Annual ADFSL Conference on Digital Forensics, Security and Law

2022 Proceedings

Timestamp Estimation From Outdoor Scenes

Tawfiq Salem
Department of Computer Information Technology, Purdue University

Jisoo Hwang
Department of Computer and Information Technology, Purdue University

Rafael Padilha
Artificial Intelligence Lab, Institute of Computing, University of Campinas

Scholarly Commons Citation

Salem, Tawfiq; Hwang, Jisoo; and Padilha, Rafael, "Timestamp Estimation From Outdoor Scenes" (2022). *Annual ADFSL Conference on Digital Forensics, Security and Law*. 2.
https://commons.erau.edu/adfsl/2022/presentations/2

TIMESTAMP ESTIMATION FROM OUTDOOR SCENES

Tawfiq Salem[1], Jisoo Hwang[1], Rafael Padilha[2]

[1] Department of Computer and Information Technology, Purdue University
West Lafayette, IN, USA
[2]Artificial Intelligence Lab (RECOD.ai), Institute of Computing, University of Campinas
Campinas, Brazil
{tsalem,hwang160}@purdue.edu, {rafael.padilha}@ic.unicamp.br

ABSTRACT

The increasing availability of smartphones allowed people to easily capture and share images on the internet. These images are often associated with metadata, including the image capture time (timestamp) and the location where the image was captured (geolocation). The metadata associated with images provides valuable information to better understand scenes and events presented in these images. The timestamp can be manipulated intentionally to provide false information to convey a twisted version of reality. Images with manipulated timestamps are often used as a cover-up for wrongdoing or broadcasting false claims and competing views on the internet. Estimating the time of capture of a photograph is a challenging task that requires a comprehensive understanding of the scene and its geographical location. In this paper, we propose a learning-based approach based on deep learning to estimate when an outdoor image was captured. We provide a detailed quantitative and qualitative evaluation of the trained models for various settings and show that the proposed approach outperforms baseline methods.

Keywords: Digital forensics, Time estimation, Scene understanding, Deep learning

1. INTRODUCTION

The appearance of an outdoor scene can drastically change depending on the time of the year and the hour of the day. As humans, if we look at an image of an outdoor scene, like the ones in Figure 1, and consider its characteristics and elements (such as sunlight or dark sky), we can roughly estimate when the image was captured. Although this process is natural for us, doing so requires an accumulated understanding of our world and how the appearance of a scene varies as time progresses.

Such variations might be as subtle as the changes in the sunlight at different times of the day or as noticeable as the changes in the color of trees' leaves over the different seasons, depending on the time of the day, the month of the year, and the geolocation where the photo was captured. Furthermore, factors like weather conditions and other visual elements (e.g., people wearing warm clothes) influence our perception of time. Even though most cameras store the timestamps of images in their metadata at the moment of creation, this information is often noisy and unreliable (Tsai et al., 2016). Several photo editing software and mobile applications overwrite the metadata when processing an image, whereas social networks often erase it during the upload process.

Manually annotating and estimating the timestamp of a collection of photos is an error-prone and infeasible task, especially as the num-

