Event-specific Image Importance (Supplementary Material)

Yufei Wang[1] Zhe Lin[2] Xiaohui Shen[2] Radomír Měch[2] Gavin Miller[2] Garrison W. Cottrell[1]
[1]University of California, San Diego [2]Adobe Research
{yuw176, gary}@ucsd.edu {zlin, xshen, rmech, gmiller}@adobe.com

1. Dataset

In the main paper, we calculated Kendall's W of the image rating from 5 workers from Amazon Mechanical Turk (AMT) in order to measure the consistency among people's rating. We further tested the statistical significance of Kendall's W. As shown in the main paper, there are different percentages of significant albums for each event type, ranging from 58% to 98%. Here, in addition to the percentage of significant albums, Table 1 shows the average Kendall's W as well as Spearman's correlation ρ for each event type. We can see that $Wedding$ albums receive on average the highest correlation/agreement, while $PersonalArtActivity$ albums receive the lowest score.

Figure 1 shows an example of the ground truth we obtained from AMT. Scores are normalized so that the range is (0,1), 1 being most important and 0 being totally irrelevant. The images are sorted by the predicted image importance by our algorithm. It's best viewed electronically. Note that all images are stretched and distorted for viewing.

As mentioned in the main paper, for each album, we used Spearman's rank correlation (ρ) and Kendall's W to evaluate the consistency from AMT workers' rating. Here we show some examples of albums in Figure 2 with their average ρ and W. Note that the average ρ and W are for an individual album.

Figure 2a-2b show two examples of albums that receive relatively high correlation and agreement from 5 AMT workers, and Figure 2c-2d show two examples of albums that receive low correlation and agreement from 5 AMT workers. Figure 2d shows an example in which all the images are of similar quality and semantics, and it's hard for people to agree on the ranking.

2. Architecture of Face Heatmap CNN

In the main paper, we mentioned that the face heatmap CNN uses the similar siamese CNN architecture to train simultaneously for 23 event types. Here in Figure 3, we show the exact architecture we used for Face Heatmap network.

3. Result

In this section, we present both quantitative results and qualitative results in addition to the main paper.

3.1. Quantitative Results

In Figure 4, we show the comparison of MAP@t%5 by six methods for each of the 23 event types. The six methods being compared are: random ranking, aesthetics, K nearest neighbors with pre-trained CNN features (KNN), single network with Euclidean loss (Euclidean), siamese network with ranking SVM loss (Ranking-SVM), and our method using Ensemble of siamese CNNs (Ensemble). We also show a "worker" method here for comparison. It is calculated as follows: for each album, we have 5 rankings from 5 workers, and we can calculate the MAP score for each worker's rating against the ground truth. Then all the MAPs over all albums are averaged for one event type. The "worker" method is to measure how workers did on those albums.

As shown, our method outperforms all the other methods in most cases, except for *Personal Art Activity*, *Architecture*, *Business Activity*, *Protest* and *Nature Trip*. Our method can even beat "worker" in some cases.

As mentioned in the main paper, the aesthetic score doesn't perform well overall, however, it has good performance on two event types: *Personal Art Activity* and *Nature Trip*. Especially for *Nature Trip*, aesthetics achieves the best performance over all methods.

In the main paper, we mentioned that we used grid search on 5-fold cross validation to decide the parameters $\{\alpha, \beta, \lambda\}$ to incorporate the face heatmap. Among 23 event types, only 10 event types showed a performance gain after face information was incorporated in the validation set, and therefore face information was only used for these 10 event types. Table 2 shows the effect of incorporating face information for these 10 event types.

In Table 3, we show the comparison of the results from single network with Euclidean loss (Euclidean), our method using Ensemble of siamese CNNs (Ensemble-CNN), and

Categories	Important Personal Event	Personal Activity	Personal Trip	Holiday
Event types and # albums	Wedding (0.486/0.548) Birthday (0.418/0.423) Graduation (0.413/0.427)	Personal Music Activity (0.425/0.447) Protest (0.418/0.446) Religious Activity (0.401/0.406) Casual Family Gather (0.383/0.369) Personal Sports (0.372/0.347) Business Activity (0.368/0.335) Group Activity (0.366/0.357) Personal Art Activity (0.339/0.280)	Architecture/Art (0.428/0.452) Theme Park (0.391/0.385) Museum (0.391/0.384) Cruise Trip (0.383/0.371) Urban Trip (0.372/0.349) Beach Trip (0.370/0.368) Show (0.366/0.336) Zoo (0.366/0.337) Nature Trip (0.357/0.321) Sports Game (0.349/0.288)	Halloween (0.395/0.397) Christmas (0.386/0.379)

Table 1: 23 Event types, and the (average Kendall's W score / average Spearman's correlation ρ) for each album. The event types fall into four categories.

