Retinal Thickness Prediction from Multi-modal Fundus Photography

Yihua Sun[1], Dawei Li[2], Seongho Kim[3], Ya Xing Wang[4], Jinyuan Wang[4], Tien Yin Wong[5,6], Hongen Liao[1(✉)], and Su Jeong Song[3(✉)]

[1] Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing, China
liao@tsinghua.edu.cn

[2] College of Future Technology, Peking University, Beijing, China

[3] Department of Ophthalmology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
sjsong7@gmail.com

[4] Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital University of Medical Science, Beijing Ophthalmology and Visual Sciences Key Laboratory, Beijing, China

[5] Tsinghua Medicine, Tsinghua University, Beijing, China

[6] Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore

Abstract. Retinal thickness map (RTM), generated from OCT volumes, provides a quantitative representation of the retina, which is then averaged into the ETDRS grid. The RTM and ETDRS grid are often used to diagnose and monitor retinal-related diseases that cause vision loss worldwide. However, OCT examinations can be available to limited patients because it is costly and time-consuming. Fundus photography (FP) is a 2D imaging technique for the retina that captures the reflection of a flash of light. However, current researches often focus on 2D patterns in FP, while its capacity of carrying thickness information is rarely explored. In this paper, we explore the capability of infrared fundus photography (IR-FP) and color fundus photography (C-FP) to provide accurate retinal thickness information. We propose a **M**ulti-**M**odal **F**undus photography enabled **R**etinal **T**hickness prediction network (**M^{2}FRT**). We predict RTM from IR-FP to overcome the limitation of acquiring RTM with OCT, which boosts mass screening with a cost-effective and efficient solution. We first introduce C-FP to provide IR-FP with complementary thickness information for more precise RTM prediction. The misalignment of images from the two modalities is tackled by the Transformer-CNN hybrid design in M^{2}FRT. Furthermore, we obtain the ETDRS grid prediction solely from C-FP using a lightweight decoder,

S. J. Song and H. Liao are the co-corresponding authors.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43990-2_55.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 585–595, 2023.
https://doi.org/10.1007/978-3-031-43990-2_55

which is optimized with the guidance of the RTM prediction task during the training phase. Our methodology utilizes the easily acquired C-FP, making it a valuable resource for providing retinal thickness quantification in clinical practice and telemedicine, thereby holding immense clinical significance.

Keywords: Retinal thickness prediction · Multi-modality · Transformer · Color fundus photography · Infrared fundus photography

1 Introduction

Retinal thickness map (RTM), generated from optical coherence tomography (OCT) volumes, provides a quantitative representation of various retina pathologic conditions [3]. The ETDRS grid is an array comprising nine values representing the averaged thickness in nine regions in RTM [5]. The RTM and ETDRS grid, are widely employed diagnostic and monitoring techniques for retinal disorders including age-related macular degeneration, glaucoma, and diabetic retinopathy [14], which are prevalent causes of visual impairment worldwide [6]. On the other hand, OCT has been a critical diagnostic tool in ophthalmology due to its exceptional sensitivity and precision in identifying major eye diseases.

However, OCT exams are only available to limited patients as it is both costly and time-consuming, which impedes the acquisition of RTM and ETDRS grid. The recent advances in deep learning [20,21] have prompted research efforts aimed at addressing this limitation. There have been attempts to predict center-involving macular edema from color fundus photographs (C-FP) [17]. Although these studies showed high sensitivity and specificity, they only provided a binary classification for the presence of macular edema. The lack of quantitative retina thickness prediction results mandated further study.

Fundus photography (FP) is widely used to image the retina, which captures the reflected signal of emitted signal from the retinal surface with a flash of light [13]. As the retina is partially-transparent, a minority of light would pass through the surface [19] and reflect back, which might carry information about the retinal thickness. This hypothesis motivates us to explore the connection between the RTM/ETDRS grid and the IR-FP/C-FP, which is rarely explored. Nonetheless, the FPs hold substantial clinical value in facilitating large-scale screening by acquiring RTM and ETDRS grid much faster and more affordable.