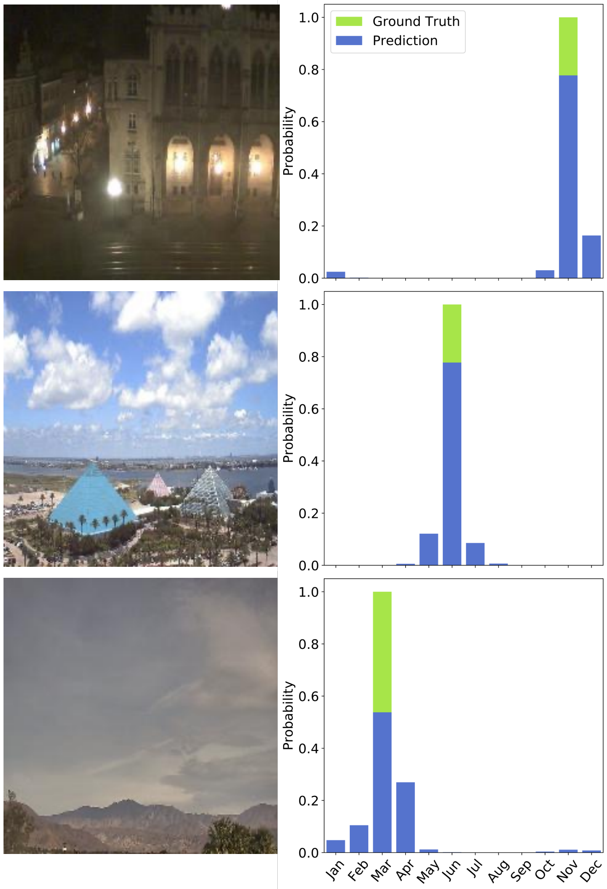

Figure 1: Given a query image, our proposed approach estimates the moment when it was captured. This figure shows how the model estimates the month in comparison to the ground truth.

ber of digital photos increases. In this sense, the development of automated methods to estimate the time-of-capture based on the content of an image can overcome such issues and improve the analysis and understanding of outdoor scenes and events taking place in those images. Furthermore, such methods can aid visual applications that benefit from accurate temporal contexts, such as semantic scene classification (Yuan, Luo, & Wu, 2010; Derpanis, Lecce, Daniilidis, & Wildes, 2012) and visual rendering from crowd-sourced images (Z. Li, Xian, Davis, & Snavely, 2020; Martin-Brualla et al., 2021). We leverage information from ground-level imagery where there are plenty of geographical contexts and temporal information. Such imagery can assist the network with learning to estimate metadata closely related to real-world scenarios (Padilha et al., 2022). By utilizing high-level scene appearances, our architecture learns and predicts the temporal information in a hierarchical way, starting from the easier time scale (i.e., month) to more granular information (i.e., week of the year and hour of the day). Doing so allows the network to better understand the relation between visual attributes and each temporal scale. As we increase granularity (e.g., from months to hours), the model uses the prediction of a higher level in the hierarchy to guide the prediction of the current stage.

In this work, we propose a high-level convolutional neural network (CNN) architecture to estimate the month, week, and hour of capture of a photograph (Figure. 3). Our network receives as additional context, the location, which allows the network to learn richer representations (Salem, Workman, & Jacobs, 2020) able to adapt the temporal analysis to geographical differences (e.g., North and South hemispheres) and varied types of scenes (e.g., seacoast or mountain regions). Our network is optimized in a novel multi-task manner, with top-level temporal information cascaded throughout the model and fed as input to estimate bottom-level ones. In our approach, the features from *month* prediction is used to estimate the *hour* of the day, whereas both are used for the *week* estimation. We quantitatively evaluate the proposed method on realistic outdoor scenes, comparing the accuracy with other baseline methods.

2. RELATED WORK

With the emergence of machine learning, many CNN-based methods have been explored to solve digital forensics problems (Ding, Zhu, Alazab, Li, & Yu, 2020). Specifically, inferring temporal information from a photograph where metadata is missing has been investigated in different ways. Several works analyze specific visual elements such as human appearance and fashion (Salem, Workman, Zhai, & Jacobs, 2016; Ginosar, Rakelly, Sachs, Yin, & Efros, 2015), visual style of objects (Vittayakorn,

