A Brain-Inspired Saliency Prediction Framework for Human-AI Cognitive Consistency in AIGC Content via Multi-Region Liquid Neurons

Shibo Wang[1], Yan Zhao[1*], Shigang Wang[1], Jian Wei[1], Shuo Li[1]

[1]College of Communication Engineering, Jilin University, Changchun 130012, China
wangsb23@mails.jlu.edu.cn, zhao_y@jlu.edu.cn

Abstract

In recent years, human-AI cognitive consistency has emerged as a crucial perspective for evaluating the perceptual quality and interpretability of AIGC (Artificial Intelligence Generated Content). This paper proposes a biologically inspired saliency prediction framework that models six core regions of the human visual system—namely V1, V2, V4, MT, LIP, and FEF—using liquid neurons to capture the dynamic saliency features aligned with human gaze behavior. To enable effective alignment between AIGC models and human cognitive mechanisms, we introduce a cross-domain dual-teacher distillation strategy and construct a large-scale multimodal dataset comprising natural images, eye-tracking data, AIGC-generated images, and their corresponding cross-attention maps. Furthermore, we propose HAMCI (Human-AI Mutual Cognitive Index), a novel metric designed to quantitatively assess the spatial and semantic alignment between predicted saliency maps and model attention distributions. The proposed method demonstrates promising performance across various saliency prediction and cognitive alignment tasks, with results comparable to or surpassing recent state-of-the-art methods in several benchmarks. The code and dataset will be released upon acceptance to facilitate future research on cognitively aligned AIGC evaluation.

Introduction

With the rapid advancement of AIGC technologies, image generation models are increasingly applied in creative design, virtual reality, and human-computer interaction (Ramesh et al. 2021; Dhariwal and Nichol 2021; Esser, Rombach, and Ommer 2021). Unlike traditional computer vision tasks, AIGC images often lack paired reference counterparts, making their quality assessment more dependent on human visual and cognitive priors (Fang et al. 2024). This challenge has motivated a new line of research toward evaluation frameworks that integrate semantic understanding and attention modeling (Li et al. 2019; Pan et al. 2021; Yang, Zhao, and Sun 2023).

To evaluate the perceptual quality of AIGC content without reference images, various image quality assessment (IQA) methods have been explored. Among them, distribution-based metrics such as FID (Heusel et al. 2017) and KID (Binkowski et al. 2018) are widely adopted. These reference-free approaches compare statistical distributions between generated and real image sets, offering a practical and scalable solution for generative model evaluation. Complementing global fidelity estimation, emerging methods incorporate spatial semantics and perceptual attention to better align with human visual perception.

Recent efforts have introduced human visual priors, particularly attention mechanisms, into IQA frameworks. Saliency-guided approaches (Wang, Liu, and Liu 2019; Zhang, Zhang, and Mou 2020), for instance, leverage predicted attention maps to highlight perceptually important regions, thereby enhancing consistency with human viewing behavior. In parallel, cognitively inspired models (Li et al. 2019; Pan et al. 2021; Yang, Zhao, and Sun 2023) integrate semantic reasoning and attention modeling to bridge low-level fidelity with high-level human cognition.

Despite these advancements, most existing approaches focus on natural scenes and rely solely on human gaze data, while overlooking internal attention mechanisms within AIGC models. Notably, diffusion-based generators such as Stable Diffusion (Rombach et al. 2022), DALL·E 3 (Ramesh et al. 2022), and Imagen (Saharia et al. 2022) inherently produce cross-attention maps that encode fine-grained correspondences between textual and visual features during synthesis. These attention maps reflect the model's internal focus across modalities and offer a rich yet underexplored signal for evaluation. Aligning model-derived attention with human gaze opens a promising direction toward cognitively consistent and interpretable AIGC assessment.

To address the gap in cognitively grounded evaluation for AIGC, we propose an integrated framework inspired by both neuroscience and deep learning. By bridging low-level human perception with high-level generative attention, our approach enables interpretable, consistent, and human-aligned saliency prediction across both real and generated content. Central to our design are biologically plausible brain-region modeling, cross-modal knowledge transfer, and a new metric for quantifying human-AI consistency. Together, these innovations lay the foundation for cognitively aligned evaluation of AIGC content.

The main contributions of this work are summarized as follows:

- We propose a saliency prediction framework that mod-

*Corresponding author.

els six key brain regions using liquid neural modules. To the best of our knowledge, this is the first work to incorporate liquid neuron-based dynamic modules into brain-inspired saliency modeling for visual cognition.

- We design a cross-domain dual-teacher distillation mechanism guided by both human gaze heatmaps and AIGC model cross-attention maps. Additionally, we introduce the HAMCI metric to quantitatively evaluate human-AI attention consistency from both spatial and semantic perspectives.
- We construct a multimodal dataset and perform experiments to evaluate the proposed framework. The results indicate promising generalization ability and cognitive consistency across saliency and attention alignment benchmarks.

Related Work

Brain-Inspired Saliency Modeling

Visual saliency prediction aims to estimate the spatial distribution of human attention when observing images, serving as a key bridge between computer vision and human perceptual behavior (Cheng et al. 2015). Early models, such as the classical approach by Itti et al.(Itti, Koch, and Niebur 1998), rely on the integration of low-level visual features. With the rise of deep learning, models like DeepGaze II(Kümmerer, Wallis, and Bethge 2016) and SALICON (Huang et al. 2015) leverage deep convolutional architectures and large-scale gaze datasets. More recently, Transformer-based architectures such as TransSalNet (Liu et al. 2021a) and TASED-Net (Linardos, Evangelopoulos, and Potamianos 2021) continue to enhance the expressive capacity of saliency models.

