
VIDEO-LEVEL THYROID ULTRASOUND DIAGNOSIS VIA YOLOv12m-SIMCLR FUSION AND DYNAMIC KEYFRAME SELECTION

Jingyu Fu, Yukai Huang

Southern University of Science and Technology
Shenzhen, China

fujy2023@mail.sustech.edu.cn, huangyukaihuangyukai@gmail.com

ABSTRACT

Thyroid ultrasound diagnosis is a dynamic process, yet static frame-level analysis often suffers from data redundancy and motion artifacts. We propose a video-level fusion framework that integrates YOLOv12m for multi-nodule detection with DeepSort for temporal tracking, ensuring spatio-temporal consistency across scanning sequences. To address severe class imbalance (malignant-to-benign ratio 4.65:1), we leverage SimCLR-based self-supervised learning to extract robust texture representations from unlabeled video data. A dynamic keyframe selection module is introduced to filter redundancy and aggregate diagnostically informative views for video-level malignancy risk stratification. Evaluated on an internal clinical dataset, our framework achieves a 90.6% malignant-case recall, significantly reducing missed diagnoses in screening scenarios. The system is deployed on a Flask-React platform to bridge the gap between algorithmic predictions and clinical decision-making, offering a scalable solution for real-time computer-aided thyroid diagnosis.

Keywords Thyroid Nodule Detection, YOLOv12m, SimCLR, Class Imbalance, Web Deployment, Dynamic Fusion Strategy

1 Introduction

Thyroid nodules represent one of the most common endocrine pathologies in clinical practice. Their reported prevalence varies significantly by diagnostic modality, ranging from 2–6% via palpation to 19–35% with ultrasound, and reaching as high as 65% in autopsy studies [1]. In recent years, the widespread adoption of high-resolution imaging technologies has led to a surge in the incidental detection of these nodules. While the vast majority are benign, approximately 10% harbor malignancy [2]. Consequently, the ability to accurately differentiate between benign and malignant lesions is critical for optimizing treatment strategies, avoiding unnecessary surgical interventions, and alleviating patient anxiety. Due to its cost-effectiveness, non-invasive nature, and real-time capabilities, ultrasound (US) has established itself as the cornerstone of thyroid nodule evaluation, serving as the primary tool for morphological assessment and risk stratification [1].

However, ultrasound diagnosis has intrinsic limitations from a signal-processing and diagnostic perspective. First, image interpretation is highly operator-dependent, leading to considerable inter-observer variability in TI-RADS grading, and the overall diagnostic accuracy heavily relies on the physician’s experience. Second, class imbalance can bias supervised models toward the majority class. Notably, the dataset used in this study is malignancy-enriched (*malignant* : *benign* $\approx$ 4.65 : 1), so the practical risk in our experiments is under-recognition of benign nodules rather than the more common scenario in population screening. Third, multi-nodule images are common in practice but many prior methods assume single-nodule inputs or require manual ROI selection, limiting applicability to real-world screening. Finally, manual annotation of US images is laborious and costly, further constraining the scale and diversity of curated datasets available for supervised training [3]. Standardized risk stratification systems (e.g., ACR TI-RADS, EU-TIRADS) partly mitigate variability by formalizing imaging criteria [4, 5], but they still rely on hand-crafted features and expert interpretation, and thus do not fully resolve challenges of generalization and scalability.

The emergence of artificial intelligence (AI) has greatly advanced computer-aided diagnosis (CAD) for thyroid ultrasound, providing objective and reproducible evaluations [6]. Studies demonstrate that CAD significantly improves diagnostic performance [7], with results comparable to those of experienced endocrinologists [2].

Various approaches have been investigated. Traditional methods rely on TI-RADS-like scoring systems [8] or hand-crafted features such as texture and morphology, combined with classifiers like support vector machines (SVMs) for benign-versus-malignant prediction [9]. However, such methods suffer from poor generalization due to reliance on manual feature engineering. Recently, deep learning methods have emerged as the dominant paradigm. CNN-based detectors such as Faster R-CNN and YOLO families have been employed for automated nodule detection, achieving high accuracy and real-time performance [10, 11, 12]. Lightweight models such as YOLO-LISA further reduce computational cost, making them more suitable for clinical deployment [12]. On the classification side, networks such as ResNet and DenseNet have achieved performance comparable to that of experienced radiologists [13].

Meanwhile, self-supervised and contrastive learning methods offer new opportunities for medical imaging tasks with limited annotated data. Frameworks such as SimCLR and Vision Transformer (ViT)-based contrastive approaches enable pretraining on unlabeled data, thereby enhancing downstream classification performance [14, 15]. However, research applying contrastive learning in thyroid ultrasound remains scarce, especially regarding the integration of multi-nodule detection with self-supervised classification.

In summary, although substantial progress has been achieved in thyroid nodule detection and classification, current research still faces several limitations. Most prior studies focus either on isolated nodule classification or on single-target detection, whereas real examinations often contain multiple nodules and substantial inter-frame redundancy.

Furthermore, many approaches rely heavily on fully supervised learning with limited annotated datasets, which restricts model generalization and robustness. In addition, the integration of multi-nodule detection with self-supervised or contrastive learning-based classification remains underexplored, leaving considerable potential for improving diagnostic accuracy and reliability in computer-aided thyroid ultrasound analysis.

Motivated by these insights, this study proposes a novel fusion framework combining YOLOv12m for multi-nodule detection with SimCLR-based self-supervised classification to enhance the diagnostic accuracy of thyroid nodules in ultrasound images. By integrating robust detection and classification capabilities, our approach addresses critical challenges such as class imbalance, subjective annotations, and the complexity of multi-nodule scenarios. The framework is further deployed on an interactive web platform, enabling real-time visualization and clinical integration, thus supporting radiologists in efficient and reliable decision-making.

The main contributions of this paper are as follows:

1. We develop a confidence-aware fusion strategy that combines detection outputs with self-supervised representations and evaluate it on a malignancy-enriched dataset (malignant:benign $\sim 4.64:1$). In internal testing, the method achieved a malignant-case recall of 0.906.
2. A multi-nodule detection strategy using YOLOv12m is proposed, effectively capturing complex nodule patterns in ultrasound images, enhancing robustness for early screening in clinical settings.
3. SimCLR’s self-supervised learning is employed to extract robust features from unlabeled data, mitigating the reliance on costly and subjective physician annotations, thus improving model generalization in imbalanced datasets.
4. A Flask+React-based web platform is introduced, enabling real-time visualization of detection boxes, classification probabilities, and tracking videos, with support for ONNX/PT model import, facilitating seamless clinical integration and interactive parameter adjustment.

The remainder of this paper is organized as follows: Section 2 reviews related works on thyroid nodule detection and classification methods. Section 3 details the proposed fusion framework, including YOLOv12m, SimCLR, and the fusion strategy. Section 4 describes the experimental setup and dataset. Section 5 presents the numerical results and performance analysis. Section 6 discusses the findings, limitations, and future directions. Section 7 summarizes the conclusions and contributions of this study.

2 Related works

Early advances in computer-aided diagnosis (CAD) for thyroid nodules have been primarily driven by deep learning. Research has evolved from CNN-based object detection frameworks to end-to-end classification networks and, more recently, to the adoption of self-supervised learning (SSL), significantly enhancing accuracy, generalizability, and clinical applicability in ultrasound-based diagnosis.

