Window-Aware Training for Data-Efficient 3D CT Foundation Model Adaptation

Ju-Yun Cheng[1][0009-0006-6819-3058], Ying-Jia Lin[1,2][0000-0003-4347-0232], and Chi-Tung Cheng[3][0000-0002-2697-4642]

[1] Department of Artificial Intelligence, Chang Gung University, Taoyuan, Taiwan
[2] Artificial Intelligence Research Center, Chang Gung University, Taoyuan, Taiwan
[3] Department of Trauma and Emergency Surgery, Chang Gung Memorial Hospital, Taoyuan, Taiwan
yjlin@cgu.edu.tw

Abstract. Computed tomography (CT) foundation models aim to learn transferable three-dimensional representations across anatomical regions and disease categories, yet large-scale CT pretraining remains costly and data-intensive. This paper presents our solution for the Task 1 Coreset Track of the CVPR 2026 Foundation Models for General CT Image Diagnosis challenge, where models are trained from scratch using only 10% of the pretraining dataset and evaluated by linear probing on 17 binary classification tasks. We propose a data-efficient self-supervised framework that combines anatomy-aware balanced coreset selection, window-aware Hounsfield Units augmentation, and multi-scale global-local contrastive learning. The coreset strategy groups CT volumes by anatomical region and dataset source to preserve diversity under the limited sampling budget. Randomly perturbed clinical CT window settings are incorporated during pretraining to capture contrast variations across tissue types. A 3D DenseNet121 encoder is trained with global-global and global-local contrastive objectives, and multi-level features are aggregated for downstream linear probing. On the validation set, the framework obtains an overall balanced accuracy of 0.706 and an AUROC of 0.716 across the 17 validation targets. These results surpass the CT-NEXUS and VoCo baselines and are comparable to 3DINO, despite using a compact backbone and fully from-scratch training.

Keywords: CT foundation model · Linear probing · Window-aware augmentation

1 Introduction

Computed Tomography (CT) foundation models [1,6,18] learn transferable three-dimensional (3D) representations from large-scale unlabeled CT volumes, providing a general feature backbone for downstream tasks across anatomical regions, disease categories, and different data sources. Compared to natural images, CT images contain more clinical information. Their voxel intensity is expressed in

Hounsfield Units (HU) [10], and different tissue structures are rendered using different window settings. For example, the contrast requirements for the lungs, soft tissues, and bones differ. However, existing self-supervised learning (SSL) techniques mostly employ data augmentation and representation learning strategies developed for natural images, which may limit the model's ability to capture the domain-specific characteristics of CT images. Building CT foundation models is challenging not only because large CT datasets are expensive to curate, but also because their composition is often uneven. When the training data is dominated by certain anatomical regions or data sources, the learned representation may also become biased toward these distributions. As a result, the model may generalize less effectively to downstream tasks involving underrepresented domains. This issue is especially relevant under a limited pre-training budget, where the selected subset largely determines the anatomical and domain diversity exposed to the model during training.

Our work is developed for the Computer Vision and Pattern Recognition (CVPR) 2026 Workshop, Task 1 Coreset Track. Participants are required to train from scratch using only 10% of the entire pre-training dataset, without the use of any pre-trained weights, in order to complete this task. Models are evaluated through linear probing (LP), in which the linear classification head is trained while the backbone remains unaltered. Their representational capabilities are assessed through 17 binary classification tasks that pertain to the abdomen and chest. Under these constraints, the challenge is not merely how to train a self-supervised model but how to preserve sufficient anatomical and domain diversity with a very limited amount of data while enabling the model to learn CT representations that can support various downstream tasks.

Related Work

Vision-Only CT Foundation Models

Several vision-only approaches have advanced self-supervised representation learning for 3D CT. VoCo [18] formulates volume contrastive learning by exploiting the relatively stable spatial relationships among organs in 3D medical images. Through sub-volume context prediction, it encourages the model to learn semantic anatomical representations and achieves strong performance across downstream tasks. 3DINO [19] adapts the teacher-student self-distillation framework of DINOv2 [15] to 3D medical imaging, combining image-level and patch-level learning objectives. Pretrained on nearly 100,000 3D scans across modalities and anatomical regions, 3DINO shows strong transferability to out-of-distribution tasks and serves as a representative method for multi-organ 3D representation learning. Despite these advances, both methods rely on large-scale pretraining datasets. While 3DINO also reports results under a partial-label downstream setting, this experiment mainly studies the label efficiency of pretrained representations during downstream training. It leaves open a different question: how to construct an effective pretraining subset when only a limited amount of unlabelled CT data can be used.

Coreset Selection

Coreset selection focuses on constructing a compact yet informative subset from a large dataset. For SSL pretraining, Kim et al. [13] introduced an open-set SSL method that selects unlabeled samples based on their latent-space distance to the target dataset. This method enhances fine-grained representation learning in the absence of labels. In the supervised classification scenario of medical images, Hong et al. proposed EVA [9], which uses a dual-window approach to track dynamic changes in sample importance during training and outperforms baselines trained on the full dataset when using only 10% of the data. Although the studies mentioned above have all proposed coreset selection strategies and shown their effectiveness in improving downstream performance, none of these methods take into account the anatomical diversity specific to CT or the need for cross-domain balance: Kim et al. [13] relies on a pre-trained SSL embedding model to measure sample distances, while EVA requires a complete labeled training process to calculate error variance; neither approach complies with this challenge's constraints of "from-scratch" development and the prohibition of pre-trained SSL models.

