
Complex Wavelet Mutual Information Loss: A Multi-Scale Loss Function for Semantic Segmentation

Renhao Lu¹

Abstract

Recent advancements in deep neural networks have significantly enhanced the performance of semantic segmentation. However, class imbalance and instance imbalance remain persistent challenges, where smaller instances and thin boundaries are often overshadowed by larger structures. To address the multiscale nature of segmented objects, various models have incorporated mechanisms such as spatial attention and feature pyramid networks. Despite these advancements, most loss functions are still primarily pixel-wise, while regional and boundary-focused loss functions often incur high computational costs or are restricted to small-scale regions. To address this limitation, we propose the complex wavelet mutual information (CWMI) loss, a novel loss function that leverages mutual information from subband images decomposed by a complex steerable pyramid. The complex steerable pyramid captures features across multiple orientations and preserves structural similarity across scales. Meanwhile, mutual information is well-suited to capturing high-dimensional directional features and offers greater noise robustness. Extensive experiments on diverse segmentation datasets demonstrate that CWMI loss achieves significant improvements in both pixel-wise accuracy and topological metrics compared to state-of-the-art methods, while introducing minimal computational overhead. Our code is available at <https://github.com/lurenhaothu/CWMI>

1. Introduction

Semantic segmentation, the process of partitioning an image into regions associated with semantic labels, plays a crucial role in applications ranging from autonomous driving to biomedical imaging. Despite significant progress driven by deep neural networks such as U-Net (Ronneberger et al., 2015) and fully convolutional networks (Long et al., 2015), challenges persist. Class imbalance, where dominant classes overshadow smaller ones, and instance imbalance, where small-scale structures are frequently ignored, remain major obstacles (Jiang et al., 2024; Kofler et al., 2023). Addressing these imbalances requires not only pixel-wise accuracy, but also the preservation of structural similarity, a property vital for ensuring spatial coherence and topological integrity in segmented outputs (Wang et al., 2004).

Although the advancements in feature extraction, such as feature pyramid networks (Lin et al., 2017) and attention mechanisms (Woo et al., 2018; Chen et al., 2021; Islam et al., 2020), have enabled models to capture multiscale contextual information, most loss functions still focus on pixel-wise optimization (Azad et al., 2023). Region-based loss functions, such as Dice loss (Milletari et al., 2016) and Tversky loss (Salehi et al., 2017), are not subject to class imbalance, but smaller instances within the same class are still easy to be overshadowed, which can be critical for preserving regional and boundary details. Zhao et al. developed Regional Mutual Information (RMI) loss, which captures statistical relationships over regions. However, RMI is constrained within a relatively small region (3×3 pixels) to avoid high computational overhead (Zhao et al., 2019). Thus, a loss function that balances computational efficiency with the ability to model structural and regional dependencies at larger scales is highly desirable.

Addressing these challenges from a frequency-domain perspective provides an innovative pathway. Patterns of varying scales in images are inherently tied to their frequency components: large-scale structures correspond to low frequencies, while finer details correspond to high frequencies. Wavelet transforms are uniquely suited for this multiscale decomposition, as they preserve both spatial and frequency information (Mallat, 1989). Among these, the steerable pyramid, proposed by (Simoncelli et al., 1992), leverages

¹Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, USA. Correspondence to: Renhao Lu <rl839@cornell.edu>.

steerable filters for redundant wavelet decomposition, enabling multiscale and multi-orientation feature extraction. Its extension, the complex steerable pyramid, further enhances this framework by using complex numbers to explicitly represent local phase information, allowing for robust extraction of structural details across scales and orientations (Portilla & Simoncelli, 2000). These properties make it a powerful tool for segmentation tasks where structural similarity is paramount.

In this paper, we introduce Complex Wavelet Mutual Information (CWMI) loss, a novel loss function that leverages the complex steerable pyramid for efficient multiscale structural information extraction. By combining the robust multiscale decomposition capabilities of the complex steerable pyramid with the statistical power of mutual information, CWMI loss explicitly incorporates local phase, orientation, and structural features into the loss calculation. This approach ensures structural coherence and boundary preservation while maintaining computational efficiency, making it well-suited for segmentation tasks with significant class and instance imbalances.

Our contributions are summarized as follows:

- We propose CWMI loss, which can maximize the mutual information in the domain of complex steerable pyramid decompositions. Such a strategy can enhance multiscale structural features for semantic segmentation, especially for tasks with significant class and instance imbalances.
- We demonstrated the superiority of CWMI with extensive experiments on four public segmentation datasets: SNEMI3D (neurite segmentation in electron microscopy slices), GlaS (gland segmentation in H&E slices), DRIVE (retinal vessel segmentation in fundus images), and MASS ROAD (road segmentation from aerial imagery). Compared with 11 state-of-art (SOTA) loss functions, CWMI showed better performance on both pixel-wise metrics and topological metrics, while introducing minimal computational overhead.

2. Related Work

Semantic segmentation has seen tremendous advancements through deep learning architectures, with U-Net and its variants becoming a cornerstone of this field. The U-Net model, proposed by (Ronneberger et al., 2015), utilizes a symmetric encoder-decoder architecture with skip connections to preserve spatial details while capturing global context. Its success has inspired numerous adaptations in its convolutional blocks (Huang et al., 2019; Diakogiannis et al., 2020) and skip connections (Zhou et al., 2018; Chen et al., 2021). Inspired by the transformer model, the attention mechanism has also been incorporated into the U-Net structure, which

has shown significant performance enhancement, as in Attention U-Net (Islam et al., 2020), TransUNet (Chen et al., 2021), Vision Mamba UNet (Ruan et al., 2024), etc. In this study, we compare our proposed CWMI loss using U-Net and Attention U-Net to test the generalization and superiority of CWMI, while the general idea is adaptable to other architectures as well.

2.1. Loss Functions for Semantic Segmentation

While the architectural advancements in segmentation models have been significant, the performance of these networks is highly influenced by the design of the loss functions. Pixel-wise cross entropy loss (CE) minimizes the log likelihood of the prediction error but is significantly prone to class imbalance. To address this issue, class balanced cross entropy (BCE), proposed by (Long et al., 2015), employs higher weights for classes with fewer pixels. Focal loss assigns higher weights to misclassified pixels with high probabilities (Ross & Dollár, 2017). Region-based losses, including Dice loss (Milletari et al., 2016), Tversky loss (Salehi et al., 2017), and Jaccard loss (Rahman & Wang, 2016), inherently handle class imbalance but fail to address instance imbalance within the same class, such as thin boundaries and small objects.

To tackle instance imbalance, weighted loss functions have been proposed. In the original U-Net paper (Ronneberger et al., 2015), the weighted cross entropy (WCE) was introduced, employing a distance-based weight map to emphasize thin boundaries between objects. However, WCE assigns weights only to boundary pixels, neglecting object pixels. The adaptive boundary weighted (ABW) loss (Liu et al., 2022) extends this approach by applying distance-based weights to both boundary and object pixels, while the Skea-Topo loss further improved the weight map based on boundary and object skeletons (Liu et al., 2024). Despite their contributions, weighted losses suffer from two major limitations: (1) the weight maps are precomputed and fixed, failing to adapt to errors during training, and (2) they often generate thicker boundaries, which preserve topology but compromise metrics like Dice score and mIoU, as observed in our qualitative results.

Several methods address instance imbalance dynamically during training, but at the cost of computational efficiency. Topology-based approaches, such as persistent homology methods (Stucki et al., 2023; Oner et al., 2023), describe image topologies and identify critical pixels but are computationally expensive, with cubic complexity to image size. The cDice loss (Shit et al., 2021) employs a soft skeletonization algorithm to detect topological errors, primarily focusing on thin-boundary objects like retinal blood vessels. Similarly, Boundary Loss (Kervadec et al., 2019) and Hausdorff Distance Loss (Karimi & Salcudean, 2019) refine boundaries but incur significant computational overhead.