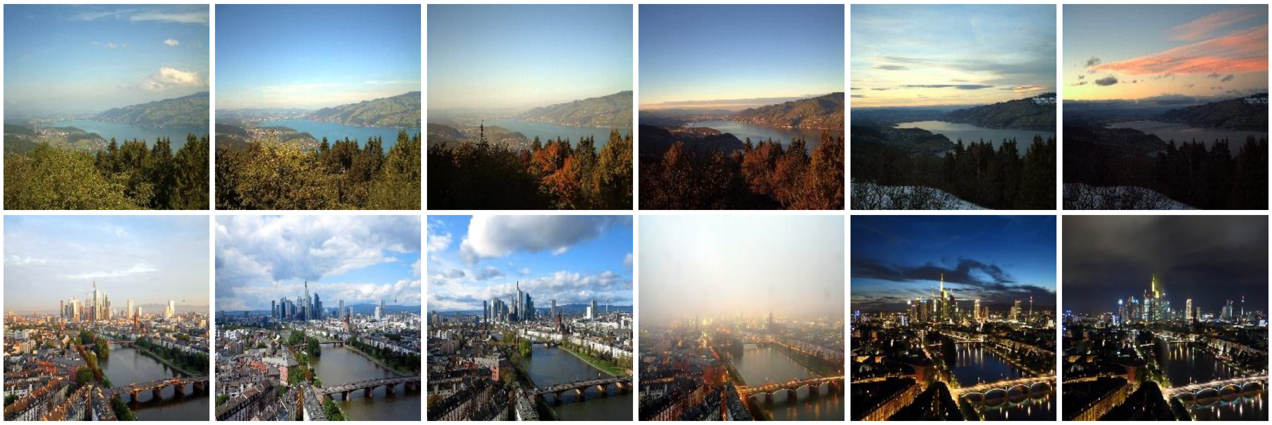

Figure 2: Sample of images from the dataset used in this experiment. These ground-level images show the changes in the scene's appearance over time. Each image is associated with the corresponding timestamp and geo-coordinates.

Berg, & Berg, 2017; Jae Lee, Efros, & Hebert, 2013), architecture (Lee, Maisonneuve, Crandall, Efros, & Sivic, 2015), and photo-generation artifacts (Martin, Doucet, & Jurie, 2014; Palermo, Hays, & Efros, 2012) to estimate when an image was captured. Even though these visual elements carry some degree of temporal information, they are not always present to reliably infer the time-of-capture of an outdoor scene. In this sense, Tsai et al. (Tsai et al., 2016) proposed a method to estimate the position of the sun based on the sky's appearance and combine it with the date and geographic location stored in the metadata of an image to estimate the hour of capture. Similarly, Kakar et al. (Kakar & Sudha, 2012) and Li et al. (X. Li, Xu, Wang, & Qu, 2017) estimate the sun azimuth angle to verify if the timestamp stored in the metadata of a photograph has been manipulated.

Instead of looking for particular visual cues or indirectly estimating time from the sun's position, other works more closely related to ours analyze the global appearances of a scene to reason about its time-of-capture. Volokitin et al. (Volokitin, Timofte, & Van Gool, 2016) use the features extracted from a pre-trained CNN to estimate the capture time (the year and hour of the day) of an outdoor image. Zhai et al. (Zhai et al., 2018) propose a CNN architecture to learn geotemporal image features that can be used to estimate the hour and month of capture for a given image. In (Laffont, Ren, Tao, Qian, & Hays, 2014), the authors show that the learned features present a high correlation to transient attributes of a scene related to the passage of time, such as season, weather, and illumination conditions. Padilha et al. (Padilha et al., 2022) presented a deep learning-based approach for verifying the timestamp associated with an image. We build upon these strategies and propose a new approach based on deep learning that incorporates visual information from ground-level imagery and geographical coordinates to directly predict the timestamp of an image. Different from previous works, our approach has been trained and optimized in an end-to-end fashion to directly estimate the timestamp of a given outdoor ground-level image.

3. CAPTURE-TIME ESTIMATION

We present a general approach for time estimation from outdoor images that could be used to estimate the time of capture and model the relationship between scene appearances and time. Such a model enables and supports many tasks to better understand and analyze scene appearance. Our objective is to develop an automated method for estimating the capture time of an image using the scene's appearance in that image as

