Reproducing the Unified Model of Saliency and Importance

Chuhan Jiao
Aalto University
Finland
chuhan.jiao@aalto.fi

ABSTRACT

The Unified Model of Saliency and Importance (UMSI) can predict human attention on both natural images and graphic designs. The performance of UMSI model is comparable to state-of-the-art models for natural image saliency and it outperforms the existing visual importance models. The pre-trained UMSI model and the inference code are available online. However, the training code of UMSI model is missing. To verify the details in the original UMSI paper, in this work we reproduce the UMSI model and run the same evaluation as in the original paper. The evaluation results show our reproduced model rivals the official UMSI model.

KEYWORDS

Visual Importance, Visual Saliency, Human-Computer Interaction

ACM Reference Format:
Chuhan Jiao. 2018. Reproducing the Unified Model of Saliency and Importance. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In graphic design, visual importance refers to the level of design elements grabbing the viewer's attention. A crucial task for designers is to make the information that they want to convey have relative higher visual importance. Therefore, a model which can predict visual importance of graphic designs can be a useful tool for designers. The model can guide designers to adjust the design elements or even re-targeting the layout of the design elements for designers.

The state-of-the-art visual saliency model trained with natural images does not have good generalization ability in visual importance prediction of graphic designs [8]. The previous visual importance model for graphic designs performs poorly in saliency prediction[14]. This model is trained on the dataset with limited types of graphic designs. However, in real world applications, there are many graphic design classes (e.g. advertisements, movie posters). The contents of design, such as layout and design elements, are different in different design classes. Thus, designers need a special model which can predict visual importance for all kinds of designs.

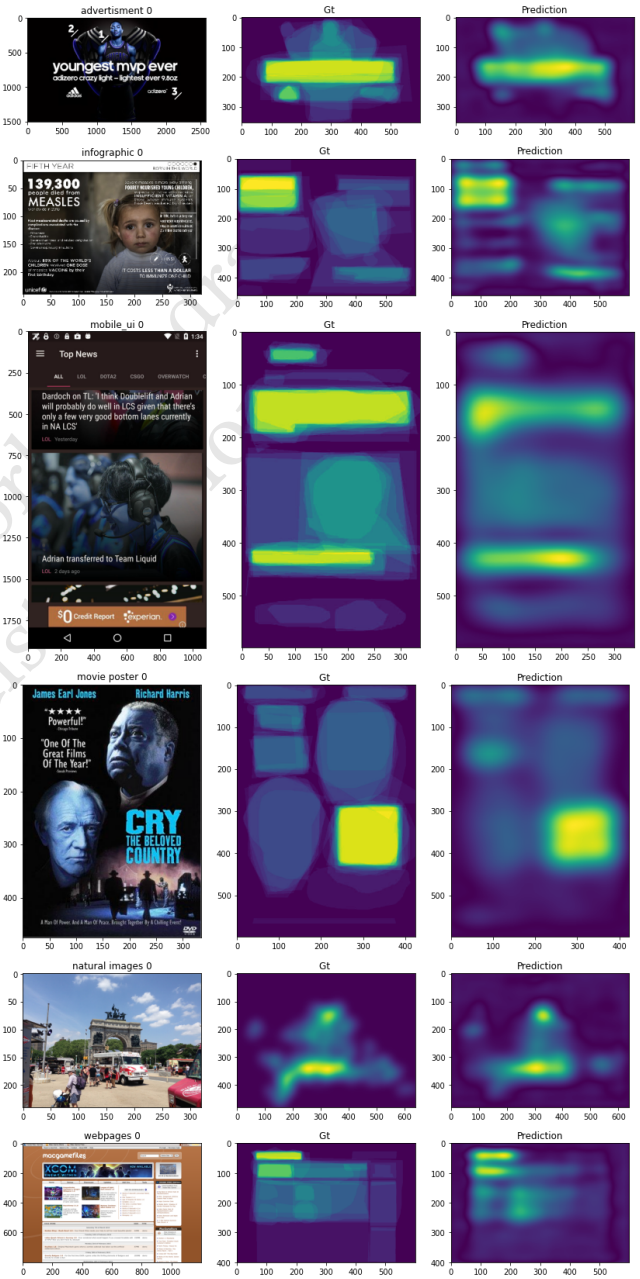

Figure 1: Visual importance and saliency predictions of our reproduced UMSI model. "gt" refers to the ground truth heatmape

