WHERE SHOULD EXCITATIONS MEET? DEEP FUSION STRATEGIES FOR MULTI-EXCITATION HYPERSPECTRAL IMAGING OF ABLATED ATRIAL TISSUE

Arpi Hunanyan[1,2]*, Nazeli Ter-Petrosyan*[1]*, Fernando Villarruel*[3]*,*
Narine Sarvazyan[1,3,4]*, Aram Butavyan*[1]*, and Varduhi Yeghiazaryan*[1]

[1]Akian College of Science and Engineering, American University of Armenia, Yerevan, Armenia
[2]Institut Polytechnique de Paris, Palaiseau, France
[3]L. A. Orbeli Institute of Physiology NAS RA, Yerevan, Armenia
[4]George Washington University, Washington, DC, United States

ABSTRACT

Multi-excitation hyperspectral imaging captures complementary tissue responses under multiple excitation wavelengths. Deep learning can exploit this diversity, yet it remains unclear at which stage excitation-specific information should be fused and which fusion mechanism is most effective. Using bovine atrial tissue, we introduce a controlled framework that systematically evaluates fixed, globally learned, and sample-dependent fusion operators at both feature and logit levels. Selected fusion variants outperform a decision-level soft majority-voting baseline. Feature-level attention remains consistently competitive, convolutional fusion is effective for rich spatial–spectral features, and globally weighted averaging provides modest logit-level gains. Overall, fusion performs best when the operator matches the representation: structured adaptive mechanisms for rich features and simpler, stable weighting for class-level evidence. The implementation and experimental configurations are available at `https://github.com/ArpiHunanyan/me-hsi-excitation-fusion`.

***Index Terms*—** Multi-excitation hyperspectral imaging, excitation fusion, feature-level fusion, decision-level fusion, deep learning, HSI classification, CNN, multi-head attention

1. INTRODUCTION

Hyperspectral imaging (HSI) provides a rich representation of biological tissue by recording a full emission spectrum at each spatial pixel. This spectral information can capture subtle biochemical and structural differences that are not visible in conventional RGB imaging. Multi-excitation hyperspectral imaging (ME-HSI) extends this setting by acquiring several hyperspectral cubes of the same sample under different excitation wavelengths. The resulting representation contains spatial, spectral, and excitation dimensions. This additional excitation axis is important because different excitation wavelengths may reveal complementary tissue responses, rather than simply adding more input channels.

Recent biomedical ME-HSI studies indicate that excitation diversity can provide complementary information. A standard treatment for atrial fibrillation (AF) is radiofrequency ablation (RFA), in which atrial tissue is intentionally lesioned to interrupt abnormal electrical activity. Multiple excitation wavelengths have been shown to improve the classification of ablated and unablated tissue compared with single-excitation settings [1, 2]. These findings indicate that excitation diversity can provide complementary tissue information and that ME-HSI should be analyzed not only as a larger input cube, but as a structured representation with a meaningful excitation axis.

The comparative ME-HSI study [2] on ablated atrial tissue is especially important for motivating this work. In that study, excitation information was mainly combined using simple aggregation strategies. That work compared joint input concatenation with independently trained excitation-specific models whose predictions were combined only at inference time through hard or soft majority voting. Soft majority voting performed best, motivating us to preserve excitation-specific processing and treat the subsequent integration of their information as an explicit fusion problem. This fusion question is not unique to ME-HSI; it is central to conventional HSI, where spectral signatures must be combined with spatial context. Prior work on ablated atrial tissue showed that model performance depends on how spectral and spatial information are used [3]. Recent HSI architectures address this by keeping complementary streams separate before fusion: S2DBFT uses spectral–spatial branches with learned attention guidance [4], while CACFTNet combines convolutional and Transformer features through covariance attention and cross-layer fusion [5]. These works suggest that effective HSI fusion should preserve structure before reweighting and combining representations, rather than collapsing all information at the input level.

Thanks to the Afeyan Family Foundation, RA HESC (24PostDoc/2-2G002), and the EU (ERA-CHAIR NAS-SAR 101087403) for funding.

Rather than treating excitation fusion as an isolated problem, it is useful to ask how deep learning has already learned to combine complementary information. Attention fuses representations through learned weighted combinations [6]. Squeeze-and-Excitation and CBAM show that feature maps can be recalibrated using learned channel or spatial importance weights [7, 8]. Selective Kernel Networks extend this idea to parallel branches, using softmax-based weights to decide which branch response should dominate [9]. Gated multimodal fusion learns how much each information source should contribute [10], while multimodal Transformers model interactions between separate streams through cross-modal attention, as in MulT [11] and attention-bottleneck fusion [12].