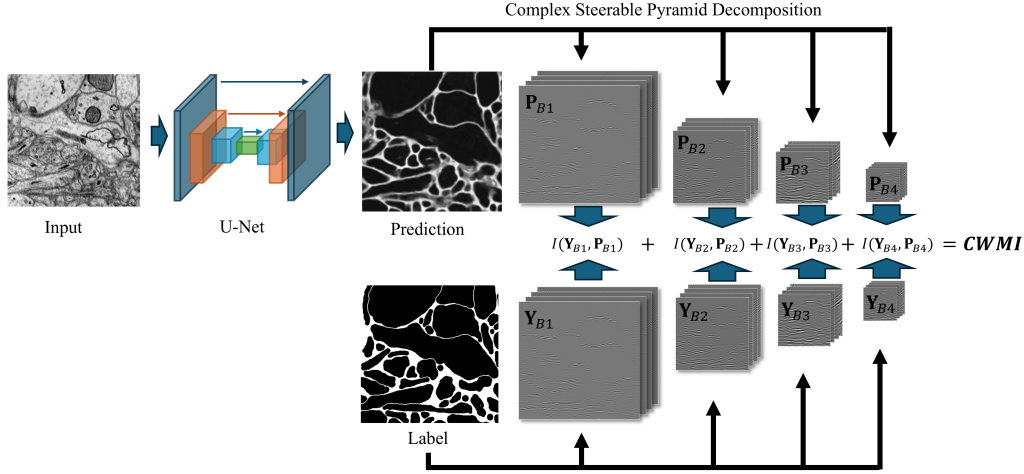


Figure 1. Illustration of the proposed Complex Wavelet Mutual Information (CWMI) Loss. The prediction and label images are decomposed using a complex steerable pyramid, which generates subbands at different scales and orientations. Mutual information is calculated for each corresponding pair of subbands, and the CWMI is computed as the sum of these mutual information values. $\mathbf{Y}_{B_n}, \mathbf{P}_{B_n}$: complex steerable decomposition of label and prediction image at level n ; $I(\mathbf{Y}_{B_n}, \mathbf{P}_{B_n})$: mutual information between \mathbf{Y}_{B_n} and \mathbf{P}_{B_n}

Region Mutual Information (RMI) loss (Zhao et al., 2019) captures pixel interdependencies over regions, but struggles with scalability for large-scale regional analysis due to the high computation cost. These losses either prioritize small regions at the expense of global accuracy or require extensive computational resources, necessitating more efficient and balanced approaches.

2.2. Wavelet-Based Loss Functions

Wavelet-based metrics were first introduced as Complex Wavelet Structural Similarity (CW-SSIM) (Sampat et al., 2009), known for their robustness to small translations and rotations. In the deep learning era, wavelet-based methods have been employed in loss functions, leveraging their ability to analyze multiscale and multiresolution features. These methods have shown promise in tasks like sketch-to-image translation (Kim & Cho, 2023), image super-resolution (Korkmaz & Tekalp, 2024), image dehazing (Yang et al., 2020), and material analysis (Prantl et al., 2022). However, to the best of our knowledge, wavelet-based loss functions have yet to be explored in semantic segmentation.

Existing wavelet-based loss functions typically rely on L_1 (Zhu et al., 2021; Korkmaz & Tekalp, 2024; Prantl et al., 2022) or L_2 distances (Kim & Cho, 2023), or structural similarity (SSIM) (Yang et al., 2020) in the decomposed domain. While effective, these methods are less suited for handling high-dimensional data with complex directional features, as in wavelet transforms, and may be vulnerable to noise. The proposed CWMI loss leverages mutual information between wavelet-based subband images, effectively capturing multiscale dependencies. As demonstrated in later ablation tests, CWMI outperforms traditional metrics like L_1 , L_2 ,

and SSIM, offering superior segmentation performance and robustness.

3. Methods

To compute the proposed Complex Wavelet Mutual Information (CWMI) loss, the prediction and ground truth label matrices are first decomposed into subbands using the complex steerable pyramid. Mutual information (MI) is then calculated for each subband. The CWMI loss is defined as the sum of the MI values across all subbands, as illustrated in Figure 1.

3.1. Complex Wavelet Decomposition

Wavelet Transform Wavelet transforms are widely used for multiscale analysis, enabling decomposition of an image into frequency subbands while preserving spatial information. Unlike traditional Fourier transforms, which analyze global frequency components, wavelet transforms provide a localized frequency representation, making them well-suited for tasks involving spatially-varying structures such as semantic segmentation. For an image $I(x, y)$, the wavelet transform is defined as:

$$W_\psi(s, t) = \int \int I(x, y) \psi_{s,t}(x, y) dx dy \quad (1)$$

where $\psi_{s,t}(x, y)$ is a scaled and translated version of the mother wavelet function $\psi(x, y)$, with s controlling the scale and t the translation.

For discrete signals, the Discrete Wavelet Transform (DWT) decomposes an image into progressively lower resolution subbands using filter banks. However, traditional wavelet

decompositions suffer from limited orientation selectivity, capturing only fixed horizontal, vertical, and diagonal directions.

Steerable Pyramid To overcome these limitations, steerable pyramid extends the wavelet framework by introducing orientation-sensitive band-pass filters, significantly enhancing orientation selectivity. Unlike DWT, which provides a non-redundant representation, the steerable pyramid offers a flexible, redundant image representation, facilitating improved multiscale analysis. This decomposition is achieved through the iterative application of steerable band-pass filters followed by downsampling. In the frequency domain, the band-pass filter for the k_{th} orientation is expressed in polar coordinates (r, θ) as:

$$B_k(r, \theta) = H(r)G_k(\theta), \quad k \in [1, K], \quad (2)$$

where $H(r)$ and $G_k(\theta)$ represent the radial and angular components, respectively:

$$H(r) = \begin{cases} \cos\left(\frac{\pi}{2} \log_2\left(\frac{2r}{\pi}\right)\right), & \frac{\pi}{4} < r < \frac{\pi}{2}, \\ 1, & r \leq \frac{\pi}{2}, \\ 0, & r \geq \frac{\pi}{4}. \end{cases} \quad (3)$$

$$G_k(\theta) = \alpha_k \left| \cos\left(\theta - \frac{\pi k}{K}\right) \right|^{K-1}, \quad (4)$$

where K is the number of orientations and $\alpha_k = 2^{k-1} \frac{(K-1)!}{\sqrt{K[2(K-1)]!}}$. Figure 2A provides an illustration of the band filter with $K = 4$.

With total recursive levels N , an image I can be decomposed as:

$$\mathbf{I} \rightarrow \begin{cases} \mathbf{I}_{H_0} \in \mathbf{R}^{H_0 \times W_0}, \\ \mathbf{I}_{B_1} \in \mathbf{R}^{K \times H_0 \times W_0}, \\ \mathbf{I}_{B_2} \in \mathbf{R}^{K \times H_1 \times W_1}, \\ \dots \\ \mathbf{I}_{B_N} \in \mathbf{R}^{K \times H_{N-1} \times W_{N-1}}, \\ \mathbf{I}_{L_0} \in \mathbf{R}^{H_{N-1} \times W_{N-1}}, \end{cases} \quad (5)$$

where \mathbf{I}_{H_0} and \mathbf{I}_{L_0} are high-frequency and low-frequency residues, and \mathbf{I}_{B_n} represents subband images at level n with k -th direction concatenated in the first dimension. Figure 2B shows an example of the decomposed output of an input image I .

