Optimizing Visual Element Placement via Visual Attention Analysis (Supplementary Material)

Category: Research

1 Data Collection Experiment

1.1 Examples of Visual Elements Used in Data Collection

To control for the effects of attractiveness of the visual elements on their own viewability, we randomized the elements used in each mission. At the start of each mission an element was randomly selected from a pool of 48 and assigned to each panel in the environment. In order to avoid bias that may be introduced by the visual element's content, we used visual elements sharing similar styles in the pool—see Figure 1 for some examples of visual elements shown to participants.

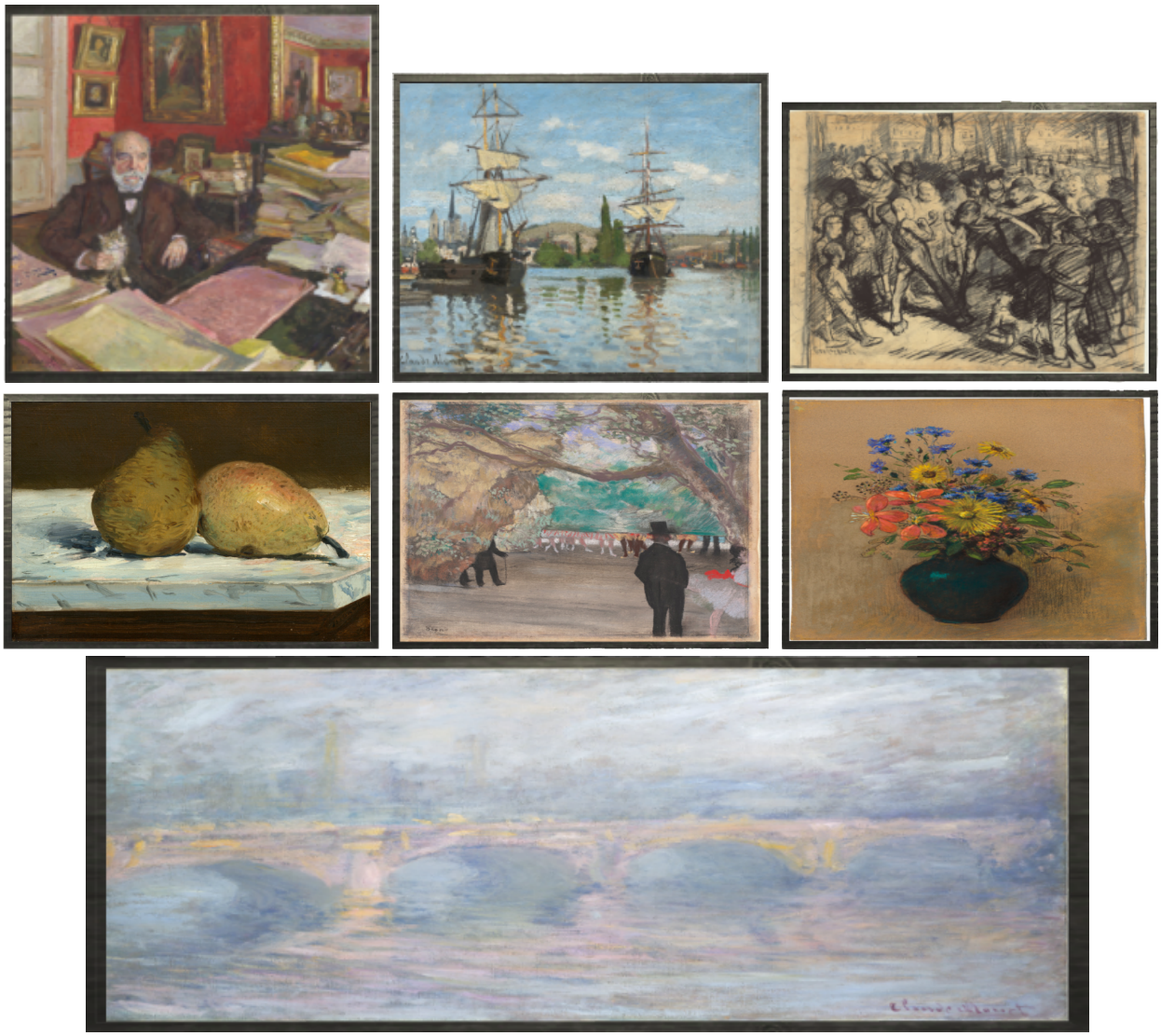
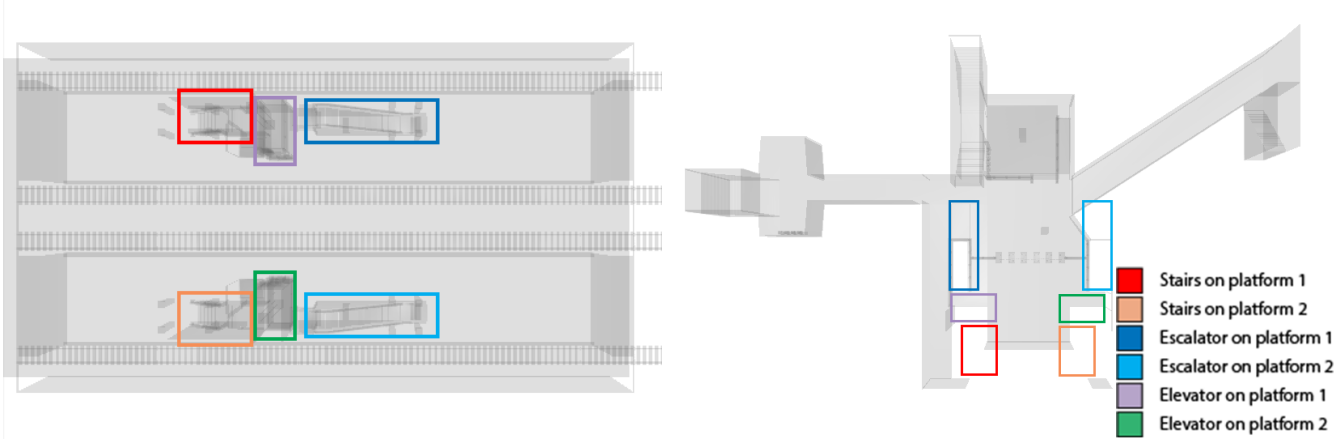