Motivation

This work is motivated by two observations. First, existing SSL pretraining pipelines largely inherit augmentation strategies from natural image literature without accounting for the fact that CT intensities are expressed in HU values and that different window settings reveal different tissue contrasts [4]. Incorporating multiple clinical CT windows as augmentations during pretraining exposes the model to the same anatomical region under varying contrast conditions, encouraging it to learn richer, CT-specific representations. Second, under a 10% data budget, naive random sampling risks creating imbalances across anatomical regions or dataset sources, which can hurt generalization across the 17 evaluation tasks. A coreset strategy that explicitly preserves anatomical diversity and cross-domain balance is better suited to from-scratch training in this multi-task setting. We summarize our contributions as follows:

1. **Anatomy-Aware Balanced Coreset Selection:** We designed and compared two coreset sampling strategies: sampling according to the proportions of the original distribution (Version 1) and balanced sampling across anatomical regions and datasets (Version 2). Through ablation experiments, we validated the advantages of the balanced sampling strategy in preserving diversity.
2. **Window-Aware HU Augmentation:** We incorporated five clinical CT windows (soft tissue, lung, bone, abdomen, and wide window) into the self-supervised pretraining augmentation pipeline, exposing the model to cross-window tissue contrast variation during pretraining and strengthening representational transfer to downstream tasks.
3. **Multi-Scale Global-Local Contrastive Learning:** Using 3D DenseNet121 as the backbone, we combine global-global and global-local contrastive losses with multi-scale feature fusion (f1–f4 concatenation with adaptive pooling) as the feature representation for LP, capturing both global anatomical structure and local fine-grained detail.

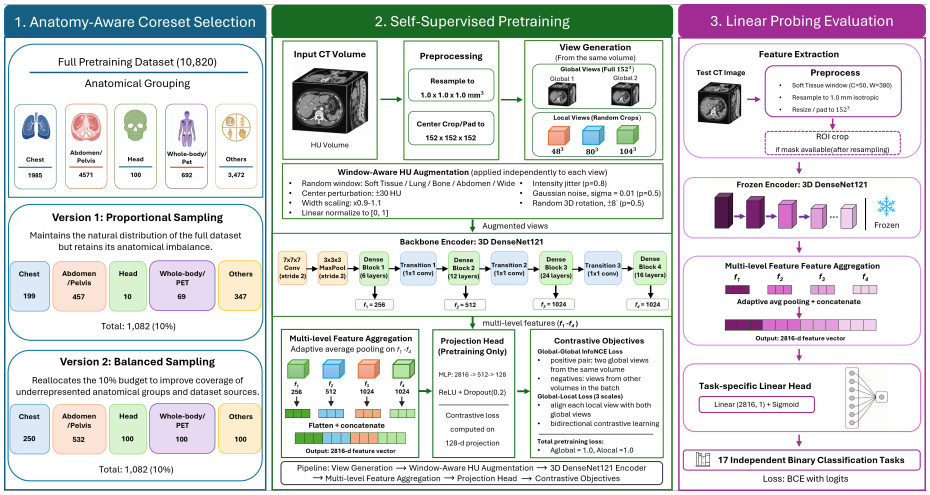

Fig. 1. Overview of the proposed framework for the Coreset Track. The pipeline consists of three stages: (1) anatomy-aware coreset selection, where the full pretraining dataset is grouped by anatomical region and sampled using proportional or balanced strategies; (2) self-supervised pretraining with window-aware HU augmentation and multi-scale global-local contrastive learning using a 3D DenseNet121 encoder; and (3) downstream LP evaluation, where the pretrained encoder is frozen and task-specific linear heads are trained for 17 independent binary classification tasks.

2 Method

This paper presents a self-supervised pretraining framework for 3D CT images, designed specifically for the data-constrained Coreset Track setting. The pipeline consists of three stages: (1) anatomy-aware coreset selection, (2) multi-scale global-local contrastive learning pretraining with window-aware augmentation, and (3) downstream evaluation via LP with a frozen backbone, as shown in Fig. 1.

2.1 Anatomy-Aware Coreset Selection

The Coreset Track requires participants to select 10% of the full pretraining dataset — 1,082 scans out of 10,820 — as the pretraining subset. We first partition the full dataset into five anatomical groups based on the primary body region covered by each scan: Chest (1,985), Abdomen/Pelvis (4,571), Head (100), Whole-body/PET (692), and Others (3,472). This grouping is not based on disease labels but on the primary anatomical region each scan corresponds to, allowing us to control the anatomical coverage of the coreset without requiring any annotation. We then compare two sampling strategies (**Version 1** and **Version 2**) built on top of this grouping.

Table 1. Anatomy-aware coreset distribution.

Anatomy	**Dataset**	**Total Count**	**Version 1**	**Version 2**
Chest	All Chest *, MSD_lung_, NIH-LYMPH-MED-	1985	199	250
Head	IntracranialHemorrhage_INSTANCE	100	10	100
Abdomen / Pelvis	amos_, KiTS, MSD_hepaticvessel_, MSD_pancreas_, MSD_colon_, MSD_liver_, NIH-LYMPH-ABD-, HCC_, WAWTACE_*, MSD_spleen_, Adrenal_Ki	4571	457	532
Whole-body / PET	psma_*, autoPET_fdg_* All	692	69	100
Others / unclear	Panorama_, mela_, lng,	3472	347	100
Total count		**10820**	**1082**	**1082**

Version 1: Proportional Sampling. The first strategy samples each anatomical group according to its original proportion in the full dataset, yielding 1,082 scans in total: Chest 199, Abdomen/Pelvis 457, Head 10, Whole-body/PET 69, and Others 347. Subsets within each group are also sampled proportionally. This approach preserves the overall distribution of the original dataset but inherits its imbalances — underrepresented groups contribute too few samples to be informative, while overrepresented groups may disproportionately influence downstream task performance.

Version 2: Balanced Sampling. The second strategy uses fixed quotas per group and distributes samples as evenly as possible across dataset sources within each group. After team discussion, we allocate Abdomen/Pelvis 532, Chest 250, Head 100, Whole-body/PET 100, and Others 100, for a total of 1,082 scans. Abdomen/Pelvis and Chest retain larger shares because they are most directly relevant to the majority of downstream tasks. Head and Whole-body/PET quotas are increased relative to their proportional allocations to prevent these minority groups from being nearly absent in the coreset. This design gives the coreset broader anatomical coverage and exposes the model to more diverse information during pretraining.

