Incomplete Multimodal Learning for Visual Acuity Prediction After Cataract Surgery Using Masked Self-Attention

Qian Zhou[1], Hua Zou[1(✉)], Haifeng Jiang[2], and Yong Wang[2]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] Aier Eye Hospital of Wuhan University, Wuhan University, Wuhan, China
zouhua@whu.edu.cn

Abstract. As the primary treatment option for cataracts, it is estimated that millions of cataract surgeries are performed each year globally. Predicting the Best Corrected Visual Acuity (BCVA) in cataract patients is crucial before surgeries to avoid medical disputes. However, accurate prediction remains a challenge in clinical practice. Traditional methods based on patient characteristics and surgical parameters have limited accuracy and often underestimate postoperative visual acuity. In this paper, we propose a novel framework for predicting visual acuity after cataract surgery using masked self-attention. Especially different from existing methods, which are based on monomodal data, our proposed method takes preoperative images and patient demographic data as input to leverage multimodal information. Furthermore, we expand our method to a more complex and challenging clinical scenario, *i.e.*, the incomplete multimodal data. Firstly, we apply efficient Transformers to extract modality-specific features. Then, an attentional fusion network is utilized to fuse the multimodal information. To address the modality-missing problem, an attention mask mechanism is proposed to improve the robustness. We evaluate our method on a collected dataset of 1960 patients who underwent cataract surgery and compare its performance with other state-of-the-art approaches. The results show that our proposed method outperforms other methods and achieves a mean absolute error of 0.122 logMAR. The percentages of the prediction errors within ± 0.10 logMAR are 94.3%. Besides, extensive experiments are conducted to investigate the effectiveness of each component in predicting visual acuity. Codes will be available at https://github.com/liyiersan/MSA.

Keywords: Incomplete Multimodal Learning · Visual Acuity Prediction · Self-Attention

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43990-2_69.

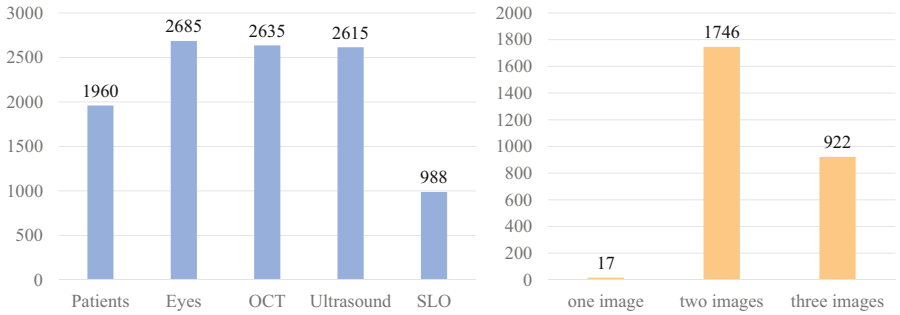
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 735–744, 2023.
https://doi.org/10.1007/978-3-031-43990-2_69

1 Introduction

Cataract has been the leading cause of vision loss. As the only treatment option, cataract surgery is one of the most commonly performed surgeries worldwide. Nevertheless, not all patients achieve complete visual recovery after surgery, which can be due to various factors, such as pre-existing eye conditions, surgical complications, and postoperative inflammation. The ability to accurately predict visual recovery can help clinicians identify high-risk patients and provide appropriate interventions to improve their visual outcomes.

Over the past decades, many efforts have been made to predict the Best Corrected Visual Acuity (BCVA) after cataract surgeries. Most of them are based on traditional approaches like retinometer [8,11,13] or visual electrophysiology [2,5,14]. These methods require specialized expertise to perform and are subject to significant variability in their results. Even though some computer-aided approaches have been proposed, most of them use traditional machine learning algorithms [1] and focus on single-modal data [9,15,16]. While they can be effective to some extent, they often have limited predictive power due to the reliance on a single source of information. In addition, traditional machine learning methods (*e.g.*, linear regression, decision trees, and support vector machines) may not be able to capture complex relationships between different modalities, such as clinical data and imaging data. What's more, clinical scenarios are more complex. For example, the multimodal data may be incomplete due to medical conditions, as shown in Fig. 1. Therefore, there is a need for more sophisticated computer-aided techniques that can integrate multiple sources of data and leverage the strengths of each to improve predictive accuracy as well as address the challenging modality-missing problem.

