

Multi-Kernel Gated Decoder Adapters for Robust Multi-Task Thyroid Ultrasound under Cross-Center Shift

Maziar Sabouri^{1,2}, Nourhan Bayasi^{2,3}, and Arman Rahmim^{1,2,3,4}

¹ Department of Astronomy and Physics, University of British Columbia, Vancouver, Canada

² Department of Basic and Translational Research, BC Cancer Research Institute, Vancouver, Canada

³ Department of Radiology, University of British Columbia, Vancouver, Canada

⁴ School of Biomedical Engineering, University of British Columbia, Vancouver, Canada

Abstract. Thyroid ultrasound (US) automation couples two competing requirements: global, geometry-driven reasoning for nodule delineation and local, texture-driven reasoning for malignancy risk assessment. Under cross-center domain shift, these cues degrade asymmetrically, yet most multi-task pipelines rely on a single shared backbone, often inducing negative transfer. In this paper, we characterize this interference across CNN (ResNet34) and medical ViT (MedSAM) backbones, and observe a consistent trend: ViTs transfer geometric priors that benefit segmentation, whereas CNNs more reliably preserve texture cues for malignancy discrimination under strong shift and artifacts. Motivated by this failure mode, we propose a lightweight family of decoder-side adapters, the Multi-Kernel Gated Adapter (MKGA) and a residual variant (ResMKGA), which refine multi-scale skip features using complementary receptive fields and apply semantic, context-conditioned gating to suppress artifact-prone content before fusion. Across two US benchmarks, the proposed adapters improve cross-center robustness: they boost out-of-domain segmentation and, in the CNN setting, yield clear gains in clinical TI-RADS diagnostic accuracy compared to standard multi-task baselines. Source code: <https://github.com/MaziarSabouri/MKGA.git>.

Keywords: Thyroid Ultrasound · Domain Shift · Multi-task Learning · Parameter-Efficient Adaptation · Gated Adapters

1 Introduction

Thyroid nodules are routinely evaluated with ultrasound (US) to guide follow-up, biopsy decisions, and risk stratification under the 5-point scoring system TI-RADS, with fine-needle aspiration cytology (FNAC) as the diagnostic reference standard [3,2]. While deep learning has achieved strong performance for nodule segmentation and malignancy prediction in curated datasets, reliable

cross-center generalization remains a major barrier to clinical adoption. In practice, thyroid US appearance varies substantially across institutions due to scanner vendor and settings, acquisition protocol, operator technique, and frequent on-image overlays such as calipers and text markup [13,5]. These shifts often break models that appear accurate in-distribution, especially when systems are expected to support both contouring (for measurement and follow-up) and risk assessment (for triage).

A key reason is that thyroid US automation couples two qualitatively different reasoning modes. Accurate segmentation is largely geometry-driven: it requires global context to delineate uncertain boundaries, enforce plausible shape, and remain stable under speckle noise and weak edges [7]. In contrast, malignancy assessment is more texture-driven, relying on subtle local cues aligned with TI-RADS descriptors (e.g., echogenicity, internal echoes, micro-calcifications) that can be sensitive to scanner-dependent statistics and overlay artifacts [6]. Critically, these cues degrade asymmetrically under domain shift: artifacts may preserve coarse shape while corrupting high-frequency texture patterns, or conversely alter apparent boundaries while leaving local texture relatively unchanged.

Despite this asymmetry, many pipelines treat segmentation and diagnosis as multi-task learning (MTL) with a single shared backbone, implicitly assuming one representation can serve both geometric and textural objectives [8]. Under cross-center shift, this assumption becomes brittle: competing supervision can drive a shared encoder toward an inductive bias that benefits one task while harming the other, resulting in negative transfer and unstable deployment behavior [15]. We argue that robust multi-task thyroid US systems must address these competing objectives via targeted feature refinement in the decoder and gradient-aware optimization, rather than relying on a naïve shared encoder.

In this work, we address this failure mode with two contributions. First, we empirically characterize multi-task interference under cross-center shift by evaluating both a CNN backbone (ResNet34) and a medical foundation ViT (MedSAM), revealing that geometry- and texture-driven objectives degrade differently and can conflict when forced through a shared encoder. Second, we propose a lightweight family of decoder-side adapters, the Multi-Kernel Gated Adapter (MKGA) and its residual variant ResMKGA, which improve artifact-robust skip fusion via multi-kernel refinement and context-conditioned gating (with residual channel recalibration in ResMKGA). To further reduce optimization conflict, we optionally apply gradient surgery (PCGrad) during multi-task training. Across ThyroidXL (in-domain) and DDTI (held-out cross-center external) evaluation, our approach strengthens cross-center robustness: it consistently improves out-of-domain segmentation and, in the CNN setting, yields clear gains in TI-RADS-based risk stratification over strong shared-backbone baselines.

