Anatomical Invariance Modeling and Semantic Alignment for Self-supervised Learning in 3D Medical Image Analysis

Yankai Jiang[1,3*], Mingze Sun[1,4*], Heng Guo[1,2], Xiaoyu Bai[1], Ke Yan[1,2], Le Lu[1] and Minfeng Xu[1,2✉]

[1]DAMO Academy, Alibaba Group
[2]Hupan Lab
[3]College of Computer Science and Technology, Zhejiang University
[4]Tsinghua Shenzhen International Graduate School, Tsinghua-Berkeley Shenzhen Institute, China
✉eric.xmf@alibaba-inc.com

Abstract

*Self-supervised learning (SSL) has recently achieved promising performance for 3D medical image analysis tasks. Most current methods follow existing SSL paradigm originally designed for photographic or natural images, which cannot explicitly and thoroughly exploit the intrinsic similar anatomical structures across varying medical images. This may in fact degrade the quality of learned deep representations by maximizing the similarity among features containing spatial misalignment information and different anatomical semantics. In this work, we propose a new self-supervised learning framework, namely **Alice**, that explicitly fulfills **A**natomical **i**nvariance modeling and semanti**c** alignm**e**nt via elaborately combining discriminative and generative objectives. **Alice** introduces a new contrastive learning strategy which encourages the similarity between views that are diversely mined but with consistent high-level semantics, in order to learn invariant anatomical features. Moreover, we design a conditional anatomical feature alignment module to complement corrupted embeddings with globally matched semantics and inter-patch topology information, conditioned by the distribution of local image content, which permits to create better contrastive pairs. Our extensive quantitative experiments on three 3D medical image analysis tasks demonstrate and validate the performance superiority of **Alice**, surpassing the previous best SSL counterpart methods and showing promising ability for united representation learning. Codes are available at https://github.com/alibaba-damo-academy/alice.*

1. Introduction

Since the advent of deep learning, the lack of high-quality annotated data has long been a thorny challenge in medical image analysis, especially for 3D tasks. Recent research efforts based on self-supervised learning (SSL) shed light on acquiring strong visual representations in an unsupervised manner [7, 8, 16, 21, 26, 42].

Nowadays, contrastive learning (CL) [3, 11, 35, 50] and masked image modeling (MIM) [2, 20, 48], together with Vision Transformers (ViTs) [15, 29], have revolutionized the field of SSL in computer vision and medical imaging, which achieve the state-of-the-art (SOTA) performance for a variety of tasks [3, 20, 40, 47, 56]. There is also a growing trend to combine CL and MIM in a self-distillation way to design more powerful SSL frameworks [23, 40, 41, 56]. Despite popularity and success, these methods still follow the self-supervised paradigm designed for specific computer vision scenarios, *e.g.*, ImageNet (ILSVRC-2012) [36], which can be less suitable or irrational when applied to medical images. Now we analyze the drawbacks and irrationalities of existing hybrid SSL approaches, which combine CL with MIM, from the following aspects:

(i) Neglecting the intrinsic similar anatomical structure across varying medical image volumes. Commonly-used computed tomography (CT) and magnetic resonance (MR) images render human anatomies with intrinsic structures. As shown in Fig. 1a, the definition of positive and negative pairs in existing siamese SSL frameworks [3, 23, 40, 41, 56] ignore the semantically consistent anatomical features across different volumes and force an incorrect constraint of instance invariance. Intuitively, utilizing the intrinsic anatomical structure across different image volumes to model the class-specific invariance can help the learned representations more robust to the size, shape, and texture
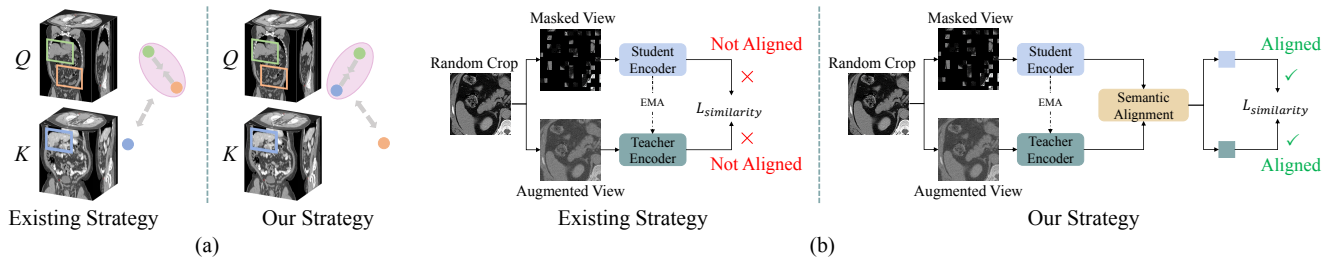
Figure 1. The motivation of our proposed method **Alice**. (a) Existing SSL methods [3, 11, 23, 40, 41, 56] simply treat image patch samples which may depict totally different anatomical information from the same CT volume as positive pairs while considering samples sharing the same semantic content class but from another volume as negatives. **Alice** leverages the consistent anatomical structures across different volumes and addresses the false positive and false negative pairs. (b) shows the defect of existing hybrid SSL [23, 41, 56] methods, which ignore the large semantic gap between masked views and augmented views. Differently, **Alice** performs anatomical information alignment and thus crafts better contrastive pairs.

variances of body parts.

(ii) Lack of anatomical semantic alignment for views extracted or sampled from the same image volume. As shown in Fig. 1b, the widely used siamese architecture in hybrid SSL approaches [12, 23, 41, 56] maximize the similarity between the representations of masked view and intact view. However, a random crop of a body part may contain different organs and human tissues (some of which are small in scale). A large masking ratio would already erase these contents and make the masked view quite distinct from the intact one. Thus, maximizing the similarity between views which incorporate totally different semantics can be harmful to the learned representations. To learn good representations for downstream tasks, SSL methods for medical images should align the anatomical features of the views which form a positive pair.

Driven by the aforementioned limitations, we present a simple, effective, and dedicated self-supervised learning framework for 3D medical segmentation tasks, **Alice**, by explicitly fulfilling **A**natomica**l i**nvariance modeling and semanti**c** alignm**e**nt through elaborately combined contrastive learning and MIM. From Fig. 1, **Alice** leverages the structural similarity across different volumetric images to explicitly learn universal consistent features from intrinsic body structures and model the anatomical invariance which is robust to the size, shape, intensity, and texture diversity of body parts caused by inter-subject variation, organ deformation, and pathological changes.

Moreover, we design a conditional anatomical feature alignment module which complements masked views with the globally matched anatomical semantics and inter-patch topology to craft better contrastive pairs, enforcing that the positive pairs encoded to semantically consistent feature representations. This process explicitly realizes anatomical semantic alignment to further strengthen the representation with spatial sensitivity and semantic discriminability.

To adequately validate the effectiveness of **Alice**, we employ 3D medical image segmentation and classification as downstream tasks. We fine-tune these widely used ViT-based medical image segmentation frameworks following [19, 40, 53] with our pre-trained weights on two publicly available benchmarks: Fast and Low-resource semi-supervised Abdominal oRgan sEgmentation in CT (FLARE 2022)*, and Beyond the Cranial Vault (BTCV) [27]. Our method achieves the current SOTA results, with 86.87% Dice on FLARE 2022 and 86.76% Dice on BTCV, surpassing previous best results by 2.22% and 1.77% respectively. We also evaluate transfer learning on a public COVID-19 classification benchmark [31]. **Alice** outperforms state-of-the-art counterpart methods by 2.52% in AUC.

Our main contributions can be summarized as:

- We investigate the irrationalities of commonly used siamese SSL frameworks applied to medical images. We propose **Alice** that is customized to leverage the anatomical similarity across volumetric medical images to model class-specific invariance.

- In **Alice**, a conditional anatomical semantic alignment module is proposed to match the most related high-level semantics between the crafted contrastive views.

- **Alice** consistently outperforms popular SSL methods on three public downstream benchmarks, showing its effectiveness and generality.

2. Related Work

Self-supervised Learning. Contrastive learning is the most popular method in self-supervised learning, which achieves remarkable performance on downstream tasks in computer vision [3, 7–9, 11, 16, 21]. The essence of contrastive learning aims to learn view-invariant representations by maximizing the similarity between the features extracted from

---

*https://flare22.grand-challenge.org/

different crops of the same image. More recently, some works explore how to learn representations by combining contrastive learning and MIM [23, 41, 56]. As pointed out in [34], these methods adopt random sampling to make different crops of the same image, which overlooks the semantic information and may generate views truly contain different image contents. A few newer investigations [34, 37, 52] attempt to leverage semantic guided information to crop semantically consistent views. However, they ignore the potential positive pairs in other images, restraining the diversity of learned representations. A key difference between these methods and **Alice** is that we mine diversified contrastive views with semantic similar contexts across varying images to encode the intrinsic structure of consistent anatomy information and facilitate the class-specific invariance.

Masked Image Modeling (MIM) accepts input image corrupted by masking and predicts the target of the masked content, which has been actively studied recently in self-supervised learning. Existing work mainly differ in their regression objectives [2, 14, 20, 43, 45, 48] or masking strategies [25, 28, 38]. In this work, **Alice** takes one step further by exploiting cooperations between contrastive learning and MIM to learn effective representations with both strong instance discriminability and local detail sensitive perceptibility, from varying or different image views.

