SPURIOUS EQUILIBRIUM IN SEGMENTATION MODELS AND RECURRENT PROCESSING IN HUMAN VISION

Kailas Dayanandan
Department of Electrical Engineering
Indian Institute of Technology
New Delhi, India
kailasd@gmail.com

Brejesh Lall
Department of Electrical Engineering
Indian Institute of Technology
New Delhi, India
brejesh@ee.iitd.ac.in

ABSTRACT

Iterative decision-making has been widely studied in human cognition and is recognized for its energy efficiency and suitability for biological computations. In contrast, instance segmentation models adopt strategies that diverge from human vision, each presenting unique strengths and limitations. In this paper, we examine the grouping problem in segmentation models and demonstrate that iterative recurrent processing facilitates the identification of diverse solutions and can enhance grouping capabilities. Our experiments further reveal that recurrent processing accelerates convergence and can generate diverse solutions that can help mitigate suboptimal spurious minima. Our work focuses on confounding cases, which have become increasingly relevant as systems are increasingly deployed in safety-critical environments.

1 INTRODUCTION

Human vision is robust generalizable and accurate compared to deep learning models. Electrophysiological research provides empirical support for the concept of dual thinking van Bergen & Kriegeskorte (2020); Chen et al. (2021), characterized by an initial rapid feed-forward stage followed by a slower, iterative refinement process VanRullen (2007). Dual-processing framework has also been substantiated by studies on visual perception Dayanandan et al. (2024); Daniel (2017), which also highlight spurious behaviors in instance segmentation models. While research on spurious correlations has predominantly focused on classification tasks Ghaznavi et al. (2024), similar challenges in instance segmentation remain relatively under-explored despite emerging findings.

Recurrent processing can efficiently perform complex computations in environments with energy and spatial constraints in human vision Kreiman & Serre (2020); however, current systems ignore them in the encoding process due to their slower operation with attention Oren et al. (2024). In this study, we investigate the underlying sources of some of the spurious errors in segmentation models Dayanandan et al. (2024) and examine the usefulness of recurrent processing in human vision for segmentation. Our approach aligns with recent methodologies focusing on generating better candidate solutions and refining them to determine the most optimal outcome Wang et al. (2023). Our key contributions include (a) a theoretical analysis of training objectives and the equilibrium they attain and a comparison with strategies in human visual processing and (b) an investigation of the role of recurrent processing alongside feed-forward networks in human vision.

2 Related Work

Deep learning models for classification identify statistical properties that are important in the training dataset that also includes correlations and are not causal which comes across as spurious features or shortcuts Izmailov et al. (2022). Recent works have looked into addressing these spurious effects in classification models Noohdani et al. (2024); Ghaznavi et al. (2024). The recent work argues that human vision is also prone to shortcuts and compares the shortcomings of human vision and deep learning based segmentation models Dayanandan et al. (2024). While incorporating sub-components can help in better performance on some of the errors to be related absence of understanding of sub-structures, the errors related to grouping conflicts are not explored.

Recurrent computation has shown to be both computationally efficient and highly effective across various tasks Wang et al. (2019); Tomar et al. (2022). Recurrent neural networks modeled after human visual processing have been shown to exhibit perceptual phenomena such as the orientation-tilt illusion Linsley et al. (2020) and the perception of illusory shapes Pang et al. (2021) akin to human vision. In recurrent processing, outputs from previous states can progressively transform the original image into a simpler representation, making it easier for the model to solve in later steps. They can help identify intersecting lines and complete patterns in deep learning models Linsley et al. (2018; 2020) and amodal completion in human vision Tang et al. (2018). While studies have highlighted the inherent limitations of fully connected networks Abbe & Boix-Adsera (2022); Abbe et al. (2022), their usefulness for generating diverse samples in human vision has not been explored especially in context of dual thinking.

3 Methods

(a) (b) (c) (d) (e)

Figure 1: (a) Grouping problem (b) Example of grouping conflict of common region from Human Confusion Dataset Dayanandan et al. (2024) (c) Segmentation model output with the confounding region (leg of the man) to be part of both instances (man and child) (d,e) Merging of components from different instances due to absence of understanding sub-structures is analyzed in existing work.