Recently, Holmberg et al. [10] presented DeepRT, a convolutional neural network (CNN) designed for predicting retinal thickness using only infrared fundus photographs (IR-FP), disregarding C-FP. Exploring the capacity of **C-FP** to provide depth information has two major advantages: 1) **More precise RTM prediction**: Different from IR-FP, C-FP is acquired using light of multiple wavelengths that penetrate different depths in the retina [19]. We assume that this can provide richer thickness information, which can lead to more precise RTM prediction when combined with IR-FP; 2) **Clinical significance**: C-FP is the most commonly used diagnostic tool in ophthalmology, and can be obtained even

using a smartphone [7]. The ability to derive thickness information from C-FP alone, without OCT scans, will make C-FP a potential tool for high functioning telemedicine platform which has the ability to diagnose, monitor treatment response, and even screen high-risk patients for diabetic macular edema (DME).

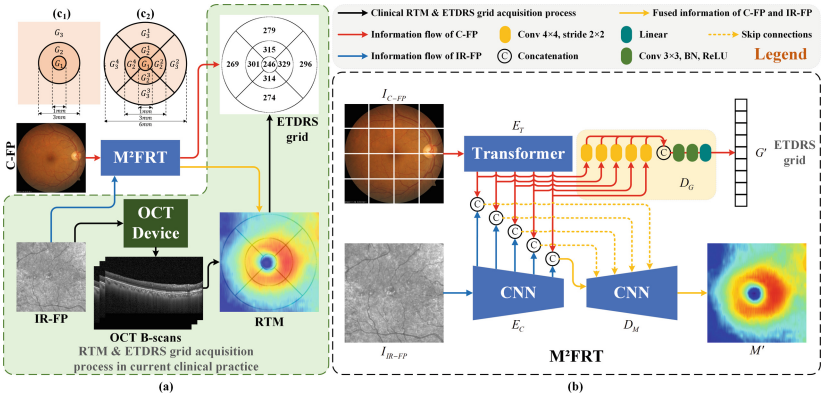

Fig. 1. (a) M^{2}FRT predicts the RTM with enhanced thickness information from mutimodal FPs without OCT scans, and the ETDRS grid can be predicted with C-FP only. (b) M^{2}FRT utilizes Transformer/CNN hybrid encoders E_T/E_C for C-FP/IR-FP, to tackle the unregistered issue and extract 2D aligned thickness information in an end-to-end learning manner. A CNN decoder D_M is employed to predict the RTM. Besides, a lightweight decoder D_G is designed to predict the ETDRS grid base on the information from C-FP only, which is guided by the RTM prediction task during training. (c_1) Areas for evaluations in the RTM prediction task, and (c_2) in the ETDRS grid (left eye) prediction task. In (c_1), G_3 is the remaining area of $G_{1,2}$ in RTM.

In this paper, we explore the capability of IR-FP and C-FP to provide accurate retinal thickness information, with a cohort of patients with DME of different grades. We propose a **M**ulti-**M**odal **F**undus photography enabled **R**etinal **T**hickness prediction network (**M^{2}FRT**). It is comprised of two separate encoders, a CNN E_C and a Transformer E_T, that encode localized information and rich depth information form IR-FP and C-FP respectively. We utilize the features extracted from E_C to facilitate the learning process of E_T in gathering 2D aligned thickness information via its attention mechanism. The enriched features are subsequently fed into a decoder to predict the RTM.

Furthermore, we obtain the ETDRS grid prediction, i.e. nine values representing averaged thickness in the predefined areas in Fig. 1 (c_2), solely from the C-FP by processing the features extracted from E_T through another lightweight decoder, which has significant clinical implications. To the best of our knowledge, we are the first to demonstrate the benefit of C-FP for RTM prediction and derive the ETDRS grid prediction solely from C-FP.

2 Methodology

In this study, we exclusively concentrate on DME to explore the predictive capacity of FPs regarding the retinal thickness. The rationale behind this is that, apart from DME, predicting retinal thickness itself has relatively less clinical value. For example, for age-related macular degeneration, the ophthalmologist needs to look for subtle changes in abnormal OCT features (e.g. subretinal fluid, pigmentary epithelial detachments [16]), rather than just the retinal thickness.