Complex Steerable Pyramid Although the steerable pyramid effectively captures amplitude information across multiple orientations, it lacks the ability to extract local phase information, which is crucial for encoding structural features such as edges and corners (Canny, 1986). To address this limitation, (Portilla & Simoncelli, 2000) introduced the complex steerable pyramid, which extends the conventional steerable pyramid by converting its decomposed images into their analytic signal representation. In

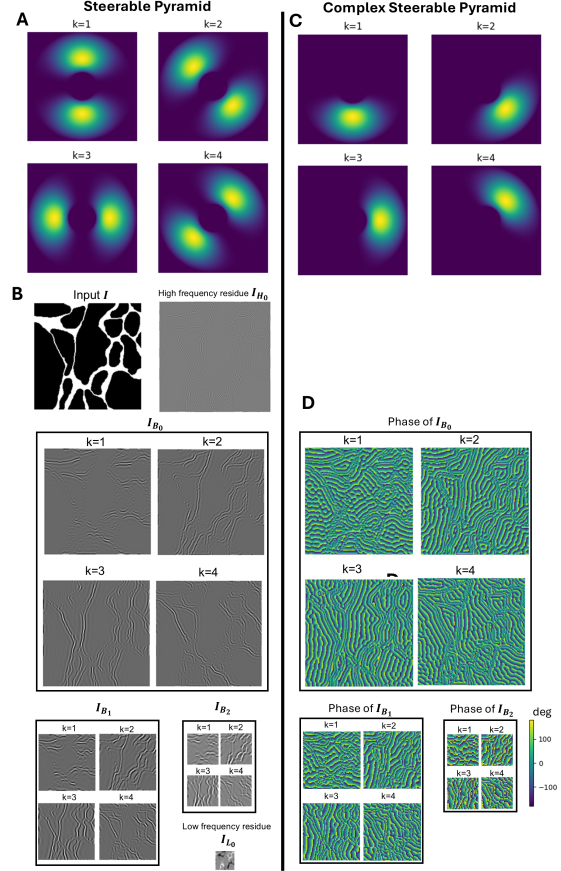


Figure 2. Steerable pyramid and complex steerable pyramid. (A) Orientation-selective band-pass filters of the steerable pyramid. (B) Example decomposition using a steerable pyramid with $N=3$, $K=4$. (C) Band-pass filters of the complex steerable pyramid, where negative frequency components are discarded. (D) Phase representation of the complex steerable pyramid output, with the real part identical to that of the steerable pyramid.

this formulation, the real part remains unchanged, while the imaginary part is obtained via the Hilbert transform of the real component. In the Fourier domain, this transformation is equivalent to discarding negative frequency components, as illustrated in Figure 2C.

For the complex steerable pyramid, the angular component $G_k(\theta)$ is modified as:

$$\tilde{G}_k(\theta) = \begin{cases} 2\alpha_k [\cos(\theta - \frac{\pi k}{K})]^{K-1}, & |\theta - \frac{\pi k}{K}| < \frac{\pi}{2}, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

This modification enables image decomposition into complex subbands, where phase information encodes critical structural features such as edges and corners (Figure 2D), while amplitude represents feature strength.

3.2. Mutual Information in Complex Wavelet Domain

According to Equation 5, the ground truth \mathbf{Y} and the prediction \mathbf{P} are decomposed into subbands \mathbf{Y}_{B_n} and $\mathbf{P}_{B_n} \in \mathbf{R}^{K \times H_{n-1} \times W_{n-1}}$ for each level $n \in [1, N]$. For each pixel (x, y) at level n , the K -directional features are treated as K -dimensional random variables.

For mutual information approximation, several studies, such as MINE (Belghazi et al., 2018) and Deep InfoMax (Hjelm et al., 2018), employ neural network-based estimators to produce tight lower bounds. While these methods are theoretically well-founded, they introduce additional training overhead, which increases algorithmic complexity and may hinder the efficiency of the loss function. Therefore, in this work, we adopt the mutual information estimation approach proposed by (Zhao et al., 2019):

$$I_l(\mathbf{Y}_{B_n}; \mathbf{P}_{B_n}) \approx -\frac{1}{2} \log \det(\mathbf{M}_n), \quad (7)$$

$$\mathbf{M}_n = \Sigma_{\mathbf{Y}_{B_n}} - \text{Cov}(\mathbf{Y}_{B_n}, \mathbf{P}_{B_n})(\Sigma_{\mathbf{P}_{B_n}}^{-1})^T \text{Cov}(\mathbf{Y}_{B_n}, \mathbf{P}_{B_n})^T \quad (8)$$

where $\Sigma_{\mathbf{Y}_{B_n}}$ and $\Sigma_{\mathbf{P}_{B_n}}$ are covariance matrices, and $\text{Cov}(\mathbf{Y}_{B_n}, \mathbf{P}_{B_n})$ is the cross-covariance.

When the ground truth and predictions are decomposed into complex subbands, this equation is extended with Hermitian transpose H :

$$\tilde{\mathbf{M}}_n = \Sigma_{\mathbf{Y}_{B_n}} - \text{Cov}(\mathbf{Y}_{B_n}, \mathbf{P}_{B_n})(\Sigma_{\mathbf{P}_{B_n}}^{-1})^H \text{Cov}(\mathbf{Y}_{B_n}, \mathbf{P}_{B_n})^H \quad (9)$$

where the covariance and cross-covariance are calculated as:

$$\Sigma_{\mathbf{Y}_{B_n}} = E[(\mathbf{Y}_{B_n} - E[\mathbf{Y}_{B_n}])(\mathbf{Y}_{B_n} - E[\mathbf{Y}_{B_n}])^H] \quad (10)$$

$$\Sigma_{\mathbf{P}_{B_n}} = E[(\mathbf{P}_{B_n} - E[\mathbf{P}_{B_n}])(\mathbf{P}_{B_n} - E[\mathbf{P}_{B_n}])^H] \quad (11)$$

$$\text{Cov}(\mathbf{Y}_{B_n}, \mathbf{P}_{B_n}) = E[(\mathbf{Y}_{B_n} - E[\mathbf{Y}_{B_n}])(\mathbf{P}_{B_n} - E[\mathbf{P}_{B_n}])^H] \quad (12)$$

Finally, the CWMI loss is computed as the sum of MI across all levels, which combines with cross entropy loss to integrate pixel-wise loss:

$$CWMI(\mathbf{Y}, \mathbf{P}) = (1-\lambda) \sum_{n=1}^N -I_l(\mathbf{Y}_{B_n}; \mathbf{P}_{B_n}) + \lambda L_{ce}(\mathbf{Y}, \mathbf{P}). \quad (13)$$

4. Experiment

4.1. Experimental Setup

Base Models To evaluate the effectiveness of the CWMI loss, we employed U-Net (Ronneberger et al., 2015) and Attention U-Net (Oktay et al., 2018) as baseline architectures.

This selection enables an assessment of CWMI’s generalization ability across both fully convolutional and attention-enhanced models. Additionally, to examine CWMI’s compatibility with recently proposed Mamba-based architectures (Yue & Li, 2024), we incorporated the Vision Mamba U-Net (VMUNet) (Ruan et al., 2024) into our experiments.

Datasets We tested CWMI on three public segmentation datasets, all characterized by class and instance imbalance: (1) SNEMI3D, a neurite segmentation dataset containing 100 1024 × 1024 grayscale images from electron microscopy slices (Arganda-Carreras et al., 2013); (2) GlaS, a gland segmentation dataset with 165 RGB images of varying sizes from histological images of colorectal cancer samples (Sirinukunwattana et al., 2017); (3) DRIVE, a retinal vessel segmentation dataset comprising 40 584 × 565 RGB images from fundus photographs (Staal et al., 2004); and (4) the Massachusetts Roads dataset (MASS ROAD), a road segmentation dataset with 1171 1500 × 1500 RGB images from aerial imagery (Mnih, 2013). We choose a subset of 120 images (ignoring images without a network of roads). For all datasets, at least two repeated three-fold cross-validations were used to ensure robust evaluation.

Baselines and Implementation Details We compared CWMI against 11 state-of-the-art (SOTA) loss functions, including pixel-wise loss functions (e.g., cross entropy, BCE (Long et al., 2015), Focal loss (Ross & Dollár, 2017)), Sensitive loss (Tang et al., 2025), region-based loss functions (e.g., Dice loss (Milletari et al., 2016), Tversky loss (Salehi et al., 2017), Jaccard loss (Rahman & Wang, 2016)), and structural/topological loss functions (e.g., WCE (Ronneberger et al., 2015), ABW loss (Liu et al., 2022), Skea-topo loss (Liu et al., 2024), RMI loss (Zhao et al., 2019), cIDice loss (Shit et al., 2021)). Hyperparameters for each baseline were tuned via grid search, with Tversky loss ($\alpha = .5, \beta = .5$) and Focal loss ($\gamma = 2.5$) as examples.

