# An Online Presentation Slide Assessment System Using Visual and Semantic Segmentation Features

**Shengzhou Yi[1], Junichiro Matsugami[2], Hiroshi Yumoto[3], and Toshihiko Yamasaki[1]**

[1]The University of Tokyo
[2]Rubato Co., Ltd.
[3]P&I Information Engineering Co., Ltd.
yishengzhou@cvm.t.u-tokyo.ac.jp, matsugami@rubato.co, yumoto@pandi.co.jp, and yamasaki@cvm.t.u-tokyo.ac.jp

## Abstract

In this study, we present a new presentation slide assessment system that can extract the structural features from any slide file formats. Our previous work used a neural network to identify novice vs. well-designed presentation slides based on visual and structural features. However, the structural feature extraction was only applicable to PowerPoint files. To solve this problem, we extract the semantic segmentation from the slide images as a new format of structural features. The proposed multi-modal Transformer extracts the features from the original images and semantic segmentation results to assess the slide design. The prediction targets are the top-10 checkpoints pointed out by the professional consultants. Class-imbalanced learning and multi-task learning methods are also applied to improve the accuracy. The proposed model only requiring the slide images achieved an average accuracy of 81.67% that is comparative to the performance of the previous work requiring the PowerPoint files.

## Introduction

Slides are commonly used in presentations to make it easy to convey their ideas. However, obtaining good slide design skill is very difficult for novices. Several evaluation systems (Kim et al. 2014; Oyama and Yamasaki 2019) were proposed to judge the quality of slides, but the ratings made by their system lacked of detailed criterion (i.e., only a single measure of goodness of the slides). A constrained generative adversarial network (GAN) was proposed to automatically optimize the layout of websites and magazines (Kikuchi et al. 2021). However, this study only considered the layout and did not consider the contents.

Our previous work (Yi, Matsugami, and Yamasaki 2022) proposed a bi-modal neural network to distinguish novice vs. well-designed presentation slides using visual and structural features. However, the used structural features required PowerPoint files to acquire the bounding box information of the slide objects. In other words, this technique restricted the file formats; therefore, the users who prefer to upload PDFs, Google Slides, or Keynote files could not use the full function of our previous model. In order to provide compatible performance for these users, we extract the semantic segmentation from the slide images as a new format of structural features to replace the previous structural features extracted from the PowerPoint files. The proposed two-level Transformer extracts the visual and structural features from the original images and semantic segmentation results, respectively. Furthermore, class-imbalanced learning and multi-task learning methods are also applied to improve the accuracy of the proposed model.

The main contributions of our work are as follows,

- A slide dataset with assessment labels was created to support the training of machine learning models.
- An online system was built for assessing novices' slides that uses both visual and structural features.
- A new method was proposed to extract structural features for the unsupported file formats in the previous work.

## Slide Improvement Dataset

We collected the data from a training course where the novices created single-page slides and improved their slides according to the consultants' advice. The slide pairs consist of the original one created by a novice without any help and the other one modified by the same person according to the advice from consultants. The dataset contains 1,080 such slide pairs in total.

By analyzing the created slides and the advice given to the novices, the consultants summarized the top-10 important and common checkpoints in slide design, including inserting a pictogram, adding a subheading, emphasizing words, emphasizing areas, adding T1 and T2, using the grid structure, itemizing the text, adding a comment, correct flow, and mutually exclusive and collectively exhaustive (MECE).

## Semantic Segmentation of Presentation Slides

In order to extract the structural features of presentation slides without PowerPoint files, we trained a model to get the semantic segmentation of slide objects from the images. The slide segmentation dataset is called SPaSe (Haurilet, Al-Halah, and Stiefelhagen 2019) that includes 2,000 slides from Slideshare-1M (Araujo et al. 2016) and their pixel-wise annotations. We examined four segmentation models in this study: PSPNet (Zhao et al. 2017), OCRNet (Yuan, Chen, and Wang 2020), BiSeNetV2 (Yu et al. 2021), and Segmenter (Strudel et al. 2021). The categories of semantic segmentation are background, structure, image, and text. OCRNet achieved the best performance among the used models

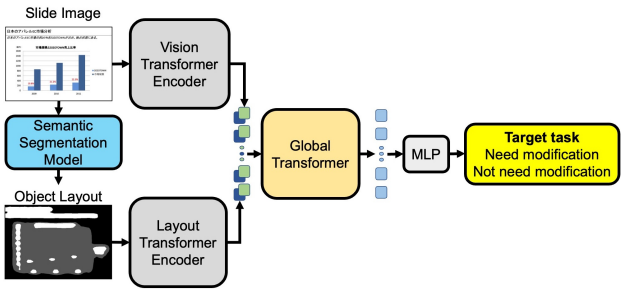


Figure 1: Two-level Transformer for slide assessment.

| Checkpoint | Sampling Points (Previous work) | Object Layout (Current work) |
|---|---|---|
| (1) Insert a pictogram | 69.06 | **83.46** |
| (2) Add a subheading | **75.82** | 73.78 |
| (3) Emphasize words | **71.74** | 66.23 |
| (4) Emphasize areas | **70.77** | 65.38 |
| (5) Add T1 and T2 | 68.75 | **83.08** |
| (6) Use grid structure | **78.04** | 70.98 |
| (7) Itemize the text | **75.36** | 71.75 |
| (8) Add a comment | **76.67** | 67.36 |
| (9) Correct flow | 53.06 | **64.58** |
| (10) MECE | 64.30 | **66.20** |
| Average | 70.36 | 71.28 |

Table 1: Validation accuracy (%) of the structural feature using the sampling points (Yi, Matsugami, and Yamasaki 2022) or object layout (semantic segmentation) without introducing class-imbalanced and multi-task learning.

with the mean Intersection over Union (mIoU) of 0.653 and the mean accuracy of 84.06%.