2 Methodology

2.1 Overview. We target robust multi-task thyroid US automation under cross-center domain shift by reducing negative transfer between geometry-driven seg-

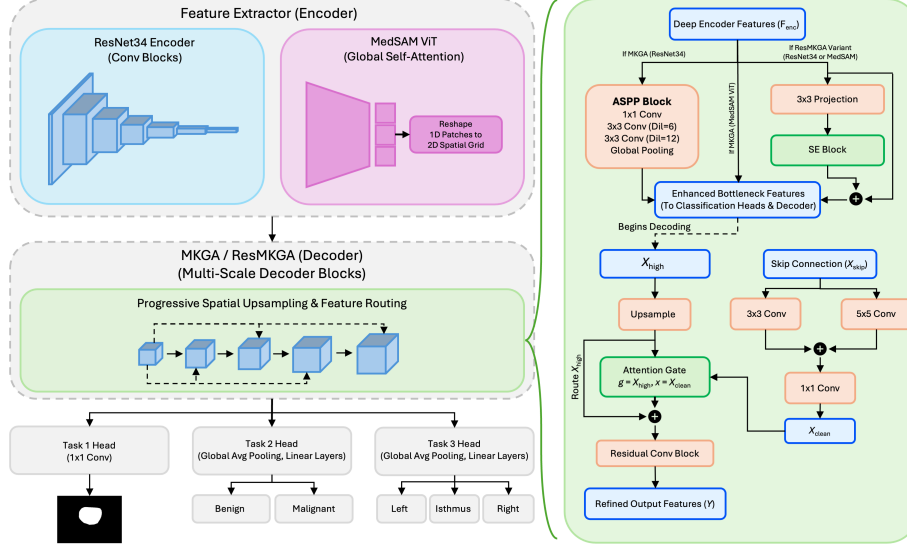


Fig. 1: Overview of the proposed MKGA and ResMKGA decoders architecture.

mentation and texture-driven classification. Our unified architecture shares a single backbone and predicts three outputs: nodule segmentation, TI-RADS malignancy, and anatomical position. To make decoding resilient to US-specific artifacts (speckle, calipers, and text overlays), we introduce a lightweight decoder adaptation family, MKGA and its residual variant ResMKGA (Fig. 1), which refine and selectively gate multi-scale skip features using semantic context. During training, we optimize all tasks jointly, and optionally apply gradient surgery (PCGrad) [22] to mitigate conflicting gradients in the shared encoder.

2.2 Backbones and Multi-Scale Decoding. We investigate two backbones for the unified multi-task framework. First, we adopt MedSAM [10], a SAM-style ViT encoder [9], motivated by its global self-attention and strong geometric priors. Since MedSAM produces a single-scale latent map, we support multi-stage decoding by constructing a progressive pseudo-skip pyramid using learnable up-sampling and projection layers to generate a sequence of aligned feature maps for the decoder stages. Second, we evaluate ResNet34 [11] for its hierarchical receptive fields and locality bias, which are well-suited to US texture cues. In this configuration, standard encoder skip features are extracted from intermediate stages and fed into the decoder. To capture global context prior to decoding, the deepest features of the ResNet34 encoder are specifically processed through an Atrous Spatial Pyramid Pooling (ASPP) bottleneck [16]. We attach three task-specific heads to the unified framework: a segmentation decoder, and two global pooling heads for TI-RADS malignancy and anatomical position classification.

2.3 Multi-Kernel Gated Adapter (MKGA). Direct skip fusion in a unified multi-task system can inject artifact-driven noise from shallow features into the

decoder, potentially destabilizing both segmentation and classification performance. We therefore introduce MKGA, a lightweight module that refines skip features using complementary receptive fields and gates skip content conditioned on semantic context before residual fusion. Let $X_{\text{skip}} \in \mathbb{R}^{H \times W \times C_s}$ denote the encoder skip feature at a given scale, and let $X_{\text{high}} \in \mathbb{R}^{H \times W \times C_h}$ denote the up-sampled deeper decoder feature aligned to the same resolution. MKGA consists of three components:

(1) *Multi-kernel skip refinement.* We apply parallel convolutions to capture multi-scale context from X_{skip} : a 3×3 convolution and a dilated 3×3 convolution with dilation $d=2$ (yielding a 5×5 receptive field). The outputs are concatenated and projected as $X_{\text{skip}}^{\text{ref}} = \phi_{1 \times 1}(\text{Conv}_{3 \times 3}(X_{\text{skip}}) \parallel \text{Conv}_{3 \times 3, d=2}(X_{\text{skip}}))$, where $\parallel$ denotes channel concatenation and $\phi_{1 \times 1}$ is a 1×1 projection followed by normalization and ReLU.