Transformers [12], which are a type of neural network architecture originally developed for natural language processing tasks, have recently shown



(a) Numbers of each category in the collected dataset. (b) Numbers of eyes with different numbers of images.

Fig. 1. Statistics for the collected dataset. (a) Number of patients, eyes, and three image modalities. OCT refers to Optical Coherence Tomography, and SLO stands for Scanning Laser Ophthalmoscopy. (b) Number of multimodal or monomodal samples. For instance, "two images" denotes the samples with two image modalities. We can see that only one-third of cases have complete multimodal images.

great promise in computer vision applications, such as image classification [4], object detection [3], and segmentation [20]. Transformers have the ability to learn from large amounts of data and can capture complex patterns and relationships in the data, which makes them well-suited for analyzing multimodal learning. Motivated by the tremendous success of Transformers in multimodal learning [10,17,18], we propose a new framework that utilizes incomplete multimodal data for predicting BCVA after cataract surgeries. In particular, our framework contains three stages: modality-specific feature extraction, attentional feature fusion, and visual acuity prediction. Firstly, for each input modality, a pre-trained efficient transformer will be used to extract features. To better leverage the clinical diagnosis keywords, an auxiliary classification loss is added to the image transformer. And to extract text features more efficiently, we apply a CLIP-like [10] input to combine discrete clinical words (*e.g.*, age, sex, and preoperative visual acuity) into sentences. Secondly, we apply an attentional transformer to fuse multimodal features. Specifically, to address the issue of missing modalities, we introduce modality embeddings and attentional masks to prevent the interference of missing modalities with the remaining modalities. Finally, a prediction head takes the fused features as input to predict the BCVA with the Mean Square Error (MSE) loss.

The main contributions of this work can be summarized as follows: (1) We develop a novel framework that uses multimodal data to predict BCVA as well as tackle the complex modality-missing issue. (2) An auxiliary classification loss is adopted to extract more comprehensive pathological features for images. Also, discrete textual words are combined into sentences to better fit the text transformer input. (3) Extensive experiments are conducted to prove the effectiveness of our method. The compared methods include incomplete multimodal learning approaches and monomodal BCVA prediction methods. We also analyze the importance of each component of our method.

2 Method

2.1 Framework Overview

As shown in Fig. 2, our framework contains three parts: modality-specific encoder, multimodal fusion network, and BCVA prediction head. During feature extraction, we take pre-trained transformers as the backbone, *i.e.*, ViT [4] as the image encoder, and CLIP [10] as the text encoder. After that, a cross-modal transformer is used to fuse features from multiple modalities, in which an attentional mask is added to tackle the missing modalities. Finally, a fully connected (FC) layer is used as the prediction head.

2.2 Monomodal Feature Extraction

Text Encoder. The text encoder is a pre-trained CLIP [10] model. The discrete physiological information is combined into a sentence format that benefits the