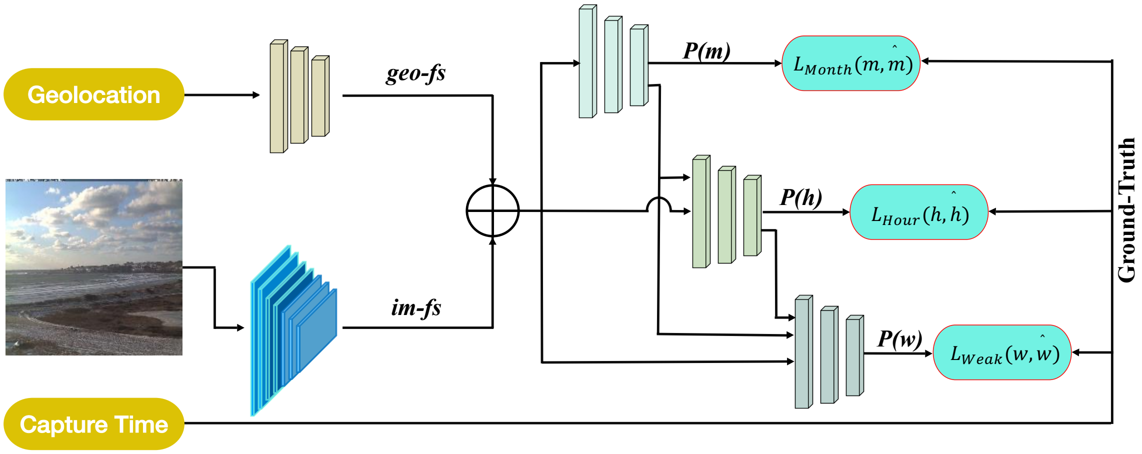

Figure 3: An overview of our network architecture.

a cue. As humans, we have the ability to draw a conclusion about the appearance of a place based on different information such as scene category, time, and geolocation of the place. For example, we can expect a forest scene to have trees with no leaves in Autumn, whereas they will look alive and have all their leaves in Summer. The geolocation of the scene also influences its appearance. We can anticipate seeing more sunny scenes in Florida than in Minnesota. In our proposed approach, we model this connection between the scene's appearance and its geolocation with time. Specifically, we train a model to learn a conditional probability distribution, $P(t|im, l)$, to estimate the capture time (t) for a given image (im), and its geolocation (l). The distribution $P(t|im, l)$ is challenging to learn because it requires memorizing the appearance of every place on Earth and how it changes over time. To learn this complex relationship, we can represent the time as an hour of the day, day, month, week, and year. We have decided to predict the hour (h) of the day, the week (w), and the month (m) as the capture time of the image. We drop the day and the year because the changes in scene appearances between different days or years are hard to be detected and would require additional information besides the image and geolocation (Palermo et al., 2012; Vittayakorn et al., 2017). In this work, the conditional probability distribution we model is $P(m, w, h|im, l)$. Given an image and its geolocation, we estimate the capture time of the given image with regard to the month (m), week of the year (w), and hour of the day (h). In our proposed architecture, we integrate the human perception that we may guess a broad range of time for the capture time, and then we narrow it down to a more specific period. We design the model first to predict the month of the year, then use this prediction to help in predicting the hour of the day and the week of the year. Therefore, the model learns three different conditional probabilities starting from estimating the month, then the hour of the day, and finally the week of the year. For predicting the hour of the day, we condition it on the month prediction, and in the same way, we condition the prediction of the week on the month and hour predictions, as explained in the following three conditional predictions.

$$P(m|im, l) \tag{1}$$

$$P(h|im, l, P(m)) \tag{2}$$

$$P(w|im, l, P(m), P(h)) \tag{3}$$

In (1), the probability of the month is conditioned on both im and l. In 2 we predict the probability over the hours of the day conditioned on the im, l, and the month prediction from (1). In the same fashion, the week prediction will also be conditioned on both the month and the hour of the day in (1) and (2).

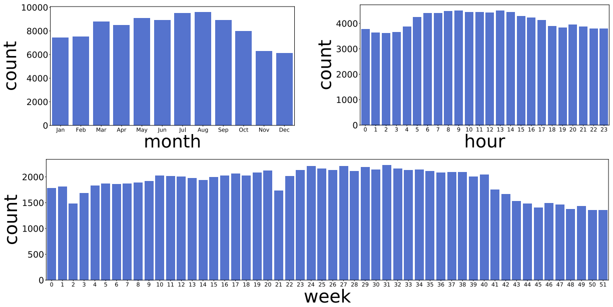

Figure 4: The temporal distribution of the subset of the AMOS dataset used in this work.

3.1 Dataset

To support our work, we use the Archive of Many Outdoor Scenes (AMOS) (Jacobs, Roman, & Pless, 2007) dataset. The AMOS data set consists of over a billion ground-level images captured from publicly available outdoor webcams across the world. In this experiment, we utilize a subset of the AMOS dataset which contains 89,280 images collected between 2011 and 2014 from 53 unique cameras (Figure 2). Those 53 cameras have been found to be more stable than other cameras (Zhai et al., 2018). We split the dataset into 94.4% for training and 5.6% for testing. The training set contains $84,288$ images, and the testing set has $4,992$. Each image is associated with the timestamp and geolocation information. The location information includes the latitude and longitude of the webcam. The time information provides the capture time in UTC time. Figure 4 presents the distribution of the images over the month, week, and hour.