For detection, Wang et al. [12] first introduced YOLOv2 into thyroid ultrasound imaging, achieving real-time detection with high accuracy. Luong et al. [11] proposed a YOLOv8-Inception-v4 hybrid framework integrating detection, segmentation, and classification, reaching 97.37% segmentation accuracy and 98.76% classification accuracy on the TDID dataset. More recently, Wang et al. [10] developed YOLO-Thyroid, incorporating coordinate attention, SIOU regression loss, and class-weighted loss, significantly improving malignant nodule recall under imbalance. Lightweight variants such as YOLO-LISA further reduce computational cost while maintaining accuracy, enabling clinical deployment Wang et al. [12].

For classification, CNN-based approaches have achieved remarkable performance. Alghanimi et al. [16] designed a CNN-ResNet50 model that achieved 91.25% accuracy on 800 ultrasound images. Tang et al. [13] proposed a distant domain transfer learning framework (DHFF), which improved cross-hospital classification accuracy by 8%. Large-scale studies also confirmed CNN reliability: Koh et al. [17] evaluated 15,375 ultrasound images using CNN ensembles, reporting AUCs between 0.898 and 0.937, comparable to expert radiologists.

To mitigate limited annotations, transfer learning and self-supervised learning (SSL) have been explored. Xu et al. [2] combined InceptionV3-based transfer learning with machine learning classifiers, achieving an AUC of 0.783. Sun et al. [15] proposed TC-ViT, integrating Vision Transformer with contrastive learning, achieving 86.9% accuracy in distinguishing TI-RADS 3 from malignant nodules. Wang et al. [18] presented a multi-view contrastive SSL method that jointly improves classification and segmentation under limited labels. Recently, modality- and anatomy-aware contrastive pretraining has been reported to improve thyroid malignancy classification performance in preliminary studies [19].

Severe imbalance exists in thyroid ultrasound datasets, with benign nodules far outnumbering malignant ones. Wang et al. [10] incorporated class-weighted loss in YOLO-Thyroid, improving malignant recall. Li et al. [20] proposed CSASN with dynamic weighted loss and cascaded attention, improving minority subtypes such as FTC and MTC. Fang et al. [21] applied SMOTE oversampling, enhancing malignant nodule recall. Sharma et al. [22] used random oversampling to increase malignant sensitivity. Yang et al. [23] employed higher malignant loss weights to reduce misdiagnosis. The recently released TN5000 dataset further highlights real-world imbalance issues in thyroid imaging Zhang et al. [24].

3 Method

3.1 Overall Architecture

To address the clinical challenges in thyroid ultrasound video analysis—specifically, multi-nodule detection and tracking in dynamic sequences, benign-malignant classification under severe class imbalance and limited labeled data, and real-time diagnostic deployment amid probe motion and imaging artifacts—we propose a real-time integrated detection-classification framework that combines supervised object detection (YOLOv12m) with self-supervised representation learning (SimCLR).

This design is motivated by the fact that thyroid ultrasound diagnosis is inherently video-based: radiologists dynamically scan the thyroid gland and integrate information from multiple views, angles, and focal planes, rather than relying on a single static image. Accordingly, the framework is designed to bridge frame-level analysis and video-level decision support, without treating adjacent ultrasound frames as fully independent observations.

The framework consists of the following core components:

1. **Frame-level Detection with YOLOv12m:** Achieves multi-nodule localization and initial malignancy probability estimation through supervised training, supporting robust detection of multiple candidate nodules per frame even in low-contrast or multifocal scenarios.
2. **Frame-level Representation Learning with SimCLR:** Performs self-supervised pretraining on unlabeled data followed by fine-tuning on labeled regions, enabling robust feature extraction of global textures and subtle patterns to generate complementary frame-level malignancy probabilities.
3. **Video-level Fusion and Classification:** Integrates frame-level predictions via probability-level fusion, dynamic video segmentation, keyframe selection based on multi-dimensional features, and temporal aggregation (enhanced by DeepSort tracking), to produce accurate patient-level malignancy risk stratification.

As shown in Fig. 1, the input ultrasound video is decomposed into frame-level data and processed in parallel by YOLOv12m and SimCLR. The frame-level outputs are dynamically fused and aggregated into video-level malignancy predictions through segmentation and keyframe mechanisms. Ultimately, the results are visualized and presented

via a Flask-React web platform, supporting real-time nodule detection, tracking visualization, keyframe storage, and interactive parameter adjustment to meet clinical application demands.

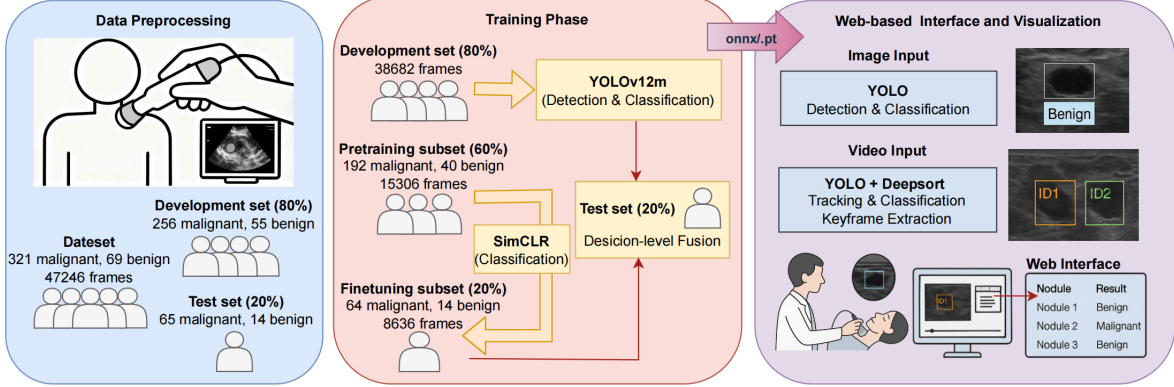


Figure 1: Overall framework of the proposed ultrasound thyroid nodule detection and classification system. The pipeline includes three main stages: (1) Data Preprocessing, where ultrasound videos are acquired and stratified into development (80%) and test (20%) sets by patient; (2) Training Phase, in which YOLOv12m performs supervised detection and classification on labeled data, and SimCLR executes self-supervised pretraining (60%) and finetuning (20%) before fusion testing; and (3) Web-based Interface and Visualization, enabling image and video inputs for real-time nodule detection, tracking (via DeepSORT), and result presentation on an interactive web platform.

3.2 Data Preprocessing

The ultrasound videos are first extracted into individual frames, resulting in a total of 47,246 images. Dataset splitting is performed at the patient level using stratified sampling to preserve the malignant-to-benign ratio (4.65:1), yielding a development set (80%, 311 patients) for model training and validation, and a held-out test set (20%, 79 patients) for final evaluation. No patient overlap exists between sets to prevent information leakage. During training, input ultrasound frames are resized to 640×640 pixels and augmented with Ultralytics-style transformations, including Mosaic augmentation, HSV jitter, random flipping, and scaling, to improve robustness to imaging variations.

3.3 Frame-level Nodule Detection

For frame-level nodule localization, we employ the YOLO [25, 26] (You Only Look Once) family of detectors, specifically YOLOv12m, as the core detection module. YOLO adopts a one-stage detection paradigm that simultaneously performs bounding-box regression, objectness estimation, and class prediction within a single feed-forward pass. This design allows efficient detection of multiple nodules in each ultrasound frame, supporting both frame-level and patient-level aggregation. YOLOv12m can also be combined with DeepSort for robust multi-nodule tracking, as detailed in Section 3.7.

As the latest state-of-the-art in the YOLO series, YOLOv12 has demonstrated superior performance across various medical detection tasks [?]. The medium-scale variant (YOLOv12m) strikes an optimal balance between accuracy and computational efficiency, making it well-suited for real-time clinical deployment on standard hardware.