Self-supervised Learning in Medical Imaging. Many works apply tailored contrastive SSL methods to medical image problems [4, 26, 39, 47] with reasonable success. Similar to MIM, image restoration is also commonly used as pre-text task to memorize spatial context from medical images. Typical attempts include inpainting tasks [1, 5, 59], Rubik's cube problem [60] and diverse context reconstruction [55]. Most recently, DiRA [17] employs discriminative [7], restorative [6], and adversarial learning [13] objectives simultaneously in a unified SSL framework for medical image analysis. Swin UNETR [40] trains a transformer-based encoder with combination of different pre-text tasks for 3D medical image segmentation. These efforts constitute important steps toward better SSL methods for medical image analysis. Nevertheless, **Alice** distinguishes itself by having two key new developments: (1) explicitly leveraging the inherent anatomical consistency between different image volumes to encode the class-specific invariance; and (2) fulfilling anatomical semantic alignment to craft better contrastive pairs.

3. Method

The overall framework of our method is illustrated in Fig. 2. Our **Alice** model consists of two branches performing masked image modeling and contrastive learning, respectively. We first mine diverse yet semantically consistent crops from varying image volumes. A query vol-

ume and a key volume are randomly picked. Then we adopt a pre-trained SAM [49] model, which performs self-supervised universal landmark detection to locate the same body part in different volumetric medical images and produce two crops, denoted as Q and K, that depict two sub-volumes from the same body part. After that, we utilize two different data augmentations, u and w, to generate two views of Q. We denote them as X_u^Q and X_w^Q. Likewise, we also utilize two different data augmentations, r and v, to generate two views of K, denoted as X_r^K and X_v^K. Here, u and r are random masking, which is similar to the "random sampling" adopted in MAE [20]. w and v are two different data augmentation operations. In the following, we elaborate on each component of **Alice** in details and describe how we optimize the relationships between and among these four views.

3.1. Network Architecture Components

Online Encoder. The online encoder E_s takes masked view pairs (X_u^Q and X_r^K) as inputs. Following MAE [20], we only feed the visible, unmasked patches to the online encoder E_s. E_s embeds visible tokens added with the positional embeddings and produces the output features (V_u^Q and V_r^K) through a sequence of transformer blocks. In this paper, we mainly consider two vision Transformer architectures (ViT [15] and Swin Transformer [29]) as the online encoder options. We adopt the strategy in [22] to allow the Swin Transformer to discard masked patches and operate only on the visible ones. After pre-training, only the online encoder E_s is used for extracting image representations in downstream tasks.

Online Decoder. In addition to the features of visible patches V_u^Q and V_r^K, the online decoder receives mask tokens as inputs. We add positional embeddings to all tokens. Then the online decoder learns to reconstruct the pixel of the masked patches. Following MAE [20], we use the normalized pixel values as targets in the reconstruction task. Our loss function ℓ_r computes the mean squared error (MSE) on masked patches between the decoder predictions ($\bar{Q}, \bar{K}$) and the original input volume crops (Q, K):

$$\ell_r = \frac{1}{n_m^Q} \sum [\Theta(\bar{Q} - Q)]^2 + \frac{1}{n_m^K} \sum [\Theta(\bar{K} - K)]^2, \quad (1)$$

where Θ is an indicator to select the prediction corresponding to masked tokens, n_m^Q and n_m^K are the number of masked patches in Q and K, respectively. This MIM objective helps the learned representations encode local context and patient-specific information of input volumes. The online decoder is set to be stacked transformer blocks.

Target Encoder. Following existing siamese frameworks [23, 41, 56], we introduce a target encoder to generate contrastive supervision for the online encoder to further strengthen the representation learned by MIM with se-
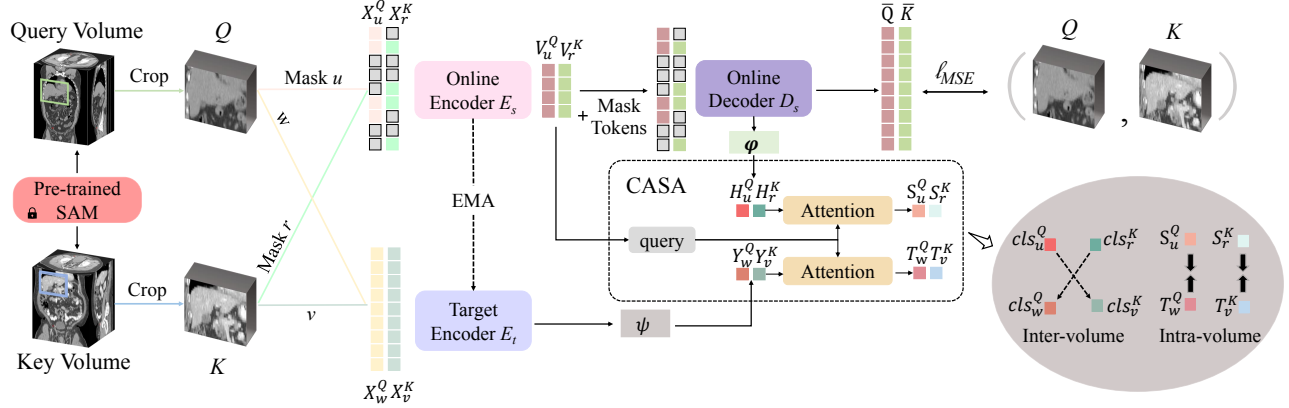
Figure 2. Overall pipeline. **Alice** contains three components: the online encoder, target encoder and online decoder. It first acquires two registered image patch crops from different CT volumes, using SAM [49]. Then four different augmented views from crops are fed into the online and target encoders respectively. The online encoder randomly masks a fraction of the image patches and operates on the remaining visible image content. The target encoder operates on the whole view. The online decoder learns to reconstruct the input volume. We propose a conditional anatomical semantic alignment (CASA) module to craft better contrastive pairs. After the pre-training, only the online encoder is kept for downstream segmentation tasks.

mantic discriminability. The target encoder shares the same architecture as the online encoder. We update parameters of the target encoder using an exponential moving average (EMA) of the online encoder weights. The target encoder takes two unmasked augmented views X_w^Q and X_v^K as inputs and embeds them into high dimensional feature representations which reserve the semantic integrity.

3.2. Anatomical Invariance Modeling

A prominent difference between previous methods and **Alice** is that we mine diversified views from different volumes to learn anatomical features that are universal across similar body parts. While existing methods [23, 40, 56] only operate on the same volume, we propose to explicitly model the anatomical invariance by maximizing the similarity between embeddings of views from Q and K.

As shown in Fig. 2, we append a projection head φ and another projection head ψ to the online decoder and target encoder respectively to produce positive feature pairs (H_u^Q, Y_v^K) and (H_r^K, Y_w^Q). φ and ψ both consist of a 3-layer MLPs with l_2-normalized bottleneck following DINO [3]. We then adopt global average pooling to these features and obtain their global visual semantics, denoted as $([cls]_u^Q, [cls]_v^K)$ and $([cls]_r^K, [cls]_w^Q)$. We encourage their high-level class-specific representations move closer in the corresponding feature space. This yields the loss:

$$\ell_{dv} = \ell_s([cls]_u^Q, [cls]_v^K) + \ell_s([cls]_r^K, [cls]_w^Q), \quad (2)$$

where ℓ_s denotes a general cosine similarity loss [10, 16]. Through optimizing such inter-volume relationship, Alice explicitly enforces anatomical invariance.

3.3. Anatomical Semantic Alignment

To further borrow the capability of semantics abstraction acquired from self-distillation, we now consider optimizing the intra-volume relationship by maximizing the similarity between views from the same volume. Different from recent works [25, 56] that use embeddings of intact views (Y_w^Q and Y_v^K) as teachers to supervise the masked views' representations (V_u^Q and V_r^K), we argue that it is unreasonable to directly encourage the masked views to have similar representations to the global views since their anatomical structure and semantic information may be significantly distinct.This dilemma motivates us to perform anatomical semantic alignment[†] between output features from the masked view and the intact view.

One straightforward way is to contrast between the output features of the online decoder and target encoder as proposed in SIM [41]. However, we empirically observe that this choice brings little improvement for downstream tasks. Directly using the features from the online decoder for self-distillation means that the online decoder has to simultaneously optimize multiple different targets. Such multi-task learning process trained a strong online decoder and thus creates an oversimple optimizing task for the online encoder since the decoder takes the most charge of recovering details and anatomical semantics. To ensure that the online encoder effectively contributes to downstream tasks, it is crucial to develop an improved training strategy that enables the online encoder to learn rich and generalized feature represen-

---

[†]Please note that anatomical semantic alignment does not involve a registration task. Our primary objective is to extract aligned features that demonstrate the highest correlation from masked views and augmented views. We do not aim to match masked images to full images in this process.
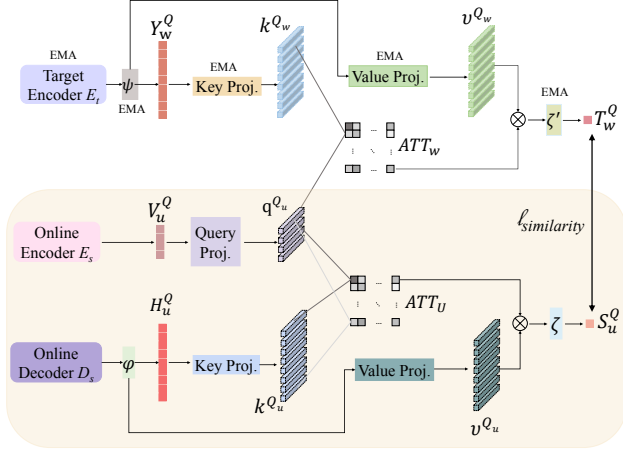
Figure 3. Diagram of the alignment process for volume crop Q in CASA. We generate a semantic-aligned feature pair of student embeddings S_u^Q and teacher embeddings T_w^Q. The processes that produce these two embeddings are symmetric. We adopt a query projection for V_u^Q, key and value projection for both H_u^Q and Y_w^Q. We compute the dot product attention between query and key matrices to allow values to match the most relevant high-level semantics given the local patch texture and topology from masked view's feature embeddings (V_u^Q).

tations. Taking this into account, we propose a conditional anatomical semantic alignment (CASA) module with learning capacity of reasoning about the encoded masked view's most semantically similar anatomical features in the embeddings of the original volume. This leads to crafting more specific and aligned high-level features for self-distillation.