2.2 Window-Aware HU Augmentation

Rather than treating CT volumes as generic grayscale images, we incorporate CT windowing directly into the view generation process during pretraining. Recent work has also explored this direction: CoralBay [6] randomly samples from multiple clinically relevant HU windows spanning soft tissue, lung, abdomen, liver, and brain views during self-supervised pretraining, encouraging the encoder to learn representations that are invariant to windowing choices and robust across

anatomical regions. Building on this idea, each time a training view is generated in our pipeline, a clinical window setting is randomly sampled and applied to the raw HU values via clipping and normalization. This means the same CT volume can appear with different tissue contrasts across views, exposing the model to intensity variation that reflects the actual clinical diversity of CT imaging.

We use five window settings: Soft Tissue, Lung, Bone, Abdomen, and Wide, covering different contrast regimes for soft tissue, pulmonary structures, skeletal anatomy, and cross-tissue observation. After window clipping, all images are linearly normalized to [0, 1].

To prevent the model from memorizing fixed window boundaries, we add random perturbations after each window is selected: the window center is shifted by a uniform offset of ± 30 HU, and the window width is scaled by a random factor between 0.9 and 1.1. This introduces contrast variability for the same tissue type across training iterations and encourages the model to learn more stable CT intensity representations. Following window normalization, each view goes through a set of lightweight augmentations. Intensity jitter is applied with 80% probability, scaling intensity values by up to $\pm 10\%$ and shifting by up to ± 0.05. Gaussian noise with a standard deviation of 0.01 is added with 50% probability. Random 3D rotation of $\pm 8^\circ$ is applied with 50% probability. Since these augmentations are applied independently to each view, the same CT volume can differ across views in window setting, intensity distribution, and spatial orientation — preventing the model from overfitting to any fixed image appearance while still maintaining recognition of the underlying anatomical structures.

2.3 Pretraining Objective

Backbone Architecture

We use 3D DenseNet121 [12] as the backbone encoder, with a block configuration of (6, 12, 24, 16), a growth rate of 32, and 64 initial feature channels. The input CT volume is first downsampled by a $7 \times 7 \times 7$ convolution followed by $3 \times 3 \times 3$ max pooling. It is then processed through four Dense Blocks, separated by three Transition Layers. Each Transition Layer applies a 1×1 convolution to reduce the channel dimension, followed by $2 \times 2 \times 2$ average pooling to reduce the spatial resolution. The four Dense Blocks produce multi-level feature maps with channel dimensions of 256, 512, 1024, and 1024, denoted as f1 to f4.

View Generation

Raw images are loaded with SimpleITK and resampled to an isotropic spacing of $1.0 \times 1.0 \times 1.0$ mm. Spatial size is standardized to $152 \times 152 \times 152$ using center cropping or padding. From this standardized volume, we generate two global views and three local views.

Global views are generated from the full 152^3 volume, with each view independently processed by the window-aware augmentation pipeline. Local views are sampled at three crop sizes, 48^3, 80^3, and 104^3, to provide multi-scale local context. To avoid sampling patches dominated by background, we use a simple HU-variance heuristic: a candidate patch is accepted when its HU standard deviation is greater than 30. This sampling procedure is repeated up to 10 times

when needed, increasing the chance that local crops contain meaningful tissue structures. After cropping, local views are processed with the same augmentation pipeline and padded back to 152^3, so all views can be passed through the same encoder.

Contrastive Learning Objective

The pre-training loss consists of two components: a global-global contrastive loss and a global-local contrastive loss, with both weights set to 1.0. In the global-global branch, two global views generated from the same CT volume are treated as a positive pair, while global views from other volumes in the batch are used as negatives [3]. A full similarity matrix is computed over the 2B global features, and the logits are optimized with cross-entropy loss under the InfoNCE formulation [14].

For global-local learning, each local view is aligned with both global views of the same volume through bidirectional contrastive losses. The three local scales are handled separately, and their losses are averaged before being combined with the global-global loss. With this design, the model not only learns consistency across the entire CT volume but also establishes a correspondence between local anatomical regions and global representations.

Projection Head and Feature Aggregation

Inspired by the hypercolumn representation [7], which combines features from different network depths, we aggregate multi-level backbone features from f1 to f4 to facilitate both pretraining and downstream feature extraction. Each feature map is processed using adaptive average pooling, flattened, and concatenated into a 2,816-dimensional feature vector. As a result, this representation preserves information from different network depths and is used as the input to subsequent modules.

In the self-supervised pretraining stage, the aggregated 2,816-dimensional feature vector is passed through a two-layer Multi-Layer Perceptron projection head. It first maps the feature vector to 512 dimensions, followed by ReLU activation and dropout with a rate of 0.2, and then projects it to 128 dimensions for contrastive learning. During LP, the projection head is discarded, and the frozen encoder directly outputs the 2,816-dimensional feature vector used to train a task-specific linear classifier.

2.4 Linear Classification Head and Loss

Feature Extraction

After self-supervised pretraining, all encoder parameters are frozen. For downstream evaluation, each CT volume follows the same spatial processing used during pretraining, including resampling to 1.0 mm isotropic spacing and resizing to $152 \times 152 \times 152$ by center cropping or padding. Intensity normalization is performed using the Soft Tissue window, with a window center of 50 and a window width of 380, corresponding to an HU range of -140 to 240.

When a foreground mask is available, as in Region of Interest (ROI)-based tasks such as adrenal hyperplasia, ROI cropping is performed after resampling based on the mask center of mass. The extracted features therefore focus on the

target anatomical region rather than the full CT volume, reducing background interference in tasks involving small organs or localized lesions.

Classification Head and Loss

The downstream evaluation is formulated as a set of independent binary classification problems. For the 17 tasks, a separate linear head is trained on top of the frozen encoder representation, with no parameter sharing across task-specific heads. The classifier produces a single logit for each input volume, and the training objective is binary cross-entropy loss with logits. At inference time, a sigmoid function is applied to convert the logit into a predicted probability.