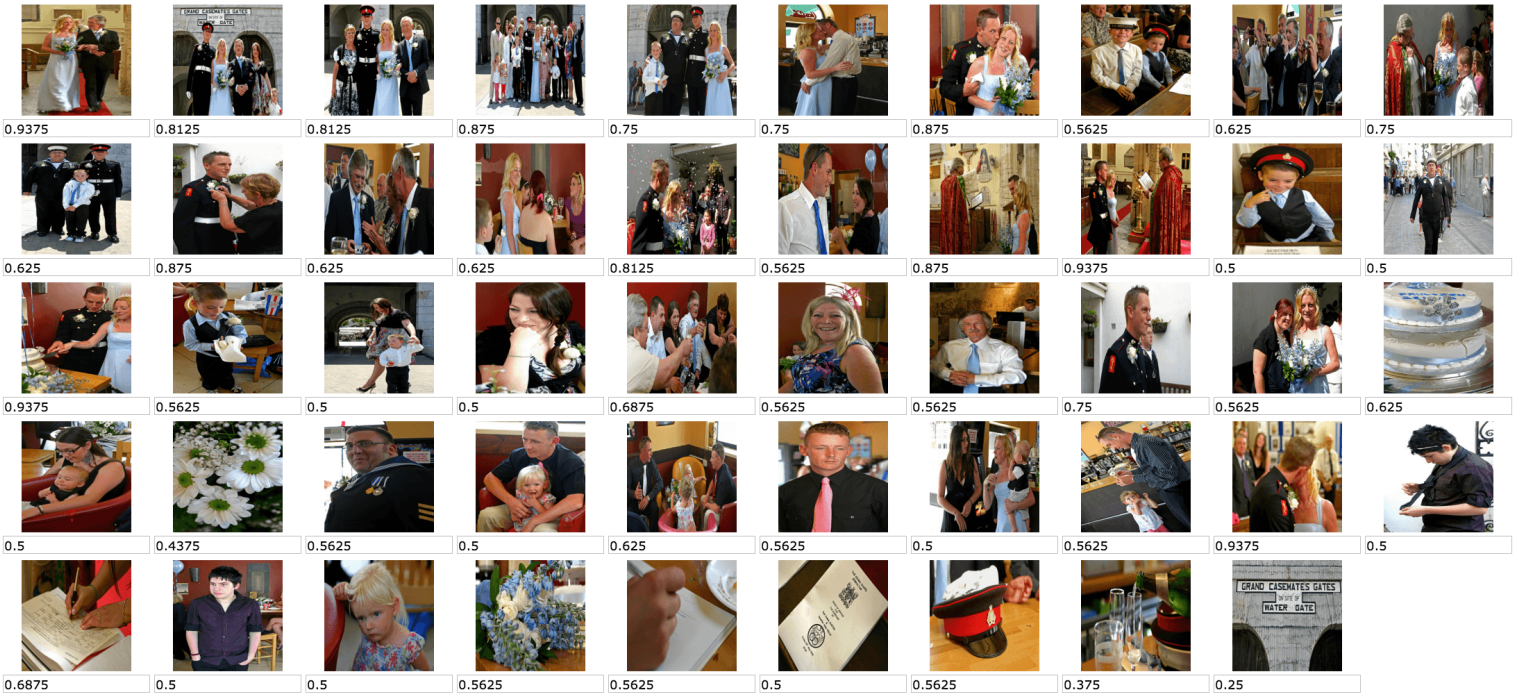


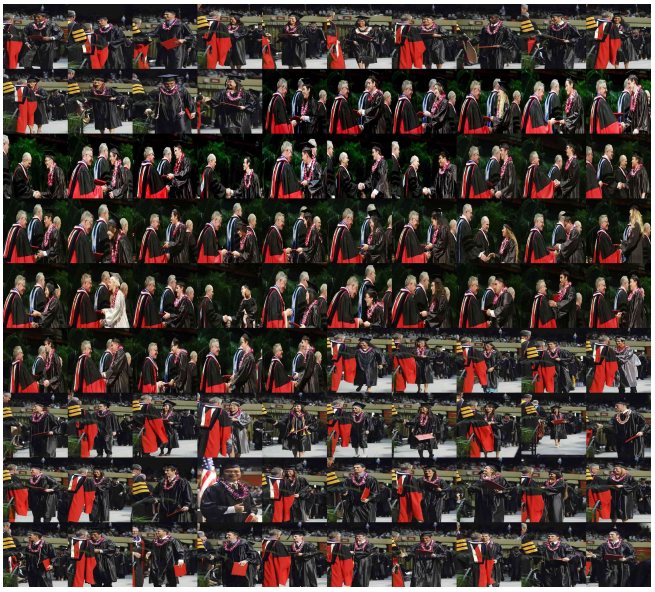
Figure 1: Example of a *Wedding* album with ground truth. Ground truth is obtained from 5 AMT workers. Here, the ground truth score of each image is given under it. The images are sorted by the predicted image importance by our algorithm. The average Spearman's correlation ρ over all possible two workers-three workers splits for this album is 0.49. It is best viewed electronically.

our method after incorporating face information (Ensemble-CNN + face). This is in addition to the result table we presented in the main paper. As mentioned in the main paper, with face information, MAP is slightly improved by about 0.1%, and the face heatmap network helped very little.

(a) A *Wedding* album. Spearman's Correlation $\rho = 0.78$, Kendall's $W = 0.64$

(b) A *Birthday* album. Spearman's Correlation $\rho = 0.61$, Kendall's $W = 0.49$

(c) A *Zoo/Botanic garden* album. Spearman's Correlation $\rho = 0.02$, Kendall's $W = 0.19$



(d) A *Graduation* album. Spearman's Correlation $\rho = -0.09$, Kendall's $W = 0.17$

Figure 2: Examples of albums in our dataset and the Spearman's Correlation ρ and Kendall's W from worker's rating for each album.

