What Does Learning About Time Tell About Outdoor Scenes?

Zeyu Zhang, Callista Baker, Noor Azam-Naseeruddin, Jingzhou Shen, Robert Pless
Department of Computer Science
George Washington University and Saint Louis University

2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) | 978-1-6654-7729-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/AIPR57179.2022.10092235

***Abstract*—In this paper, we explore the potential of utilizing time-stamps as labels for Deep Learning from webcams, surveillance cameras, and other fixed viewpoint image situations. Specifically, we explore if learning to classify images by the time they were taken uncovers interesting patterns and behaviors in the scenes captured by these cameras. We describe approaches to building datasets with large quantities of images and their accompanying labels, making them suitable for large-scale deep learning approaches. We share our results from the initial deep learning experiments.**

I. Introduction

One of the reasons to place cameras outdoors is to understand the patterns of behaviors and change that are visible. In urban environments, these patterns capture typical behaviors: people coming to a coffee shop, parking at their workplace, etc. Understanding these patterns is critical to next generation surveillance and urban planning paradigms.

But these patterns may be difficult to detect automatically, especially if there are patterns that you don't know to look for (e.g. the bird that lands on the roof at the same time each day), so it is interesting to think about ways of automatically learning these patterns that define the passing of time. Our approach to this problem is to think about the converse of this problem. In scenarios where we have a large collection of images of one scene with time-stamps, we hypothesize that those patterns of change can be used to infer the time the image was taken — and furthermore that learning to classify images by time of day will create representations that highlight interesting things about the scene. In this sense, we are using a time-stamp as a label which is a proxy for understanding the scene, similar to the way that image colorization was used as a proxy for other visual understanding problems in single images [11].

To support this effort we need a large collection of time-stamped imagery. There have been a few datasets created before, including the AMOS ("Archive of Many Outdoor Scenes") dataset [6], [5], [16], and the smaller but better stabilized Webcam Clip Art dataset [10], and the SkyFinder dataset [15], which annotates ground vs. sky pixels and includes weather data. These datasets are now quite old, difficult to get access to, and relatively low frame rate (often one image per 30 minutes).

To address these limitations, we have created a strategy to create a new, large, meta-index of webcam imagery. This is based on the recognition that a large collection of sites already create long-term webcam imagery archives. We characterize these archives, understand the structure of the URLs that provide links to their images. This allows us to keep the native temporal and image resolution without adding to the network bandwidth, and simplifies our task of sharing the data. This meta-index captures images that are up to 12 Megapixels, have

