Learning a Dynamic Map of Visual Appearance

Tawfiq Salem
Miami University

Scott Workman
DZYNE Technologies

Nathan Jacobs
University of Kentucky

Abstract

The appearance of the world varies dramatically not only from place to place but also from hour to hour and month to month. Every day billions of images capture this complex relationship, many of which are associated with precise time and location metadata. We propose to use these images to construct a global-scale, dynamic map of visual appearance attributes. Such a map enables fine-grained understanding of the expected appearance at any geographic location and time. Our approach integrates dense overhead imagery with location and time metadata into a general framework capable of mapping a wide variety of visual attributes. A key feature of our approach is that it requires no manual data annotation. We demonstrate how this approach can support various applications, including image-driven mapping, image geolocalization, and metadata verification.

1. Introduction

Recent concern about "fake news" has lead to a significant interest in verifying that imagery is real and unmanipulated. Early work on this problem focused on low-level image statistics [4, 6], but this approach is unable to detect the falsification of image metadata. Matzen and Snavely [21] introduce an approach for finding anomalous timestamps, but their method is based on visual correspondences and requires overlapping imagery. Recent work has begun to look at this problem more thoroughly, with new datasets [10] and proposals for comprehensive systems [3]. However, no previous work provides the dynamic map of visual attributes that is necessary for detecting time/location metadata falsification.

We propose to use visual attributes estimated from ground-level images, such as those shown in Figure 1, to learn a dynamic map of visual attributes. Beyond metadata verification, there are numerous applications for such a map, including geolocalizing images, providing contextual information for autonomous vehicles, and supporting further studies on the relationship between the visual environment and human health and happiness [26].

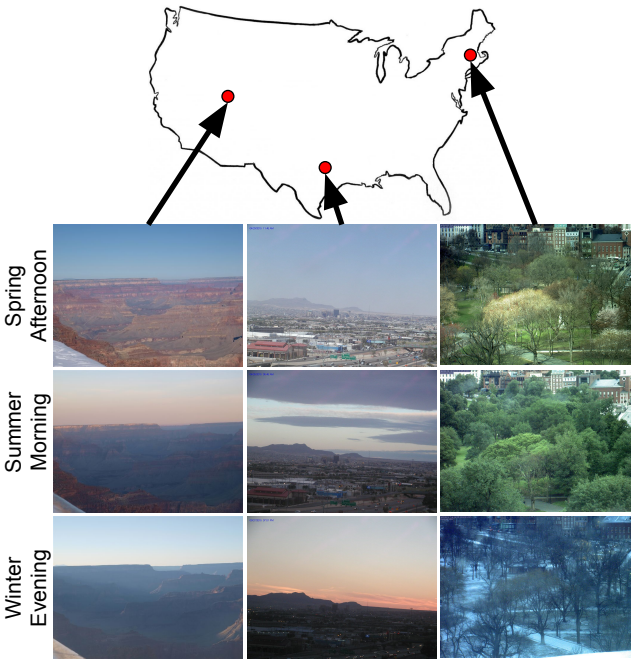


Figure 1: Visual appearance changes dramatically due to differences in location and time. Our work takes advantage of sparsely distributed ground-level image data, with associated location and time metadata, in conjunction with overhead imagery to construct dynamic maps of visual appearance attributes.

Predicting visual attributes directly from location and time is difficult because of the complexity of the distribution. It would, for example, require memorizing the location of every road and building in the area of interest. To overcome this, our model combines overhead imagery with location and time using a multi-modal convolutional neural network. The result is a model capable of generating a worldwide, dynamic map of visual attributes that captures both local and global patterns.

We focus on two visual attributes: the scene category [44], such as whether the image views an attic or a zoo, and transient attributes [15], which consist of time-varying

properties such as sunny and foggy. We selected these because they are well known, easy to understand, and have very different spatiotemporal characteristics. The former is relatively stable over time, but can change rapidly with respect to location, especially in urban areas. The latter has regular, dramatic changes throughout the day and with respect to the season.

Our approach has several useful properties: it does not require any manually annotated training data; it can model differences in visual attributes at large and small spatial scales; it captures spatiotemporal trends, but does not require overhead imagery at every time; and is extendable to a wide range of visual attributes. To evaluate our approach, we created a large dataset of paired ground-level and overhead images each with location and time metadata, which we call Cross-View Time (CVT). Using CVT, we demonstrate the effectiveness of our dynamic mapping approach on several tasks. In each case, our full model, which combines overhead imagery and metadata, is superior.

2. Related Work

Recent advances in computer vision have enabled the estimation of a wide variety of visual attributes, including scene category [44], weather conditions [15], and demographics [7]. As these techniques mature, many application areas have developed that require an understanding of the relationship between visual attributes, geographic location, and time.

2.1. Image-driven mapping

Typically image-based methods for generating maps start by extracting visual attributes from large-scale geotagged image collections and then apply a form of spatial smoothing, such as locally weighted averaging. Examples include methods for mapping land cover [17], scenicness [41], snowfall [34], facial appearance [2], and a variety of other visual attributes [33].

Integrating overhead imagery with image-driven mapping reduces the need for spatial smoothing, resulting in higher quality maps. This has been demonstrated for a variety of visual attributes, including building properties [40], natural beauty [39], scene layouts [42], soundscapes [25], object distributions [9, 24], and land use [28]. Recent work has taken this to the extreme by synthesizing complete ground-level images [5, 23, 42].