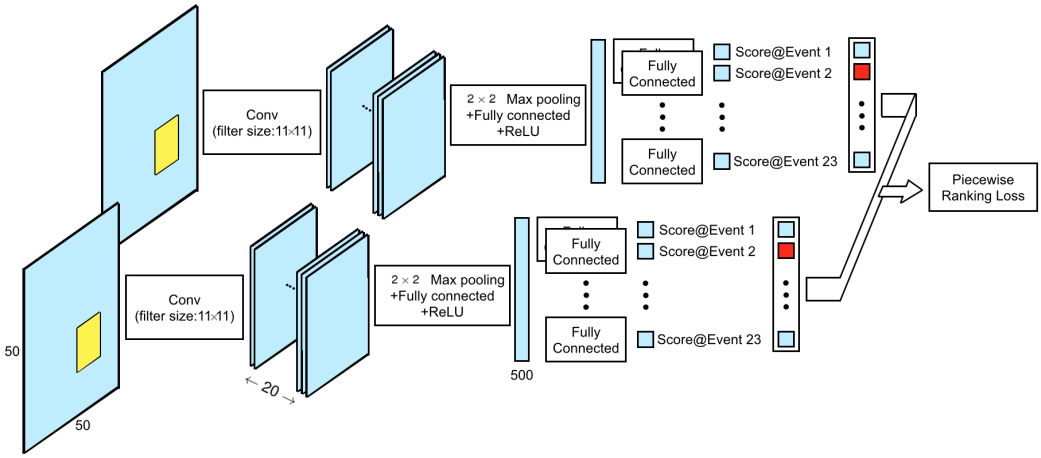


Figure 3: Face Heatmap CNN architecture

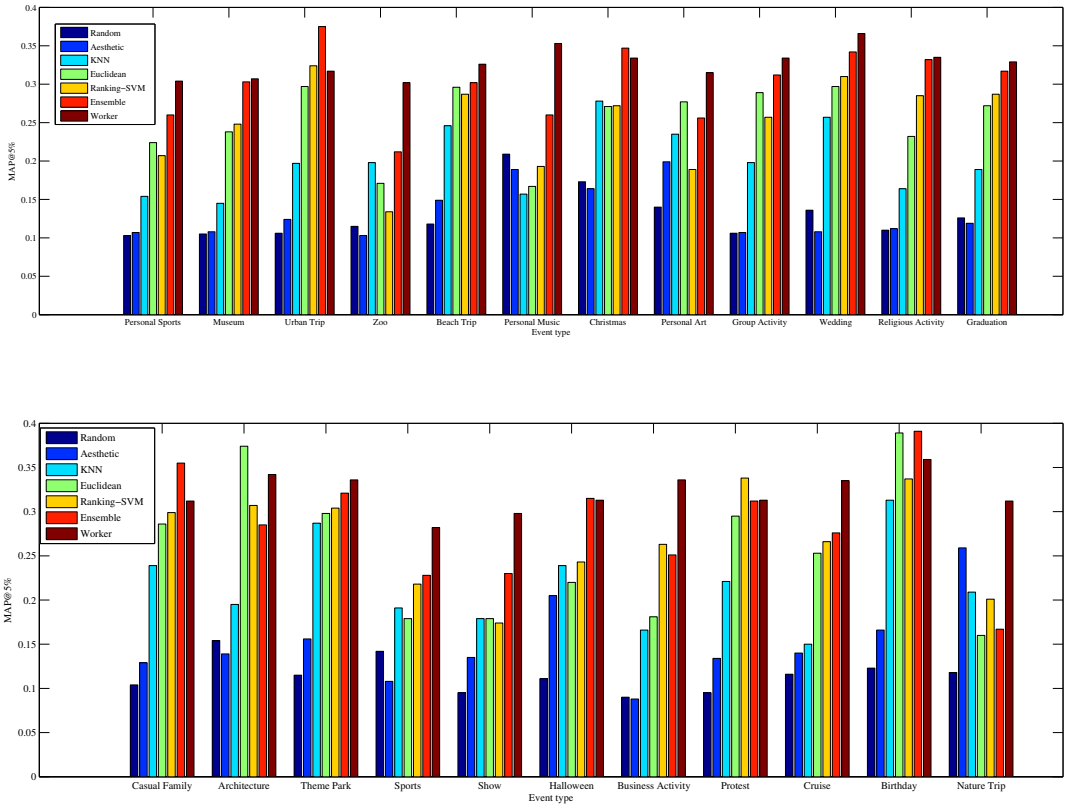


Figure 4: Comparison of six methods for 23 event types respectively. Individual worker's performance is also included as comparison. Results of MAP@t%5 are shown.

t%	5	15	25
Beach Trip	0.353(+0.051)	0.455(+0.022)	0.555(+0.011)
Nature Trip	0.167(+0.008)	0.272(+0.008)	0.369(+0.007)
Group Activity	0.315(+0.003)	0.489(+0.001)	0.586(+0.003)
Halloween	0.315(+0.000)	0.424(+0.001)	0.529(+0.002)
Personal Art Activity	0.256(+0.000)	0.361(0.002)	0.449(+0.000)
Religious Activity	0.320(-0.012)	0.416(0.000)	0.503(+0.005)
Graduation	0.317(+0.001)	0.444(0.002)	0.548(+0.001)
Sports	0.228(+0.001)	0.322(0.002)	0.420(+0.002)
Show	0.232(+0.002)	0.356(0.002)	0.473(+0.001)
Museum	0.293(-0.010)	0.367(-0.010)	0.453(-0.006)

Table 2: For a given event type, MAP@$t\%$ for the Ensemble-CNN after using the face information. The difference between before v.s. after face information is shown in parentheses. All the 10 event types for which face information is used are shown here.