3 Experiments

3.1 Dataset and evaluation metrics

The data used in this study is divided into two parts: the pre-training dataset and the downstream evaluation dataset. The pre-training dataset consists of 10,820 CT volumes, covering the chest, abdomen, head, whole-body PET, and various other anatomical regions. Subject to the constraints of the Coreset Track, we selected 10% (1,082 instances) for self-supervised training from scratch. The sampling strategy is described in Section 2.1.

Regarding evaluation, the downstream evaluation dataset consists of three sources, totaling 17 independent binary classification tasks. AMOS-clf provides abdominal CT scans and 15 disease labels, including ascites, cholecystitis, fatty liver, gallstones, colorectal cancer, hepatic calcifications, hepatic cysts, liver lesions, splenomegaly, adrenal hyperplasia, hydronephrosis, kidney stones, renal cysts, atherosclerosis, and lymphadenopathy. Some tasks provide organ foreground masks, which can be used for ROI-based feature extraction. The COVID-CT dataset provides over 300 chest CT images, while the LUNA25 dataset contains 1,200 chest CT images with labels indicating whether lung nodules are benign or malignant.

Each task is evaluated using two metrics: Balanced Accuracy (BalAcc, the average of sensitivity and specificity) and Area Under the Receiver Operating Characteristic curve (AUROC). The former is less sensitive to class imbalance, while the latter reflects the model's overall discriminative ability at different decision thresholds. The official ranking includes the inference time for each case, and the final score is calculated by averaging the individual rankings across the three metrics.

3.2 Implementation details

Pre-training and downstream feature extraction share most of the preprocessing steps, but the order differs. During the pre-training stage, raw CT volumes are read with SimpleITK and resampled to $1.0 \times 1.0 \times 1.0$ mm isotropic spacing. Each volume is standardized to $152 \times 152 \times 152$: dimensions smaller than 152 are padded with a constant value of -1024 HU (air), and the padded volume is

center-cropped to the target size. This standardized volume serves as the input to the window-aware augmentation pipeline described in Section 2.2. In the feature extraction stage, we normalize the raw HU values using a soft-tissue window first, followed by resampling and cropping. This distinction stems from the contrasting requirements of the two stages. During pre-training, to encourage the model to discern variations across multiple windows within the same volume, the spatial size is fixed prior to window modulation. During inference, since only a single window is applied, performing normalization earlier does not alter the output; instead, it ensures consistency with the configuration of the official test Docker image for subsequent processing. For tasks where organ masks are provided, ROI cropping is performed as described in Section 2.4. Masks are resampled using the nearest-neighbor method to prevent interpolation from distorting their discrete label properties, while the volume itself is processed using trilinear interpolation.

The preprocessing transforms and data loading are built on the MONAI framework [2]. Handling thousands of memory-intensive 3D CT images on a single GPU introduces both I/O and Video Random Access Memory (VRAM) constraints. We configured the data loader with four workers and pinned memory to maximize data transfer efficiency and GPU utilization. To counter the physical batch size limit of 3 imposed by VRAM, 5-step gradient accumulation was integrated to stabilize training with an effective batch size of 15.

Environment settings The development environments and requirements are presented in Table 2.

Table 2. Development environments and requirements.

Item	**Setting**
System	Ubuntu 24.04.3 LTS
CPU	Intel(R) Core(TM) Ultra 7 265K, 20 threads
RAM	62 GiB
GPU	1 × NVIDIA GeForce RTX 5090
CUDA version	13.0
Programming language	Python 3.12.3
Deep learning framework	PyTorch 2.12.0+cu130

Training protocols This section describes the implementation details of the pre-training phase, including data augmentation, sampling strategies, and model selection criteria.

The data augmentation techniques used during training are the window-aware augmentation described in Section 2.2, followed by intensity jitter, Gaussian noise, and small-angle rotation. We did not use flipping or large-angle rotation because the left-right structure of CT images (such as the relative positions

of the liver and spleen) is significant. Using flipping might cause the model to learn incorrect invariance, so we selected augmentation techniques more suitable for medical images.

For data sampling, volumes are uniformly drawn from the coreset. For each volume within the same batch, two global views and three local views at different scales are generated. No additional class balancing is performed because no labels were used during the pretraining phase.

The model selection criterion is based on the training loss for each epoch (the sum of the global-global and global-local losses). If the loss for the current epoch is lower than the current best value, the corresponding encoder weights are retained as the final model weight. Subsequent downstream experiments use this version rather than the model from the last epoch at the end of training.

Table 3. Training Protocols for Task 1 – Coreset Track.

Item	**Setting**
Pretrained backbone	None, trained from scratch
Backbone tuning strategy	Full (self-supervised contrastive learning)
Batch size (pretraining)	3 (gradient accumulation $\times$ 5, effective batch size = 15)
Patch / volume size	152^3 (global), 48^3 / 80^3 / 104^3 (local)
Total pretraining epochs	80
Optimizer	AdamW (weight decay = 1×10^{-4})
Initial learning rate	3×10^{-4}
Lr decay schedule	Cosine annealing ($T_{\max} = 80$, $\eta_{\min} = 1 \times 10^{-6}$), no warm-up
Gradient clipping	max norm = 5.0
Pretraining time	9–10 hours
Feature extraction + LP time	$\sim$4.5 hours
Loss function	InfoNCE (global-global + global-local, $\lambda = 1.0$, temperature = 0.2)
Number of model parameters	Encoder: 11.24M; encoder + projection head (pretraining only): 12.75M

3.3 Quantitative results on the validation set

The quantitative experiments were organized into four parts. First, different backbone architectures were evaluated under the Coreset Version 1 setting to analyze the effect of model architecture and window-aware training. Second, the selected DenseNet121 backbone was compared under two coreset sampling strategies to examine the impact of subset construction. Third, we conducted an additional ablation comparing single-window and multi-window variants to check which method is better. Based on the previous experiments, the optimal combination was found to be DenseNet + V2 + single window. We then compared it with other baselines.