In this work, we perform image-driven mapping using overhead imagery, with location and time as additional context, resulting in high-resolution, dynamic maps of visual attributes. Most previous work has either ignored time, or merely used it to filter images outside of a time interval prior to spatial smoothing. Our work is similar to [37], but we focus on mapping visual attributes.

2.2. Image geolocalization

Recently, there has been a significant interest in the problem of image geolocalization, i.e., estimating the geographic location of the camera, or an object in the scene, given visual attributes extracted from the image [11, 35]. More recent work has shown that learning a feature mapping between ground-level and overhead image viewpoints enables image localization in regions without nearby ground-level images [18, 19, 36, 38]. From this work, we see that image geolocalization requires the ability to extract visual attributes from ground-level images and an understanding of the geospatial distribution of these attributes. The former motivates our focus on generating high-quality, dynamic maps of visual attributes.

2.3. Location context aids image understanding

Studies have shown that additional context can aid visual understanding. Tang et al. [29] use the location an image was captured to improve classification accuracy. Luo et al. [20] use overhead imagery as additional context to improve event recognition in ground-level photos. Zhai et al. [43] describe methods for learning image features using location and time metadata. Lee et al. [16] use map data to learn to estimate geo-informative attributes such as population density and elevation. Wang et al. [32] use location information along with weather conditions to learn a feature representation for facial attribute classification. One potential use of our dynamic mapping approach would be as a model of the context needed for such image understanding applications.

3. Cross-View Time (CVT) Dataset

In an effort to support dynamic image-driven mapping, we introduce a new large-scale dataset that contains geotagged ground-level images, corresponding capture time, and co-located overhead images. We refer to our dataset as the Cross-View Time (CVT) dataset. It is similar to previous cross-view datasets [31, 36, 38], but ours is unique in providing timestamps for all images.

Our dataset is built from two sources of ground-level images. The first source is the Archive of Many Outdoor Scenes (AMOS) [13], a collection of over a billion images captured from public outdoor webcams around the world. This subset [22] includes images captured between the years 2013 and 2014, from 50 webcams, totaling 98 633 images. Each image is associated with the location of the webcam and a timestamp (UTC) indicating when the image was captured. The second source is a subset of the Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M) [30]. This subset [43] contains geotagged outdoor images, with timestamps, captured by smartphones.

We combined images from both of these sources to form

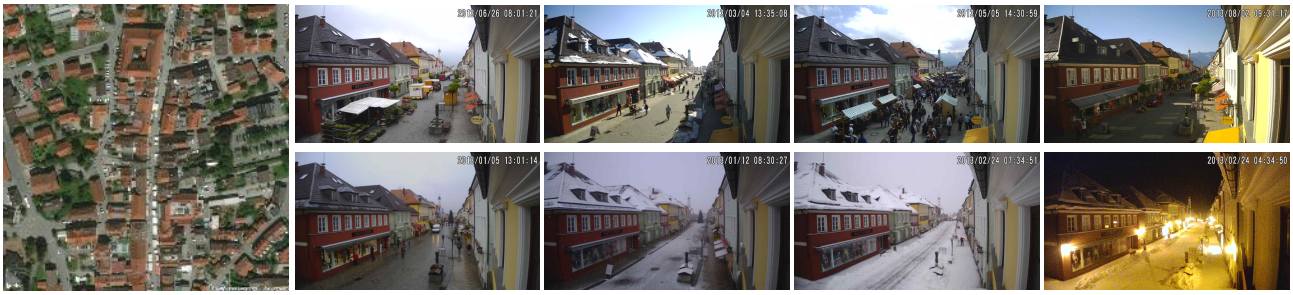


Figure 2: An overhead image and the corresponding ground-level images from our CVT dataset.

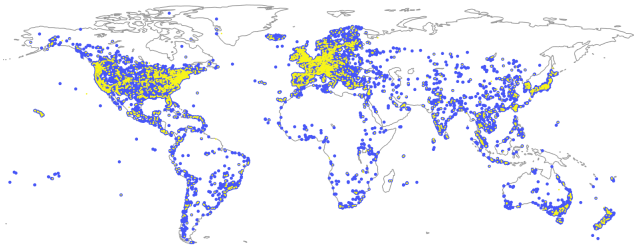


Figure 3: The spatial distribution of the CVT dataset. The blue (yellow) dots represent the training (testing) data.

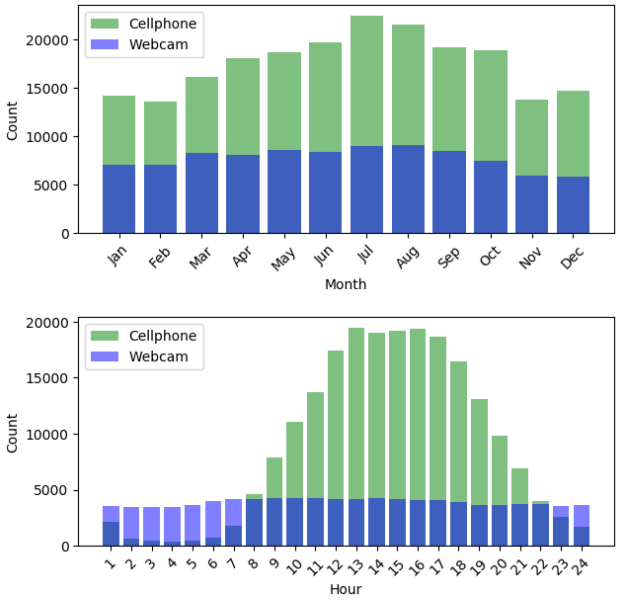


Figure 4: The temporal distribution of the CVT dataset.

a hybrid dataset containing 305 011 ground-level images. For each image, we also downloaded an orthorectified overhead image from Bing Maps (800×800, 0.60 meters/pixel), centered on the geographic location. We randomly selected 25 000 ground-level images, and the corresponding overhead images, and reserved them for testing. This resulted in a training dataset of 280 011 image pairs. Figure 2 shows example images from the CVT dataset.