Structurally, the YOLOv12m detection network consists of three major components:

- **Backbone:** A CSP-structured convolutional network with Residual Efficient Layer Aggregation Networks (R-ELAN) and Spatial Pyramid Pooling-Fast (SPPF) modules to efficiently expand the receptive field while maintaining stability.
- **Neck:** A combination of Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) for multi-scale feature fusion, critical for detecting nodules of varying sizes and contrasts.
- **Head:** Predicts bounding box coordinates, objectness confidence, and malignancy probability in a single forward pass.

Each input ultrasound frame is resized to 640×640 and fed into YOLOv12m to generate a set of predicted bounding boxes $\{b_i\}$, each associated with an objectness confidence c_i and malignancy probability p_i . The overall detection loss

integrates three components—localization, objectness, and classification—as:

$$\mathcal{L}_{det} = \lambda_{box}\mathcal{L}_{box} + \lambda_{obj}\mathcal{L}_{obj} + \lambda_{cls}\mathcal{L}_{cls}, \quad (1)$$

where $\mathcal{L}_{box}$ measures bounding-box alignment (e.g., CIOU), and $\mathcal{L}_{obj}$, $\mathcal{L}_{cls}$ are binary cross-entropy terms for objectness and malignancy classification, respectively.

Detected nodules from YOLOv12m are further tracked across frames using DeepSort to maintain temporal consistency, with full implementation details provided in Section 3.6.1. During inference, predictions below a confidence threshold $t_c = 0.25$ are filtered, and non-maximum suppression (NMS) with IoU threshold $t_{NMS} = 0.7$ eliminates redundant boxes.

3.4 Frame-level Self-Supervised Classification

To supplement YOLOv12m’s detection-based probabilities and reduce reliance on annotated data, we employ SimCLR [27, 28], a contrastive self-supervised learning framework designed to extract robust feature representations from unlabeled thyroid ultrasound frames. By leveraging large-scale unlabeled data, SimCLR enhances the model’s generalization ability and robustness to subjective annotation noise, which are particularly valuable in medical imaging where high-quality labels are scarce.

SimCLR extracts robust, high-level representations from unlabeled thyroid ultrasound frames using a ResNet-50 backbone encoder, a projection head, and the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss. Strong data augmentations—including random resized cropping, horizontal flipping, color jitter, grayscale conversion, and Gaussian blur—are applied to generate multiple views of each frame. The NT-Xent loss enforces intra-frame view similarity while maximizing inter-frame separation:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1, k \neq i}^{2N} \exp(\text{sim}(z_i, z_k)/\tau)}, \quad (2)$$

where $\text{sim}(z_i, z_j) = \frac{z_i \cdot z_j}{\|z_i\| \|z_j\|}$ denotes cosine similarity, τ is the temperature parameter (set to 0.5), and N is the batch size. This design exploits the abundance of unlabeled ultrasound data to learn invariant texture and structural patterns, enhancing generalization and resilience to annotation variability or imaging artifacts commonly encountered in clinical practice.

The training pipeline includes two stages: pre-training and fine-tuning. During pre-training, the ResNet-50 encoder and projection head are optimized on unlabeled ultrasound frames to learn invariant and discriminative representations. In the fine-tuning stage, the pretrained encoder is partially frozen, and a lightweight classification head is trained on labeled data to predict benign and malignant classes:

$$\mathcal{L}_{cls} = -[y \log(p) + (1 - y) \log(1 - p)], \quad (3)$$

where y is the ground truth label and p is the predicted malignancy probability.

By exploiting unlabeled data through contrastive pre-training, SimCLR effectively captures global semantic and texture-level distinctions in ultrasound frames, providing stable and generalizable feature representations for downstream classification.

SimCLR is used here primarily as a representation-learning module rather than as an independent clinical classifier; its role is to provide complementary texture-sensitive features for the downstream fusion stage.

3.5 Dynamic Fusion Mechanism

YOLOv12m excels at spatial localization but its classification head relies on limited labeled data, which can cause bias in multi-nodule or complex background scenarios. In contrast, SimCLR provides semantically rich and stable representations from unlabeled data, capturing global textural cues that are often overlooked by detection-based models. However, it lacks spatial awareness and may misinterpret contextual noise as pathological texture.

To harness their complementary strengths—YOLOv12m’s precise localization and SimCLR’s discriminative representation—we propose a dynamic fusion mechanism that adaptively integrates frame-level predictions. Rather than using simple averaging, we employ a two-stage decision rule consisting of (i) confidence-based model prioritization and (ii) weighted score fusion for uncertain cases. The priority rule layer determines the dominant model when confidence levels are extremely high or low, ensuring decisive outputs. The weighted fusion layer operates in uncertain cases, providing smooth transitions between models and enhancing robustness.

Let C_Y and C_S denote the confidence scores of YOLOv12m and SimCLR, respectively, and y_Y, y_S their predicted malignancy probabilities. The final fused prediction $\hat{y}$ is defined as:

$$\hat{y} = \begin{cases} y_S, & \text{if } C_S > \tau_S \text{ and } |C_S - C_Y| > \delta, \\ y_Y, & \text{if } C_Y < \tau_Y, \\ \alpha y_Y + (1 - \alpha) y_S, & \text{otherwise,} \end{cases} \quad (4)$$

where τ_S and τ_Y are confidence thresholds, δ controls divergence sensitivity, and α is an adaptive weighting coefficient.

The overall process can be summarized as:

```
if simclr_prob > threshold_malignant:
    return simclr_prob, "SimCLR_dominant"
elif yolo_prob < threshold_benign:
    return yolo_prob, "YOLO_benign"
elif abs(yolo_prob - simclr_prob) > threshold_divergence:
    return max_confidence(simclr_prob, yolo_prob)
else:
    return alpha * yolo_prob + (1 - alpha) * simclr_prob
```

This hierarchical design—prioritizing SimCLR for texture-driven malignancy, YOLOv12m for localization-driven benign exclusion, and weighted blending elsewhere—ensures adaptive dominance, mitigates individual model biases, and delivers stable frame-level outputs for downstream video aggregation.

3.6 Video-level Classification and Keyframe Selection

Although frame-level diagnosis is common in the literature, ultrasound examination is inherently a continuous scanning process, providing multiple complementary views of the same nodule. Aggregating cross-frame information yields several advantages: (i) suppressing single-frame noise such as speckle or shadow artifacts; (ii) stabilizing predictions under probe motion or transient defocusing; (iii) capturing appearance variations between adjacent views that single images cannot reflect. Instead of learning a fully temporal model, we adopt keyframe-based aggregation to reduce redundancy, preserve clinically informative views, and limit the annotation burden associated with video-level supervision. This approach focuses on clinically meaningful moments, reduces temporal redundancy, and better reflects real-world diagnostic practice, where only a limited number of representative frames are archived and reviewed.

Therefore, this work establishes a video-level classification framework integrating multi-target tracking, multi-dimensional feature extraction, dynamic segmentation, keyframe selection, and video-level fusion.

3.6.1 Multi-target Tracking via DeepSort

To achieve stable temporal association of nodules across continuous ultrasound frames, we introduce a multi-target tracking strategy based on DeepSort. In dynamic scanning, the same nodule may appear at slightly different positions or scales due to probe motion and patient respiration, which makes frame-wise detection unstable. Therefore, it is necessary to track the same nodule over time to ensure spatial-temporal consistency and reliable feature aggregation.

