Recognizing and Curating Photo Albums via Event-Specific Image Importance (Supplementary Material)

Yufei Wang[1]
yuw176@ucsd.edu
Zhe Lin[2]
zlin@adobe.com
Xiaohui Shen[2]
xshen@adobe.com
Radomír Měch[2]
rmech@adobe.com
Gavin Miller[2]
gmiller@adobe.com
Garrison W. Cottrell[1]
gary@ucsd.edu

[1] University of California, San Diego
9500 Gilman Drive,
San Diego, USA

[2] Adobe Research
345 Park Avenue,
San Jose, USA

1 The ML-CUFED Dataset

Figure 1: Example of two birthday albums (both have the photo uploader's tag "birthday").

In Figure 1 we show an example of the need to collect the multi-label ML-CUFED dataset with because of albums with ambiguous or multiple event types. The two albums in Figure 1 are both labeled as birthday events in CUFED, but they can also fall into the category of casual family/friends gathering. These two event types are not mutually exclusive. Moreover, intuitively, we would consider the album on the right to be a more typical birthday event, with distinguishable elements such as birthday hats and cakes, while the album on the left is more of a casual family/friends gathering rather than an obvious birthday event. Therefore, collecting the event types and their proportion in one album from more peoples' views is necessary.

Categories	Event Types
All Event Types	Wedding, Birthday, Graduation, Protest, Personal Music Activity, Religious Activity, Casual Family/Friends Gathering, Group Activity, Personal Sports, Business Activity, Personal Art Activity, Architecture/Art, Urban Trip, Cruise Trip, Nature Trip, Theme Park, Zoo, Museum, Beach Trip, Show, Sports Game, Christmas, Halloween
Top 10 event types of two-label albums	(Personal Sports, Sports): 68, (Urban Trip, Architecture/Art): 27, (Zoo, Nature Trip): 22, (Show, Personal Music Activity): 22, (Casual Family/Friends Gathering, Group Activity): 17, (Birthday, Casual Family/Friend Gather): 16, (Halloween, Group Activity): 12, (Beach Trip, Cruise Trip): 8, (Show, Group Activity): 8

Table 1: 23 Event types of ML-CUFED, and most frequent event type pairs of 2-label albums with their occurrence.

Table 1 shows all the 23 event types in ML-CUFED dataset, and the most frequent event type pairs of 2-label albums. Overall, there are 363 albums with multiple labels, about 20% of the ML-CUFED dataset.

In Figure 2, we show three examples of albums with multiple labels. These albums contain a mixture of different event types. For example, Figure 2(b) is a Christmas night event in a theme park (the fourth image in the second row shows "Merry Christmas" with the Christmas lights, better seen if zoomed in), therefore the multi-label: (Christmas & Theme Park) is more reasonable than the single label Christmas.

2 Details of Joint Album Recognition-curation System

2.1 Architecture of Event Curation Network

In Figure 3, we show the Curation-Siamese network used in the training stage for image importance prediction in the main paper. Similar to [3], we use a siamese network to predict the importance difference between an input image pair from an album given the ground-truth event type.

In [3], the siamese network predicts the absolute image score for each input image first, and then calculates the importance score difference, and a Piecewise Ranking Loss (PRL) is used as objective (shown in Figure 3 as Piecewise Ranking Loss(1)). This pathway is preserved in our architecture, and is denoted as **Pathway1**.

Unlike the architecture in [3], we add another pathway to directly predict the score difference between the image pair (as shown in the dotted box in the middle in Figure 3). We denote this extra pathway as **Pathway2**. This pathway concatenates the image features extracted from both input images (*fc7* layer features for AlexNet, or the 500-unit fully connected layer features after the *pool5* layer when using ResNet), and adds a 300-unit fully connected layer on top of the concatenated features, followed by a ReLU nonlinearity and dropout layer with 0.5 dropout rate. Then, the score difference between the image pair is directly predicted. The piecewise ranking loss is also used for this pathway, denoted Piecewise Ranking Loss (2).

In Pathway1, the siamese networks only see the two images separately, and predict the absolute importance score independently. However, Pathway2 adds a single network that sees both of the images, and directly predicts the score difference.