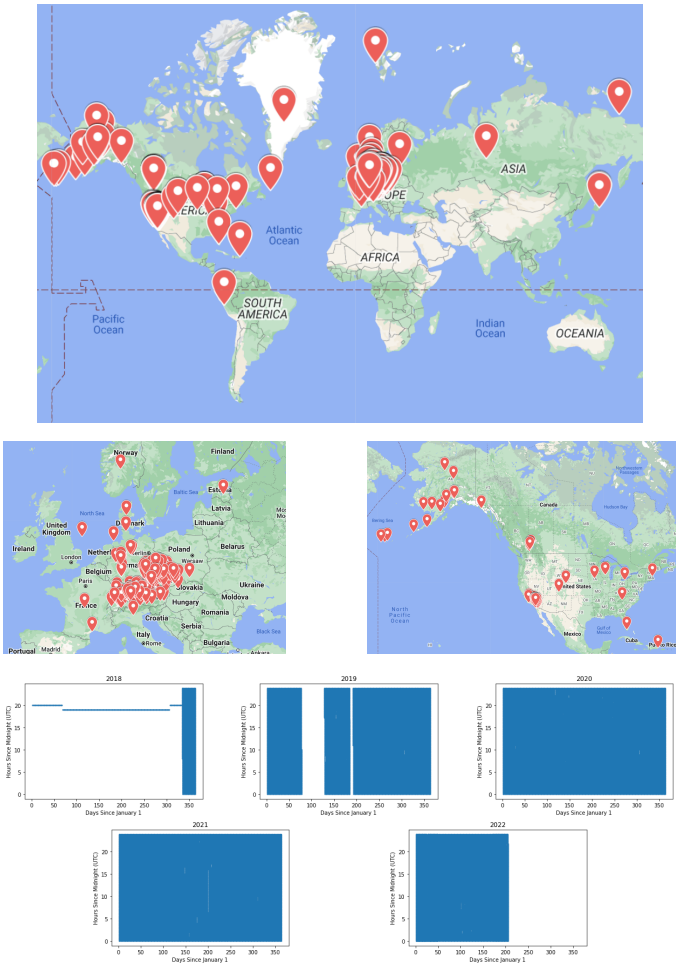

Fig. 1. We index existing webcam archives that retain outdoor webcam images for many years and publicly share those images, as a resource for long-term scene understanding research. Our index comprises 377 different webcams, with especially dense representation in regions of Europe and North America. For each of these cameras, we index when images were captured — for one camera, the data availability plots are shown for 2018-2022 in the bottom set of figures. Data availability for cameras like this is typical, as webcam providers start capturing images, they often changes the frequency with which they capture or archive imagery, and they often have substantial downtimes.

978-1-5386-5541-2/18/$31.00 ©2018 IEEE

images every minute and have archive durations lasting from 2 to 15 years.

We train deep learning algorithms on images from single scene to classify imagery by time of day and time of year. Because the network is trained for a single scene, the classification often becomes very good, and because there are so many images from that scene, the classification generalizes well to unseen images from that camera. Because the images usually come with time-stamps, this classification is less interesting than the set of features in the scene that were most important for that classification. Depending on the scene, those features include natural processes from lighting (shadow direction, dawn and dusk times), natural weather processes (consistent fog and cloud patterns), and human patterns (like a restaurant consistently rolling down an awning, or a city square bringing in).

To automatically find these features, we create neural network visualizations that highlight the image regions most salient for the time-of-day and day-of-year classifications. We share preliminary results across a number of different scenes and highlight how choices in Deep Learning visualization affect the interpretability of the results. Our initial results highlight limitations of this approach, primarily that weather/lighting cues dominate the scene, with very strong cues about time of day, making it harder to find the human activities than we expect. This suggests that further work is needed to find patterns of human behavior, such as explicitly extracting cues unrelated to lighting.

II. Related work

1) Webcam research in computer vision: The first webcam viewed a critical element in the patterns of human life, a coffee pot [18], a coffee pot at the University of Cambridge, which became network accessible in 1993, and shared images until 2001. Some of the earliest works used outdoor webcams to explore extracting intrinsic images from natural lighting variations [**?**], and analysing traffic flow in cities [17].

Larger scale efforts to understand statistics of variations across many webcams include finding low-dimensional linear structures [6], geo-location cues [7], and understanding transient patterns of weather [9], [1], [12]. Webcams have been used to characterize long term patterns of human behavior at the city scale in terms of urban traffic density [20] and in characterizes pedestrian and bicycling behavior [4].

Explicit datasets that have been shared to supported webcam research include AMOS ("Archive of Many Outdoor Scenes") dataset [6], [5], [16], and Webcam ClipArt, a smaller collection of cameras with mostly high quality and whose images are approximately aligned [10], and the SkyFinder dataset [15], which makes great efforts to exactly align the imagery and provides pixel-specific annotations of sky and not-sky pixels.

2) Webcam or Outdoor Time-Lapse and Estimating Time: There is quite limited work in looking at long term outdoor imagery that relates to estimating time of day. Recent work seeks to automatically detect and read analog clocks that are visible in outdoor imagery [19], and validating a time-stamp of outdoor imagery [13], and estimating location/time of cameras based on tracking shadows [8]. To our knowledge, there is no prior work trying to estimate time-of-day from a single webcam image.

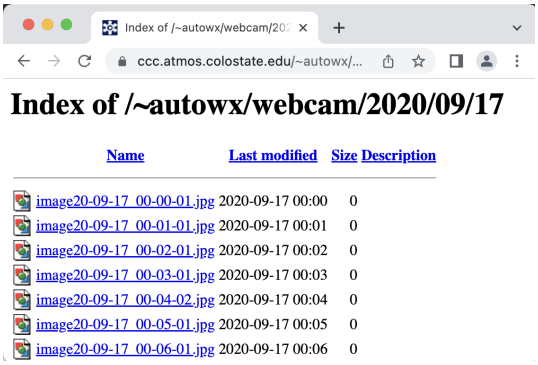

Fig. 2. An example web page that is found by search for open directory access pages that share image that contain the word "webcam".