In parallel, several studies have sought to enhance the biological plausibility of saliency models by incorporating principles from neuroscience (Wang, Li, and Chen 2021; Li, Chen, and Yu 2018). For instance, NeuroSaliency (Pan et al. 2022) simulates the response patterns of cortical regions such as V1 and IT to improve spatial selectivity. BioViNet (Zhang et al. 2022) further models temporal dynamics inspired by biological vision systems to support dynamic saliency prediction in videos. While these biologically driven models offer cognitive insights into attention mechanisms, they primarily focus on natural scenes and have not explored the connection between saliency prediction and the internal attention mechanisms of AIGC models.

Quality Assessment of AI-Generated Content

With the rapid advancement of image generation models, particularly diffusion-based architectures such as Stable Diffusion (Rombach et al. 2022), DALL·E (Ramesh et al. 2022), and Imagen (Saharia et al. 2022), evaluating the perceptual quality of AIGC-generated images has become an increasingly critical research problem. Conventional metrics such as PSNR and SSIM primarily assess low-level pixel-wise differences, but often fail to reflect semantic consistency or structural plausibility in generative outputs. To address these limitations, perceptual metrics such as LPIPS (Zhang et al. 2018a) and FID (Heusel et al. 2017) have been proposed, which evaluate feature-level similarity and distributional alignment within learned embedding spaces.

Beyond these, recent studies have explored leveraging attention information from generative models for quality assessment. For example, GazeGAN (Liu et al. 2020) guides the generation process using human gaze heatmaps to enhance perceptual focus, while AttentionDistill (Xu et al. 2022) incorporates cross-attention maps during generation to improve the discriminative capacity of quality evaluators. These works highlight the potential of internal attention maps as informative signals for perceptual evaluation, but an integrated framework for aligning human gaze with model attention remains underdeveloped.

Cross-Domain Knowledge Distillation

Knowledge distillation (KD) has become a widely adopted technique for model compression and transfer, especially in classification and detection tasks. The seminal work by Hinton et al. (Zhang et al. 2018b; Gupta et al. 2021; Shen et al. 2021) pioneered cross-modal distillation (Gupta et al. 2021), aiming to fuse complementary knowledge across domains and modalities. In saliency modeling, preliminary attempts to enhance model robustness have incorporated dual-source distillation; for example, SaliencyDistill (Jiang et al. 2022) proposes such a strategy. However, integrating biologically inspired attention mechanisms with internal attention maps of generative models as dual teachers for cross-domain alignment remains an unexplored direction. Building upon this motivation, our work proposes a cognitively inspired distillation strategy that unifies human gaze and AIGC attention patterns to enable human-AI cognitive consistency evaluation.

Method

Overview of the Framework

This work proposes a saliency prediction framework based on six brain-region liquid neuron modules as shown in Fig. 1, aiming to simulate the multi-stage dynamic processing mechanism of the human visual cortex and construct a biologically inspired human attention proxy model. The model first extracts multi-scale spatial and semantic features via a Transformer backbone visual encoder, serving as the perceptual basis for subsequent neuron modeling. The core components are six liquid neuron modules corresponding to key visual areas V1, V2, V4, MT, LIP, and FEF. Each module employs continuous-time dynamic modeling to simulate functional characteristics of these brain regions, including edge detection, boundary integration, color and shape recognition, motion direction perception, sparse attention allocation, and task modulation. The FEF module integrates higher-level features and behavioral control, acting as the key output stage for generating saliency maps.

Neuro-Inspired Visual Transformer

To effectively extract multi-scale spatial representations and capture high-level semantics from AIGC-generated images, we adopt a lightweight Transformer-based visual encoder as