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

To solve this problem, Fosco et al. proposed [9] the UMSI model. The UMSI model has ability to predict the visual saliency of natural images and different kinds of design in real-time, including infographics, posters, mobile UIs, advertisements and Webpages. The model has shown promising results in terms of efficiency and its generalization ability towards different design classes. However, the training code is not available online. To verify the details of the training process in their paper. In this work, reproduced the UMSI model for predicting visual importance. The evaluation results show that the reproduced UMSI model is comparable to the official UMSI model in different tasks. In graphic design classification, our reproduced model performs better than the official model.

The remainder of the paper is organized as follows. Section 2 presents the related works, while Section 3 introduces UMSI model and training details in reproducing. Section 4 demonstrates the evaluation results of our reproduced UMSI model in different tasks. Section 5 discusses the failure cases of reproduced model and proposes possible directions for future works. Finally, Section 6 draws the conclusion.

2 RELATED WORK

Most prior works estimate visual saliency by measuring eye tracking [12]. However, implement eye tracking for each design to get accurate prediction is time consuming and not helpful in real-time interactive design.

With the help of deep learning, many computer vision researchers have studied visual saliency [7, 10]. These methods have produced high performance on existing saliency datasets with natural images[4, 1, 11, 5, 2]. However, these models which are trained on purely nature images cannot directly be used on graphic designs. Bylinskii et al. [3] first introduced predictors to both graphic designs and data visualizations based on neural networks. Fosco et al. [9] followed up with the first visual importance prediction model which performed well across different types of designs and natural image and proposed manually labeled dataset Imp1K including 5 classes of designs. The architecture of their model modification of a strong semantic segmentation model DeepLabv3+ [6]. The model first trained on SALCON dataset which consists of natural images, then fine-tuned on Imp1K dataset. In their experimental result, the UMSI model trained by this training procedure demonstrates strong generalization ability. However, the code of training procedure is unavailable. To reproduce this training procedure, in this work we follow the details described in their work reproducing the UMSI model.

3 UNIFIED MODEL OF SALIENCY AND IMPORTANCE

3.1 Model Architecture

Unified Model of Saliency and Importance (UMSI) [9] is the state-of-the-art model in predicting visual importance across graphic design types. The architecture of UMSI model is presented in Figure 2.. Given an image as input, the outputs of UMSI are the heatmap representing the importance value of every pixel and the predicted label of the type of the input. Similar to other visual saliency predictors,

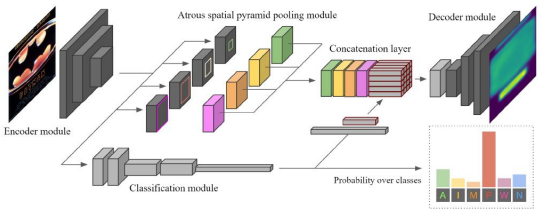

Figure 2: The model architecture of UMSI [9].

UMSI model is designed based on encoder-decoder architecture. In addition, UMSI model has a classification submodule.

The Xception-based encoder is comprised of depthwise separable convolutions in order to extract feature maps of input images. The following Atrous Separable Pyramid Pooling (ASPP) layer aggregates the multi-scale information of the image. The output of the encoder is also used for classification. The classification module is parallel with ASPP. In the classification module, the input feature map is processed by convolutional layers and then flatten to a 1d tensor. This tensor produces the class probabilities after Softmax. In addition, this tensor is resized and then concatenated with the output of ASPP. By concatenating with the classification information, the model can learn the general heatmap trend of different design classes. This concatenated tensor is passed to the decoder. The decoder is a set of convolutions for up-sampling the input tensor to the original image size.

3.2 Loss Function

Several evaluation metrics have been proposed for visual saliency and visual importance tasks. In order to get a better evaluation, in saliency and importance models, the loss functions are usually a combination between different metrics. In reproducing, we use the same loss function as in [9]. Kullback-Leibler divergence (KL) and Pearson's Correlation Coefficient (CC) losses are used in visual importance prediction. In addition, binary cross-entropy loss is used for classification module. The loss function is defined as follows:

$$\begin{aligned} loss(p\_map, gt\_map, p\_label, gt\_label) = \\ \alpha KL(p\_map, gt\_map) \\ +\beta CC(p\_map, gt\_map) \\ +\gamma BinaryCrossEntropy(p\_label, gt\_label)) \end{aligned} \tag{1}$$

where p_map, gt_map, p_label and gt_label refer to the prediction importance heatmap, the ground truth importance heatmap, the prediction label and the ground truth label respectively. α, β and γ are coefficients which balance the three loss functions.