3.2 Architectural Details

Our proposed approach for modeling the relationship between the scene appearances and time, as in (1), (2), and (3) has two phases. The first phase represents the feature extraction, where the goal is to learn representative features for the two inputs, the *image features (im-fs)* and *geolocation (geo-fs)*. We concatenate the two feature representations *(im-fs, geo-fs)* into one vector representation. Then we provide it as an input to the second phase where the objective is to learn to predict the right capture time for the given image. In the second phase, we have three prediction heads, individually estimating the month (P(m)), hour of the day (P(h)), and week of the year (P(w)) as described in Figure 3. Because the top-level temporal information is fed as an input to estimate bottom-level ones, we refer to this model as the *Cascaded* model.

For evaluation and comparison purposes, we also train two other models as baselines. The first one is *Cascaded without location* model, in which we only use the image as an input without including geolocation. The second baseline is the *Not cascaded* model, in which the prediction heads are independent without leveraging previously estimated temporal information. To extract the features from the ground-level input image, the DenseNet-121 (Huang, Liu, Van Der Maaten, & Weinberger, 2017) model pretrained on ImageNet is used as a base model with all the layers being trainable during training. The extracted features are flattened and then fed into two fully connected layers with 256 and 128 neurons, respectively. After each fully connected layer, ReLU activation and batch normalization are applied.

In the location branch, the location features are extracted using two fully connected layers with 256 and 128 neurons, and the ReLU activation function is applied to both layers. Each fully connected layer is followed by dropout with a rate of 0.5 and 0.3, respectively. Then batch normalization is performed after every dropout. Finally, ground and location features are concatenated and used as input to the three classifiers(month, hour, and week).

Depending on whether the cascading technique is utilized, the input for each classifier is provided in a different way. First, with the cascading models, the output from the month classifier as well as the combined features (or ground features only in the case of the Without location model) are concatenated and fed into the hour branch. The output from the hour prediction is then added to these concatenated features, which are provided as the input for the week classifier. The non-cascaded model also receives the combined features of ground and location input. However, each classifier makes a prediction independently. The output from the previous classifier does not cascade into the next classifier as its input.

Each classifier branch consists of three fully connected layers. The first fully connected layer has 256, and the second has 128 neurons. We use the ReLU activation function on the two layers. Then we apply dropout with a 0.5 ratio on the first and 0.3 on the second, and the batch normalization gets applied on both layers. Finally, the last fully connected layer includes linear activation (with 12, 53, and 24 neurons for the month, week, and hours predictions, respectively) followed by the Softmax function.

3.3 Implementation Details

To implement the proposed architecture, we use Keras 2.2.4 with TensorFlow. We preprocess ground-level images and location data prior to passing them into the models. The input images are augmented by zooming and horizontally flipping. Input images are resized to 224×224 and scaled to [0,1]. The location input (latitude and longitude) is converted to ECEF coordinates (Earth-centered-Earth-fixed) and normalized to [-1,1]. All the models are trained for 100 epochs and optimized using Adam optimizer (Kingma & Ba, 2014) with an initial learning rate of 0.01 and a batch size of 64. We apply L2 regularization with the value of 0.0001 for each fully connected layer.

4. EVALUATION

In this section, we evaluate the accuracy of the different models. Then we illustrate how their accuracy changes over the different thresholds.

Timestamp	Network	Top 1	Top 3	Top 5
Month	Not cascaded	49.76	85.88	93.63
	Cascaded w/o location	50.78	85.86	**93.53**
	Cascaded	**53.31**	**86.44**	92.81
Week	Not cascaded	13.64	40.10	58.73
	Cascaded w/o location	14.72	40.36	58.69
	Cascaded	**15.32**	**43.37**	**62.78**
Hour	Not cascaded	22.00	57.01	75.76
	Cascaded w/o location	24.02	60.20	77.14
	Cascaded	**26.66**	**62.08**	**78.35**

Table 1: Time estimation accuracy (%) yielded for considered networks (Cascaded, Not cascaded, and Cascaded without location input).

4.1 Quantitative Evaluation

Using the test set, we evaluate how well our models estimate the capture time (month, week, and hour) of a given image. Table 1 shows the accuracy of each network. Our results show that the Cascaded network performs the best, followed by the Cascaded without location and Not cascaded models. With the Cascaded network, which outperforms the other models, we compute Top-K accuracy with different thresholds and present it in Figure 5.

