribute maps over time for the transient attribute *stressful*. In each, yellow (blue) corresponds to a higher (lower) value for the corresponding attribute.

place. This is unlike a satellite weather map, which would tell us what the conditions are at a particular time. While we do lose some information, this is accounted for by including geographic location and time as additional context. In practice it is best if the overhead image is captured relatively close in time (within a few years) to account for major land use and land cover changes.

**Limitations** One of the limitations of this study is the reliance on social media imagery. This means that our visual appearance maps will exhibit biases about when people prefer to take pictures, or are willing to share pictures. For example, we are likely undersampling cold and stormy weather conditions and oversampling sunsets. This is part of the motivation for incorporating imagery from the AMOS dataset. This, at least, doesn't have the same temporal bias because the webcams collect images on a regular interval, regardless of conditions. However, these are sparsely distributed spatially and, at least in our dataset, outnumbered by the social media imagery. Despite this, we were still able to demonstrate effective learning and this problem could be overcome as more data becomes available. Another limitation is that our current approach cannot model longer-term, year-over-year trends in visual attributes. This results because our representation of time only reflects the month and time of day, not the year.

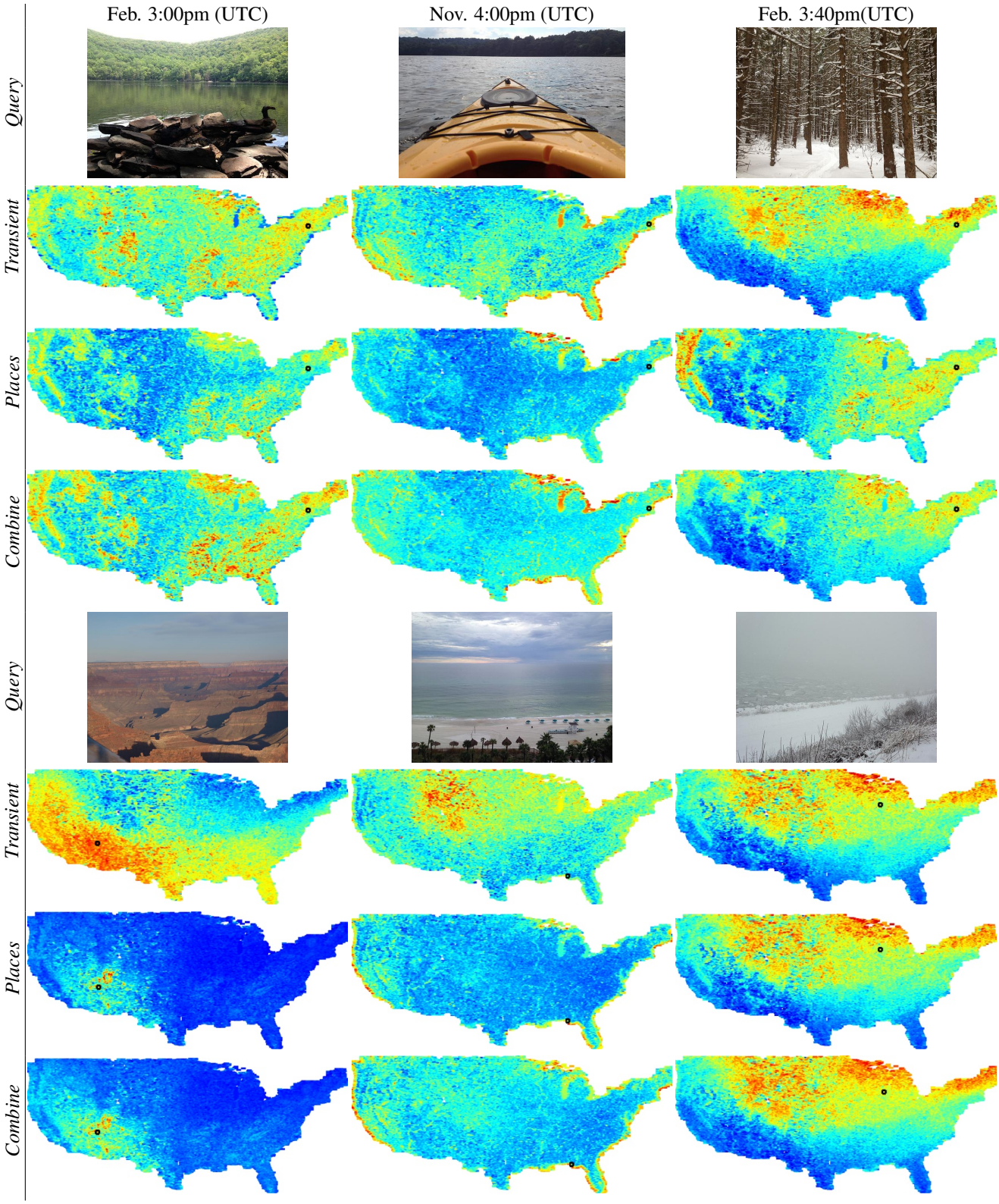


Figure 3: Given a query ground-level image (top), we show localization results (bottom) for different scoring strategies, visualized as a heatmap. Red (blue) represents a higher (lower) likelihood that the image was captured at that location.