The KL divergence loss measures the distance between the distribution of the prediction map and the distribution of ground truth map. The lower KL score refers to the better model performance.

The CC loss measures the correlation between the prediction map and the ground truth map. The CC score is close to 1 for the good prediction map.

3.3 Datasets

For training and testing our model, the same as [9], we use three popular datasets in training. Visual importance datasets Imp1k [9] and GDI [14], and visual saliency dataset SALICON [10].

- SALICON: SALICON is the largest visual saliceny dataset which consists of 10000 training images, 5000 validation images, and 5000 test images. These images are all natural images selected from COCO dataset [13]. The training set and validation set have corresponding ground truth saliency heatmap and fixation information. The ground truth for the test set is not available. The predictions are evaluated on SALICON website.
- Imp1k: Imp1k is a visual importance dataset. The dataset contains 1000 images of 5 different types of graphic designs (advertisements, infographics, movie posters, mobile UIs and webpages). There are 800 training images and 200 validation images.
- GDI: Similar to Imp1k, GDI is a visual importance dataset for graphic designs as well. Different from Imp1k, all 1078 images in GDI are mostly advertisements and posters.

3.4 Model Training

In reproducing, we follow the training procedure and training details in the original paper [9]. To make the model have a good performance on both natural images and graphic designs, we use SALICON with no fixation information and Imp1k in training.

3.4.1 Training Details. In the beginning, the model is trained on SALICON training dataset for 10-15 epochs. After this, the model has a good ability on predicting the visual saliency of natural images. The model weights after each epoch are saved. The model weights with the lowest validation loss is selected for fine-tuning on Imp1k. The training loss is shown in Figure 3.

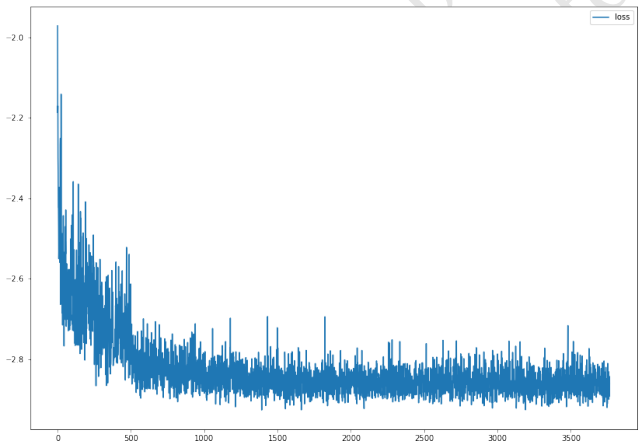

Figure 3: The training loss on SALICON dataset

In fine-tuning, the model is trained on Imp1k for another 10-15 epochs. In epoch epochs, to keep the model remember the knowledge of natural images, randomly selected 160 images from SALICON training set are added in Imp1k training set. This number of natural images can maintain the class balance. Figure 4 shows the loss of fine-tuning. In the whole training procedure, the weight of

