Visualization Assessment: A Machine Learning Approach

Xin Fu[1,2], Yun Wang[2], Haoyu Dong[2], Weiwei Cui[2], Haidong Zhang[2]
[1] Wuhan University *† [2] Microsoft Research ‡

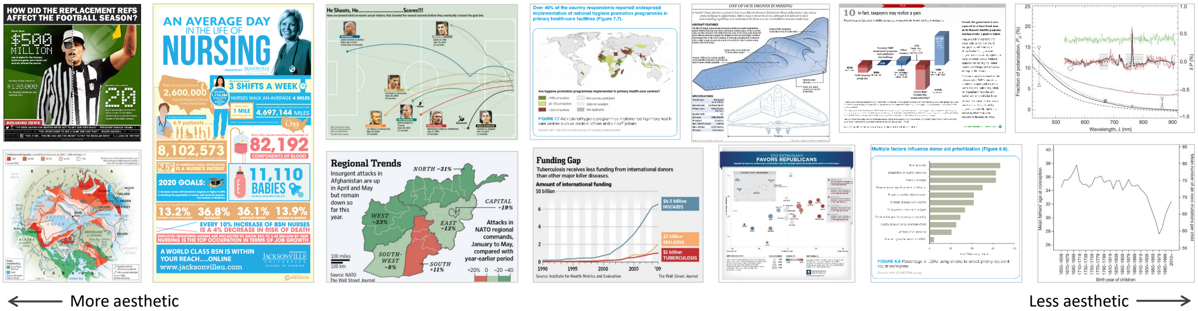


Figure 1: Sample visualization images from *Vis410* ranked by our machine learning model regarding aesthetics (decreasing from left to right) show effectiveness of our method.

Abstract

Researchers assess visualizations from multiple aspects, such as aesthetics, memorability, engagement, and efficiency. However, these assessments are mostly carried out through user studies. There is a lack of automatic visualization assessment approaches, which hinders further applications like visualization recommendation, indexing, and generation. In this paper, we propose automating the visualization assessment process with modern machine learning approaches. We utilize a semi-supervised learning method, which first employs Variational Autoencoder (VAE) to learn effective features from visualizations, subsequently training machine learning models for different assessment tasks. Then, we can automatically assess new visualization images by predicting their scores or rankings with the trained model. To evaluate our method, we run two different assessment tasks, namely, aesthetics and memorability, on different visualization datasets. Experiments show that our method can learn effective visual features and achieves good performance on these assessment tasks.

Keywords: Visualization, automated design, visualization assessment, presentation

1 Introduction

Expressive visualizations served for presentation purposes have become prevalent in news articles, presentation slides, scientific publications, and so on [8,24,25,30]. As a result, researchers in InfoVis community have started to pay attention to a wide variety of assessment perspectives, including memorability [4,5], aesthetics [14], engagement [16], and enjoyment [29]. These have become important criteria for visualization assessment. Recent work has summarized underlying factors and gained insights on good visualizations from different aspects through in-lab user studies or crowd-sourcing experiments [4,5,14,16]. For example, Borkin et al. [5] conduct a study using Amazon's Mechanical Turk to analyze contributing attributes to memorability scores of visualizations; Harrison et al. [14] study how quickly aesthetic impressions are formed.

While the findings from empirical studies are inspiring and thought-provoking for designers to improve visualization design, it is still inadequate, if not impossible, to derive standard and practicable measurements for assessing expressive visualizations. Specific assessment tasks, such as estimating the memorability of given visualizations, or comparing the aesthetic scores of two visualizations, are still beyond attainment. This hinders further visualization applications, such as recommendation, indexing, generation, and mixed-initiative authoring of expressive visualizations.

In this paper, we explore the automatic assessment of expressive visualizations with machine learning algorithms. We combine an unsupervised module to extract effective features representing the visualization images, and a supervised module to train task-oriented learning models for various assessment tasks based on the features extracted from prior module.

More specifically, since there is no large-scale visualization dataset with human labels for assessment tasks, we first leverage Variational Autoencoder (VAE) [23] to make use of the unlabeled data to extract effective feature embeddings, which avoids the need for huge datasets and expert-engineered features. With the obtained features, we can train task-specific machine learning models for visualization assessment tasks. We can therefore benefit from previous user experiment datasets and develop automatic models to tackle various assessment tasks.