During the test stage, only one test image is fed into the trained network, with one importance score as the prediction from Pathway1. Though not used in the test stage, Pathway2 helps with the training of the network shared between both pathways, and effectively improves the performance of the network.

	MAP@t%				P@t%			
t%	5	10	15	20	5	10	15	20
AlexNet-Pathway1 [3]	0.298	0.362	0.417	0.469	0.199	0.300	0.354	0.407
AlexNet-Pathway1&2	0.305	0.368	0.421	0.472	0.211	0.307	0.362	0.412
ResNet-Pathway1 [3]	0.305	0.376	0.427	0.477	0.202	0.309	0.368	0.423
ResNet-Pathway1&2	**0.310**	**0.382**	**0.432**	**0.481**	**0.206**	**0.311**	**0.372**	**0.428**

Table 2: Comparison of the Curation-Siamese with only Pathway1, as used in [3], and the two pathway model used in this paper. Note that all the results shown here are obtained assuming ground-truth event types are known during the test stage.

In Table 2, we show the performance gained by using the two pathway model in Figure 3 versus training with only Pathway1. We show the comparison for both AlexNet and ResNet. There is a steady improvement with the use of Pathway2. The performance gain is about 0.5% on both MAP and Precision, which is similar to the gain from the use of 2-stage learning in [3].

2.2 Architecture of Recognition LSTM Network

The architecture of the Recog-LSTM network for album-wise event type prediction is shown in Figure 4. The album's images are first fed into the trained CNN for single images. For AlexNet, *fc7* features are extracted, and for ResNet, *pool5* features are extracted. The dimensionality of the features is then reduced to 512 with PCA, and the sequence of compressed features is fed into the LSTM network. The LSTM network we use is the same as the one described in [7]. The dimensionality of the hidden units is 512. The hidden-unit features for all the time frames are then averaged by the mean pooling layer over time. The mean hidden features are used as the features of the whole album, and are fed into the prediction layer for the final event type prediction.

AdaDelta is used to train the network. There are 1404 training albums in ML-CUFED. To overcome the overfitting problem, we subsample 20 sub-albums from one album. The sub-albums contain no less than 75% images of the original album.

(a) An example of a Birthday & Casual Family Gathering album. It was originally labeled as Birthday in CUFED.

(b) An example of a Christmas & Theme Park album. It is originally labeled as Christmas in CUFED.

(c) An example of an Urban Trip & Architecture/Art album. It is originally labeled as Architecture/Art in CUFED.

Figure 2: Examples of albums with multi-label in ML-CUFED, the original labels in CUFED are also shown. It is better to view digitally.

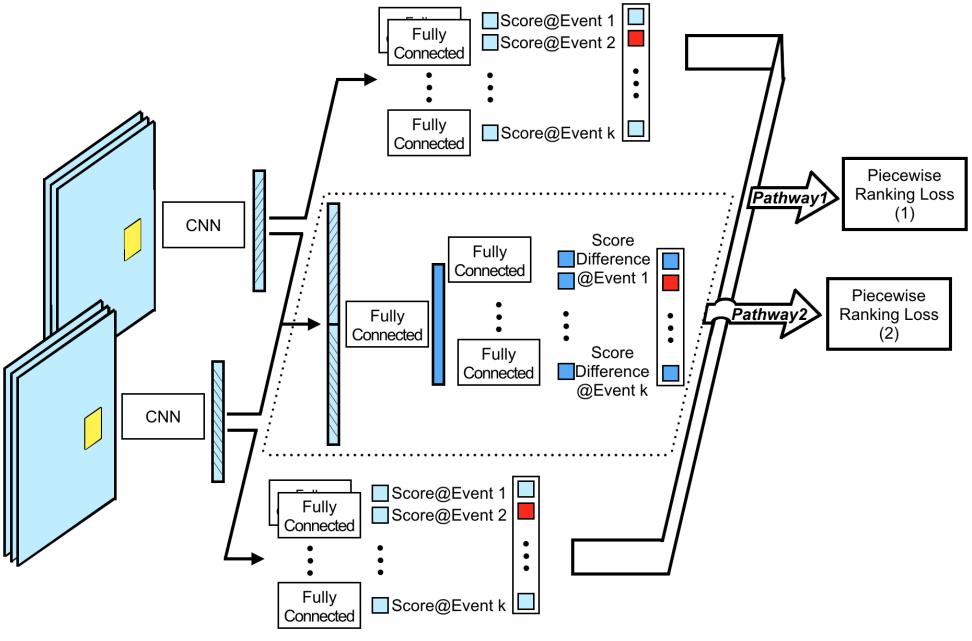


Figure 3: Architecture of the event curation siamese network (Curation-Siamese) during training. The "CNN" parts are the standard siamese network, the middle pathway that predicts score differences directly is novel in this application.