(2) *Context-conditioned gating.* To suppress irrelevant skip activations, we use an additive attention gate conditioned on X_{high} . The attention map is computed as $\alpha = \sigma(\psi(\delta(W_g(X_{\text{high}}) + W_s(X_{\text{skip}}^{\text{ref}}))))$, where σ is Sigmoid, δ is ReLU, and ψ, W_g, W_s are 1×1 convolutions. The gated skip feature is $X_{\text{skip}}^{\text{gate}} = \alpha \odot X_{\text{skip}}^{\text{ref}}$, with $\odot$ denoting element-wise multiplication.

(3) *Residual fusion.* We fuse semantic and gated skip features via concatenation followed by a lightweight residual refinement block: $Y_{\text{MKGA}} = \mathcal{F}_{\text{res}}(X_{\text{high}} \parallel X_{\text{skip}}^{\text{gate}})$, where $\mathcal{F}_{\text{res}}(\cdot)$ denotes two sequential 3×3 convolutions, each followed by normalization and ReLU.

2.4 Residual Bottleneck Variant (ResMKGA). To stabilize the deepest latent representation where multi-task objectives often conflict under domain shift, we introduce ResMKGA. This module applies a residual correction to the encoder bottleneck feature F_{enc} before decoding. Specifically, we compute $X_{\text{high}}^{\text{enh}} = F_{\text{enc}} + \text{SE}(\phi_{3 \times 3}(F_{\text{enc}}))$, where $\phi_{3 \times 3}$ is a 3×3 projection and $\text{SE}(\cdot)$ denotes a Squeeze-and-Excitation block [17] with a two-layer bottleneck MLP for channel recalibration. For ResNet34, this serves as a lightweight alternative to heavy multi-scale context heads; for MedSAM, it acts as a task-specific adapter that aligns the foundation-model features to the broader requirements of multi-task thyroid US decoding and classification.

2.5 Training Objectives. We optimize the unified model with task-appropriate losses. For segmentation, we use a compound Dice and pixel-wise Cross-Entropy objective, $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}^{\text{pixel}}$. For malignancy and anatomical position classification, we use image-level Cross-Entropy, $\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{CE}}^{\text{image}}$. The overall multi-task objective is a weighted sum $\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda_{\text{mal}} \mathcal{L}_{\text{mal}} + \lambda_{\text{pos}} \mathcal{L}_{\text{pos}}$, with λ set in our experiments (and PCGrad optionally applied to reduce gradient conflict).

3 Experiments and Results

3.1 Datasets and Evaluation Protocol. We evaluate cross-center robustness using an in-domain split from the training center and an external out-of-domain benchmark from a different center. All models are trained using ThyroidXL [12] (11,635 images, 4,093 patients), which provides nodule masks, TI-RADS labels,

and anatomical position (left/right/isthmus). We perform a patient-level split with 80% development and 20% in-domain test. The development set is further split into training/validation (80/20) for model selection. External generalization is assessed on DDTI [13], which is used *only* for testing and contains substantial artifacts, including embedded calipers and text overlays in some cases. Following preprocessing, we split dual-view acquisitions into single-view samples and remove border artifacts, yielding 660 images. DDTI provides masks and TI-RADS labels but does not include position labels. For malignancy risk, we binarize TI-RADS into Low-Risk (≤ 3) vs. High-Risk (≥ 4) for both datasets [3,18,19].

3.2 Implementation Details. Images are zero-padded to square to preserve aspect ratio and resized to 512×512 ; intensities are normalized using ImageNet statistics. To model inter-center variability during training, we apply appearance perturbations (Gaussian noise, Gaussian blur, multiplicative noise) and mild geometric transforms (affine, scale $< 10\%$, rotation $\pm 15^\circ$, elastic deformation). Horizontal flipping is disabled to preserve laterality cues for the positioning task. Models are trained in PyTorch using AdamW ($\text{lr} = 10^{-4}$, batch size = 16) for up to 100 epochs with early stopping (patience = 15) based on validation loss.

3.3 Tasks, Metrics, and Statistical Testing. We evaluate (i) segmentation, (ii) binary TI-RADS malignancy classification, and (iii) 3-way anatomical positioning (ThyroidXL only). Segmentation uses Dice and IoU; classification uses accuracy, F1, and ROC AUC. We test improvements over baselines using Wilcoxon signed-rank (segmentation), McNemar (accuracy), and DeLong (ROC AUC; multi-class positioning via flattened one-hot encoding), with Benjamini–Hochberg FDR correction and significance at adjusted $p < 0.05$.

