Text as Illumination: Spatial Contrastive Retinex Learning for Language-guided Medical Image Segmentation

Jian Shi[1], Cheng Zhen[1], Pingping Zhang[2(✉)], Rui Xu[1], Yanan Lv[3], Yili Ma[3], Huan Bi[4], Haojie Li[5], and Huchuan Lu[2]

[1] School of Software Technology & DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian, China
[2] School of Future Technology, Dalian University of Technology, Dalian, China
[3] Central Hospital of Dalian University of Technology, Dalian, China
[4] Cancer Hospital of Dalian University of Technology, Dalian, China
[5] College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China
zhpp@dlut.edu.cn

Abstract. Language-guided Medical Image Segmentation (LMIS) has shown great potential to improve the delineation of anatomical structures and lesions by integrating clinical textual information. Existing methods generally rely on either implicit interaction between textual and visual features or auxiliary coarse-grained supervision for cross-modal alignment. However, these methods lack explicit and fine-grained constraints to ensure semantic consistency, causing a mismatch between language and the segmentation outputs. To address this issue, we propose Text-as-Illumination Retinex Network (TIRNet), a novel Retinex-inspired framework that treats text embeddings as semantic illumination for feature modulation, thereby improving semantic consistency in LMIS. TIRNet introduces two key blocks integrated at each decoder stage: (1) the Retinex-inspired Text Modulation Block (RTMB), which employs positive and negative illumination maps to enhance text-relevant foreground features and suppress background interference; and (2) the Consistent Detail Compensation Block (CDCB), which selectively recovers high-frequency details via a consistency-gated mechanism conditioned on illumination reliability. Furthermore, we propose a Multi-Scale Illumination Supervision Loss (MSIS-Loss), comprising a Region-Grounded Contrastive Loss (RGC-Loss) that enforces cross-modal similarity to be concentrated in text-relevant foreground regions and suppressed in background regions, and a Background Suppression Loss (BS-Loss) that provides pixel-level supervision for negative illumination maps, jointly ensuring a precise cross-modal alignment at each decoder stage. Extensive experiments on the MosMedData+ and QaTa-COV19 datasets demonstrate that TIRNet achieves state-of-the-art performance in LMIS. The code is available at: https://github.com/anaanaa/TIRNet.

Keywords: Language-guided Medical Image Segmentation · Retinex-inspired Feature Modulation · Region-Grounded Contrastive Learning.

1 Introduction

Medical image segmentation serves as a fundamental yet challenging task in computer-aided diagnosis, focusing on the precise delineation of key regions of interest, such as lesions and anatomical structures [1, 30]. Although purely vision-based deep learning methods have achieved considerable progress, their success relies on large-scale datasets with expert-level pixel-wise annotations. However, acquiring these annotations is expensive and labor-intensive, leading to a scarcity of high-quality labeled data. Notably, routine clinical reports generated in standard medical practice provide rich complementary semantic priors. Motivated by this observation, Language-guided Medical Image Segmentation (LMIS) [9, 25, 27] provides a promising way to leverage textual priors.

Existing LMIS methods generally follow two representative paradigms: implicit interaction methods that fuse textual and visual features via feature merging or attention [12, 6], and coarse supervision methods that introduce auxiliary constraints for cross-modal alignment [26, 28]. However, the former lacks explicit constraints to ensure semantic consistency across modalities, while the latter provides only coarse-grained supervision, limiting its ability to model fine spatial correspondence. For instance, LViT [12] integrates text embeddings via a pixel-level attention module, but the interaction between textual and visual features remains implicit and lacks explicit region-level alignment. TeViA [26] introduces explicit supervision by aligning foreground visual representations extracted from the segmentation head with text through a momentum-updated prototype. However, this coarse supervision is insufficient for modeling fine spatial correspondence between textual semantics and pixel-wise predictions. As shown in Fig. 1, both LViT and TeViA struggle to effectively focus on the foreground regions.