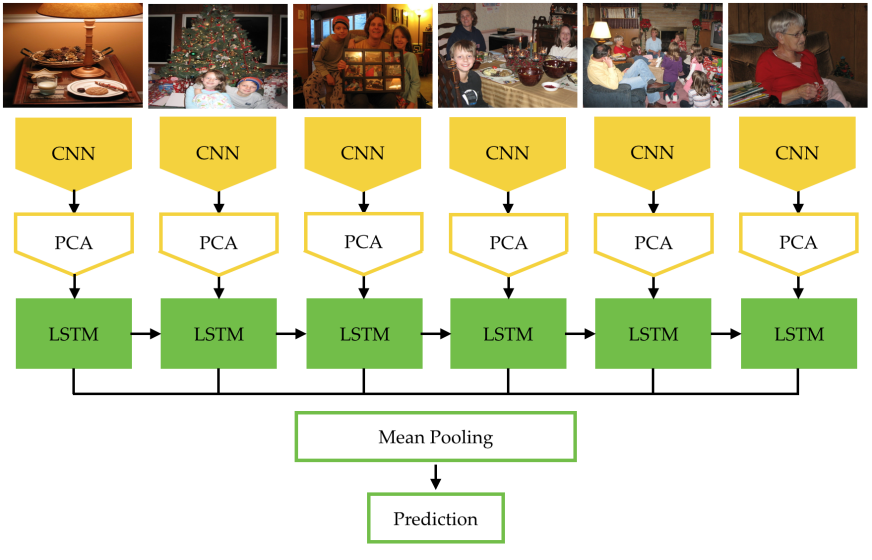


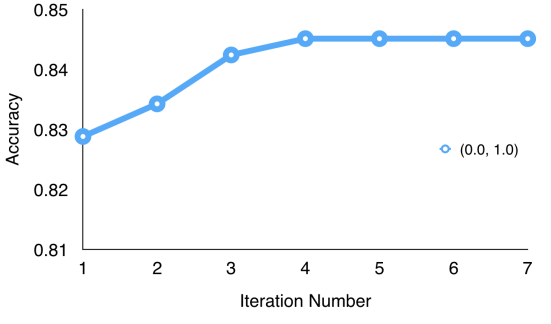
Figure 4: Architecture of the LSTM network for album-wise event recognition (Recog-LSTM)

3 Results

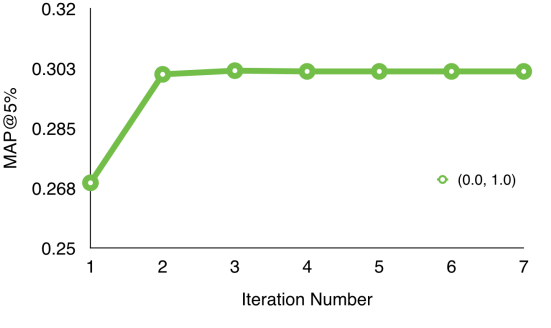
3.1 Performance over iterations

In the main paper, we described the curation-recognition procedure which iteratively updates the album-wise event type prediction and the image-wise importance score prediction. There are two hyper-parameters for the iterative procedure: $\theta = (m, \alpha)$. Here, m is a threshold (a fraction of the maximum probability) used to eliminate event types with low probability by setting their probability to 0. These are then ignored by the image importance prediction procedure; α is the emphasis we put on the image importance score for event type prediction. When $m = 0$, all event types are considered for image importance calculation; when $m = 1$, only one event type with highest probability is considered.