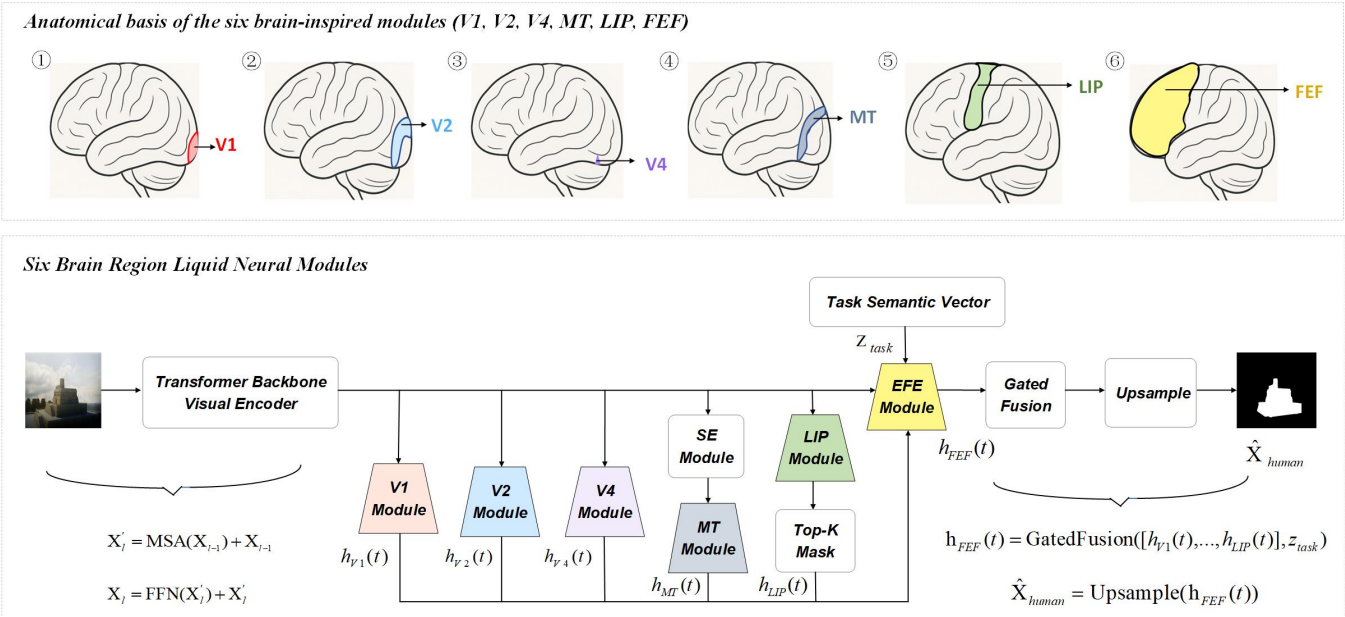

Figure 1: Overview of the saliency prediction pipeline with brain-inspired liquid neuron modules. Transformer-based encoders first extract visual features, which are then processed in parallel by six liquid neuron modules simulating cortical areas (V1, V2, V4, MT, LIP, FEF), each modeling distinct functional dynamics. The outputs are fused via a gated mechanism and upsampled to generate a biologically plausible saliency map aligned with human cognition.

the backbone of our perception module. Compared to convolutional networks, Transformers offer enhanced capacity for modeling long-range dependencies via global self-attention (Dosovitskiy et al. 2021; Touvron et al. 2021; Liu et al. 2021b), which is essential for understanding complex visual contexts in multimodal generation.

Given an input image $\mathbf{I} \in R^{3\times H\times W}$, we first apply patch embedding and positional encoding to transform it into a token sequence $\mathbf{X}_0$. This sequence is then passed through L stacked Transformer encoder layers, each composed of a multi-head self-attention (MSA) block and a feed-forward network (FFN):

$$\mathbf{X}_\ell^{'} = MSA(\mathbf{X}_{\ell-1}) + \mathbf{X}_{\ell-1}, \tag{1}$$

$$\mathbf{X}_\ell = FFN(\mathbf{X}_\ell^{'}) + \mathbf{X}_\ell^{'}. \tag{2}$$

This process yields hierarchical features $\{\mathbf{X}_1, \ldots, \mathbf{X}_L\}$, which are subsequently routed into brain-inspired modules to simulate perceptual attention at varying cognitive depths.

To further enhance dynamic attention modeling, we incorporate Liquid Neural Networks (LNNs) (Hasani et al. 2021; Grigsby, Hasani, and Rus 2023; Rus et al. 2023) into the system. LNNs are a class of neural models based on continuous-time dynamical systems, where each neuron evolves according to a parameterized differential equation. Unlike traditional static neurons, LNN units maintain internal states, enabling state-dependent and history-aware behavior.

This biologically plausible mechanism is particularly suitable for mimicking the temporal integration characteristics of human vision. LNNs naturally capture context-dependent attention shifts over time and exhibit strong robustness and generalization, especially under distributional shifts (Hasani et al. 2022). These properties make them a compelling choice for cognitively aligned saliency and attention modeling in generative settings.

Six Brain-Region Liquid Neuron Modules

Liquid Neuron Dynamic Modeling Foundation To simulate the temporal dynamics of the human neural system, we introduce the design concept of Liquid Time-Constant Networks (LTC) (Hasani et al. 2021), which models neuron state evolution using input-driven continuous-time differential equations. Unlike traditional neurons (RNN, GRU, LSTM), LTC neurons learn time constants as dynamic gates, better reflecting biological neuron memory and response mechanisms.

In visual perception tasks, LTC effectively models dynamic responses to blurred edges, texture transitions, and latent motion trends in static images by adapting across multiple time scales, alleviating response delays or forgetting. Thus, we build neural modules simulating multi-brain-region temporal dynamics based on LTC. The state change of each brain region module r follows:

$$\frac{d\mathbf{h}_r(t)}{dt} = -\frac{1}{\tau_r(\mathbf{x}_r(t))}\mathbf{h}_r(t) + f_r(\mathbf{x}_r(t), \mathbf{h}_r(t)), \tag{3}$$

where $\mathbf{h}_r(t)$ denotes the hidden state of region r at time t; $\mathbf{x}_r(t)$ is the input features to region r from the encoder or other modules; $\tau_r(\mathbf{x}_r(t))$ is the input-dependent dynamic time constant controlling response speed and forgetting; $f_r(\cdot)$ is a nonlinear update function involving convolution, gating, and activation.

This structure is discretized by numerical integration for forward propagation, enabling dynamic evolution of neuron states, enhancing perception of blurred areas, complex textures, and latent dynamics in images, consistent with continuous and dynamic human attention characteristics.

V1 Module: Early Edge Detection V1, the first visual cortex stage, handles local edge detection, orientation selectivity, and spatial frequency encoding, with fast neuron responses and small time constants sensitive to high-frequency information. Fixed small-time-constant liquid neurons are incorporated to emulate the rapid response and sparse activation behavior characteristic of V1:

$$\frac{d\mathbf{h}_{V1}(t)}{dt} = -\frac{1}{\tau_{V1}}\mathbf{h}_{V1}(t) + \tanh\big(\mathbf{W}_{V1} * \mathbf{x}(t)\big), \quad (4)$$

where $\mathbf{x}(t) \in R^{C_{in}\times H\times W}$ is the initial feature map from the Transformer backbone; $\mathbf{W}_{V1} \in R^{C_{out}\times C_{in}\times k\times k}$ is a convolution kernel simulating simple cell filters; τ_{V1} is fixed small to represent rapid response; $\tanh(\cdot)$ ensures nonlinearity and suppresses excessive responses.

This module rapidly captures local edges and details, serving as the foundation for saliency detection, ensuring spatial resolution and detail accuracy.