Our paper makes the first attempt to automate visualization assessment. To understand the effectiveness of our approach, we evaluate our method on two representative assessment tasks, namely, aesthetics and memorability. We leverage three public datasets to train the models. Experiments on the datasets demonstrate the effectiveness of our method. We believe our research can inspire further study on automatic assessment of expressive visualizations. Many applications such as visualization recommendation, generation, and indexing can be developed based on visualization assessment techniques. Our main contributions are as follows:

- We propose exploiting computational models to automatically assess expressive visualization images.
- We implement a VAE that can learn representations for visualizations without human supervision. The extracted features from VAE can be adopted in different assessment tasks with targeted machine learning models.
- We experiment on two visualization datasets for two different assessment tasks, namely, memorability and aesthetics. The results demonstrate the effectiveness of our approach.

[*] Work done during X. Fu's internship at Microsoft Research.
[†] fuxin@whu.edu.cn
[‡] {wangyun, hadong, weiweicu, haizhang}@microsoft.com

2 Related Work

2.1 Assessment of Visualization Design and Encoding

Visualization design has gone beyond standard visual charts in the InfoVis community [21, 26, 33, 34]. Consequently, there is a need to understand the quality of visualizations from multiple aspects, such as effectiveness [2], confidence [32], enjoyment [29], memorability [10, 29], readability [2], aesthetics [2, 3], and learnability [10], through empirical experiments. For example, Borkin et al. [4, 5] pioneered the study of visualization memorability, showing that visualizations are intrinsically memorable with consistency across people. Harrison et al. [14] conduct online experiments to investigate people's judgments of the aesthetic appeal of infographics. Hung et al. [16] design a questionnaire to assess user engagement of information visualizations from 11 different characteristics. Bateman et al. [1] measure accuracy and long-term recall of embellished charts and plain ones, finding that the accuracy for embellished charts are no worse but the recall is significantly better. Haroz et al. [13] test the memorability, speed of finding information, and engagement performance for pictographs. These research works have shown the growing interest and importance of visualization assessment. But there is still a lack of approaches to automatically assess visualization designs and it hinders the application of these research findings.

Another thread of research seeks machine learning approaches to recommend visualization automatically. To generate suggestions for data exploration, the systems use machine learning algorithms, such as learning-to-rank and decision tree, to incorporate visualization design knowledge or fit to empirical user data and promote effective visual encoding [19, 27, 28]. However, these systems often rely on handcrafted features, which is a labor-intensive and biased process. Moreover, they only support a limited number of standard chart types and encodings, which have well-defined and rather small design space. We are still short of a method to assess visualization images in real world, which are much more freeform, complex, and flexible. In our work, we leverage machine learning models to learn from a large set of visualization images and extract features to support different assessment tasks when given new visualization images.

2.2 Assessment of Generic Images

Generic images refer to natural photos taken in the real world, which are usually available in large quantities. In the last decade, a number of researchers in the field of computer vision and multimedia have tried to assess such images from various perspectives, such as aesthetics [9, 31], memorability [7, 17, 20], and interestingness [12].

In such generic image assessment tasks, handcrafted image descriptors (e.g. HOG, SIFT) and visual factors (e.g. color, saliency) are traditionally used to train machine learning models to predict their labels. With the development of deep learning algorithms, CNNs are introduced into image assessment tasks. More accurate estimation results have been achieved [20, 31].

However, visualization images show different characteristics. They are designed and made up of visual elements such as shapes, icons, text descriptions, annotations, etc., while generic images typically contain irregular graphics, such as natural landscapes, objects, people, etc. Such discrepancies make their spatial and visual properties different. Therefore, we need to train vision models on visualization images to capture distinct features from them. However, the success achieved in computer vision tasks are largely attributed to the availability of massive well-labeled datasets. For visualizations assessment, there is no such huge labeled dataset for us to train an advanced deep CNN. As a first step, we make use of previous user studies by utilizing a semi-supervised method.

3 Method

The overall pipeline of our machine learning-based assessment is shown in Fig. 2. We incorporate both unsupervised and supervised stages to make use of unlabeled and labeled visualizations.