We utilized steerable pyramids with four decomposition levels and four orientations in all experiments. $N = 4$ and $K = 4$ of the steerable pyramid decomposition and a regularization parameter of $\lambda = 0.1$ were determined by the ablation experiments, and applied in all CWMI experiments. For the implementation of Equation 9, although PyTorch supports complex matrix calculations, our experiments indicated that its efficiency remains suboptimal. Consequently, we computed using their real representations, which are mathematically equivalent (Golub & Van Loan, 2013). To assess the significance of the complex steerable pyramid, we compared CWMI with a real-number-only variant (CWMI-Real), implemented according to Equations 4 and 8. Adam optimizer with a StepLR scheduler (initial learning rate 1×10^{-4} , decay rate 0.8, step size 10) was used. U-Net and Attention U-Net models were trained for 50 epochs; VMUNet models were trained for 100 epochs due to their slower convergence. All models were trained with a batch

Complex Wavelet Mutual Information Loss: A Multi-Scale Loss Function for Semantic Segmentation

SENMI3D										
Methods	mIoU \uparrow	mDice \uparrow	UNet VI \downarrow	ARI \uparrow	HD \downarrow	mIoU \uparrow	mDice \uparrow	AttenUNet VI \downarrow	ARI \uparrow	HD \downarrow
CE	.753 \pm .005	.851 \pm .004	2.082 \pm .270	.500 \pm .039	1.143 \pm .176	.751 \pm .009	.850 \pm .007	1.914 \pm .361	.525 \pm .056	1.178 \pm .261
BCE	.736 \pm .006	.841 \pm .004	1.756 \pm .156	.561 \pm .020	1.413 \pm .088	.735 \pm .010	.840 \pm .007	1.835 \pm .150	.553 \pm .023	1.417 \pm .224
Dice	.767 \pm .004	.862 \pm .003	1.406 \pm .085	.604 \pm .017	.919 \pm .154	.768 \pm .005	.862 \pm .004	1.507 \pm .086	.583 \pm .016	.804\pm.072
Focal	.728 \pm .007	.835 \pm .005	1.751 \pm .078	.556 \pm .016	1.725 \pm .225	.725 \pm .005	.833 \pm .003	1.911 \pm .189	.540 \pm .019	1.803 \pm .107
Jaccard	.766 \pm .004	.861 \pm .003	1.357 \pm .114	.614 \pm .019	.904 \pm .087	.762 \pm .003	.858 \pm .002	1.360 \pm .079	.609 \pm .007	1.000 \pm .125
Tversky	.765 \pm .004	.861 \pm .003	1.328 \pm .077	.617 \pm .017	.990 \pm .183	.765 \pm .003	.860 \pm .002	1.303 \pm .057	.621 \pm .009	.969 \pm .101
WCE	.714 \pm .005	.825 \pm .003	1.987 \pm .398	.525 \pm .042	1.575 \pm .250	.713 \pm .005	.825 \pm .004	1.818 \pm .034	.543 \pm .009	1.613 \pm .192
ABW	.605 \pm .066	.739 \pm .057	3.637 \pm 1.703	.299 \pm .187	4.605 \pm 2.926	.616 \pm .072	.749 \pm .063	3.077 \pm 1.303	.366 \pm .142	4.045 \pm 3.220
Skea-topo	.572 \pm .149	.671 \pm .197	3.822 \pm 2.560	.300 \pm .259	1.618 \pm .207	.602 \pm .095	.729 \pm .096	3.759 \pm 2.324	.306 \pm .243	4.889 \pm 5.928
RMI	.764 \pm .006	.859 \pm .005	1.443 \pm .207	.587 \pm .037	.943 \pm .130	.764 \pm .008	.859 \pm .006	1.404 \pm .124	.585 \pm .028	.958 \pm .170
clDice	.706 \pm .014	.819 \pm .011	2.049 \pm .351	.502 \pm .052	1.489 \pm .218	.701 \pm .022	.815 \pm .016	1.887 \pm .161	.518 \pm .034	1.720 \pm .244
Sensitive	.763 \pm .006	.858 \pm .005	1.543 \pm .069	.576 \pm .023	.948 \pm .143	.757 \pm .006	.854 \pm .004	1.822 \pm .313	.527 \pm .049	1.016 \pm .139
CWMI-Real	.776 \pm .004**	.867 \pm .003**	1.205 \pm .059**	.634 \pm .013*	.807 \pm .117*	.775 \pm .004**	.867 \pm .003**	1.193 \pm .070**	.634 \pm .014*	.824 \pm .103
CWMI	.778\pm.004***	.869\pm.003***	1.162\pm.068***	.638\pm.015*	.739\pm.095**	.777\pm.006**	.868\pm.005**	1.162\pm.085**	.639\pm.019*	.807 \pm .069
GlaS										
Methods	mIoU \uparrow	mDice \uparrow	UNet VI \downarrow	ARI \uparrow	HD \downarrow	mIoU \uparrow	mDice \uparrow	AttenUNet VI \downarrow	ARI \uparrow	HD \downarrow
CE	.637 \pm .204	.724 \pm .192	.933 \pm .043	.400 \pm .333	6.633 \pm 4.654	.640 \pm .209	.725 \pm .195	.906 \pm .107	.406 \pm .327	7.914 \pm 6.637
BCE	.614 \pm .226	.702 \pm .213	1.032 \pm .152	.355 \pm .355	4.165 \pm 1.929	.619 \pm .222	.708 \pm .208	.986 \pm .139	.365 \pm .358	4.354 \pm 1.869
Dice	.632 \pm .201	.719 \pm .189	.887 \pm .080	.396 \pm .322	7.004 \pm 4.670	.636 \pm .197	.723 \pm .187	.954 \pm .056	.396 \pm .322	7.335 \pm 4.843
Focal	.581 \pm .253	.672 \pm .241	1.093 \pm .103	.330 \pm .356	4.447 \pm 2.023	.581 \pm .252	.672 \pm .240	1.043 \pm .109	.336 \pm .369	5.040 \pm 2.802
Jaccard	.637 \pm .205	.723 \pm .193	.947 \pm .076	.397 \pm .335	6.979 \pm 4.511	.638 \pm .202	.725 \pm .190	.888 \pm .076	.406 \pm .331	6.758 \pm 4.431
Tversky	.634 \pm .205	.722 \pm .193	.953 \pm .031	.394 \pm .332	6.705 \pm 4.398	.627 \pm .195	.717 \pm .186	.924 \pm .102	.384 \pm .314	7.521 \pm 4.493
WCE	.829 \pm .010	.902 \pm .007	.832 \pm .075	.721 \pm .008	2.264 \pm .210	.823 \pm .016	.898 \pm .011	.873 \pm .097	.705 \pm .023	2.590 \pm .602
ABW	.760 \pm .012	.857 \pm .008	1.468 \pm .079	.619 \pm .027	3.527 \pm .508	.762 \pm .019	.858 \pm .013	1.472 \pm .103	.614 \pm .037	3.623 \pm .887
Skea-topo	.788 \pm .008	.876 \pm .005	1.259 \pm .042	.658 \pm .002	2.985 \pm .765	.784 \pm .016	.873 \pm .010	1.299 \pm .086	.645 \pm .022	3.363 \pm .609
RMI	.839 \pm .011	.907 \pm .007	.820 \pm .082	.726 \pm .024	2.644 \pm .372	.835 \pm .015	.905 \pm .010	.822 \pm .083	.720 \pm .039	2.701 \pm .366