3.4 Adaptation Settings & Baselines. To evaluate clinical deployment scenarios, we consider two regimes: (i) *Frozen backbone*, training only the decoder and task heads to isolate MKGA’s contribution as a Parameter-Efficient Fine-Tuning (PEFT) module and probe feature transferability; and (ii) *Unfrozen adaptation*, permitting end-to-end optimization to test architectural synergy. Specifically, we fully fine-tune ResNet34 and apply LoRA [14] (ranks $r \in \{4, 16, 32\}$) to MedSAM’s attention layers. Multi-source domain generalization frameworks (e.g., IRM [20], DomainBed [21]) are inapplicable to our single-center cohort and thus excluded from baselines.

3.5 Cross-Center Segmentation Robustness. Table 1 shows that most methods perform well in-domain on ThyroidXL (Dice ≈ 0.80 – 0.87 ; best in-domain Dice: ResNet34+MKGA (Frozen), 0.869 ± 0.130), but all exhibit a cross-center generalization gap on DDTI. For CNNs, naïve end-to-end fine-tuning is brittle: ResNet34 (Unfrozen) drops from 0.861 ± 0.130 (in-domain) to 0.590 ± 0.262 (external). In contrast, adding MKGA/ResMKGA markedly improves external stability (DDTI Dice 0.659 ± 0.236 and 0.671 ± 0.256 , respectively; $p < 0.05$ vs. ResNet34 unfrozen). While PCGrad improves the unfrozen baseline to 0.642 (Table 2), ResMKGA still surpasses this PCGrad-enhanced baseline ($p < 0.05$), indicating that artifact-aware skip refinement and context-conditioned gating provide architectural robustness that reduces reliance on optimization-only fixes for segmentation. For ViTs, MedSAM+ResMKGA+LoRA ($r=4$) achieves the

Table 1: Segmentation and diagnostic (TI-RADS) performance evaluated on ThyroidXL and DDTI, and positioning on ThyroidXL only. Best in **Bold**.

Model Architecture	ThyroidXL (In-domain Test Set)									DDTI (External Test Set)				
	Segmentation			TI-RADS			Position			Segmentation		TI-RADS		
	Dice	IoU	Acc	F1	AUC	Acc	F1	AUC		Dice	IoU	Acc	F1	AUC
<i>Convolutional Neural Networks (ResNet34)</i>														
ResNet34 (Frozen)	0.857	0.764	0.855	0.857	0.920	0.855	0.852	0.947		0.631	0.512	0.357	0.396	0.556
ResNet34 (Unfrozen)	0.861	0.772	0.872	0.873	0.931	0.864	0.863	0.953		0.590	0.466	0.406	0.466	0.577
ResNet34 + MKGA (Frozen)	0.869	0.784	0.873	0.873	0.915	0.870	0.868	0.952		0.671	0.551	0.430	0.489	0.600
ResNet34 + MKGA (Unfrozen)	0.862	0.775	0.872	0.872	0.922	0.875	0.870	0.956		0.659	0.533	0.632	0.683	0.642
ResNet34 + ResMKGA (Frozen)	0.853	0.761	0.862	0.862	0.923	0.856	0.855	0.949		0.596	0.474	0.676	0.716	0.609
ResNet34 + ResMKGA (Unfrozen)	0.865	0.777	0.871	0.872	0.934	0.865	0.866	0.951		0.671	0.553	0.492	0.557	0.620
<i>Vision Transformers (MedSAM)</i>														
MedSAM (Frozen)	0.799	0.690	0.674	0.679	0.699	0.727	0.699	0.832		0.637	0.507	0.687	0.709	0.411
MedSAM + LoRA ($r = 4$)	0.826	0.727	0.769	0.767	0.811	0.831	0.822	0.928		0.631	0.504	0.426	0.497	0.499
MedSAM + LoRA ($r = 16$)	0.832	0.732	0.799	0.800	0.862	0.824	0.816	0.925		0.649	0.518	0.370	0.426	0.491
MedSAM + LoRA ($r = 32$)	0.826	0.726	0.753	0.752	0.792	0.818	0.801	0.926		0.634	0.510	0.395	0.461	0.511
MedSAM + MKGA (Frozen)	0.837	0.740	0.753	0.755	0.797	0.791	0.781	0.909		0.643	0.516	0.406	0.467	0.522
MedSAM + MKGA + LoRA ($r = 4$)	0.850	0.755	0.799	0.797	0.851	0.833	0.829	0.930		0.651	0.526	0.430	0.502	0.455
MedSAM + MKGA + LoRA ($r = 16$)	0.846	0.752	0.802	0.801	0.861	0.830	0.827	0.935		0.652	0.526	0.435	0.507	0.485
MedSAM + MKGA + LoRA ($r = 32$)	0.853	0.760	0.798	0.788	0.835	0.827	0.821	0.929		0.646	0.519	0.470	0.543	0.515
MedSAM + ResMKGA (Frozen)	0.836	0.739	0.761	0.759	0.803	0.802	0.789	0.911		0.649	0.523	0.497	0.568	0.528
MedSAM + ResMKGA + LoRA ($r = 4$)	0.853	0.760	0.800	0.796	0.851	0.821	0.813	0.935		0.675	0.548	0.463	0.537	0.478
MedSAM + ResMKGA + LoRA ($r = 16$)	0.855	0.762	0.776	0.781	0.850	0.816	0.801	0.931		0.663	0.542	0.359	0.414	0.512
MedSAM + ResMKGA + LoRA ($r = 32$)	0.850	0.758	0.803	0.795	0.858	0.830	0.822	0.934		0.655	0.531	0.514	0.585	0.509