V2 Module: Boundary and Blur Integration V2 integrates edges from V1, strengthening boundary perception and processing blurry and gradient regions. We design input-dependent learnable time constants enabling dynamic response rate adjustment to image complexity:

$$\tau_{V2}(\mathbf{x}, \mathbf{h}_{V2}) = Softplus\big(\mathbf{W}_\tau * \mathbf{x} + \mathbf{U}_\tau * \mathbf{h}_{V2}\big), \quad (5)$$

$$\begin{aligned}\tfrac{d\mathbf{h}_{V2}(t)}{dt} = \; & -\tfrac{1}{\tau_{V2}(\mathbf{x},\mathbf{h}_{V2})}\mathbf{h}_{V2}(t) \\ & +\sigma\big(\mathbf{W} * \mathbf{x}(t) + \mathbf{U} * \mathbf{h}_{V2}(t)\big)\end{aligned} \quad (6)$$

where $\mathbf{W}_\tau, \mathbf{U}_\tau$ modulate the time constants, $\mathbf{W}, \mathbf{U}$ are convolution kernels, and $\sigma(\cdot)$ is an activation function (e.g., ReLU or sigmoid).

Dynamic response adjustment helps V2 naturally process blurred boundaries and gradients, reducing artifacts and fragments in saliency maps, enhancing visual continuity and structural plausibility.

V4 Module: Color and Shape Integration V4 is a high-level visual area responsive to color, shape, and region grouping, exhibiting multi-frequency oscillations. We simulate multi-frequency processing via parallel multi-time-constant mechanisms, dynamically weighting multiple frequency branches:

$$\mathbf{h}_{V4}(t) = \sum_{i=1}^{K} g_i(\mathbf{x}(t)) \cdot ODESolve_{\tau_i}(f_i, \mathbf{x}(t)), \quad (7)$$

where K is the number of frequency branches; τ_i are distinct time constants representing oscillation bands; $g_i(\mathbf{x})$ are softmax-normalized weights dynamically adjusting branch influence; $ODESolve_{\tau_i}$ is the state update corresponding to τ_i.

This mechanism enables concurrent focus on large-scale color regions and shape groups, improving regional coherence and semantic consistency in saliency maps.

MT Module: Motion Direction and Latent Trend MT is primarily responsible for processing motion direction and speed, exhibiting high sensitivity to visual motion cues (Born and Bradley 2005; Newsome and Pare 1989). While MT mainly responds to dynamic stimuli, in the context of static images, it can leverage SE modules (Hu, Shen, and Sun 2018) to capture latent directional and flow information, thereby enriching the spatial attention mechanism.

$$\frac{d\mathbf{h}_{MT}(t)}{dt} = -\frac{1}{\tau_{MT}}\mathbf{h}_{MT}(t) + SE(\mathbf{x}(t)) \odot (\mathbf{W} * \mathbf{x}(t)), \quad (8)$$

where SE$(\cdot)$ dynamically weights channels, τ_{MT} is fixed, and $\odot$ denotes channel-wise product.

This improves sensitivity to perspective lines and guiding arrangements, enriching spatial directionality and flow perception in saliency maps.

LIP Module: Sparse Attention Resource Allocation We employ a Top-K masking mechanism to enforce sparse focus on salient locations, filtering irrelevant information:

$$\mathbf{m}_{LIP}(t) = TopKMask(\mathbf{x}(t), k), \quad (9)$$

$$\begin{aligned}\tfrac{d\mathbf{h}_{LIP}(t)}{dt} = \; & -\tfrac{1}{\tau_{LIP}}\mathbf{m}_{LIP}(t) \odot (\mathbf{W} * \mathbf{x}(t)) \\ & +\mathbf{h}_{LIP}(t)\end{aligned} \quad (10)$$

where k controls selected salient points, and $\mathbf{m}_{LIP}(t)$ is the corresponding binary sparse mask.

This improves focus in saliency maps, avoids attention diffusion, and aligns attention distribution with human sparse cognitive patterns.

FEF Module: Task-Oriented Fusion FEF integrates visual information from multiple brain regions and task instructions, controlling attention, focus, and behavior decisions. It fuses the five regional states with a task semantic vector, dynamically modulating final attention:

$$\begin{aligned}HAMCI = \; & \gamma_1 \cdot IoU(\hat{\mathbf{X}}_{student}, \mathbf{S}^{ca}) \\ & +\gamma_2 \cdot KL(\hat{\mathbf{X}}_{student} \parallel \mathbf{S}^{ca}) \\ & +\gamma_3 \cdot \Big|\mathcal{H}(\hat{\mathbf{X}}_{student}) - \mathcal{H}(\mathbf{S}^{ca})\Big|\end{aligned} \quad (11)$$

$$\hat{\mathbf{X}}_{human} = Upsample(\mathbf{h}_{FEF}(t)), \quad (12)$$

where $\mathbf{z}_{task}$ is the task semantic embedding, and GatedFusion uses gating to weight brain region responses.

This closes the perception-to-action feedback loop, generating high-quality task-consistent saliency maps.

Cross-Domain Dual-Teacher Distillation

To bridge the cognitive gap between real and AIGC images, we propose a cross-domain dual-teacher distillation strategy that aligns attention in the student model across domains. Two complementary supervisory signals—one from human gaze heatmaps and the other from cross-attention maps of AIGC models—guide the student to learn human-like saliency perception while maintaining compatibility with generative attention patterns:

- **Teacher 1**: Trained on real images and fixation maps with six-region liquid neuron architecture, outputs highly consistent human attention $\mathbf{X}_{T1}$ supervising "human-like attention" learning;

- **Teacher 2**: Independently trained on AIGC images and their Cross-Attention maps, extracts internal generative attention features, outputting $\mathbf{X}_{T2}$ guiding "generation attention structure" alignment.