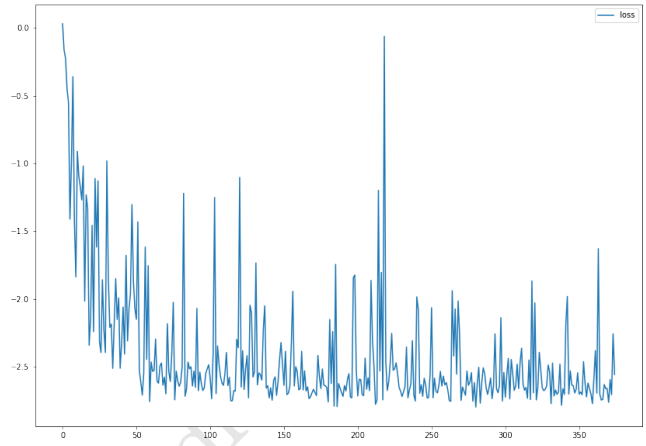

Figure 4: The training loss in fine-tuning on Imp1k dataset

KL loss $\alpha = 10$, and the weight of CC loss $\beta = -3$. For the classification loss, the weight is 0 when the model is training on SALICON dataset. Since sending too many natural images to the classification module can cause class imbalance. In fine-tuning, the weight $\gamma = 5$ for binary cross-entropy loss. We train the model on RTX 2060 Super GPU, which has limited RAM. Thus, the batch size is 4 in our reproducing, different from batch size 8 in the official UMSI model. In addition, we only use two drop out layers with rate 0.3 in our model. The rest of training parameters are the same as [9]. We use Adam optimizer with an initial learning rate of 0.0001. The learning rate decays 10 times every three epochs.

4 MODEL EVALUATION

In this section, we compare our reproduced UMSI model with the official UMSI model in different tasks on different dataset.

4.1 Evaluation Metrics

4.1.1 Importance and saliency metrics. Except from KL and CC, we use coefficient of determination (R^2) and Root-Mean-Square Error (*RMSE*) evaluating the performance of importance and saliency as well.

- R^2: R^2 evaluates the goodness-of-fit of the prediction versus the ground truth. The higher R^2 value indicates the better performance.
- *RMSE*: RMSE measures the deviation between the prediction and the ground truth. RMSE is always non-negative, and a value of 0 indicates a perfect fit to the data. The lower RMSE score refers to the better performance.

4.2 Saliency Evaluation

We first compare our reproduced UMSI model with the official UMSI model in visual saliency prediction. Due to our computer has limited computational power, predicting heatmaps of all 5000 images in SALICON test set and upload for evaluation is impossible. Instead of using the test set, the reproduced model is evaluated on SALICON

validation set. Table 1. shows the result of comparison. Although

Methods	$R^2 \uparrow$	RMSE ↓	CC ↑	KL ↓	ACC ↑
UMSI[9]	**0.635**	**0.096**	**0.880**	**0.196**	**0.999**
UMSI-reproduced(ours)	0.593	0.098	0.851	0.358	0.996

Table 1: Evaluation of visual saliency prediction on SALICON dataset. We compare our reproduced UMSI model with the official UMSI model on validation set of SALICON.

the official UMSI model outperforms our reproduced model on saliency prediction. Our reproduced model gets the comparably result to the official UMSI model.

4.3 Visual Importance Evaluation

In addition to saliency prediction evaluation, we evaluation visual importance prediction on both GDI and Imp1k dataset.

For GDI dataset, we use the entire dataset for the evaluation. Table 2. presents the result on GDI dataset. The performance of our reproduced UMSI model is close to the official UMSI model.

Methods	$R^2 \uparrow$	RMSE ↓	CC ↑	KL ↓
UMSI[9]	**0.447**	**0.192**	**0.818**	**0.186**
UMSI-reproduced(ours)	0.408	0.204	0.795	0.220

Table 2: Evaluation of visual importance prediction on GDI dataset. We compare our reproduced UMSI model with the official UMSI model on the entire GDI dataset. While the official UMSI model outperforms our reproduced UMSI model on GDI dataset, our reproduced model also gets competitive results.

For Imp1k dataset, we use the validation set for the evaluation. Table 3. shows the evaluation result. Our reproduced UMSI model rivals the official UMSI model.

Methods	$R^2 \uparrow$	RMSE ↓	CC ↑	KL ↓	ACC ↑
UMSI[9]	0.080	0.141	**0.839**	**0.149**	0.935
UMSI-reproduced(ours)	**0.117**	**0.139**	0.812	0.192	**0.950**

Table 3: Evaluation of visual importance prediction on Imp1k dataset. We compare our reproduced UMSI model with the official UMSI model on validation set of Imp1k. The result shows our reproduced model rivals the official UMSI model.

4.4 Classification Evaluation

For classification evaluation, we use the validation set of both Imp1k and SALICON. Imp1k provides 5 types of graphic designs and SALICON provides natural images.

Table 4. shows the classification accuray per class. The result of average accuracy of graphic designs and natural images can be found in Table 1. and Table 3. The official model performs well in the class of natural images. In graphic design classification, our model gets 95% average accuracy which is slightly better than the official model 93%.