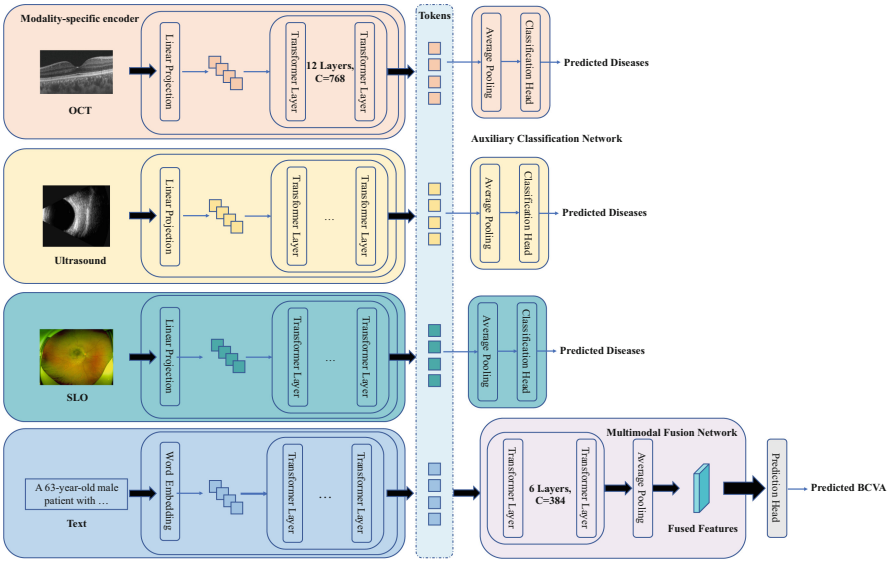


Fig. 2. Pipeline of the proposed framework. The modality-specific encoders utilize vanilla multi-head self-attention. In contrast, the multimodal fusion network employs masked multi-head self-attention. Notably, the fusion network takes features (*i.e.*, tokens) from all modalities as input, whereas each auxiliary classification network receives features from a single modality as input.

text encoder. For instance, the text "male, 67 years old, preoperative visual acuity 0.52 logMAR" will be combined into the sentence "A 67-year-old male patient with preoperative visual acuity of 0.52 logMAR". By combining the text in this way, the physiological texts of all patients are fed into the text transformer in a unified manner. Compared with directly concatenating the texts and inputting them into the model, the combined sentences are more in line with the real scene and are easier to be understood by the model to extract key semantic information. Moreover, the CLIP text encoder is trained on a large text corpus, enabling excellent generalization performance. Thus, during the training of the overall model, the weights of the text encoder can be fixed.

Image Encoder. The image encoder adopts ViT [4]. Since ViT is trained on natural images and may not be directly applicable to medical images, it can not be fixed during the training. However, the pre-trained weights can be utilized to expedite convergence. In addition, there are diagnostic keywords given by ophthalmologists for each image in the dataset. It is not appropriate to directly use the CLIP text encoder to extract medical features from these keywords since CLIP is trained on natural language texts. To this end, we introduce an auxiliary classification loss in the image encoder. Specifically, for each input image, a multi-label classification network is incorporated after the image encoder to predict

the diseases contained in the image. The classification network is composed of an average pooling layer and a fully connected layer. For simplicity, we adopt the binary cross-entropy loss as the auxiliary classification loss as Eq. (1).

$$L_{CLS} = \sum W^i L^i_{BCE} \tag{1}$$

where W^i equals 1 if the i-th modality is available and equals 0, otherwise.

By adding the classification loss, the final loss to train the whole model is:

$$L = L_{MSE} + \alpha L_{CLS} \tag{2}$$

where L_{MSE} is the mean square error loss between the predicted BCVA and the ground truth, α is a hyper-parameter and set to 0.5.

2.3 Multimodal Feature Fusion

We use a cross-modal Transformer as the multimodal fusion network. The obtained modality-specific tokens are projected into the same dimension and concatenated into an input sequence. In contrast to the modality-specific Transformer, besides adding positional embeddings to all input tokens, we also add learnable modality-specific embeddings to all tokens indicating the modality information.

Complete Multimodal Learning. In the collected dataset, all cases have corresponding text modality, and almost all cases have corresponding OCT modality. Therefore, we built a complete multimodal prediction model based on the text modality and OCT modality. In such a situation, only the OCT encoder and text encoder will be preserved, while the SLO encoder and Ultrasound encoder will be dropped. Since transformers can handle sequences of any input length, we don't need to make any changes to the fusion network. Using complete multimodal learning, we can compare our methods to other BCVA prediction approaches which do not consider the incomplete multimodal scenario.