In this work, we study excitation fusion through two questions: where should excitation-specific information be combined, and how should it be aggregated? By varying the fusion stage and operator, we show that operator rankings change with the representation being fused: attention is consistently competitive at feature level, whereas averaging and globally learned weighting are more reliable at logit level.

2. DATASET

Sample collection: As reported in [13], bovine left atrial tissue provides a model close to human myocardium due to its similar optical properties, relevant to HSI acquisition. Consequently, the experiments were executed using bovine hearts obtained from local butchers in the Yerevan surroundings. Three bovine heart samples were provided by a local butcher, immediately after sacrifice of the animals (during November–December 2024). The hearts were dissected on-site, extracting the atria which were immediately transported on-ice to the lab (Fig. 1). All samples were obtained post-mortem from animals sacrificed for food purposes; no live animal experiments were performed during this phase. A total of eight ME-HSI samples were acquired.

RFA procedure: Once in the lab, left atria were separated into pieces for RFA execution, using an EPT-1000 XP machine at 8 W power for 15 s, connected to a 4 mm Blazer catheter (Boston Scientific) which was perpendicularly placed on the tissue during ablation. Immediately after RFA execution, the lesion center was marked by inserting a needle to create a dot in the tissue, which was later used for the TTC analysis (Fig. 1).

HSI acquisition setup: The imaging system comprised a self-assembled light source with a 300 W xenon lamp outfitted with four filters centered at 360, 370, 380 and 390 nm (all with 10 nm FWHM; Thorlabs), together with a Nuance FX (Perkin Elmer) hyperspectral camera. Four cubes were obtained, with the above-mentioned λ_{exc} and 31 emission wavelengths ($\lambda_{\text{em}} = 420$–$720$ nm with 10 nm step), while spatial dimensions varied depending on tissue size. All imaging experiments were completed within six hours after animal sacrifice to limit post-mortem tissue changes.

Histological analysis: Upon completion of the imaging, tissue slices were cut from the centers of the marked lesions and stained with 1.0% TTC solution for 20 minutes at room temperature, to determine the lateral size of each lesion and to generate ground truth masks necessary for classification tasks (Fig. 1).

Ground truth mask creation: To differentiate tissue pixels from background, a threshold signal was applied to the images, and the pixels with a signal below this threshold were labeled 'background'. The lesions' center coordinates were obtained by finding the needle dots in the original images. From each center a circle of a diameter equal to the lateral size obtained by TTC was drawn. As the tissue slicing breaks tissue fibers, resulting in an overestimation of the lesion's size, the outer 30% of each circle was labeled 'unsure', while the inner 70% was labeled 'ablated'. Also, for the cases where the ablation centers were unclear in the images, completely 'unsure' circles were drawn around the estimated center coordinates. Pixels not included in any of the previously mentioned classes were labeled as 'unablated' tissue.

For the models, 'background' pixels were not included, while 'ablated' and 'unablated' pixels were included in the training and test samples. 'Unsure' pixels were not included in the training and/or test samples, but were predicted by the models.

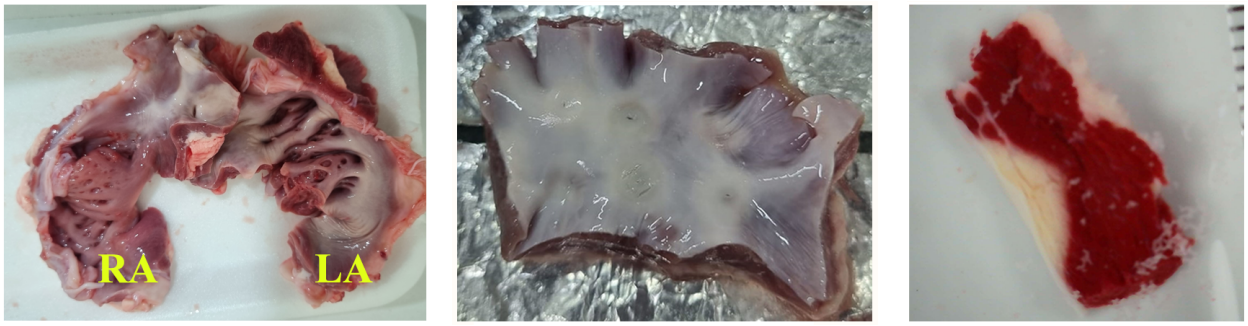
ME-HSI data preprocessing: Data preprocessing consisted of exposure time correction, lamp emission correction, and pixel-wise min–max normalization, intended to reduce the effect of different absolute intensities between pixels of the same image.