Figure 1: Examples of visual elements with different sizes used.

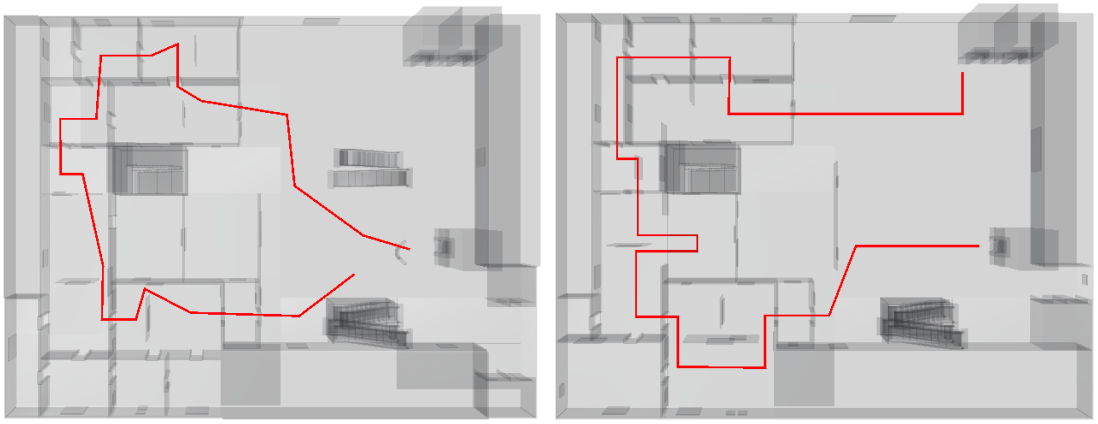
1.2 Scene Layout

Figure 2 shows the layouts of *Subway L1* and *L2* used in our data collection experiment. During the data collection experiments participants were required to complete missions in these two floored scenes. In *Museum L1* and *L2*, participants were automatically moved from one floor to the other after the mission was complete. However in the *Subway* scenes participants could willingly move from one floor to the other using the stairs, elevators or escalators as can be seen in Figure 2.

