Keep It Frozen: Domain-Routed Conditional Residual Modulation for Multi-Domain Vision Transformers

Supplementary Material

1. Methodology

In this section, we provide further details for the proposed methodology of DCRM-ViT model.

1.1. Image Encoder

The input image is first divided into non-overlapping patches of size $p \times p$. These patches are then linearly embedded into vectors and augmented with positional embeddings to retain spatial information. Given an image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ with height H, width W, and channels C, the image is divided into $N_p = \frac{HW}{p^2}$ patches. Each patch $\mathbf{p}_i \in \mathbb{R}^{p \times p \times C}$ is projected to a vector $\mathbf{e}_i \in \mathbb{R}^D$ using a linear layer, where D is the embedding dimension. The set of patch embeddings is given by $\mathbf{E} = \{\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_{N_p-1}\}$, where $\mathbf{e}_0$ is a learnable class token added to aggregate global image information. The embeddings and positional embeddings are fed into DR. and a series of RMB modules, generating new token representations that encode the image features.

1.2. Soft Parameter Sharing

We implement a soft-parameter sharing mechanism to effectively utilize knowledge from both domains. The final RMB parameters are computed by weighting the domain-specific parameters according to the domain probabilities. This can be mathematically written as $\theta_A = p_{\text{medical}}\theta_{A,\text{medical}} + p_{\text{natural}}\theta_{A,\text{natural}}$. This approach enables the model to handle images with features from different domains or ambiguous characteristics, thus enhancing its robustness and generalization.

1.3. Regularization of Dynamic Parameters

To prevent the dynamically generated parameters and domain attention biases from deviating excessively from the base parameters, we introduce a regularization term in the loss function as $\mathcal{L}_{\text{reg}} = \lambda \left(\|\theta_A - \theta_{A,\text{base}}\|^2 + \|\mathbf{B}_d - \mathbf{B}_{\text{base}}\|^2 \right)$. Here, $\theta_{A,\text{base}}$ and $\mathbf{B}_{\text{base}}$ are the base RMB parameters and attention biases (initialized from the pre-trained weights of ViT), and λ is a regularization coefficient. This regularization ensures that the model retains the foundational knowledge from the pre-trained model while adapting to domain-specific nuances.

1.4. Drop Path Regularization

Drop path regularization is implemented to enhance the generalization capability of the DCRM-ViT model and prevent overfitting. This technique randomly drops Transformer blocks during the training phase, encouraging the model to develop redundant paths for information processing. This increases its fault tolerance and reduces dependence on any single path during inference. The drop path regularization can be mathematically expressed as:

$$\mathbf{h}_{\text{drop}} = \begin{cases} \mathbf{h} & \text{with probability } (1-p) \\ 0 & \text{with probability } p \end{cases} \tag{1}$$

where p is the drop path rate.

1.5. Robustness to Noise.

The robustness of domain probabilities to noise in medical data is a crucial concern due to the inherently noisy nature of such data. Here, the domain classifier is trained not only on clean images but also on medical images, learning the discriminative nature between both domains. In case we remove the noise completely, then the model will probably get overfit and will not be able to perform well when shown some real-time clinical medical images, which definitely would contain the noise. So, the model should be adapted to the nature of the medical domain accordingly. Therefore, we have already incorporated the Drop Path (as explained above) and regularization approach within the domain classifier to prevent overfitting to noisy or outlier data points, enhancing the generalization capabilities of the DR. module across noisy inputs. Table 1 also shows the comparison of different modules used in our model along with their different characteristics.

1.6. Zero-Shot Transformation

This adaptation process is formulated as an optimization problem that aims to minimize the squared euclidean distance between the transformed feature representations of the source models and the corresponding CLIP embeddings, with an added regularization term to mitigate overfitting. Here, the source model corresponds to the model that lacks direct zero-shot capability. The mathematical expression for this optimization function includes a Frobenius norm of W as the regularization term:

$$\min_{W,b} \left\{ \sum_{i=1}^{N} \|W f_{\text{source}}(\mathbf{x}_i) + b - f_{\text{CLIP}}(\mathbf{x}_i)\|_2^2 + \lambda \|W\|_F^2 \right\} \tag{2}$$

Where λ is a regularization parameter that helps balance the fit and complexity of the model to enhance generalization. The transformation W and bias b are updated iteratively using stochastic gradient descent, with the update rules defined

Figure 1. DCRM-ViT results using ultrasound and natural images, showing labels and confidence levels

as:

$$W^{(t+1)} = W^{(t)} - \eta \left(\nabla_W L(W, b; \mathbf{x}^{(t)}) + \lambda W^{(t)} \right) \quad (3)$$

$$b^{(t+1)} = b^{(t)} - \eta \nabla_b L(W, b; \mathbf{x}^{(t)}) \quad (4)$$

$$L(W, b; \mathbf{x}) = \sum_{i=1}^{n} \| W f_{\text{source}}(\mathbf{x}_i) + b - f_{\text{CLIP}}(\mathbf{x}_i) \|_2^2 \quad (5)$$