Incomplete Multimodal Learning. Not all cases have complete modalities for the three image modalities (*i.e.*, OCT, SLO, and Ultrasound). For the missing modalities, one possible way is to simply represent them by 0 values [18]. However, the 0 values will be regarded as noise by the model as they do not contain any useful information. To avoid model degradation, we add attentional masks in the vanilla self-attention to exclude the interactions among missing modalities and available modalities.

The proposed attentional masks are easy to implement. As shown in Eq. (3), self-attention in transformers is mainly matrix multiplication.

$$Attn(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_z}})V \tag{3}$$

in which, Q, K, and V are queries, keys, and values obtained from tokens, respectively. d_z is the projection dimension. To avoid interactions between irrelevant tokens, the masked self-attention is computed as:

$$Attn_{Mask}(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_z}} + M)V \tag{4}$$

in which, M is the mask matrix. For each element in M, it will be 0 if the interactions should be included, or it will be negative infinity to avoid unnecessary interactions. By adding negative infinity, the results of softmax will be very close to 0. Figure 3 shows an example of masks when modalities are missing.

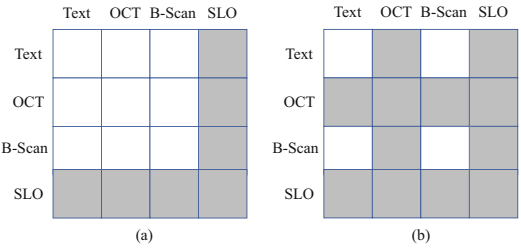


Fig. 3. An example of the attentional mask. (a) means only SLO is missing; (b) represents both OCT and SLO are missing. White cells represent a value of 0, and gray cells represent a value of negative infinity. (Color figure online)

2.4 Implementation Details

To validate the effectiveness of the proposed method, we have conducted extensive experiments implemented with Pytorch and 8×RTX 3090 GPUs. The input images are resized to 224 × 224. The Adam optimizer is adopted with an initial learning rate of 0.001 and $\beta_1 = 0.9$, $\beta_2 = 0.99$. The mini-batch size is set to 32. We train the model for 100 epochs in total, and the learning rate will be decayed by 0.1 every 20 epochs. Besides, we randomly split all samples to 80% for training and 20% for testing. All experiments are conducted with 5-fold cross-validation to produce more solid results.

3 Experiments

3.1 Datasets and Evaluation Metrics

The collected dataset consists of 1960 patients (2685 eyes) having cataract surgeries at Aier eye hospital of Wuhan University. The collected modalities are texts and images. The images contain 2635 Optical Coherence Tomography (OCT), 2615 Ultrasound, and 988 Scanning Laser Ophthalmoscopy (SLO). The textual

information includes sex, age, preoperative and postoperative visual acuity. For each image, three ophthalmologists will label it to obtain the diagnosis (*i.e.*, clinical diagnosis keywords) of 14 retinal diseases. The retinal diseases include normal, vitreous opacity, posterior staphyloma, stellate vitreous degeneration, pathological myopia changes, retinal atrophy, macular degeneration, epiretinal membrane, ellipsoid band partially missing, retinoschisis, retinal hemorrhage, macular edema, macular hole, and retinitis pigmentosa. We use Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and prediction accuracy as the metrics. For simplicity, we consider predictions to be accurate if the prediction errors are within $\pm$ 0.10 logMAR.

3.2 Quantitative Performance

We have compared the results on the collected dataset with other approaches. The compared methods include state-of-the-art methods which aim to predict BCVA using OCT like CTT-Net [15], Wei *et al.* [16], and other algorithms considering incomplete multimodal learning such as Huang *et al.* [6], Ma *et al.* [7], and Zhao *et al.* [19]. For the former methods, we compare our method with them using the complete data. As for the latter approaches, they are not proposed for BCVA prediction. Therefore, we finetune them so that they can be applied to incomplete BCVA prediction data.