	MAP@$t\%$						P@$t\%$					
t%	5	10	15	20	25	30	5	10	15	20	25	30
Euclidean	0.266	0.329	0.389	0.444	0.494	0.540	0.173	0.260	0.328	0.391	0.439	0.485
Ensemble-CNN	0.305	0.364	0.417	0.471	0.519	0.563	0.216	0.301	0.360	0.411	0.459	0.504
Ensemble-CNN + face	0.306	0.364	0.418	0.472	0.520	0.563	0.215	0.303	0.360	0.413	0.460	0.503

Table 3: Comparison of predictions using different methods that weren't shown in the main paper. Evaluation metric here is MAP@$t\%$ and P@$t\%$.

3.2. Qualitative Results

In addition to the visual example of our method's performance in the main paper, we show more examples of our method. Here we present 52 examples randomly selected from all 23 event types from Figure 5a to Figure 5w. For each album, we show top 10-20% images of the album from three methods. For each method, the top photos returned are arranged in chronological order. (Each album has different size, while we want to constrain the number of images we show to make it easier to view.) First row is the ground truth we acquired from AMT worker; second row is our prediction using Ensemble-CNN which we introduced in the main paper; third row is the result from random selection. Note that the images are distorted for viewing.

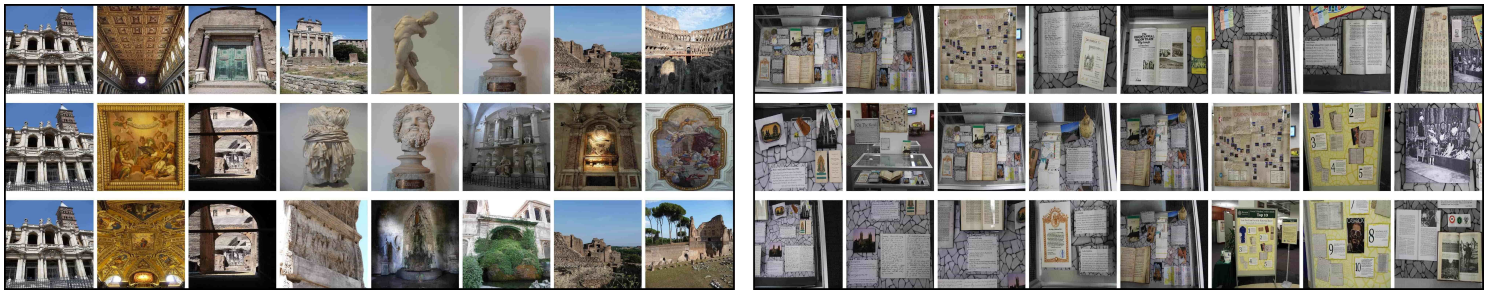
We can see that for most albums that have strong narrative structure or albums that consist of images that vary much in quality or semantics, our method's results are close to, though do not perfectly match the ground truth result; on the contrary, the results from random selection are obviously less appealing (for example, second album in Figure 5v, both albums in 5j, first album in 5p, etc.). For instance, in first two albums in Figure 5a, our method captures the important moments of the wedding events, similar to those people picked (in the ground truth); however random selection has many images that are less important, for example, photos of people eating, or photos of guests talking, while not looking at the camera.

There are also some albums in which most of the images are of similar quality or semantics, for example, albums in Figure 5q and 5r.

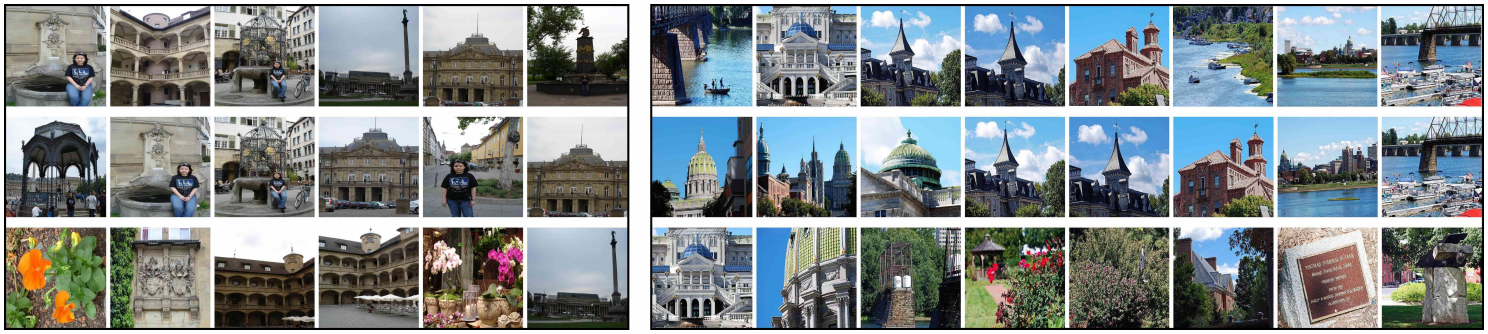
(a) Examples of 4 *Wedding* albums. Top 20%, 20%, 10%, 20% of images are shown respectively.