These hyper-parameters are determined using a 111-album validation set and running a grid search on choices of $\theta = (m, \alpha)$. Using ResNet features, the hyper-parameters are as follows: for ML-CUFED, $\theta = (0.0, 1.0)$. For PEC, $\theta = (1.0, 1.4)$. Using fine-tuned AlexNet features: $\theta = (0.3, 1.9)$ for ML-CUFED, and $\theta = (0.6, 1.1)$ for PEC. In Figure 5, we show the system performance with respect to the iteration number on ML-CUFED using ResNet features.



(a) Album-wise event recognition accuracy v.s. iteration number for hyper-parameters $(m, \alpha) = (0.0, 1.0)$ of the iterative curation-recognition procedure.



(b) Image importance prediction accuracy (MAP@5%) v.s. iteration number for hyper-parameters $(m, \alpha) = (0.0, 1.0)$.

Figure 5: Performance of our joint system with respect to iteration number on ML-CUFED using ResNet features.

As shown in Figure 5, both album-wise event recognition performance and image importance score prediction performance improve over iterations, and converges after a small number of iterations.

3.2 Event-specific Image Importance with AlexNet

In the main paper, we show the image importance prediction using different methods with ResNet. Here, we show the prediction results with AlexNet. Similar to the results with ResNet, CNN-Noevent performs a little better than CNN-Noevent (test). CNN-LSTM-Iterative greatly outperforms CNN-Noevent, with a steady 3% MAP increment. There is also a steady 3% increment for P at t < 20%. CNN-LSTM-Iterative closely approaches the upper bound (CNN-GTEvent), with a more notable performance gap between CNN-LSTM-Iterative and CNN-Noevent. However, compared with the results with ResNet, the improvement here is smaller: with AlexNet, 70% of the gap that exists between CNN-Noevent (test) and the results using the ground truth event type (CNN-GTEvent) is crossed by CNN-LSTM-Iterative, while with ResNet, the 79% of the gap is crossed. With AlexNet, 56% of the gap is crossed, while with ResNet, the 62% of it is crossed. This is because the ResNet features achieve better event type recognition performance (as in Table 2 in the main paper), and better event type recognition in turn helps improve the image importance score prediction result.

	MAP@t%						**P@t%**					
t%	5	10	15	20	25	30	5	10	15	20	25	30
Random	0.113	0.161	0.211	0.256	0.303	0.350	0.044	0.090	0.142	0.193	0.243	0.298
CNN-Noevent(test)	0.251	0.303	0.358	0.414	0.462	0.508	0.142	0.211	0.284	0.335	0.384	0.436
CNN-Noevent	0.258	0.316	0.369	0.425	0.475	0.519	0.168	0.245	0.307	0.373	0.422	0.468
CNN-LSTM-Iterative	**0.278**	**0.347**	**0.400**	**0.453**	**0.502**	**0.547**	**0.191**	**0.280**	**0.340**	**0.394**	**0.450**	**0.491**
CNN-GTEvent	0.305	0.372	0.424	0.476	0.522	0.565	0.218	0.304	0.361	0.417	0.461	0.504

Table 3: Comparison of event-specific image importance predictions using different methods with AlexNet. The evaluation metric here is MAP@t% and $P@t$%. We also show the score using a random ranking as a lower bound. We also provide a CNN-GTEvent result which uses ground-truth event type information when testing as an upper-bound.

3.3 More Qualitative Results on ML-CUFED

In this section, we show more examples of the qualitative results of our algorithm.

In Figure 7 and 8, several albums in ML-CUFED are shown. Figure 7 is an example of test results from AlexNet, and Figure 8 is using ResNet. For the examples shown here, the album-wise event type prediction is incorrect with the CNN recognition method, which simply averages the results from the classification of single images, but with the proposed CNN-LSTM-Iterative algorithm, the event type prediction is corrected. Also, we can see the ground-truth and predicted importance ranking of the images in each album. We also show the baseline image importance prediction results in the middle row. This baseline is achieved without event type prediction by just averaging the importance prediction across all event types. Note that there are equal ranks for multiple images in the ground-truth importance ranking. This is because the importance scores from 5 votes of Amazon Mechanical Turk(AMT) workers have a lot of ties. Therefore, the ranks shown here are the median ranking of all the images with the same score, and thus the ranks for the images with same ground-truth score are the same.