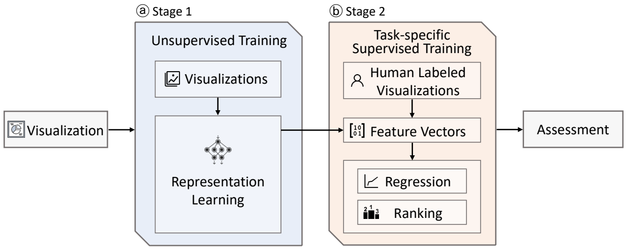


Figure 2: The pipeline of our semi-supervised assessment approach for visualization assessment.

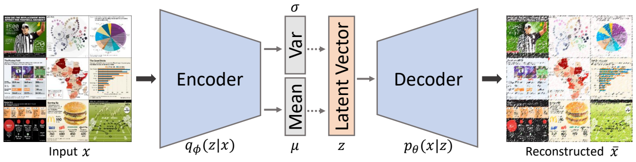


Figure 3: Overview of the Variational Autoencoder model.

The first stage (Fig. 2a) is intended for unsupervised representation learning. Specifically, we fed a large amount of unlabeled visualizations into a VAE, which can automatically learn representations from these images. Subsequently, the trained VAE can encode any new input visualizations into low-dimensional feature vectors.

The second stage (Fig. 2b) utilizes the attained feature vectors from VAE to train supervised machine learning models for different assessment tasks. Hence, our model can extract features and assess any visualization images automatically.

3.1 Learning Representations using VAE

An overview of the VAE model is shown in Fig. 3. Typically, a VAE first transforms the input image x to a latent vector $z = Encoder(x) \sim q_\phi(z|x)$ with an encoder network, and then a decoder network is used to decode z back to an image $\bar{x}$ that will be as similar as the original one: $\bar{x} = Decoder(z) \sim p_\theta(x|z)$.

Specifically, the encoder model $q_\phi(z|x)$ with variational parameter ϕ is parametrized as a multivariate normal distribution following rule $q_\phi(z|x) = \mathcal{N}(z; \mu, \mathrm{diag}(\sigma))$, where μ and σ denotes vectors of means and variance respectively. Both μ and σ are learned through the encoder network: $(\mu, \log\sigma) = \mathrm{EncoderNeuralNet}_\phi(x)$. Then, the latent vector can be constructed by $z = \mu + \sigma \odot \mathcal{N}(0, I)$, so that z has the characteristic of being independent unit Gaussian random variables, i.e., $z \sim \mathcal{N}(0, I)$. Hence, instead of deterministic mapping image x, the probabilistic encoder $q_\phi(z|x)$ will produce a robust vector conditioned on the input, and thus we use the vector z as the low-dimensional representation of x.

We design our model based on the general VAE architecture. The details of the model are described in the supplementary materials.

The training procedure of our model follows the approach of the Variational Autoencoder [23], where the loss function is the sum of two terms: the reconstruction loss $\mathcal{L}_R$, and the Kullback-Leibler (KL) divergence loss $\mathcal{L}_{KL}$.

$$\begin{aligned}\mathcal{L}_{VAE} &= \mathcal{L}_R + \mathcal{L}_{KL} \\ &= -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})]\end{aligned} \tag{1}$$

Therefore, after training on massive visualization images, the VAE can effectively encode new input images into low-dimensional vectors z. As described in previous research [11, 22], such feature vectors are high-level representations of the original image, and thus can be used in machine learning models to tackle different assessment tasks.

3.2 Task-Specific Visualization Assessment

The aforementioned trained VAE can encode each visualization image into an effective low-dimensional feature vector. With feature vectors obtained from labeled visualization images, we further train task-specific assessment model based on machine learning techniques.

Because previous user studies provide a number of labeled visualizations regarding aesthetics and memorability, we select these two tasks. Concerning aesthetics, we usually grade or rank each given visualization image based on our impression [14]. For memorability tasks, scores are typically calculated based on the experimental results [4, 5]. Generally, these assessment tasks can be categorized and formulated as regression and ranking problems, and thus corresponding machine learning methods can be applied.

The regression problem requires mapping from input data to target numeric labels. In our assessment task, we need a model to estimate the score (memorability or aesthetics) of each visualization image. Following previous work [17, 20], we choose Support Vector Regressor (SVR), a robust regression model, to learn a non-linear function to map features obtained from VAE to assessment score.