In standard clinical settings, the ophthalmologist will acquire the C-FP upon patients' arrival. If RTM is deemed necessary for diagnosis, a separate device will capture IR-FP and conduct OCT scanning. Figure 1 (a) illustrates the acquisition process of RTM using OCT, where each B-scan is registered with the 2D positions in IR-FP. The ETDRS grid is an array comprising nine values indicating the average thickness (μm) in nine predefined regions in RTM (Fig. 1 (c_2)).

Dovetailed with the clinical settings, M^{2}FRT aims to predict the RTM corresponding to the IR-FP, utilizing enriched depth information from pre-collected C-FP. The RTM requires precise pixel-wise correspondences to the IR-FP, while the ETDRS grid is a regional concept. Therefore, we can manage to derive an ETDRS grid prediction using only easier acquired C-FP, even in the absence of IR-FP, which holds importance within clinical scenarios and telemedicine.

As mentioned above, the FPs from the two modalities are captured by different machines. So, the FPs are not registered and have a distinct field of view (FoV). The recent advances in vision Transformers [4,12,18] have inspired us to address this challenge, because the multi-head attention mechanism is location-agnostic, but rather leverages patch embedding and position encoding to introduce positional information.

2.1 Encoder

The overall pipeline of M^{2}FRT is presented in Fig. 1 (b). The notations used for the images in the modality of IR-FP and C-FP are $I_{IR\text{-}FP}$ and $I_{C\text{-}FP}$, respectively. The objective is to predict the thickness map M in the FoV of $I_{IR\text{-}FP}$ and ETDRS grid G, which represents the central area of the retina and is the major concern in clinical practices.

The convolution and concatenation pose "hard" operations on the spatial dimensions. Thus, whether we concatenate $I_{IR\text{-}FP}$ and $I_{C\text{-}FP}$ as input or in the feature space under a CNN backbone, the misalignment of $I_{IR\text{-}FP}$ and $I_{C\text{-}FP}$ will deteriorate the performance for M prediction. In contrast, the spatial information is "softly" incorporated into the Transformer architecture, where the subsequent operations in the feature space are location-agnostic.

Therefore, we utilize a CNN encoder E_C from U-Net [15] to extract features from $I_{IR\text{-}FP}$, and a Transformer encoder E_T from 2D ViT/UNETR [4,8] to extract features from $I_{C\text{-}FP}$. Notably, the deep features extracted by E_T are spatially perturbed. M^{2}FRT leverages attention mechanisms in E_T to gather 2D aligned thickness information from $I_{C\text{-}FP}$, guided by the features extracted from

$I_{IR\text{-}FP}$ by E_C. The extracted multi-level features from $I_{IR\text{-}FP}$ and $I_{C\text{-}FP}$ are denoted as $f_{IR\text{-}FP}$ and $f_{C\text{-}FP}$ respectively, as shown in the following equations:

$$f_{IR\text{-}FP} = E_C(I_{IR\text{-}FP}), \quad f_{C\text{-}FP} = E_T(I_{C\text{-}FP}). \tag{1}$$

2.2 Decoder

M^{2}FRT extracts 2D aligned depth information from C-FP, which enrich the depth representations acquired from IR-FP in an end-to-end learning manner. The extracted features are fused by concatenation and passed to the decoder D_M to generate the thickness map prediction M', where $M' = D_M(f_{IR\text{-}FP}, f_{C\text{-}FP})$.

With fine-grained thickness information extracted for the RTM prediction task, the encoded features obtained from E_T are ready to be decoded to predict the ETDRS grid using a lightweight decoder D_G. In D_G, the features from multiple levels are combined using a series of convolutions and concatenations. Then the final prediction for G is generated by a linear projection. The predicted ETDRS grid is denoted as G', where $G' = D_G(f_{C\text{-}FP})$.

2.3 Loss Functions

The loss functions $\mathcal{L}_1^M$ and $\mathcal{L}_1^G$ are employed in the prediction of the RTM and ETDRS grid using L_1 criteria, respectively, as shown in the following equations,

$$\mathcal{L}_1^M = \|M - M'\|_1, \; \mathcal{L}_1^G = \frac{1}{9}\sum_{i=1}^{9}\left|G^{(i)} - G'^{(i)}\right|, \tag{2}$$

where $G^{(i)}$ and $G'^{(i)}$ are the i-th number in the ETDRS grid ground truth G and prediction G'. The final loss function is $\mathcal{L} = \mathcal{L}_1^M + \mathcal{L}_1^G$.