Ablation 1 – Backbone Comparison Using the Version 1 coreset based on proportional sampling, four backbone architectures were evaluated: DenseNet121 [12], ResNet34-SE [8,11], ViT3D [5], and VoCo [18]. All models were trained with the same pretraining subset and the same augmentation pipeline.

Table 4. Backbone comparison under Coreset V1. Mean BalAcc and AUROC across the 17 validation targets, grouped by anatomical region, along with mean per-case running time (feature extraction + head optimization) and encoder parameter count. For each column within a region, the best value is shown in **bold**.

Anatomical region	Backbone	BalAcc	AUROC	Runtime (s)	# Params (M)
		No Window			
Chest	ResNet34-SE	0.650	0.645	1.017	63.63
	ViT3D	0.670	**0.680**	1.061	26.83
	VoCo	0.645	0.635	1.048	27.09
	DenseNet121	**0.675**	0.650	**1.010**	**11.24**
Abdomen	ResNet34-SE	0.697	0.683	1.017	63.63
	ViT3D	0.688	0.689	1.061	26.83
	VoCo	0.695	0.683	1.048	27.09
	DenseNet121	**0.706**	**0.703**	**1.010**	**11.24**
Overall	ResNet34-SE	0.691	0.678	1.017	63.63
	ViT3D	0.686	0.688	1.061	26.83
	VoCo	0.689	0.678	1.048	27.09
	DenseNet121	**0.702**	**0.697**	**1.010**	**11.24**
		With Window			
Chest	ResNet34-window	0.680	0.650	**1.032**	63.63
	VoCo-window	0.680	0.650	1.108	27.09
	ViT3D	0.660	0.650	1.261	26.83
	DenseNet121-window	**0.685**	**0.660**	1.039	**11.24**
Abdomen	ResNet34-window	0.695	0.695	**1.032**	63.63
	VoCo-window	0.687	0.683	1.108	27.09
	ViT3D	0.691	0.693	1.261	26.83
	DenseNet121-window	**0.717**	**0.718**	1.039	**11.24**
Overall	ResNet34-window	0.694	0.690	**1.032**	63.63
	VoCo-window	0.686	0.679	1.108	27.09
	ViT3D	0.688	0.688	1.261	26.83
	DenseNet121-window	**0.714**	**0.711**	1.039	**11.24**

We used a Version 1 coreset constructed using proportional sampling to evaluate several backbone network architectures under both window-free and windowed settings. In the no window scenario, DenseNet121 performed best.

After introducing windowed training, DenseNet121-window maintained its lead across all anatomical regions. Its BalAcc/AUROC for the chest reached 0.685/0.660, for the abdomen 0.717/0.718, and for the overall set 0.714/0.711. Compared to the window-agnostic DenseNet121 configuration, the BalAcc metric improved from 0.702 to 0.714, and the AUROC metric improved from 0.697 to 0.711. This confirmed that window-aware training can help the model distinguish

different diseases. Subsequent experiments therefore adopt the window-aware DenseNet121 as the fixed backbone.

Ablation 2 – Coreset Strategy Comparison After finalizing the backbone model as DenseNet121, we compared the two coreset sampling strategies mentioned in Section 2.1.

Table 5. Coreset Version 1 vs. Version 2 (DenseNet121 backbone). Mean BalAcc and AUROC across the 17 validation targets, grouped by anatomical region. For each metric within a region, the best score is shown in **bold**.

Anatomical region	Coreset	BalAcc	AUROC
Chest	**V1 (Proportional)**	**0.685**	0.660
	V2 (Balanced)	0.640	**0.670**
Abdomen	V1 (Proportional)	**0.717**	0.718
	V2 (Balanced)	0.715	**0.723**
Overall	V1 (Proportional)	**0.714**	0.711
	V2 (Balanced)	0.706	**0.716**

As shown in Table 5, the two strategies trade off between the metrics: proportional sampling (V1) attains higher BalAcc, whereas balanced sampling (V2) attains higher AUROC. We prioritize AUROC because it reflects threshold-independent class separability and is more robust to class imbalance: a higher AUROC indicates that V2 separates positive and negative cases more reliably regardless of the operating point, which the linear probe can subsequently tune. We therefore adopt V2, trading a small drop in overall BalAcc ($0.714 \rightarrow 0.706$) for improved AUROC ($0.711 \rightarrow 0.716$) and better-separated features.

In the chest subset, V2 improved the AUROC from 0.660 to 0.670. But the balanced accuracy (BalAcc) decreased from 0.685 to 0.640. This discrepancy may be related to the limited number of chest samples in the balanced coreset. In the abdominal subset, the differences were small: V1's BalAcc/AUROC was 0.717/0.718, while V2's BalAcc/AUROC was 0.715/0.723. Although the V2 version did not improve all metrics, in a data-constrained setting, it can improve AUROC performance. Furthermore, considering the diversity of this model's future applications, we believe that V2 is more suitable for medical imaging applications.

Ablation 3 – Single vs. Multi-Window Comparison To further analyze the influence of CT window configuration, we compared single-window and multi-window feature extraction under the selected DenseNet121-based framework.

The results show that the single-window method performs more stable than the multi-window method. Although the multi-window approach exposes the

Table 6. Single-window and multi-window feature extraction method comparison. For each metric within a region, the best score is shown in **bold**.

Anatomical region	method	BalAcc	AUROC
Chest	**single**	**0.640**	**0.670**
	multiple	**0.640**	0.635
Abdomen	**single**	0.715	**0.723**
	multiple	**0.717**	0.707
Overall	**single**	0.706	**0.716**
	multiple	**0.708**	0.698

model to more information across different contrast levels, the results were not actually better.

One possible reason is that the multi-window approach applies the same image to multiple window positions and then combines the results. In effect, this distributes the signal across different contrast spaces, which may cause features to be more dispersed and affect the decision criteria of downstream classifiers. In contrast, the single-window approach consistently applies the same contrast settings to each image, resulting in greater feature consistency and, consequently, more stable model training.

Result – Comparison with Official Baselines Comparing our best-performing (DenseNet121 + Coreset V2 + Single-window) combination with three official baselines: CT-NEXUS [17,16], 3DINO [19], and VoCo [18].