DeepSort is chosen because it combines motion modeling and appearance matching, enabling robust tracking even under temporary occlusions or blurred frames. It predicts object trajectories using a Kalman filter and matches detections through a combination of Mahalanobis distance (for position consistency) and cosine similarity (for appearance consistency). In our implementation, YOLOv12m provides frame-level detections (x, y, w, h, c_i) , while DeepSort assigns consistent tracking IDs by integrating motion prediction and feature embedding matching. This allows each nodule to maintain a unique identity throughout the video, ensuring that extracted temporal features originate from the same anatomical target rather than random detections. Such tracking-based consistency is crucial for subsequent feature extraction, keyframe selection, and video-level diagnosis.

3.6.2 Multi-dimensional Feature Extraction

To select diagnostically meaningful keyframes and support video-level fusion, we extract a comprehensive set of features describing both the visual and morphological characteristics of each tracked nodule. The goal is to identify frames that are clear, well-focused, and representative—those with strong contrast against the background, stable structure, and high detection confidence—while filtering out redundant or low-quality frames.

Five complementary categories of features are extracted from each tracked bounding box:

- Statistical: Sharpness, mean intensity, grayscale ratio, background similarity (SSIM), average detection confidence, and nodule count.
- Geometric: Area ratio, aspect ratio, circularity, and bounding box stability.
- Texture (GLCM-based): Contrast, dissimilarity, homogeneity, and energy, capturing microstructural texture patterns.
- Wavelet: LH, HL, HH subband energies from a Haar transform to describe multi-scale texture variations.
- Radiomics: A comprehensive set of descriptors, including first-order statistics, GLCM, GLRLM, GLSZM, GLDM, NGTDM, and 2D shape features, providing fine-grained intensity and structure representation.

Each frame yields a ~ 70 -dimensional feature vector, which is standardized and reduced to ~ 15 principal components via PCA to enhance computational efficiency and mitigate redundancy. Frames without valid detections are filled with zeros to preserve temporal continuity. These extracted features serve as the basis for dynamic video segmentation and keyframe selection, allowing the model to focus on clinically informative moments rather than treating all frames equally.

3.6.3 Dynamic Video Segmentation and Keyframe Selection

Although ultrasound video provides rich temporal information, raw sequences often contain extensive redundancy (e.g., near-static probe positions) and transient artifacts such as motion blur or acoustic shadows. Naively aggregating all frames not only increases computational cost but also weakens the diagnostic signal. Moreover, in real clinical workflows, radiologists typically review only a few representative, high-quality images rather than the entire sequence. To bridge this gap, we design a dynamic video segmentation and keyframe selection module that (i) filters redundant and noisy frames to improve temporal aggregation robustness, and (ii) extracts concise, clinically actionable frames to support report generation.

Dynamic Segmentation

Each tracked nodule sequence is automatically divided into visually consistent sub-segments. Segmentation is triggered when substantial changes in visual appearance or nodule configuration are detected, determined by the following complementary criteria:

1. Cosine similarity drop: For adjacent frames F_t and F_{t+1} , we compute the cosine similarity between their PCA-reduced feature vectors:

$$S_t = \frac{F_t \cdot F_{t+1}}{\|F_t\| \|F_{t+1}\|}. \quad (5)$$

A segmentation boundary is triggered when S_t falls below a dynamic threshold τ_d , indicating a significant visual change such as texture variation or illumination shift.

2. Tracking ID transition: When the tracking ID of a detected nodule changes, it implies the emergence or disappearance of a nodule in the field of view, prompting a new segment.

A minimum segment length of 10 frames prevents over-fragmentation due to transient artifacts.

Keyframe Selection

Within each segment, candidate frames are scored using a weighted combination of image clarity, nodule prominence, contrast, and feature stability:

$$\text{Score}_i = \sum_k w_k \cdot f_{i,k}, \quad (6)$$

where $f_{i,k}$ denotes normalized feature values and w_k the corresponding importance weights. The scoring emphasizes sharpness and area ratio to highlight clear, well-exposed nodules, while incorporating intensity and texture features (e.g., GLCM contrast, wavelet energy) to ensure diagnostic representativeness. Frames with the highest composite scores are selected as keyframes, balancing image quality and morphological informativeness. This mechanism not only filters out redundant or blurred frames but also aligns with the clinical decision-making process, where physicians rely on a handful of representative views to form diagnostic judgments.

The resulting keyframes serve as the foundation for subsequent video-level fusion, providing stable, interpretable, and computationally efficient inputs for malignancy assessment.

3.6.4 Video-level Fusion Classification

After keyframes are selected from each segment, they are aggregated to produce a robust, video-level malignancy prediction. This stage integrates complementary spatial-temporal cues from multiple representative frames while mitigating inconsistencies caused by noise, motion, or probe variations. Compared with single-frame inference, video-level classification better mirrors the clinical workflow, where radiologists synthesize multi-view evidence across the entire ultrasound scan.

The process consists of two steps: (i) fusing detection- and representation-based predictions at the keyframe level, and (ii) aggregating these fused results into a unified video-level malignancy score.

Keyframe-level fusion

For each selected keyframe, we obtain a detection-based probability p_Y from YOLOv12m and a representation-based probability p_S from SimCLR. The two sources are combined via the dynamic fusion mechanism described in Section 3.5 to yield a fused probability $\tilde{p}$ that balances detection confidence and learned representation similarity.

Video-level Aggregation

To derive the final video-level malignancy score per tracked nodule, we first remove outlier keyframes whose $\tilde{p}_i$ values deviate excessively (e.g., due to motion blur or false detection) using the Inter-Quartile Range (IQR) rule:

$$\text{trimmed} = \{\tilde{p}_i \mid Q_1 - 1.5 \cdot \text{IQR} \leq \tilde{p}_i \leq Q_3 + 1.5 \cdot \text{IQR}\}, \quad (7)$$

where Q_1, Q_3 are the 25th and 75th percentiles. If the trimmed set is empty, the median of the original values is used.

The video-level score is then computed as the mean of the trimmed probabilities:

$$P_{\text{video}} = \frac{1}{N'} \sum_{i=1}^{N'} \tilde{p}_i^{\text{trimmed}}, \quad (8)$$

where N' is the number of valid keyframes. This outlier-robust averaging effectively suppresses frame-level noise and yields a stable, scan-level decision without additional weighting parameters.

Finally, P_{video} is thresholded at 0.5 to produce the binary malignancy label per nodule. All selected keyframes are visualized on the clinical interface with bounding boxes, tracking IDs, and individual $\tilde{p}_i$ values, allowing radiologists to review representative multi-view evidence alongside the aggregated risk score.

By integrating detection, representation learning, tracking, and keyframe-based aggregation into a unified pipeline, the proposed framework bridges the gap between frame-level algorithmic predictions and video-level clinical decision-making, providing a practical and deployable solution for real-world thyroid ultrasound diagnosis.

3.7 Web-based Clinical Interface and Deployment

The entire framework is deployed on a **Flask-React** web platform integrating YOLOv12m, SimCLR, and the dynamic fusion module. The system supports real-time video upload, automatic keyframe extraction, and visualization of nodule bounding boxes and malignancy probabilities. Clinicians can interactively adjust thresholds, review keyframes, and download annotated outputs for medical documentation, bridging algorithmic research with real-world clinical workflows.

4 Experiments

4.1 Dataset overview

The dataset utilized in this research consists of thyroid ultrasound video sequences curated from institutional clinical databases. To ensure patient privacy and data security, all images were fully de-identified before processing. The cohort comprises 390 patients, including 321 malignant and 69 benign cases, representing a malignant-to-benign ratio of approximately 4.65:1. This distribution reflects the malignancy-enriched nature of the specific clinical subsets used for training robust diagnostic models.