3. METHODS

Each ME-HSI sample provides E distinct views of the same tissue, one for each excitation wavelength. Rather than combining them immediately, we preserve each view in an excitation-specific branch and make fusion the central architectural choice: when should the branches meet, and how should their representations be combined?

We use two backbone families to test whether the fusion behavior is consistent across different feature-extraction regimes. The MLP serves as a compact spectral baseline, mapping each excitation-specific spectrum through fully connected layers and relying mainly on per-pixel spectral information. In contrast, LeViT [14] is a hybrid convolution–attention backbone that processes each excitation spatially, combining local image structure with broader contextual interactions to learn richer spatial–spectral features.

Both models were previously effective for ME-HSI classification [2], while remaining lighter and easier to control than graph-based approaches requiring additional superpixel construction. This makes them suitable for a systematic fusion study, where the main variable should be the fusion stage and



(a) (b) (c)

Fig. 1: Representative images of the radiofrequency ablation (RFA) process on bovine atrial tissue. (a) Dissected bovine atria showing the right and left atria (RA and LA, respectively). (b) Left atrial tissue showing six RFA lesions, with their centers marked by needle punctures. (c) TTC-stained tissue slice taken from the center of an RFA lesion, showing necrotic tissue in white and viable tissue in red.

Table 1: Evaluated excitation-fusion operators. Weight type distinguishes fixed, globally learned, sample-dependent, and learned joint mixing; permutation invariance indicates whether reordering the excitations leaves Y unchanged. Here $\widetilde{X}_e \in \mathbb{R}^D$ is the normalized representation of excitation e, $\widetilde{X} \in \mathbb{R}^{E\times D}$ their stack, and $Y \in \mathbb{R}^D$ the fused output; g are pre-softmax scores, $w = \mathrm{softmax}(g)$ the weights, W_f, b_f, w_g, b_g, q learned parameters, k the kernel size, and MHA multi-head attention.

Operator	Mathematical Formulation	Weight Type	Permutation Invariant
AVG	$Y = \frac{1}{E}\sum_{e=1}^{E} \widetilde{X}_e$	Fixed uniform	Yes
SUM	$Y = \sum_{e=1}^{E} \widetilde{X}_e$	Fixed uniform	Yes
FC	$Y = W_f\,\mathrm{vec}(\widetilde{X}) + b_f$	Learned mixing	No
Gate	$g_e = w_g^\top \widetilde{X}_e + b_g$ $w = \mathrm{softmax}(g)$ $Y = \sum_{e=1}^{E} w_e \widetilde{X}_e$	Sample-dependent	Yes
CNN	$g = \mathrm{Conv1D}_{k=3}(\widetilde{X}^\top)$ $w = \mathrm{softmax}(g)$ $Y = \sum_{e=1}^{E} w_e \widetilde{X}_e$	Sample-dependent	No
WAVG	$g \in \mathbb{R}^E$, g learned $w = \mathrm{softmax}(g)$ $Y = \sum_{e=1}^{E} w_e \widetilde{X}_e$	Globally learned	No
ATT	$q \in \mathbb{R}^D$, q learned $Y = \mathrm{MHA}(q, \widetilde{X}, \widetilde{X})$	Sample-dependent	Yes

fusion operator rather than extra preprocessing complexity.

Within this controlled fusion framework, the excitation branches form parallel paths through the network, and the fusion point determines when these paths first interact. Moving this point shifts the information being combined from rich intermediate representations to final class evidence. Building on prior results showing that decision-level soft majority voting outperformed direct input combination [2], we compare two fusion stages: feature-level fusion, performed before classification, and logit-level fusion, performed after each branch has produced its own class scores. Here, Φ denotes the feature extractor, Ψ the classification head, and $\mathcal{B}$ the complete backbone including classification. The notation $E\times$ indicates E fully independent excitation-specific branches with no shared parameters; thus, feature-level fusion follows $E \times \Phi \rightarrow$ fusion $\rightarrow \Psi$, whereas logit-level fusion follows $E \times \mathcal{B} \rightarrow$ fusion.

