Understanding Squeeze-and-Excitation Layers for Medical Image Segmentation

Miguel L. Martins
INESC-TEC
Faculty of Sciences of the University of Porto
Porto, Portugal
miguel.l.martins@inesctec.pt

Miguel T. Coimbra
INESC-TEC
Faculty of Sciences of the University of Porto
Porto, Portugal
mcoimbra@fc.up.pt

Francesco Renna
INESC-TEC
Faculty of Sciences of the University of Porto
Porto, Portugal
francesco.renna@fc.up.pt

***Abstract*—The U-Net is one of the most fundamental architectural advancements in the deep learning era. It is a crucial tool for image segmentation, especially for biomedical modalities. The research community seems to interpret the effectiveness of neural architectural search (such as the nn-U-Net) as evidence that architectural enhancements proposed since its debut are mostly unnecessary. We argue that there are still network-in-network primitives that can be leveraged to further enhance its performance, focusing on the squeeze-and-excitation (SE) pathway specifically in this paper. Specifically, we study its use of global descriptors, since it should be at odds with the spatial resolution required for dense-prediction tasks. It is theorized in the literature that performance is probably gained from some implicit ability of the learned excitations to filter supposedly uninformative channels during training.**

We explain this almost unreasonable success through an analysis of the empirical estimates of the excitation covariance matrix. Our analysis also directly contradicts the above conjecture — the most effective SE approach actually displayed the less extreme filtering behaviour, weighing all channels much closer to the mean (0.5). Our experiments are conducted in three diverse, staple biomedical modalities: dermoscopy, colonoscopy, and ultrasound.

***Index Terms*—U-Net, Semantic Segmentation, Biomedical Imaging, Squeeze-and-Excitation, Attention**

I. Introduction

Introduced in 2015 [1], the U-Net quickly became one of the most effective deep learning architectures for biomedical applications [2]. Today, the machine learning community still finds success for this architecture for many tasks, from reconstruction [3], super-resolution [4], to generative diffusion models [5].

Naturally, several enhancements were proposed to the original U-Net, from Transformer encoders, enhancing skip connection design and connectivity, to the addition of several attention models to complement the limitations of convolutional layers [2].

Notwithstanding, recently it has been proposed that much of these research efforts were not effective given a proper hyperparameter configuration (i.e, depth, number of filters, data-augmentation, etc.) of the baseline U-Net [6]. This motivated the proposal of neural architectural search algorithms such as the nn-U-Net, which is currently at the forefront in terms of performance for the medical imaging decathlon [6], [7].

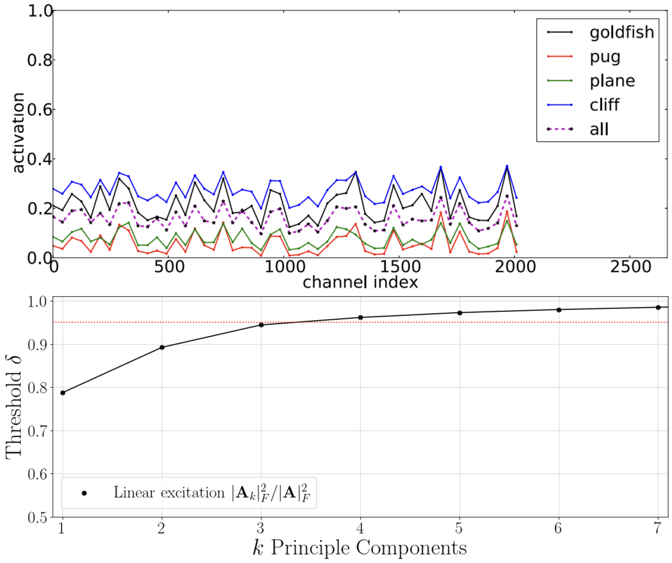

Fig. 1. (Top) Excitation analysis of excitation magnitude on a ImageNet-1000 classification task, adapted from [8]. (Bottom) The number of principal components that capture $\delta = 0.95$ of the excitation variance on a binary segmentation task on the ISIC-18 dataset.