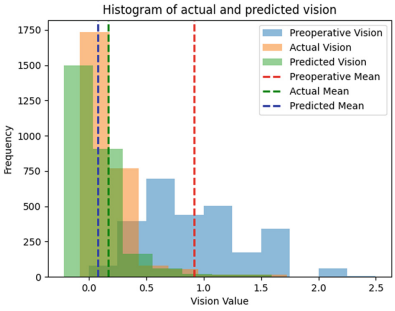


Fig. 4. Distribution of preoperative, predictive, and postoperative visual acuity. Actual vision means the actual postoperative visual acuity.

From Table 1, we can see that our proposed framework achieves the best performance. Specifically, we have improved CTT-Net [15] and shown a sharp rise compared to Wei *et al.* [16] on the complete dataset. Even though CCT-Net uses both text and oct modalities, the utilization of text is still limited. Wei *et al.* [16] directly ignores the textual information and only uses some simple frameworks to predict BCVA, thus achieving the worst results. When using incomplete data, the performance is also improved greatly, and we still achieve the best performance. Huang *et al.* [6] try to apply image synthesis to solve the modality-missing problem. However, the collected images in our dataset are not

Table 1. Quantitative prediction results on the collected dataset. Complete means text and OCT modalities are available and complete. Incomplete means text and three image modalities are available, but the image modalities may be incomplete.

Dataset	Methods	MAE (↓)	SMAPE (↓)	Accuracy (↑)
Complete	CTT-Net [15] (OCT+Text)	0.168 ± 0.014	85.236 ± 3.277	0.887 ± 0.022
	CTT-Net [15] (OCT)	0.174 ± 0.013	89.635 ± 2.881	0.872 ± 0.016
	Wei *et al.* [16] (OCT)	0.237 ± 0.093	93.587 ± 3.236	0.723 ± 0.056
	Ours (OCT)	0.153 ± 0.012	65.615 ± 1.690	0.901 ± 0.018
	Ours (OCT + Text)	0.142 ± 0.009	62.550 ± 1.668	0.923 ± 0.014
Incomplete	Huang *et al.* [6]	0.176 ± 0.054	88.672 ± 3.051	0.854 ± 0.017
	Ma *et al.* [7]	0.139 ± 0.013	61.722 ± 2.007	0.917 ± 0.015
	Zhao *et al.* [19]	0.133 ± 0.021	59.673 ± 2.362	0.921 ± 0.021
	Ours	$\mathbf{0.122 \pm 0.007}$	$\mathbf{57.165 \pm 1.610}$	$\mathbf{0.943 \pm 0.012}$

Table 2. Ablation study results on incomplete datasets.

Model	MAE (↓)	SMAPE (↓)	Accuracy (↑)
Baseline	0.176 ± 0.014	87.642 ± 3.023	0.874 ± 0.014
Baseline + combined sentences	0.163 ± 0.012	78.932 ± 3.672	0.893 ± 0.011
Baseline + cls loss	0.157 ± 0.009	71.023 ± 2.346	0.912 ± 0.015
Baseline + attentional mask	0.145 ± 0.010	62.328 ± 2.064	0.925 ± 0.013
Baseline + all	$\mathbf{0.122 \pm 0.007}$	$\mathbf{57.165 \pm 1.610}$	$\mathbf{0.943 \pm 0.012}$

aligned, and image synthesis may not work in such a situation. Zhao *et al.* [19] propose to learn common representations of all modalities, and this idea works in cases of minorly missing modalities but not in our dataset. Ma *et al.* [7] achieve similar performance to our proposed method due to their robust design. The results show that our model is capable of extracting modality-specific features as well as fusing them in an effective way. Besides, the performance also shows the robustness of our proposed method. Note that almost all results on the incomplete multimodal dataset outperform results on the complete multimodal dataset. This is due to the incomplete multimodal dataset will always contain OCT and text modalities, thus having more information in the input. Figure 4 shows the distribution of preoperative, predictive, and postoperative visual acuity. We can see that the predicted visual acuity largely overlaps with the true postoperative visual acuity, demonstrating the effectiveness of our method.