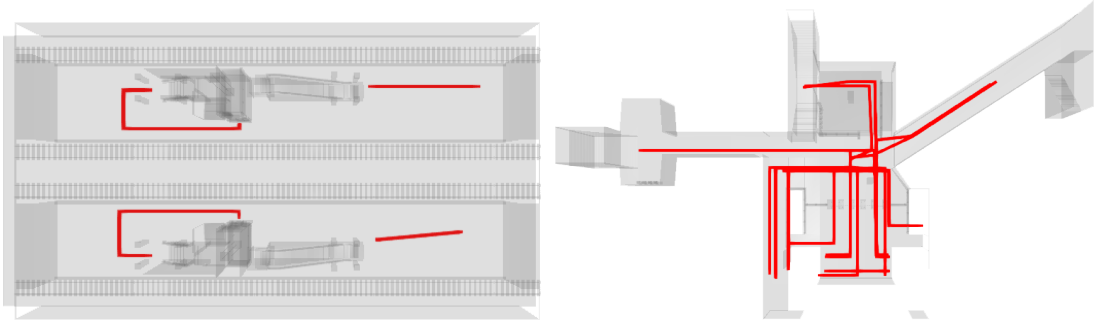


(a) *Subway L1* (b) *Subway L2*

Figure 2: Users can move freely between *Subway L1* and *L2* using the stairs, elevators or escalators marked with colored rectangles. For example, the red rectangles in (a) and (b) correspond to the stairs that connect between *Subway L1* and *L2*.



(a) *Museum L1* (b) *Museum L2*



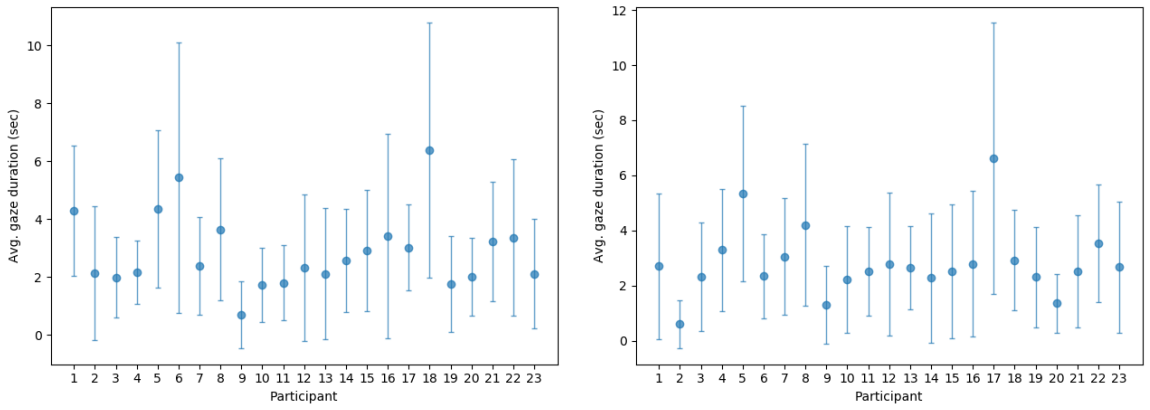
(a) *Subway L1* (b) *Subway L2*

Figure 3: *Museum* and *Subway* navigation paths used for computing the *distance from nearest path* feature for training the regressor.