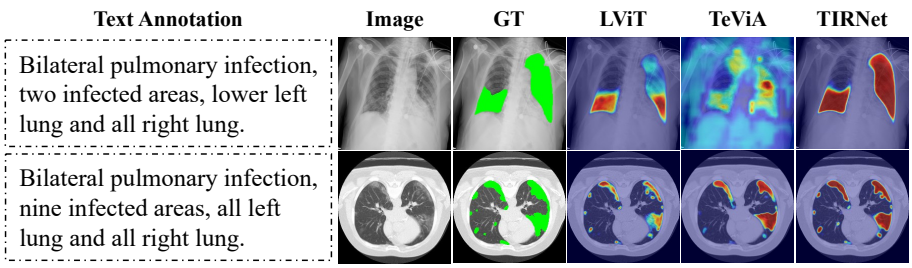

Fig. 1. Limitations of existing methods. Both implicit interaction (LViT [12]) and coarse supervision (TeViA [26]) methods fail to effectively focus on foregrounds, while our TIRNet achieves accurate localization.

To address the above issues, we present Text-as-Illumination Retinex Network (TIRNet), a novel framework for LMIS. Specifically, TIRNet integrates two key blocks at each decoder stage. The Retinex-inspired Text Modulation Block (RTMB) derives positive and negative illumination maps from cross-modal sim-

ilarities to explicitly amplify foregrounds and suppress backgrounds. The Consistent Detail Compensation Block (CDCB) selectively recovers high-frequency details via an illumination-guided consistency gate. Furthermore, we propose a Multi-Scale Illumination Supervision Loss (MSIS-Loss) to enforce explicit alignment by employing a Region-Grounded Contrastive Loss (RGC-Loss) to maximize foreground-background margins and a pixel-level Background Suppression Loss (BS-Loss) to explicitly suppress background responses. Extensive experiments on the MosMedData+ and QaTa-COV19 datasets demonstrate that TIRNet achieves state-of-the-art performance in LMIS.

Our main contributions are summarized as follows:

- We propose TIRNet for LMIS, which draws inspiration from Retinex theory by treating text embeddings as semantic illumination, enabling explicit foreground amplification and background suppression in decoder features.
- We design the RTMB to derive complementary illumination maps for feature modulation, and the CDCB to selectively recover high-frequency details via an illumination-guided consistency gate.
- We introduce a MSIS-Loss, combining RGC-Loss and BS-Loss, to enforce cross-modal semantic alignment.
- Experiments on the MosMedData+ and QaTa-COV19 datasets demonstrate that TIRNet achieves superior performance over existing methods.

2 Method

As shown in Fig. 2, TIRNet adopts a U-shaped encoder-decoder architecture. A CNN-based visual encoder extracts multi-scale visual features $\{\mathbf{F}_{\text{enc}}^{s}\}_{s=1}^{4}$, while a frozen CLIP [16] text encoder extracts token-level text features $\mathbf{E}_{\text{text}} \in \mathbb{R}^{L\times D}$, which are refined by a 1D convolution and mean-pooled into a global semantic embedding $\mathbf{e} \in \mathbb{R}^{D}$. At each decoder stage, the RTMB uses $\mathbf{e}$ as semantic illumination to modulate features, which are then fed to the CDCB to recover high-frequency details from the corresponding encoder stage. During training, similarity maps from each RTMB are supervised by the proposed MSIS-Loss.

2.1 Retinex-inspired Text Modulation Block

Retinex theory [11] models an image as $\mathbf{S} = \mathbf{R} \odot \mathbf{L}$, where the reflectance $\mathbf{R}$ represents the intrinsic structural content and the illumination $\mathbf{L}$ modulates spatial visibility. Recent Retinex-inspired methods have shown that Retinex-based decomposition and illumination-aware modeling can effectively guide feature enhancement, ranging from low-level image restoration to multispectral object detection [2, 13]. Inspired by this illumination-guided representation mechanism, we propose the Retinex-inspired Text Modulation Block (RTMB), which adapts Retinex decomposition to cross-modal feature modulation. In RTMB, the visual decoder feature $\mathbf{F}_{\text{dec}}$ acts as the reflectance $\mathbf{R}$, while the text embedding serves as a semantic illumination field $\mathbf{L}$. This text-conditioned modulation enhances