clDice	.816 \pm .017	.894 \pm .012	.920 \pm .107	.698 \pm .026	2.809 \pm .476	.800 \pm .013	.884 \pm .009	1.004 \pm .113	.684 \pm .026	3.146 \pm .286
Sensitive	.821 \pm .014	.897 \pm .009	.941 \pm .101	.694 \pm .033	2.916 \pm .394	.824 \pm .018	.899 \pm .011	.880 \pm .108	.702 \pm .041	2.757 \pm .398
CWMI-Real	.838 \pm .014	.907 \pm .010	.798 \pm .068	.727 \pm .027	2.813 \pm .653	.842 \pm .022	.909 \pm .015	.788 \pm .063	.724 \pm .038	2.758 \pm .764
CWMI	.843\pm.016	.910\pm.011	.761\pm.081*	.735\pm.026	2.569\pm.466	.844\pm.007	.911\pm.005	.755\pm.106	.737\pm.015	2.569\pm.547
DRIVE										
Methods	mIoU \uparrow	mDice \uparrow	UNet VI \downarrow	ARI \uparrow	HD \downarrow	mIoU \uparrow	mDice \uparrow	AttenUNet VI \downarrow	ARI \uparrow	HD \downarrow
CE	.770 \pm .014	.856 \pm .011	1.379 \pm .136	.406 \pm .059	2.344 \pm .398	.757 \pm .017	.845 \pm .014	1.436 \pm .126	.372 \pm .059	2.932 \pm .875
BCE	.742 \pm .004	.835 \pm .003	1.501 \pm .175	.399 \pm .088	2.355 \pm .434	.746 \pm .012	.838 \pm .010	1.443 \pm .106	.427 \pm .033	2.169 \pm .264
Dice	.779 \pm .018	.863 \pm .014	1.293 \pm .116	.471 \pm .045	1.696 \pm .275	.776 \pm .015	.861 \pm .012	1.335 \pm .136	.445 \pm .079	2.146 \pm .832
Focal	.737 \pm .008	.832 \pm .006	1.404 \pm .168	.471 \pm .113	2.147 \pm .331	.745 \pm .011	.837 \pm .009	1.410 \pm .139	.450 \pm .075	2.141 \pm .400
Jaccard	.761 \pm .013	.850 \pm .011	1.289 \pm .181	.502 \pm .088	1.988 \pm .641	.766 \pm .020	.854 \pm .015	1.271 \pm .145	.521 \pm .051	1.623 \pm .313
Tversky	.752 \pm .020	.843 \pm .016	1.309 \pm .086	.517 \pm .052	1.638 \pm .369	.768 \pm .014	.856 \pm .011	1.265 \pm .129	.521 \pm .048	1.702 \pm .298
WCE	.741 \pm .013	.835 \pm .011	1.417 \pm .069	.462 \pm .019	2.100 \pm .498	.733 \pm .009	.828 \pm .007	1.441 \pm .031	.444 \pm .039	2.326 \pm .848
ABW	—	—	—	—	—	—	—	—	—	—
Skea-topo	—	—	—	—	—	.599 \pm .121	.674 \pm .169	1.602 \pm .113	.265 \pm .250	5.791 \pm 4.492
RMI	.787 \pm .012	.869 \pm .010	1.282 \pm .127	.449 \pm .064	1.797 \pm .489	.785 \pm .009	.867 \pm .007	1.279 \pm .096	.448 \pm .047	1.884 \pm .340
clDice	.482 \pm .094	.598 \pm .085	2.010 \pm .119	.207 \pm .129	9.813 \pm 8.153	.509 \pm .063	.617 \pm .066	1.912 \pm .157	.238 \pm .161	15.704 \pm 11.080
Sensitive	.763 \pm .008	.850 \pm .007	1.440 \pm .060	.376 \pm .039	2.824 \pm .562	.757 \pm .018	.846 \pm .015	1.444 \pm .040	.380 \pm .028	3.134 \pm .782
CWMI-Real	.788 \pm .016	.870 \pm .012	1.104 \pm .175*	.582 \pm .075	1.309 \pm .299*	.787 \pm .026	.870 \pm .019	1.072 \pm .113**	.594 \pm .057*	1.398 \pm .468
CWMI	.798\pm.012*	.878\pm.009*	1.032\pm.176*	.613\pm.082*	1.079\pm.382*	.795\pm.015*	.875\pm.012*	1.007\pm.162**	.622\pm.062**	1.219\pm.410*
MASS ROAD										
Methods	mIoU \uparrow	mDice \uparrow	UNet VI \downarrow	ARI \uparrow	HD \downarrow	mIoU \uparrow	mDice \uparrow	AttenUNet VI \downarrow	ARI \uparrow	HD \downarrow
CE	.733 \pm .015	.826 \pm .013	3.473 \pm .608	.194 \pm .083	1.927 \pm 3.301	.640 \pm .209	.725 \pm .195	.906 \pm .107	.406 \pm .327	7.914 \pm 6.637
BCE	.689 \pm .011	.794 \pm .009	1.708 \pm .265	.560 \pm .059	12.792 \pm 3.684	.619 \pm .222	.708 \pm .208	.986 \pm .139	.365 \pm .358	4.354 \pm 1.869
Dice	.761 \pm .011	.849 \pm .009	1.601 \pm .138	.558 \pm .042	11.189 \pm 4.532	.636 \pm .197	.723 \pm .187	.954 \pm .056	.396 \pm .322	7.335 \pm 4.843
Focal	.671 \pm .015	.780 \pm .013	1.717 \pm .150	.556 \pm .042	11.576 \pm 1.950	.581 \pm .252	.672 \pm .240	1.043 \pm .109	.336 \pm .369	5.040 \pm 2.802
Jaccard	.759 \pm .008	.848 \pm .006	1.332 \pm .041	.630 \pm .008	11.274 \pm 2.805	.638 \pm .202	.725 \pm .190	.888 \pm .076	.406 \pm .331	6.758 \pm 4.431
Tversky	.755 \pm .009	.845 \pm .007	1.379 \pm .196	.618 \pm .061	11.501 \pm 4.470	.627 \pm .195	.717 \pm .186	.924 \pm .102	.384 \pm .314	7.521 \pm 4.493
WCE	.650 \pm .016	.762 \pm .014	1.726 \pm .103	.556 \pm .004	1.845 \pm .470	.823 \pm .016	.898 \pm .011	.873 \pm .097	.705 \pm .023	2.590 \pm .602
ABW	.685 \pm .014	.791 \pm .012	1.678 \pm .124	.566 \pm .030	11.654 \pm 4.661	.762 \pm .019	.858 \pm .013	1.472 \pm .103	.614 \pm .037	3.623 \pm .887
Skea-topo	.620 \pm .095	.734 \pm .085	2.236 \pm .824	.463 \pm .159	11.653 \pm 5.620	.784 \pm .016	.873 \pm .010	1.299 \pm .086	.645 \pm .022	3.363 \pm .609
RMI	.766 \pm .008	.852 \pm .006	1.824 \pm .263	.499 \pm .065	11.632 \pm 3.918	.839 \pm .013	.908 \pm .008	.819 \pm .071	.727 \pm .032	2.587 \pm .435
clDice	.645 \pm .051	.755 \pm .045	2.162 \pm .467	.476 \pm .081	15.461 \pm 4.999	.800 \pm .013	.884 \pm .009	1.004 \pm .113	.684 \pm .026	3.146 \pm .286
Sensitive	.734 \pm .022	.827 \pm .018	2.933 \pm .982	.305 \pm .159	11.292 \pm 3.828	.824 \pm .018	.899 \pm .011	.880 \pm .108	.702 \pm .041	2.757 \pm .398
CWMI-Real	.765 \pm .010	.853 \pm .008	1.231 \pm .183	.650 \pm .050	1.836 \pm 4.994	.846 \pm .012	.913 \pm .008	.766 \pm .067***	.736 \pm .021*	2.504 \pm .528
CWMI	.767\pm.010	.854\pm.008	1.148\pm.166*	.667\pm.047*	1.326\pm4.109	.839\pm.015	.908\pm.010	.762\pm.084***	.728\pm.035**	2.747\pm.680