Figure 3 shows the spatial distribution of the training images (blue dots) and testing images (yellow dots). Visual analysis of the distribution reveals that the images are captured from all over the world, with more images from Europe and the United States. Furthermore, examining the capture time associated with each image shows that the images cover a wide range of times. Figure 4 visualizes the distribution over month and hour for both ground-level image sources. We observe that the webcam images are captured more uniformly across time than the cellphone images. The dataset is available at our project website.[1]

4. Dynamic Visual Appearance Mapping

We present a general approach for dynamic visual appearance mapping that could be used to model a broad range of attributes and support many tasks.

[1] https://tsalem.github.io/DynamicMaps/

4.1. Problem Statement

Our objective is to construct a map that represents the expected appearance at any geographic location and time. The expected appearance is defined using a set of visual attributes, which could be low level, such as a color histogram, or high level, such as the scene category. For a given visual attribute, a, such a map can be modeled as a conditional probability distribution, $P(a|t, l)$, given the time, t, and location, l, of the viewer. The distribution $P(a|t, l)$ is challenging to learn because it essentially requires memorizing the Earth and how it changes over time.

We assume we are given a set of ground-level images, $\{I_i\}$, each with associated capture time, $\{t_i\}$, and geolocation metadata, $\{l_i\}$. Furthermore, we assume we have the ability to calculate, or estimate with sufficient accuracy, each visual attribute from all images. The computed visual attributes, $\{a_i\}$, can be considered samples from the proba-

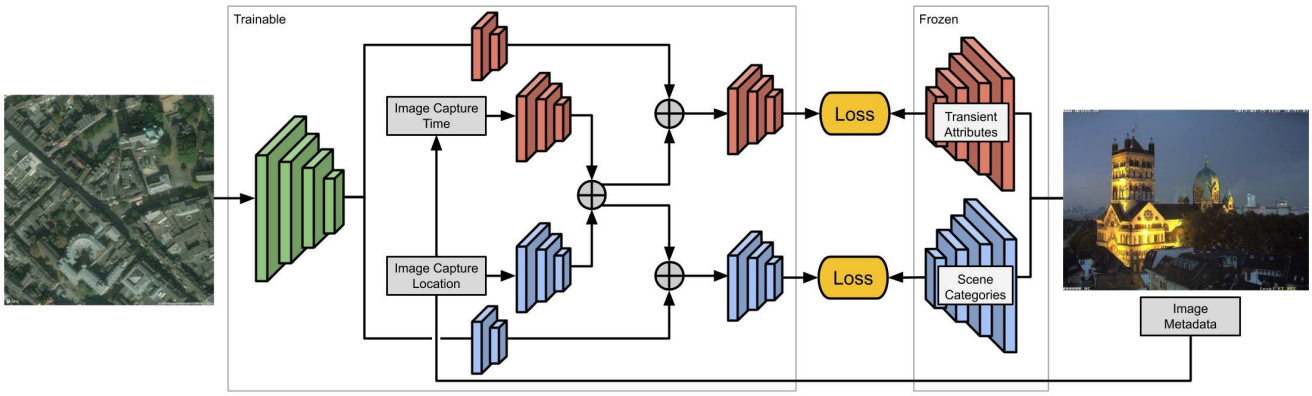


Figure 5: An overview of our network architecture, which includes the network we train to predict visual attributes (left) and the (frozen) networks we use to extract visual attributes from the ground-level images (right).

bility distribution, $P(a|t, l)$, and used for model fitting.

4.2. Approach Overview

To overcome the difficulty of directly modeling $P(a|t, l)$, we also condition the distribution on an overhead image, $I(l)$, of the location. Specifically, we define a conditional probability distribution, $P(a|t, l, I(l))$. In doing so, the network no longer has to memorize, for example, the location of every road, river, and building in the world. We implement this using a mixture of convolutional and fully-connected neural networks to map from the conditioning variables to the parameters of distributions over a visual attribute, $P(a|F(t, l, I(l); \Theta))$, where Θ represents the parameters of all neural networks.

See Figure 5 for an overview of our complete architecture, which, in this case, simultaneously predicts two visual attributes. From the left, we first construct a feature embedding for each conditioning variable using a set of *context* neural networks. We combine these context features to predict the visual attributes using a per-attribute, *estimator* network. From the right, a set of pre-trained networks extract visual attributes from the ground-level images. These networks are only used for extracting visual attributes and are not trained in our framework.

This macro-architecture was carefully designed to balance several criteria. Most importantly, the overhead image is not dependent on time. This means that an overhead image is not required for every timestamp, t, of interest. An overhead image *is* required for each location, but this is not a significant limitation given the wide availability of high-resolution satellite and aerial imagery. In addition, at inference time, feature extraction for the satellite image only needs to happen once, because the extraction process is not time or attribute dependent.