Table 2: Impact of PCGrad on multi-task optimization. Best in **Bold**.

Model Architecture	ThyroidXL (In-domain Test Set)									DDTI (External Test Set)				
	Segmentation			TI-RADS			Position			Segmentation		TI-RADS		
	Dice	IoU	Acc	F1	AUC	Acc	F1	AUC		Dice	IoU	Acc	F1	AUC
<i>Convolutional Neural Networks (ResNet34)</i>														
ResNet34 (Unfrozen)	0.861	0.772	0.872	0.873	0.931	0.864	0.863	0.953		0.590	0.466	0.406	0.466	0.577
ResNet34 + PCGrad (Unfrozen)	0.859	0.768	0.872	0.874	0.934	0.873	0.872	0.959		0.642	0.517	0.594	0.653	0.633
ResNet34 + MKGA (Unfrozen)	0.862	0.775	0.872	0.872	0.922	0.875	0.870	0.956		0.659	0.533	0.632	0.683	0.642
ResNet34 + MKGA + PCGrad (Unfrozen)	0.867	0.779	0.869	0.870	0.929	0.873	0.870	0.954		0.635	0.510	0.554	0.616	0.637
ResNet34 + ResMKGA (Unfrozen)	0.865	0.777	0.871	0.872	0.934	0.865	0.866	0.951		0.671	0.553	0.492	0.557	0.620
ResNet34 + ResMKGA + PCGrad (Unfrozen)	0.867	0.779	0.858	0.838	0.926	0.873	0.871	0.957		0.645	0.521	0.601	0.659	0.603
<i>Vision Transformers (MedSAM)</i>														
MedSAM + LoRA ($r = 4$)	0.826	0.727	0.769	0.767	0.811	0.831	0.822	0.928		0.631	0.504	0.426	0.497	0.499
MedSAM + LoRA ($r = 4$) + PCGrad	0.835	0.737	0.798	0.800	0.865	0.825	0.819	0.934		0.629	0.501	0.366	0.426	0.491
MedSAM + ResMKGA + LoRA ($r = 4$)	0.853	0.760	0.800	0.796	0.851	0.821	0.813	0.935		0.675	0.548	0.463	0.537	0.478
MedSAM + ResMKGA + LoRA ($r = 4$) + PCGrad	0.858	0.767	0.799	0.800	0.858	0.827	0.825	0.935		0.658	0.536	0.512	0.583	0.437

best external Dice (0.675 ± 0.226), yet its advantage over lightweight ResNet34+ResMKGA (Unfrozen) and ResNet34+MKGA (Frozen) is not significant ($p > 0.05$), suggesting the decoder adapters are the main driver of robustness across backbones. Consistent with this, increasing LoRA rank degrades external Dice (0.675 at $r=4$ to 0.655 at $r=32$), supporting the hypothesis that higher-capacity adaptation overfits to source appearance rather than preserving transferable geometric structure. Qualitative examples are shown in Fig. 2.