Table 1. Quantitative results of different loss functions across the four datasets and two neural network models. The **bold** numbers indicate the best performance for each metric, while the

size of 10. Early stopping based on mIoU was employed to select the best model. Training was conducted on an NVIDIA A100 GPU using the Google Colab runtime.

Data Augmentation and Evaluation Metrics Random flips and rotations were applied to all datasets to improve generalization. For SNEMI3D and MASS ROAD, images were randomly cropped to 512×512 , while for GlaS, images were cropped to 448×576 to standardize input sizes. No cropping was performed for DRIVE due to its uniform image dimensions.

Performance was evaluated using five metrics: mIoU and mDice for regional precision, variation of information (VI) (Nunez-Iglesias et al., 2013) and adjusted Rand index (ARI) (Vinh et al., 2009) for clustering precision, and Hausdorff distance (HD) for boundary and topological accuracy. These metrics provide a comprehensive assessment of both regional overlap and structural fidelity.

SNEMI3D VMUNet					
Methods	mIoU \uparrow	mDice \uparrow	VI \downarrow	ARI \uparrow	HD \downarrow
CE	.723 \pm .026	.829 \pm .020	2.577 \pm 1.211	.414 \pm .173	1.709 \pm .235
BCE	.739 \pm .004	.843 \pm .002	1.499 \pm .057	.594 \pm .010	1.274 \pm .231
Dice	.725 \pm .029	.831 \pm .021	2.735 \pm 1.213	.402 \pm .165	1.767 \pm .720
Focal	.725 \pm .010	.833 \pm .007	1.677 \pm .141	.563 \pm .025	1.699 \pm .372
Jaccard	.725 \pm .019	.831 \pm .014	1.953 \pm .541	.499 \pm .075	1.716 \pm .566
Tversky	.737 \pm .016	.840 \pm .012	1.619 \pm .324	.555 \pm .045	1.560 \pm .408
WCE	.714 \pm .005	.826 \pm .004	1.743 \pm .026	.552 \pm .006	1.432 \pm .169
ABW	.678 \pm .003	.800 \pm .003	2.076 \pm .030	.490 \pm .005	1.572 \pm .021
RMI	.762 \pm .010	.857 \pm .007	1.419 \pm .183	.588 \pm .030	1.071 \pm .306
cdDice	.714 \pm .002	.824 \pm .001	1.754 \pm .062	.542 \pm .003	1.990 \pm .262
sensitive	—	—	—	—	—
CWMI	.783\pm.004*	.872\pm.003*	.982\pm.106*	.660\pm.022*	.629\pm.073

Table 2. Quantitative results of different loss functions on SNEMI3D dataset and VMUNet model. The **bold** numbers indicate the best performance for each metric, while the underlined numbers denote the second-best performance. A hyphen (“-”) indicates cases where the model did not converge. For the loss metrics CWMI, the statistical significance is marked with asterisks: * $p < .05$, where p is the maximum p-value of the student’s t tests against all the baseline loss functions.

4.2. Quantitative and qualitative results

As shown in Table 1, the proposed CWMI loss outperforms other loss functions across the majority of evaluation metrics for all datasets using both U-Net and Attention U-Net architectures. Statistically, CWMI achieves significantly superior performance on all metrics for the SNEMI3D and DRIVE datasets, as well as on clustering-based metrics (VI and ARI) for the GlaS and MASS ROAD datasets. Moreover, CWMI consistently outperforms its real-valued variant, underscoring the importance of incorporating phase information within the steerable pyramid decomposition. For the Mamba-based architecture, VMUNet, CWMI also significantly outperforms baseline loss functions on the SNEMI3D dataset (Table 2). Qualitative results from SNEMI3D (Figure 3), GlaS (Figure 4), DRIVE (Figure 5), and MASS ROAD (Figure 6) further illustrate CWMI’s effectiveness

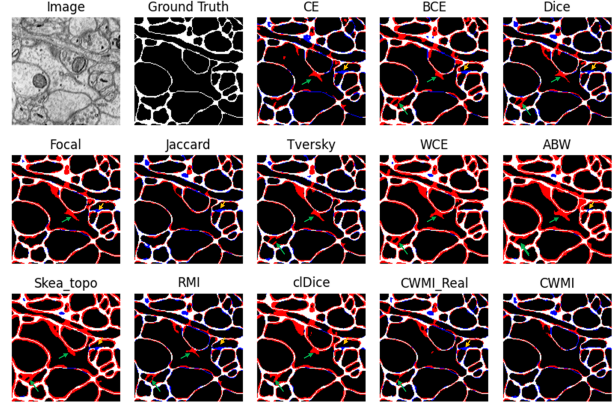


Figure 3. Qualitative results of different loss functions on the SNEMI3D dataset. Red: false positive regions; Blue: false negative regions. Green arrow: challenging false positive and Orange arrow: challenging false negative that are successfully addressed by CWMI.

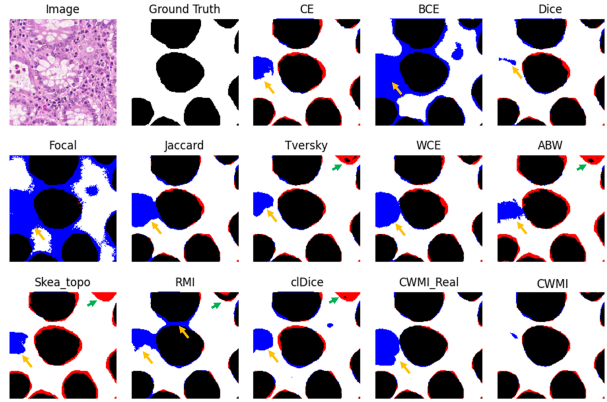


Figure 4. Qualitative results of different loss functions on the GlaS dataset.

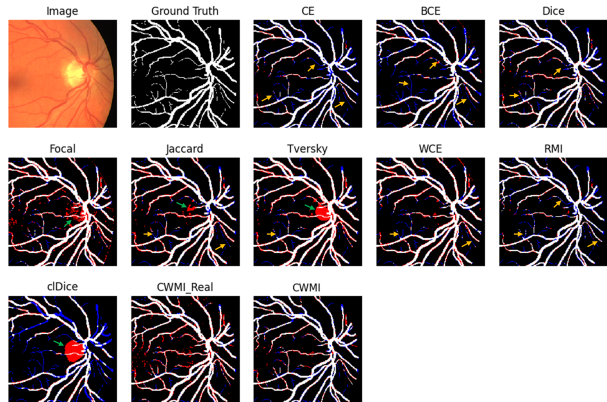


Figure 5. Qualitative results of different loss functions on the DRIVE dataset.

in mitigating challenging false positive and false negative segmentation errors that persist with other state-of-the-art loss functions.

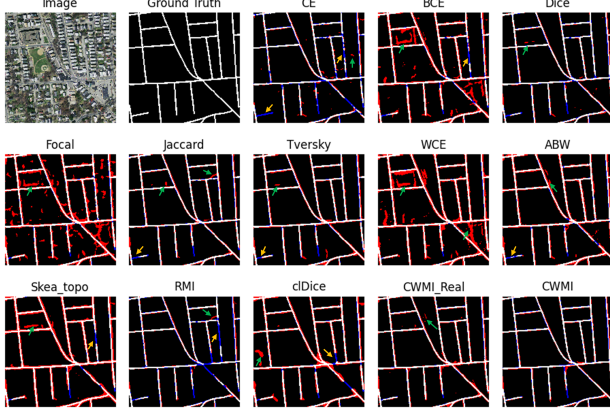


Figure 6. Qualitative results of different loss functions on the MASS ROAD dataset.

4.3. Ablation studies

In this section, all ablation experiments were performed on the SNEMI3D dataset with U-Net model.

Mutual Information (MI) vs L_1 , L_2 Distance and Structural Similarity (SSIM) As previously discussed, various metrics based on wavelet transforms have been developed for evaluating image similarity and guiding loss functions. To assess the advantages of mutual information (MI) over commonly used metrics such as L_1 , L_2 , and SSIM, we conducted a comparative analysis. The experimental results, summarized in Table 3, demonstrate that MI consistently outperforms L_1 , L_2 distance, and SSIM across multiple evaluation metrics. Unlike L_1 and L_2 , which focus on pixel-wise intensity differences, and SSIM, which emphasizes structural similarity, MI captures joint statistical dependencies between features in each direction. This ability provides a more robust representation of structural differences between predictions and labels, particularly in complex segmentation tasks.

Methods	mIoU \uparrow	mDice \uparrow	VI \downarrow	ARI \uparrow	HD \downarrow
L_1	.773 \pm .002	.866 \pm .001	1.22 \pm .05	.632 \pm .007	.70 \pm .08
L_2	.774 \pm .003	.867 \pm .002	1.23 \pm .06	.633 \pm .011	.76 \pm .13
SSIM	.468 \pm .032	.613 \pm .020	5.96 \pm .96	.081 \pm .030	9.31 \pm 1.64
CWMI	.779\pm.004	.870\pm.003	1.16\pm.09	.640\pm.005	.69\pm.04

Table 3. Quantitative results comparing mutual information (MI), L_1 , L_2 distance, and structural similarity (SSIM) based on U-Net with SNEMI3D. The **bold** numbers indicate the best performance for each metric.