A total of 47,246 frames were extracted from the video sequences (34,271 malignant and 7,249 benign). Each frame is accompanied by a standardized metadata file that includes a unique patient identifier, the ground-truth pathology label, and region-of-interest (ROI) coordinates (x, y, w, h) . The dataset structure is designed to support both frame-level

representation learning and temporal-based patient-level diagnosis. All data collection procedures were conducted in accordance with the Declaration of Helsinki. The study protocol was approved by the institutional review board, and the requirement for informed consent was waived due to the retrospective nature of the study.

To prevent information leakage, data splitting was performed at the patient level using stratified sampling to maintain the malignant-to-benign ratio (4.65:1). The dataset was divided as follows:

- **Test set (20%)**: 79 patients (65 malignant, 14 benign), 8,564 frames (7,041 malignant, 1,523 benign, ratio 4.64:1).
- **Development set (80%)**: 311 patients (256 malignant, 55 benign), 38,682 frames (27,230 malignant, 5,726 benign, ratio 4.76:1).

For SimCLR experiments, the development set was further split:

- **Pretraining subset (60%)**: 232 patients, 15,306 frames (192 malignant, 40 benign, ratio 4.80:1).
- **Finetuning subset (20%)**: 78 patients, 8,636 frames (64 malignant, 14 benign, ratio 4.57:1), with five-fold cross-validation (stratified by patient ID and class). Because benign cases were the minority class in this internal cohort, we augmented benign samples using horizontal flipping and random scaling (oversampling factor = 2) to partially mitigate class imbalance.
- **Test subset (20%)**: 79 patients, 8,568 frames (65 malignant, 14 benign, ratio 4.64:1, aligned with the test set).

No patient overlap exists across pretraining, finetuning, and test subsets, with consistent ratios (4.57-4.80, < 10%variation), ensuring reliable evaluation.

4.2 Data Augmentation

To mitigate the class imbalance (malignant:benign $\sim 4.65:1$) and improve robustness under diverse ultrasound imaging conditions, we adopted a unified yet task-specific augmentation pipeline across all models. The design objective was twofold: (i) to increase benign sample diversity and alleviate bias toward the dominant malignant class, and (ii) to enhance invariance to probe motion, scale, and speckle noise commonly observed in real thyroid scans.

All input frames were resized and normalized:

$$I_{\text{norm}}(i, j) = \frac{I(i, j) - \min(I)}{\max(I) - \min(I)}. \quad (9)$$

Spatial augmentations. Random horizontal flipping, rotation, and scaling simulated probe movement and scanning angle variations, while Mosaic augmentation (YOLOv12m) and RandomResizedCrop (SimCLR) diversified spatial context to improve small-nodule detection. These transformations mimic clinical scanning variability, encouraging models to focus on nodule appearance rather than positional bias.

Photometric augmentations. HSV jittering, brightness/contrast perturbation, and Gaussian blur simulated device- and operator-dependent variations in gain and focus. For SimCLR, additional grayscale conversion and blur were used during pretraining to strengthen contrastive feature learning under limited labeling.

Sampling strategy. To alleviate the 4.65:1 malignant-to-benign imbalance, benign samples were oversampled (factor = 2) using the above augmentations, while malignant cases were kept in their natural proportion to prevent overfitting to synthetic patterns. This balanced representation improved benign recall and reduced diagnostic bias toward malignancy.

Overall, this augmentation design produced more stable, generalizable feature representations across patients, devices, and scanning conditions.

4.3 Implementation Details

All experiments were implemented in PyTorch 2.4.1 with CUDA 12.4 support, using Python 3.8.20 on an Ubuntu 22.04 system equipped with an NVIDIA RTX 4090 (24GB) GPU, 16 vCPU Intel(R) Xeon(R) Gold 6430. Data splitting was performed at the patient level using stratified sampling to maintain the malignant-to-benign ratio (4.65:1), ensuring no information leakage.

All random seeds were fixed for dataset shuffling and weight initialization. The same data splits and augmentation pipeline were consistently applied across models to ensure fair comparison and reproducibility. Training configurations were optimized per task, with consistent hardware and optimization strategies for fair comparison.

4.4 Evaluation Metrics

To comprehensively evaluate model performance on imbalanced thyroid ultrasound data (malignant:benign 4.65:1), we designed metrics for classification and detection tasks. Classification metrics, derived from the confusion matrix (true positives TP, false positives FP, true negatives TN, false negatives FN), include:

1. **Precision:** Measures the proportion of predicted malignant samples that are truly malignant, critical for reducing false positives.
2. **Recall (Sensitivity):** Reflects the detection rate of true malignant samples, prioritizing the minimization of missed diagnoses.
3. **Accuracy:** Measures overall prediction correctness, though less reliable for imbalanced data.
4. **F1-score:** Balances precision and recall, making it suitable for evaluating performance on imbalanced datasets.
5. **AUC:** Represents the area under the receiver operating characteristic (ROC) curve, indicating the model’s overall discriminative ability across different thresholds.

Per-class metrics are reported for both malignant and benign nodules so that performance on the minority benign class is not obscured by overall accuracy.

For detection tasks, we use mean average precision (mAP) based on intersection over union (IoU). Average precision (AP) is calculated from the Precision-Recall curve, and mAP is the mean AP across classes (benign and malignant). We report mAP@0.5 (IoU threshold 0.5) and mAP@0.5:0.95 (IoU range from 0.5 to 0.95 in steps of 0.05), with a focus on per-class precision and recall, particularly malignant recall, to minimize missed diagnoses.

Additionally, inference latency (ms per frame) is measured to assess the system’s real-time feasibility for clinical applications.

5 Results

To comprehensively evaluate the proposed framework, we conducted a multi-stage validation process. We first optimized the detection backbone and data strategy, then assessed the effectiveness of the fusion classification mechanism under severe class imbalance, and finally validated the system’s diagnostic reliability at the patient level.

5.1 Establishing the Optimal Detection Baseline

To construct a robust foundation for the proposed fusion framework, we first evaluated the performance of YOLOv12m against its predecessors (YOLOv8n, YOLOv11m) and investigated the impact of nodule composition strategies. As detailed in Table 1, YOLOv12m achieved competitive detection performance, with the best balance between recall and overall detection accuracy among the evaluated detector variants. The metric labels in Table 1 should be rechecked for consistency before publication. While YOLOv8n exhibited slightly higher precision (0.758), YOLOv12m provided a better balance between precision and recall (0.760 and 0.720, respectively), which is critical for medical object detection where missed detections are unacceptable.

Table 1: Performance Comparison of YOLO Variants on Thyroid Ultrasound Nodule Detection

Model	Precision	Recall	mAP50	mAP50:95	AP50-95	Inference Time (ms)
YOLOv8n	0.758	0.678	0.742	0.485	0.716	0.3
YOLOv11m	0.775	0.675	0.692	0.453	0.721	3.4
YOLOv12m	0.760	0.720	0.731	0.484	0.739	4.4

To further assess model robustness under realistic clinical conditions, we compared a *Filtered Single-Nodule* strategy—restricting both training and validation to images containing only one nodule—with an *Unrestricted Multi-Nodule* strategy that allows multiple nodules per image. As detailed in Table 2, despite the Single-Nodule dataset appearing "cleaner" with 6,491 instances, the Multi-Nodule strategy (9,199 instances) yielded a decisive performance leap. Specifically, the overall recall increased from 0.628 to 0.720 (+14.6%), and the recall for benign nodules saw a remarkable surge of 44.4% (from 0.435 to 0.628).