(b) Examples of 4 *Graduation* albums. Top 10%, 10%, 10%, 20% of images are shown respectively.

(c) Examples of 4 *Birthday* albums. Top 20%, 20%, 20%, 20% of images are shown respectively.



(d) Examples of 2 *Museum* albums. Top 15%, 20% of images are shown respectively.



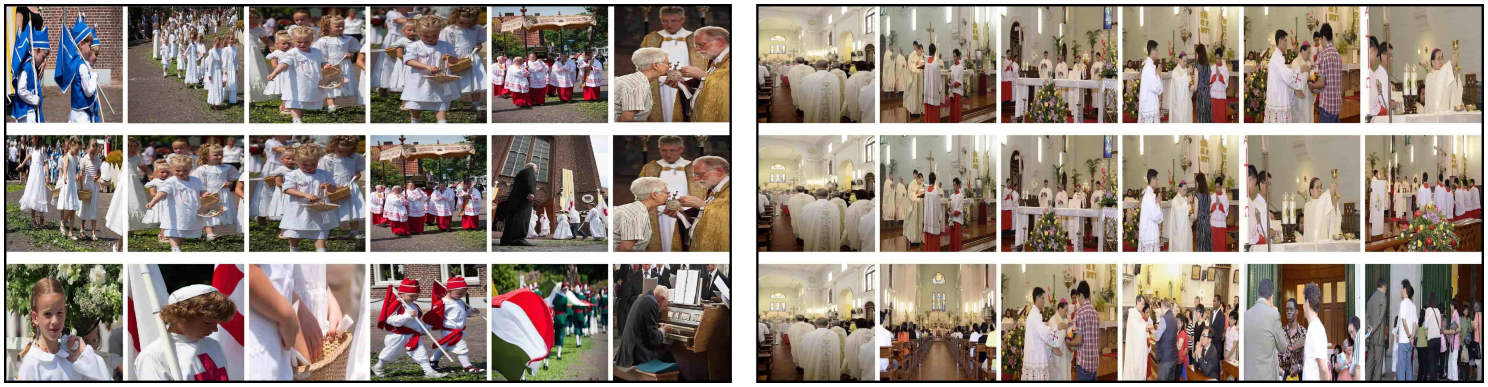
(e) Examples of 2 *Urban Trip* albums. Top 20%, 10% of images are shown respectively.

(f) Examples of 2 *Personal Sports* albums. Top 20%, 20% of images are shown respectively.

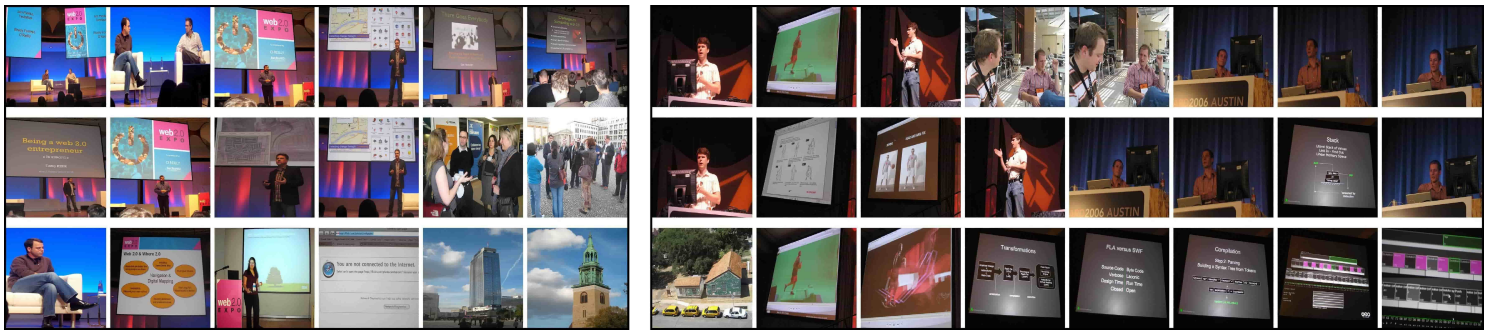
(g) Examples of 2 *Cruise Trip* albums. Top 10%, 20% of images are shown respectively.