4.3. Network Architecture Details

We propose a novel macro-architecture for modeling a dynamic visual appearance map. In this section, we define the specific neural network architectures and hyper-parameters we used for evaluation.

Visual Attributes We focus on two visual attributes: *Places* [44], which is a categorical distribution over 365 scene categories, and *Transient* [15], which is a multi-label attribute with 40 values that each reflect the degree of presence of different time-varying attributes, such as *sunny*, *cloudy*, or *gloomy*. To extract the *Places* attributes, we use the pre-trained VGG-16 [27] network. To extract the *Transient* attributes, we use a ResNet-50 [12] model that we trained using the Transient Attributes Database [15].

Context Networks The context networks encode every conditioning variable, i.e., time, geographic location, and overhead image, to a 128-dimensional feature vector. For the time and geolocation inputs, we use two similar encoding networks, each consisting of three fully connected layers with a ReLU activation. The layers have 256, 512, and 128 neurons respectively. The geographic location is represented in earth-centered earth-fixed coordinates, scaled to the range $[-1, 1]$. The time is factored into two components: the month of the year and the hour of the day. Each is scaled to the range $[-1, 1]$. For the overhead image, we use a ResNet-50 model to extract the 2048-dimensional feature vector from the last global average pooling layer. This feature is passed to a per-attribute head. Each head consists of two fully connected layers that are randomly initialized using the Xavier scheme [8]. The layers of each head have 256 and 128 neurons respectively, each with a ReLU activation.

Estimator Networks For each visual attribute there is a separate estimator network, with only fully connected layers, that directly predicts the visual attribute. The input for these is the concatenation of the outputs of the context

	Places		*Transient*	
Model	Top-1	Top-5	Within-0.1	Within-0.2
loc (k-NN)	17.68	40.26	50.96	77.10
time (k-NN)	5.84	17.82	48.75	75.77
time+loc (k-NN)	19.08	41.15	51.84	77.51
loc (CNN)	12.70	32.45	48.50	75.45
time (CNN)	4.45	16.91	47.37	75.34
time+loc (CNN)	17.05	35.50	54.69	79.15
sat (CNN)	15.16	38.40	49.87	76.55
sat+loc (CNN)	16.98	41.46	50.57	77.17
sat+time (CNN)	19.66	40.78	56.14	79.79
sat+time+loc (CNN)	**21.58**	**44.00**	**56.91**	**80.55**

Table 1: A comparison of the prediction accuracy of our full approach, *sat+time+loc*, with various baseline models.

networks. For each estimator network, the first two layers (which have ReLU activations) contain 256 and 512 neurons, respectively. The third layer represents the output, with the number of neurons depending on the visual attribute. In this case, there are 365 output neurons for the *Places* estimator, with a *softmax* activation, and 40 for the *Transient* estimator, with a *sigmoid* activation.

4.4. Implementation Details

We jointly optimize all estimator and context networks with losses that reflect the quality of our prediction of the visual attributes extracted from ground-level images, $\{I_i\}$. For the *Places* estimator, the loss function is the KL divergence between attributes estimated from the ground-level image and the network output. For the *Transient* estimator, the loss function is the mean squared error (MSE). These losses are optimized using Adam [14] with mini-batches of size 32. We applied L_2 regularization with scale 0.0005 and trained all models for 10 epochs with learning rate 0.001.

All networks were implemented using TensorFlow [1] and will be shared with the community. Input images are resized to 224×224 and scaled to $[-1, 1]$. We pre-trained the overhead context network to directly predict *Places* and *ImageNet* categories of co-located ground-level images, minimizing the KL divergence for each attribute. The weights are then frozen and only the added attribute-specific heads are trainable.

For extracting *Transient* attributes from the ground-level images, we train a ResNet-50 using the Transient Attributes Database [15] with the MSE loss. The weights were initialized randomly using the Xavier scheme, and optimized using Adam [14] until convergence with learning rate 0.001 and batch size 64. The resulting model achieves 3.04% MSE on the test set, improving upon the 4.3% MSE presented in the original work [15].

5. Evaluation

We evaluate our approach using the CVT dataset quantitatively, qualitatively, and on a variety of applications. We use Top-1 and Top-5 classification accuracy as the metric for evaluating quality of the *Places* attribute predictions. For the *Transient* attribute we use the percent of attribute predictions within a threshold (0.1 or 0.2) of the ground truth. In both cases, these are averaged across the full test set.

5.1. Exploratory Dataset Analysis

To better understand the relationship between location, time, and these attributes, we conducted a preliminary study without using overhead imagery. For the *Places* attribute, we use a k-NN classifier ($k = 30$) to explore this relationship. As features we used time (linear) and latitude/longitude (degrees). We scaled the time using grid-search to optimize the accuracy when using all features. The resulting classifier obtained 19.08% accuracy on the test set (see Table 1). If we remove the time feature, the accuracy drops a small amount to 17.68%. If we remove both location features, the accuracy is 5.84%, which is better than ignoring all features (1.96%). From this, we can see that the *Places* attribute is highly dependent on location but less-so on time. We were surprised that the time feature by itself resulted in such high accuracy. We suspect that this is due to differences in the types of pictures taken at different times of year.