Here, η denotes the learning rate and t represents the iteration number. After alignment, the effectiveness of the adaptation is quantified by computing the cosine similarity between the transformed feature vectors and the CLIP embeddings. This metric provides a measure of how well the adapted model features align with the multimodal semantic space of CLIP:

$$\text{cosine\_similarity} = \frac{\langle W f_{\text{source}}(\mathbf{x}), f_{\text{CLIP}}(\mathbf{x}) \rangle}{\| W f_{\text{source}}(\mathbf{x}) \| \| f_{\text{CLIP}}(\mathbf{x}) \|} \quad (6)$$

A high cosine similarity indicates a successful alignment, affirming the model's capability to perform zero-shot classification by interpreting textual descriptions associated with images. This enables these traditionally non-zero-shot models to recognize and categorize images without explicit prior training on specific class labels. This adaptation extends the utility of these models for advanced applications where labels are scarce or unavailable, enhancing their applicability in diverse real-world scenarios.

2. Datasets

This paper incorporates ultrasound, CT, and MRI, as well as natural imagery datasets to comprehensively evaluate the effectiveness of the models. Further details for each modality is provided below.

2.1. Ultrasound Datasets

FPUS23. The FPUS23 dataset [14] utilizes a simulated 23-week gestation fetus phantom to overcome the ethical challenges associated with actual patient data. This dataset, comprising 15,728 ultrasound images captured with the Philips Epiq-7 system, is enhanced by Anatomically Intelligent Ultrasound (AIUS) technology. It includes images categorized into four classes: Head, Abdomen, Arms, and Legs, demonstrating its diversity and applicability for anatomical studies. We filtered the images from this dataset so that each image would have a single class like arm or an abdomen, and then applied various augmentations to increase the size of the dataset.

Fetal Planes DB. The Fetal Planes (FP) dataset [4] features over 12,400 annotated ultrasound images from 1,792 patients. These images are sorted into six categories reflecting key fetal anatomical planes used in prenatal screenings: Abdomen, Brain, Femur, Thorax, and the maternal cervix. Additional categorization into sub-planes, including trans-thalamic, trans-cerebellum, and trans-ventricular, provides detailed insights into brain anatomy, although these are not the primary focus of this paper.

BUS-UCLM. [18] is a curated set of 683 breast ultrasound images from 38 patients acquired at Ciudad Real General University Hospital on a Siemens ACUSON S2000 with an 18L6 HD probe. Each image has an expert RGB mask (black = normal, green = benign, red = malignant) and having a total class count of 419 normal, 174 benign, 90 malignant samples, respectively.

BUID. [1] is a collection of 780 PNG images (500×500 px) from 600 female patients collected at Baheya Hospital (Cairo) using LOGIQ E9 / LOGIQ E9 Agile systems. It contains normal, benign, and malignant classes with per-image freehand mask ground truths. It has a class count of 487 benign, 210 malignant, and 133 normal, respectively.

BUS-BRA.[7] is a public dataset of 1,875 anonymized breast ultrasound images from 1,064 female patients acquired on four scanners at Brazil's National Institute of Cancer. It provides biopsy-proven labels (722 benign, 342 malignant), expert lesion masks, and standardized 5- and 10-fold cross-validation splits.

2.2. CT/MRI Datasets

ACDC. [2] is a cine-MRI dataset of 150 patients evenly split into five diagnostic groups (NOR, MINF, DCM, HCM, ARV), acquired over six years on 1.5T/3T Siemens scanners with SSFP sequences. The training set includes 100 labeled subjects, and the test set contains 50. Expert annotations are provided at the ED / ES for the LV / RV cavities and the LV myocardium.

MMWHS. [21] is a challenge dataset that provides 120 clinical 3D cardiac volumes (60 CT and 60 MRI) covering the whole heart. It targets segmentation of standard heart substructures and enables consistent benchmarking across modalities.

2.3. Natural Imagery Datasets

For natural-image evaluation, we use CIFAR-10 [12], Caltech-101 [6], Natural Images [16], Food-101 [3], SUN397 [19], and Stanford Cars [11].