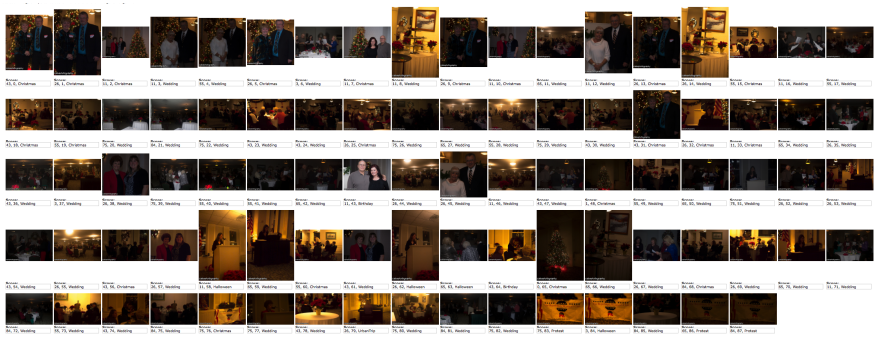
Only a fraction of images in the albums are shown here due to limited space. We deliberately choose both images with high ground-truth importance and low ground-truth importance for each album to show the overall quality of the albums.

For example, in the third example of a Cruise Trip album in Figure 7, there are many images of the iceberg and the sea similar to the last image shown here. If only CNN-recognition is used, and the prediction is produced by averaging the prediction of every image in the album, the album is recognized as a Beach Trip. However, after we assign different importance scores to images, as shown in the ranking of images, this album is correctly recognized as a Cruise Trip. Also, by comparing the image importance ranking between the second and third rows, corresponding to baseline importance prediction and CNN-LSTM-iterative importance prediction method, we can see that predicting the event type as a cruise increases the importance score of the first three images, which are more relevant to the event type, while decreasing the importance score of the photo of the cat and the selfie-like photo. The image rankings of the proposed method (in the third row) is obviously closer to the ground-truth ranking than the baseline method (in the second row). This demonstrates the advantages of the joint recognition-curation algorithm. In other words, our full method is able to rank important images (which are more indicative of event types) at the top.

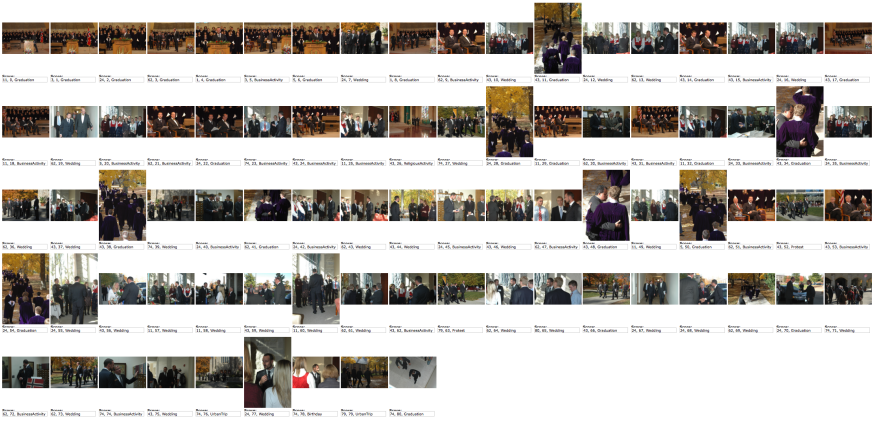
In Figure 6, we also show two examples of entire albums which are correctly recognized with our model but wrongly recognized with CNN-recognition baseline. For each image, we show three types of information under it: ground-truth importance ranking, predicted importance ranking, and single image event type prediction.

In Figure 9 and 10, more examples with correct event type prediction in ML-CUFED are shown. Figure 9 is from the network using AlexNet, and 10 is from the network using ResNet. We can also see how the images are ranked with predicted importance by the baseline algorithm and the proposed joint event recognition-curation algorithm. We can see many examples of better importance prediction results with the joint algorithm using the event type prediction, such as the first and second image in the first Birthday album in Figure 9; and the first three images in the third Wedding album in Figure 10.

In Figure 11 and Figure 12, we show some examples of incorrect event type prediction in ML-CUFED. Figure 11 is from AlexNet, and Figure 12 is from ResNet. In Figure 11, the ground-truth event type of the three example albums are book signing event (business activity), ball (group activity), and graduation party (graduation) respectively. In Figure 12, the ground-truth event type of the three example albums are Korean traditional wedding, Christmas family party, and casual friends gathering respectively.