For the Chest subset, our model attains the second-best AUROC (0.670), behind only 3DINO and ahead of CT-NEXUS and VoCo. Its BalAcc (0.640), however, is lower than that of 3DINO and VoCo. This gap between AUROC and BalAcc is informative: the model can still rank chest cases well, but the selected operating point does not translate that ranking into equally reliable class decisions. A reason is the smaller share of chest-specific samples in the balanced coreset, which may leave the decision boundary less stable for chest-related targets.

On the Abdomen subset, the model reports a BalAcc of 0.715 and an AUROC of 0.723, the second-best on both metrics, behind only 3DINO and ahead of CT-NEXUS and VoCo. The stronger abdomen results suggest that the anatomy-aware coreset and CT window-aware augmentation fit abdomen-related tasks particularly well, where organ-level context and intensity variation are central to the prediction.

Across the 17 validation targets, our model obtains a BalAcc of 0.706 and an AUROC of 0.716, ranking second on both metrics, trailing only 3DINO while staying above CT-NEXUS and VoCo. Relative to CT-NEXUS, BalAcc increases by 0.084 and AUROC by 0.104. Relative to VoCo, the gains are 0.016 and 0.032. Taken together, these results suggest that even with only a 10% coreset, a com-

Table 7. Final comparison with official baselines. Mean BalAcc and AUROC across the 17 validation targets, grouped by anatomical region. For each metric within a region, the best score is shown in **bold** and the second best is <u>underlined</u>.

Anatomical region	Method	BalAcc	AUROC	# Params (M)
Chest	CT-NEXUS	0.604	0.570	∼90*
	3DINO	**0.701**	**0.705**	∼86−300*
	VoCo	<u>0.649</u>	0.635	∼25−30*
	Ours	0.640	<u>0.670</u>	**11.24**
Abdomen	CT-NEXUS	0.624	0.618	∼90*
	3DINO	**0.722**	**0.740**	∼86−300*
	VoCo	0.695	0.691	∼25−30*
	Ours	<u>0.715</u>	<u>0.723</u>	**11.24**
Overall	CT-NEXUS	0.622	0.612	∼90*
	3DINO	**0.719**	**0.736**	∼86−300*
	VoCo	0.690	0.684	∼25−30*
	Ours	<u>0.706</u>	<u>0.716</u>	**11.24**

*Parameter counts for CT-NEXUS, 3DINO, and VoCo are estimated from their publicly available architecture configurations (ResidualEncoderUNet, 3D ViT, and Swin-UNETR with feature_size=48, respectively); exact counts were not independently verified by re-instantiating each model. The parameter count for our method (DenseNet121) was computed directly from the model definition.

pact DenseNet121 encoder trained from scratch can still learn useful CT representations, with the clearest benefit seen on abdomen-related tasks.

3.4 Qualitative Analysis of the Validation Set

Abdominal Task Analysis: Fatty Liver.
For the abdominal imaging task, fatty liver was selected as a representative target. As shown in Fig. 2(a), the easy case is a correctly classified negative case, with a predicted probability of $P(\text{class} = 1) = 0.001$. The liver parenchyma in this slice appears relatively homogeneous, and no obvious diffuse hypodense change is observed compared with the surrounding abdominal structures. Therefore, the model confidently ruled out fatty liver in this case.

In contrast, Fig. 2(b) shows a hard case for the same target. Although the ground truth is negative, the model incorrectly predicted a positive label with $P(\text{class} = 1) = 0.562$. This probability is close to the decision boundary, suggesting that the model was uncertain in this case. One possible reason is that the liver margin is partially compressed, leading to visually uneven intensity and grayscale distribution within the liver region. In addition, the stomach and adjacent soft tissues may interfere with the liver appearance under the soft tissue window, causing the model to misinterpret localized attenuation variation as fatty liver-related features.

Chest Task Analysis: COVID-CT.
For the chest imaging task, COVID-CT was selected as the representative target. Figure 2(c) shows an easy negative case that was correctly classified by the