We take the process of aligning features from X_u^Q (masked view from volume crop Q) and X_w^Q (augmented intact view from volume crop Q) as an example to elaborate. Fig. 2 shows the high-level idea of the anatomical semantic alignment. We use V_u^Q from the online encoder as a criterion query, and generate a contrastive pair from outputs of online decoder and target encoder with aligned anatomical semantics constrained by the criterion query. Concretely, we generate teacher embeddings T_w^Q and student embeddings S_u^Q, which are aligned and capture the most semantically similar anatomical information in the original volume as expressed in the masked view's embeddings V_u^Q. Here, the anatomical features from V_u^Q serve as the guidance for semantic alignment.

Fig. 3 shows the process of generating the student embeddings S_u^Q (in yellow part) and teacher embeddings T_w^Q. Consider generating S_u^Q first, we project V_u^Q into the query $q^{Q_u} \in \mathbf{R}^{N_m^Q \times C}$ matric: $q^{Q_u} = LN(V_u^Q)W_q$. Then we project H_u^Q, which encodes the reconstructed features, into key $k^{Q_u} \in \mathbf{R}^{N \times C}$ and value $\nu^{Q_u} \in \mathbf{R}^{N \times C}$ matrices: $k^{Q_u} = LN(H_u^Q)W_k$, $\nu^{Q_u} = LN(H_u^Q)W_\nu$.

Here C is the projection dimension, LN is a LayerNorm

layer and W_q, W_k and W_ν are projection matrices. The query matrices q^{Q_u} are used to seek from the key matrices k^{Q_u} to attend to semantics with the highest relevance. The value matrices ν^{Q_u} represent the anatomical features from which we aggregate only certain high-level global information depending on V_u^Q. Since V_u^Q is extracted from masked views, we aim to adaptively highlights the relevant features from the reconstructed views to selectively obtain anatomical semantics based on their relevance to V_u^Q.

In order to learn flexible conditioning between the online encoder outputs V_u^Q and the reconstructed features H_u^Q, we compute scaled dot-product attention [44] from the query-projected matrices q^{Q_u} to the key-projected matrices k^{Q_u}. The dot product attention gives relevancy weights $ATT_u \in \mathbf{R}^{N_m^Q \times N}$ from local fragmentary patch information to each high-level global anatomical semantics:

$$ATT_u(V_u^Q, H_u^Q) = softmax\left(\frac{q^{Q_u} \cdot k^{Q_u T}}{\sqrt{C}}\right). \quad (3)$$

We then leverage ATT_u to aggregate the value-projected matrices ν^{Q_u} and get the student embeddings through a projection layer ζ as follows:

$$S_u^Q = \zeta(ATT_u \cdot \nu^{Q_u}). \quad (4)$$

The process of generating the teacher embeddings T_w^Q is similar and the mere difference is that we use the target encoder's outputs Y_w^Q to produce key and value matrices. As T_w^Q and S_u^Q are guided by the same query matrices, they learn to encode globally matched anatomical semantics and inter-patch topology information from different views conditioned by the distribution of masked view's local image content. We then maximize the similarity between semantic-aligned teacher embeddings T_w^Q and student embeddings S_u^Q to produce more suitably learned representations.

The process of aligning features from X_r^K (masked view from volume crop K) and X_v^K (augmented intact view from volume crop K) is exactly the same as the above. We also obtain teacher embeddings T_v^K and student embeddings S_r^K from the volume K. Now we have positive feature pairs (S_u^Q, T_w^Q) and (S_r^K, T_v^K). The semantically aligned features in each pair encompass the same high-level anatomical information, we aim to reach a consensus among their representations by maximizing the similarity between them. We therefore define the intra-volume loss as:

$$\ell_{st} = \ell_s(S_u^Q, T_w^Q) + \ell_s(S_r^K, T_v^K), \quad (5)$$

By involving the inter-volume contrast (in Sec. 3.2) and intra-volume contrast (in Sec. 3.3), we explicitly model the anatomical invariance. Combining such advantage with the MIM objective, the online encoder learns to capture

both high-level discriminative anatomical features and fine-grained localization-sensitive context details.

Note that, in Sec. 3.2, we do not adopt semantic alignment between views from different volumes, as they should share the same high-level class information but be distinct in local texture and shape. There is no direct relevance between the masked view from volume crop Q and the intact view from volume crop K.

4. Experiments

4.1. Datasets & Evaluation Metrics

Pre-training Datasets. We use a total of 2,000 unlabeled CT scans from the Fast and Low-resource semi-supervised Abdominal oRgan sEgmentation in CT (FLARE 2022) challenge dataset[‡] to train **Alice** by self-supervised learning. Any (available) forms of annotations or labels of these 2,000 CTs are **not** employed during the pre-training stage.

Downstream Datasets & Evaluation Metrics. Three public datasets are used for downstream evaluation. For segmentation task: (1) In addition to 2,000 unlabeled CT scans, FLARE 2022 also provides a downstream training set including 50 labeled CT scans with pancreatic diseases. The offline test set incorporates 20 CT scans of patients with liver, kidney, spleen, or pancreas diseases. The segmentation targets are 13 organs: liver, spleen, pancreas, right kidney, left kidney, stomach, gallbladder, esophagus, aorta, inferior vena cava, right adrenal gland, left adrenal gland, and duodenum. (2) The Beyond the Cranial Vault (BTCV) abdomen challenge dataset [27] contains 30 subjects of abdominal CT scans where 13 organs (not the same as FLARE 2022) are annotated by interpreters under the supervision of radiologists at Vanderbilt University Medical Center. Following [47], we employ two settings of test set configurations: offline test set and online test set, for BTCV dataset. For classification task: (3) We conduct experiments on a public benchmark MosMedData: Chest CT Scans with COVID-19 Related Findings [31]. This dataset contains a total of 1,110 lung CT scans with COVID-19 related findings, as well as without such findings. We randomly split 70% of the dataset for training, 10% for validation and the rest 20% for testing. Note that all downstream datasets do **not** have any intersection with the dataset used for pre-training. For quantitative evaluation, we adopt two segmentation metrics of Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD), and one classification metric of the area under the receiver operator curve (AUC).

4.2. Implementation Details

Pre-training Setup. We use ViT-B [15] and Swin-B [29] as the default backbones for online encoder. Other than adopting two different random crops for the online encoder and

target encoder as common practice, we utilize SAM [49] to first locate/align the same body part, then we use a default input volume crop size of $192 \times 192 \times 64$ to generate respective views of consistent anatomies. This avoids the large disparity between the inputs of online/target encoders when randomly cropped regions are far apart spatially, or scarcely semantically relevant. For MIM tasks, we apply augmentations and masking strategy following MAE [20] to the input of the online encoder. For the input of the target encoder, strong data augmentations, *e.g.* random resized rotation, flipping, intensity scaling and shifting, are adopted to avoid a trivial solution. We keep the same projection head structure as used in [11, 16]. The training loss is a summation of ℓ_r, ℓ_{dv}, and ℓ_{st}. During pre-training, we adopt a lightweight decoder to reduce computing overhead and only use the online encoder for downstream segmentation tasks. We employ AdamW optimizer [30] with the momentum set to $\beta_1 = 0.9$, $\beta_2 = 0.95$ and cosine learning rate schedule with a warmup of 100 epochs. The pre-training process uses a batch size of 8 per GPU and an initial learning rate of $5e^{-5}$ for 100K iterations. We implement **Alice** model in PyTorch [33]. All pre-training experiments are conducted on 8 NVIDIA A100 GPUs.

Downstream Training Setup. For segmentation tasks, we apply our pre-trained encoder weights to various ViT-based segmentation networks designed for medical tasks of UN-ETR [19], nnFormer [53], and Swin UNETR [40], by following their settings. We compare different SSL methods within [19, 40, 53]. Five-fold cross validation is used to train and evaluate models for all FLARE 2022 and BTCV experiments. For classification task, we utilizes $2,000$ unlabeled CT scans from FLARE 2022 for pre-training to evaluate the adaptability of Alice for different scenes (COVID-19 classification). Ten-fold cross validation is used for COVID-19 dataset. Detailed training hyperparameters for these downstream tasks can be found in the supplemental material.