Contribution of different components of complex steerable pyramid To better understand the role of individual components in the CWMI loss, we conducted an ablation experiment evaluating the performance of variants based on the mutual information loss of magnitude-only, phase-only, and real-only representations, as shown in Table 4. The CWMI-Phase variant consistently underperformed across all metrics, suggesting that phase information alone

is insufficient for reliable segmentation. The CWMI-Mag variant, by contrast, achieved the best performance on clustering metrics (VI and ARI), indicating its strong capacity to capture regional structures. The full CWMI, combining both magnitude and phase via the complex representation, yielded the best results on pixel-wise accuracy metrics (mIoU and mDice) and topological metric (HD), while ranking second on VI and ARI, demonstrating a well-rounded performance across different dimensions of segmentation quality. The CWMI-Real variant performed worse than both CWMI and CWMI-Mag on most metrics, although it outperformed CWMI-Phase, further supporting the necessity of incorporating both magnitude and phase components. These results highlight the critical role of magnitude in structural segmentation, while also demonstrating that the full complex formulation of CWMI achieves superior generalization across diverse evaluation criteria.

Methods	mIoU \uparrow	mDice \uparrow	VI \downarrow	ARI \uparrow	HD \downarrow
CWMI-Real	.776 \pm .004	.868 \pm .003	1.211 \pm .048	.635 \pm .011	.774 \pm .065
CWMI-Mag	.775 \pm .004	.867 \pm .003	.908\pm.051	.660\pm.009	.735 \pm .065
CWMI-Phase	.694 \pm .008	.810 \pm .006	2.487 \pm .534	.458 \pm .061	2.000 \pm .258
CWMI	.778\pm.004	.869\pm.003	1.164 \pm .079	.638 \pm .016	.720\pm.066

Table 4. Quantitative comparison of mutual information loss across different components of the complex steerable pyramid decomposition. CWMI-Real: real component; CWMI-Mag: magnitude component; CWMI-Phase: phase component; CWMI: full complex representation (combining real and imaginary components). Results are based on a U-Net model trained on the SNEMI3D dataset. **Bold** values indicate the best performance, and underlined values denote the second-best for each evaluation metric.

Impact of Regularization Parameter λ To evaluate the sensitivity of CWMI to the regularization weight λ , we experimented with $\lambda = 0.1, 0.5, 0.9$. As shown in Table 5, performance across all evaluation metrics remained largely stable, with only marginal variations. This robustness suggests that CWMI provides complementary information to the cross-entropy term and is not overly sensitive to the regularization strength. Additionally, we found the inclusion of the cross-entropy term to be crucial for directional learning, as mutual information is inherently symmetric and cannot distinguish between correct and inverted label assignments.

Impact of Decomposition Level N and Number of Orientations K Theoretically, the steerable pyramid can decompose an image into very high levels with an infinite number of orientations, provided the input image is sufficiently large. However, does deeper decomposition or a higher number of orientations improve feature extraction and loss computation? To address this, we analyzed the impact of N (decomposition level) and K (number of orientations) on the performance of CWMI, as shown in Table 3. Interestingly, both N and K achieved optimal performance at relatively low values, suggesting that the critical information for segmentation is concentrated in relatively high-frequency regions. This observation supports the idea

that emphasizing high-frequency subbands refines small instances and narrow boundaries, while lower-frequency components from deeper decompositions contribute less informative features compared to higher-frequency ones.

Knowing that the first four layers of decomposition are crucial, we further performed a layer ablation experiment (Table 3, Layer Ablation), where only selected layers were used to compute mutual information while discarding the remaining layers. Our results show that the third layer outperforms other layers in region-based (mIoU, mDice) and cluster-based metrics (VI, ARI), while the fourth layer achieves the best performance in topological metrics (HD). These findings suggest that features extracted from multiple layers are essential for achieving high performance across all evaluation metrics, highlighting the importance of incorporating information from both mid- and high-frequency subbands in the segmentation task.

λ ablation					
	mIoU \uparrow	mDice \uparrow	VI \downarrow	ARI \uparrow	HD \downarrow
$\lambda=0.1$.778 \pm .004	.869 \pm .003	1.188 \pm .075	.636 \pm .008	.703 \pm .034
$\lambda=0.5$.778 \pm .006	.869 \pm .004	1.176 \pm .076	.639 \pm .014	.729 \pm .078
$\lambda=0.9$.777 \pm .003	.869 \pm .002	1.219 \pm .073	.636 \pm .013	.740 \pm .061
K ablation					
	mIoU \uparrow	mDice \uparrow	VI \downarrow	ARI \uparrow	HD \downarrow
$K=2$.777 \pm .006	.868 \pm .004	1.18 \pm .10	.637 \pm .028	.79 \pm .02
$K=4$.779 \pm .004	.870 \pm .003	1.16 \pm .09	.640 \pm .005	.69 \pm .04
$K=8$.776 \pm .003	.868 \pm .002	1.20 \pm .06	.635 \pm .009	.76 \pm .08
$K=12$.772 \pm .005	.865 \pm .003	1.33 \pm .05	.621 \pm .004	.81 \pm .08
N ablation					
	mIoU \uparrow	mDice \uparrow	VI \downarrow	ARI \uparrow	HD \downarrow
$N=2$.760 \pm .002	.857 \pm .001	1.30 \pm .04	.618 \pm .009	1.20 \pm .12
$N=4$.777 \pm .002	.869 \pm .001	1.21 \pm .10	.634 \pm .017	.79 \pm .190
$N=6$.768 \pm .003	.862 \pm .002	1.29 \pm .05	.618 \pm .004	.88 \pm .18
Layer ablation					
	mIoU \uparrow	mDice \uparrow	VI \downarrow	ARI \uparrow	HD \downarrow
1st layer	.738 \pm .008	.843 \pm .006	1.61 \pm .07	.576 \pm .015	1.38 \pm .14
2nd layer	.762 \pm .008	.858 \pm .006	1.25 \pm .08	.624 \pm .016	1.12 \pm .19
3rd layer	.774 \pm .003	.867 \pm .002	1.23 \pm .10	.638 \pm .020	.82 \pm .12
4th layer	.766 \pm .004	.861 \pm .003	1.39 \pm .09	.614 \pm .012	.76 \pm .04

Table 5. The impact of regularization parameter λ , decomposition level N , and orientation number K on CWMI performance on U-Net with SNEMI3D. In the layer ablation test, only the selected layer is computed for mutual information and all other layers are discard. The **bold** numbers indicate the best performance for each metric.

Computational complexity analysis For an input image of size $H \times W$ with K orientation decompositions, the computational complexity of the CWMI loss function is analyzed as follows.

First, the forward Fourier transform has a time complexity of $O(HW \log(HW))$. For the first layer of decomposition, the operations include: Band-pass filtering: $O(HWK)$; Inverse Fourier transform: $O(HW \log(HW)K)$; and mutual information computation: $O(HWK^2)$. Summing these terms, the total complexity for the first decomposition layer is: $O(HW(K \log(HW) + K^2))$. For subsequent decomposition layers, the image size reduces by a factor of four at each step, meaning the second layer pro-

cesses an image of size $HW/4$, and the third layer processes $HW/16$, and so on. Thus, the total computational complexity follows a geometric series with a superior bound $O(\frac{4}{3}HW(K \log(HW) + K^2))$

From our ablation experiments, the optimal number of orientations is $K = 4$, which is relatively small. Hence the complexity of CWMI is linear to $O(HW \log(HW))$, which remains scalable for high-resolution images. Compared to topology-aware losses such as RMI, cIDice, or Hausdorff Distance Loss, CWMI achieves better structural and boundary preservation with lower computational overhead, as shown in Table 4.

	Epoch Time (s)	Δt to CE (s)
CE	2.04	.00
BCE	2.02	-0.02
Dice	2.01	-0.03
Focal	2.01	-0.03
Jaccard	2.02	-0.02
Tversky	2.01	-0.03
WCE	2.15	0.12
ABW	2.33	0.30
Skea-topo	2.72	0.69
RMI	2.28	0.25
cIDice	2.37	0.34
CWMI-Real	2.18	0.15
CWMI	2.27	0.23

Table 6. Training time per epoch and relative change to CE baseline for various loss functions on U-Net model with SNEMI3D dataset.