III. A meta-archive of webcam imagery

In this section we explain an approach to create and curate a list of existing webcam imagery archives. This process consisted of finding existing webcam archives, indexing them them to find and time-stamp the URLs of imagery in them, and adding meta-data to make the archive useful.

A. Webcam Archive Discovery

Our approach to discovering existing webcam archives was not automated. We used extensive human effort to discover webcam archives with a variety of explicit searches. In order to find a small collection of distinct webcams per search we used a collection of searching strategies:

- Include: "Index of" in the search. This appears on websites that permit directory access, and this directory access is often necessary for us to be able to parse long URLs to determine time and date information.
- Include: site:.ca, site:.fr, etc. in order to limit the search results to specific countries.
- Include: "webcam archive", or "live stream archive" to limit results to pages that are likely to have webcam archive images.
- Include: "inurl:2019/07" or similar patterns consistent with common year/month directory structures to identify long term webcam archives.

Using such searches, we have to date found 377 different webcams that have existing long term archives, and whose image URLs are publicly available to download. Figure 2 shows an example of a webpage that we found using similar searches, a directory-accessible webpage that provides web-viewable image directories. Visible inspection of this shows that images are archived every minute.

B. Webcam Archive URL scraping

For such sites, we use a python script to recursively search the directory tree for all images from the same webcam, and

Example image URLs

```
https://www.avo.alaska.edu/webcam/arch//Redoubt_2/2020/09/10/redoubt-2-20200910T085100Z.jpg
https://ok.water.usgs.gov/scripts/webcam/archive/202009/20200911/USGS07196320_20200911_101505.jpg
http://209.97.184.57/webcamarchive/2019-11-November/13-Wednesday/11%3A53.jpg
```

Fig. 3. The image URLs had significant variability making some hand-tuning necessary to parse each to record time and date of the images.

to extract the date and time in UTC format. The latitude and longitude of each camera was determined manually, and the interpretation of the time listed in the filename was manually examined to determine if it was likely to be listed in local time or UTC time.

The script was designed to handle a variety directory structures. Some examples of URLs highlight the variability that we observed. For each camera we hand coded regular expression pattern matchers to extract the time and date of each image, and used the estimated GPS coordinate of the camera to translate local times to UTC, if necessary.

In other cases, the webdirectory was not visible through HTTP, but looking at the URLs through their visualization infrastructure had clear patterns that allowed us to generate the URLs.

C. Statistics to Date

We have found and indexed webcam archives from 377 different outdoor webcams. The oldest images for which we currently have URLs date back to 2005. The median camera refresh reate is 10 minutes, and the highest frame rate archive contains images every minute. The median duration of images for the archive of a single camera is 1600 days. The cameras are concentrated largely in the United States and Western Europe. We believe there are likely more webcam archives for cameras in these regions, but our approach to finding webcams was also limited by our language skills (largely English) and intuitions of how to best search for these.

IV. Network Training and Sample Results

One possible use of a very large, time-stamped dataset is to take advantage of time as an always available proxy label for interesting behaviors that happen in the scene. In this section we share initial directions of this research.

A. Data preparation and method

We select two webcam scenes for this task. One is the parking lot at an university which has 47000 images, and another one is a river port with a sculpture which contains 71000 images. Some example images are shown in Figure 4. Images from the camera are partitioned by the hour they were taken (so all images taken between 2:00 and 2:59 are in one class), so the ground truth is known for all images.

Next we randomly select a few years in the dataset, and divide the dataset to our training dataset and test dataset. Then for the training set, we divide them to training set and validation set. The proportion is 7:3. In our paper, We use images from 2017,2018 and 2020 in the training/validation set, and images from 2019 as the test set.