3.3 Ablation Study

As shown in Table 2, the proposed framework mainly benefits from the auxiliary classification loss and attentional fusion mask. Our analysis is as follows: Images provide more valuable information than text, making the auxiliary classification loss more effective than text combining. In missing multimodal learning, feature

fusion takes precedence, and the masked self-attention mechanism contributes the most. Additionally, we conducted experiments to evaluate each modality's effectiveness. The results can be seen in the supplementary material.

4 Conclusion

In this paper, we present a new framework for BCVA prediction on the collected incomplete multimodal dataset. We take full advantage of multimodal information through our framework. The text modality is better utilized through the combination of the words. Moreover, image modality is explored effectively by the auxiliary classification loss. The attentional mask addresses the modality-missing issue. Extensive experiments have proved the effectiveness and superiority of our method.

Acknowledgements. This work is partially supported by Bingtuan Science and Technology Program (No. 2022DB005 and 2019BC008) and Key Research and Development Program of Hubei Province (2022BCA009).

References

1. Alexeeff, S.E., et al.: Development and validation of machine learning models: electronic health record data to predict visual acuity after cataract surgery. Perm. J. **25**, 188 (2021)
2. An, J., Zhang, L., Wang, Y., Zhang, Z.: The success of cataract surgery and the preoperative measurement of retinal function by electrophysiological techniques. J. Ophthalmol. **2015**, 401281 (2015)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Forshaw, T.R.J., Ahmed, H.J., Kjær, T.W., Andréasson, S., Sørensen, T.L.: Full-field electroretinography in age-related macular degeneration: can retinal electrophysiology predict the subjective visual outcome of cataract surgery? Acta Ophthalmol. **98**(7), 693–700 (2020)
6. Huang, Z., Lin, L., Cheng, P., Peng, L., Tang, X.: Multi-modal brain tumor segmentation via missing modality synthesis and modality-level attention fusion. arXiv preprint arXiv:2203.04586 (2022)
7. Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X.: Are multimodal transformers robust to missing modality? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18177–18186 (2022)
8. Mimouni, M., Shapira, Y., Jadon, J., Frenkel, S., Blumenthal, E.Z.: Assessing visual function behind cataract: preoperative predictive value of the Heine lambda 100 Retinometer. Eur. J. Ophthalmol. **27**(5), 559–564 (2017)
9. Obata, S., et al.: Prediction of postoperative visual acuity after vitrectomy for macular hole using deep learning-based artificial intelligence. Graefe's Archive for Clinical and Experimental Ophthalmology, pp. 1–11 (2021)

10. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
11. Tharp, A., Cantor, L., Yung, C.W., Shoemaker, J.: Prospective comparison of the Heine Retinometer with the mentor Guyton-Minkowski potential acuity meter for the assessment of potential visual acuity before cataract surgery (1994)
12. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)
13. Wald, C.S., Unterlauft, J.D., Rehak, M., Girbardt, C.: Retinometer predicts visual outcome in Descemet membrane endothelial keratoplasty. Graefes Arch. Clin. Exp. Ophthalmol. **260**(7), 2283–2290 (2022)
14. Wang, H., et al.: Electrophysiology as a prognostic indicator of visual recovery in diabetic patients undergoing cataract surgery. Graefes Arch. Clin. Exp. Ophthalmol. **259**, 1879–1887 (2021)
15. Wang, J., et al.: CTT-Net: a multi-view cross-token transformer for cataract post-operative visual acuity prediction. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 835–839. IEEE (2022)
16. WeiL, L., et al.: An optical coherence tomography-based deep learning algorithm for visual acuity prediction of highly myopic eyes after cataract surgery. Front. Cell Develop. Biol. **9**, 652848 (2021)
17. Xu, J., et al.: GroupViT: semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18134–18144 (2022)
18. Zhang, Y., et al.: mmFormer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. LNCS, vol. 13435. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_11
19. Zhao, J., Li, R., Jin, Q.: Missing modality imagination network for emotion recognition with uncertain missing modalities. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 2608–2618 (2021)
20. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)