5. Conclusion

In this study, we introduced Complex Wavelet Mutual Information (CWMI) loss, a novel loss function for semantic segmentation that leverages the multiscale, multi-orientation decomposition capabilities of the complex steerable pyramid. By integrating mutual information across wavelet subbands, CWMI effectively captures high-dimensional dependencies and local structural features, including critical phase information, which are essential for accurate segmentation. Extensive experiments on four challenging datasets demonstrate that CWMI consistently outperforms state-of-the-art loss functions across most evaluation metrics, particularly in segmenting small instances and narrow boundaries, while introducing minimal computational overhead. These results highlight CWMI as a robust and versatile loss function that effectively addresses key challenges in segmentation, such as class and instance imbalance, boundary precision, and topological consistency.

Beyond semantic segmentation, the core principles of CWMI—multiscale feature extraction and structural consistency—suggest its potential applicability to a broader range of computer vision and machine learning tasks, such as image-to-image translation and super-resolution, which we leave for future exploration.

Although extensively studied on 2D images, extending it to multi-class and 3D segmentation is theoretically feasible but necessitates further validation in future research.

Acknowledgements

This study was conducted without external funding. I would like to thank the reviewers for their constructive feedback, which significantly improved the quality of this work. I am especially grateful to my wife, Di Duan, for her unwavering support of my passion for machine learning. I thank Xinzi He for valuable discussions that contributed to the revision of this study.

Impact Statement

The proposed Complex Wavelet Mutual Information (CWMI) loss introduces a novel approach to structural-aware learning in deep neural networks. By leveraging multiscale decomposition through the complex steerable pyramid and mutual information across frequency subbands, CWMI enables improved segmentation performance, particularly for small-scale structures and thin boundaries. Our empirical results demonstrate significant improvements in both pixel-wise and topological accuracy across multiple datasets. Beyond segmentation, CWMI has the potential to generalize to a wide range of real-world applications, including medical imaging, autonomous driving, satellite-based environmental monitoring, and industrial defect detection.

Overall, CWMI provides a **computationally efficient, structure-aware** loss function that enhances segmentation performance while maintaining practical scalability. By extending its application beyond segmentation, we aim to contribute to the broader field of generative modeling, object detection, and self-supervised learning in deep neural networks.

References

- Arganda-Carreras, I., Seung, H. S., Vishwanathan, A., and Berger, D. R. Snemi3d: 3d segmentation of neurites in em images (isbi 2013). <https://snemi3d.grand-challenge.org/>, 2013. Accessed: 2024-12-20.
- Azad, R., Heidary, M., Yilmaz, K., Hüttemann, M., Karim-ijafarbigloo, S., Wu, Y., Schmeink, A., and Merhof, D. Loss functions in the era of semantic segmentation: A survey and outlook. *arXiv preprint arXiv:2312.05391*, 2023.
- Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.
- Canny, J. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.
- Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- Diakogiannis, F. I., Waldner, F., Caccetta, P., and Wu, C. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.
- Golub, G. H. and Van Loan, C. F. *Matrix computations*. JHU press, 2013.
- Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
- Huang, M., Huang, C., Yuan, J., and Kong, D. Fixed-point deformable u-net for pancreas ct segmentation. In *Proceedings of the Third International Symposium on Image Computing and Digital Medicine*, pp. 283–287, 2019.
- Islam, M., Vibashan, V., Jose, V. J. M., Wijethilake, N., Utkarsh, U., and Ren, H. Brain tumor segmentation and survival prediction using 3d attention unet. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I* 5, pp. 262–272. Springer, 2020.
- Jiang, W., Li, Y., Yi, Z., Chen, M., and Wang, J. Multi-instance imbalance semantic segmentation by instance-dependent attention and adaptive hard instance mining. *Knowledge-Based Systems*, 304:112554, 2024.
- Karimi, D. and Salcudean, S. E. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on medical imaging*, 39(2):499–513, 2019.
- Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., and Ayed, I. B. Boundary loss for highly unbalanced segmentation. In *International conference on medical imaging with deep learning*, pp. 285–296. PMLR, 2019.
- Kim, M. W. and Cho, N. I. Whfl: Wavelet-domain high frequency loss for sketch-to-image translation. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, pp. 744–754, 2023.
- Kofler, F., Shit, S., Ezhov, I., Fidon, L., Horvath, I., Al-Maskari, R., Li, H. B., Bhatia, H., Loehr, T., Piraud, M., et al. Blob loss: Instance imbalance aware loss functions

- for semantic segmentation. In *International Conference on Information Processing in Medical Imaging*, pp. 755–767. Springer, 2023.
- Korkmaz, C. and Tekalp, A. M. Training transformer models by wavelet losses improves quantitative and visual performance in single image super-resolution. *arXiv preprint arXiv:2404.11273*, 2024.
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.
- Liu, C., Ma, B., Ban, X., Xie, Y., Wang, H., Xue, W., Ma, J., and Xu, K. Enhancing boundary segmentation for topological accuracy with skeleton-based methods. *arXiv preprint arXiv:2404.18539*, 2024.
- Liu, W., Chen, J., Liu, C., Ban, X., Ma, B., Wang, H., Xue, W., and Guo, Y. Boundary learning by using weighted propagation in convolution network. *Journal of Computational Science*, 62:101709, 2022.
- Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.
- Mallat, S. G. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.
- Milletari, F., Navab, N., and Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.
- Mnih, V. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013.
- Nunez-Iglesias, J., Kennedy, R., Parag, T., Shi, J., and Chklovskii, D. B. Machine learning of hierarchical clustering to segment 2d and 3d images. *PloS one*, 8(8): e71715, 2013.
- Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- Oner, D., Garin, A., Koziński, M., Hess, K., and Fua, P. Persistent homology with improved locality information for more effective delineation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10588–10595, 2023.
- Portilla, J. and Simoncelli, E. P. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40:49–70, 2000.
- Prantl, L., Bender, J., Kugelstadt, T., and Thuerey, N. Wavelet-based loss for high-frequency interface dynamics. *arXiv preprint arXiv:2209.02316*, 2022.
- Rahman, M. A. and Wang, Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pp. 234–244. Springer, 2016.
- Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.
- Ross, T.-Y. and Dollár, G. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2980–2988, 2017.
- Ruan, J., Li, J., and Xiang, S. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.
- Salehi, S. S. M., Erdogmus, D., and Gholipour, A. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pp. 379–387. Springer, 2017.
- Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., and Markey, M. K. Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, 18(11):2385–2401, 2009.
- Shit, S., Paetzold, J. C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J. P., Bauer, U., and Menze, B. H. cldice-a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16560–16569, 2021.
- Simoncelli, E. P., Freeman, W. T., Adelson, E. H., and Heeger, D. J. Shiftable multiscale transforms. *IEEE transactions on Information Theory*, 38(2):587–607, 1992.
- Sirinukunwattana, K., Pluim, J. P., Chen, H., Qi, X., Heng, P.-A., Guo, Y. B., Wang, L. Y., Matuszewski, B. J., Bruni, E., Sanchez, U., et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

- Staal, J., Abràmoff, M. D., Niemeijer, M., Viergever, M. A., and Van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.
- Stucki, N., Paetzold, J. C., Shit, S., Menze, B., and Bauer, U. Topologically faithful image segmentation via induced matching of persistence barcodes. In *International Conference on Machine Learning*, pp. 32698–32727. PMLR, 2023.
- Tang, Q., Liu, F., Zhang, D., Jiang, J., Tang, X., and Chen, C. P. Increase the sensitivity of moderate examples for semantic image segmentation. *Image and Vision Computing*, 154:105357, 2025.
- Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pp. 1073–1080, 2009.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.
- Yang, H.-H., Yang, C.-H. H., and Tsai, Y.-C. J. Y-net: Multi-scale feature aggregation network with wavelet structure similarity loss function for single image dehazing. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2628–2632. IEEE, 2020.
- Yue, Y. and Li, Z. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*, 2024.
- Zhao, S., Wang, Y., Yang, Z., and Cai, D. Region mutual information loss for semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.
- Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pp. 3–11. Springer, 2018.
- Zhu, Q., Wang, H., and Zhang, R. Wavelet loss function for auto-encoder. *IEEE Access*, 9:27101–27108, 2021.