For the *Transient* attributes, we used a similar setup. The only change was using a k-NN regression model. Table 1 shows that the difference between features is less dramatic than it was for the *Places* attributes. Instead, we focus on the impact of removing the location and time features on the individual attributes. When removing the location feature, we found, for example, that the accuracy for some attributes went down more than 6% (e.g., *busy*, *fog*, *gloomy*) while for others it went up more than 2% (e.g., *dawndusk*, *dark*, *night*). For the time feature, we found that the accuracy went down for all attributes, with some going down significantly (e.g., *winter*, *snow*, *lush*) but others only marginally (e.g., *rain*, *sunrisesunset*, *sentimental*).

These results highlight that the relationship between visual attributes, location, and time is complex and that our dataset enables us to translate intuitive notions into concrete experimental results.

5.2. Quantitative Evaluation

We trained several variants of our full model, *sat+time+loc*. For each, we omit either one or two of the conditioning variables but retain all other aspects. We use the same training data, training approach, and micro-architectures. In total, we trained six baseline models: *loc*,

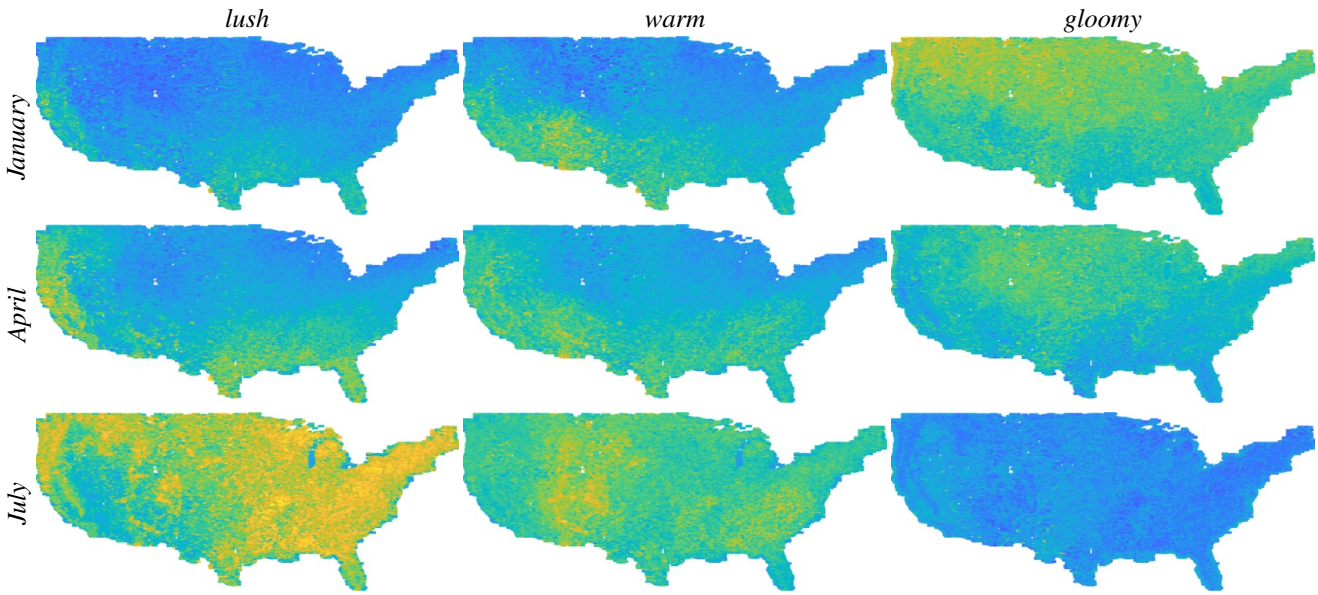


Figure 6: Dynamic visual attribute maps for different *Transient* attributes. In each, yellow (blue) corresponds to a higher (lower) value for the corresponding attribute. Each attribute exhibits unique spatial and temporal patterns, which closely match the authors' personal travel experiences.

time, *sat*, *time+loc*, *sat+loc*, and *sat+time*. We evaluate the accuracy of all methods on the test set.

Table 1 shows the accuracy for all approaches on both visual attributes. We find that our method has the highest accuracy. However, the ranking of baseline models changes depending on the visual attribute. For example, the accuracy for the *sat+loc* model is relatively worse for the *Transient* attribute than the *Places* attribute. This makes sense because the former is highly dependent on when an image was captured and the latter is more stable over time. We also note the significant improvement, for both attributes, obtained by including overhead imagery in the model. For example, the *time+loc* model is significantly worse than our full model.

5.3. Examples of Visual Attribute Maps

Figure 6 shows several example attribute maps rendered from our model. To construct these we use the CVUSA dataset [38], which contains overhead imagery across the continental United States. Specifically, we use a subset of 488 243 overhead images associated with the Flickr images in the dataset. For each overhead image, we compute visual attributes using our full model, *sat+time+loc*. We specify the time of day as 4pm, and vary the month.