The complete training framework is illustrated in Fig. 2. The student takes AIGC images as input and aligns with the output from both teachers, preserving biologically inspired attention while adapting to AIGC structure preferences, thereby enabling cross-domain cognitive consistency modeling. The overall training process optimizes the joint loss:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{dyn} + \beta_1\mathcal{L}_{distill}^{(T1)} + \beta_2\mathcal{L}_{distill}^{(T2)}, \quad (13)$$

where each term is defined as:

- **Liquid Neuron Dynamics Regularization**:

$$\mathcal{L}_{dyn} = \sum_{t=1}^{T-1} \|\mathbf{h}(t+1) - \mathbf{h}(t)\|_2^2, \quad (14)$$

constraining the temporal smoothness of neuron states across brain regions;

- **Teacher 1 Distillation Loss**:

$$\mathcal{L}_{distill}^{(T1)} = KL\big(\hat{\mathbf{X}}_{student} \parallel \mathbf{X}_{T1}\big), \quad (15)$$

Maintaining a biologically inspired human attention structure;

- **Teacher 2 Distillation Loss**:

$$\mathcal{L}_{distill}^{(T2)} = \left\|\hat{\mathbf{X}}_{student} - \mathbf{X}_{T2}\right\|_2^2, \quad (16)$$

aligning with the spatial attention patterns of the generative model.

Hyperparameters α, β_1, β_2 balance losses and are tuned on validation data. This distillation enhances attention adaptation on real and generated images, providing an integrated prediction baseline for cognitive consistency evaluation.

Human-AI Consistency Index (HAMCI)

To quantify the cognitive matching between the predicted human attention map $\hat{\mathbf{X}}_{student}$ from the student model and the internal Cross-Attention map $\mathbf{S}^{ca}$ from the AIGC model, we propose the Human-AI Mutual Consistency Index (HAMCI). HAMCI systematically evaluates attention alignment on AIGC content by considering spatial overlap, distribution consistency, and sparsity differences:

$$\begin{aligned} HAMCI = {} & \gamma_1 \cdot IoU(\hat{\mathbf{X}}_{student}, \mathbf{S}^{ca}) \\ & + \gamma_2 \cdot KL(\hat{\mathbf{X}}_{student} \parallel \mathbf{S}^{ca}) \\ & + \gamma_3 \cdot \left|\mathcal{H}(\hat{\mathbf{X}}_{student}) - \mathcal{H}(\mathbf{S}^{ca})\right| \end{aligned} \quad (17)$$

where IoU measures spatial intersection of salient regions; KL measures distributional difference; $\mathcal{H}(\cdot)$ is Shannon entropy; γ_i are weighting parameters determined by cross-validation.

Compared with traditional saliency metrics, HAMCI requires no manual annotation, directly comparing generated attention maps with AIGC internal mechanisms, thus providing a integrated evaluation of model interpretability and cognitive rationality.

Experiments

Datasets and Evaluation Metrics

To fully reflect human visual attention and the AIGC generation mechanism, two subdomain datasets are constructed and used:

- **REAL-EYE (Real Image Domain)**: This dataset is sourced from SALICON (Jiang et al. 2015), MIT1003 (Judd et al. 2009), and CAT2000 (Borji, Sihite, and Itti 2015), comprising approximately 12,000 natural scene images and corresponding human eye-tracking heatmaps. The dataset is split into 8,000 training images and 4,000 testing images. It is used to train a teacher model based on real gaze data (Teacher 1), providing the student model with authentic visual attention supervision signals.
- **AIGC-EYE (Generated Image Domain)**: Collected from Stable Diffusion 2.1 (Rombach et al. 2022), DALL·E 3 (Ramesh et al. 2022), and Midjourney v5.2 (Midjourney 2023), this dataset contains 6,000 high-quality generated images and their corresponding Cross-Attention Maps. It includes two typical scene categories: central single-object and multi-object complex scenes. The training set and test set contain 4,200 and 1,800 images, respectively. This dataset trains a teacher model based on the generative model's attention mechanism (Teacher 2) to capture the internal focus patterns of generated content.

We evaluate performance using standard saliency metrics (NSS (Peters et al. 2005), SIM (Judd, Durand, and Torralba 2012), and CC (Riche et al. 2013)) and introduce HAMCI to measure the spatial and semantic consistency between predicted saliency and cross-attention maps. Lower HAMCI scores indicate better human–model alignment and higher cognitive consistency.

Experimental Setup

The model is implemented with PyTorch 1.13.1, and both training and testing are conducted on a workstation equipped with an AMD Ryzen 7 7800X3D CPU and an NVIDIA RTX 3090 GPU (24GB VRAM). Training uses the AdamW optimizer with an initial learning rate of 1×10^{-4}, combined with a learning rate scheduling strategy for gradual decay. The batch size is set to 24, and the dropout rate is 0.1. The liquid neuron modules adopt the fourth-order Runge-Kutta numerical integration method to ensure the accuracy and stability of temporal dynamics computation. The total training epochs are 200, with early stopping based on validation performance to prevent overfitting.