Table 2: Multi-Nodule vs. Single-Nodule Performance Comparison

Metric	Single	Multi	Improvement
Precision	0.699	0.760	+8.7%
Recall	0.628	0.720	+14.6%
mAP50	0.662	0.731	+10.4%
mAP50:95	0.438	0.484	+10.5%

Note: Single-nodule dataset: 6,492 images, 6,491 instances; Multi-nodule dataset: 8,564 images, 9,199 instances.

These results highlight a unique advantage of the YOLO architecture in thyroid ultrasound analysis: by training on images where multiple nodules co-exist, the model learns to better distinguish suspicious lesions from background artifacts and adjacent anatomical structures. Although the unrestricted setting introduces more complex and heterogeneous training samples, it more accurately reflects real-world ultrasound scenarios and better leverages YOLO’s inherent multi-object detection capability. Therefore, YOLOv12m trained without single-nodule constraints was selected as the baseline detector for subsequent experiments.

5.2 Limitations of Conventional Supervised Classification

To isolate the intrinsic difficulty of benign–malignant discrimination from detection-related factors such as localization errors or small lesion size, we evaluated several conventional supervised CNN classifiers—including ResNet50, EfficientNet-B0, and DenseNet—on a cropped ROI-level nodule dataset. In this setting, each input corresponds to a manually annotated nodule region, and the task is reduced to pure image-level classification.

As shown in Table 3, although certain models achieved relatively high overall accuracy—for example, EfficientNet-B0 with a 2:1 weighted loss reached an accuracy of 0.86—most supervised classifiers exhibited pronounced performance imbalance between benign and malignant classes. Specifically, the unweighted ResNet50 achieved high malignant precision (0.90) but suffered from a low benign recall of 0.57, indicating a strong bias toward predicting malignancy. Introducing aggressive class weighting partially alleviated this issue but resulted in unstable trade-offs; for instance, ResNet50 with a 3:1 weighting improved benign recall to 0.59 while reducing malignant recall to 0.82.

Table 3: Classification Performance Comparison

Model	Overall Accuracy
ResNet50 (Unweighted)	0.78
ResNet50 (Weighted 2:1)	0.79
ResNet50 (Weighted 3:1)	0.78
ResNet50 (Focal Loss)	0.69
EfficientNet-B0 (Weighted 1.5:1)	0.85
EfficientNet-B0 (Weighted 2:1)	0.86
DenseNet (Weighted 1.5:1)	0.79

These results suggest that even under a simplified ROI-level setting, conventional supervised CNNs struggle to maintain balanced sensitivity across classes on severely imbalanced data. More importantly, we observed that ROI-level classification remained inferior to detection-based predictions produced by YOLOv12m operating on full ultrasound images. This disparity indicates that isolating nodules from their surrounding thyroid tissue may discard diagnostically relevant contextual cues, such as relative echogenicity and spatial consistency.

Motivated by these limitations, we incorporate a SimCLR-based self-supervised module to learn complementary feature representations from large-scale unlabeled data. By alleviating label dependency and enriching the representation space beyond localized ROI appearance, the self-supervised features provide a more robust foundation for downstream fusion-based diagnosis.

5.3 Efficacy of the Fusion Framework

Table 4 presents the core results of this study, comparing the proposed Fusion Model against standalone YOLOv12m and SimCLR. The fusion strategy was designed to prioritize malignant sensitivity (recall) to minimize missed diagnoses in clinical screening. The Fusion Model achieved an outstanding Malignant Recall of 90.6%, representing a substantial improvement over standalone SimCLR (76.4%) and YOLOv12m (83.5%). By dynamically integrating SimCLR’s

texture-sensitive probabilities with YOLO’s objectness scores, the system successfully recovered malignant cases that were missed by the object detector alone. This demonstrates that combining global detection with self-supervised texture representation can enhance sensitivity while maintaining high overall accuracy.

Table 4: Performance Comparison of YOLOv12m, SimCLR, and Fusion Model

Model	Accuracy	Precision	Recall	F1-Score
YOLOv12m	0.802	0.917	0.835	0.874
SimCLR	0.674	0.826	0.764	0.794
Fusion	0.823	0.882	0.906	0.894

5.4 Patient-Level Validation and Keyframe Selection

While frame-level metrics provide granular insight into model stability, clinical diagnosis is inherently a patient-centered process. Radiologists do not diagnose isolated frames; they synthesize information from a continuous video stream to form a holistic judgment. To validate the system’s translational potential, we aggregated predictions at the patient level, moving beyond simple image classification to video-based diagnostic triage. Table 5 (All Frames) and Table 6 (Keyframes Only) compare the diagnostic performance using two aggregation strategies. The results underscore the critical importance of the Dynamic Keyframe Selection module.

Table 5: Patient-level performance using **all frames** after IQR trimming.

Model	Accuracy	Precision	Recall	F1-Score
YOLOv12m	0.8101	0.9032	0.8615	0.8819
SimCLR	0.6582	0.8167	0.7538	0.7840
Fusion	0.8481	0.8732	0.9538	0.9118

Table 6: Patient-level performance using **only keyframes** after IQR trimming .

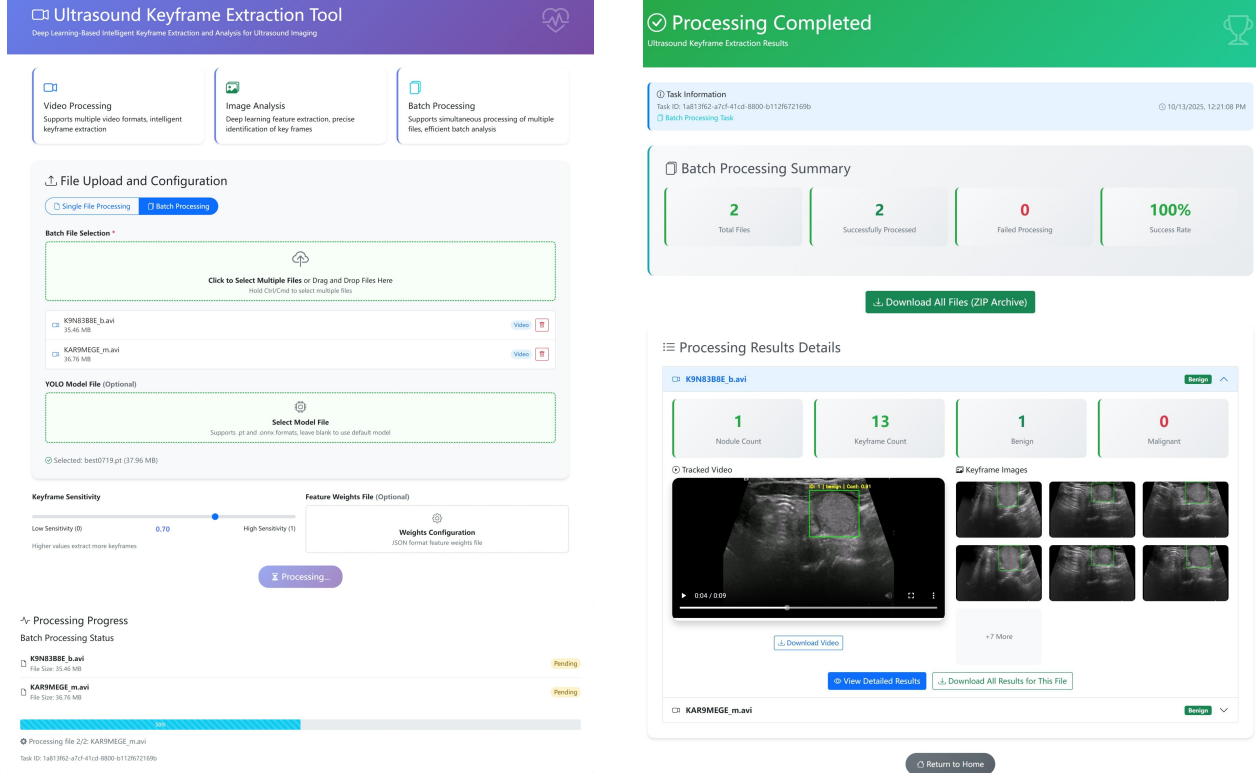
Model	Accuracy	Precision	Recall	F1-Score
YOLOv12m	0.8718	0.8857	0.9688	0.9254
SimCLR	0.6410	0.8103	0.7344	0.7705
Fusion	0.8846	0.8767	1.0000	0.9343

When aggregating predictions across "All Frames" (Table 5), the Fusion Model achieves the highest recall (0.9538) and F1-score (0.9118), indicating strong sensitivity to malignant cases while maintaining competitive accuracy (0.8481) and precision (0.8732). Compared to standalone YOLOv12m and SimCLR, the fusion strategy effectively combines global detection and texture-aware features to improve detection robustness across the full video sequence.