1.3 Navigation Paths

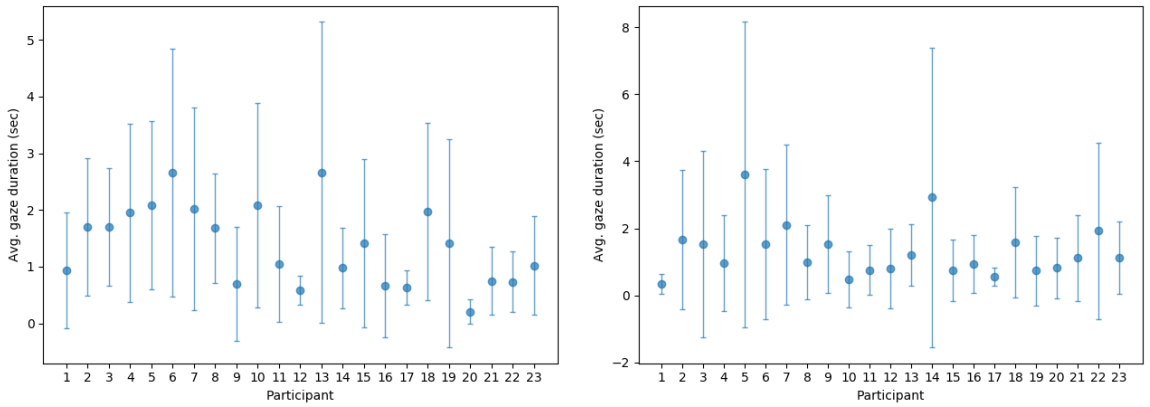
In order to compute the *distance from nearest path* feature, we define a set of paths that the participants are likely to traverse based on the missions. Because the *Subway* scenes are more goal-oriented, we specify 16 paths according to our predefined missions (4 in *L*1 and 12 in *L*2). As for the *Museum* scenes, we define multiple paths starting from the main elevator to one of the entrances of the inner-gallery and finally out from the other entrance of the inner-gallery. These paths are shown in Figure 3.

Each segment of the path is a straight line. To find the distance from a visual element to a path, we calculate the distance between the visual element's centroid and the nearest segment of the path. We assume that the visual element and the path have the same height. The same paths are used for training the regressors and for the optimization.



(a) *Museum L1* (b) *Museum L2*

Figure 4: The average and standard deviation total gaze duration for visual elements viewed by each participant in the data collection session. (a) shows the average total gaze duration for elements viewed in *Museum L1* scene, while (b) shows the average total gaze duration for elements viewed in *Museum L2* scene.



(a) *Subway Placement 1* (b) *Subway Placement 2*

Figure 5: The average and standard deviation total gaze duration for visual elements viewed by each participant in the data collection session. (a) shows the average total gaze duration for elements viewed in the first *Subway* scene, while (b) shows the average total gaze duration for elements viewed in the second *Subway* scene.

1.4 Dispersion of Gaze duration Across Participants

Figures 4 and 5 show the average total gaze duration that visual elements received for each participant while navigating the *Museum* and *Subway* scenes respectively in the data collection experiment. The average computation excluded visual elements that received zero gaze duration from the participant. Samples from both the test and training data sets were used in this computation.

1.5 User Feedback

We spoke to the participants after the data collection experiments. Two participants reported feeling dizzy while they were navigating in the virtual space. Their data was discarded and 23 participants remained in the final data set. We tried to alleviate dizziness by giving participants the opportunity to take a break after each mission. Nineteen of the 23 participants acknowledged their familiarity with the navigation set-up; and fifteen reported having previously played a VR game at least once. Consequently, participants were observed to become quickly familiarized with navigating our environments.

As for the difficulty of the missions, the majority of participants felt that the *Museum* levels were much easier than the *Subway* levels. They expressed their confusion in navigating the real-world Kenmore subway station we based our *Subway* design on as well. Unlike with the *Subway* scene, only a few participants recognized or were familiar with the Museum of Modern Art we modeled as *Museum*.

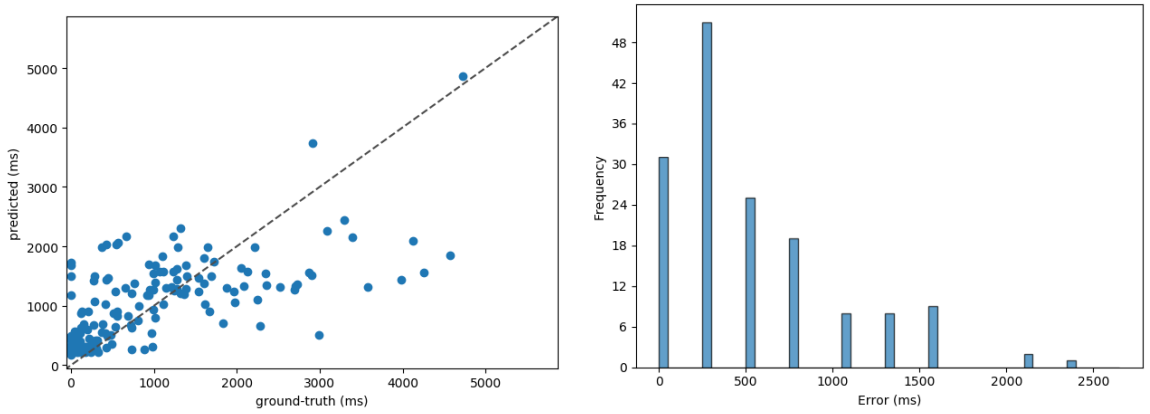
2 Analysis of Gaze Duration Prediction

In our preliminary trials, we separated our dataset according to scene type. The *Subway* dataset was split into 904 samples for training and 226 samples for the test set, while the *Museum* dataset was split into $1,755$ samples for training and 439 for testing.