Because the independently parameterized branches process excitation-specific inputs, their representations may differ in scale and distribution. We therefore normalize each excitation representation independently across its feature dimensions before fusion, bringing the representations to a comparable scale while preserving excitation-specific information.

At feature level, the normalized representations are stacked as $\widetilde{X} \in \mathbb{R}^{E\times D}$, where E is the number of excitations and D is the feature dimension; $D = 2048$ for MLP and $D = 512$ for LeViT. At logit level, each branch produces class scores, giving $\widetilde{X} \in \mathbb{R}^{E\times C}$, where C is the number of classes.

Once the branches have reached their meeting point, the next question is how their information should be combined. We therefore compare fusion operators ranging from simple fixed aggregation to globally learned and sample-dependent mechanisms. Table 1 summarizes the operators under their abbreviated names (**AVG**, **SUM**, **FC**, **Gate**, **CNN**, **WAVG**, and **ATT**), together with their mathematical formulation, weighting strategy, and permutation-invariance property.

The operators differ not only in how their weights are learned, but also in whether they depend on excitation order. **SUM** and **AVG** use fixed uniform weights and are permutation invariant, whereas **FC** learns an order-sensitive joint mixing of all representations. **Gate**, **CNN**, and **WAVG** share the same weighted-sum structure, but obtain their weights differently: **Gate** predicts sample-dependent weights while remaining permutation invariant, **CNN** derives sample-dependent weights from local relations along the ordered

excitation axis, and **WAVG** learns one global weight for each fixed excitation position.

Finally, since few contemporary fusion studies seem complete without an attention mechanism, **ATT** is included as a structured, sample-dependent alternative. It learns which excitation representations should contribute most to the fused output while remaining permutation invariant, as no excitation-order encoding is introduced.

We evaluate ME-HSI classification with a training protocol based on simulated user clicks: five clicks per class (ablated, unablated), each covering 0.2% of the image area, for 2% total training coverage. Performance is reported as overall accuracy (OA), average accuracy (AA), and Cohen's κ, averaged over the eight ME-HSI samples and five click splits. Experiments run on Ubuntu 24.04 with an Intel Core i9-14900K CPU and an NVIDIA GeForce RTX 4090 GPU.

4. RESULTS AND DISCUSSION

Table 2: Controlled comparison of excitation-fusion strategies, given as the mean over eight ME-HSI samples and five splits, with standard deviation below each value. Metrics are overall accuracy (OA), average accuracy (AA), and Cohen's κ. $E\times$ denotes E independent excitation branches, Φ the feature extractor, Ψ the classification head, and $\mathcal{B}$ the full backbone; hence $E \times \Phi \rightarrow \text{Operator} \rightarrow \Psi$ is feature-level and $E \times \mathcal{B} \rightarrow \text{Operator}$ logit-level fusion. **Soft MV** is the soft majority-voting baseline [2]. Each metric column is shaded independently from its minimum to maximum value, with darker cells indicating higher values.

Fusion Configuration	LeViT Family			MLP Family		
	OA	AA	κ	OA	AA	κ
$E \times \mathcal{B} \rightarrow$ Soft MV	.800 ± .121	.851 ± .068	.371 ± .206	.787 ± .118	.838 ± .072	.352 ± .211
$E \times \Phi \rightarrow$ AVG $\rightarrow \Psi$	.797 ± .118	.832 ± .090	.352 ± .222	.782 ± .111	.836 ± .063	.336 ± .197
$E \times \Phi \rightarrow$ SUM $\rightarrow \Psi$	.802 ± .108	.847 ± .071	.359 ± .201	.785 ± .117	.835 ± .066	.339 ± .192
$E \times \Phi \rightarrow$ FC $\rightarrow \Psi$	.791 ± .155	.818 ± .119	.329 ± .197	.783 ± .112	.832 ± .067	.330 ± .189
$E \times \Phi \rightarrow$ ATT $\rightarrow \Psi$	.805 ± .108	.858 ± .065	.415 ± .187	.789 ± .123	.842 ± .075	.391 ± .206
$E \times \Phi \rightarrow$ Gate $\rightarrow \Psi$	.765 ± .150	.827 ± .106	.357 ± .205	.777 ± .108	.835 ± .066	.371 ± .202
$E \times \Phi \rightarrow$ CNN $\rightarrow \Psi$	.818 ± .106	.854 ± .088	.426 ± .206	.782 ± .124	.832 ± .074	.391 ± .222
$E \times \Phi \rightarrow$ WAVG $\rightarrow \Psi$	.786 ± .159	.854 ± .084	.422 ± .225	.780 ± .112	.846 ± .061	.385 ± .199
$E \times \mathcal{B} \rightarrow$ AVG	.815 ± .110	.869 ± .062	.439 ± .195	.784 ± .119	.848 ± .070	.401 ± .217
$E \times \mathcal{B} \rightarrow$ SUM	.798 ± .126	.857 ± .077	.424 ± .215	.780 ± .121	.843 ± .070	.391 ± .210
$E \times \mathcal{B} \rightarrow$ FC	.796 ± .151	.847 ± .094	.415 ± .210	.785 ± .120	.846 ± .072	.398 ± .210
$E \times \mathcal{B} \rightarrow$ ATT	.803 ± .157	.836 ± .119	.422 ± .244	.789 ± .122	.845 ± .072	.399 ± .208
$E \times \mathcal{B} \rightarrow$ Gate	.801 ± .150	.813 ± .139	.379 ± .254	.783 ± .126	.842 ± .069	.390 ± .193
$E \times \mathcal{B} \rightarrow$ CNN	.776 ± .147	.805 ± .115	.312 ± .206	.783 ± .122	.830 ± .071	.341 ± .201
$E \times \mathcal{B} \rightarrow$ WAVG	.815 ± .113	.869 ± .064	.446 ± .206	.793 ± .111	.851 ± .066	.404 ± .210