Instance segmentation can be formally defined as partitioning an image into mutually exclusive pixel subsets, each corresponding to a distinct object instance. However, our previous study found that this criterion is not always strictly followed despite its fundamental logical basis Dayanandan et al. (2024). This paper investigates how spurious minima contribute to such inconsistencies, particularly in confounding cases. Let us consider two different regions of size P_1 and P_2, and a common region b with two possible valid groupings (P_1, b) and (b, P_2). We investigate the best possible case for this grouping by splitting the common region b. In this case, we consider the ratio $x \in [0, 1]$, that represents the split in the common region (Fig.1a). We set that bx is assigned to P_1 and $b(1-x)$ is assigned to P_2, so that b is completely assigned between them. In case of real world segmentation, as sub-components cannot be split (for e.g. in fig.1b the leg can be assigned to man or child, but not both), the leg in , there could be two ground truth possibilities with b completely assigned to either P_1 or P_2.

3.1 Theoretical Contributions

In this section, we consider feed-forward deep learning models, there is no iterative procedure and there is only a single decision made. We assume a general case where the prior probability of two ground truths (P_1, b) and (b, P_2) are different and later analyze case for an equal probability for both ground truths.

Theorem 3.1. *If there are two regions P_1 and P_2 and a common region b, such that (P_1, b) and (b, P_2) are valid groupings with a prior probability of m and n for them being correct. Let x be the ratio of pixels assigned to P_1, then the average IoU score is convex for $x \in [0, 1]$.*

The average IoU in the general case can be written as below.

$$I_G(x) = m\left\{\frac{P_1 + xb}{P_1 + b} + \frac{P_2}{P_2 + (1-x)b}\right\} + n\left\{\frac{P_1}{P_1 + xb} + \frac{P_2 + (1-x)b}{P_2 + b}\right\} \tag{1}$$

$$I'_G(x) = \frac{P_1 bm}{(b(1-x) + P_1)^2} + \frac{bm}{P_1 + b} - \frac{P_2 bn}{(P_2 + xb)^2} - \frac{bn}{P_2 + b} \tag{2}$$

$$I''_G(x) = \frac{2P_1 b^2 m}{(b(1-x) + P_1)^3} + \frac{2P_2 b^2 n}{(bx + P_2)^3} \tag{3}$$

Since x is a fractions with values in $[0, 1]$, $1 - x$ is also positive. P_1, P_2, b are area and m, n are probabilities, all the terms in equation 3 are positive hence the second differential is positive for x in $[0, 1]$. This implies that the function $I_G(x)$ is convex for x in $[0, 1]$, and there is only one minima even for a general case, with a possibility of one or two maximas at the ends of the interval. The above result also implies that the maximum average IoU will lie at the ends of this interval ($x = 0$ or $x = 1$) and any split in common region would lead to sub-optimal score. The maxima will depend on m (for either of scenarios $m + n = 1$) and P_1, P_2 and b. From equation 1, we obtain

$$I_G(x = 0) = m\left\{\frac{P_1}{P_1 + b} + \frac{P_2}{P_2 + b}\right\} + 2n \tag{4}$$

$$I_G(x = 1) = 2m + n\left\{\frac{P_1}{P_1 + b} + \frac{P_2}{P_2 + b}\right\} \tag{5}$$

The value of x for the maximum IoU is either $x = 0$ or $x = 1$, which can be selected from the maximum of $I_G(x = 0)$ and $I_G(x = 1)$ above and any other value of x is sub-optimal.

Corollary 3.2. *If there are two individual regions of same size P and a common region b, such that (P, b) and (b, P) are valid groupings with a prior probability of m and n for them being correct. Let x be the ratio of pixels assigned to P, then the maximum average IoU will be at $x = 0$ if $m > n$ and at $x = 1$ if $m < n$.*

We should choose $x = 0$ when $I_G(x = 0) > I_G(x = 1)$. Setting $P_1 = P_2 = P$ in eq. 4 and 5, this implies $\left\{\frac{mP}{P+b} + n\right\} > \left\{\frac{nP}{P+b} + m\right\}$. Multiplying by $(P + b)$ and evaluating implies that, we should choose $x = 0$ when $nb > mb$ or $n > m$. Hence we should choose $x = 0$ (completely assign to right instance fig.1a) when $n > m$ or we should choose the more likely combination.