(a) A Christmas album which is wrongly recognized as Wedding with CNN-recognition baseline, and is corrected recognized with CNN-LSTM-Iterative.



(b) A Graduation album which is wrongly recognized as Wedding with CNN-recognition baseline, and is corrected recognized with CNN-LSTM-Iterative.

Figure 6: Two albums in ML-CUFED, which are recognized wrongly with CNN-recognition baseline, but are corrected with CNN-LSTM-Iterative. Under each image, we show: (ground-truth importance ranking, predicted importance ranking, single image event type prediction). Images are sorted in predicted importance order. Better viewed in digital and zoomed in.

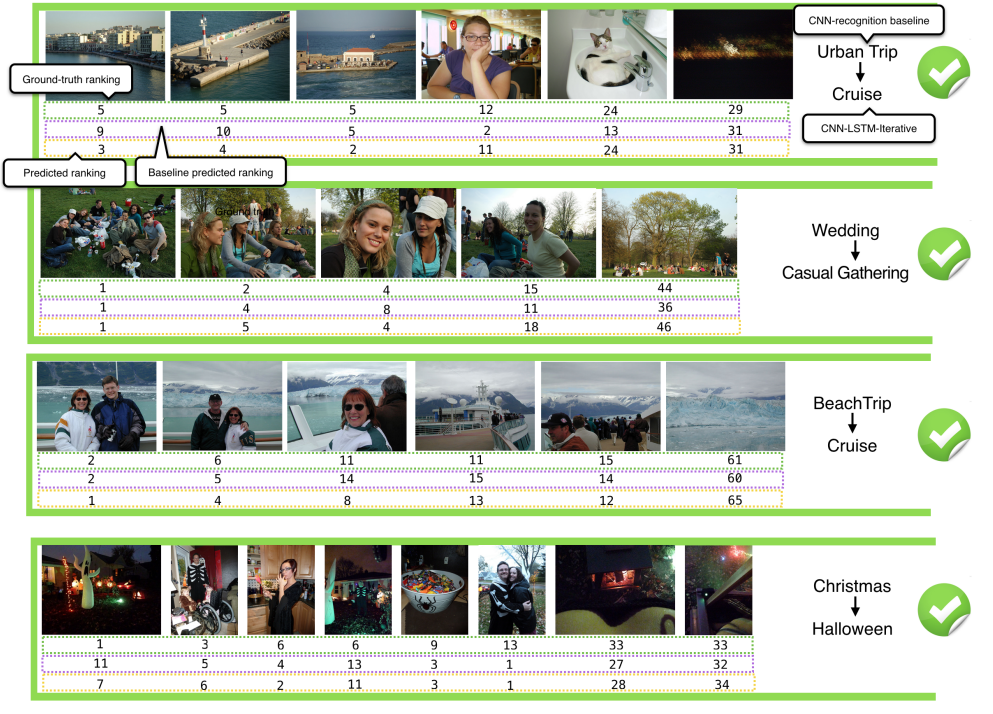


Figure 7: Examples of recognition-curation result from ML-CUFED using AlexNet. These examples were incorrectly categorized by the CNN-recognition method, but correctly categorized by CNN-LSTM-Iterative, as shown to the right of the album examples. Below the images, we show the ground-truth ranking of each image in the album, the baseline predicted image importance ranking, and the predicted importance ranking of each image in the three rows respectively. Baseline image importance prediction is done by not using event type prediction and just averaging the predicted importance score over all event types.

Figure 8: Examples of recognition-curation result from ML-CUFED using ResNet. These examples were incorrectly categorized by the CNN-recognition method, but correctly categorized by the CNN-LSTM-Iterative.

Figure 9: More examples of the recognition-curation results from ML-CUFED using AlexNet. The event types of albums are correctly recognized, as shown to the right of each album. Below each album, we show the ground-truth ranking, the baseline predicted importance ranking (not using event type prediction and averaging the predicted importance score over all event types), and the predicted importance ranking with proposed iterative method for each image in the three rows respectively.