3.6 Cross-Center Malignancy (TI-RADS) Generalization. Table 1 reports TI-RADS malignancy classification. Unlike segmentation, external performance is dominated by a *texture failure mode* for ViT-based encoders: MedSAM variants perform well in-domain (e.g., MedSAM+ResMKGA+LoRA $r=4$: Acc 0.800, AUC 0.851) but collapse on DDTI (AUC ≈ 0.48 –0.50), consistent

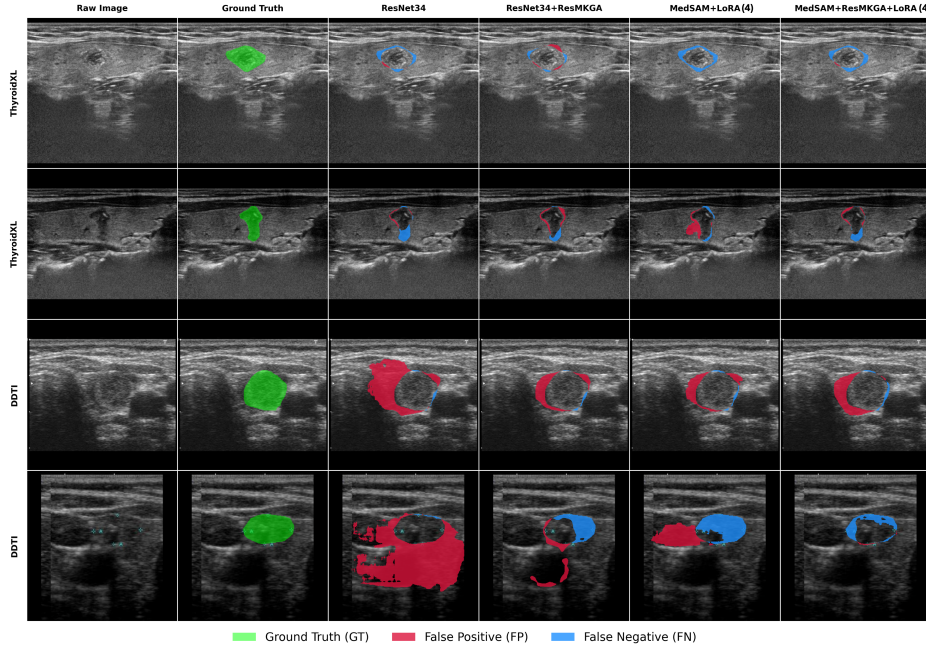


Fig. 2: Qualitative segmentation comparison on ThyroidXL and DDTI datasets.

with TI-RADS relying on subtle local, high-frequency cues that are strongly perturbed by cross-center artifacts (calipers/text) and scanner-dependent statistics. We attribute this failure to fundamental inductive biases: ViT self-attention acts as a spatial low-pass filter. While this aids coarse geometric reasoning for segmentation, it intrinsically smooths over the high-frequency micro-textures (e.g., microcalcifications) essential for diagnostic classification, and parameter-efficient LoRA cannot rewrite this core inductive bias. In contrast, the CNN backbone transfers substantially better under this shift. While ResNet34 (Unfrozen) achieves only 0.577 AUC on DDTI, ResNet34+MKGA reaches 0.642 AUC and yields a large, statistically significant gain in diagnostic decisions (Acc 0.406 $\rightarrow$ 0.632, McNemar $p < 0.001$), even though the AUC increase is not significant (DeLong $p=0.45$). PCGrad improves the unfrozen baseline (AUC 0.577 $\rightarrow$ 0.633) but does not provide consistent additional benefit once MKGA/ResMKGA are introduced, indicating that decoder-side multi-kernel refinement and context-conditioned gating already suppress artifact-driven activations while preserving discriminative texture features.

3.7 Anatomical Positioning. We evaluate anatomical position only on ThyroidXL, since DDTI lacks position labels (Table 1). CNN-based models outperform ViT variants on this task, with ResNet34+MKGA (Unfrozen) achieving the best in-domain performance (Acc 0.875, F1 0.870, AUC 0.956), consistent with CNN hierarchies better preserving global anatomical layout than patch-

Table 3: Ablation studies isolating the architectural components of the proposed MKGA and ResMKGA modules. Best in **Bold**.

Model Architecture	ThyroidXL (In-domain Test Set)									DDTI (External Test Set)				
	Segmentation			TI-RADS			Position			Segmentation		TI-RADS		
	Dice	IoU	Acc	F1	AUC	Acc	F1	AUC		Dice	IoU	Acc	F1	AUC
<i>Core MKGA Ablations (ResNet34 Backbone)</i>														
ResNet34 + MKGA (NoGate)	0.862	0.775	0.869	0.869	0.930	0.861	0.859	0.953		0.673	0.552	0.499	0.569	0.589
ResNet34 + MKGA (NoMulti)	0.855	0.768	0.862	0.863	0.922	0.874	0.871	0.955		0.629	0.509	0.359	0.402	0.565
ResNet34 + MKGA (K1_3)	0.868	0.780	0.874	0.874	0.924	0.872	0.870	0.953		0.652	0.527	0.479	0.547	0.591
ResNet34 + MKGA (K3_7)	0.862	0.774	0.869	0.870	0.922	0.859	0.855	0.948		0.604	0.477	0.497	0.558	0.661
ResNet34 + MKGA (K3_5)	0.862	0.775	0.872	0.872	0.922	0.875	0.870	0.956		0.659	0.533	0.632	0.683	0.642
<i>Residual MKGA Variant Ablations (ResNet34 Backbone)</i>														
ResNet34 + ResMKGA (NoSE)	0.864	0.774	0.873	0.874	0.939	0.873	0.868	0.955		0.639	0.524	0.548	0.613	0.615
ResNet34 + ResMKGA	0.865	0.777	0.871	0.872	0.934	0.865	0.866	0.951		0.671	0.553	0.492	0.557	0.620