When objective scores are hard to get, we rank the data by relatively comparing them. We address such a task by training a learning-to-rank model to predict the order of input visualizations. Specifically, we adopt the LambdaRank algorithm provided by LightGBM [18] to learn the order labeled by users, and predict the ordinal score of each visualization image.

Our proposed model can attain high-level and task-independent features from visualization image via VAE, and these features can be further exploited depending on the type of the task. Thus, apart from the two tasks we addressed here, our method has the potential to be extended to other assessment tasks in the future.

4 Experiments

To understand the effectiveness of our proposed method, we evaluate on two assessment datasets. We describe the evaluation metrics and our results with comparisons against other conventional approaches.

4.1 Datasets

We adopt three different visualization datasets for training, *Vis6K*, *Aes330*, and *Vis410*, from multiple sources. We explain how we use these datasets in Sect. 4.3.

Vis6K: Bylinskii et al. [6] scraped a dataset containing more than 60K visualization images. We filtered images which have abnormal aspect ratio, and randomly selected 6,000 images, which is almost 20 times the size of the following labeled datasets. It allows our VAE model to learn effective representations for visualizations.

Aes330: The dataset for aesthetics assessment is from a previous work [14], which has 330 visualizations evaluated by 1,278 on online participants with two-stage ratings (1 to 9). The average ratings are calculated as the aesthetic assessment score.

Vis410: Borkin et al. [5] assess 410 visualizations through a user study. In our experiments, we use the memorability scores from this dataset for memorability assessment. Additionally, we ask users to label the aesthetic rankings, as described later in Sect. 4.3.

4.2 Metrics

To evaluate the performance of our method, we adopt two metrics used in previous assessment works [4, 17, 20].

Spearman's rank correlation coefficient (SRCC) measures consistency between the predicted and ground truth rankings, within the range $[-1, +1]$ where 1 represents prefect agreement and -1 indicates the maximum disagreement. It means that higher ρ values indicate better prediction result, where N is the total number of testing samples, $\hat{r}_i$ is a rank of the i^{th} ground truth score, and r_i the rank of the i^{th} prediction. This rank correlation is defined as follows:

$$\rho(\hat{r}, r) = 1 - \frac{6\sum_i^N (\hat{r}_i - r_i)^2}{N(N^2-1)} \qquad (2)$$

Mean squared error (MSE) is used as a secondary means for regression models. *SRCC* shows monotonic relationships between the reference and observations but does not reflect the absolute numerical errors between them. In the *MSE* equation: $\mathrm{MSE}(\hat{y}, y) = \frac{1}{N}\sum_{i=1}^N (\hat{y}_i - y_i)^2$, y_i is the prediction and $\hat{y}_i$ the ground truth evaluation score, and N the number of tested samples.

In the following experiments, regression and ranking models are trained on different datasets according to the type of assessment label. We apply *SRCC* and *MSE* on regression results (i.e. predicted assessment scores), to evaluate both relative relationships and absolute errors between the estimation and ground truth. For ranking models, where relative orders are predicted, *SRCC* is suitable for computing consistency.

4.3 Performance Evaluation

We evaluate the performance of our method on memorability and aesthetics with *Vis410* and *Aes330*, respectively. Because the labels of *Vis410* and *Aes330* are scores and ratings, we naturally treat them as a regression problem and evaluate them with *SRCC* and *MSE*. Furthermore, we collect ranking labels for *Vis410* regarding aesthetics. The benefits are three-fold: we can (1) evaluate our ranking model, (2) verify our method on the same dataset for different tasks, and (3) understand the relationship between visualization memorability and aesthetics. We collected the rankings from 4 users with different backgrounds (aged between 22 to 28, one female and three males) to independently rank the 410 visualizations based on their own judgment of aesthetics. Among them, two have art or design background and two have no design experience. The average ranking is calculated for each image.