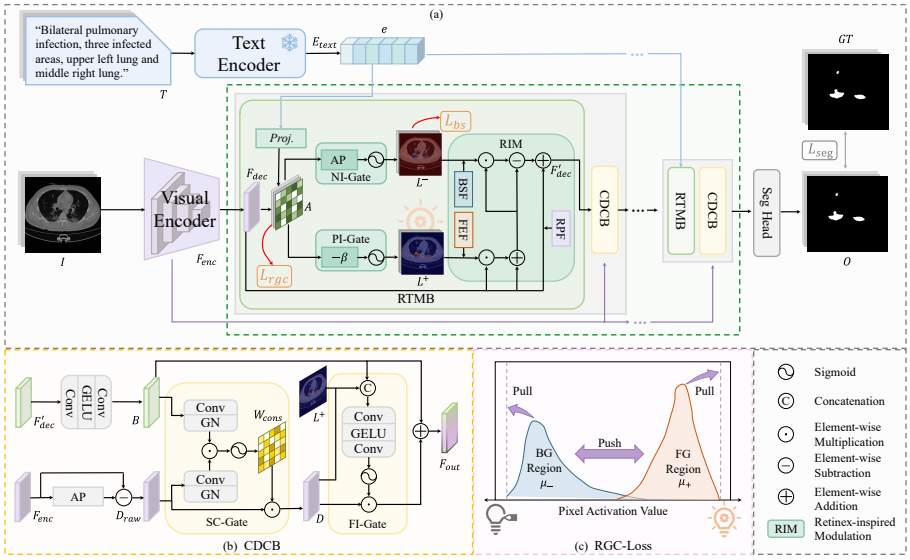

Fig. 2. Overview of the TIRNet. It integrates RTMB and CDCB into each decoder stage. The RGC-Loss maximizes cross-modal similarity in text-relevant foregrounds and suppresses background activations to enlarge the foreground-background margin.

target regions that are semantically aligned with the text embedding while suppressing irrelevant background responses, producing the modulated decoder feature $\mathbf{F}'_{\text{dec}}$ with an improved foreground-background separability.

At the s-th decoder stage, the RTMB takes the decoder feature $\mathbf{F}^{(s)}_{\text{dec}} \in \mathbb{R}^{C_s \times H_s \times W_s}$ and the global text embedding $\mathbf{e}$ as inputs. For notational simplicity, we omit the superscript (s) in the following equations. Given the decoder feature $\mathbf{F}_{\text{dec}} \in \mathbb{R}^{C \times H \times W}$ and $\mathbf{e}$, we firstly project $\mathbf{e}$ into the image feature space via a learnable linear projection $\mathbf{W}_p \in \mathbb{R}^{C \times D}$:

$$\mathbf{t} = \mathbf{W}_p \mathbf{e}, \quad \mathbf{t} \in \mathbb{R}^C. \tag{1}$$

Then we compute the text-guided relevance score at each spatial location:

$$A_{h,w} = \left\langle \frac{\mathbf{f}_{h,w}}{\|\mathbf{f}_{h,w}\|_2}, \frac{\mathbf{t}}{\|\mathbf{t}\|_2} \right\rangle, \tag{2}$$

where $\mathbf{f}_{h,w} \in \mathbb{R}^C$ is the decoder feature vector at spatial location (h, w), and $\langle \cdot, \cdot \rangle$ is the inner product. The map $\mathbf{A} \in \mathbb{R}^{1 \times H \times W}$ aggregates channel-wise normalized visual-text correlations into a semantic relevance score at each spatial location.

Based on this semantic relevance map $\mathbf{A}$, we derive two complementary illumination maps through two parallel gating pathways. The Positive Illumination Gate (PI-Gate) applies a 3×3 Average-Pooling (AP) followed by a sigmoid activation function σ to produce a spatially coherent foreground illumination:

$$\mathbf{L}^+ = \sigma\big(\text{AP}(\mathbf{A})\big) \in [0, 1]^{1 \times H \times W}. \tag{3}$$

Conversely, the Negative Illumination Gate (NI-Gate) inverts the similarity with a learnable sharpness β to capture text-irrelevant backgrounds:

$$\mathbf{L}^{-} = \sigma\big(-\beta\mathbf{A}\big) \in [0,1]^{1\times H\times W}, \tag{4}$$

where β is strictly positive. The negation ensures that low-relevance locations receive high activations in $\mathbf{L}^{-}$, while the learnable β adaptively controls the foreground–background transition sharpness.