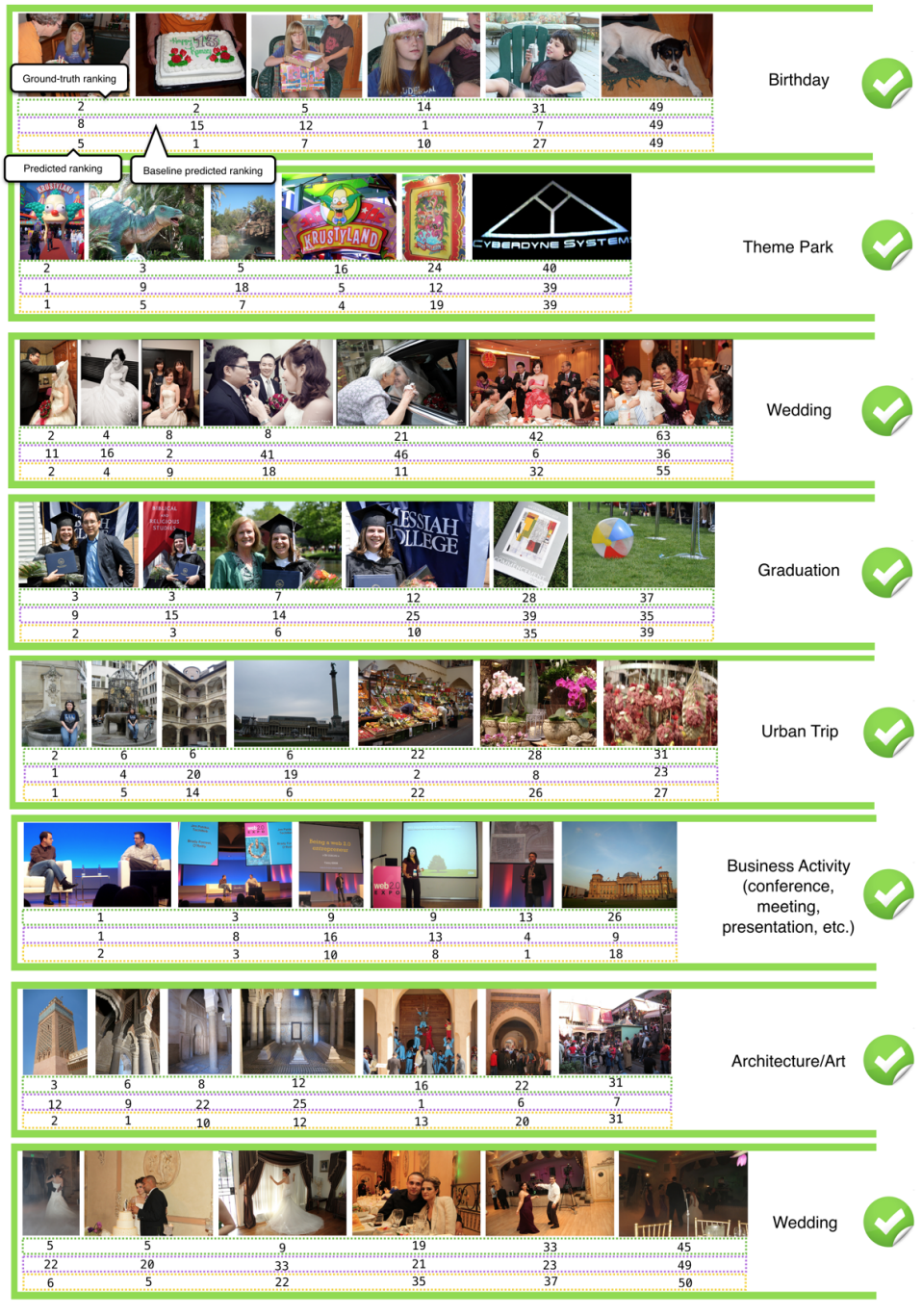


Figure 10: More examples of recognition-curation results from ML-CUFED using ResNet. The event types of albums are correctly recognized, as shown in the right of each album.

Figure 11: Examples of recognition-curation result from ML-CUFED using AlexNet, whose event types are predicted incorrectly. The predicted event type and the ground-truth event type are shown in the right of each album. The ground-truth event type is shown in parenthesis.

Figure 12: Examples of recognition-curation result from ML-CUFED using ResNet, whose event types are predicted incorrectly.