4.3. Results

FLARE 2022 Abdominal Organ Segmentation. We conduct extensive comparison between **Alice** and existing SOTA methods. We first fix three ViT-based medical segmentation framework, and compare **Alice** with the random initialization (Rand. init.) and other advanced SSL methods designed for computer vision and medical tasks. The evaluation results on the offline test set are shown in Table 1. Compared with strong Transformer-based contrastive learning methods MoCo v3 and DINO, **Alice** outperforms them by at least absolute 3.74% in DSC. Compared with MIM methods MAE and SemMAE, the DSC score of Alice surpasses that of MAE and SemMAE by a large margin. Furthermore, **Alice** also achieves at least 2.22% improvement of DSC relative to hybrid SSL methods IBOT and CMAE,

Backbone	UNETR		Swin UNETR		nnFormer	
Method	DSC	NSD	DSC	NSD	DSC	NSD
Rand. init.	80.95±3.48	85.23±3.74	81.04±3.29	85.60±3.55	81.33±3.05	86.05±3.31
MoCo v3 [11]	82.01±3.27	86.49±3.52	82.63±3.18	86.92±3.44	82.88±2.82	86.89±3.10
DINO [3]	82.07±3.15	86.31±3.40	82.69±2.99	87.08±3.17	82.95±2.74	87.45±2.92
IBOT [56]	83.15±3.21	87.61±3.48	83.77±3.14	88.34±3.30	84.04±2.81	88.46±3.11
SIM [41]	83.04±2.97	87.37±3.36	83.59±2.74	88.57±2.90	83.96±2.64	88.61±2.83
MAE [20]	83.09±2.92	87.42±3.43	83.62±2.66	88.55±2.92	84.01±2.59	88.63±2.75
SemMAE [28]	83.13±2.87	87.90±3.31	-	-	-	-
CMAE [23]	83.59±2.76	88.25±2.98	84.25±2.59	89.17±2.88	84.47±2.42	89.44±2.65
medical MAE [57]	83.11±2.91	87.45±3.39	83.66±2.64	88.57±2.86	84.03±2.57	88.67±2.71
Tang *et al.* [40]	82.97±3.22	87.51±3.50	83.14±3.01	88.74±3.37	83.59±2.89	89.51±3.23
Alice	**85.81±2.05**	**90.03±2.28**	**86.75±1.89**	**91.22±2.12**	**86.87±1.84**	**91.28±2.09**

Table 1. Average DSC and NSD of 13 organs obtained using different ViT-based SSL strategies on the FLARE 2022 offline test set. We fix the adopted segmentation baselines as UNETR, nnFormer, and Swin UNETR. "-" means there is no direct adaptation on the corresponding Swin Transformer based backbones.

which combine contrastive learning and MIM. Notably, Alice is also superior to SOTA ViT based SSL methods, medical MAE [57] and pre-training framework proposed in [40], which are tailored for medical image analysis.

Method	Pre-trained Parts	DSC	NSD
MoCo v2 [9]	3D ResNet E.	81.96±3.29	86.58±3.45
BYOL [16]	3D ResNet E.	81.94±3.32	86.60±3.50
ContrastiveCrop [34]	3D ResNet E.	82.55±3.05	87.26±3.15
LoGo [52]	3D ResNet E.	82.53±3.08	87.08±3.17
PCRL [55]	3D UNet E.&D.	83.41±2.81	87.84±3.00
PCRLv2 [54]	3D nsUNet E.	84.45±2.49	89.11±2.68
PGL [46]	3D ResNet E.	83.16±2.85	87.93±2.99
Chaitanya *et al.* [4]	UNet E.&D.	82.29±2.90	86.72±3.14
nnU-Net [24]	-	83.62±2.78	88.51±2.85
TransVW [18]	3D UNet E.&D.	82.70±3.02	87.26±3.18
SAM [49]	3D UNet E.&D.	82.48±3.06	87.06±3.21
Models Gen. [59]	V-Net E.&D.	83.35±2.82	88.12±2.95
DiRA [17]	3D UNet E.&D.	82.57±2.88	87.23±3.01
Alice	nnFormer E.	**86.87±1.84**	**91.28±2.09**

Table 2. Average DSC and NSD of 13 organs obtained using different CNN-based SSL strategies and models on the FLARE 2022 offline test set. "E." means only encoder is pre-trained while "E.&D." means both encoder and decoder are pre-trained.

We also compare Alice with strong CNN-based SSL methods and baselines. The results are shown in Table 2. Alice significantly outperforms the other CNN-based SOTA methods. **Alice** achieves better performance compared with PCRL, TransVW, SAM, DiRA, and Models genesis, which additionally pre-train a decoder. Notably, **Alice** outperforms the method of Chaitanya *et al.* [4] which also utilizes inter-volume slices. Chaitanya *et al.* [4] assume that all volume images have been aligned and use simple grouping

Method	Ensemble	Offline Test Set	Online Test Set
MoCo v2 [9]	1	82.05±2.82	-
MoCo v3 [11]	1	82.02±2.77	-
DINO [3]	1	82.61±1.79	-
MAE [20]	1	83.16±2.14	-
IBOT [56]	1	83.28±2.26	-
CMAE [23]	1	83.47±1.33	-
PCRL [55]	1	82.73±2.42	-
PCRLv2 [54]	1	83.55±1.49	-
PGL [46]	1	82.57±2.60	-
Chaitanya *et al.* [4]	1	82.74±2.12	-
medical MAE [57]	1	83.23±2.05	-
SAM [49]	1	82.00±3.01	84.07
DoDnet [51]	5	-	86.44
UNETR [19]	5	-	85.55
PaNN [58]	5	-	85.00
nnU-Net [24]	10	-	87.62
nnFormer [53]	5	82.88±2.59	-
DiRA [17]	1	83.14±2.04	-
Tang *et al.* [40]	1	82.58±2.20	84.72
UniMiSS [47]	1	84.99±1.57	87.05
Alice	1	**86.76±0.98**	**88.58**

Table 3. Average DSC of 13 organs on the BTCV offline and online test set. Most results of online test set are directly obtained from BTCV test leaderboard. Result 84.72 of Tang *et al.* is drawn from their paper. Results of MoCo v2, MoCo v3, DINO, PCRL, PGL and UniMiSS are drawn from the original paper of UniMiSS. The segmentation backbone for **Alice** is nnFormer.

of image slices in 3D volumes to craft contrastive patches, which will lead to unaligned pairs. Differently, we adopt a learning-based module SAM [49] to get aligned sample pairs from different volumes online and we further align feature level semantics by CASA module. Our Alice is

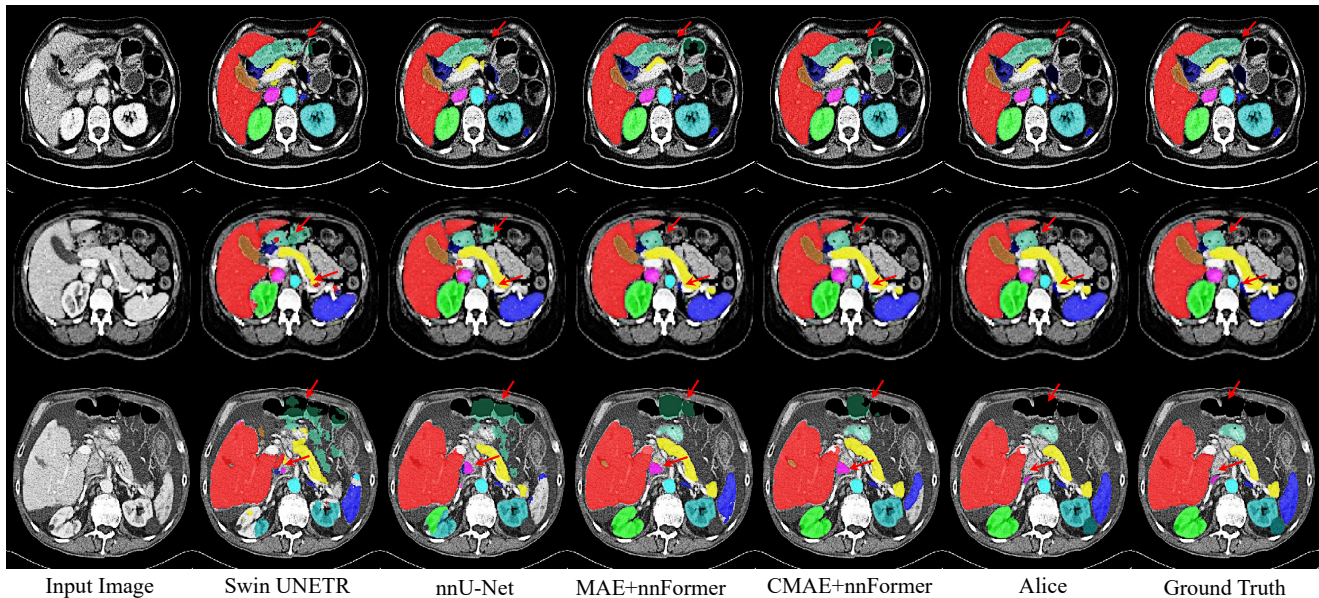| Input Image | Swin UNETR | nnU-Net | MAE+nnFormer | CMAE+nnFormer | Alice | Ground Truth |

Figure 4. Qualitative visualizations on FLARE 2022 offline test set. We compare Alice with other advanced SSL methods MAE, CMAE, self-supervised Swin UNETR, and strong CNN baseline nnU-Net. We present the visualizations of BTCV in the supplemental material.