Retinex-inspired Modulation. The decoder feature is modulated as:

$$\mathbf{F}'_{\text{dec}} = \mathbf{F}_{\text{dec}} \odot \big(\mathbf{1} + \rho\mathbf{L}^{+}\big) \odot \big(\mathbf{1} - \kappa\mathbf{L}^{-}\big) + \gamma\mathbf{F}_{\text{dec}}, \tag{5}$$

where ρ, κ, and γ denote the learnable Foreground Enhancement Factor (FEF), Background Suppression Factor (BSF), and Residual Preservation Factor (RPF).

2.2 Consistent Detail Compensation Block

Although the RTMB effectively disentangles the foreground and background, the illuminated regions still lack high-frequency details essential for precise boundary delineation due to feature downsampling. To address this issue, the Consistent Detail Compensation Block (CDCB) selectively recovers fine-grained structural information from the encoder under the guidance of illumination reliability.

Given the refined feature $\mathbf{F}'_{\text{dec}}$, we derive a base representation $\mathbf{B} = f_{\text{base}}(\mathbf{F}'_{\text{dec}})$ via a 3×3 conv–GELU–3×3 conv sequence. To isolate high-frequency details from the encoder feature $\mathbf{F}_{\text{enc}}$, we compute the residual $\mathbf{D}_{\text{raw}}$ between the original feature and its low-pass filtered counterpart, which extracts edge and texture information complementary to the decoder stream [24]:

$$\mathbf{D}_{\text{raw}} = \mathbf{F}_{\text{enc}} - \text{AP}(\mathbf{F}_{\text{enc}}). \tag{6}$$

Then, we introduce a Semantic Consistency Gate (SC-Gate) that selectively integrates encoder details semantically aligned with the decoder representation:

$$\mathbf{W}_{\text{cons}} = \sigma\Big(\delta \sum_{c=1}^{C} \bar{\mathbf{q}}_c \odot \bar{\mathbf{k}}_c\Big), \tag{7}$$

where $\bar{\mathbf{q}} = \text{GN}(g_q(\mathbf{B}))$ and $\bar{\mathbf{k}} = \text{GN}(g_k(\mathbf{D}_{\text{raw}}))$. g_q and g_k are 1×1 convolution layers. $\text{GN}(\cdot)$ denotes Group Normalization [22]. δ is a learnable parameter. Consequently, the modulated detail is obtained as $\mathbf{D} = \mathbf{W}_{\text{cons}} \odot \mathbf{D}_{\text{raw}}$.

Finally, a Fusion Illumination Gate (FI-Gate) controls detail injection using $\mathbf{L}^{+}$ as a cross-modal reliable signal:

$$\mathbf{G} = \sigma\big(f_{\text{gate}}([\mathbf{B};\, \mathbf{D};\, \mathbf{L}^{+}])\big), \tag{8}$$

where $[\cdot;\cdot]$ is a channel-wise concatenation. f_{gate} is a 3×3 conv–GELU–1×1 conv sequence. Guided by $\mathbf{L}^{+}$, the gate selectively preserves fine structural details within the target foreground while suppressing task-irrelevant background interference. The output aggregates the base representation with the gated detail as $\mathbf{F}_{\text{out}} = \mathbf{B} + \mathbf{G} \odot \mathbf{D}$.

2.3 Multi-Scale Illumination Supervision Loss

To enforce cross-modal alignment across all decoder stages, we propose a Multi-Scale Illumination Supervision Loss (MSIS-Loss) over the stage-wise illumination cues $\{\mathbf{A}^{(s)}, \mathbf{L}^{-(s)}\}_{s=1}^{S}$. The MSIS-Loss consists of two complementary terms: a Region-Grounded Contrastive Loss (RGC-Loss) and a Background Suppression Loss (BS-Loss). Unlike conventional instance-level [4, 18] or pixel-level [19, 21] contrastive losses, the RGC-Loss aggregates cross-modal similarities within anatomically meaningful foreground and background regions. This spatially structured supervision effectively bridges the granularity gap between global text semantics and local pixel predictions. Specifically, we downsample the ground-truth to $\mathbf{G}^{(s)}$ at scale s, partitioning the spatial domain into a positive region $\Omega_{+}^{(s)} = \{(i,j) \mid \mathbf{G}_{i,j}^{(s)} = 1\}$ and a negative region $\Omega_{-}^{(s)} = \{(i,j) \mid \mathbf{G}_{i,j}^{(s)} = 0\}$. We then compute the exponentiated similarity $\mathbf{P}^{(s)} = \exp(\mathbf{A}^{(s)})$ and aggregate the mean activation intensity within each region:

$$\mu_{+}^{(s)} = \frac{\sum_{i,j} \mathbf{G}_{i,j}^{(s)} \mathbf{P}_{i,j}^{(s)}}{|\Omega_{+}^{(s)}|}, \quad \mu_{-}^{(s)} = \frac{\sum_{i,j} (1 - \mathbf{G}_{i,j}^{(s)}) \mathbf{P}_{i,j}^{(s)}}{|\Omega_{-}^{(s)}|}. \tag{9}$$

The contrastive loss maximizes foreground-background margins:

$$\mathcal{L}_{\text{rgc}}^{(s)} = -\log \frac{\mu_{+}^{(s)}}{\mu_{+}^{(s)} + \mu_{-}^{(s)}}. \tag{10}$$

Furthermore, we formulate a BS-Loss that explicitly supervises $\mathbf{L}^{-(s)}$ with the inverse ground-truth:

$$\mathcal{L}_{\text{bs}}^{(s)} = \text{BCE}\big(\mathbf{L}^{-(s)},\ 1 - \mathbf{G}^{(s)}\big). \tag{11}$$

This loss ensures that the negative illumination correctly captures the non-target background.

We aggregate the stage-wise losses as:

$$\mathcal{L}_{\text{rgc}} = \frac{1}{S} \sum_{s=1}^{S} \mathcal{L}_{\text{rgc}}^{(s)}, \quad \mathcal{L}_{\text{bs}} = \frac{1}{S} \sum_{s=1}^{S} \mathcal{L}_{\text{bs}}^{(s)}. \tag{12}$$

The overall training loss is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{rgc}} \mathcal{L}_{\text{rgc}} + \lambda_{\text{bs}} \mathcal{L}_{\text{bs}}, \tag{13}$$

where $\mathcal{L}_{\text{seg}}$ is the segmentation loss, combining a Dice loss and a binary cross-entropy loss. The loss weights are experimentally set to λ_{rgc}=1.0 and λ_{bs}=0.5.

3 Experiments

Datasets and Metrics. We evaluate TIRNet on two public datasets: MosMedData+ [15] with 2,729 CT scans and QaTa-COV19 [5] with 9,258 chest X-rays.

Table 1. Quantitative comparison on MosMedData+ and QaTa-COV19 datasets. The dashed line separates uni-modal methods (top) from multi-modal methods (bottom).

Method	Params (M)	Flops (G)	MosMedData+				QaTa-COV19			
			m-Dice	m-IoU	g-Dice	g-IoU	m-Dice	m-IoU	g-Dice	g-IoU
U-Net [17]	14.80	50.3	62.52	49.24	72.38	56.71	78.46	68.59	86.66	76.47
U-Net++ [29]	74.50	94.6	69.74	57.39	80.19	66.93	79.55	70.21	87.76	78.19
nnUNet [10]	19.10	412.7	73.18	61.02	80.46	67.31	79.13	69.69	87.37	77.57
Swin-UNet [3]	82.30	67.3	65.94	52.35	77.85	63.73	77.29	67.55	86.44	76.12
GLoRIA [8]	45.60	60.8	70.11	56.75	79.31	65.71	79.51	70.36	88.25	78.97
LAVT [25]	118.60	83.8	73.66	60.74	80.66	67.58	78.87	69.35	87.56	77.88
TGANet [20]	19.80	41.9	70.73	57.90	80.06	66.75	80.41	71.24	88.45	79.29
LViT [12]	29.70	54.1	74.52	61.15	76.64	62.12	83.40	74.78	90.63	82.86
LGA [7]	8.24	381.1	74.53	61.10	80.55	67.43	83.46	74.50	89.85	81.56
CMIRNet [23]	239.80	134.5	73.88	60.25	79.64	66.17	79.63	69.87	87.45	77.71
AT [27]	44.00	22.4	72.41	58.63	78.53	64.65	79.57	69.48	87.19	77.30
TeViA [26]	146.80	11.2	72.32	59.06	79.47	65.94	84.12	75.69	90.96	83.41
TIRNet (Ours)	41.10	24.1	**75.41**	**62.77**	**81.38**	**68.61**	**84.77**	**76.47**	**91.23**	**83.88**