6 Discussion

This study presents a unified fusion framework that integrates YOLOv12m-based object detection with SimCLR-based self-supervised representation learning for thyroid ultrasound nodule analysis under severe class imbalance. Rather than pursuing increasingly complex network architectures, the proposed approach focuses on combining complementary sources of diagnostic information: YOLOv12m emphasizes spatial localization and structural cues within full ultrasound frames, while SimCLR captures global texture and appearance patterns learned from large-scale unlabeled data. By leveraging these complementary strengths, the framework aims to enhance diagnostic robustness under realistic clinical constraints, including limited annotations, heterogeneous image quality, and real-time deployment requirements.

A key finding of this work is that strong detection performance alone does not necessarily translate into reliable diagnostic classification. Although YOLOv12m demonstrates excellent localization accuracy and real-time efficiency, its malignancy predictions are primarily driven by objectness confidence and bounding-box features, which are not explicitly optimized for benign–malignant discrimination. In contrast, SimCLR learns texture- and appearance-sensitive representations that are more invariant to annotation noise and class imbalance, but lacks spatial awareness and temporal stability when applied independently. The proposed fusion framework effectively exploits the complementary failure modes of these two models: YOLOv12m provides reliable localization and anatomical context, while SimCLR



(a) Main UI dashboard showcasing interactive features for thyroid ultrasound analysis.

(b) Batch processing results displaying detection boxes and malignancy probabilities.

Figure 2: Visual outputs of the web interface: (a) Main UI dashboard for system interaction and (b) batch processing results validating model performance.

contributes discriminative semantic cues beyond geometric information. This complementary integration explains the consistent improvements in recall and F1-score observed in the fusion model compared with either component alone.

Another important observation is that isolating nodules into cropped regions of interest does not necessarily simplify the diagnostic task. ROI-level classification removes contextual information such as relative echogenicity, background tissue appearance, and spatial consistency within the thyroid gland, all of which are routinely considered by radiologists during interpretation. Detection-based inference on full ultrasound frames preserves this global context, contributing to more stable predictions. By further integrating representation-level information learned from unlabeled data, the fusion strategy mitigates dataset bias and improves robustness without sacrificing spatial interpretability.

At the patient level, ultrasound data are inherently acquired as continuous video sequences rather than isolated frames. However, such videos often contain substantial redundancy, near-duplicate views, and transient low-quality frames caused by probe motion, defocusing, or acoustic artifacts. Naively aggregating predictions from all frames may dilute diagnostically relevant information and introduce unnecessary noise into patient-level decisions.

To address this challenge, the proposed framework introduces a dynamic keyframe selection strategy that identifies a compact set of clear, representative, and diagnostically informative frames from each ultrasound video. By filtering out redundant and low-quality frames, keyframe selection serves as an effective form of information compression, preserving clinically meaningful content while improving the stability of video-level aggregation. The observed improvements in patient-level performance demonstrate that keyframe-based aggregation is not merely an engineering optimization, but a clinically meaningful design choice.

This strategy closely aligns with real-world clinical workflows, where radiologists typically rely on a limited number of representative images rather than reviewing every frame of a scan. By basing patient-level malignancy assessment on selected keyframes instead of all frames, the system not only enhances robustness and consistency, but also produces concise visual evidence that can be directly reviewed and interpreted by clinicians. Importantly, these selected keyframes and corresponding predictions are integrated into a web-based graphical interface, enabling real-time visualization,

interactive review, and result archiving. This design bridges algorithmic inference with practical clinical usage, facilitating both diagnostic decision-making and real-world deployment.

Despite these promising results, several limitations should be acknowledged. First, the dataset was collected from a single institution, and multi-center validation is necessary to assess generalizability across different scanners, operators, and patient populations. Second, the current fusion strategy is rule-based rather than end-to-end optimized, leaving room for future work on learnable fusion mechanisms that may further improve adaptability. Finally, this study focuses exclusively on B-mode ultrasound; incorporating additional modalities such as elastography, Doppler information, or clinical metadata may further enhance diagnostic reliability.

Overall, this work demonstrates that combining detection-based localization with self-supervised representation learning provides a practical, interpretable, and clinically aligned solution for thyroid ultrasound analysis under real-world constraints. By emphasizing robustness, patient-level reasoning, and deployability through a web-based interface, the proposed framework offers a solid foundation for future extensions toward comprehensive computer-aided diagnosis systems.

7 Conclusion

This study presents a novel fusion framework that integrates YOLOv12m’s object detection capabilities with SimCLR’s self-supervised feature learning for thyroid ultrasound nodule detection and classification. The framework demonstrates strong performance on a highly imbalanced clinical dataset.

The workflow ensures patient-level data independence through rigorous dataset partitioning, enables efficient multi-nodule detection via YOLOv12m, and extracts robust texture and structural features from unlabeled data using SimCLR, reducing reliance on subjective physician annotations. A confidence-aware fusion strategy combines priority rules with weighted averaging to integrate detection and classification outputs, yielding high recall and stable diagnostic predictions.

The model is deployed on an interactive Flask+React web platform, providing real-time visualization of detection boxes, malignancy probabilities, and keyframes, which enhances radiologists’ decision-making efficiency in clinical screening.

Clinically, this framework minimizes the risk of missed malignant nodules, stabilizes predictions under probe motion and imaging artifacts, and aligns with real-world workflows by focusing on representative keyframes rather than all frames. SimCLR’s self-supervised approach also addresses the bottleneck of scarce annotated data, offering a scalable solution for medical imaging tasks.

Future work will explore optimizing the fusion mechanism, integrating multimodal imaging data, and extending the approach to other ultrasound applications such as breast and liver examinations. These developments are expected to further enhance diagnostic accuracy, clinical applicability, and potential for translational deployment. Overall, the proposed framework provides an efficient, interpretable, and clinically valuable tool for thyroid ultrasound diagnosis with significant potential for widespread adoption and intellectual property protection.