5 DISCUSSION

In this section, we analyze the possible reasons behind the slightly worse performance of our reproduced UMSI model. In additionm we discuss the common failure cases predicted by our reproduced UMSI model and the official UMSI model.

5.1 The Possible Reasons Behind the Slightly Worse Performance

In general, the performance of our reproduced UMSI model is similar to the official UMSI model. However, our reproduced model performs slightly worse. There are some possible reasons for this.

(1) Batch size: The max batch size can be used in our computer is 4. The model may be stuck in the local optimum when using a small batch size in training.
(2) Over-fitting: Since the detail of dropout setting in the official UMSI model is unclear. In reproducing, we only use two dropout layers. Trying different setting to avoid over-fitting may improve the performance of our model.

5.2 Failure Cases

Figure 5. shows the examples which only fail in visual importance prediction. Figure 6. presents the failure case in classification only. While, Figure 7. demonstrates the examples which fail in both visual importance prediction and classification. The visual importance

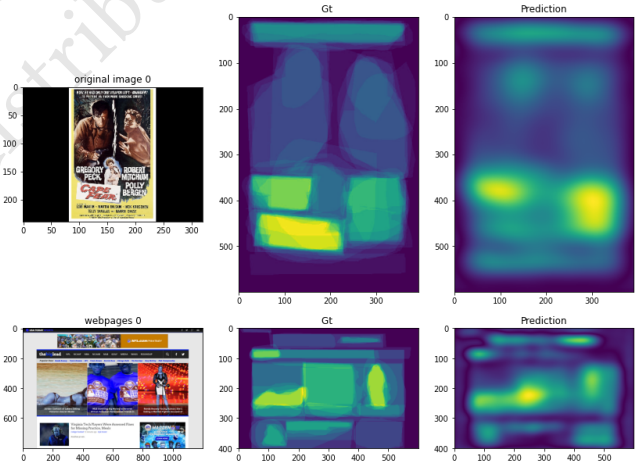

Figure 5: Example failure cases in visual importance prediction. "gt" refers to the ground truth heatmap.

prediction is easy to fail, when the design has many design elements or there are many overlays. In movie poster, infographics, as well as other types of designs, the title always have relative higher visual importance value and the title usually has the biggest text size. For example, the top images in Figure 5. has three zones with text and the text in these zones are in almost the same size. Our model confuses the text boxes with the real title in this case.

Visual saliency and visual importance are incompatible in some parts. Visual saliency datasets are collected by eye-tracking, the large elements or attractive elements in images have the highest visual saliency score. However, in visual importance datasets, the

Methods	Advertisements	Infograhics	Mobile UIs	Movie Posters	Webpages	Natural Images
UMSI[9]	0.96	0.99	0.995	0.995	0.995	**0.999**
UMSI-reproduced (ours)	**0.97**	**0.995**	0.995	0.995	0.995	0.996

Table 4: Evaluation result of classification on Imp1k and SALICON dataset. We compare our reproduced UMSI model with the official UMSI model on the validation set of Imp1k and the validation set of SALICON.

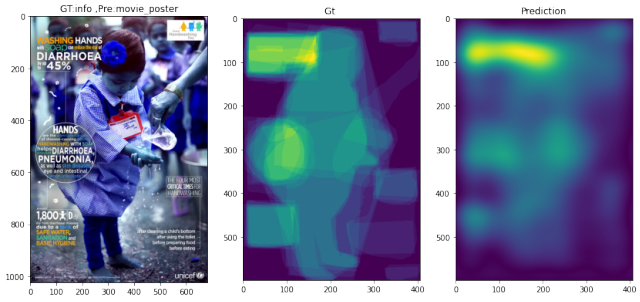

Figure 6: Example failure case in classification. "gt" refers to the ground truth heatmap.

highest importance value usually assigns to the title or text. UMSI model has knowledge in both saliency and importance. Since SALICON dataset has more images than imp1k, the model learned more knowledge of saliency. For instance, in the first and the third examples in Figure 7., our model gives the highest visual importance value to those design elements which supposed to have a relative higher saliency value, instead of the title. Balancing the saliency knowledge and the importance knowledge in training is a worthwhile direction in the future work.

6 CONCLUSION

In this work, reproduced the state-of-the art visual importance model - UMSI. We compare the reproduced UMSI model with the official UMSI model, the results show that our reproduced model rivals the official model. Moreover, we analysis the common failure cases in UMSI model and propose suggestions to the future works.