We contend that despite these impressive results, one should not simply discard any additions to the baseline architecture as being unecessary "bells and whistles". In fact, as initially shown by [9], network-in-network patterns such as the squeeze-and-excitation (SE), not only have a proven track record for classification tasks [8], but also generalize remarkably well for segmentation tasks. Unfortunately, their effectiveness is not well understood at a fundamental level by the machine learning community. The most common conjecture is that SE modules probably filter uninformative channels [9]–[12]. However, we contend that this statement lacks any concrete meaning: does this imply setting the

This work was supported by European funds within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU) with reference 2024.07584.IACDC/2024, by the Horizon Europe Framework Programme (Grant Agreement Nº: 101095359) and by the UK Research and Innovation (Grant Agreement Nº: 10058099). Miguel L. Martins' work is funded by the individual FCT grant reference number 2021.06503.BD.

response of entire channels deterministically to 0, or is this filtering somehow dependent on the input? Is this behavior consistent across different SE architectures? Why does such an aggressive bottleneck work in dense prediction tasks that require high spatial resolution? Empirically, each excitation becomes more class-specific with the depth of the network for classification [8] (see Fig. 1), does this phenomenon also occur in segmentation?

A. Contributions

In this work, we lay the first stepping-stones towards understanding the almost unreasonable effectiveness of these approaches by studing several SE architectures [9], [13]–[15], by using them to augment a fixed baseline U-Net for three different binary medical image segmentation modalities.

Since each pixel has its own class, using class-conditional information becomes infeasible in contrast with classification tasks [8] (see Fig. 1). Instead, we study the characteristics of the empirical covariance matrix of the excitations and how their principal components relate to the number of classes for each segmentation task.

We find that the skip-connections effectively propagate semantic information by implicitly reducing the dimensionality of encoder layers closer to the input. Our results indicate that the most successful architectures offer pretty mild recalibration with some degree of instance variability, and directly contradict the argument that excitations are filtering each channels (at least, aggressively so).

B. Related work

Squeeze-and-Excitation (SE) Networks were introduced in [8] in order to address the limited receptive field while simultaneously modeling inter-channel dynamics in intermediate convolutional layers for the context of image classification. The activations are recalibrated (i.e., excited) according to a channel attention function, g, which is a non-linear map of a vector of global channel descriptors. This naturally induces a global receptive field for each layer of a convolutional neural network (CNN). Several works extend this framework with more intricate statistical descriptors or by redesigning the attention function g [12]. The style-based recalibration [13] (SRM) includes an additional global standard deviation pooling, alongside a channel-wise dense layer in the computation of g. In [14], global average pooling (GAP) is shown to be statistically equivalent to the lowest frequency component of the 2-dimensional Discrete Cosine Transform (DCT). The authors showed better performance compared to SE [8] by allowing more filters from the DCT basis to pool the responses in a multi-resolution way. A more extensive survey can be found in [12].

Surprisingly, although g typically describes each channel by means of global statistics, SE has shown to improve performance in segmentation for medical imaging tasks. In [9] showed that SE layers the performance of a baseline U-Net. Recently, [15] showed improved results when computing g as a function of the expected Hölder exponent for each channel, a quantity that relates to the fractal dimension.

Other research efforts were conducted to further expand the SE module for medical image segmentation [9], [16], but these often include spatial attention functions and/or designed for 3D segmentation, and are thus out-of-scope for the preliminary postulates of this paper.

II. Methods

We will denote a CNN encoder given an input tensor $\mathbf{X} \in \mathbb{I}_0$ as $f_l(\mathbf{X}) = f_l \circ \ldots \circ f_2 \circ f_1(\mathbf{X}))$, so that $f_l : \mathbb{I}_{l-1} \to \mathbb{I}_l$ for $\mathbb{I}_l \equiv \mathbb{R}^{H_l \times W_l \times C_l}$. We will refer to f_l as the l-th layer of f, so that $l \in \{1, \ldots, L\}$. f_{lhwc} is the c-th channel, at position (h, w) of the l-th layer. Assume these indexing symbols are consistent throughout and that their omission implies tensor slicing, e.g. $f_{lc} \equiv f_{l::c}$. We will study four representative SE solutions in the context of image segmentation that have found success in the medical imaging literature.

A. Squeeze-and-Excite (SE) [8]

Squeeze-and-excitation networks were originally studied by [9], and we will onward use the the acronym cSE (channel SE) for consistency with medical image segmentation literature, even though cSE and SE are functionally equivalent.