CNN-based Method	AUC on COVID-19	ViT-based Method	AUC on COVID-19
BYOL [16]	85.74±5.04	DINO [3]	86.87±4.35
Peng *et al.* [34]	87.02±3.11	IBOT [56]	87.55±3.63
LoGo [52]	86.95±3.59	SIM [41]	87.67±2.95
PCRL [55]	87.31±2.88	MAE [20]	86.62±3.27
PCRLv2 [54]	88.36±2.51	SemMAE [28]	86.94±3.44
PGL [46]	86.08±4.72	CMAE [23]	87.73±3.02
DiRA [17]	87.43±3.55	**Alice**	**90.88±1.29**

Table 4. Classification performance of using different pre-training strategies on the COVID-19 screening test set. CNN-based SSL methods take the 3D ResNet as their encoder backbone. ViT-based SSL methods take the ViT-B as their encoder backbone.

more suitable to handle real-world complex problems like changes in the range of CT scans and temporal changes.

Fig. 4 shows the qualitative results and demonstrate the merits of **Alice**. Most competing methods suffer from segmentation target incompleteness related failures and misclassification of background regions as organs (false positives). **Alice** produces sharper boundaries and generates results that are more consistent with the ground truth in comparison with all other models. This success is attributed to the advantage of leveraging the intrinsic anatomical invariance across varying volumes and semantic alignment between contrastive views, which help the learned representation robust to organ deformation and pathological changes.

BTCV Multi-organ Segmentation. We also compare **Alice** with other SOTA SSL methods on the BTCV offline and online test set. As shown in Table 3, **Alice**, without using any ensemble strategy, still achieves the competitive performance with the best DSC on both offline and online test set. Note that compared to SOTA methods designed for medical images, namely UniMiSS and the self-supervised Swin UN-

ETR (Tang *et al.*), using over $5,000$ 3D CT scans for pre-training, our **Alice** outperforms them using only 40% of the data. This effectiveness can be explained that our approach is capable of learning representations that is robust to the size, shape, intensity, and texture diversity of body parts by modeling the anatomical invariance. During pre-training, we map the high-level embeddings of the same organs from varying volumes to the same point, which helps the model in downstream tasks reduce misclassification failures and noticeably improves the performance.

COVID-19 Classification. We compare **Alice** with the state-of-the-arts including representative CNN-based SSL methods and ViT-based SSL methods. The results are shown in Table 4. Compared with CNN-based SSL methods BYOL, PCRL, PCRLv2, PGL and DiRA, **Alice** outperforms them at least absolute 2.52% in AUC. Notably, **Alice** achieves much better results than LoGo [52] and Peng *et al.* [34], which also design specific strategies to generate semantic-aligned contrastive view pairs. However, these two methods only operate within each image independently and ignore the inter-volume consistency. Compared against strong ViT-based SSL methods, **Alice** significantly outperforms them at least absolute 3.15% in AUC. It proves the effectiveness of modeling anatomical invariance and performing semantic alignment to assist the SSL process. Moreover, as an inter-scene evaluation, Table 4 reveals **Alice** has the ability to generalize to other scenarios. More details are in our Supplementary Material.

Method	Inter-Volume Relationship ℓ_{dv}	FLARE 2022	
		DSC	NSD
IBOT [56]	$\times$	84.04±2.81	88.46±3.11
	$\checkmark$	85.41±2.42	89.97±2.70
Alice	$\times$	85.63±2.11	90.01±2.32
	$\checkmark$	**86.87±1.84**	**91.28±2.09**

Table 5. Ablation study of modeling the inter-volume anatomical invariance. The segmentation backbone is nnFormer. "$\times$" means inputing views from the same volume and not using ℓ_{dv}, while "$\checkmark$" means inputing views from different volumes and using ℓ_{dv}.

Method	FLARE 2022	
	DSC	NSD
IBOT (*w/.* vs. *w/o.*)	85.34±2.63 vs. 84.04±2.81	89.88±2.89 vs. 88.46±3.11
SIM (*w/.* vs. *w/o.*)	85.21±2.36 vs. 83.96±2.64	89.89±2.67 vs. 88.61±2.83
Alice (*w/.* vs. *w/o.*)	**86.87±1.84** vs. 85.66±2.18	**91.28±2.09** vs. 90.02±2.32

Table 6. Ablation study of the CASA on FLARE 2022 benchmark. We apply the CASA on different siamese SSL architectures which take masked view and unmasked view as inputs. "*w/.* vs. *w/o.*" denotes the comparison between using CASA or without using CASA. The segmentation backbone for **Alice** is nnFormer.

4.4. Ablation Study

Significance in Modeling Inter-volume Anatomical Invariance. We investigate the effect of leveraging the intrinsic anatomical structures of different volumes to model the anatomical invariance. We feed the same body part of query volume and key volume to another SSL method IBOT and add the inter-volume relationship contrastive loss ℓ_{dv} for it. Table 5 shows that involving feature contrasting between mined consistent views from varying aligned CT volumes can substantially improve the 3D segmentation accuracy in the downstream task for both IBOT and **Alice**. The performance gain on DSC is 1.37% and 1.24%, respectively. This shows that modeling the anatomical invariance among CT volumes to learn intrinsic high-level semantics during pretraining benefits the learned representations for downstream tasks.

Effectiveness of Anatomical Semantic Alignment. Table 6 shows the ablation study on the effectiveness of the CASA module. Our approach can be applied to most general siamese architecture based SSL methods. All methods achieve more than 1.21% absolute DSC gain on the FLARE 2022 dataset by adopting CASA. This shows applying our anatomical semantic alignment to harvest better contrastive pairs from masked and unmasked views yields significant improvements on various siamese SSL methods.

The impact of whether to use negative samples. In Eq. 2, we utilize BYOL-style cosine loss [16] as our default choice for contrastive learning. This loss only maximizes the similarity between positive views and eliminates the use of negative pairs. Another widely used similarity loss function is

Method	loss function ℓ_s	FLARE 2022	
		DSC	NSD
Alice	InfoNCE loss [7, 21, 32]	86.83±1.88	91.20±2.12
	BYOL-style cosine loss [16]	**86.87±1.84**	**91.28±2.09**

Table 7. Ablation study on whether to use negative samples. The segmentation backbone is nnFormer. InfoNCE loss seeks to simultaneously pull close positive views and push away negative samples while BYOL-style cosine loss only maximizes the similarity between positive views and eliminates the use of negative pairs.

the InfoNCE loss [7, 21, 32], which aims to simultaneously pull close positive views and push away negative samples. The key distinction between these two widely used types of similarity loss functions lies in the utilization of negative samples. **Alice** is compatible with a wide range of SSL techniques and is independent of the specific training losses used in those techniques. Thus we conduct experiments to investigate the influence of whether negative pairs are used. We discuss two widely used similarity loss functions: InfoNCE loss (exploits both positive and negative samples) and BYOL-style cosine loss (does not exploit negative samples). When using InfoNCE loss for contrastive learning, we utilize views from crops of different body parts in the same batch to compose negative pairs. Table 7 shows the ablation study on whether to use negative samples. We observed that using cosine loss in **Alice** achieves slightly higher performance than InfoNCE loss on FLARE 2022 test set. Thus, we do not utilize negative samples and use cosine loss as the default for the contrastive learning branch in **Alice**. The discussion of why cosine loss does not need negative pairs can be found in BYOL's literature [16].

5. Conclusions

In this work, we propose a novel self-supervised learning method (**Alice**) for improving the learned image representation of contrastive learning and MIM by modeling the class-specific invariance of intrinsic anatomical semantics in 3D medical images. We also introduce a conditional anatomical semantic alignment module that generates better contrastive pairs with more consistent high-level information. Extensive quantitative experiments reporting superior results validate the effectiveness of our method.

References

[1] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *ICCV*, pages 3478–3488, 2021. 3

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 1, 2, 4, 7, 8

[4] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *NeurIPS*, 33:12546–12558, 2020. 3, 7

[5] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 2019. 3

[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, pages 1691–1703. PMLR, 2020. 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 1, 2, 3, 9

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *NeurIPS*, 33:22243–22255, 2020. 1

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 7

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 4

[11] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021. 1, 2, 6, 7

[12] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *ECCV*, pages 108–124. Springer, 2022. 2

[13] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *NeurIPS*, 32, 2019. 3

[14] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1, 3, 6

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020. 1, 2, 4, 6, 7, 8, 9

[17] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *CVPR*, pages 20824–20834, 2022. 3, 7, 8

[18] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Trans. Medical Imaging*, 40(10):2857–2868, 2021. 7

[19] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *WACV*, pages 574–584, 2022. 2, 6, 7

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 1, 3, 6, 7, 8

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1, 2, 9

[22] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. Green hierarchical vision transformer for masked image modeling. *arXiv preprint arXiv:2205.13515*, 2022. 3

[23] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *arXiv preprint arXiv:2207.13532*, 2022. 1, 2, 3, 4, 7, 8

[24] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 7

[25] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. *arXiv preprint arXiv:2203.12719*, 2022. 3, 4

[26] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 1, 3

[27] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015. 2, 6

[28] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *arXiv preprint arXiv:2206.10207*, 2022. 3, 7, 8

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1, 3, 6

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[31] Sergey P Morozov, AE Andreychenko, NA Pavlov, AV Vladzymyrskyy, NV Ledikhova, VA Gombolevskiy, Ivan A Blokhin, PB Gelezhe, AV Gonchar, and V Yu Chernina.

Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020. 2, 6

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 9

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 6

[34] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *CVPR*, pages 16031–16040, 2022. 3, 7, 8

[35] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S Ryoo. Self-supervised video transformer. In *CVPR*, pages 2874–2884, 2022. 1

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1

[37] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *CVPR*, pages 11058–11067, 2021. 3

[38] Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *ICML*, pages 20026–20040. PMLR, 2022. 3

[39] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *NeurIPS*, 33:18158–18172, 2020. 3

[40] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *CVPR*, pages 20730–20740, 2022. 1, 2, 3, 4, 6, 7

[41] Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022. 1, 2, 3, 4, 7, 8

[42] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *NeurIPS*, 33:6827–6839, 2020. 1

[43] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017. 3

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 5

[45] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, pages 14668–14678, 2022. 3

[46] Yutong Xie, Jianpeng Zhang, Zehui Liao, Yong Xia, and Chunhua Shen. Pgl: prior-guided local self-supervised learning for 3d medical image segmentation. *arXiv preprint arXiv:2011.12640*, 2020. 7, 8

[47] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *ECCV*, pages 558–575. Springer, 2022. 1, 3, 6, 7

[48] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022. 1, 3

[49] Ke Yan, Jinzheng Cai, Dakai Jin, Shun Miao, Dazhou Guo, Adam P Harrison, Youbao Tang, Jing Xiao, Jingjing Lu, and Le Lu. Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. *IEEE Trans. Medical Imaging*, 2022. 3, 4, 6, 7

[50] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *CVPR*, pages 8354–8363, 2022. 1

[51] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *CVPR*, pages 1195–1204, 2021. 7

[52] Tong Zhang, Congpei Qiu, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Leverage your local and global representations: A new self-supervised learning strategy. In *CVPR*, pages 16580–16589, 2022. 3, 7, 8

[53] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021. 2, 6, 7

[54] Hong-Yu Zhou, Chixiang Lu, Chaoqi Chen, Sibei Yang, and Yizhou Yu. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 7, 8

[55] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *ICCV*, pages 3499–3509, 2021. 3, 7, 8

[56] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 1, 2, 3, 4, 7, 8, 9

[57] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image analysis. *arXiv preprint arXiv:2203.05573*, 2022. 7

[58] Yuyin Zhou, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, and Alan L Yuille. Prior-aware neural network for partially-supervised multi-organ segmentation. In *ICCV*, pages 10672–10681, 2019. 7

[59] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021. 3, 7

[60] Jiuwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S Kevin Zhou, and Yefeng Zheng. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical Image Analysis*, 64:101746, 2020. 3

Supplementary Materials for
Anatomical Invariance Modeling and Semantic Alignment for Self-supervised Learning in 3D Medical Image Analysis

Yankai Jiang[1,3*], Mingze Sun[1,4*], Heng Guo[1,2], Xiaoyu Bai[1], Ke Yan[1,2], Le Lu[1] and Minfeng Xu[1,2✉]

[1]DAMO Academy, Alibaba Group
[2]Hupan Lab
[3]College of Computer Science and Technology, Zhejiang University
[4]Tsinghua Shenzhen International Graduate School, Tsinghua-Berkeley Shenzhen Institute, China
✉eric.xmf@alibaba-inc.com

A. Implementation Details

A.1. Preprocessing pipeline for pre-training dataset

The FLARE 2022 dataset is collected from more than 20 medical groups under the license permission, including MSD [16], KiTS [9], AbdomenCT-1K [13], and TCIA [5]. It provides a training set including 2000 unlabelled CT scans with liver, kidney, spleen, or pancreas diseases. We split 10% of the unlabelled CT scans for validation in the pre-training stage, and thus the number of training and validation volumes are 1800 and 200, respectively. **Alice** is pre-trained using only unlabelled images (any annotations were not utilized). First, we clip the CT image intensities from -125 to 255, and then normalize them to the range of 0 to 1. We adopt SAM [20] to locate aligned body parts. The results of landmarks on query and key volumes aligned by SAM are shown in Fig. 5 and Fig. 6. We use a default input volume crop size of $192 \times 192 \times 64$ to generate respective views of consistent anatomies according to the aligned landmarks on each query and key volume pair. In this way, **Alice** is pre-trained via a diverse set of human body compositions, and learns a general-purpose representation from different medical groups' data that can be leveraged for a wide range of downstream tasks.

A.2. End-to-end fine-tuning settings for downstream datasets

We apply our pre-trained online encoder weights to various ViT-based segmentation networks designed for medical tasks of UNETR, nnFormer, and Swin UNETR, by follow-

ing most of their settings. The detail settings are shown in Tab. 8.

B. Results on Downstream Tasks

In this section, we show more results on 3D classification downstream task.

B.1. Dataset

We conduct experiments on a public benchmark MosMedData: Chest CT Scans with COVID-19 Related Findings [14]. This dataset consists of lung CT scans with COVID-19 related findings, as well as without such findings. We use the associated radiological findings of the CT scans as labels and formulate this task as a 2 classes classification to predict presence of viral pneumonia. This dataset contains a total of 1110 CTs. We randomly split 70% of the dataset for training, 10% for validation and the rest 20% for testing. We adopt the ten-fold cross-validation method.

B.2. Preprocessing pipeline

We first rotate the CT volumes by 90 degrees to fix the orientation. We adopt a threshold between -1000 and 400 to clip the CT intensities, and then normalize the Hounsfield units (HU) values to be between 0 and 1. All volumes are resized to $128 \times 128 \times 64$. We use the online data augmentation, including random rotation, scaling, flipping, adding white Gaussian noise, Gaussian blurring, adjusting rightness and contrast, simulation of low resolution.

B.3. Setup

To perform classification, we extract the pretrained online encoder and appended a FC layer with the output chan-

Config	FLARE 2022			BTCV
	UNETR	Swin UNETR	nnFormer	nnFormer
optimizer	AdamW	AdamW	SGD	SGD
base learning rate	$1e^{-4}$	$4e^{-4}$	0.01	0.01
weight decay	$1e^{-5}$	$1e^{-5}$	$3e^{-5}$	$3e^{-5}$
optimizer momentum	0.9	0.9	0.99	0.99
batch size	8	8	8	8
learning rate schedule	cosine decay	cosine decay	"poly" decay	"poly" decay
warmup epochs	50	50	40	40
training epochs	1000	1000	1000	1000
augmentation	random flip, rotation, intensities shifting	random flip, rotation, intensities shifting	scaling, gaussian blur, mirroring	scaling, gaussian blur, mirroring
Spacing	$0.76 \times 0.76 \times 1.5$	$0.76 \times 0.76 \times 1.5$	$0.76 \times 0.76 \times 1.5$	$0.76 \times 0.76 \times 2$
Crop size	$96 \times 96 \times 96$	$96 \times 96 \times 96$	$128 \times 128 \times 96$	$128 \times 128 \times 96$

Table 8. End-to-end fine-tuning settings for FLARE 2022 and BTCV datasets.

Method	Backbone	COVID-19		
		20%	50%	100%
Rand. init.	3D ResNet	73.55±9.33	76.64±7.11	84.73±5.13
MoCo v2 [3]		76.73±9.16	77.94±7.03	85.86±4.92
BYOL [6]		76.69±9.20	77.88±7.15	85.74±5.04
ContrastiveCrop [15]		78.65±8.36	80.61±6.28	87.02±3.11
LoGo [21]		78.60±8.82	80.53±6.75	86.95±3.59
PCRL [23]		79.44±8.44	81.17±6.21	87.31±2.88
PGL [19]		76.77±9.09	78.02±6.93	86.08±4.72
DiRA [7]		78.06±9.04	79.15±6.86	87.43±3.55
Rand. init.	ViT-B	72.80±9.25	76.76±6.90	85.05±4.94
MoCo v3 [4]		77.62±9.17	78.91±6.62	86.32±4.66
DINO [2]		78.49±8.77	80.33±6.28	86.87±4.35
IBOT [24]		79.53±8.05	81.42±5.53	87.55±3.63
SIM [18]		79.85±7.87	81.60±5.05	87.67±2.95
MAE [8]		78.25±8.02	79.78±6.84	86.62±3.27
SemMAE [11]		78.57±7.60	80.47±5.67	86.94±3.44
CMAE [10]		80.05±7.08	81.65±4.92	87.73±3.02
Tang et al. [17]		79.59±7.59	81.52±5.05	87.70±3.07
Alice		**83.30±6.04**	**85.23±3.91**	**90.88±1.29**

Table 9. Classification performance of using different pre-training strategies on the COVID19 screening test set. CNN-based SSL methods take the 3D ResNet as their encoder backbone. ViT-based SSL methods take the ViT-B as their encoder backbone. We adopt three label settings (using 20%, 50%, and 100% labeled training data).

nel as the number of classes for prediction. We train the classification model using the AdamW optimizer with a warm-up cosine scheduler of 400 iterations. We use a batch-size of 8 per GPU, an initial learning rate of $5e^{-5}$, a momentum of 0.9 and a decay of $1e^{-5}$ for 10K iterations. We utilize the cross entropy loss. The classification performance is measured by the area under the receiver operator curve (AUC).