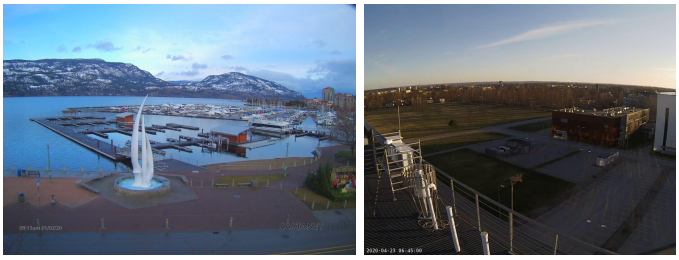

Fig. 4. Example images from two scenes river port (left) and parking lot (right) in our experiments. Both were chosen because the scenes contain significant human activity, (although that is sometimes small in the image).

We tried both ResNet and Vision transformer for this task. For ResNet, we use the ResNet-101 and start with weights pretrained on ImageNet [3]. The images are pre-processed so that timestamps are not visible, then blurred with a Gaussian filter with kernel size of 5×5, and standard deviation of 0.5. Images are then resized to 224×224 and normalized to have zero mean and unit variance.

In the other experiment we use the ViT-B/32 [2], which also pretrained on ImageNet. Images are pre-processed as same as the ResNet experiment despite the images are resized to 448×448 In both cases we train the network using the standard cross-entropy loss function with batch size 24. We use cosine scheduler with warm restart [14] to schedule the learning rate. We set the minimum learning rate to 0.0001, number of iterations for the first restart is 24 and T_mult is 2. The learning rate start from 0.01 for ViT and 0.001 for ResNet.

B. Results and Visualizations

Here we show results of the training process, and some results of visualizations that try to highlight the most important features. Figure 5 shows results from a scene containing a parking lot. Over the course of the day, the parking lot fills in the morning and empties in the evening, so there are human patterns that are defined by the time of day. We show the confusion matrix for the 24 classes (each out of the day). On the training data, the confusion matrix is nearly diagonal because the accuracy is very high, indicating some degree of over-training.

The bottom left of the figure shows results on testing data taking from the same year as the training data (but from days that were not used during training). The daylight hours are very accurately predicted, but there is less variation at night, so the results are less accurate then. We also tested on images from the following year (bottom right) and see that the variations over long time scales in this scene are small enough that the prediction results are largely similar.

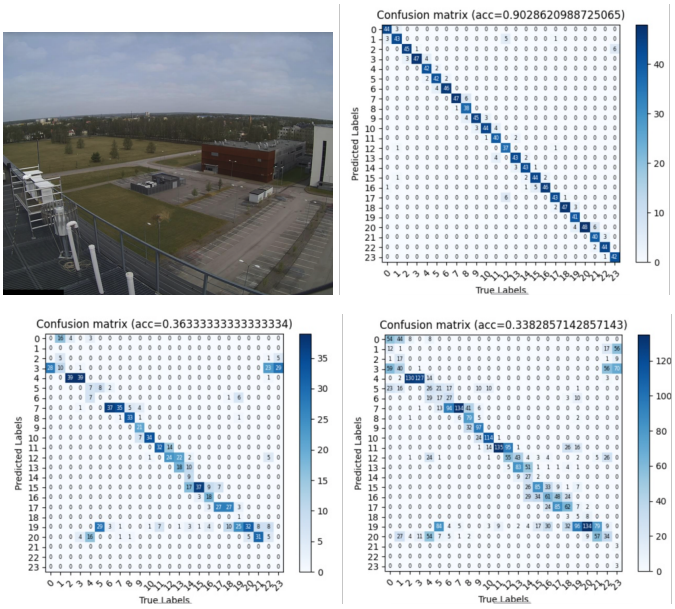

Fig. 5. For the scene shown in the top left, we summarize prediction accuracy by showing confusion matrices. On training data (top right), the accuracy is very high and there are no trends in what errors there are. On test data from the same year (bottom left) and test data from a different year (bottom right), the accuracy is quite high during daylight hours, but worse at night.