(h) Examples of 2 *Protest* albums. Top 20%, 15% of images are shown respectively.

(i) Examples of 2 *Christmas* albums. Top 10%, 15% of images are shown respectively.



(j) Examples of 2 *Religious Activity* albums. Top 20%, 20% of images are shown respectively.

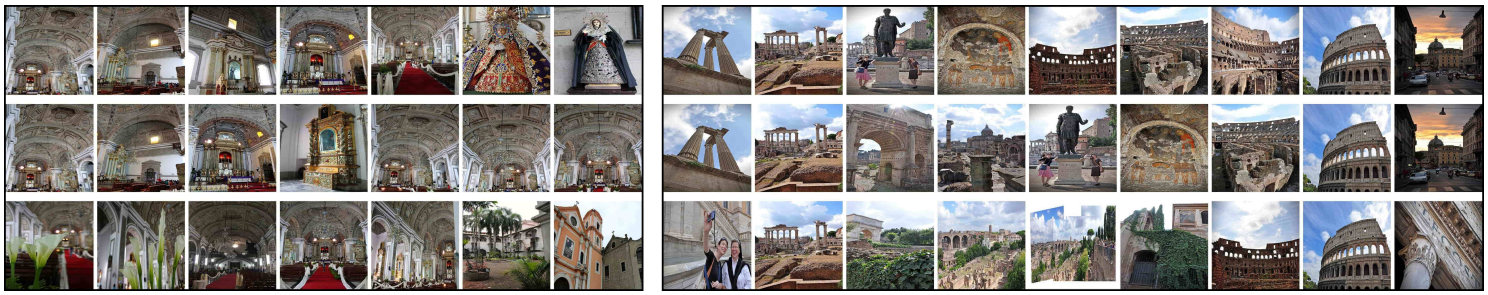


(k) Examples of 2 *Business Activity* albums. Top 20%, 20% of images are shown respectively.

(l) Examples of 2 *Sports Game* albums. Top 20%, 20% of images are shown respectively.

(m) Examples of 2 *Theme Park* albums. Top 20%, 20% of images are shown respectively.

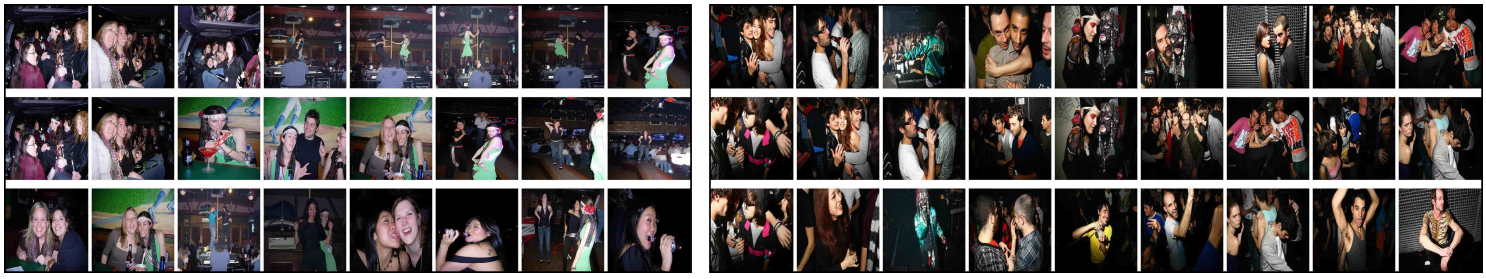
(n) Examples of 2 *Show/Parade* albums. Top 20%, 20% of images are shown respectively.