REFERENCES

[1] Ali Borji and Laurent Itti. "CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research". In: *CVPR 2015 workshop on "Future of Datasets"* (2015). arXiv preprint arXiv:1505.03581.

[2] Ali Borji, Dicky N Sihite, and Laurent Itti. "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study". In: *Image Processing, IEEE Transactions on* 22.1 (2013), pp. 55–69.

[3] Zoya Bylinskii et al. "Learning visual importance for graphic designs and data visualizations". In: *Proceedings of the 30th Annual ACM symposium on user interface software and technology*. 2017, pp. 57–69.

[4] Zoya Bylinskii et al. *MIT Saliency Benchmark*. http://saliency.mit.edu/.

[5] Zoya Bylinskii et al. "What do different evaluation metrics tell us about saliency models?" In: *arXiv preprint arXiv:1604.03605* (2016).

[6] Liang-Chieh Chen et al. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: *ECCV*. 2018.

[7] Marcella Cornia et al. "A Deep Multi-Level Network for Saliency Prediction". In: *International Conference on Pattern Recognition (ICPR)*. 2016.

[8] Marcella Cornia et al. "Predicting human eye fixations via an lstm-based saliency attentive model". In: *IEEE Transactions on Image Processing* 27.10 (2018), pp. 5142–5154.

[9] Camilo Fosco et al. "Predicting Visual Importance Across Graphic Design Types". In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 2020, pp. 249–260.

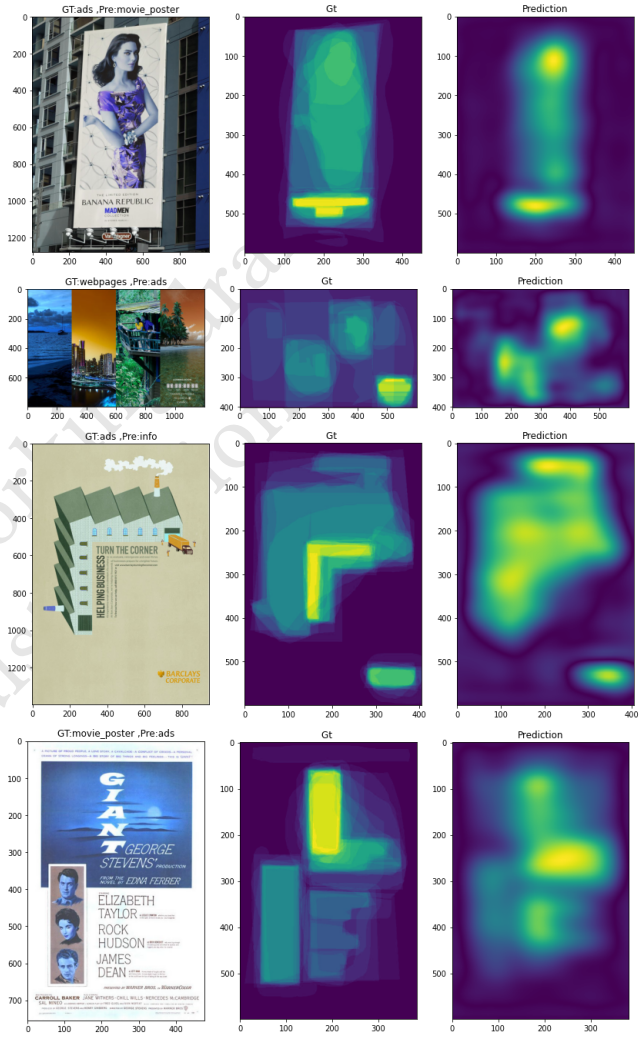

Figure 7: Examples fail in both visual prediction and classification. "gt" refers to the ground truth heatmap.

[10] Xun Huang et al. "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 262–270.

[11] Tilke Judd, Frédo Durand, and Antonio Torralba. "A Benchmark of Computational Models of Saliency to Predict Human Fixations". In: *MIT Technical Report*. 2012.

[12] Luis A Leiva et al. "Understanding Visual Saliency in Mobile User Interfaces". In: *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. 2020, pp. 1–12.

[13] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[14] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. "Learning layouts for single-pagegraphic designs". In: *IEEE transactions on visualization and computer graphics* 20.8 (2014), pp. 1200–1213.