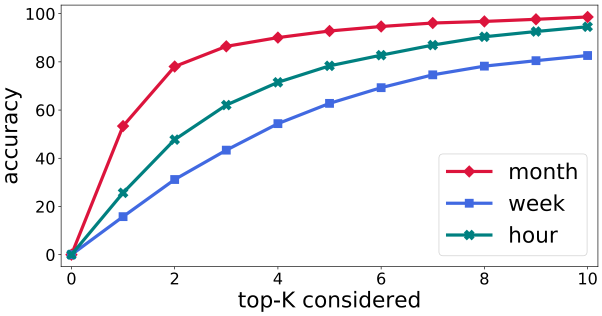

Figure 5: Time estimation accuracy for month, week, and hour obtained with the proposed approach (Cascaded).

4.2 Time Estimation

In this experiment, we explore how our models learn the relationship between scene appearances and time. When we provide ground-level images, our models estimate when those images were taken in terms of month, week, and hour. For instance, as can be seen in Figure 1, our models predict the month of capture, yielding high probabilities around the ground-truth month.

4.3 Capturing Temporal Patterns

We analyze how our models identify and estimate temporal trends in scene appearances that shift over a period of time (e.g., during the day or across the year). We find how the models can capture the patterns in two different scenarios and compare their results. As presented in Figure 6, our models can estimate the different times in the same place while showing the temporal patterns. The models can generate curves of

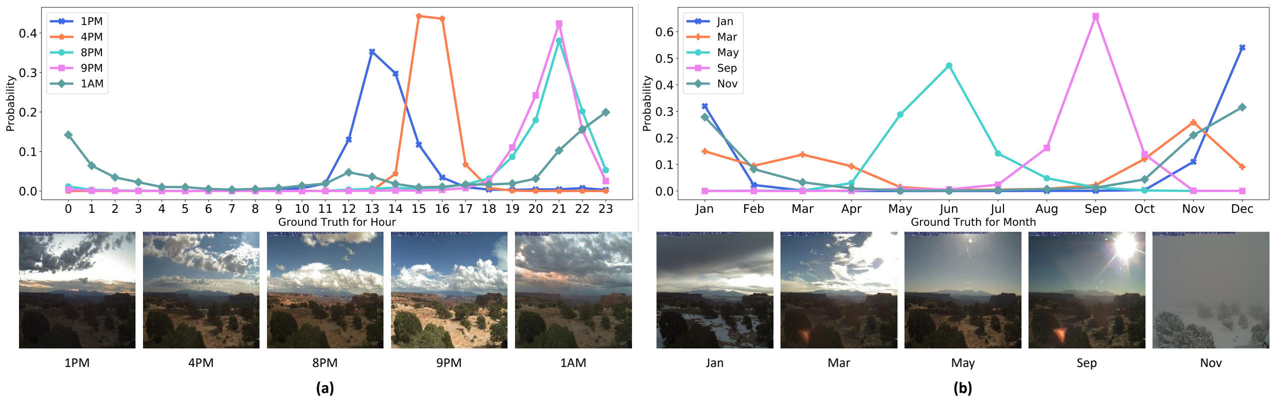

Figure 6: Visualization of consistency probability. Both (a) and (b) show the same location; (a) presents images captured at different hours on a fixed date in September and (b) displays pictures from different months in a year on a fixed hour, at 3PM (UTC). Our models are able to learn various temporal patterns that may change as the appearance of the place alters over time.

consistency probability, with each of them reaching the highest point on or near the ground-truth of the capture time.

5. CONCLUSION

We introduced a high-level CNN model that estimates the time-of-capture (month, week, and hour of the day) of ground-level photos. We trained the model in a multi-task manner with a novel cascading technique, where top-level temporal features are fed to estimate low-level attributes, boosting the network performance in time estimation. Despite the accuracy improvement, a potential drawback of the cascading approach could be the propagation of errors from top-to-bottom-level temporal attributes. As for future work, we will explore the impact of such errors on the overall performance, as well as include additional contexts (e.g., co-located satellite imagery, weather measurements) as input to the model for more accurate prediction.