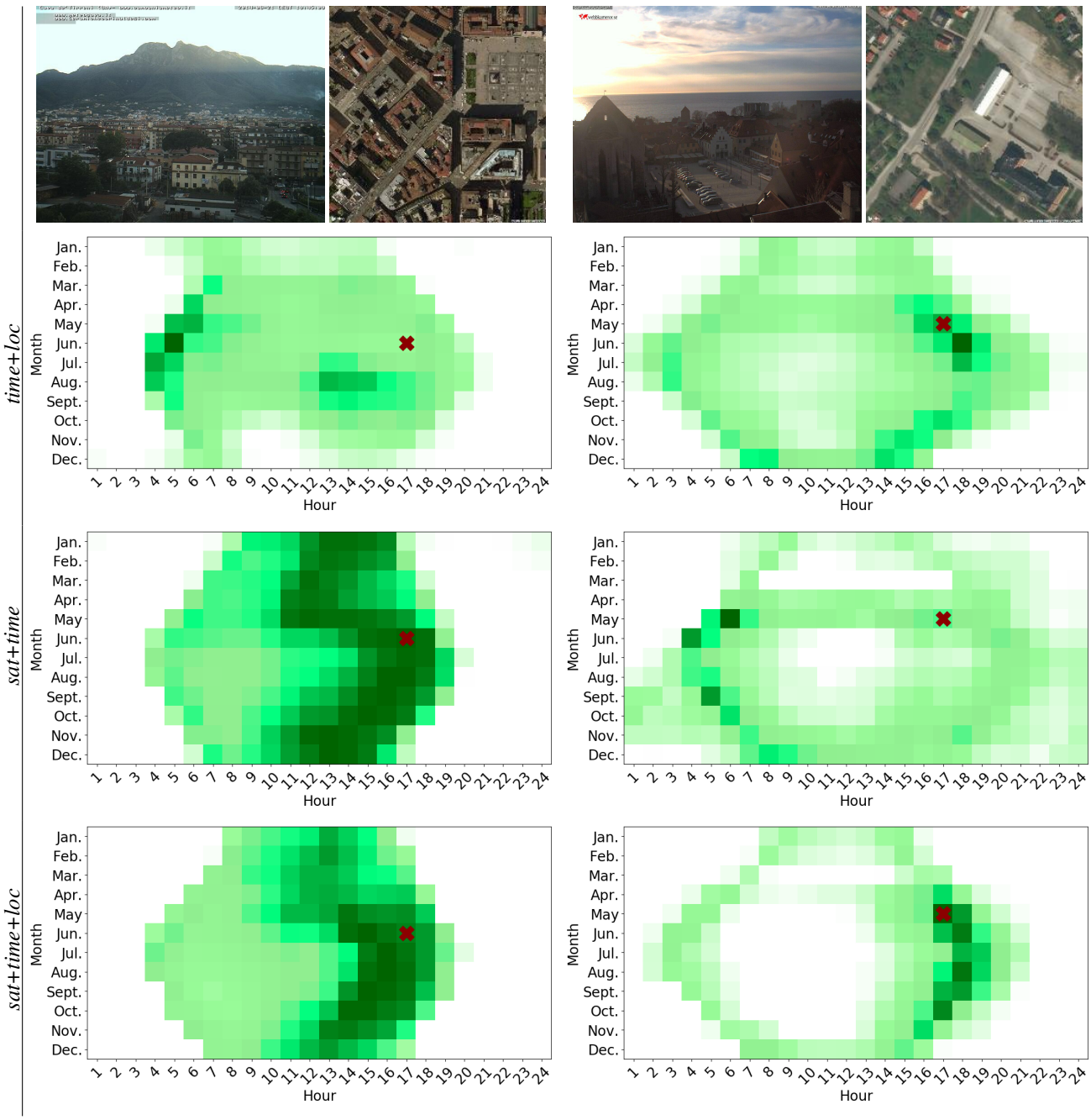


Figure 4: Different examples highlighting temporal patterns learned by our model. (top) For each example, we show the original image and the overhead image of its location. (bottom) For every possible hour and month, we use the different models (left) to predict the visual attributes. The heatmaps show the distance between the true and predicted visual attributes, with dark green (white) representing smaller (larger) distances.