The gains over **Soft MV** are selective rather than uniform across fusion operators and backbones. As shown in Table 2, only feature-level **CNN** and **ATT** improve LeViT over **Soft MV** across OA, AA, and κ simultaneously, whereas for MLP this is achieved only by **ATT**. At logit level, LeViT improves with **AVG** and **WAVG**, while MLP improves with **ATT** and **WAVG**. The largest gains are obtained with logit-level **WAVG**, improving LeViT by +1.5% OA, +1.8% AA, and +0.075 κ, and MLP by +0.6% OA, +1.3% AA, and +0.052 κ, although the gains remain modest.

The limited sample size (three hearts, eight ME-HSI samples) and large standard deviations reduce the robustness and generalizability of operator rankings; therefore, we interpret the results as recurring trends rather than definitive performance differences. This contrast suggests that adaptive fusion is most effective while excitation representations retain spatial–spectral detail, whereas once they are compressed into class scores, simpler and more stable aggregation becomes preferable.

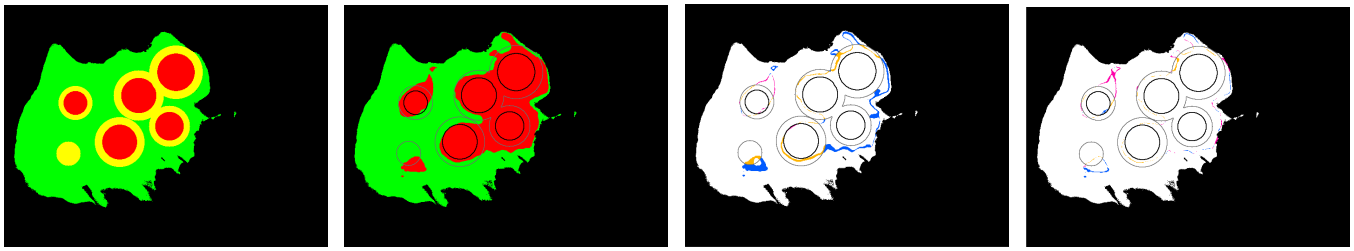
Beyond the gains over the baseline, the rankings reveal a broader pattern: fusion works best when the operator matches the structure of the representation it receives.

SUM remains consistently mid-ranking because, with $E = 4$, it is simply $4\times$ **AVG** after normalization and therefore changes mainly the output scale rather than the fused information. At logit level, **AVG** is effective because the excitation outputs are aligned as class evidence, allowing parameter-free averaging to reduce branch-specific noise without altering class semantics. At feature level, however, independently learned feature dimensions need not be equally informative or aligned, making uniform averaging less effective. For the MLP, whose representations are primarily spectral, **AVG** remains comparable in OA and AA, although its lower κ indicates that it is not uniformly competitive.

The generally weaker feature-level performance of **FC** and **Gate** suggests that linear mixing or scoring without structured cross-excitation interaction is insufficient in this setting. **FC** jointly mixes all excitation features through a single unconstrained projection, whereas **Gate** scores each excitation independently with a shared linear function before softmax weighting. Although **FC** becomes more competitive at logit level, where the representation is much smaller, the behavior of **Gate** remains less consistent.