B.4. Results

We compare **Alice** with the state-of-the-arts including representative CNN-based SSL methods and ViT-based SSL methods. The results are shown in Tab. 9. **Alice** noticeably surpasses the other state-of-the-art SSL frameworks. We show when using 20% of labeled training data, **Alice** achieves approximately 11% improvement comparing to training from scratch. When employing all labeled training data, the self-supervised pre-training shows 5.83% higher AUC. In practice, the AUC number 85.05 of learning from scratch with entire dataset can be achieved by us-

ing pre-trained weights from **Alice** with 50% training data, which indicates that **Alice** can reduce the annotation effort by at least 50% for this task.

Compared with the state-of-the-art CNN-based SSL methods MoCov2, BYOL, PCRL, PGL and DiRA, **Alice** outperforms them at least absolute 3.86% and 3.45% in AUC when using 20% and 100% labeled training data, respectively. Notably, **Alice** achieves much better results than LoGo and ContrastiveCrop, which also design specific strategies to generate semantic-aligned contrastive view pairs. However, these two methods only operate within each image independently and ignore the inter-volume consistency. The superiority performance of **Alice** also reveals the effectiveness of our anatomical semantic alignment strategy.

Compared against strong ViT-based SSL methods, **Alice** significantly outperforms them on all three label settings. The performance gains over the second, third and fourth top-ranked methods are 3.15%, 3.18%, 3.21% and 3.33% on AUC when 100% labels are available. It proves the effectiveness of modeling anatomical invariance and performing semantic alignment to assist the SSL process. Besides, we find that contrastive learning tends to benefit classification task more than MIM, which is consist with many previous studies [10, 18, 12]. Contrastive learning naturally endows the pretained model with strong instance discriminability, while MIM focuses more on learning local relations in input image for fulfilling the reconstruction task [12]. We also notice that the ViT-based methods tend to outperform the CNN-based methods when the number of training data scales up. It reflects that the Vision Transformer is a competitive architecture and the SSL is vital for it to achieve good performance.

C. Ablation Studies

We have conducted additional ablation experiments to further validate the design choices we have made in **Alice**.

C.1. Masking for the target encoder

We investigate whether adopting random masking for the target encoder affects the model performance. As shown

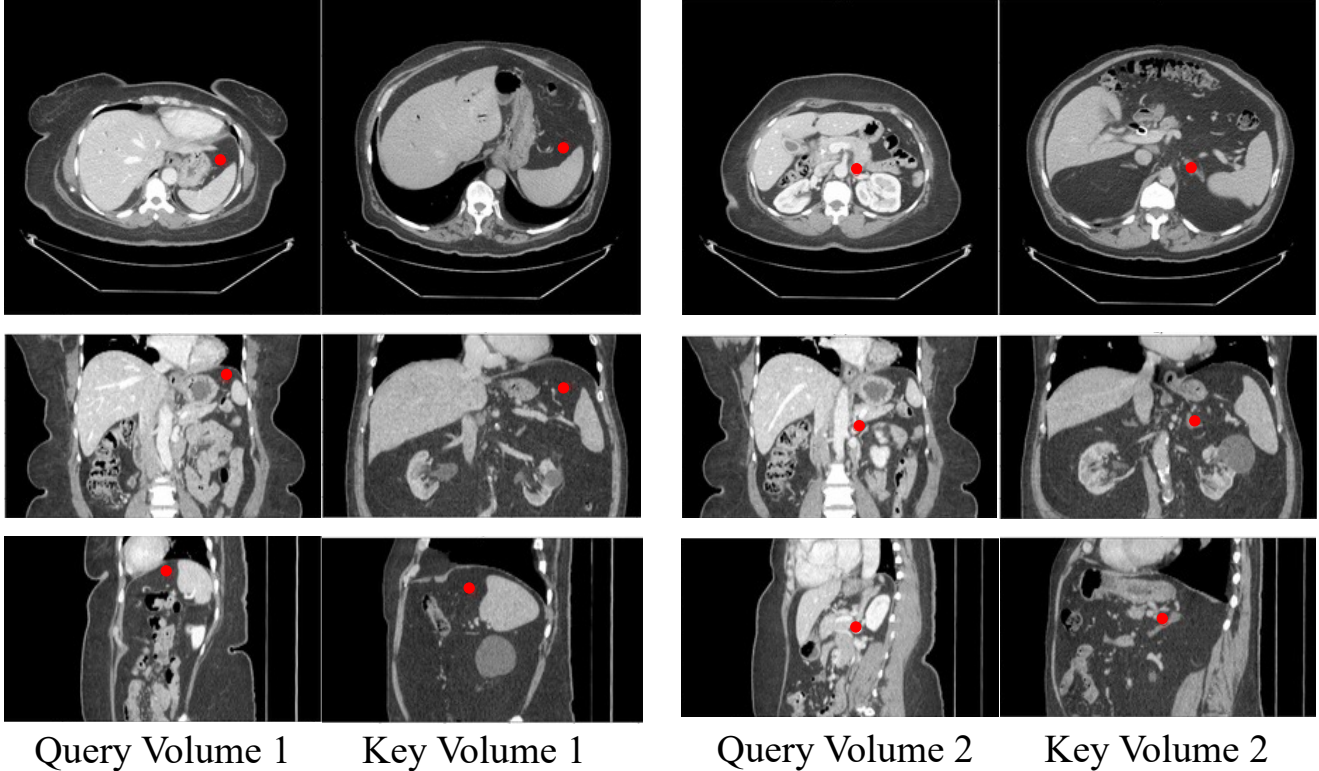| Query Volume 1 | Key Volume 1 | Query Volume 2 | Key Volume 2 |

Figure 5. An example of anatomical location matching via SAM [20]. We randomly select an anatomical point in a query CT image, and then use SAM to find its matched point in a key volume from another patient. The red points are selected points in query volume or detected points in key volume.

Masking ratio	DSC on FLARE 2022	DSC on BTCV (offline)
0.75	84.45±2.63	84.30±1.66
0.5	85.22±2.44	85.06±1.35
0.25	86.01±2.27	85.94±1.27
0	**86.87±1.84**	**86.76±0.98**

Table 10. Experiment on whether adopting masking to the input of target encoder. We test different masking ratio settings on FLARE 2022 dataset and BTCV dataset. The segmentation backbone is nnFormer.

Method	MIM ℓ_r	Inter-Volume ℓ_{dv}	Intra-Volume ℓ_{st}	DSC on FLARE 2022
nnFormer baseline	×	×	×	81.33±3.05
	×	✓	×	83.17±2.82
	✓	✓	×	85.66±2.18
	✓	×	✓	85.63±2.11
Alice	✓	✓	✓	**86.87±1.84**

Table 11. Ablation study of different combinations of self-supervised objectives in **Alice** on FLARE 2022 benchmark. The segmentation backbone is nnFormer.

in Tab. 10, we observe that using the intact views yields the best results on FLARE 2022 dataset and BTCV dataset. The target encoder provides the online encoder with the contrastive supervision. If target encoder also takes random masked input with degenerated semantic information, the anatomical alignment process will be sub-optimal since the teacher embedding from CASA module may hardly access gloabl information from the original volume crop. Thus, the target encoder in **Alice** uses the whole intact views as inputs.

C.2. Efficacy of combining MIM and CL

We perform experiments on pre-training with different combinations of self-supervised objectives to study the ef-

fectiveness of MIM and contrastive learning. Tab. 11 shows the results on FLARE 2022 test set. Overall, employing all objectives achieves best Dice of 86.87%.