Theorem 3.3. *If there are two regions P_1 and P_2 such that $Area(P_1) = Area(P_2)$ and a common region b, such that (P_1, b) and (b, P_2) are valid groupings with equal probability, splitting common region by half [$(P_1, b/2)$, $(b/2, P_2)$] gives minimum (worst) average IoU score, and assigning common region to one of [(P_1, b), (P_2)] or [(P_1), (b, P_2)] gives maximum (best) average IoU score for single stage prediction.*

Setting $P_1 = P_2 = P$ and $m = n = \dfrac{1}{2}$ in eq.1 gives

$$I(x) = 0.5\left\{\frac{P + xb}{P + b} + \frac{P}{P + (1-x)b}\right\} + 0.5\left\{\frac{P}{P + xb} + \frac{P + (1-x)b}{P + b}\right\} \tag{6}$$

However, we can directly set $P_1 = P_2 = P$ and $m = n = \frac{1}{2}$ in eq.2 gives

$$I'(x) = 0.5\left\{ \frac{Pb}{(b\cdot(1-x)+P)^2} - \frac{Pb}{(bx+P)^2} \right\} \tag{7}$$

The roots for the equation $I'(x) = 0$ is $x = \frac{1}{2}$. Differentiating the $I'(x)$ and setting $x = \frac{1}{2}$

$$I''(x) = \frac{Pb^2}{(bx+P)^3} + \frac{Pb^2}{(b\cdot(1-x)+P)^3} \tag{8}$$

$$I''(x = \frac{1}{2}) = \frac{2Pb^2}{(\frac{b}{2}+P)^3} \quad > 0 \quad as \quad P, b > 0 \tag{9}$$

$I(x)$ in convex for x in [0,1] as $I''(x) > 0$ (eq. 8), Further $I(x)$ has minima at $x = \frac{1}{2}$.This implies that $I(x)$ increases in both directions with maxima at either of the edges. The value of $I(x)$ (eq. 10) at the edges are same, hence maxima exists at both $x = 0$ and $x = 1$.

$$I_{Max}(x) = I(x = 0) = I(x = 1) = \frac{P}{P+b} + 1 \tag{10}$$

Theorem shows that assigning completely to one instance maximizes average IoU score. This also implies that the maximum average IoU lie at the ends of this interval ($x = 0$ or $x = 1$) and any split in common region would lead to sub-optimal score.

Corollary 3.4. *If there are two regions P_1 and P_2 such that $Area(P_1) = Area(P_2)$ and a common region b, such that (P_1, b) and (b, P_2) are valid groupings with equal probability, then the IoU average score is symmetric around $x = \frac{1}{2}$ for x in [0,1].*

This is intuitive as the values of IoU averaged across instances and scenarios. We can show that $I(x = \frac{1}{2} + y) = I(x = \frac{1}{2} - y)$, hence IoU is symmetric around $x = \frac{1}{2}$ for x in $[0, 1]$.

$$I(x = \frac{1}{2} + y) = I(x = \frac{1}{2} - y) = 0.5\left\{ 1 + \frac{P}{P + (\frac{1}{2} - y)b} + \frac{P}{P + (\frac{1}{2} + y)b} \right\} \tag{11}$$

Theorem 3.5. *If there are two regions P_1 and P_2 such that $Area(P_1) = Area(P_2)$ and a common region b, such that (P_1, b) and (b, P_2) are valid groupings with equal probability, then the average IoU score for assigning common region to one of the instance is equal to the score of assigning it completely to both instances ($\big[(P_1, b), (P_2, b)\big]$).*

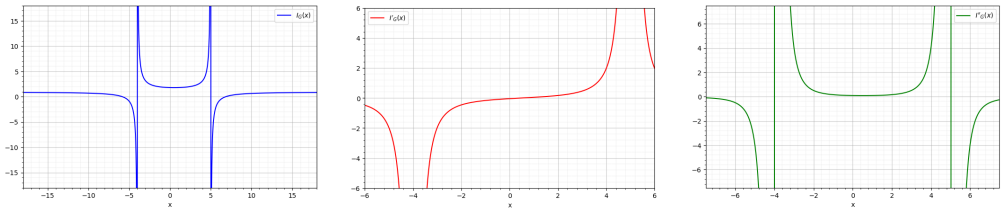
$$I_D(x) = 0.5\left\{ \frac{P+b}{P+b} + \frac{P}{P+b} \right\} + 0.5\left\{ \frac{P}{P+b} + \frac{P+b}{P+b} \right\} = \frac{P}{P+b} + 1 = I_{Max}(x) \tag{12}$$