We trained two regressors each using the *Subway* and *Museum* scene datasets separately. The *Subway* Epsilon-Support Vector Regressor (ε-SVR) was trained with an ε of 0.01 and a error term parameter C of $1,000$ on the *Subway* training set. The *Museum* ε-SVR was trained with an ε of 0.1 and C of 1. The decision trees were trained on their respective training sets with a maximum depth of 3 for the *Subway* and a maximum depth of 5 for the *Museum*. The *Subway* random forest regressor was trained using the *Subway* dataset with 4 trees with a maximum depth of 6. While the *Museum* random forest regressor was trained using the *Museum* dataset with 9 trees with a maximum depth of 5. We used these hyper- parameters as they produced the highest accuracy using grid-search with 10-fold cross-validation.

	Subway		***Museum***		***Subway & Museum***	
	RMSE(%)	**RMSE(ms)**	**RMSE(%)**	**RMSE(ms)**	**RMSE(%)**	**RMSE(ms)**
Support Vector Regressor	7.75	2,058	13.46	2,294	*7.54*	*2,002*
Decision Tree	7.15	1,899	13.41	2,286	*6.77*	*1,797*
Random Forest	7.13	1,894	13.53	2,307	*6.70*	*1,780*

Table 1: Prediction error of our *test set* of 226 and 439 *Subway* and *Museum* examples respectively using different types of regressors. The ms error was computed by scaling up the $[0,1]$ output of our regressors by the maximum values (2,656 ms) for *Subway* and (1,705 ms) for *Museum*.



(a) Scatter plot of ground truth and predicted gaze duration

(b) Mean absolute error (MAE) distribution

Figure 6: Visualizing the random forest regressor's prediction. In (a), the predicted gaze durations are plotted against the ground-truth gaze durations. (b) visualizes the mean absolute error (MAE) of the ground-truth and predicted values in (a) as a histogram. As shown in (b), most error (80%) is less than 1,000 ms.

In a 10-fold cross-validation done on the 904 sample *Subway training set*, we obtained a root mean squared error of 1,714 ms, 1,560 ms and 1,575 ms for the support vector machine, decision tree and random forest respectively. While a 10-fold cross-validation done on the 1,755 *Museum training set* resulted in root mean squared errors of 2,532 ms, 2,307 ms and 2,304 ms for the support vector machine, decision tree and random forest respectively. Table 1 shows the Root Mean Square Error (RMSE) achieved by the 6 regressors each on their respective test set.

Table 1 shows the root mean square errors on the test sets. Combining the *Subway* and *Museum* into a 3,045 training set and a 762 test set resulted in overall lower root mean square error regressors. In practice, it is simpler to train and apply one regressor for all scene types as apposed to training and applying one regressor per scene. Moreover, a generalizable predictor is more preferable compared to a specialty predictor. Consequentially, we chose to use a combined data set to train and test our final predictor.

Figure 6 visualizes our regressor's performance. An element's ground-truth gaze duration was computed using the average gaze duration of the element in the test set. There are a total of 157 unique elements in our 762 sample test set. These 157 elements' ground-truth (average gaze duration) and predicted (using our random forest regressor) gaze durations were used to produce the scatter plot and histogram in Figure 6. Because our target gaze durations were scaled to $[0,1]$ prior to training, our regressor predicts gaze durations within that range. As a result, the output of our regressors should be scaled up by the maximum gaze duration in the training set (2,655 ms). Figure 6 shows the predicted gaze durations scaled by that factor.