In contrast, **ATT** preserves the excitation structure and uses a learned query to emphasize informative representations for each sample, while **CNN** additionally exploits local relations along the ordered excitation axis. This is especially effective with LeViT, whose features retain richer spatial–spectral information. Notably, the operators that exploit the ordering of the excitation axis, **CNN** at feature level and **WAVG** at both stages, outperform their permutation-invariant counterpart **Gate**, which may indicate useful structure in the 360–390 nm sequence.

Overall, **WAVG** is strongest at logit level, where it yields



(a) GT (b) $E \times \mathcal{B} \rightarrow$ **Soft MV** (c) $E \times \Phi \rightarrow$ **WAVG** $\rightarrow \Psi$ (d) $E \times \mathcal{B} \rightarrow$ **WAVG**

Fig. 2: Qualitative comparison of LeViT fusion strategies. (a) Ground truth (GT): 'ablated' tissue in red, 'unablated' tissue in green, 'unsure' regions in yellow, and 'background' in black. (b) **Soft MV** prediction, with 'ablated' and 'unablated' tissue in red and green; GT 'ablated' and 'unsure' boundaries are overlaid in black and gray. (c), (d) Difference maps of feature-level and logit-level **WAVG** with respect to **Soft MV**: white denotes identical predictions, blue errors corrected by **WAVG**, magenta correct **Soft MV** predictions turned into errors, yellow disagreement inside GT 'unsure' regions, and black background.

modest but consistent improvements across both backbones.

Fig. 2 illustrates how the fusion stage affects the spatial prediction pattern. Feature-level **WAVG** produces more localized corrections relative to **Soft MV**, but also introduces additional disagreements near lesion boundaries, whereas logit-level **WAVG** remains closer to the baseline while reducing some isolated errors. This example suggests that feature fusion is more adaptive spatially, while logit fusion provides a more conservative aggregation of class evidence.

5. CONCLUSION

This work proposes a controlled framework for evaluating feature- and logit-level fusion strategies in ME-HSI. Operator rankings vary with fusion stage and backbone: LeViT shows clearer benefits from feature-level **ATT** and **CNN**, whereas MLP improvements are smaller and less consistent. At logit level, **WAVG** shows the most consistent, although modest, gains across backbones. In contrast, unrestricted linear mixing and independent gating are less effective for high-dimensional feature fusion. The qualitative example illustrates stage-dependent behavior, with feature-level fusion producing more localized changes and logit-level fusion yielding more conservative predictions.

6. REFERENCES

[1] Hunanyan et al., "Deep learning classification of ablated atrial tissue in multi-excitation hyperspectral images," in *WHISPERS*, 2025, pp. 1–5.

[2] Ter-Petrosyan et al., "Comparative analysis of multi-excitation hyperspectral image configurations for ablated atrial tissue classification," in *WHISPERS*, 2025, pp. 1–5.

[3] Gabrielyan et al., "Comparative analysis of deep learning methods for classification of ablated regions in hyperspectral images of atrial tissue," *IEEE Access*, vol. 13, pp. 35029–35047, 2025.

[4] Zhang et al., "S2dbft: Spectral–spatial dual-branch fusion transformer for hyperspectral image classification," *IEEE TGRS*, vol. 63, pp. 1–17, 2025.

[5] Cheng et al., "Cacftnet: A hybrid cov-attention and cross-layer fusion transformer network for hyperspectral image classification," *IEEE TGRS*, vol. 62, pp. 1–17, 2024.

[6] Vaswani et al., "Attention is all you need," in *NIPS*, 2017, vol. 30.

[7] Hu et al., "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[8] Woo et al., "Cbam: Convolutional block attention module," in *ECCV*, September 2018.

[9] Li et al., "Selective kernel networks," in *CVPR*, 2019, pp. 510–519.

[10] Arevalo et al., "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[11] Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *ACL*, 2019, pp. 6558–6569.

[12] Nagrani et al., "Attention bottlenecks for multimodal fusion," in *NeurIPS*, 2021, vol. 34, pp. 14200–14213.

[13] Muselimyan et al., "Anatomical and optical properties of atrial tissue: Search for a suitable animal model," *Cardiovasc Eng Technol*, vol. 8, no. 4, pp. 505–514, 2017.

[14] Graham et al., "LeViT: A vision transformer in ConvNet's clothing for faster inference," in *ICCV*, 2021, pp. 12259–12269.