3.4 More Qualitative Results on PEC

In Figure 13, We also show some examples of event recognition result on PEC dataset of our CNN-LSTM-Iterative system using ResNet. As in Section 3.3, we show two examples which are incorrectly categorized by the CNN-recognition method, but correctly categorized by the CNN-LSTM-Iterative in Figure 13(a); four examples which are correctly recognized in Figure 13(b); one example which is wrongly recognized by CNN-LSTM-Iterative in Figure 13(c).

There is no ground-truth image importance score in PEC, therefore in Figure 13 we only show the image importance rank predicted by CNN-LSTM-Iterative. For each album, we only show a fraction of images in it, but we deliberately choose both images with high predicted importance and low predicted importance.

We also show the PEC results using AlexNet in Figure 14.

3.5 Mapping from PEC Label to ML-CUFED Label

In the main paper, to show the generalizability of our algorithm, we showed our results when tested on the PEC dataset [1]. In PEC, there are several event types that are not contained in ML-CUFED, such as Saint Patrick's Day, Easter, and Skiing, and there are two event types that can map to single event type in ML-CUFED: Children's Birthday and Birthday can be mapped to single Birthday event in ML-CUFED. Therefore, we provide the mapping from PEC label to ML-CUFED label in Table 4 here. There are 9 event types in PEC after merging.

PEC	(Children's) Birthday	Christmas	Concert	Graduation	Exhibition
ML-CUFED	Birthday	Christmas	Show	Graduation	Museum
PEC	Halloween	Hiking, Road Trip	Wedding	Cruise	
ML-CUFED	Halloween	Nature Trip	Wedding	Cruise	

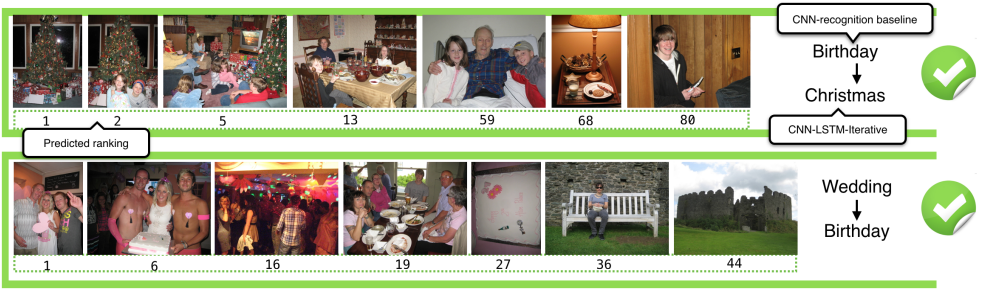
Table 4: Event type matching from PEC Dataset to ML-CUFED Dataset.

(a) Examples of two albums that are incorrectly categorized by CNN-recognition method, but correctly categorized by CNN-LSTM-Iterative.

(b) Examples of four albums that are correctly recognized.

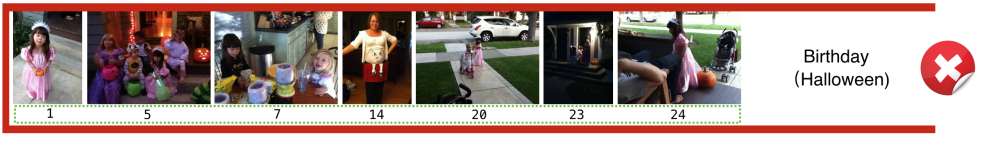
(c) An album whose event type is predicted incorrectly.

Figure 13: Examples of album recognition result on PEC dataset using ResNet. Rank of the predicted image importance is also shown for each image.



(a) Examples of two albums that are incorrectly categorized by CNN-recognition method, but correctly categorized by CNN-LSTM-Iterative.

(b) Examples of four albums that are correctly recognized.



(c) An album whose event type is predicted incorrectly.

Figure 14: Examples of album recognition result on PEC dataset using AlexNet.

References

[1] L. Bossard, M. Guillaumin, and L. Van. Event recognition in photo collections with a stopwatch hmm. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013.

[2] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[3] Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, and W. Cottrell, G. Event-specific image importance. In *Proc. Computer Vision and Pattern Recognition Conference (CVPR)*, 2016.