Moreover, we utilized Monte-Carlo simulation to provide statistical evidence that our regressors can correctly predict the gaze durations of visual elements using their 12-feature vector. On the test set, we computed the correlation coefficient $r = 0.64$ between the predicted and ground-truth gaze durations previously mentioned. By shuffling the target values (gaze durations) of the samples in our training set, we created 10,000 surrogate data sets. We trained a random forest regressor using each surrogate set and computed the correlation coefficient r_i (where $i \in [1,10,000]$) using the values predicted by these regressors. We estimated the p-value of the correlation coefficient r by comparing it with correlation coefficients $r_1, r_2, ...r_{10,000}$ given by random forest regressors trained with surrogate data sets. The p-value for the original regressor was computed by estimating the probability that the regressors trained with surrogate data sets outperformed the regressor. The p-value was estimated to be less than 0.0001. As a result, we can conclude that the prediction performance of our regressor was significantly better than that of random prediction. In addition, we found $r_{svr} = 0.44$, $r_{dt} = 0.61$ and $r_{rf} = 0.64$ to be the correlation coefficients that are produced by comparing the support vector, decision tree and random forest regressor's predictions with the average gaze duration of each element in the test set.

Figure 8 shows additional results comparing the ground-truth gaze duration with the predicted gaze duration for *Museum L1*, *Museum L2*, and *Subway L1 & L2*. Please refer to the figure caption for more details.

We observed relatively high prediction error on gaze durations for visual elements placed at the corner in the scenes while the prediction performed reasonably well in terms of the overall mean absolute error (e.g., 71.99 ms for *Museum L1*) as shown in Figure 8. For example, in *Museum L1* (first row), errors of the elements near the upper-right corner have a ground-truth gaze duration of 4,515 ms and 4,250 ms, but are predicted by the regressor to have a gaze duration of 797 ms and 1,328 ms. We observed in our data collection process for the *Museum* scenes that participants usually started by cycling the inner-gallery. Once done, they usually spent the remaining time of the session viewing the isolated elements. Because there were only a few isolated elements, they received a larger chunk of the participant's time knowing that he had viewed all the other elements in the museum.

Also notice that in Figure 8, the mean absolute error is lower for *Subway L1* and *Subway L2* than for the two *Museum* scenes. The relatively

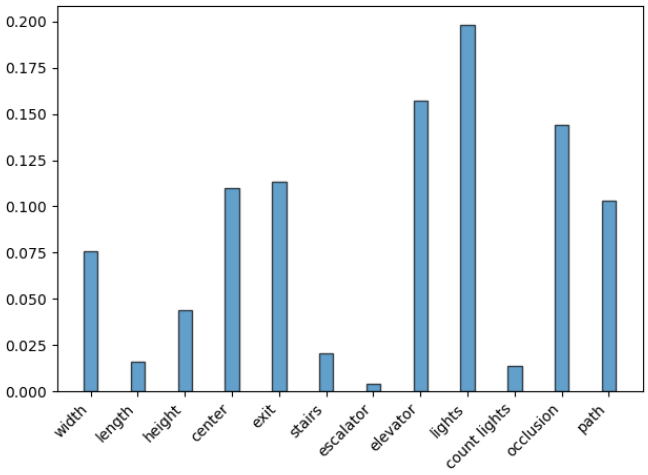


Figure 7: The Gini importance [1] of our 12 features. Higher Gini importance for a feature indicates the feature splits the data set with decreased node impurity (or is a great split). The feature importances are normalized and thus their values sum to 1.

high error is likely caused by our model's inability to predict human paths accurately. In the *Subway* scenes, we presented the participants with clear missions instead of allowing them to roam freely like in the *Museum* scenes. This made it easier for us to define navigational paths in the *Subway* scenes and therefore improve the accuracy in those scenes. Automatically estimating user navigation paths from our eye gaze data set could be a useful extension to our work.

To show a comparison between different features for predicting visual attention, we show the Gini importance of the features in Figure 7, which depicts the importance of location-based features (e.g., distances from center, exit, elevator, navigational paths) versus element-based features (e.g., dimensions).