The output of each encoder layer f_l is characterized by a ‘’squeeze” function $g : \mathbb{I}_l \to \mathbb{R}^{C_l}$ such that:

$$g(f_l(\mathbf{X})) = \sigma(\mathbf{W_2}\,\mathrm{ReLU}(\mathbf{W}_1\mathrm{GAP}(f_l(\mathbf{X})))), \tag{1}$$

where GAP stands for spatial global average pooling, $\mathbf{W}_1 \in \mathbb{R}^{\lfloor \frac{C_l}{s^*} \rfloor \times C_l}$, $\mathbf{W}_2 \in \mathbb{R}^{C_l \times \lfloor \frac{C_l}{s^*} \rfloor}$, for some $1 \leq s^* < C_l$, and σ is the sigmoid activation function. The output of the encoder at level l for $l < L - 1$ thus becomes:

$$f_l^{\text{Excited}}(\mathbf{X}) = f_l(\mathbf{X}) \odot g(f_l(\mathbf{X})), \tag{2}$$

where $\odot$ is the (broadcastable) element-wise product.

B. Style-based Recalibration (SRM) [13]

Style-based Recalibration (SRM) integrates the standard deviation as a proxy for style in the context of style-transfer. Denoting a global standard-deviation pooling layer as GSP, the squeeze function is defined as:

$$g(f_l(\mathbf{X})) = \phi([\mathrm{GAP}((f_l(\mathbf{X}))), \mathrm{GSP}((f_l(\mathbf{X})))]), \tag{3}$$

where $\phi : \mathbb{R}^{C_l \times 2} \to \mathbb{R}^{C_l}$ is a learnable linear map. An additional batch-normalization layer is applied to g.

C. Frequency Channel Attention (FCA) [14]

Frequency channel attention splits the C_l channels of each layer l in k groups such that $[f_l(\mathbf{X})_1, \ldots, f_l(\mathbf{X})_k] = f_l(\mathbf{X})$.

Then the two-dimensional discrete cosine transform (DCT) filters are pre-computed and stored on a tensor $\mathbf{B}$ such that

$$\mathbf{B}_{l,h,w}^{i,j} = \cos\left(\frac{\pi h}{H_l}\left(i + \frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W_l}\left(j + \frac{1}{2}\right)\right). \tag{4}$$

The squeeze operation is performed for each of the k groups

$$g(f_l(\mathbf{X}))_k = \sum_h \sum_w f_l(\mathbf{X})_k B_{l,h,w}^{i_k,j_k}, \tag{5}$$

i_k, j_k depend on k since there is a distinct DCT basis for each of the k groups $g(f_l(\mathbf{X})) := \sigma(\mathbf{W}[g(f_l(\mathbf{X}))_1, \ldots, g(f_l(\mathbf{X}))_k])$, so that $\mathbf{W} \in \mathbb{R}^{C_l \times C_l}$.

D. Monofractal Recalibration [15]

Monofractal recalibration [15] considers each activation map f_{lc} has a two-dimensional self-similar measure μ_{lc} characterized by distinct singularity exponent $\alpha_{lc}(\cdot)$ which can be determined as follow:

$$\alpha(x) = \frac{\sum_r \left(\log \mu(B_r(x)) - \frac{1}{|R|}\sum_{r'} \log \mu(B_{r'}(x))\right)}{\sum_r \left(\log r - \frac{1}{|R|}\sum_{r'} \log r'\right)}. \quad (6)$$

For discrete, finitely many $r > 1 \in R$, $B_r(x)$ is a square of side r centered around $x \in \mathbb{R}^{H \times W}$. Then let

$$\alpha_{lc} = [\alpha(f_{lhwc}(\mathbf{X})) : \text{for each } (h, w) \text{ at resolution } l]. \quad (7)$$

Then, the squeeze function is such that $g(f_l(\mathbf{X})) := g(\tilde{\alpha}_l))$, where $\tilde{\alpha}_l$ is the output of a batch-normalization layer over α_l.

E. The Linear Excitation Threshold

As argued in the original paper SE [8], for the task of image classification, empirically the activations induced by g become more class-specific as $l \to L$, becoming (at least) visually separable before the output layer (see Fig. 1). However, for image segmentation, each pixel maps to its own class, and this relationship becomes impossible to estimate given the lack of spatial resolution on the excitation maps g.

A linear estimation of the the covariance of the excitations $g(f_l)$ is thus set forth, so we solve for a surrogate $\mathbf{A} \in \mathbb{R}^{C_l \times C_l}$ such that $\mathbf{A} := \frac{1}{n-1} g(f_l)^T g(f_l)$ is the empirical covariance matrix and n is the size of a validation set. Note that this aligns with recent trends in deep learning theory that state that, under mild assumptions, deep representations tend to display linear behaviour after training [17], [18].