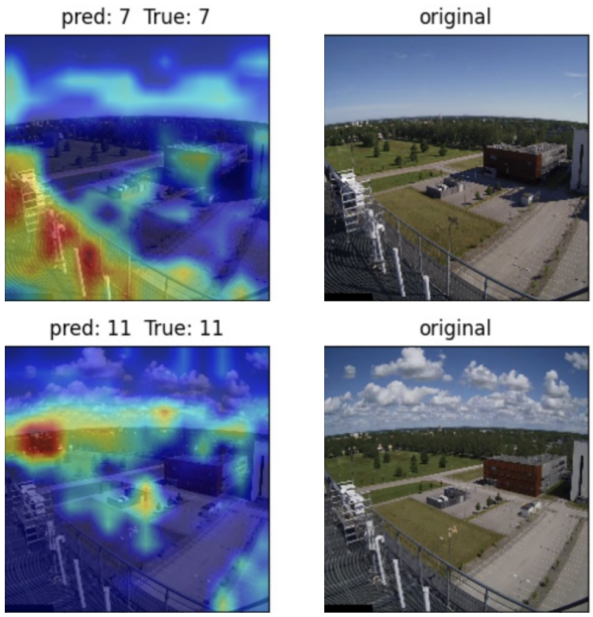

Fig. 6. Saliency visualization showing the image regions that have the highest impact on the classification. (Top) An image from the 7am hour correctly classified, and (bottom) an image from the 11am hour correctly classified.

Figure 6 shows the Class Activation Maps [21] highlighting the image regions most responsible for the predicted class. Although this scene has a strong variation in appearance due to human activity (the cars in the parking lot, among other things), the 7am classification is driven largely by the bright morning appearance of poles on the rooftop near the webcam, and the 11am classification is driven by shadows of poles in the parking lot and the visibility of buildings in the distance.

Visualizations using ResNet as the basic machine learning model are shown in Figure 7. We show the CAM saliency map for a sequence of images captured at different times of day. The model always focuses on the tall statue, but at different times of day focuses on the shadows in the foreground and parts of the harbor. Inspection of many images like this make it clear that the status appearance changes consistently, the shadows move consistently, but the appearance of the building in the harbor was more difficult to interpret.

Visualizations using Vision Transformer architecture are shown in Figure 8. The class activation maps for the final layer of the ViT are less strongly correlated with location than the ResNet architecture. Possible explanations consistent with these salience visualizations might include the thinness of clouds in the sky, the amount and location of unshaded areas on the ground and the amount of sunlight reflected off the lake, but extensive observation of many images like this failed to find very coherent cues.

V. Discussion and Conclusion

We have shown results of learning to classify an image by the hour of day in which the image was taken. For two example scenes, we find that we can reliably predict the time of day in the daytime, and have more mixed results in the evening. But predicting the time an image was taken is rarely a useful task in itself.

Our hypothesis was that the network would learn to tell the time of day based on varying patterns in the scene that were caused by human activity, and therefore the time-label would be a proxy for learning interested features of human activity in the scene. In these two cameras we have shown results that highlight that the lighting cues seem to dominate the explanations of the time of day, probably because the lighting cues are very strong (in terms of the magnitude of the image changes that they cause) and, in the case of long cast shadows, for example, very specific. Initial efforts to find cameras where this wasn't the case were not successful.

This does not show that using time cannot be a good proxy for human behavior. We believe that Deep Learning that integrates time-sequence modeling (e.g. LSTM), and/or otherwise focusses on images captured with a much higher frame rate may be more successful. A scene where the mail-truck reliable comes around 1pm will not have this be a consistent cue if an entire hour is considered as one class, or if images are captured only once per hour.

The processing of finding cameras and indexing the images was also instructive. Future work in this area should consider several issues. First, anecdotally, about 5% of the webcam archives that we discovered have since changed their directory structures or permissions over the course of the 6 months doing this project. This is consistent with the "lifetime" of a webcam being approximately 10 years (even for good ones!). We think this is likely because most webcams are sharing images "for fun" and the institutional support to maintain them is not infinite.

Second, there are other sources of long term time-lapse imagery that are not shared as large directories of .jpg files.