REFERENCES

[1] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

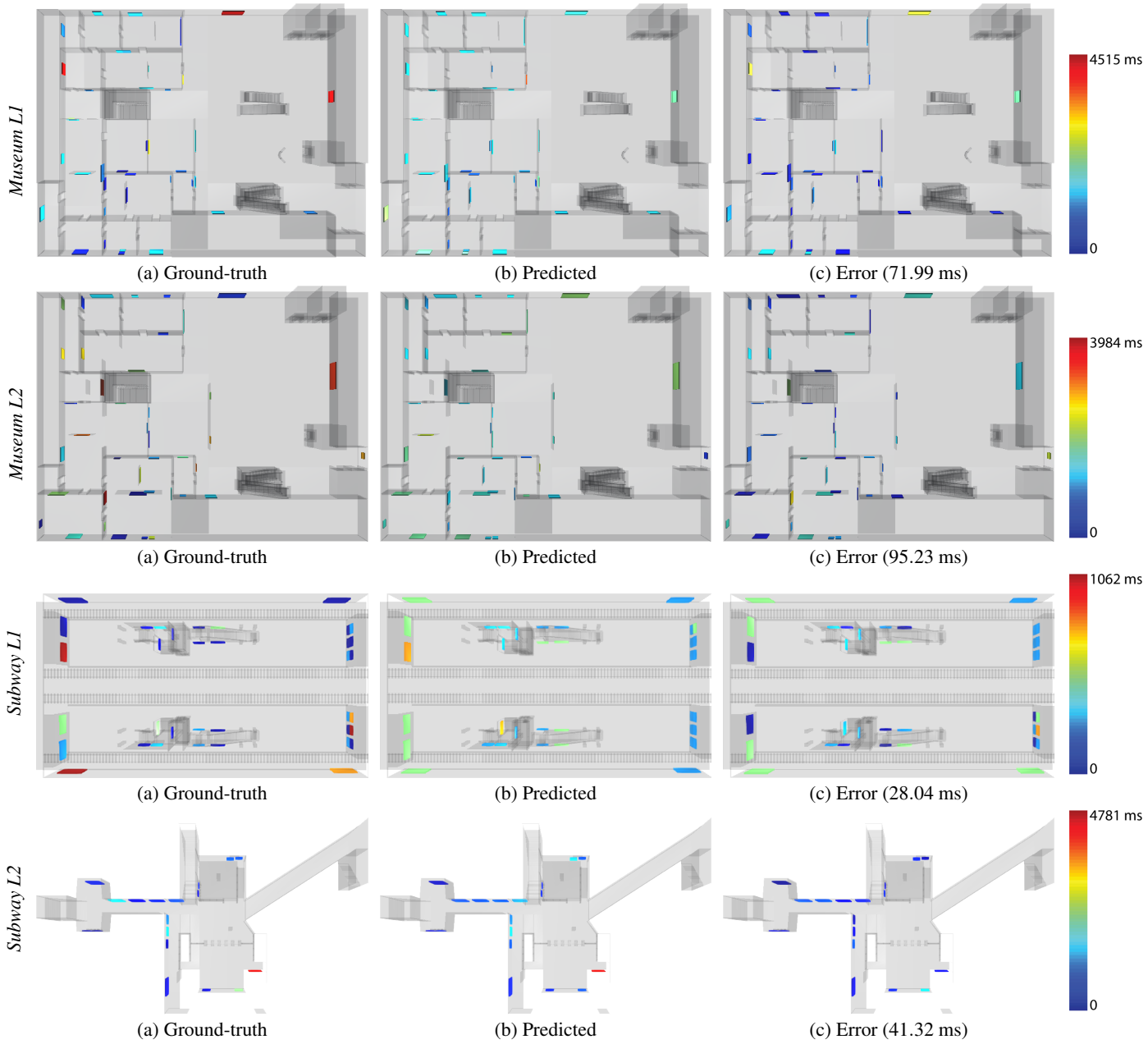


Figure 8: The error distribution in each level by comparing the ground-truth gaze duration with the predicted gaze duration. The results of *Museum L1*, *Museum L2*, *Subway L1* and *Subway L2* are shown. For each scene, the (a) ground-truth gaze duration (defined as the average gaze duration of the visual elements in our test set), (b) the predicted gaze duration by the regressor, as well as (c) the mean absolute error (MAE) between the ground-truth gaze duration and the predicted gaze duration are shown. The ground-truth gaze duration was computed from 203 samples in *Museum L1*, 218 samples in *Museum L2*, 143 samples in *Subway L1* and 87 samples in *Subway L2*.