Acknowledgements: We thank Dr. Nathan Jacobs for the helpful feedback he provided at different stages during this study.

REFERENCES

Derpanis, K. G., Lecce, M., Daniilidis, K., & Wildes, R. P. (2012). Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *IEEE conference on computer vision and pattern recognition* (pp. 1306–1313).

Ding, F., Zhu, G., Alazab, M., Li, X., & Yu, K. (2020). Deep-learning-empowered digital forensics for edge consumer electronics in 5g hetnets. *In IEEE consumer electronics magazine*.

Ginosar, S., Rakelly, K., Sachs, S., Yin, B., & Efros, A. A. (2015). A century of portraits: A visual historical record of american high school yearbooks. In *IEEE international conference on computer vision workshops* (pp. 1–7).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Jacobs, N., Roman, N., & Pless, R. (2007). Consistent temporal variations in many outdoor scenes. In *IEEE conference on computer vision and pattern recognition* (pp. 1–6).

Jae Lee, Y., Efros, A. A., & Hebert, M. (2013).

Style-aware mid-level representation for discovering visual connections in space and time. In *IEEE international conference on computer vision* (pp. 1857–1864).

Kakar, P., & Sudha, N. (2012). Verifying temporal data in geotagged images via sun azimuth estimation. *IEEE Transactions on Information Forensics and Security*, *7*(3), 1029–1039.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *In arXiv preprint arXiv:1412.6980*.

Laffont, P.-Y., Ren, Z., Tao, X., Qian, C., & Hays, J. (2014). Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics*, *33*(4), 149.

Lee, S., Maisonneuve, N., Crandall, D., Efros, A., & Sivic, J. (2015). Linking past to present: Discovering style in two centuries of architecture. In *IEEE international conference on computational photography.*

Li, X., Xu, W., Wang, S., & Qu, X. (2017). Are you lying: Validating the time-location of outdoor images. In *International conference on applied cryptography and network security* (pp. 103–123).

Li, Z., Xian, W., Davis, A., & Snavely, N. (2020). Crowdsampling the plenoptic function. In *European conference on computer vision* (pp. 178–196).

Martin, P., Doucet, A., & Jurie, F. (2014). Dating color images with ordinal classification. In *ACM international conference on multimedia retrieval* (p. 447).

Martin-Brualla, R., Radwan, N., Sajjadi, M. S., Barron, J. T., Dosovitskiy, A., & Duckworth, D. (2021). Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *IEEE conference on computer vision and pattern recognition* (pp. 7210–7219).

Padilha, R., Salem, T., Workman, S., Andaló, F. A., Rocha, A., & Jacobs, N. (2022). Content-aware detection of temporal metadata manipulation. *IEEE Transactions on Information Forensics and Security*, *17*, 1316–1327.

Palermo, F., Hays, J., & Efros, A. A. (2012). Dating historical color images. In *European conference on computer vision* (pp. 499–512).

Salem, T., Workman, S., & Jacobs, N. (2020). Learning a dynamic map of visual appearance. In *IEEE conference on computer vision and pattern recognition* (pp. 12435–12444).

Salem, T., Workman, S., Zhai, M., & Jacobs, N. (2016). Analyzing human appearance as a cue for dating images. In *IEEE winter conference on applications of computer vision* (pp. 1–8).

Tsai, T.-H., Jhou, W.-C., Cheng, W.-H., Hu, M.-C., Shen, I.-C., Lim, T., . . . Hidayati, S. C. (2016). Photo sundial: estimating the time of capture in consumer photos. *Neurocomputing*, *177*, 529–542.

Vittayakorn, S., Berg, A. C., & Berg, T. L. (2017). When was that made? In *IEEE winter conference on applications of computer vision* (pp. 715–724).

Volokitin, A., Timofte, R., & Van Gool, L. (2016). Deep features or not: Temperature and time prediction in outdoor scenes. In *IEEE conference on computer vision and pattern recognition workshops* (pp. 63–71).

Yuan, J., Luo, J., & Wu, Y. (2010). Mining compositional features from gps and visual cues for event recognition in photo collections. *IEEE Transactions on Multimedia*, *12*(7), 705–716.

Zhai, M., Salem, T., Greenwell, C., Workman, S., Pless, R., & Jacobs, N. (2018). Learning geo-temporal image features. In *British machine vision conference.*