Denote $\mathbf{A}_k$ as the singular value decomposition of $\mathbf{A}$ using only its top k singular values, then for ϵ small enough we define the *linear excitation threshold* ($\delta \in (0, 1]$) is given by

$$\text{argmin}_k \frac{|\mathbf{A}_k|_F^2}{|\mathbf{A}|_F^2} \geq \delta. \quad (8)$$

In other words, for a given δ, if $k \to C$, we say that g accurately captures class dependencies, in the same spirit of original the empirical analysis of [8] (illustrated in Fig. 1).

TABLE I
Mean ± standard deviation Dice score (%) of the cross-validation experiments. $\cdot^{\dagger}$ and $\cdot^{\ddagger}$ signify that the null-hypothesis of the pairwise t-test with regards to the U-Net baseline is rejected with $p \leq 0.05$ and $p \leq 0.01$, respectively. Best mean results are boldfaced.

Model	ISIC18	Kvasir-SEG	BUSI
U-Net [1]	85.40 ± 0.25	72.22 ± 1.82	62.20 ± 2.40
+cSE [8]	$85.94 \pm 0.36^{\dagger}$	**72.72 ± 1.52**	65.36 ± 1.36
+SRM [13]	84.33 ± 1.27	61.13 ± 3.42	$68.09 \pm 3.14^{\dagger}$
+FCA [14]	86.19 ± 0.75	70.00 ± 2.51	66.27 ± 2.48
+Mono [15]	$\mathbf{86.24 \pm 0.27}^{\ddagger}$	71.86 ± 2.37	$\mathbf{69.00 \pm 2.53}^{\ddagger}$

F. Materials

We select three staple public medical image segmentation datasets spanning a diverse set of imaging modalities: dermoscopy, colonoscopy, and ultrasound. Speficially, the 2018 International Skin Lesion Collaboration (ISIC-18) [19], the 2020 KvasirSeg [20] for colonic polyp detection, and the 2020 Breast Ultrasound Images (BUSI) dataset [21], containing masks for ultrasound scans from absent (i.e., normal), benign and malign masses.

III. Experimental Methodology

We follow the U-Net implementation used in the experimental section of [2], [15]. It has $L = 3$ encoder/decoder pairs that output 32, 64, and 128 channels, respectively. The bottleneck has 256 channels. All convolution layers use rectified linear unit (ReLU). A batch size of 16 was used to guarantee that channel attention methods that use batch-normalization layers are not at a disadvantage [13], [15]. We use Adam for gradient descent over 400 epochs, and a scheduler would reduce the learning rate (initially set to 1×10^{-4}) by a factor of 0.5 should a plateau be detected over the span of 5 epochs. Data augmentation was limited to random flips.

Each model was evaluated under a 10-fold cross-validation for each dataset. The image / mask pairs were normalized to [0, 1] and then down-scaled using bilinear interpolation to a spatial resolution of 224×224. In all cases, 10% of the out-of-fold data was set aside as the validation set for early stopping. To understand the performance of each model we display the Dice score statistic in Table I. Stratified sampling according to the frames' pathological class was employed to generate the folds for the BUSI dataset, mitigating data-leakage or imbalance conditioned on the sampling bias of malignant frames. We set $k = 16$ for FCA [14] and $R \equiv \{2, 4, 8\}$ for Monofractal recalibration [15], since they are the best-reported configurations reported by the authors.

A. Excitation analysis

The validation activation responses were collected for each of the attention modules. We set $\delta = 0.95$ and approximate $\mathbf{A}_k$ by principle component analysis. We solve (8) iteratively by increasing the number of principal components k until the ratio exceeds δ. Additionally, we also inspect the excitation responses at the neuron level by fitting a Gaussian whose sufficient statistics are computed channel-wise across all samples in the validation sets (see Fig. 2).

B. Findings and Discussion

Excitations do not necessarily filter non-informative channels: This common analogy [2], [9], [12] simply is not replicated in our results. Although Monofractal displays the highest performance for the ISIC-18 dataset (see Table I), its excitations weigh each channel much less aggressively than all other SE functions (see Fig. 2).

