Dual-Contrastive Dual-Consistency Dual-Transformer: A Semi-Supervised Approach to Medical Image Segmentation

Ziyang Wang[*]
Department of Computer Science
University of Oxford, UK
ziyang.wang@cs.ox.ac.uk

Congying Ma[*]
Department of Computer Science
University of Bath, UK
congying.ma@bath.edu

Abstract

Medical image segmentation serves as a crucial underpinning for a myriad of clinical applications. The advent of deep learning techniques has significantly propelled advancements in this field. However, challenges persist due to the limited availability of labelled medical imaging data and the substantial cost of data annotation. This paper introduces a novel semi-supervised learning strategy, amalgamating pseudo-labelling and contrastive learning with a consistency regularization framework. This innovative approach incorporates a modified contrastive learning strategy and a confidence-aware pseudo-labeling strategy, both of which are integrated into a dual-segmentation network ensemble learning structure. Inspired by the recent success of self-attention mechanisms, we harness the power of the Vision Transofmer(ViT) within our proposed semi-supervised framework, and conduct a comprehensive comparison among various combinations of ViT and Convolutional Neural Network(CNN) with the proposed strategy. The efficacy of our proposed method is validated using a publicly available medical image segmentation dataset, where it demonstrates state-of-the-art performance against established methods. The proposed method, all baseline methods, and dataset are available at https://github.com/ziyangwang007/CV-SSL-MIS.

1. Introduction

Medical image segmentation is a fundamental facet of computer-aided diagnosis, facilitating precise identification and quantification of anatomical structures and pathological regions within medical imaging. The segmented images gleaned from this process are instrumental in a broad range of clinical applications, including treatment planning, disease monitoring, and surgical navigation. Over the past decade, deep learning techniques, primarily CNN, have profoundly transformed the field, attaining state-of-the-art performance in a myriad of segmentation tasks [6, 24, 27, 33]. Despite these advancements, CNN-based segmentation networks exhibit certain limitations in modelling global context information due to their local receptive fields and inherent spatial invariance, as outlined in recent studies [6,22,23, 38, 42]. As a result, contemporary research has shifted its focus towards self-attention-based approaches [12, 54, 56], examining the potential of ViT to model long-range dependencies for medical image segmentation [3, 5, 23, 51].

Despite advancements in segmentation network, applying deep learning methods to medical image segmentation is hindered by the limited availability of labelled medical imaging data. Obtaining such data is labour-intensive and time-consuming, often requiring manual annotation by expert radiologists [1]. This constraint has driven research towards developing semi-supervised learning methods, which can exploit both labelled and unlabelled data to improve segmentation performance [18, 20, 21, 31, 37], among which series of consistency-based concepts have been proposed. Aimed at further improving performance under different data situations, other studies proposed self-supervised learning [4, 15] and multi-task learning [7, 28], which can further improve feature learning performance. The incorporation of consistency learning and contrastive learning have been tackled in recent studies, employing a basic contrastive learning framework or taking contrastive learning as a pre-trainning strategy [19]. Deploying consistency learning and contrastive learning simultaneously with corresponding backbone networks for image segmentation with semi-supervised fashion, however, should be further explored.

In this paper, we introduce a novel approach to semi-supervised medical image segmentation: the **D**ual-contrastive **D**ual-consistency regularization ensemble **D**ual-Vision Transformer, named 3D-ViT. This method amalgamates the strengths of consistency and contrastive learning to optimize the utilization of limited annotations. To de-

[*] Ziyang Wang and Congying Ma are joint first authors, and contribute equally to this work.

velop a contrastive learning strategy within a consistency learning framework, the primary goal is to encourage the consistency between sub-networks to the large extent. The 3D-ViT incorporates a confidence-aware pseudo-labelling strategy through and a robust contrastive learning strategy, both of which are based on domain-adaptive data augmentations. The key contributions of this paper are as follows:

1. **Dual-Consistency Learning:** The introduced dual-consistency learning reinforces prediction uniformity across sub-networks and perturbed input images. For a best compatibility with the contrastive learning, ensembled pseudo labels, generated via weakly augmented images and masked with a confidence threshold, are utilized for unlabelled data.