The assignment $\big[(P_1, b), (b, P_2)\big]$ is preferred by deep learning models, as it gives maximum average IoU score for both ground truths (eq.10 and 12), though it will be a spurious equilibrium not which is partially wrong in both scenarios.

Observations : We can observe that the function $I_G(x)$ is convex for x in $[0, 1]$ in Fig.2a and the second differential is positive for x in $[0, 1]$ in Fig.2c as in Theorem 3.1 implying that the there is only one minima, with a possibility of one or two maximas at the ends of the interval.

4 EXPERIMENTS

We use a simulation to show the theoretical results. We consider two shapes similar to dumbbell that are overlapped and the information is insufficient to correctly identify the correct assignment ($m = \frac{1}{2}$). This is similar to the coarse-to-fine approach in human vision, where the initial inference is made on the coarser image with lesser information.



(a) $I_G(x)$ (b) $I'_G(x)$ (c) $I''_G(x)$

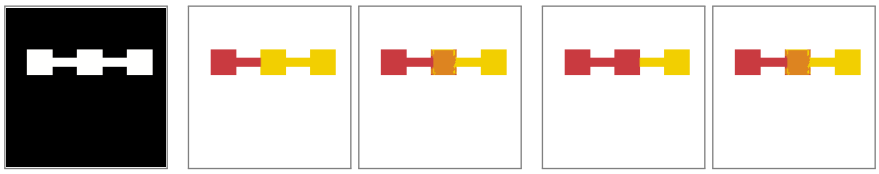
Figure 2: Plot for $P_1 = P_2 = 4$ and $b = 1$ for different values of x for m=0.5. There are no discontinuities in the interval in $x \in [0, 1]$

4.1 Simulations

We consider three implementations (a) single stage decision making similar to deep learning, where one single decision is made (b) Two decisions are made simultaneously (c) Two decisions are made sequentially one after the other with inputs from the previous decision. We use inputs containing a common region center square patch which can be assigned to either left of right instances shown in red and yellow (fig.4). We use a differential form of maximum function to find the maximum score of two decisions in second and third setting.

4.1.1 Single Stage Decision Making

We use a UNet model for our experiment. We observed that the common region is split between instances during intermediate steps in training, and the common region is assigned to both instances after training over a significant number of epochs. This results in the deep learning model being partially wrong in both ground truth scenarios as in Theorem 3.5. In human vision, the common region is assigned to one instance and evaluated to ensure correctness iteratively, and assigning to only one randomly can be completely correct 50% of the time with maximum average IoU score or the optimum value as in Theorem 3.5.



(a) Input (b) Ground Truth and Prediction (c) Ground Truth and Prediction

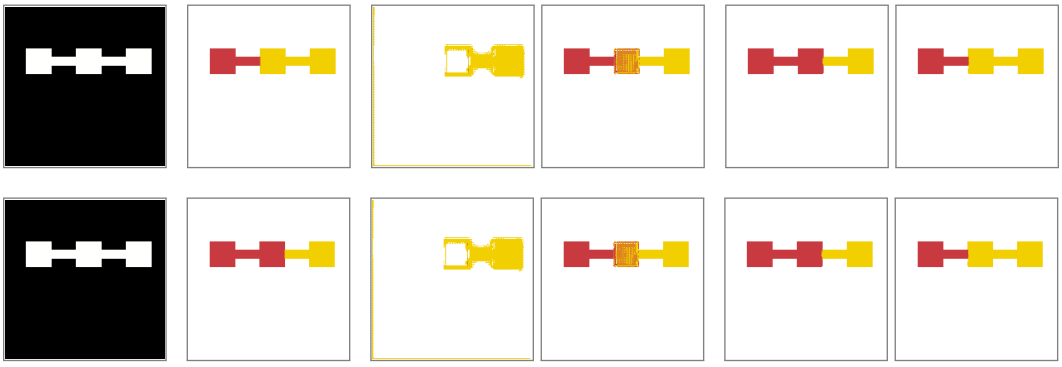
Figure 3: Single prediction in deep learning models for different ground truths. We can observe that the common region is in orange as it predicted for both left and right instances or the spurious equilibrium and learned in 10 epoch and training for 100 epochs did not change the behavior.