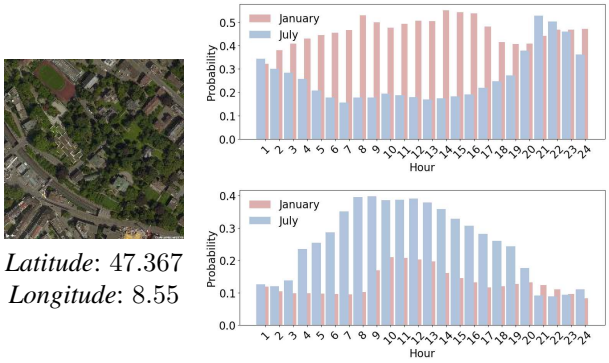
The trends we observe are in line with our expectations. For example, for the transient attribute *lush*, which refers to vegetation growing, January has low values (blue) in the northernmost regions. However, the highest estimates (yellow) include regions like Florida and California. The lushness estimate progressively increases from January through April, achieving its highest value in July. Similarly, the *warm* attribute is highest in the southwest during both winter and spring, but reaches higher overall values in the summer months. Meanwhile, the *gloomy* attribute is highest during winter, with a bias towards the Pacific Northwest, and decreases during the summer.

Figure 7 shows an example of how the estimated attribute varies over time. Our proposed model captures changes in the different attributes not only over months of the year but also over hours of the day. In Figure 7 (top, right) the *cold* attribute during a day in January is higher than a day in July, whereas in Figure 7 (bottom, right) the *warm* attribute is opposite. These results demonstrate that our model has captured temporal trends.

6. Applications

We show how our dynamic mapping approach can be used to support three image-understanding applications: localization, retrieval, and metadata verification. Together, they demonstrate that combining overhead imagery, location, and time is critical for correctly modeling the dynamic distribution of visual attributes.

A key component of each application is computing the distance between the visual attributes of a ground-level image and the visual attributes predicted by our model. For the *Places* attribute we use the KL divergence and for the *Transient* attribute we use the L_2 distance. We also define *Combine* which is a weighted average of these two, with λ



Latitude: 47.367
Longitude: 8.55

Figure 7: For a given location and corresponding overhead image, (top, right) shows the predictions from our model for the *cold* attribute. Similarly, (bottom, right) shows the warm attribute. Both examples show that our model has learned dynamic patterns of visual attributes.

	Transient		*Places*		*Combine*	
Context	Top-1%	Top-5%	Top-1%	Top-5%	Top-1%	Top-5%
sat	4.80	15.30	18.80	42.00	21.60	42.60
sat+loc	5.50	15.40	23.00	45.40	23.90	45.00
sat+time	13.10	22.50	23.90	43.60	24.90	44.00
sat+time+loc	13.70	25.00	28.70	47.60	**31.20**	**49.30**

Table 2: Localization accuracy of different models and distance measures.

as the weight for *Places* and $1 - \lambda$ for *Transient*. The value of λ is selected empirically for each application.

6.1. Application: Image Localization

We evaluated the accuracy of our models on the task of image geolocalization, using a set of 1000 ground-level query images randomly sampled from the test set. To localize an image, we first extract its visual attributes. Then, we predict the visual attributes for all 1000 overhead images. As context, we use the location of the corresponding overhead image and the capture time of the ground-level image. We compute the distance between these predicted attributes and the attributes extracted from the image. We use $\lambda = 0.58$ when computing the *Combine* distance.

Table 2 shows the results of this experiment. Each number represents the percentage of query images that were correctly localized within the Top-k% of candidate locations. For a given threshold, a higher percentage localized is better. This experiment shows that our full model outperforms the baselines and that using the *Combine* distance results in the highest accuracy. It also shows that the time attribute is essential when localizing using the *Transient* feature. In all cases, using only the imagery, which is the current state of the art, results in the lowest accuracy.

	Transient		*Places*		*Combine*	
Context	Top-1%	Top-5%	Top-1%	Top-5%	Top-1%	Top-5%
time	13.4	48.90	10.85	45.40	13.25	48.85
loc+time	31.50	81.50	27.20	74.90	36.20	82.10
sat+time	34.50	81.65	31.65	79.55	37.50	83.30
sat+time+loc	32.95	82.30	33.60	79.85	**40.30**	**84.35**

Table 3: Time verification accuracy of various baselines and two thresholds. Our approach with the *Combine* distance outperforms all other methods.

6.2. Application: Image Retrieval

In this qualitative application, we show how we can use our model to retrieve a set of ground-level images that would be likely to be observed at a given location and time. We start with an overhead image, specify a time of interest, and predict the visual attributes. We use the *Combine* distance defined in the previous section to find the closest ground-level images. In Figure 8, we show examples of images retrieved using this process. We observe that the ground-level images contain the expected scene type and appear to be from the appropriate time of day. For example, the top left overhead image contains a bridge and the closest ground-level images are visually consistent at both input timestamps.

6.3. Application: Metadata Verification

We focus on verifying the time that an image, with known location, was captured. For a given ground-level image, we first extract its visual attributes and then predict the visual attributes for a range of different times. We compute the distance between the actual and predicted attributes resulting in a distance for each possible time. Figure 9 shows heatmaps of these distances for two test examples, using our full model and the *Combine* distance. These show that our model is able to identify a small set of likely times.

We conducted a quantitative evaluation on a sample of 2000 images. For each image, we compute the distances as described above and then rank the times based on distance. Ideally, the correct time will have the lowest distance. In Table 3, we show the percent of images for which the correct time was within the Top-k% of possible times. The results show that the *Combine* distance outperforms both *Places* and *Transient*. While this approach does not fully solve the problem of detecting metadata falsification, it demonstrates that our model could be an important part of the solution.