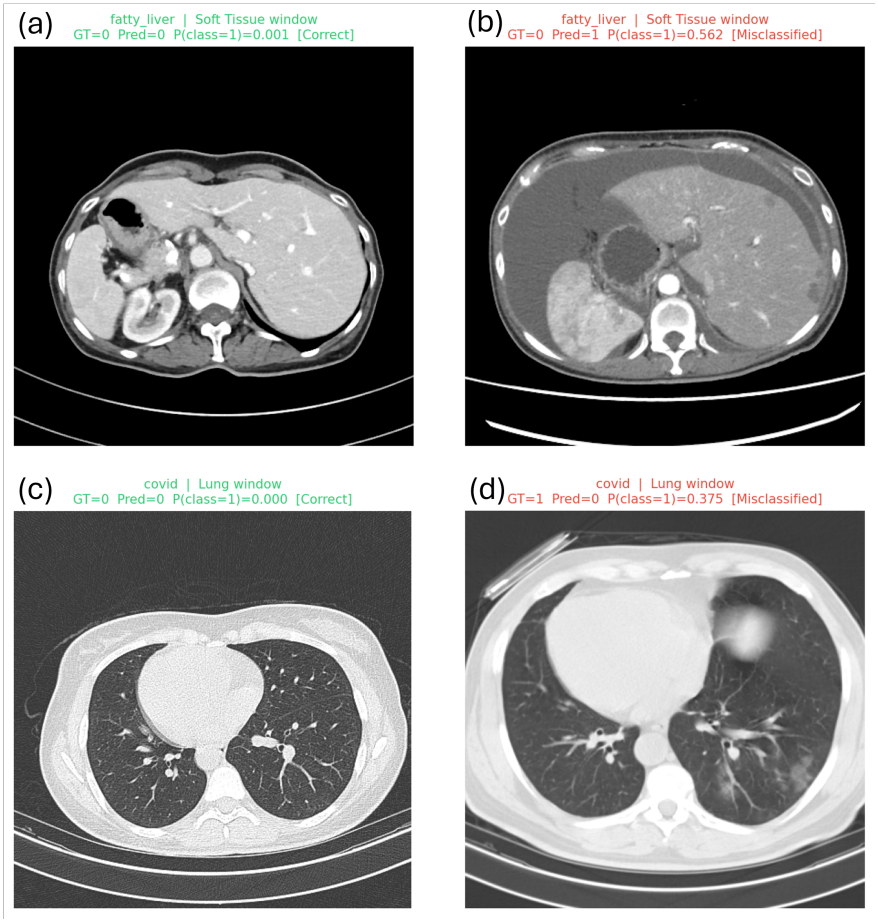

Fig. 2. Qualitative analysis on the validation set. (a) An easy and correctly classified case for fatty liver. (b) A hard and misclassified case for fatty liver. (c) An easy and correctly classified case for COVID-CT. (d) A hard and misclassified case for COVID-CT. Each panel shows a representative axial slice together with the ground-truth label, predicted label, and predicted probability.

model, with $P(\text{class} = 1) = 0.000$. The lung fields are clear, and no apparent ground-glass opacity, infiltration, or other visible abnormality is observed. This allows the model to make a stable negative prediction.

Figure 2(d) shows a false-negative case, where the ground truth is positive but the model predicted a negative label with $P(\text{class} = 1) = 0.375$. A subtle ground-glass opacity can be observed near the peripheral region of the lower lung field, but the lesion is small and has relatively low contrast. This makes the abnormal finding difficult to distinguish from normal lung texture. Moreover, the image quality is relatively limited, and motion-related blurring near the cardiac border may further obscure the boundary between lung structures and adjacent soft tissues. During global feature aggregation, such localized abnormal features may be diluted by the large amount of normal lung tissue in the volume. This result suggests that the model still has room for improvement in capturing small lesions and local details.

3.5 Results on the final testing set

On the final testing set, our team, *ilalab*, took first place in the Coreset Track, with a Balanced Accuracy of 0.616, an AUROC of 0.647, and an average execution time of 4.307 seconds per case. Considering that the Coreset Track allows the use of only 10% of the pretraining data and that the model must be trained entirely from scratch, this is an encouraging result under the data-constrained setting.

Although the training conditions for the Coreset Track and the All-data Track differ, making a completely fair comparison impossible, if we simply refer to the final ranking scores, our method would likely place around fourth in the All-data Track. In other words, even when using only a small amount of data, our method remains competitive. It also shows that the combination of anatomy-aware coreset selection, window-aware CT augmentation, and multi-scale global-local contrastive learning can still learn useful CT representations under data-constrained conditions.

3.6 Limitation and future work

Based on the results, our method shows competitive performance under the Coreset Track setting, but some tasks remain challenging. While current coreset selection relies on rule-based grouping and sampling based on anatomical regions and dataset sources, it does not use the image content itself to assess the representativeness or difficulty of samples. Though this design aligns with the constraints of a from-scratch approach and unlabelled data selection and avoids reliance on external pre-trained models, differences in scanning protocols, contrast phases, or disease distributions may still exist within the images. Future work could leverage more granular image features or HU distribution information without relying on labels. By doing so, coreset selection could not only maintain anatomical balance but also better capture image diversity.

Ranking Results: CT Foundation Model (Task 1)

All-data Track

Rank	Team name	Balanced Acc	AUROC	Runtime
1	fomofo	0.654	0.714	9.414
1	raidium	0.639	0.707	4.393
3	cemrg	0.643	0.705	4.467
4	maest	0.600	0.624	6.273
5	xupy	0.566	0.599	10.86
6	sdfmu	0.160	0.170	33.83

Coreset Track

Rank	Team	Balanced Acc	AUROC	Runtime
1	ilalab	0.616	0.647	4.307
2	ahus	0.606	0.638	4.66

Fig. 3. Final testing ranking results for Task 1. Our team, *ilalab*, ranked first in the Coreset Track. The results show that the proposed method remains competitive even when only 10% of the pretraining data is used.

Second, in the LP stage, this study uses global feature representations from a frozen encoder and employs multi-level aggregation to concatenate features from different depths into fixed-length vectors. While this design is simple and effective, it may still lack sufficient sensitivity for small lesions or localised abnormalities. Qualitative analysis of COVID-CT misclassification cases shows that localised ground-glass opacities have a small extent, which may be diluted by a large amount of normal lung tissue signals during global feature aggregation. Similar issues may also arise in tasks such as the classification of malignant lung nodules, liver cysts, or other small ROIs. In the future, we could explore incorporating attention-based aggregation strategies, whereby the model can more effectively capture local lesion features while retaining global structural information.

In addition, although multi-level feature aggregation is designed to combine shallow and deep representations, its individual contribution has not yet been isolated in this study. Future work could compare the current f1–f4 concatenated representation with simpler alternatives, such as using only the final feature map or using different combinations of intermediate features, to better understand the role of multi-level aggregation in downstream LP performance.

Regarding window-aware augmentation, the current approach primarily uses five fixed clinical CT windows with random perturbations in centre and width. While this method exposes the model to different tissue contrasts during pretraining, the window combinations are still manually defined. Moreover, our single-window and multi-window comparison suggests that using more window

settings during feature extraction does not always lead to better downstream performance. Future work could therefore explore task-adaptive or anatomy-adaptive window selection to apply more appropriate intensity transformations for different anatomical regions, disease targets, or data sources.

4 Conclusion

This study participated in Task 1 of the Coreset Track at the CVPR 2026 Workshop, exploring how to build transferable 3D CT representations through data subset design and CT-specific augmentation within the Coreset budget. Different from pre-training methods that rely on complete, large-scale datasets, we focus on the more constrained setting of the Coreset Track: using only about 10% of unlabeled CT volumes and training the model from scratch.