Our VAE model is first trained on the unlabeled *Vis6K* dataset and converges after 500 epochs. Then, the model is used to extract latent feature vectors on both *Vis410* and *Aes330* datasets. We randomly split the datasets 5 times, where 80% and 20% are used for training and testing respectively. We average the predicted results from regression and ranking models. For comparison, we choose the following methods in our experiments:

- **HOG**: Histogram of oriented gradients (HOG) has been used as a handcrafted feature in many generic image assessment tasks [17]. We extract 512-dimensional feature vector from the input visualizations, and then machine learning (ML) models (SVR, LambdaRank) are trained to make prediction.
- **PCA**: We use principal component analysis (PCA) to transform the input into a low-dimensional vector. The output dimension of PCA is set to 512. We first train a PCA model on the unlabeled *Vis6K* dataset, and then apply it to the labeled datasets for ML models.
- **ResNet**: We use ResNet-18 [15] pre-trained on ImageNet to extract 512-dimensional feature vectors from its penultimate layer. ML models are then applied.
- **CNN**: We build a simple end-to-end CNN which directly predicts the score given an input visualization image. Thus, it can not be applied to rank the aesthetics of *Vis410* dataset.
- **VAE**: The VAE is first trained on *Vis6K* and then used to map new inputs into 512-dimensional feature vectors on testing datasets, followed by ML models.

Results Analysis As shown in Table 1, our method has achieved better results (higher rank correlation and lower absolute error) on both aesthetics and memorability assessment tasks, even though the VAE is not directly trained on these two datasets. ResNet is trained on millions of generic images, but the learned prior knowledge doesn't not directly contribute to the assessment of visualizations. For the supervised regression task, the end-to-end CNN trained on these two assessment datasets performs poorly, due to insufficiently labeled training data. The correlations between the predicted and ground truth results are visualized in Fig. 4a and Fig. 4b. We also find results on *Aes330* are lower by about 20% compared to *Vis410*. After further investigation, we notice that all the visualizations in this dataset are manually chosen with quite

Table 1: Results on *Aes330* and *Vis410* datasets for two assessment tasks (aesthetics and memorability) using both regression and ranking models. The higher *SRCC* (Spearman's rank correlation) and lower *MSE* (Mean squared error) indicate better performance. VAE performs better compared to a few other methods.

Task	**Aesthetics**			**Memorability**	
Dataset	**Aes330**		**Vis410**	**Vis410**	
Metric	**SRCC↑**	**MSE↓**	**SRCC↑**	**SRCC↑**	**MSE↓**
HOG	0.281	0.685	0.425	0.474	0.524
PCA	0.221	0.741	0.537	0.591	0.452
ResNet	0.272	0.667	0.124	0.352	0.582
CNN	0.122	1.004	N/A	0.394	0.682
VAE	**0.365**	**0.636**	**0.645**	**0.651**	**0.401**

similar quality, so the aesthetic ratings do not vary too much (75.1% falls between 4 and 6 out of 1 to 9).

Memorability vs. Aesthetics We further explore the relationship between memorability and aesthetics on *Vis410*. The rank correlation between memorability and aesthetics is displayed in Fig. 4c. We observe $\rho = 0.461$, which means a high aesthetic ranking doesn't guarantee a memorable visualization, and vice versa.

Human Consistency Are aesthetic scores consistent between different users? Are machine predicted scores comparable to human users? To answer these questions, we separate the users who ranked *Vis410* into two groups and compare their results of the ranking for multiple times. We observe an average rank correlation of 0.694 between the two user groups, as shown in Fig. 4d. It is close to the result in previous studies for memorability [4] , where $\rho = 0.62$ is obtained. Our method has achieved $\rho = 0.645$ for aesthetics on the same dataset, close to the human consistency 0.694 here. Thus, the relatively high correlation in our study further demonstrates the stable consistency of aesthetics rank of visualizations.

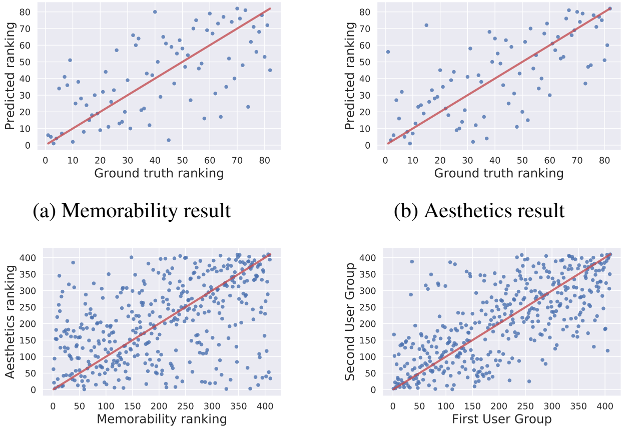
Feature Visualization Our VAE model trained on *Vis6K* can encode any visualization into a 512 dimensional feature vector. We further use t-SNE to cast the high dimensional features onto a 2D plane, and visualize them in Fig. 5. The visualization samples with lower memorability score gather at the top right corner while those with higher scores tend to appear at the bottom left. This pattern intuitively shows us that our VAE model encodes visualizations effectively and can facilitate subsequent assessment task.