7. Conclusion

We introduced a novel method for constructing dynamic visual attribute maps. In several large scale experiments, we demonstrated the practical utility of the model and highlighted the importance of including time, location, and an overhead image of the location as conditioning variables.

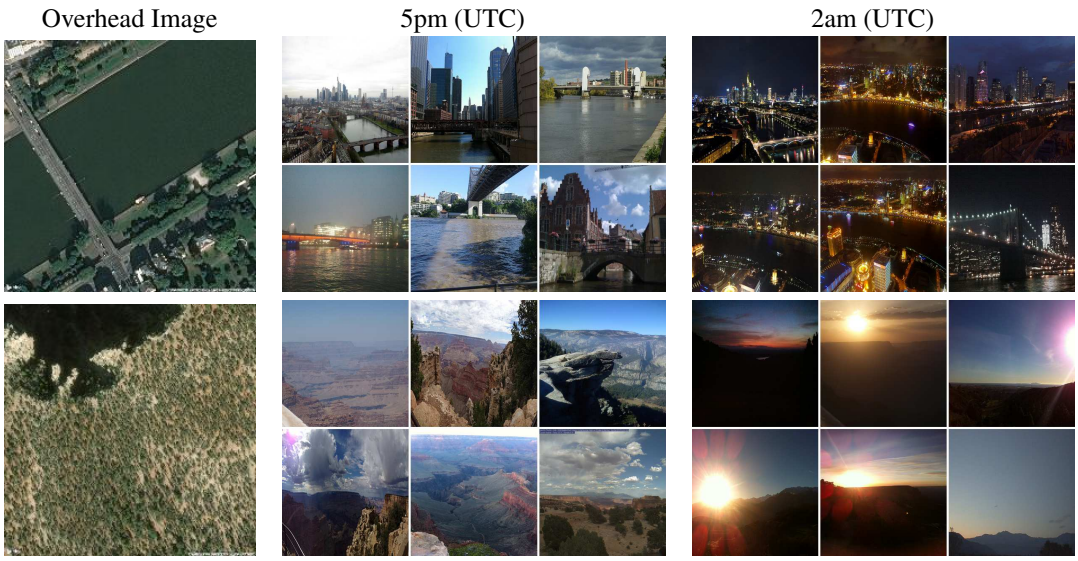


Figure 8: For each overhead image, we predict the visual attributes using our full model and compute the average distance between them and those of the ground-level images in the test set. (left) The overhead images of two query locations. The closest images when using August at 5pm as input (middle) and when using August at 2am (right).

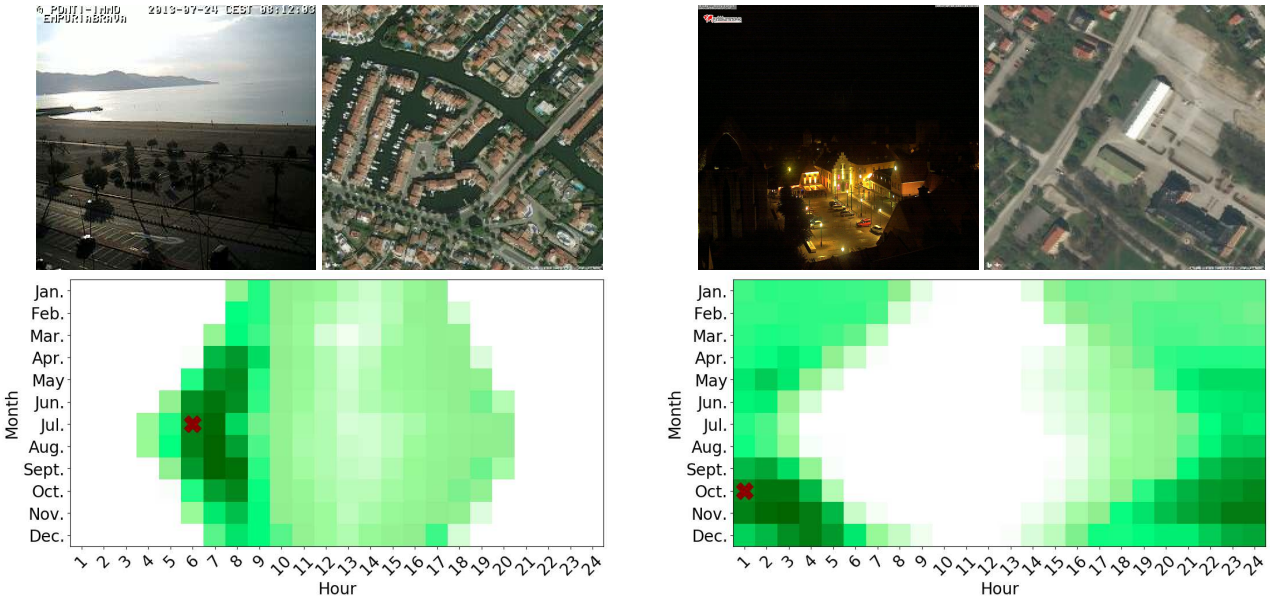


Figure 9: Two examples highlighting temporal patterns learned by our model. For each example, we show the original image and the overhead image of its location. For every possible hour and month, we use our full model to predict the visual attributes. The heatmap shows the distance between the true and predicted visual attributes, with dark green (white) representing smaller (larger) distances.