Model Comparison Experiments

To evaluate the effectiveness of our model, we compare it with six representative saliency prediction methods, including weakly supervised, CNN-based, and Transformer-based approaches. UNISAL (Kümmerer, Wallis, and Bethge 2020), representing weak supervision; DeepGaze II (Kümmerer, Wallis, and Bethge 2016) and SALICON (Huang et al. 2015) as CNN-based models; EML-NET (Jia and Bruce 2020) which fuses multiple CNN

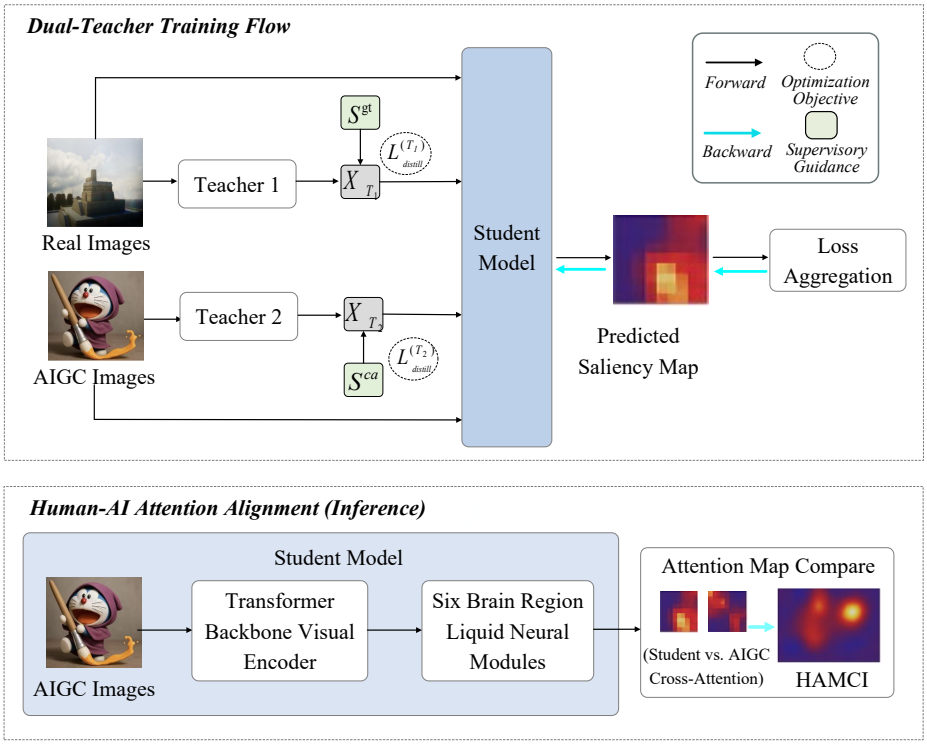

Figure 2: Workflow of the student model training process.

Model	NSS ↑	SIM ↑	CC ↑	HAMCI ↓
Ours	**2.650±0.001**	**0.710±0.002**	**0.670±0.000**	**0.290±0.001**
UNISAL (Kümmerer, Wallis, and Bethge 2020)	2.100±0.003	0.640±0.002	0.570±0.004	0.430±0.003
DeepGaze II (Kümmerer, Wallis, and Bethge 2016)	2.290±0.002	0.660±0.003	0.600±0.002	0.390±0.004
SALICON (Huang et al. 2015)	2.240±0.001	0.650±0.004	0.590±0.003	0.410±0.003
TASED-Net (Min and Korhonen 2019)	2.330±0.002	0.670±0.001	0.620±0.003	0.380±0.002
EML-NET (Jia and Bruce 2020)	2.390±0.001	0.680±0.003	0.630±0.002	0.370±0.004
TransSalNet (Liu et al. 2021a)	2.460±0.004	0.690±0.002	0.640±0.001	0.350±0.003

Table 1: Comparison of saliency models on the AIGC-EYE dataset. All values are reported with ± standard deviation. Metrics include NSS, SIM, CC, and HAMCI. The best results are highlighted in bold.

features; and TASED-Net (Min and Korhonen 2019) and TransSalNet (Liu et al. 2021a) as Transformer-based approaches. All models are tested on the AIGC-EYE dataset to assess their performance in modeling human-AI cognitive consistency.

As shown in Table 1, our model consistently outperforms six mainstream state-of-the-art methods across all metrics. Compared to the strongest baseline, TransSalNet, it achieves an NSS improvement of 0.19, with SIM and CC increasing by 0.02 and 0.03, respectively, indicating enhanced spatial precision and structural alignment. The HAMCI score decreases by 0.06, reflecting improved alignment with the cross-attention patterns of AIGC models and stronger human-AI cognitive consistency.

Visualization results in Fig. 3 further demonstrate the superiority of our model across diverse scenarios. In complex multi-object scenes, traditional models often produce scattered or blurry saliency maps, whereas our model accurately localizes key objects with clear structural saliency. In central single-object scenes, it effectively avoids excessive center bias and preserves object boundaries, exhibiting closer alignment with human gaze. These results underscore the robustness, perceptual accuracy, and cross-domain generalization capability of our approach.

Ablation Study on Brain Modules

To systematically evaluate the independent contributions of the six brain-region liquid neuron modules (V1, V2, V4, MT, LIP, FEF) to model performance, ablation experiments progressively introducing these modules are designed. The results are as follows:

As shown in Table 2, the progressive integration of

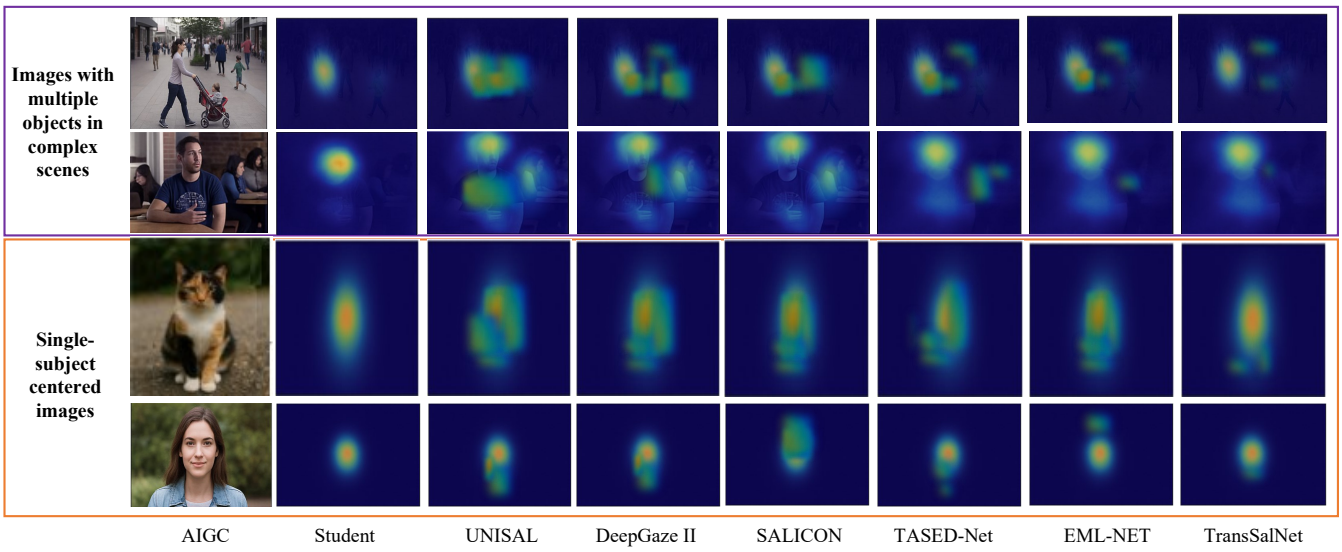

Figure 3: Visual Comparison of Saliency Prediction and Human-AI Consistency.

Configuration	NSS ↑	SIM ↑	CC ↑	HAMCI ↓
-	2.01	0.60	0.55	0.46
+V1+V2	2.26	0.64	0.59	0.40
+V4	2.38	0.67	0.61	0.36
+MT	2.47	0.68	0.63	0.34
+LIP	2.54	0.69	0.65	0.31
+FEF (Ours)	**2.65**	**0.71**	**0.67**	**0.29**

Table 2: Performance evaluation with progressive addition of brain modules.

six brain-region liquid neuron modules (V1, V2, V4, MT, LIP, FEF) consistently improves saliency prediction performance. Early modules (V1, V2) significantly enhance spatial accuracy and distribution alignment (NSS↑ from 2.01 to 2.26; HAMCI↓ to 0.40). V4 improves structural modeling (CC↑ to 0.61), while MT captures motion-related cues. LIP further sharpens spatial focus and target selection. With FEF integrated, the model achieves optimal results across all metrics, with HAMCI reaching 0.29, demonstrating the effectiveness of multi-region collaboration in modeling human-AI cognitive consistency.

Dual-Teacher Distillation Experiment Analysis

To evaluate the impact of the distillation mechanism on model performance and human-AI cognitive alignment, we compare four training strategies: NoDist (no distillation), Teacher 1 (real gaze only), Teacher 2 (AIGC attention only), and DT-Fuse (dual-teacher fusion). Results are shown in Table 3.

Experimental results confirm the effectiveness of distillation strategies in improving both saliency performance and human-AI cognitive alignment. Compared to NoDist, both Teacher 1 and Teacher 2 significantly enhance NSS, SIM, and CC scores. Teacher 1 better captures human-like attention, while Teacher 2 improves adaptation to generative structures. The dual-teacher strategy (DT-Fuse) combines their strengths and achieves the best overall performance, including the lowest HAMCI score, indicating superior cognitive consistency and cross-domain generalization.

Strategy	NSS ↑	SIM ↑	CC ↑	HAMCI ↓
NoDist	2.43	0.67	0.61	0.37
Teacher1	2.53	0.69	0.63	0.32
Teacher2	2.49	0.68	0.62	0.34
DT-Fuse	**2.65**	**0.71**	**0.67**	**0.29**

Table 3: Distillation strategies comparison.

Conclusion

This study presents a brain-inspired saliency prediction framework that combines functional modeling of six visual cortical areas with a dual-teacher distillation strategy. The model integrates neural dynamics to simulate hierarchical visual processing and aligns human gaze with AIGC model attention for cognitively consistent saliency prediction. Extensive experiments on cross-domain datasets demonstrate superior performance across standard saliency metrics and the proposed HAMCI index, validating the effectiveness of the framework in aligning human and AI attention. Future work will extend this framework to dynamic scenes involving temporal saliency and attentional shifts, and explore its applicability to broader tasks such as generative content evaluation and cognitive health analysis, aiming to promote more interpretable and human-aligned AIGC systems.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62571215 and 62271226.

References

Binkowski, M.; Sutherland, D.; Arbel, M.; and Gretton, A. 2018. Demystifying MMD GANs. In *Proc. Int. Conf. Learning Representations (ICLR)*.

Borji, A.; Sihite, D.; and Itti, L. 2015. CAT2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*.

Born, R. T.; and Bradley, D. C. 2005. Visual processing of motion. *Neuron*, 55(2): 179–206.

Cheng, M.-M.; Mitra, N. J.; Huang, X.; Torr, P. H.; and Hu, S.-M. 2015. Deep contrast learning for salient object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 478–487.

Dhariwal, P.; and Nichol, A. Q. 2021. Diffusion Models Beat GANs on Image Synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 12873–12883.

Fang, X.; Wang, S.; Lv, X.; and Yan, J. 2024. PCQA: A Strong Baseline for AIGC Quality Assessment Based on Prompt Condition. In *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, 1–6.

Grigsby, A.; Hasani, R.; and Rus, D. 2023. Liquid Neural Networks for Vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Gupta, A.; Das, S.; Tarlow, D.; and Malik, J. 2021. Cross-Modal Knowledge Distillation for Few-Shot Action Recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 11487–11496.

Hasani, R.; Lechner, M.; Amini, A.; and Rus, D. 2022. Closed-form Continuous-time Neural Networks. *Nature Machine Intelligence*, 4(7): 410–421.