3 Experiments

3.1 Experimental Setup

Dataset. A total of 967 retinal images were gathered from 361 distinct patients diagnosed with DME of different grades who underwent intravitreal injections. The dataset is collected in Kangbuk Samsung Hospital (IRB Approval Number: KBSMC 2022-12-016-001) between 2020 and 2021. The averaged retinal thickness (μm) in the dataset is 275.92 ± 20.91 (mean $\pm$ std.). For each patient, 31 B-scans are obtained by a Heidelberg OCT device, which are used to calculate the retinal thickness between the internal limiting membrane and the Bruch's membrane. The segmentations of the membrane layers are directly exported from the OCT machine. Images with poor fixation or OCTs with major segmentation errors are excluded by an experienced ophthalmologist.

Data Pre-processing. For IR-FP, we center-crop the area corresponding to the OCT scanning area with a resolution of 544×544, and then calculate RTM

Table 1. Quantitative comparison of different methods for RTM prediction, with MAE (μm) and PSNR (dB). The top-2 methods are highlighted in bold and underlined. By incorporating multi-modal FP as input, networks can access more comprehensive thickness information, resulting in improved performance. The most efficient way to tackle the unregistered problem is to utilize encoders E_C and E_T for IR-FP and C-FP respectively. Asterisks indicate M^{2}FRT outperforms the baselines with p-values<0.01.

Inputs	Methods			RTM		G_1		G_2		G_3	
				MAE↓	PSNR↑	MAE↓	PSNR↑	MAE↓	PSNR↑	MAE↓	PSNR↑
IR-FP	UNet++ [22]			29.28*	25.93*	64.52*	22.09*	31.71*	27.68*	27.92*	26.11*
	DeepRT [10]			25.49*	28.04*	66.57*	21.88*	28.96*	28.93*	23.78*	28.81*
	E_T, D_M			27.21*	27.87*	40.42*	27.64*	29.23	29.46	26.48*	28.05*
	U-Net [15] (E_C, D_M)			27.84*	27.38*	60.98*	22.63*	30.90*	28.43*	26.40*	27.86*
IR-FP & C-FP	U-Net [15]			25.16*	28.36*	44.07*	26.32*	28.92*	29.19*	23.95*	28.72*
	IR-FP Enc.	C-FP Enc.	Dec.	–	–	–	–	–	–	–	–
	E_T	E_T	D_M	26.44*	27.99*	40.01	27.88	28.54	29.54	25.68*	28.19*
	E_T	E_C	D_M	25.33*	28.49*	39.52	28.05	28.58	<u>29.76</u>	24.33*	28.76*
	E_C	E_C	D_M	24.80*	28.54*	43.19*	26.87*	28.93*	29.32*	23.54*	28.92*
	E_C	E_T	D_M	<u>23.82</u>	<u>28.91</u>	<u>38.92</u>	**28.16**	**27.80**	**29.78**	<u>22.65</u>	<u>29.28</u>
	E_C	E_T	D_M, D_G	**23.80**	**28.92**	**38.60**	<u>28.12</u>	<u>28.29</u>	29.64	**22.54**	**29.33**

Enc.: Encoder; Dec.: Decoder. M^{2}FRT is comprised of E_C, E_T, D_M, D_G.

ground truth within. With respect to the B-scans, the retinal thickness is calculated for 31 lines in the 2D IR-FP, and then linearly interpolated to match the resolution of IR-FP. For C-FP, we resize it to 544×544 from an original resolution of 3608×3608. The dataset is randomly split into training and test datasets at the patient level. The training/test dataset consisted of 657/310 images from 252/109 patients, respectively.

Implementation Details. The M^{2}FRT is implemented with PyTorch [2] and MONAI [1], and detailed configurations are in the supplementary material. Random flipping and rotation are utilized for data augmentation. We use the Adam [11] optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$ for training for 300 epoches. The initial learning rate is 0.001 and exponentially decayed with $\gamma = 0.999$.