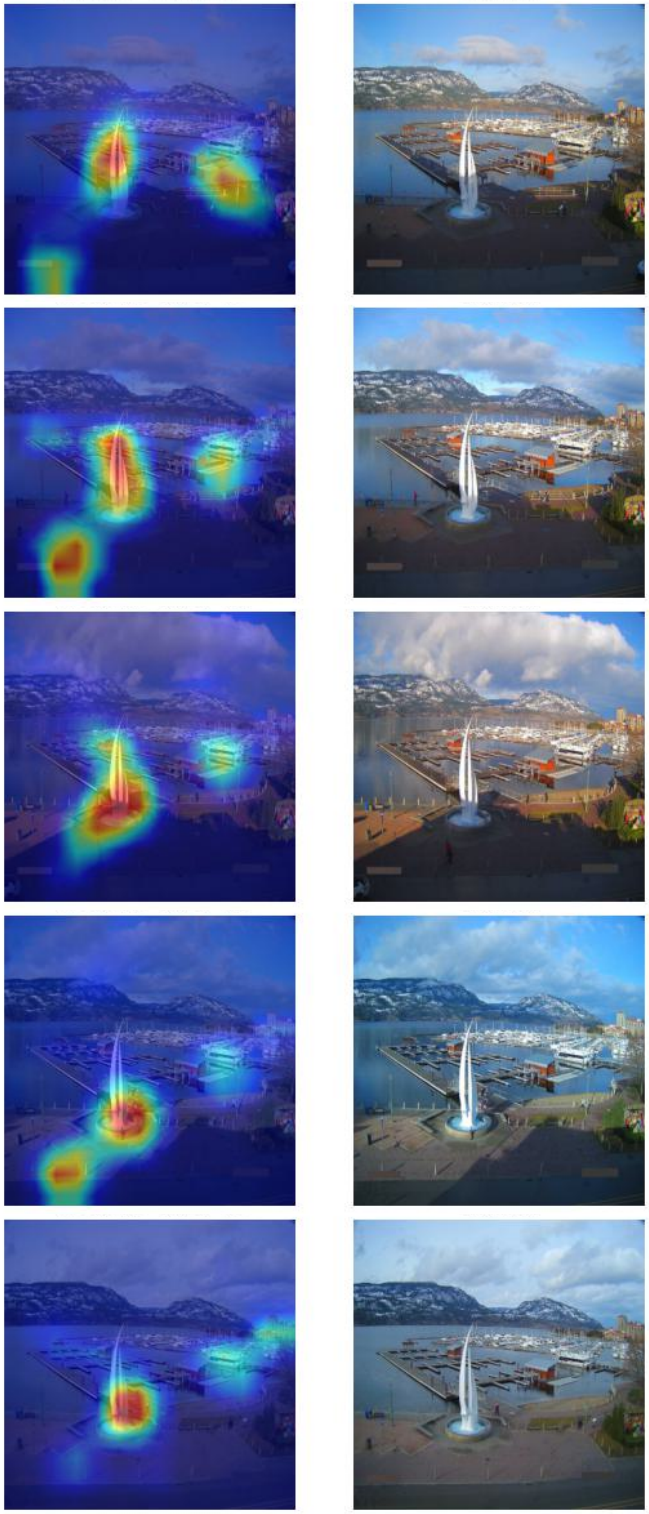

Fig. 7. Images on the left is the class activation maps of test images on trained ResNet. The activation is highlighted by red. Images on the right is the original images. These images are from 9 AM to 1 PM within same day. The shadow of building moves from left to right in these images

Increasingly, webcams are live-streams and not archived at all, or a daily summary video is created. Working to integrate this image sources could dramatically increase the number of possible scenes (with a cost of introducing video artifacts that