Hasani, R.; Lechner, M.; Amini, A.; Rus, D.; Grosu, R.; and Rus, D. 2021. Liquid Time-Constant Networks. In *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proc. Advances in Neural Inf. Process. Syst. (NeurIPS)*, 6626–6637.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huang, X.; Shen, C.; Boix, X.; and Zhao, Q. 2015. SALICON: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 1072–1080.

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11): 1254–1259.

Jia, P.; and Bruce, N. D. B. 2020. EML-NET: An expandable multi-layer network for saliency prediction. *IEEE Transactions on Image Processing*, 29: 5006–5018.

Jiang, J.; Zhang, L.; Xie, W.; Chen, Z.; and Wang, G. 2022. SaliencyDistill: Dual-Teacher Knowledge Distillation for Robust Saliency Prediction. In *Proc. AAAI Conf. Artif. Intell.*

Jiang, M.; Huang, S.; Duan, J.; and Zhao, Q. 2015. SALICON: Saliency in context. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 1072–1080.

Judd, T.; Durand, F.; and Torralba, A. 2012. Benchmarking saliency models on natural images. In *Proc. IEEE European Conf. Computer Vision (ECCV)*, 547–561. Springer.

Judd, T.; Ehinger, K. A.; Durand, F.; and Torralba, A. 2009. Learning to predict where humans look. *IEEE International Conference on Computer Vision (ICCV)*, 2106–2113.

Kümmerer, M.; Wallis, T. S.; and Bethge, M. 2016. DeepGaze II: Reading fixations from deep features trained on object recognition. In *arXiv preprint arXiv:1610.01563*.

Kümmerer, M.; Wallis, T. S.; and Bethge, M. 2020. UNISAL: A unified model for predicting saliency and scanpaths. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 14594–14604.

Li, W.; Chen, L.; and Yu, X. 2018. Neuromorphic Visual Attention System for Saliency Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10): 2707–2718.

Li, Z.; Yang, C.; He, X.; Zhang, H.; and Gao, W. 2019. Attention-based no-reference image quality assessment. *IEEE Transactions on Multimedia*, 21(11): 2965–2979.

Linardos, A.; Evangelopoulos, G.; and Potamianos, A. 2021. TASED-Net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6619–6628.

Liu, Y.; Pan, J.; Xu, H.; Wang, W.; and Xu, Q. 2021a. TransSalNet: Visual saliency prediction with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4433–4442.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Liu, Z.; Xiao, T.; Liu, W.; Li, Y.; Wang, H.; Wang, C.; and Jia, J. 2020. GazeGAN: A Generative Adversarial Model for Gaze Prediction in the Wild. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 528–544. Springer.

Midjourney. 2023. Midjourney V5.2 Update Announcement. https://docs.midjourney.com/docs/release-notes. Accessed: 2025-07-10.

Min, J.; and Korhonen, J. 2019. TASED-Net: Temporally attentive salient edge detection network. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2506–2510.

Newsome, W. T.; and Pare, E. B. 1989. Neuronal correlates of a perceptual decision. *Nature*, 341(6237): 52–54.

Pan, J.; Canton-Ferrer, C.; McGuinness, K.; and O'Connor, N. E. 2021. SalGAN: Visual saliency prediction with adversarial networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1–9.

Pan, J.; Liu, Y.; Xu, H.; Wang, W.; and Xu, Q. 2022. NeuroSaliency: Simulating cortical visual processing for human-like attention prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Peters, R. J.; Iyer, A.; Itti, L.; and Koch, C. 2005. Components of bottom-up gaze allocation in natural images. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, volume 1, 23–30. IEEE.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. *Proc. Int. Conf. Mach. Learn. (ICML)*.

Riche, N.; Mancas, M.; Le Meur, O.; Duvinage, M.; Gosselin, B.; and Zhang, D. 2013. Saliency and human fixations: State-of-the-art and study of comparison metrics. *IEEE Transactions on Image Processing*, 22(1): 55–69.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 10684–10695.

Rus, D.; Hasani, R.; Lechner, M.; and Amini, A. 2023. Liquid Neural Networks. *Communications of the ACM*, 66(4): 30–32.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Salimans, S.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Proc. Int. Conf. Machine Learning (ICML)*, 17468–17491.

Shen, Y.; Xu, H.; Yu, G.; and Luo, P. 2021. Multi-teacher knowledge distillation for object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9): 4003–4014.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training Data-efficient Image Transformers & Distillation through Attention. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Wang, M.; Li, Q.; and Chen, Z. 2021. BioViNet: A biologically inspired dynamic neural network for visual saliency prediction. In *Proc. IEEE Int. Conf. Computer Vision Workshops (ICCV Workshops)*, 1234–1243.

Wang, X.; Liu, J.; and Liu, J. 2019. Saliency-Guided No-Reference Image Quality Assessment. *IEEE Transactions on Multimedia*, 21(12): 3012–3024.

Xu, J.; Zhang, Z.; Liu, J.; and Wang, Z. 2022. AttentionDistill: Distilling Cross-Attention for No-Reference Image Quality Assessment. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 331–347. Springer.

Yang, L.; Zhao, X.; and Sun, J. 2023. CogniT-IQA: Cognitive-Inspired No-Reference Image Quality Assessment Integrating Human Visual Attention. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, H.; Li, Z.; Li, Q.; and Zhao, Q. 2022. BioViNet: Biologically inspired video saliency prediction via neural dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1749–1758.

Zhang, L.; Zhang, H.; and Mou, X. 2020. Saliency-Aware Image Quality Assessment via Multi-Task Learning. *IEEE Transactions on Image Processing*, 29: 7994–8007.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 586–595.

Zhang, Y.; Xiang, T.; Li, T. M.; and Hospedales, T. M. 2018b. Deep mutual learning. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 4320–4328.