Observation : The single stage decision making is similar to the general setting in deep learning applications which optimizes the output for all samples or for all possible ground truths. This results in the choice of the spurious equilibrium. The model chooses an optimum based on it's training criteria or loss function resulting in spurious equilibrium that will never be fully correct and reproduces output combination that is not present in the training set.

4.1.2 Dual Decisions

We extended this experiment to have two decisions by generating four outputs and considering them in sets of two to denote two sets of decisions. In order to ensure that the system generates both possible responses, we use the loss which is an approximation of minimum loss of either of these decisions. This ensures that the model generates the combinations in all scenarios and the minimum loss in this case will be zero. A UNet without recurrent processing could not generate the two ground truths in the two sets of decisions (Fig.4).

Observation : In dual decision with a simple loss function the considers the maximum score of two decisions, the simple UNet could not generate correct outputs, as it trains the better initialized decision set to the spurious equilibrium. Any change from this equilibrium affects in both decisions which worsens the maximum score and hence reverts back to the spurious equilibrium in the next step. The network does not directly have knowledge the other set of decisions to make a prediction dissimilar to it, and the training loss affects both the decisions together and are not separated due to which it remains in the spurious equilibrium.



(a) Input (b) Ground Truth (c) Dual Predictions (d) Recurrent Predictions

Figure 4: Dual prediction with recurrent processing generates appropriate outputs, whereas for just dual decisions the common region is in orange as it predicted for both left and right instances or the spurious equilibrium. We trained recurrent processing with 10 epochs and without recurrent processing for 100 epochs as the model could not learn any meaningful patterns in 10 epochs. In the case of a dual decision, we can observe that the common region is in orange or predicted for both left and right instances or the spurious equilibrium for one of the decisions and the other decision does not learn anything meaningful.

4.1.3 Recurrent Dual Decisions

In our experiments we used the recurrent UNet from Wang et al. (2019). The model has a UNet backbone with a GRU (Gated Recurrent Unit) for recurrent processing, which generate two outputs, i.e. first step and then the second step that also takes input from the first step. This model architecture learns the required behaviour within fewer epochs, showing the relevance of recurrent processing in such decision making settings (Fig.4).

Observation : In a recurrent neural network, the decision is sequential, as the first decision $D_1 = f(x)$ depends only on input x and the second decision has the access to the first decision (or the history of decisions in general) $D_2 = f'(x, D_1)$. In case of an traditional convolutional neural network, the decisions $D_1 = f(x)$ and $D_2 = f'(x)$ are independent and the loss function is ineffective to learn the required behaviour during the training process. This also indicates the importance of recurrent processes in human vision for dual thinking which relies on multiple decisions.

Recurrent dual decision simulates the feed-forward processing followed by recurrent processing observed in the electrophysiological studies and our simulation finds them to be easy to train for evaluation of multiple approaches as proposed in studies on dual thinking in visual perception Dayanandan et al. (2024). A recurrent neural network with a simple loss function that minimizes the loss in either of the decisions learns the correct behavior with recurrent processing, which is similar to human vision, whereas dual decisions without recurrent processing could not learn this pattern during training.

5 Discussion

Human vision must ensure accurate inference to prevent harm in safety-critical situations, such as driving, robotic surgery, or encountering potential threats. Unlike deep learning models that make

decisions based solely on pre-learned representations, human vision actively gathers additional information when initial data is insufficient. By focusing on critical regions and iteratively refining its inference through recurrent processing, human vision improves accuracy. Deep learning models are trained using ground truth and optimized to minimize loss in each training batch, focusing on reducing batch-level loss rather than ensuring the best prediction for each sample. During training, a spurious equilibrium, analogous to a Nash equilibrium, can emerge, wherein any deviation from this suboptimal state results in an increase in loss, driving the network to revert to its prior state. As shown in Fig. 4c for dual predictions, the model converges on a spurious equilibrium for one of the decisions. Any perturbation during training influences both decisions, leading to an overall increase in loss and consequently forcing the model to return to its previous state. In such an equilibrium, the model may achieve the best overall metric but never attain complete correctness, as partial assignments remain inherently flawed. This contrasts with human vision, which prioritizes the most accurate prediction for each instance.