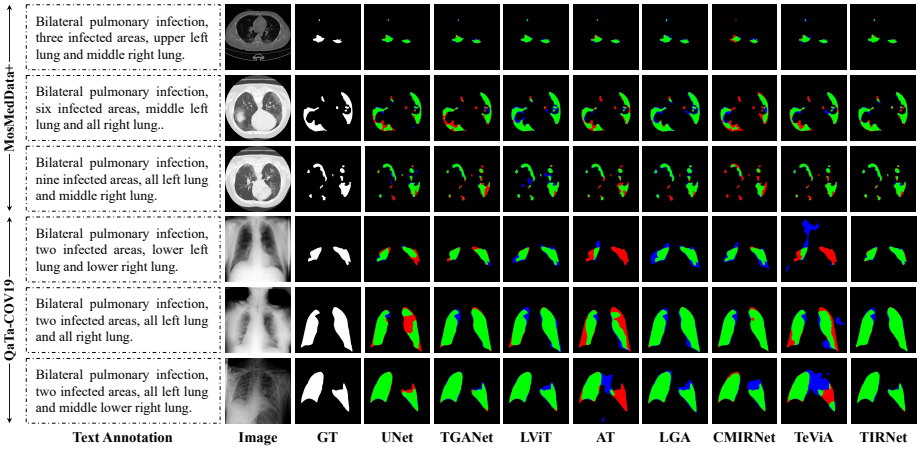

Fig. 3. Visual comparison of segmentation results of different methods (Green: True Positive, Red: False Negative, Blue: False Positive).

Following the data split protocol of LViT [12], we conduct extensive experiments on both datasets. To quantitatively assess performance, we report both sample-level averages (m-Dice, m-IoU), and global metrics (g-Dice, g-IoU) aggregated over the entire dataset.

Implementation Details. TIRNet was implemented in PyTorch and trained on a single NVIDIA Tesla V100 GPU. Input images were resized to 224×224 with a batch size of 32. We used the AdamW [14] optimizer with an initial learning rate

Table 2. Ablation study results of key blocks and the MSIS-Loss.

RTMB	CDCB	MSIS-Loss	m-Dice	m-IoU	g-Dice	g-IoU
✗	✗	✗	78.44	68.67	86.90	76.83
✓	✗	✗	83.13	74.45	90.48	82.61
✓	✗	✓	83.65	75.08	90.72	83.02
✗	✓	✗	79.97	70.53	87.71	78.11
✓	✓	✗	84.36	75.96	91.13	83.70
✓	✓	✓	**84.77**	**76.47**	**91.23**	**83.88**

Table 3. Sensitivity of hyperparameters λ_{rgc} and λ_{bs}.

$\lambda_{\text{rgc}}/\lambda_{\text{bs}}$	m-Dice	m-IoU	g-Dice	g-IoU
0/0	84.36	75.96	91.13	83.70
1/0	84.54	76.11	91.16	83.75
0/0.5	84.20	75.81	91.22	83.85
0.5/0.5	84.35	75.96	91.10	83.65
1/1	84.39	75.94	91.04	83.55
1/0.5	**84.77**	**76.47**	**91.23**	**83.88**