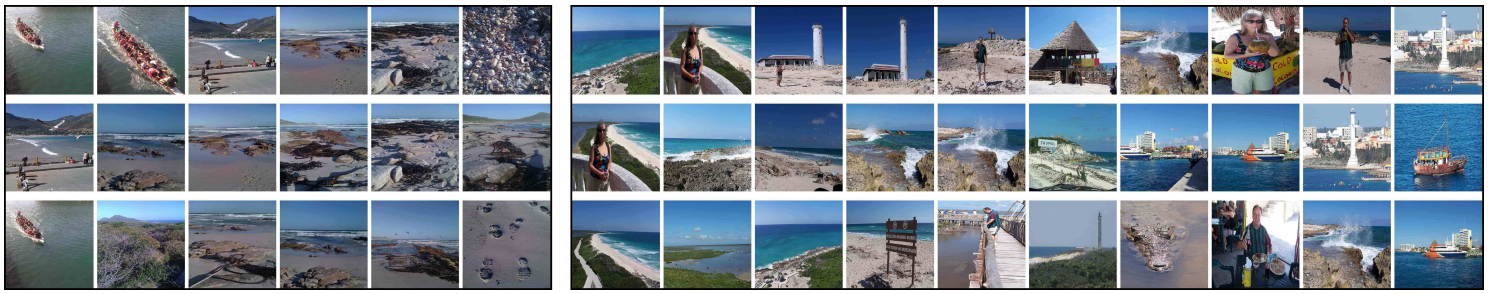
(o) Examples of 2 *Architecture/Art* albums. Top 20%, 20% of images are shown respectively.

(p) Examples of 2 *Personal Art Activity* albums. Top 20%, 15% of images are shown respectively.

(q) Examples of 2 *Zoo* albums. Top 20%, 15% of images are shown respectively.



(r) Examples of 2 *Group Activity* albums. Top 20%, 15% of images are shown respectively.

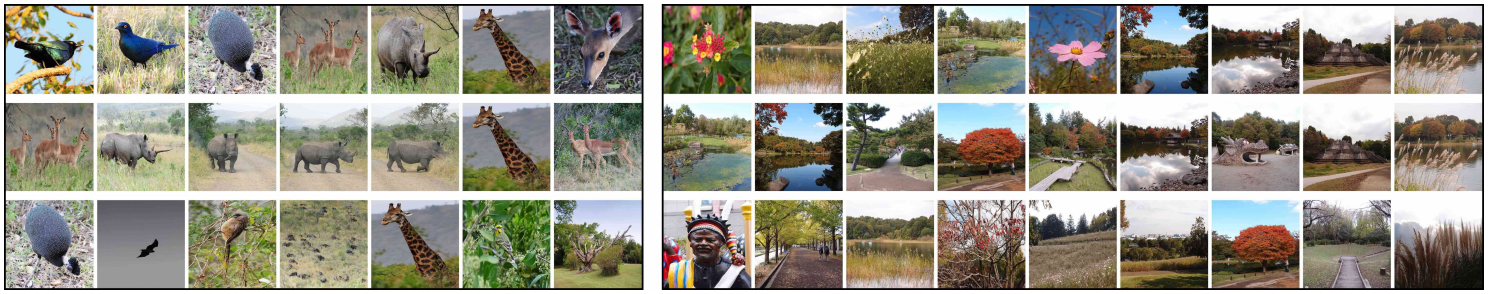


(s) Examples of 2 *Beach Trip* albums. Top 20%, 15% of images are shown respectively.

(t) Examples of 2 *Personal Music Activity* albums. Top 15%, 15% of images are shown respectively.

(u) Examples of 2 *Halloween* albums. Top 20%, 20% of images are shown respectively.

(v) Examples of 2 *Casual Family/Friends Activity* albums. Top 20%, 20% of images are shown respectively.



(w) Examples of 2 *Nature Trip* albums. Top 20%, 15% of images are shown respectively.

Figure 5: Example of results. For each album, top 10-20% images of the album from three methods are shown. Images are arranged in chronological order. For each album, first row is the ground truth we acquired from AMT workers; second row is our prediction using Ensemble-CNN which we introduced in the main paper; third row is the result from random selection.