Such a model has many potential uses, including image-driven mapping, image localization, and metadata verification. In future work, we plan to focus on adapting this model to more directly support the application of metadata verification and to include additional visual attributes.

Acknowledgements: We gratefully acknowledge the financial support of an NSF CAREER grant (IIS-1553116), the University of Kentucky Center for Computational Sciences, and a Google Faculty Research Award. Thanks to Armin Hadzic for helpful feedback on the manuscript.

References

[1] M. Abadi et al. Tensorflow: A system for large-scale machine learning. In *USENIX Symposium on Operating Systems Design and Implementation*, 2016. 5

[2] Zachary Bessinger, Chris Stauffer, and Nathan Jacobs. Who goes there? Approaches to mapping facial appearance diversity. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2016. 2

[3] Aparna Bharati, Daniel Moreira, Joel Brogan, Patricia Hale, Kevin Bowyer, Patrick Flynn, Anderson Rocha, and Walter Scheirer. Beyond pixels: Image provenance analysis leveraging metadata. In *IEEE Winter Conference on Applications of Computer Vision*, 2019. 1

[4] Tiziano Bianchi and Alessandro Piva. Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE Transactions on Information Forensics and Security*, 7(3):1003–1017, 2012. 1

[5] Xueqing Deng, Yi Zhu, and Shawn Newsam. What is it like down there? Generating dense ground-level views and image features from overhead imagery using conditional generative adversarial networks. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2018. 2

[6] Hany Farid. Image forgery detection. *IEEE Signal Processing Magazine*, 26(2):16–25, 2009. 1

[7] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, 2017. 2

[8] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010. 4

[9] Connor Greenwell, Scott Workman, and Nathan Jacobs. What goes where: Predicting object distributions from above. In *IEEE International Geoscience and Remote Sensing Symposium*, 2018. 2

[10] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *IEEE Winter Conference on Applications of Computer Vision*, 2019. 1

[11] James Hays and Alexei A Efros. IM2GPS: Estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 2016. 4

[13] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Consistent temporal variations in many outdoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 2

[14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014. 5

[15] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics*, 33(4):149, 2014. 1, 2, 4, 5

[16] Stefan Lee, Haipeng Zhang, and David J Crandall. Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2015. 2

[17] Daniel Leung and Shawn Newsam. Proximate sensing: Inferring what-is-where from georeferenced photo collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 2

[18] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2

[19] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[20] Jiebo Luo, Jie Yu, Dhiraj Joshi, and Wei Hao. Event recognition: Viewing the world with a third eye. In *ACM International Conference on Multimedia*, 2008. 2

[21] Kevin Matzen and Noah Snavely. Scene chronology. In *European Conference on Computer Vision*, 2014. 1

[22] Radu P Mihail, Scott Workman, Zach Bessinger, and Nathan Jacobs. Sky segmentation in the wild: An empirical study. In *IEEE Winter Conference on Applications of Computer Vision*, 2016. 2

[23] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional GANs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[24] Tawfiq Salem, Connor Greenwell, Hunter Blanton, and Nathan Jacobs. Learning to map nearly anything. In *IEEE International Geoscience and Remote Sensing Symposium*, 2019. 2

[25] Tawfiq Salem, Menghua Zhai, Scott Workman, and Nathan Jacobs. A multimodal approach to mapping soundscapes. In *IEEE International Geoscience and Remote Sensing Symposium*, 2018. 2

[26] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. Quantifying the impact of scenic environments on health. *Scientific reports*, 5:16899, 2015. 1

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 4

[28] Shivangi Srivastava, John E Vargas-Muñoz, and Devis Tuia. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote Sensing of Environment*, 228:129–143, 2019. 2

[29] Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *IEEE International Conference on Computer Vision*, 2015. 2

[30] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2

[31] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[32] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *IEEE International Conference on Computer Vision*, 2016. 2

[33] Jingya Wang, Mohammed Korayem, Saul Blanco, and David J Crandall. Tracking natural events through social media and computer vision. In *ACM International Conference on Multimedia*, 2016. 2

[34] Jingya Wang, Mohammed Korayem, and David Crandall. Observing the natural world with flickr. In *ICCV Workshop on Computer Vision for Converging Perspectives*, 2013. 2

[35] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, 2016. 2

[36] Scott Workman and Nathan Jacobs. On the location dependence of convolutional neural network features. In *IEEE/ISPRS Workshop: EARTHVISION: Looking From Above: When Earth Observation Meets Vision*, 2015. 2

[37] Scott Workman and Nathan Jacobs. Dynamic traffic modeling from overhead imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[38] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision*, 2015. 2, 6

[39] Scott Workman, Richard Souvenir, and Nathan Jacobs. Understanding and mapping natural beauty. In *IEEE International Conference on Computer Vision*, 2017. 2

[40] Scott Workman, Menghua Zhai, David Crandall, and Nathan Jacobs. A unified model for near and remote sensing. In *IEEE International Conference on Computer Vision*, 2017. 2

[41] Ling Xie and Shawn Newsam. IM2MAP: Deriving maps from georeferenced community contributed photo collections. In *ACM SIGMM International Workshop on Social Media*, 2011. 2

[42] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[43] Menghua Zhai, Tawfiq Salem, Connor Greenwell, Scott Workman, Robert Pless, and Nathan Jacobs. Learning geo-temporal image features. In *British Machine Vision Conference*, 2018. 2

[44] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 1, 2, 4