D. More explanations on the feature alignment module

The feature alignment module (CASA) is the one contribution of our paper. We have also thought about using traditional image registration methods (e.g. deformable many-one registration) or some unsupervised learning-based medical image registration methods [1, 22] for alignment. However, a large masking ratio would already erase many image contents and make the masked view quite distinct from the intact one. We found existing medical image registration methods can not work well to solve this problem. Driven
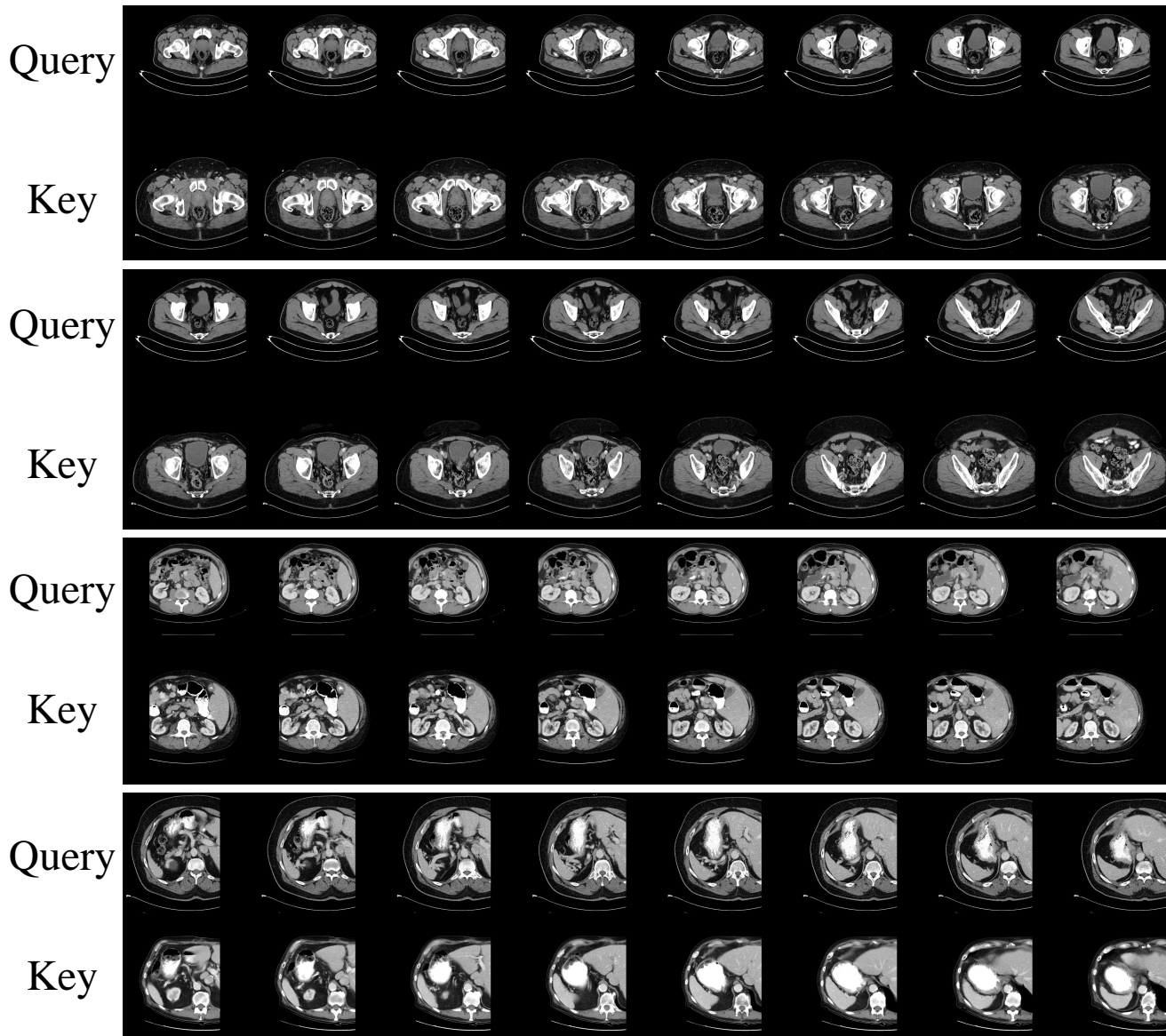
Figure 6. Random anatomy matching results of SAM [20]. We show different query and key crops.

by this limitation, we propose the CASA module to perform alignment in the feature space. "anatomical semantic alignment" is not performing a registration task. So the feature alignment module (CASA) in our **Alice** is not trying to match masked images to full images. It aims to extract aligned features (most correlated) from masked views and augmented views. This module is essentially a self-attention based feature extractor, not a registration algorithm. We compared our module with SIM [18] and CMAE [10] which adopt a specific decoder to generate aligned features. Our method outperforms these two methods in all tasks.

E. BTCV Quantitative Comparisons

In this section, we provide the quantitative comparisons on BTCV offline test set. Note that the ground truth labels of online test set are not accessible. As shown in Fig. 7, **Alice** successfully identifies all organs with high accuracy while it is easy to see that Swin UNETR and nnU-Net produce some under-segmentation and over-segmentation errors. Moreover, as can be seen from the comparison results in the last row of Fig. 7, Swin UNETR and nnU-Net misclassify the spleen (red) as liver (pink) while **Alice** makes the right organ classification. Such superiority of **Alice** owes to the effectiveness of modeling anatomical invariance.

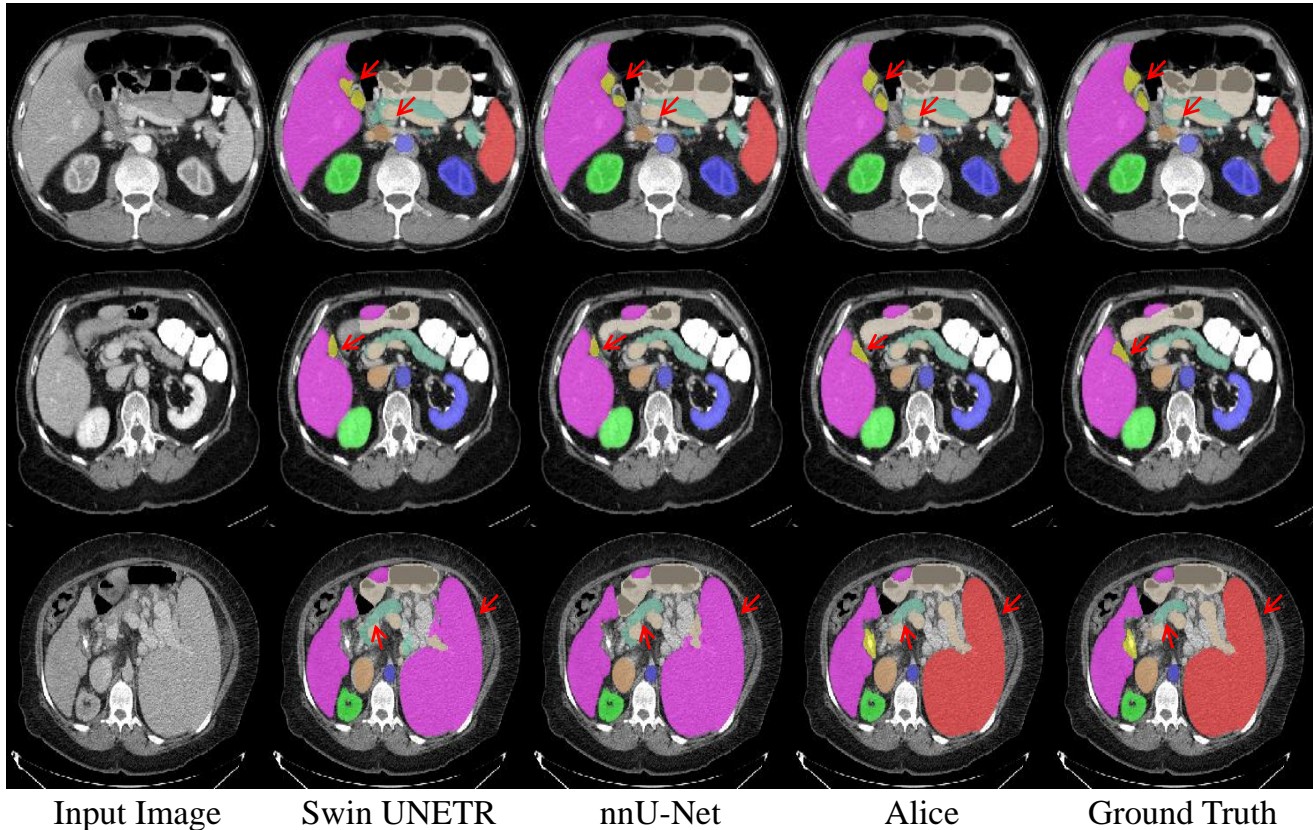| Input Image | Swin UNETR | nnU-Net | Alice | Ground Truth |

Figure 7. Qualitative visualizations on BTCV offline test set. We compare **Alice** with state-off-the-art segmentation methods, namely Swin UNETR and nnU-Net. The segmentation backbone for **Alice** is nnFormer.

References

[1] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018. 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021. 2

[5] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013. 1

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020. 2

[7] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *CVPR*, pages 20824–20834, 2022. 2

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2

[9] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021. 1

[10] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *arXiv preprint arXiv:2207.13532*, 2022. 2, 4

[11] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided

masking for learning masked autoencoders. *arXiv preprint arXiv:2206.10207*, 2022. 2

[12] Siyuan Li, Di Wu, Fang Wu, Zelin Zang, Baigui Sun, Hao Li, Xuansong Xie, Stan Li, et al. Architecture-agnostic masked image modeling–from vit back to cnn. *arXiv preprint arXiv:2205.13943*, 2022. 2

[13] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE TPAMI*, 2021. 1

[14] Sergey P Morozov, AE Andreychenko, NA Pavlov, AV Vladzymyrskyy, NV Ledikhova, VA Gombolevskiy, Ivan A Blokhin, PB Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020. 1

[15] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *CVPR*, pages 16031–16040, 2022. 2

[16] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 1

[17] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *CVPR*, pages 20730–20740, 2022. 2

[18] Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022. 2, 4

[19] Yutong Xie, Jianpeng Zhang, Zehui Liao, Yong Xia, and Chunhua Shen. Pgl: prior-guided local self-supervised learning for 3d medical image segmentation. *arXiv preprint arXiv:2011.12640*, 2020. 2

[20] Ke Yan, Jinzheng Cai, Dakai Jin, Shun Miao, Dazhou Guo, Adam P Harrison, Youbao Tang, Jing Xiao, Jingjing Lu, and Le Lu. Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. *IEEE Trans. Medical Imaging*, 2022. 1, 3, 4

[21] Tong Zhang, Congpei Qiu, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Leverage your local and global representations: A new self-supervised learning strategy. In *CVPR*, pages 16580–16589, 2022. 2

[22] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10600–10610, 2019. 3

[23] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *ICCV*, pages 3499–3509, 2021. 2

[24] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2