Performance Metrics. For the RTM predictions, we use mean absolute error (MAE) and peak signal-to-noise ratio (PSNR) for evaluation in the areas $G_{1,2,3}$ as shown in Fig. 1 (c_1), where the peak signal is set to $800\mu m$. For the ETDRS grid predictions, we calculate the MAE of the predictions of the nine grids, as shown in Fig. 1 (c_2). For the right eye, the grid must be mirrored horizontally, i.e., $G_3^2 \leftrightarrow G_3^4$ and $G_2^2 \leftrightarrow G_2^4$. The Wilcoxon signed-rank test is employed to compare the performance of M^{2}FRT with the baselines.

3.2 Quantitative and Qualitative Evaluations on RTM Predictions

To better illustrate the problem and our solution, we begin with the most concise design, U-Net [15]. In Table 1, the MAE/PSNR for U-Net with IR-FP as input are 27.84 µm/27.38 dB. By concatenating multi-modal IR-FP and C-FP as input to the U-Net, the performance improved to 25.16 µm/28.36 dB, indicating that C-FP has the potential of containing additional thickness information.

However, the multi-modal FPs are unregistered and have a distinct FoV, in which case a mere concatenation of these inputs would diminish the network's

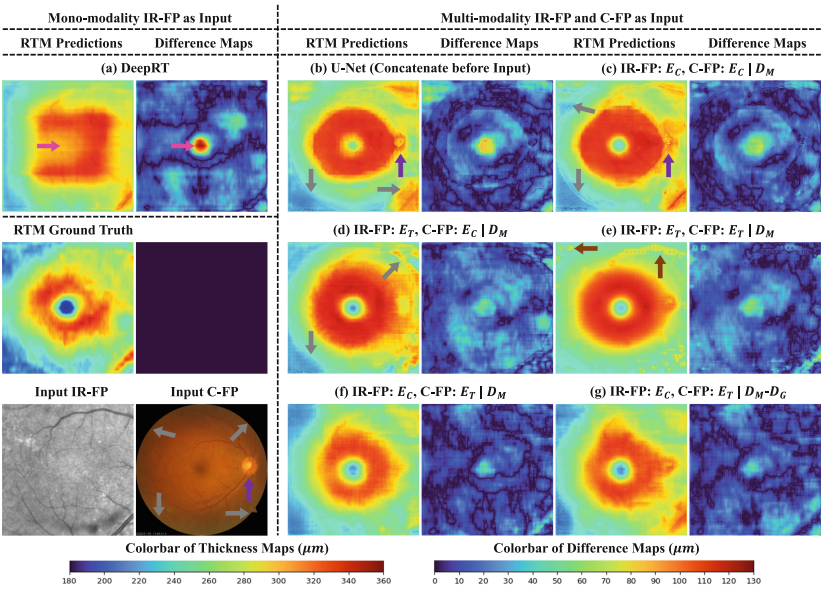

Fig. 2. The use of IR-FP alone to predict RTM will cause larger errors, particularly in the central area (pink arrows). Artifacts can be generated by the annular boundary (grey arrows) and vessels (purple arrows) from C-FP when misaligned information is roughly fused. Additionally, the dual E_T design weakens the localized 2D correspondence, where patch embedding can generate artifacts (brown arrows). Our proposed methods (f) and (g), with E_C and E_T extracting features for IR-FP and C-FP respectively, better leverage 2D aligned thickness information and lead to lower errors. (Color figure online)

capacity to effectively exploit the thickness information from paired 2D positions. A simple solution is to encode the multi-modal FPs with two separated convolutional encoders E_C, where the features are deeply fused along the downsampling path. The unregistered problem is eased by the higher-level features with a larger receptive field, and the MAE/PSNR are improved to 24.80 µm/28.54 dB.

After all, 2D convolution and concatenation pose a "hard" operation to the spatial dimensions, which is still interfered with by the unregistered problem. As shown in Fig. 2, in (b) and (c), there are artifacts in the RTM predictions caused by the annular boundary (grey arrows) and misaligned vessels from C-FP (purple arrows). On the contrary, the attention operations in Transformer are location-agnostic, where the spatial information is more "softly" introduced into the network by patch embedding and position encoding [4].