References

- [1] Diana S Dean and Hossein Gharib. Epidemiology of thyroid nodules. *Best practice & research Clinical endocrinology & metabolism*, 22(6):901–911, 2008.
- [2] Yan Xu, Mingmin Xu, Zhe Geng, Jie Liu, and Bin Meng. Thyroid nodule classification in ultrasound imaging using deep transfer learning. *BMC Cancer*, 25(1):544, March 2025.
- [3] Xiangdong Guo, Haifeng Zhao, and Zhenyu Tang. An Improved Deep Learning Approach for Thyroid Nodule Diagnosis. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 296–299, April 2020. ISSN: 1945-8452.
- [4] Ge-Ge Wu, Wen-Zhi Lv, Rui Yin, Jian-Wei Xu, Yu-Jing Yan, Rui-Xue Chen, Jia-Yu Wang, Bo Zhang, Xin-Wu Cui, and Christoph F Dietrich. Deep learning based on acr ti-rads can improve the differential diagnosis of thyroid nodules. *Frontiers in Oncology*, 11:575166, 2021.
- [5] Franklin N Tessler, William D Middleton, Edward G Grant, Jenny K Hoang, Lincoln L Berland, Sharlene A Teefey, John J Cronan, Michael D Beland, Terry S Desser, Mary C Frates, et al. Acr thyroid imaging, reporting and data system (ti-rads): white paper of the acr ti-rads committee. *Journal of the American college of radiology*, 14(5):587–595, 2017.

- [6] Xiaowen Liang, Jinsui Yu, Jianyi Liao, and Zhiyi Chen. Convolutional neural network for breast and thyroid nodules diagnosis in ultrasound imaging. *BioMed Research International*, 2020(1):1763803, 2020.
- [7] Ming-Hsun Wu, Kuen-Yuan Chen, Shyang-Rong Shih, Ming-Chih Ho, Hao-Chih Tai, King-Jen Chang, Argon Chen, and Chiung-Nien Chen. Multi-reader multi-case study for performance evaluation of high-risk thyroid ultrasound with computer-aided detection. *Cancers*, 12(2):373, 2020.
- [8] Bryan R Haugen, Erik K Alexander, Keith C Bible, Gerard M Doherty, Susan J Mandel, Yuri E Nikiforov, Furio Pacini, Gregory W Randolph, Anna M Sawka, Martin Schlumberger, et al. 2015 american thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the american thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid*, 26(1):1–133, 2016.
- [9] Yu-Len Huang and Dar-Ren Chen. Support vector machines in sonography: application to decision making in the diagnosis of breast cancer. *Clinical imaging*, 29(3):179–184, 2005.
- [10] Shidan Wang, Zi-An Zhao, Yuze Chen, Ye-Jiao Mao, and James Chung-Wai Cheung. Enhancing Thyroid Nodule Detection in Ultrasound Images: A Novel YOLOv8 Architecture with a C2fA Module and Optimized Loss Functions. *Technologies*, 13(1):28, January 2025. Publisher: Multidisciplinary Digital Publishing Institute.
- [11] Duong Trong Luong, Do Duy Thai, Phan Minh Nghia, Nguyen Thong Nhat, Dang Le Dat, and Do Thanh Tuan. Anomaly Segmentation and Classification in Thyroid Ultrasound Images Using YOLOv8-GSO-Inception-V4. In *2025 10th International Conference on Applying New Technology in Green Buildings (ATiGB)*, pages 701–706, July 2025.
- [12] Lei Wang, Shujian Yang, Shan Yang, Cheng Zhao, Guangye Tian, Yuxiu Gao, Yongjian Chen, and Yun Lu. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the yolov2 neural network. *World journal of surgical oncology*, 17(1):12, 2019.
- [13] Fenghe Tang, Jianrui Ding, Lingtao Wang, and Chunping Ning. A novel distant domain transfer learning framework for thyroid image classification. *Neural Processing Letters*, 55(3):2175–2191, 2023.
- [14] Jihong Ouyang, Dong Mao, Zeqi Guo, Siguang Liu, Dong Xu, and Wenting Wang. Contrastive self-supervised learning for diabetic retinopathy early detection. *Medical & Biological Engineering & Computing*, 61(9):2441–2452, 2023.
- [15] Jiawei Sun, Bobo Wu, Tong Zhao, Liugang Gao, Kai Xie, Tao Lin, Jianfeng Sui, Xiaoqin Li, Xiaojin Wu, and Xinye Ni. Classification for thyroid nodule using ViT with contrastive learning in ultrasound images. *Computers in Biology and Medicine*, 152:106444, January 2023.
- [16] Ghufuran B. Alghanimi, Hadeel K. Aljobouri, and Khaleel Akeash Al-shimmari. CNN and ResNet50 Model Design for Improved Ultrasound Thyroid Nodules Detection. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETIS)*, pages 1000–1004, January 2024.
- [17] Jieun Koh, Eunjung Lee, Kyunghwa Han, Eun-Kyung Kim, Eun Ju Son, Yu-Mee Sohn, Mirinae Seo, Mi-ri Kwon, Jung Hyun Yoon, Jin Hwa Lee, Young Mi Park, Sungwon Kim, Jung Hee Shin, and Jin Young Kwak. Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. *Scientific Reports*, 10(1):15245, September 2020. Publisher: Nature Publishing Group.
- [18] Jian Wang, Xin Yang, Xiaohong Jia, Wufeng Xue, Rusi Chen, Yanlin Chen, Xiliang Zhu, Lian Liu, Yan Cao, Jianqiao Zhou, et al. Thyroid ultrasound diagnosis improvement via multi-view self-supervised learning and two-stage pre-training. *Computers in Biology and Medicine*, 171:108087, 2024.
- [19] Eich Takaya and Ryusei Inamori. ModAn-MulSupCon: Modality-and Anatomy-Aware Multi-Label Supervised Contrastive Pretraining for Medical Imaging, August 2025. arXiv:2508.18613 [eess].
- [20] Peiqi Li, Yincheng Gao, Renxing Li, Haojie Yang, Yunyun Liu, Boji Liu, Jiahui Ni, Ying Zhang, Yulu Wu, Xiaowei Fang, Lehang Guo, Liping Sun, and Jiangang Chen. Diagnosis for Less-Prevalent Thyroid Carcinoma Subtype Using a Dual-Branch Attention Deep Network with Ultrasound Images, September 2025. arXiv:2505.02211 [eess].
- [21] Mingshuang Fang, Qingfeng Ma, and Binxiong Xu. Automated multi-class classification of thyroid nodules in ultrasound imaging using transformer-based segmentation and hybrid feature learning. *Journal of Radiation Research and Applied Sciences*, 18(4):101782, 2025.
- [22] Rohit Sharma, Gautam Kumar Mahanti, Ganapati Panda, and Abhishek Singh. Thyroid nodules classification using weighted average ensemble and d-critic based topos methods for ultrasound images. *Current Medical Imaging*, 20(1):E050423215446, 2024.

- [23] Changan Yang, Muhammad Awais Ashraf, Mudassar Riaz, Pascal Umwanzavugaye, Kavimbi Chipusu, Hongyuan Huang, and Yueqin Xu. Improving diagnostic precision in thyroid nodule segmentation from ultrasound images with a self-attention mechanism-based Swin U-Net model. *Frontiers in Oncology*, 15:1456563, February 2025.
- [24] Huan Zhang, Qianglin Liu, Xiaolin Han, Lijuan Niu, and Weidong Sun. TN5000: An Ultrasound Image Dataset for Thyroid Nodule Detection and Classification. *Scientific Data*, 12(1):1437, August 2025. Publisher: Nature Publishing Group.
- [25] Momina Liaqat Ali and Zhou Zhang. The yolo framework: A comprehensive review of evolution, applications, and benchmarks in object detection. *Computers*, 13(12), 2024.
- [26] Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716, 2023.
- [27] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML’20. JMLR.org, 2020.
- [28] Tobias Uelwer, Jan Robine, Stefan Sylvius Wagner, Marc Höftmann, Eric Upschulte, Sebastian Konietzny, Maike Behrendt, and Stefan Harmeling. A survey on self-supervised methods for visual representation learning. *Mach. Learn.*, 114(4), March 2025.