To preserve sufficient anatomical diversity within the limited data, we divided the full pre-training dataset into chest, abdomen/pelvis, head, whole-body/PET, and others based on the primary scan region and compared two coreset strategies: proportional sampling and balanced sampling. The results show that the two sampling strategies emphasize different aspects of performance. Proportional sampling tends to provide higher BalAcc, while balanced sampling improves AUROC and feature separability. This indicates that, when data is limited, the composition of the selected subset directly influences the range and stability of representations the model can learn.

Regarding the pretraining method, we incorporated the HU windowing characteristics of CT into the self-supervised learning process by designing window-aware HU augmentation, enabling the model to be exposed to different tissue contrasts in soft tissue, lung, bone, abdomen, and wide-window scenarios during training. We then employed multi-scale global-local contrastive learning to allow the model to simultaneously learn the global structure of the entire CT volume and the local features of local crops at different scales. Using a 3D DenseNet121 backbone combined with multi-level feature aggregation, the model integrates features from different depths into a fixed representation for subsequent LP evaluation across 17 binary classification tasks.

On the validation set, our method achieved an overall BalAcc of 0.706 and an AUROC of 0.716. Compared to the CT-NEXUS and VoCo baselines, our method performs better on the overall metrics, while 3DINO still provides the strongest validation performance. On the final testing set, our team, *ilalab*, ranked first in the Coreset Track, with a BalAcc of 0.616, an AUROC of 0.647, and an average runtime of 4.307 seconds per case. Therefore, even when using only 10% of the pretraining data and training entirely from scratch, the proposed system can still achieve competitive performance when paired with appropriate data selection and CT-specific augmentation strategies.

The main finding of this study is that, in a coreset CT pretraining setting, model performance depends on more than just the backbone architecture itself. How data is selected, how HU intensity is transformed, and how global and local features are aligned are all factors that influence representation learning. Future

model improvements could focus on content-aware data configuration, adaptive window selection, and more detailed ablation studies on multi-level feature aggregation. Incorporating methods such as attention-based aggregation may further enhance the model's ability to detect small lesions and local abnormalities.

Acknowledgements We thank all the data owners for making the medical images publicly available and CodaLab [20] for hosting the challenge platform. This work was partially supported by the National Science and Technology Council, Taiwan, under Grant No. NSTC 114-2222-E-182-001-MY2.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Blankemeier, L., Kumar, A., Cohen, J.P., Liu, J., Liu, L., Van Veen, D., Gardezi, S.J.S., Yu, H., Paschali, M., Chen, Z., et al.: Merlin: a computed tomography vision–language foundation model and dataset. Nature pp. 1–11 (2026) 1
2. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A.: Monai: An open-source framework for deep learning in healthcare (2022), https://arxiv.org/abs/2211.02701 9
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (13–18 Jul 2020), https://proceedings.mlr.press/v119/chen20j.html 7
4. DenOtter, T.D., Schubert, J.: Hounsfield unit. In: StatPearls. StatPearls Publishing, Treasure Island (FL) (2023), https://www.ncbi.nlm.nih.gov/books/NBK547721 3
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy 11
6. Gatopoulos, I., Känzig, N., Otálora, S., Tang, F.: Coralbay: A self-supervised ct foundation model. arXiv preprint arXiv:2606.03888 (2026) 1, 5
7. Hariharan, B., Arbelaez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015) 7

8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 11
9. Hong, Y., Zhang, X., Zhang, X., Zhou, J.T.: Evolution-aware variance (eva) coreset selection for medical image classification. In: Proceedings of the 32nd ACM International Conference on Multimedia (2024). https://doi.org/10.1145/3664647.3681592 3
10. Hounsfield, G.N.: Computerized transverse axial scanning (tomography): Part 1. description of system. The British journal of radiology **46**(552), 1016–1022 (1973) 2
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 11
12. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017) 6, 11
13. Kim, S., Bae, S., Yun, S.Y.: Coreset sampling from open-set for fine-grained self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7537–7547 (June 2023) 3
14. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2019), https://arxiv.org/abs/1807.03748 7
15. Oquab, M., Darcet, T., Moutakanni, T., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024) 2
16. Vaish, P., Meister, F., Heimann, T., Brune, C., Wolterink, J.M.: Consistent view alignment improves foundation models for 3d medical image segmentation (2025), https://arxiv.org/abs/2509.13846 13
17. Vaish, P., Meister, F., Heimann, T., Brune, C., Wolterink, J.M.: CT-NEXUS: Consistent view alignment improves foundation models for 3d medical image segmentation. https://github.com/kmin940/CT-NEXUS (Sep 2025), version 1.0.0; released September 19, 2025; accessed June 29, 2026 13
18. Wu, L., Zhuang, J., Chen, H.: Large-scale 3D medical image pre-training with geometric context priors. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 1, 2, 11, 13
19. Xu, T., et al.: A generalizable 3D framework and model for self-supervised learning in medical imaging. npj Digital Medicine **8**(1), 639 (2025) 2, 13
20. Xu, Z., Escalera, S., Pavao, A., et al.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 19

Table 8. Checklist Table. Please fill out this checklist table in the answer column. (**Delete this Table in the camera-ready submission**)

Requirements	Answer
A meaningful title	Yes
The number of authors ($\leq$6)	Yes
Author affiliations and ORCID	Yes
Corresponding author email is presented	Yes
Task and track (LP All / LP Coreset / EAO VO / EAO VL) are clearly stated in the abstract	Yes
Validation BalAcc and AUROC are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	Yes
Pre-processing	Sec. 3.2
Strategies for data augmentation	Sec. 2.2
Strategy for class imbalance	Sec. 2.1 / Table 1
Strategies to improve model inference (under 40s/case)	Sec. 3.2 / Table 3
Post-processing	N/A
Environment setting table is provided	Table 2
Training protocol table is provided (Task 1 and/or Task 2)	Table 3
Ablation study	Sec. 3.3 / Tables 4–6
Efficiency evaluation results (running time) are provided	Table 4, Figure 3
Qualitative example (success and failure) is provided	Fig. 2
Limitation and future work are presented	Yes
Reference format is consistent	Yes
Main text $\geq$ 8 pages (not including references and appendix)	Yes