Therefore, we employ distinct encoders of a CNN E_C and a Transformer E_T to IR-FP and C-FP respectively. The attention mechanism in E_T is encouraged to gather 2D aligned thickness information from the perturbed patch embeddings, with the guidance from the decoder D_M and the $\mathcal{L}_1^M$ loss function. With this CNN-Transformer hybrid design, the MAE/PSNR performance are

Table 2. Quantitative comparison of different methods with MAE (μm). Our method incorporates pixel-wise supervision from the RTM prediction branch, and improves the MAE results. Asterisks indicate M^{2}FRT outperforms the baselines with p-values<0.05.

Methods	Mean Absolute Error (μm)									
	ETDRS Grid	G_1	G_2^1	G_2^2	G_2^3	G_2^4	G_3^1	G_3^2	G_3^3	G_3^4
ResNet-50 [9]	25.12*	37.46*	27.22	29.73*	25.89	26.43*	19.12*	22.27	17.65	20.36*
ResNet-101 [9]	25.06*	36.41*	27.05	29.29	25.76	26.18*	19.32	22.60*	17.86	21.11*
E_T, D_G	24.42*	34.88	26.70	29.29*	25.63	25.13	18.76	22.18*	**17.38**	19.85*
E_T, D_G, E_C, D_M	**23.84**	**34.36**	**26.15**	**28.24**	**25.25**	**24.53**	**18.50**	**21.42**	17.44	**18.71**

M^{2}FRT is comprised of E_T, D_G, E_C, D_M.

improved to 23.82 µm/28.91 dB in Table 1, and the network produced the best visual quality and smaller errors in Fig. 2 (f) and (g).

Since IR-FP acts as a localizer for the OCT scan and RTM, spatially perturbing the features from IR-FP with E_T is not appropriate for the accurate prediction of RTM, and thus not yielding better quantitative results, as shown in Table 1. In Fig. 2 (d), the annular boundary artifacts from C-FP still exist (grey arrows). When both encoders are substituted by E_T, in Fig. 2 (e), the 2D localizing information is degraded, in which case, there will be artifacts caused by the patch embedding (brown arrows).

Our proposed M^{2}FRT utilizes a combination of multi-modal IR-FP and C-FP to predict the RTM. M^{2}FRT outperforms the state-of-the-art (SOTA) RTM prediction technique, DeepRT [10], which uses mono IR-FP as input. Besides, methods with multi-modal FPs surpass methods with mono IR-FP as input, especially in the central G_1 area, as shown in Table 1 and pink arrows in Fig. 2. The results demonstrate that C-FP has the ability to provide complementary depth information with IR-FP. The effectiveness of our methodology is validated through the ablation study on the encoders and decoders, as presented in Table 1.

Additionally, when E_T is guided to gather aligned features for RTM using the attention mechanism, the deep features from E_T are ready to be decoded by D_G for ETDRS grid predictions, which involves computing the averaged thickness in nine predefined regions. Notably, the ETDRS grid prediction task does not have a significant impact on the performance of the RTM prediction (the last two rows in Table 1), while the ETDRS grid prediction task can benefit from the supervision provided by the RTM prediction task, which will be discussed in Sect. 3.3.

3.3 Quantitative Evaluations on ETDRS Grid Predictions

Following the clinical settings, we predict the full RTM based on the IR-FP localizer in place of the OCT scanning procedure, which can boost mass screening. We gather enriched thickness information from C-FP and improve the performance with a hybrid CNN-Transformer design, as elaborated in Sect. 3.2.

In addition to identifying 2D disease patterns in C-FP, predicting the ETDRS grid solely from C-FP can exploit additional information in the C-FP and hold significant clinical value for rapid diagnosis, especially in the field of telemedicine. To achieve this, we can adopt a conventional learning-based method to predict the nine numbers in the ETDRS grid, i.e. ResNet [9], as shown in Table 2.

However, simply approximating the nine numbers will neglect the fine-grained thickness information. To address this issue, following the design in Sect. 3.2, the encoder E_T for C-FP is guided by the encoder E_C from the IR-FP part for detailed RTM predictions. Therefore, E_T has been trained to extract fine-grained depth information from C-FP, which can be decoded for the averaged thickness for ETDRS grid predictions with D_G. The fine-grained thickness supervision from the RTM prediction task can benefit the ETDRS grid prediction task, as shown in the last two rows of Table 2. Besides, our proposed M^{2}FRT outperforms its ablation and other baselines, as shown in Table 2. We can also observe from Table 1 and 2 that the central thickness in G_1 area is more challenging to predict than the surrounding area for the RTM and ETDRS grid prediction task.