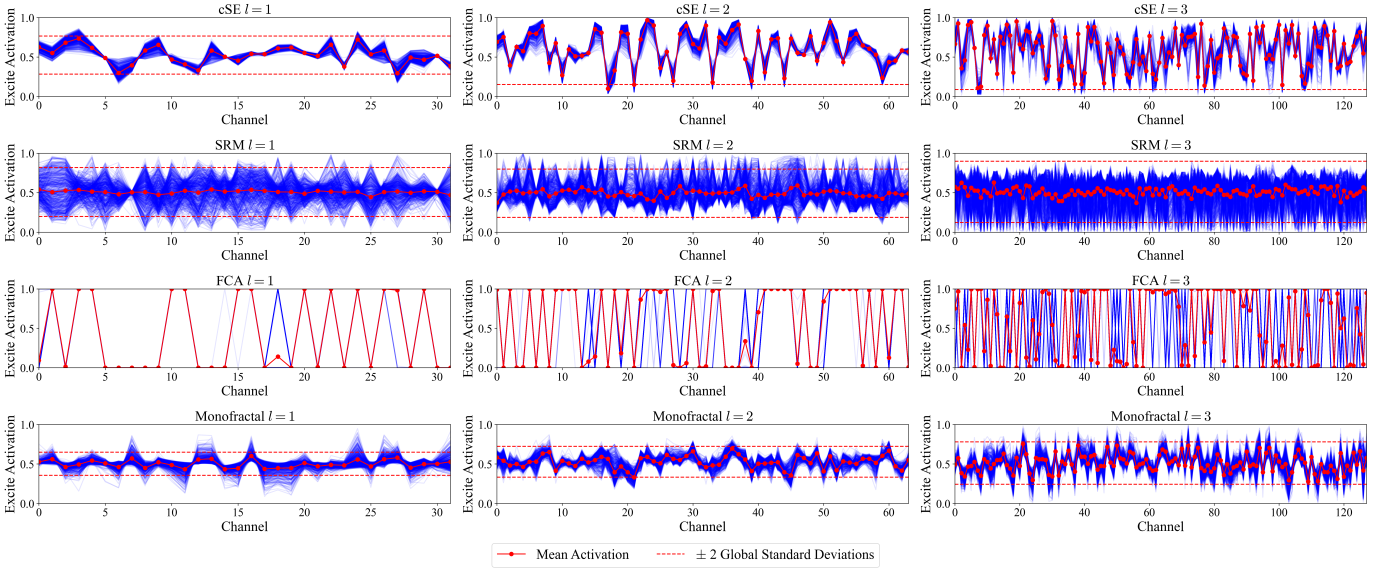

Fig. 2. Test set excitation responses of cSE, SRM, FCA, and Monofractal recalibration for $l \in \{1, 2, 3\}$ in the first fold of the ISIC dataset. Each blue segment marks a response of the layer given an instance. SRM displays very heterogeneous behaviour per instance. On the opposite end of the spectrum, FCA is almost instance-agnostic. cSE and Monofractal recalibration display similar instance variability, but markedly different excitation responses around the average.

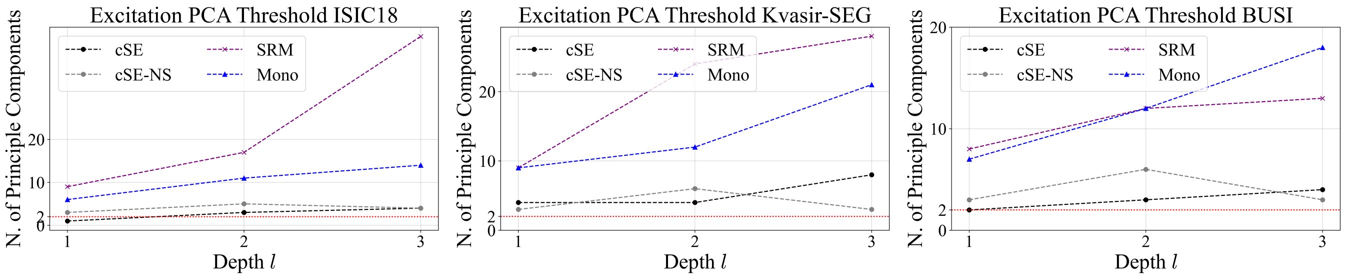

Fig. 3. Number of principle components required to explain 95% of the variance for each model for every encoder depth l for each dataset. cSE-NS stands for cSE in an U-Net without skip-connections. FCA is not included due to its almost heterogeneous behavior. The U-Shape architecture makes it so that become less class-dependent as $l \to L$.