## Assessment Model of Presentation Slides

### Two-Level Transformer Encoder

Figure 1 presents the architecture of the proposed two-level Transformer. It contains two uni-modal Vision Transformer encoders to extract visual and structural features from the slide image and semantic segmentation, respectively. After concatenating these features, the global Transformer extracts the bi-modal features, and the multi-layer perceptron (MLP) classifies the slide samples into the positive class (with corresponding problems) or the negative class (without corresponding problems).

### Class-Imbalanced and Multi-Task Learning

We applied several class-imbalanced learning methods in our experiments, including re-weighting the class-wise sampling probabilities using the inverse class frequency (RW), the class-balanced loss (Cui et al. 2019), and the deferred re-weighting strategy (DRW) (Cao et al. 2019).

We designed an auxiliary task, which is to recognize whether a slide is before or after the novices' modification. It is strongly related to the target task, because the slides after modification have a lower possibility of having the considered designed problems than the slides before modification.

| Method | Average accuracy |
|---|---|
| Visual only | **77.72** |
| Structural only | 71.28 |
| Visual+Structural | 79.10 |
| Visual+Structural +Imbalanced learning | 81.12 |
| Visual+Structural +Multi-task learning | 79.86 |
| Visual+Structural +Imbalanced learning +Multi-task learning | **81.67** |

Table 2: Average validation accuracy (%) of the feature, class-imbalanced, and multi-task learning.

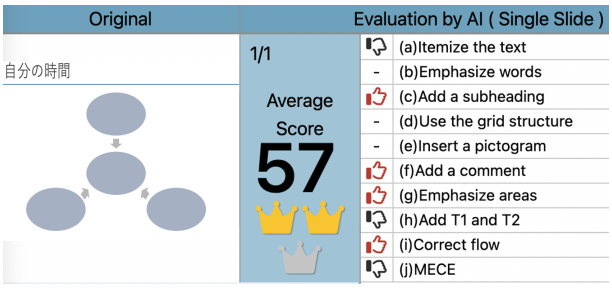


Figure 2: Analysis example of the slide assessment system.

The prediction model for the source and target tasks were trained at the same time in multi-task learning.

## Experiments

We performed binary classification for each checkpoint separately. Table 1 shows the prediction accuracy of using the structural features based on the one-hot vectors of sampling points in our previous work (Yi, Matsugami, and Yamasaki 2022) and the proposed semantic segmentation representing the object layout. It is shown that the proposed semantic segmentation slightly outperformed the sampling points on average. Furthermore, our proposed model does not restrict the file formats; therefore, the proposed structural features can be extracted any formats of the presentation slide files.

We combined the optimal settings of class-imbalanced and multi-task learning methods together using bi-modal features for each checkpoint. Table 2 shows that both of them improved the performance of the proposed model, and the final average accuracy of 81.67% is comparative to the average accuracy of 81.79% in our previous work (Yi, Matsugami, and Yamasaki 2022) requiring PowerPoint files.

## Demo System

Figure 2 shows an analysis example using our online slide assessment system. From the analysis results, we can easily find what design problems should be tackled by the user. The final score is the average evaluation score for all checkpoints. Meanwhile, the system can compare the first uploaded slide and the latest version modified by the same user.

## Acknowledgements

We would like to thank Mr. Yukiyoshi Katsumizu and Mr. TakuyaYamamoto of P&I Information Engineering Co., Ltd. for their contribution.

## References

Araujo, A.; Chaves, J.; Lakshman, H.; Angst, R.; and Girod, B. 2016. Large-scale query-by-image video retrieval using bloom filters. *arXiv preprint arXiv:1604.07939*, 1–7.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in neural information processing systems*, 1567–1578.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9268–9277.

Haurilet, M.; Al-Halah, Z.; and Stiefelhagen, R. 2019. Spase-multi-label page segmentation for presentation slides. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 726–734. IEEE.

Kikuchi, K.; Simo-Serra, E.; Otani, M.; and Yamaguchi, K. 2021. Constrained graphic layout generation via latent optimization. In *Proceedings of the 29th ACM International Conference on Multimedia*, 88–96.

Kim, S.; Jung, W.; Han, K.; Lee, J. G.; and Mun, Y. Y. 2014. Quality-Based Automatic Classification for Presentation Slides. In *ECIR*, 638–643.

Oyama, S.; and Yamasaki, T. 2019. Visual clarity analysis and improvement support for presentation slides. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 421–428. IEEE.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7262–7272.

Yi, S.; Matsugami, J.; and Yamasaki, T. 2022. Assessment System of Presentation Slide Design Using Visual and Structural Features. *IEICE Transactions on Information and Systems*, 105(3): 587–596.

Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; and Sang, N. 2021. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129(11): 3051–3068.

Yuan, Y.; Chen, X.; and Wang, J. 2020. Object-contextual representations for semantic segmentation. In *European conference on computer vision*, 173–190. Springer.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.