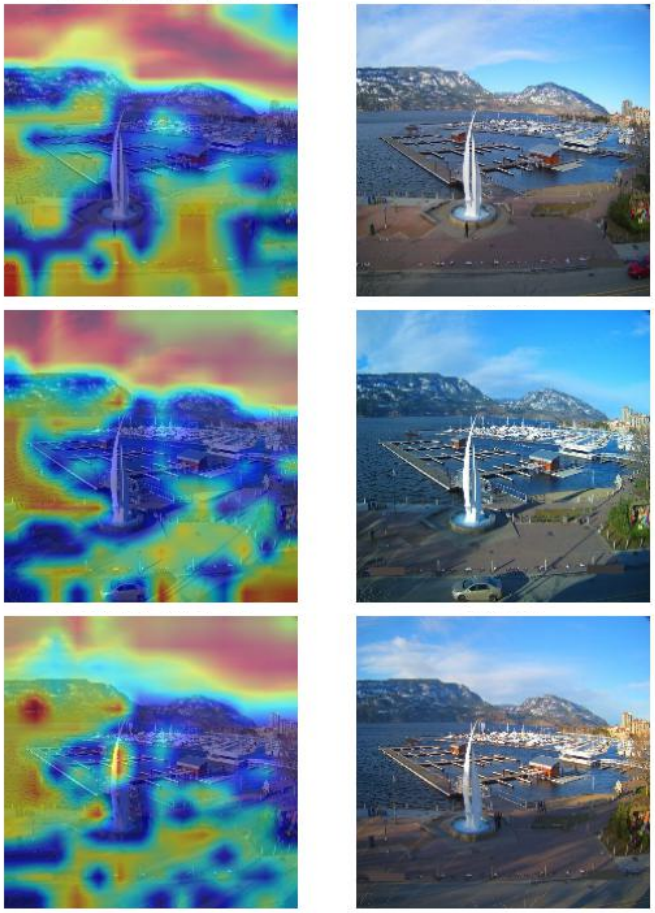

Fig. 8. Class activation maps of time predictions using the ViT architecture show much less consistent structure, because the ViT architecture does not preserves spatial structure after the first layer of the network.

are consistent over multiple frames).

Third, we have found that it is vital to build visualization tools to search through the dataset and explore data availability. Visualizations such as the annual summary of which images are available in the bottom of Figure 1 were unexpectedly valuable debugging tools across every step of our process.

REFERENCES

[1] R. Baltenberger, M. Zhai, C. Greenwell, S. Workman, and N. Jacobs. A fast method for estimating transient scene attributes. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] J. A. Hipp, D. Adlakha, A. A. Eyler, R. Gernes, A. Kargol, A. H. Stylianou, and R. Pless. Learning from outdoor webcams: surveillance of physical activity across environments. In *Seeing cities through big data*, pages 471–490. Springer, 2017.

[5] N. Jacobs, W. Burgin, R. Speyer, D. Ross, and R. Pless. Adventures in archiving and using three years of webcam images. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 39–46. IEEE, 2009.

[6] N. Jacobs, N. Roman, and R. Pless. Consistent temporal variations in many outdoor scenes. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6. IEEE, 2007.

[7] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating static cameras. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–6. IEEE, 2007.

[8] I. N. Junejo and H. Foroosh. Estimating geo-temporal location of stationary cameras using shadow trajectories. In *European conference on computer vision*, pages 318–331. Springer, 2008.

[9] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on graphics (TOG)*, 33(4):1–11, 2014.

[10] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan. Webcam clip art: Appearance and illuminant transfer from time-lapse sequences. *ACM Transactions on Graphics (TOG)*, 28(5):1–10, 2009.

[11] G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6874–6883, 2017.

[12] P. Lepetit, L. Barthes, C. Mallet, and N. Viltard. Learning to compare visibility on webcam images. In *Proceedings of the 10th International Conference on Climate Informatics*, pages 91–97, 2020.

[13] X. Li, W. Xu, S. Wang, and X. Qu. Are you lying: Validating the time-location of outdoor images. In *International Conference on Applied Cryptography and Network Security*, pages 103–123. Springer, 2017.

[14] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[15] R. P. Mihail, S. Workman, Z. Bessinger, and N. Jacobs. Sky segmentation in the wild: An empirical study. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–6. IEEE, 2016.

[16] J. D. O'Sullivan, A. Stylianou, A. Abrams, and R. Pless. Democratizing the visualization of 500 million webcam images. In *2014 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–5. IEEE, 2014.

[17] S. Santini. Analysis of traffic flow in urban areas using web cameras. In *Proceedings Fifth IEEE Workshop on Applications of Computer Vision*, pages 140–145. IEEE, 2000.

[18] Q. Stafford-Fraser. On site: The life and times of the first web cam. *Communications of the ACM*, 44(7):25–26, 2001.

[19] C. Yang, W. Xie, and A. Zisserman. It's about time: Analog clock reading in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2508–2517, 2022.

[20] S. Zhang, G. Wu, J. P. Costeira, and J. M. Moura. Understanding traffic density from large-scale web camera data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5898–5907, 2017.

[21] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.