***Skip connections affect the covariance of g depending on** l:* We tested removing the skip-connections (see Fig. 3, cSE-NS) and found that this would break the monotonic quasi-linear relationship between and l and k. This behavior is suggestive of the original findings presented in [8] for classification. Since $C \ll C_l$, this is reflected in the dimensionality of the linear manifold where the approximation of $\mathrm{Cov}(g)$ resides, for a given δ.

Balanced instance variability is likely advantageous: Both cSE and Monofractal displayed the best results in our experiments. In contrast, as it can be observed in Fig. 2, when there was an excessive degree of instance variability (SRM), or an almost deterministic excitation behaviour (FCA), performance was usually either sub-par, and even detrimental when comparing SRM to baseline, for two out of the three datasets.

IV. Conclusion

In this paper we layed the foundation for understanding squeeze-and-excitation layers for the context of image segmentation. We propose an analysis of the empirical covariance of the excitation layers which contradicted common conjectures in the literature around these channel-attention functions. Specifically, our evidence suggest that their effectiveness is not linked to any type of apparent filtering of certain presumably uninformative channels. Our findings also suggest that a balanced instance variability correlates with increased downstream performance.

We intend to extend this work by analyzing these dynamics during the entirety of the training procedure. Moreover, we also intend to study the viability of extending the nn-U-Net [6] to include the most promising squeeze-and-excitation primitives.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, “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.

[2] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, “Medical image segmentation review: The success of u-net,” *arXiv preprint arXiv:2211.14830*, 2022.

[3] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

[4] J. Tang, L. Niu, L. Liu, H. Dai, and Y. Ding, “Vmg: Rethinking u-net architecture for video super-resolution,” *IEEE Transactions on Broadcasting*, 2024.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[6] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[7] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, and P. F. Jaeger, “nnu-net revisited: A call for rigorous validation in 3d medical image segmentation,” *arXiv preprint arXiv:2404.09556*, 2024.

[8] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[9] A. G. Roy, N. Navab, and C. Wachinger, “Recalibrating fully convolutional networks with spatial and channel “squeeze and excitation” blocks,” *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 540–549, 2018.

[10] S. Pereira, A. Pinto, J. Amorim, A. Ribeiro, V. Alves, and C. A. Silva, “Adaptive feature recombination and recalibration for semantic segmentation with fully convolutional networks,” *IEEE transactions on medical imaging*, vol. 38, no. 12, pp. 2914–2925, 2019.

[11] R. Azad, A. R. Fayjie, C. Kauffmann, I. Ben Ayed, M. Pedersoli, and J. Dolz, “On the texture bias for few-shot cnn segmentation,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2674–2683, 2021.

[12] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, “Attention mechanisms in computer vision: A survey,” *Computational visual media*, vol. 8, no. 3, pp. 331–368, 2022.

[13] H. Lee, H.-E. Kim, and H. Nam, “Srm: A style-based recalibration module for convolutional neural networks,” in *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 1854–1862, 2019.

[14] Z. Qin, P. Zhang, F. Wu, and X. Li, “Fcanet: Frequency channel attention networks,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 783–792, 2021.

[15] M. L. Martins, M. T. Coimbra, and F. Renna, “Singularity strength re-calibration of fully convolutional neural networks for biomedical image segmentation,” in *Proceedings of the 2024 32nd European Signal Processing Conference (EUSIPCO)*, 2024.

[16] A.-M. Rickmann, A. G. Roy, I. Sarasua, and C. Wachinger, “Recalibrating 3d convnets with project & excite,” *IEEE transactions on medical imaging*, vol. 39, no. 7, pp. 2461–2471, 2020.

[17] K. Park, Y. J. Choe, and V. Veitch, “The linear representation hypothesis and the geometry of large language models,” *arXiv preprint arXiv:2311.03658*, 2023.

[18] M. Huh, B. Cheung, T. Wang, and P. Isola, “The platonic representation hypothesis,” *arXiv preprint arXiv:2405.07987*, 2024.

[19] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, *et al.*, “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic),” *arXiv preprint arXiv:1902.03368*, 2019.

[20] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, “Kvasir-seg: A segmented polyp dataset,” in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pp. 451–462, Springer, 2020.

[21] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, “Dataset of breast ultrasound images,” *Data in brief*, vol. 28, p. 104863, 2020.