Algorithm 1 Training Procedure for Model Adaptation

Initialization: Start with pre-trained transformer parameters θ, initial RMB parameters ϕ_{base}, DR. module parameters ω_{base}, and learning rates α and β.
for each task T **do**
  Domain Probability Estimation: Use the domain classifier D to compute $D(\mathbf{x})$ for images in $\mathcal{D}_T^{\text{task}}$ and $\mathcal{D}_T^{\text{dom}}$.
  Inner Loop:
    Generate RMB parameters ϕ using the PSN unit and domain probabilities.
    Fine-tune ϕ on $\mathcal{D}_T^{\text{task}}$.
    Update ϕ'_T for task T.
  Outer Loop:
    Compute $\mathcal{L}_{\text{total}}$ on $\mathcal{D}_T^{\text{dom}}$.
    Update ω, ϕ, and α.
end for
Repeat:
  Iterate over all tasks, performing inner and outer loop updates until convergence.

3. Qualitative Results

Figure 1 illustrates the correctly classified examples from a batch of test data spanning both fetal-ultrasound (arm, brain, femur) and natural-image (flower, car, dog, cat) domains. In all cases, the model's top-1 prediction matches the ground truth with high confidence (0.93–0.98), demonstrating robust domain-agnostic recognition that complements the quantitative gains reported earlier. We also showed the attention map outputs in Figure 2.

4. Experimental Setup

The experiments are conducted on NVIDIA A100 GPU with 40 GB of VRAM and 128 GB of RAM. Adam [10] optimizer was used for all model training, whereas SGD [17], along with the CosineAnnealingLR scheduler, was used to adapt the self-supervised models to zero-shot configurations. We used different models like MAE [8], DINOv2 [13], and CLIP [15], as well as the low-rank methods like Tip-Adapter [20], AdaptFormer [5], and LoRA [9]. The batch size of 32 was used for all the datasets. We use the inner-loop learning rate to be relatively high with value of 1×10^{-3} so that each RMB can rapidly adapt to its task-specific data. In contrast, we use a bit lower learning rate of 1×10^{-4} as the outer loop updates the parameters more conservatively. We also provide our pseudo-algorithm in Algorithm 1.

For validation and performance testing, a split of 60-20-20 is used, where 60% of the data from each dataset is used for training ($\mathcal{D}_T^{\text{dom}}$, $\mathcal{D}_T^{\text{task}}$), 20% for validation, and the remaining 20% for testing. The difference between $\mathcal{D}_T^{\text{dom}}$ and $\mathcal{D}_T^{\text{task}}$ is that the former is used as a binary classification problem where the classes are either fetal or natural domain. In contrast, the latter refers to the main task classes like fetal arm, head, abdomen, etc. We evaluate the performance of each model using two primary metrics: accuracy and F1 score. Accuracy measures the proportion of correct predictions made by the model out of all predictions. In contrast, the F1 Score assesses the model's precision and recall balance, which is particularly useful in scenarios with imbalanced datasets.

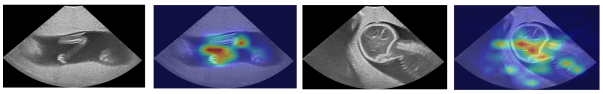

Figure 2. Visualization of attention maps

5. Extended Ablations

To enlighten the impact of drop path regularization on the performance of the DCRM-ViT model, we conducted three experiments across two datasets, FPUS23 and FP. The results, as delineated in Table 3, illustrate that a moderate drop path rate of 0.1 enhances both accuracy and F1 scores, particularly on the FP dataset. Conversely, increasing the drop path rate to 0.3 decreased performance metrics, underscoring the critical need for precise tuning of regularization parameters to maintain the delicate balance between model robustness and learning efficiency. Table 2 also shows the different ablations conducted for the RMB depth and the backbone replacement.

Additional conditional baseline. To test whether a simpler conditional design is sufficient, we also compare against a strong baseline that appends learned domain tokens/prompts to the ViT input and uses a standard high-rank adapter in each block, with trainable parameters matched to DCRM-ViT. This baseline improves over static adapters but remains below DCRM-ViT, showing that static domain prompting is weaker than our per-sample DR.→PSN conditioning. On FPUS23, DCRM-ViT outperforms this matched baseline by +12.3% accuracy.

6. Limitations and Future Work

The granularity of DCRM-ViT's DR. module may not be sufficient for complex medical scenarios, as it primarily distinguishes between broad image categories without delving into finer medical subdomains. This limits its diagnostic capabilities in environments that require detailed differentiation among medical conditions. Moreover, DCRM-ViT's

Table 1. Comparison of module types in the DCRM-ViT framework

Module Type	Primary Role	Input Sensitivity	Mechanism	Benefit
DR.	Domain-aware modulation	Varies per input (e.g., fetal vs. general images)	Gate-channel architecture with PSN-generated parameters	Enables input-dependent adaptation across domains
RMB	Task-specific fine-tuning	Fixed per task	Lightweight residual bottleneck modules	Enables parameter-efficient customization and improved classification performance

Table 2. Compact ablation on RMB depth (left block) and backbone choice (right block) in DCRM-ViT. The left value shows accuracy while right value corresponds to F1 score.