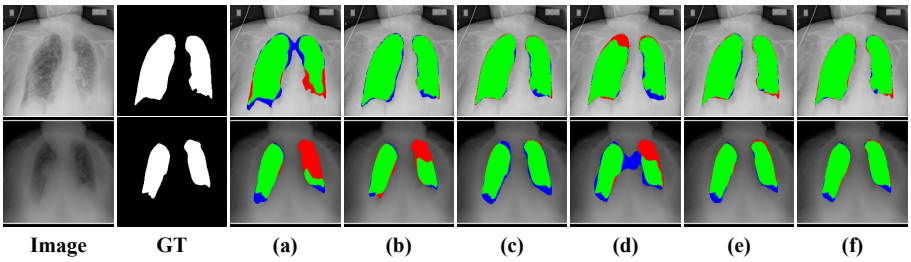

Fig. 4. Qualitative visualization of the component ablation study. Panels (a)–(f) correspond to the methods in the 1st–6th rows of Table 2, respectively.

of 3×10^{-4}, decayed to 1×10^{-6} via cosine annealing. Data augmentation included random flipping, rotation, scaling, Gaussian blur, and photometric distortions.

Comparative Experiments. We compare TIRNet with existing methods on MosMedData+ and QaTa-COV19, with quantitative results reported in Table 1. On MosMedData+, TIRNet improves over LViT [12] by 0.89% in m-Dice and 1.62% in m-IoU, and surpasses TeViA [26] by 1.91% in g-Dice and 2.67% in g-IoU. On QaTa-COV19, TIRNet also achieves the best overall performance, exceeding LViT [12] by 1.37% in m-Dice and 1.69% in m-IoU, and improving over TeViA [26] by 0.27% in g-Dice and 0.47% in g-IoU. The visual comparisons in Fig. 3 further demonstrate that TIRNet produces more accurate and robust segmentation results across diverse challenging cases.

Ablation Studies. To assess each component in TIRNet, we conducted ablation studies on the QaTa-COV19 dataset, with results summarized in Table 2. Starting from the baseline, including the RTMB yields a 4.69% increase in m-Dice, showing that semantic illumination effectively enhances features. The CDCB alone contributes a 1.53% gain in m-Dice, confirming its role in recovering high-frequency structural details. Adding the MSIS-Loss to the RTMB further improves m-Dice by 0.52%, highlighting the need for explicit cross-modal alignment. Overall, the full TIRNet achieves the best performance, outperforming the baseline by 6.33% in m-Dice. Fig. 4 shows qualitative results for the component ablations in Table 2. These results indicate that each component in TIRNet contributes to more accurate and semantically consistent segmentation.

In addition, we investigate the sensitivity of the hyperparameters λ_{rgc} and λ_{bs}. We evaluate different combinations of these hyperparameters, as reported

in Table 3. The results indicate that the optimal performance is achieved when λ_{rgc} and λ_{bs} are set to 1 and 0.5, respectively.

4 Conclusion

In this paper, we propose TIRNet, which maps Retinex theory to LMIS. Treating clinical text embeddings as semantic illumination, it separates foreground from similar backgrounds and reduces semantic mismatch in cross-modal interaction. TIRNet consists of two key blocks, the RTMB and the CDCB. The RTMB uses complementary positive and negative illumination maps to enhance features and suppress interference. The CDCB restores high-frequency details via an illumination-guided consistency gate. Furthermore, the MSIS-Loss aligns modalities by enlarging the foreground-background margin and matching negative illumination to the background distribution. Experiments on the MosMedData+ and QaTa-COV19 datasets demonstrate that TIRNet achieves state-of-the-art performance, validating the effectiveness of mapping the Retinex illumination formulation to the cross-modal segmentation setting.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (No.62576069), Liaoning Province Artificial Intelligence Innovation Development Plan (No.2023JH26/10200016), Dalian Science and Technology Innovation Fund (No.2023JJ11CG001), Fundamental Research Funds for the Central Universities (No.DUT25YG262, DUT24YG135) and Natural Science Foundation of Liaoning Province (No.2025-MS-025).

Disclosure of Interests. The authors declare no competing interests.