Finally, sample images from the testing set of *Vis410* are displayed in Fig. 1, sorted by aesthetics order estimated by our method. The aesthetics rankings are generally consistent with our own judgments, which shows our method can effectively assess visualizations.

5 Discussion

Our study has demonstrated the capability of machine learning algorithms to understand and evaluate visualization images in the wild. We envision that many creative applications will be enabled by automatic visualization assessment. For example, it can further support mixed-initiative authoring systems to assist designers by giving feedback and optimizing their visualization designs; it can be used as a discriminator to facilitate automatic visualization generation by generative models such as GAN; it can be extended as a tool for designers and researchers to understand the design space of visualizations in the wild; it can be embedded into visualization searching and recommendation systems to select and present best results to from different assessment perspectives.

Our method has achieved a rank correlation that reaches near human consistency on memorability and aesthetics. However, it still remains unclear how extensible it is to other assessment tasks. For instance, interactive visualizations require users to have in-depth analysis; personal difference may affect *learnability* of visualizations to a great extent. These issues might stop machine learning models from capturing effective features. We hope to gather more data and experiment on different assessment tasks in the future.



(a) Memorability result

(b) Aesthetics result

(c) Aesthetics vs memorability

(d) Human consistency

Figure 4: Rank correlations for different comparisons on *Vis410*, where x and y axes represent rankings for each sample point, and the yellow line indicates y=x: (a) predicted memorability result ($\rho = 0.637$); (b) predicted aesthetics result ($\rho = 0.652$); (c) correlation between aesthetics and memorability ($\rho = 0.461$); (d) consistency between two user groups ($\rho = 0.694$).

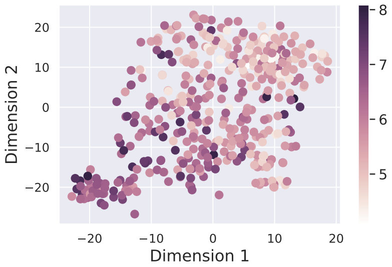


Figure 5: Visualization of feature vectors in a 2D plane using t-SNE. They are extracted by VAE on *Vis410*. The darker the color, the higher the ground truth memorability score.

Given the relatively small labeled datasets, our approach performs best among all the tested methods. However, when the labeled datasets are large enough, other supervised methods may perform better. We plan to collect a larger amount of well-labeled data to analyze the underlying facts that influence the intrinsic attributes of an visualization image. We hope researchers and designers could gain more insights into the characteristics of such images through statistical analysis.

6 Conclusion

Visualization assessment is crucial to visualization research. However, expressive visualization assessments are mostly carried out through user studies. We propose a machine learning approach for automatic visualizations assessment. We first learn low-dimensional representations from visualization images using Variational Autoencoder in an unsupervised manner. Then, we exploit the learned VAE to extract efficient features to facilitate the training of machine learning models for target assessment tasks. Our method has been evaluated through two assessment tasks, aesthetics and memorability. We envision our technique to be applied to multiple domains and foster future research.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable suggestions.

References

[1] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful Junk?: The Effects of Visual Embellishment on Comprehension and Memorability of Charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2573–2582, New York, NY, USA, 2010. ACM.

[2] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, T. Schreck, D. Weiskopf, and D. A. Keim. Quality Metrics for Information Visualization. *Computer Graphics Forum*, 37(3):625–662, 2018.

[3] R. Borgo, A. Abdul-Rahman, F. Mohamed, P. W. Grant, I. Reppa, L. Floridi, and M. Chen. An Empirical Study on Using Visual Embellishments in Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2759–2768, Dec. 2012.

[4] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):519–528, Jan. 2016.

[5] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What Makes a Visualization Memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, Dec. 2013.

[6] Z. Bylinskii, S. Alsheikh, S. Madan, A. Recasens, K. Zhong, H. Pfister, F. Durand, and A. Oliva. Understanding Infographics through Textual and Visual Tag Prediction. *arXiv:1709.09215 [cs]*, Sept. 2017.