based attention for coarse laterality cues. Adding PCGrad yields at most a negligible change (e.g., AUC 0.957 for ResNet34+ResMKGA+PCGrad; Table 2), and differences among the ResNet34 baselines and our decoder-adapter variants are not statistically significant ($p > 0.05$). Importantly, this indicates that MKGA/ResMKGA improve robustness in the harder cross-center tasks (segmentation and malignancy) *without* degrading positioning, supporting targeted decoder refinement as a structural solution to multi-task interference rather than relying on gradient surgery.

3.8 Ablation Studies. Table 3 isolates the architectural drivers of MKGA and ResMKGA on a ResNet34 backbone and reveals a consistent segmentation–diagnosis trade-off under cross-center shift. *Gating vs. raw skip fusion:* Removing the context-conditioned gate (NoGate) slightly improves external segmentation (Dice 0.673 vs. 0.659, $p < 0.001$) but sharply degrades TI-RADS generalization (AUC 0.589 vs. 0.642; Acc 0.499 vs. 0.632, $p < 0.001$), indicating that gating is primarily a *texture-preservation* mechanism that filters artifact-driven activations harmful for diagnosis, even if raw skip cues suffice for boundary delineation. *Multi-kernel refinement:* Removing the multi-kernel block (NoMulti) harms both tasks (DDTI Dice 0.629, $p = 0.011$; AUC 0.565) and causes a major collapse in diagnostic accuracy (Acc 0.359, $p < 0.001$), confirming that complementary receptive fields are necessary to capture multi-scale TI-RADS texture cues under shift. *Kernel design:* Small receptive fields 1×1 & 3×3 (K1_3) retain segmentation (Dice 0.652, $p = 0.15$) but fail on diagnosis (Acc 0.479, $p < 0.001$), while overly large fields 3×3 & 7×7 (K3_7) are unstable (Dice 0.604, $p < 0.001$; Acc 0.497, $p < 0.001$) despite a marginally higher AUC (0.661, not significant). The proposed 3×3 & 5×5 (K3_5) setting provides the best balance, sustaining strong segmentation and high diagnostic robustness. *Bottleneck stabilization (ResMKGA):* Ablating SE (NoSE) reduces external segmentation (Dice 0.639 vs. 0.671, $p = 0.003$) with negligible AUC change (0.615 vs. 0.620), suggesting residual channel recalibration mainly stabilizes deep features to preserve boundary coherence under heavy shift.

4 Conclusion

In this study, we study cross-center robustness for multi-task thyroid ultrasound and show that domain shift amplifies negative transfer between geometry-driven segmentation and texture-driven TI-RADS cues. We find MedSAM to transfer well for segmentation, while ResNet34 is more reliable for malignancy and positioning under artifact-heavy shift. To mitigate this within a unified model, we introduce lightweight decoder adapters (MKGA/ResMKGA) that refine and gate multi-scale skip features to suppress artifact-driven activations. These modules strengthen out-of-domain segmentation and improve cross-center TI-RADS discrimination in the CNN setting, supporting targeted decoder refinement as a practical, parameter-efficient strategy for robust clinical deployment. While focused on thyroid ultrasound, vendor-specific artifact interference is systemic across medical imaging. We hypothesize MKGA generalizes to other artifact-heavy modalities, including echocardiography, breast, and musculoskeletal ultrasound, which future work will prospectively validate.

Acknowledgments. The authors have no competing interests.

Disclosure of Interests. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Horizons Grant DH-2025-00119.