References

1. Bozorgpour, A., Kolahi, S.G., Azad, R., Hacihaliloglu, I., Merhof, D.: CENet: Context enhancement network for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 120–129 (2025)
2. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage Retinex-based Transformer for low-light image enhancement. In: IEEE/CVF International Conference on Computer Vision. pp. 12504–12513 (2023)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-Unet: Unet-like pure Transformer for medical image segmentation. In: European Conference on Computer Vision Workshops. pp. 205–218 (2022)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607 (2020)
5. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for Covid-19 detection using chest X-Ray images. In: IEEE International Conference on Image Processing. pp. 2306–2310 (2022)

6. Guo, Y., Zeng, X., Zeng, P., Fei, Y., Wen, L., Zhou, J., Wang, Y.: Common vision-language attention for text-guided medical image segmentation of pneumonia. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 192–201 (2024)
7. Hu, J., Li, Y., Sun, H., Song, Y., Zhang, C., Lin, L., Chen, Y.W.: LGA: A language guide adapter for advancing the SAM model's capabilities in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 610–620 (2024)
8. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021)
9. Huang, X., Li, H., Cao, M., Chen, L., You, C., An, D.: Cross-modal conditioned reconstruction for language-guided medical image segmentation. IEEE Transactions on Medical Imaging **44**(4), 1821–1835 (2024)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
11. Land, E.H.: The Retinex theory of color vision. Scientific American **237**(6), 108–129 (1977)
12. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: LViT: Language meets Vision Transformer in medical image segmentation. IEEE Transactions on Medical Imaging **43**(1), 96–107 (2023)
13. Liu, Z., Geng, K., Cheng, X., Wang, Z., Yin, G., Sun, Y., Ma, T.: RetinexDet: Enhancing multispectral object detection via retinex state space duality and wavelet-based frequency adaptive fusion. IEEE Transactions on Industrial Informatics **21**(12), 9619–9630 (2025)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
15. Morozov, S.P., Andreychenko, A.E., Blokhin, I.A., Gelezhe, P.B., Gonchar, A.P., Nikolaev, A.E., Pavlov, N.A., Chernina, V.Y., Gombolevskiy, V.A.: MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic. Digital Diagnostics **1**(1), 49–59 (2020)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241 (2015)
18. Rui, S., Chen, L., Tang, Z., Wang, L., Liu, M., Zhang, S., Wang, X.: Multi-modal vision pre-training for medical image analysis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5164–5174 (2025)
19. Shi, J., Sun, B., Ye, X., Wang, Z., Luo, X., Liu, J., Gao, H., Li, H.: Semantic decomposition network with contrastive and structural constraints for dental plaque segmentation. IEEE Transactions on Medical Imaging **42**(4), 935–946 (2022)
20. Tomar, N.K., Jha, D., Bagci, U., Ali, S.: TGANet: Text-guided attention for improved polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 151–160 (2022)

21. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L.: Exploring cross-image pixel contrast for semantic segmentation. In: IEEE/CVF International Conference on Computer Vision. pp. 7303–7313 (2021)
22. Wu, Y., He, K.: Group normalization. In: European Conference on Computer Vision. pp. 3–19 (2018)
23. Xu, M., Xiao, T., Liu, Y., Tang, H., Hu, Y., Nie, L.: CMIRNet: Cross-modal interactive reasoning network for referring image segmentation. IEEE Transactions on Circuits and Systems for Video Technology **35**(4), 3234–3249 (2024)
24. Yan, T., Wan, Z., Zhang, P., Cheng, G., Lu, H.: TransY-Net: Learning fully transformer networks for change detection of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–12 (2023)
25. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: LAVT: Language-aware Vision Transformer for referring image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022)
26. Zeng, Q., Luo, H., Lu, Z., Xie, Y., Wang, Z., Zhang, Y., Xia, Y.: Harnessing text insights with visual alignment for medical image segmentation. IEEE Transactions on Medical Imaging **45**(2), 477–489 (2026)
27. Zhong, Y., Xu, M., Liang, K., Chen, K., Wu, M.: Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest X-ray images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 724–733 (2023)
28. Zhou, X., Song, Q., Nie, J., Feng, Y., Liu, H., Liang, F., Chen, L., Xie, J.: Hybrid cross-modality fusion network for medical image segmentation with contrastive learning. Engineering Applications of Artificial Intelligence **144**, 110073 (2025)
29. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. In: International Workshop on Deep Learning in Medical Image Analysis. pp. 3–11 (2018)
30. Zhu, W., Chen, X., Qiu, P., Farazi, M., Sotiras, A., Razi, A., Wang, Y.: SelfReg-UNet: Self-regularized UNet for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 601–611 (2024)