Dataset	# RMB layers			Backbone		
	8	12	14	ResNet-50	VGG-16	ViT-B/16
FPUS23	61.4±1.3 / 0.61±0.04	63.4±0.9 / 0.69±0.03	62.7±1.1 / 0.65±0.04	60.5±1.5 / 0.61±0.05	59.2±1.2 / 0.60±0.03	63.4±0.8 / 0.62±0.02
Fetal Planes	87.2±0.7 / 0.87±0.02	89.2±0.6 / 0.90±0.01	88.2±0.8 / 0.88±0.03	87.0±0.9 / 0.87±0.04	85.5±1.0 / 0.86±0.05	89.2±0.5 / 0.88±0.02

Table 3. Impact of drop path regularization on model performance across FPUS23 and FP datasets.

Drop Path Rate	FPUS23		Fetal Planes	
	Accuracy (%)	F1 Score	Accuracy (%)	F1 Score
0.0	63.1 ± 1.4	0.67 ± 0.05	88.5 ± 0.7	0.88 ± 0.02
0.1	63.4 ± 0.8	0.69 ± 0.03	89.0 ± 0.4	0.89 ± 0.01
0.2	62.9 ± 1.2	0.68 ± 0.06	89.3 ± 0.9	0.90 ± 0.03
0.3	62.3 ± 1.0	0.67 ± 0.04	88.7 ± 0.5	0.89 ± 0.04

performance heavily depends on the quality of its underlying pre-trained backbone ViT models. By having inadequate or non-representative pre-training, the model's effectiveness can be affected in specialized medical tasks, affecting its adaptability and precision. Moreover, our evaluation focuses on static 2D scans, which shows that DCRM-ViT has not yet been tested on video sequences or other medical imaging modalities such as X-ray. Finally, DCRM-ViT is based on classification and segmentation tasks, whereas many clinical applications (e.g., landmark detection) require more dense predictions.

To refine DCRM-ViT, we aim to improve the ability of the DR. module to recognize and adapt to more specific medical modalities such as PET, X-Ray, and other scans, which will cover a broad spectrum of adaptability in the medical domain. Additionally, the current structure of DCRM-ViT follows a linear approach, where we tend to explore other parallel options to combine the different modules, like having separate RMB branches (e.g., one branch for natural features, one for the MRI domain, and one for fetal), followed by a fusion mechanism. Furthermore, we will generalize the DR. + RMB integration for dense prediction by embedding domain-aware calibration directly into landmark-detection heads, thereby enabling pixel-level refinement of fetal structures. Finally, to enable video analysis, we also look forward to incorporating temporal feature fusion, for example, via a lightweight spatio-temporal transformer or recurrent memory tokens to enforce consistency across video frames and improve robustness to motion blur.

7. Key Insights.

There is a notable difference in the zero-shot performance between fetal and natural image datasets. However, it is worth noting that this accuracy has been reported in zero-shot settings while using the feature transformation approach. This lower performance can be improved further by using several targeted improvements, like expansion of training data to cover more variations of the medical domain, implementation of a contrastive pre-training approach specifically for the fetal domain, and modifying the transformation approach by using CLIP embedding space tailored for the fetal domain rather than the natural imagery domain. Moreover, during the feature transformation, we used the normal CLIP embedding space, which limits the capability of DCRM-ViT.

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 2

[2] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018. 2

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 3

[4] Xavier P Burgos-Artizzu, David Coronado-Gutiérrez, Brenda Valenzuela-Alcaraz, Elisenda Bonet-Carne, Elisenda Eixarch, Fatima Crispi, and Eduard Gratacós. Evaluation of deep convolutional neural networks for automatic classification of

common maternal fetal ultrasound planes. *Scientific Reports*, 10(1):10200, 2020. 2

[5] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 3

[6] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 3

[7] Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4):3110–3123, 2024. 2

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 3

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[11] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 3

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *arXiv preprint arXiv:1807.10108*, 2009. 3

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[14] Bharath Srinivas Prabakaran, Paul Hamelmann, Erik Ostrowski, and Muhammad Shafique. Fpus23: An ultrasound fetus phantom dataset with deep neural network evaluations for fetus orientations, fetal planes, and anatomical features. *IEEE Access*, 2023. 2

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[16] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018. 3

[17] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 3

[18] Noelia Vallez, Gloria Bueno, Oscar Deniz, Miguel Angel Rienda, and Carlos Pastor. Bus-uclm: Breast ultrasound lesion segmentation dataset. *Scientific Data*, 12(1):242, 2025. 2

[19] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 3

[20] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 3

[21] Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58:101537, 2019. 2