References

1. Siegel, R.L., Kratzer, T.B., Giaquinto, A.N., Sung, H., Jemal, A.: Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians* **75**(1), 10–45 (2025). <https://doi.org/10.3322/caac.21871>
2. H. Huang, Y. Dong, X. Jia, J. Zhou, D. Ni, J. Cheng, and R. Huang. Personalized diagnostic tool for thyroid cancer classification using multi-view ultrasound. *arXiv preprint arXiv:2207.00496*, 2022.
3. Tessler, F.N., Middleton, W.D., Grant, E.G., et al.: Thyroid Imaging Reporting and Data System (TI-RADS): A User’s Guide *Radiology* **287**(1), 29–36 (2017). <https://doi.org/10.1148/radiol.2017171240>
4. Alyami, J., Almutairi, F.F., Aldoassary, S., Albeshry, A., Almontashri, A., Abounassif, M., Alamri, M.: Interobserver variability in ultrasound assessment of thyroid nodules *Medicine (Baltimore)* **101**(41), e31106 (2022). <https://doi.org/10.1097/MD.00000000000031106>
5. Ogut, E.: Artificial Intelligence in Clinical Medicine: Challenges Across Diagnostic Imaging, Clinical Decision Support, Surgery, Pathology, and Drug Discovery. *Clin. Pract.* **15**(9), 169 (2025). <https://doi.org/10.3390/clinpract15090169>
6. Kang, Q., Lao, Q., Li, Y., Jiang, Z., Qiu, Y., Zhang, S., Li, K.: Thyroid nodule segmentation and classification in ultrasound images through intra- and inter-task consistent learning. *Med. Image Anal.* **79**, 102443 (2022). <https://doi.org/10.1016/j.media.2022.102443>

7. Zheng, T., Qin, H., Cui, Y., Wang, R., Zhao, W., Zhang, S., Geng, S., Zhao, L.: Segmentation of thyroid glands and nodules in ultrasound images using the improved U-Net architecture. *BMC Med. Imaging* **23**(1), 56 (2023). <https://doi.org/10.1186/s12880-023-01011-8>
8. Lakkapragada, A., et al.: Mitigating negative transfer in multi-task learning. In: Proc. AAAI (2023) — discusses how standard multi-task optimization can suffer from negative transfer and various methods to mitigate it. [https://doi.org/10.48550/arXiv.2211.12999:contentReference\[oaicite:0\]index=0](https://doi.org/10.48550/arXiv.2211.12999:contentReference[oaicite:0]index=0)
9. Kirillov, A., et al.: Segment Anything (SAM): Foundation model for universal image segmentation. arXiv:2304.02643 (2023) — introduces SAM, a ViT-based segmentation model demonstrating strong global shape modeling but noted limitations on domain-specific tasks such as medical imaging. :contentReference[oaicite:1]index=1
10. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. *Nature Commun.* **15**(1) (2024). <https://doi.org/10.1038/s41467-024-44824-z>
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv:1512.03385 (2015). <https://arxiv.org/abs/1512.03385>
12. V. H. Duong, H. Vu, H. D. Phan, D. Q. Nguyen, D. H. Pham, Q. T. Le, B. S. Nguyen, T. D. Do, V. S. Dinh, T. C. Nguyen, H. H. Pham, and D. H. Ngo, “ThyroidXL: Advancing Thyroid Nodule Diagnosis with an Expert-Labeled, Pathology-Validated Dataset,” in *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, LNCS 15974, Springer Nature Switzerland, September 2025, pp. 616–626.
13. L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero, “An open access thyroid ultrasound image database,” in *Proc. SPIE 9287, 10th International Symposium on Medical Information Processing and Analysis*, 92870W, 28 January 2015. <https://doi.org/10.1117/12.2073532>
14. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” *arXiv preprint arXiv:2106.09685*, 2021. <https://arxiv.org/abs/2106.09685>
15. Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, “Characterizing and Avoiding Negative Transfer,” *arXiv preprint arXiv:1811.09751*, 2019.
16. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” *arXiv preprint arXiv:1606.00915*, 2017.
17. J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.
18. O. Abdelrazik, M. Elsayed, N. Wahab, N. Rajpoot, and A. Shephard, “A Deep Learning Framework for Thyroid Nodule Segmentation and Malignancy Classification from Ultrasound Images,” 2025, arXiv:2511.11937. <https://arxiv.org/abs/2511.11937>.
19. A. Kunapinun, D. Songsaeng, S. Buathong, M. N. Dailey, C. Keatmanee, and M. Ekpanyapong, “Explainable Automated TI-RADS Evaluation of Thyroid Nodules,” **Sensors**, vol. 23, no. 16, p. 7289, 2023. doi:10.3390/s23167289.
20. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: “Invariant Risk Minimization.” arXiv:1907.02893 (2020). <https://arxiv.org/abs/1907.02893>
21. Gulrajani, I., Lopez-Paz, D.: “In Search of Lost Domain Generalization,” arXiv:2007.01434 (2020). <https://arxiv.org/abs/2007.01434>

22. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: “Gradient Surgery for Multi-Task Learning, ” arXiv:2001.06782 (2020). <https://arxiv.org/abs/2001.06782>