[7] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva. Intrinsic and extrinsic effects on image memorability. *Vision Research*, 116:165–178, Nov. 2015.

[8] S. Chen, J. Li, G. Andrienko, N. Andrienko, Y. Wang, P. H. Nguyen, and C. Turkay. Supporting Story Synthesis: Bridging the Gap between Visual Analytics and Storytelling. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2019.

[9] Y. Deng, C. C. Loy, and X. Tang. Image Aesthetic Assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, July 2017.

[10] A. Figueiras. A Review of Visualization Assessment in Terms of User Performance and Experience. In *2018 22nd International Conference Information Visualisation (IV)*, pages 145–152, July 2018.

[11] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taïga, F. Visin, D. Vázquez, and A. C. Courville. PixelVAE: A Latent Variable Model for Natural Images. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[12] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool. The Interestingness of Images. In *2013 IEEE International Conference on Computer Vision*, pages 1633–1640, Dec. 2013.

[13] S. Haroz, R. Kosara, and S. L. Franconeri. ISOTYPE Visualization: Working Memory, Performance, and Engagement with Pictographs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1191–1200, New York, NY, USA, 2015. ACM.

[14] L. Harrison, K. Reinecke, and R. Chang. Infographic Aesthetics: Designing for the First Impression. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1187–1190, New York, NY, USA, 2015. ACM.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[16] Y.-H. Hung and P. Parsons. Assessing User Engagement in Information Visualization. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pages 1708–1717, New York, NY, USA, 2017. ACM.

[17] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *CVPR 2011*, pages 145–152, June 2011.

[18] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 3149–3157, USA, 2017. Curran Associates Inc.

[19] A. Key, B. Howe, D. Perry, and C. Aragon. VizDeck: Self-organizing Dashboards for Visual Analytics. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 681–684, New York, NY, USA, 2012. ACM.

[20] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva. Understanding and Predicting Image Memorability at a Large Scale. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2390–2398, Dec. 2015.

[21] N. W. Kim, E. Schweickart, Z. Liu, M. Dontcheva, W. Li, J. Popovic, and H. Pfister. Data-Driven Guides: Supporting Expressive Design for Information Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):491–500, Jan. 2017.

[22] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. Semi-supervised Learning with Deep Generative Models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3581–3589, Cambridge, MA, USA, 2014. MIT Press.

[23] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[24] R. Kosara. Presentation-Oriented Visualization Techniques. *IEEE Computer Graphics and Applications*, 36(1):80–85, Jan. 2016.

[25] R. Kosara and J. Mackinlay. Storytelling: The Next Step for Visualization. *Computer*, 46(5):44–50, May 2013.

[26] Z. Liu, J. Thompson, A. Wilson, M. Dontcheva, J. Delorey, S. Grigg, B. Kerr, and J. Stasko. Data Illustrator: Augmenting Vector Design Tools with Lazy Data Binding for Expressive Visualization Authoring. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 123:1–123:13, New York, NY, USA, 2018. ACM.

[27] Y. Luo, X. Qin, N. Tang, and G. Li. DeepEye: Towards Automatic Data Visualization. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 101–112, Apr. 2018.

[28] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):438–448, Jan. 2019.

[29] B. Saket, A. Endert, and J. Stasko. Beyond Usability and Performance: A Review of User Experience-focused Evaluations in Visualization. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, pages 133–142, New York, NY, USA, 2016. ACM.

[30] E. Segel and J. Heer. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, Nov. 2010.

[31] H. Talebi and P. Milanfar. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, Aug. 2018.

[32] E. Wall, M. Agnihotri, L. Matzen, K. Divis, M. Haass, A. Endert, and J. Stasko. A heuristic approach to value-driven evaluation of visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):491–500, Jan 2019.

[33] Y. Wang, H. Zhang, H. Huang, X. Chen, Q. Yin, Z. Hou, D. Zhang, Q. Luo, and H. Qu. InfoNice: Easy Creation of Information Graphics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 335:1–335:12, New York, NY, USA, 2018. ACM.

[34] Z. Wang, S. Wang, M. Farinella, D. Murray-Rust, N. Henry Riche, and B. Bach. Comparing Effectiveness and Engagement of Data Comics and Infographics. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 253:1–253:12, New York, NY, USA, 2019. ACM.