6 Conclusion

Human vision focuses on ensuring the accuracy of individual instances, whereas deep learning models aim to maximize overall accuracy, often leading to spurious equilibria. Theoretical analysis highlights the existence of spurious optima, which prevents neural networks from converging to the correct equilibrium. Simulations show that recurrent computation enables faster training in fewer epochs while feed-forward networks struggle to achieve the desired behavior. Our study also supports the role of sub-component formation in human vision by showing that splitting a cohesive group of pixels that are likely to occur together and assigning them to different instances is inherently suboptimal. These findings emphasize key differences in decision-making between human vision and deep learning models and highlight the importance of recurrent processing in human perception.

References

Emmanuel Abbe and Enric Boix-Adsera. On the non-universality of deep learning: quantifying the cost of symmetry. *Advances in Neural Information Processing Systems*, 35:17188–17201, 2022.

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pp. 4782–4887. PMLR, 2022.

Chen Chen, Kerstin Hammernik, Cheng Ouyang, Chen Qin, Wenjia Bai, and Daniel Rueckert. Cooperative training and latent space data augmentation for robust medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 149–159. Springer, 2021.

Kahneman Daniel. Thinking, fast and slow, 2017.

Kailas Dayanandan, Nikhil Kumar, Anand Sinha, and Brejesh Lall. Dual thinking and logical processing – are multi-modal large language models closing the gap with human vision ? *arXiv preprint arXiv:2406.06967*, 2024.

Mahdi Ghaznavi, Hesam Asadollahzadeh, Fahimeh Hosseini Noohdani, Soroush Vafaie Tabar, Hosein Hasani, Taha Akbari Alvanagh, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Trained models tell us how to make them robust to spurious correlation without group annotation. *arXiv preprint arXiv:2410.05345*, 2024.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35: 38516–38532, 2022.

Gabriel Kreiman and Thomas Serre. Beyond the feedforward sweep: feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, 1464(1):222–241, 2020.

Drew Linsley, Junkyung Kim, Vijay Veerabadran, Charles Windolf, and Thomas Serre. Learning long-range spatial dependencies with horizontal gated recurrent units. *Advances in neural information processing systems*, 31, 2018.

Drew Linsley, Junkyung Kim, Alekh Ashok, and Thomas Serre. Recurrent neural circuits for contour detection. In *International Conference on Learning Representations*, 2020.

Fahimeh Hosseini Noohdani, Parsa Hosseini, Aryan Yazdan Parast, Hamidreza Yaghoubi Araghi, and Mahdieh Soleymani Baghshah. Decompose-and-compose: A compositional approach to mitigating spurious correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27662–27671, 2024.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state rnns. *arXiv preprint arXiv:2401.06104*, 2024.

Zhaoyang Pang, Callum Biggs O'May, Bhavin Choksi, and Rufin VanRullen. Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *Neural Networks*, 144:164–175, 2021.

Hanlin Tang, Martin Schrimpf, William Lotter, Charlotte Moerman, Ana Paredes, Josue Ortega Caro, Walter Hardesty, David Cox, and Gabriel Kreiman. Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840, 2018.

Nikhil Kumar Tomar, Debesh Jha, Michael A Riegler, Håvard D Johansen, Dag Johansen, Jens Rittscher, Pål Halvorsen, and Sharib Ali. Fanet: A feedback attention network for improved biomedical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Ruben S van Bergen and Nikolaus Kriegeskorte. Going in circles is the way forward: the role of recurrence in visual inference. *Current Opinion in Neurobiology*, 65:176–193, 2020.

Rufin VanRullen. The power of the feed-forward sweep. *Advances in Cognitive Psychology*, 3(1-2): 167, 2007.

Wei Wang, Kaicheng Yu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Recurrent u-net for resource-constrained segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2142–2151, 2019.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.