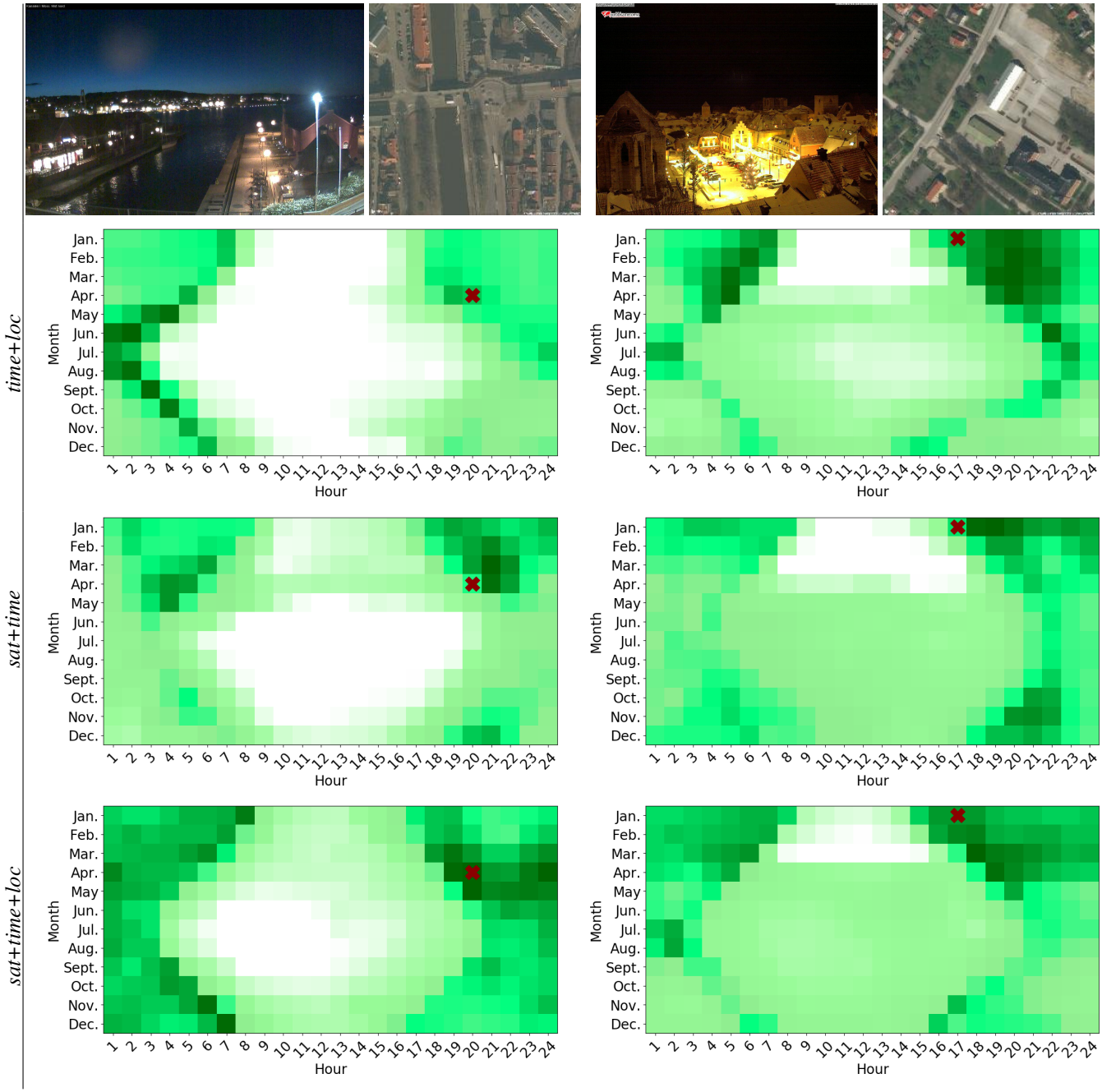


Figure 5: Different examples highlighting temporal patterns learned by our model. (top) For each example, we show the original image and the overhead image of its location. (bottom) For every possible hour and month, we use the different models (left) to predict the visual attributes. The heatmaps show the distance between the true and predicted visual attributes, with dark green (white) representing smaller (larger) distances.