4 Conclusion

In this paper, we demonstrate the advantages of leveraging multi-modal information from C-FP for RTM prediction with respect to IR-FP, which overcomes the limitations of OCT and has the potential to enhance mass screening. Additionally, we propose a novel method for predicting the ETDRS grids solely from C-FP, which has significant clinical importance for fast diagnosis, telemedicine, etc. Our results indicate that additional fine-grained supervision from the RTM prediction task is beneficial for ETDRS grid prediction, where the ETDRS grid is decoded from the encoder of C-FP by a lightweight decoder during the training procedure of the RTM prediction task. Further research could be conducted for: 1) Predicting RTM of multiple retinal layers simultaneously, and 2) Improving RTM prediction's resolution and detail by acquiring finer OCT as ground truth.

Acknowledgments. The authors acknowledge supports from National Key Research and Development Program of China (2022YFC2405200), National Natural Science Foundation of China (82027807, U22A2051), Beijing Municipal Natural Science Foundation (7212202), Institute for Intelligent Healthcare, Tsinghua University (2022ZLB001), and Tsinghua-Foshan Innovation Special Fund (2021THFS0104). We would like to thank Hee Guan Khor for discussions on experiments and writing, and Zhuxin Xiong for discussions on data pre-processing.

References

1. Medical open network for artificial intelligence (MONAI). https://monai.io/
2. PyTorch. https://pytorch.org/
3. Bhende, M., Shetty, S., Parthasarathy, M.K., Ramya, S.: Optical coherence tomography: a guide to interpretation of common macular diseases. Indian J. Ophthalmol. **66**(1), 20–35 (2018)

4. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
5. Early Treatment Diabetic Retinopathy Study Research Group: grading diabetic retinopathy from stereoscopic color fundus photographs-an extension of the modified airlie house classification: ETDRS report number 10. Ophthalmology **98**(5, Supplement), pp. 786–806 (1991)
6. Flaxman, S.R., et al.: Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. Lancet Glob. Health **5**(12), e1221–e1234 (2017)
7. Haddock, L.J., Kim, D.Y., Mukai, S.: Simple, inexpensive technique for high-quality smartphone fundus photography in human and animal eyes. J. Ophthalmol. **2013**, 518479 (2013)
8. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 574–584 (2022)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
10. Holmberg, O.G., et al.: Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. Nat. Mach. Intell. **2**(11), 719–726 (2020)
11. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (2015)
12. Li, J., Chen, J., Tang, Y., Wang, C., Landman, B.A., Zhou, S.K.: Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. Med. Image Anal. **85**, 102762 (2023)
13. Panwar, N., Huang, P., Lee, J., Keane, P.A., Chuan, T.S., Richhariya, A., Teoh, S., Lim, T.H., Agrawal, R.: Fundus photography in the 21st century-a review of recent technological advances and their implications for worldwide healthcare. Telemedicine and e-Health **22**(3), 198–208 (2016)
14. Röhlig, M., Prakasam, R.K., Stüwe, J., Schmidt, C., Stachs, O., Schumann, H.: Enhanced grid-based visual analysis of retinal layer thickness with optical coherence tomography. Information **10**(9) (2019)
15. Ronneberger, O., Fischer, P., Brox, T.: U-NET: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241 (2015)
16. Schmidt-Erfurth, U., Waldstein, S.M., Deak, G.G., Kundi, M., Simader, C.: Pigment epithelial detachment followed by retinal cystoid degeneration leads to vision loss in treatment of neovascular age-related macular degeneration. Ophthalmology **122**(4), 822–832 (2015)
17. Varadarajan, A.V., et al.: Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. Nat. Commun. **11**(1), 130 (2020)
18. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
19. Wang, L.V., Wu, H.I.: Biomedical Optics: Principles and Imaging. John Wiley & Sons (2012)
20. Zhou, S.K., et al.: A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE **109**(5), 820–838 (2021)

21. Zhou, S.K., Rueckert, D., Fichtinger, G.: Handbook of Medical Image Computing and Computer Assisted Intervention. Academic Press (2019)
22. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1