2. **Dual-Contrastive Learning:** We propose an enhanced contrastive learning strategy via two pairs of projectors, seamlessly integrated into semi-supervised frameworks to bolster feature learning performance, which acts on image-level feature maps and exploits both labelled and unlabelled data via comparisons between weak and strong augmentations.

3. **Dual-Transformer:** Building upon the successes of self-attention mechanisms [3], we investigate ViT's potential in segmentation tasks. For a fair and comprehensive comparison, we probe both ViT-based and CNN-based segmentation networks with identical architecture and diverse learning schemes, showcasing that our proposed 3D-ViT surpasses CNN or combination of CNN & ViT with an ablation study. The framework is robust over diverse combinations of sub-networks, surpassing most selected baseline methods.

4. **Performance:** We evaluate the efficacy of the proposed 3D-ViT on a public cardiac segmentation dataset [1]. We compare its performance against various existing methods [10, 31, 36, 37, 39, 40, 44–47, 53, 55], demonstrating that the 3D-ViT outperforms other semi-supervised methods. We conduct our experiments under identical hyper-parameter settings and data conditions, employing multiple evaluation metrics. The code for 3D-ViT and all baseline methods are available for further research and exploration.

2. Related Work

Segmentation Network: Image semantic segmentation has been extensively studied in recent years, with the primary focus on developing deep-learning-based networks for accurate and robust segmentation. The development of CNN-based segmentation networks such as FCN [24], DeepLab [6], U-Net [33], and V-Net [27] have significantly contributed to the success of medical image segmentation. Various advanced techniques to further improve performance, such as 3D CNN [11], Atrous CNN [43], residual connections [13], attention mechanisms [29], and densely connected [16], have also been explored in segmentation networks, achieving state-of-the-art results on CT, MRI, and ultrasound tasks [27, 33, 48]. Recent research has also focused on applying self-attention mechanisms [38], for image processing tasks, such as ViT [12]. ViT-based networks thus have been further explored for dense prediction downstream tasks, such as Swin-Transformer [23], and SegFormer [51]. Within the realm of medical image analysis, most studies have sought to integrate traditional CNN techniques with ViT, resulting in networks like TransUNet [5] and SwinUNet [3,23]. In this paper, we construct segmentation networks based on classical encoder-decoder U-shaped networks, enriched with ViT layers. We further incorporate CNN layers to provide a fair comparison and demonstrate the efficacy of our proposed 3D-ViT.

Contrastive Learning: Contrastive learning has emerged as a powerful tool for learning robust and discriminative feature representations, and it is a form of self-supervised learning [30]. The key idea is to learn an embedding space where similar images are pulled closer together, while dissimilar images are pushed apart. This approach has been successful in various tasks, including self-supervised learning for feature representation [9, 14], and unsupervised domain adaptation [17]. In medical image analysis, contrastive learning has been employed to tackle challenges associated with limited annotations and to enhance feature learning capabilities, thereby improving performance [4, 15, 52]. Researchers have also investigated various strategies for contrastive learning, such as different augmentation techniques, similarity metrics, and negative sample mining, to further enhance the effectiveness of the networks [25, 49]. Our proposed 3D-ViT utilizes two pairs of projectors for two comparisons for contrastive learning on image-level feature information to fully improve feature learning performance.

Semi-Supervised Learning: Semi-supervised learning has gained significant attention as a means to address the scarcity of labelled data in medical imaging applications. Early works in this field include the use of temporal ensembling [20], Mean Teacher [37], uncertainty-aware mean teacher [53], and Deep Co-Training [31]. More recent approaches have leveraged adversarial training [18, 44] and pseudo-label-based methods such as cross pseudo supervision [10] to tackle unlabelled data under consistency-aware concerns. The primary goal of semi-supervised approaches is to enforce consistent inference for perturbed versions of the same unlabelled data or networks. Recent studies have also explored semi-supervised learning to the feature learning power of ViT-based segmentation networks [26, 47, 48]. The proposed 3D-ViT extends the con-

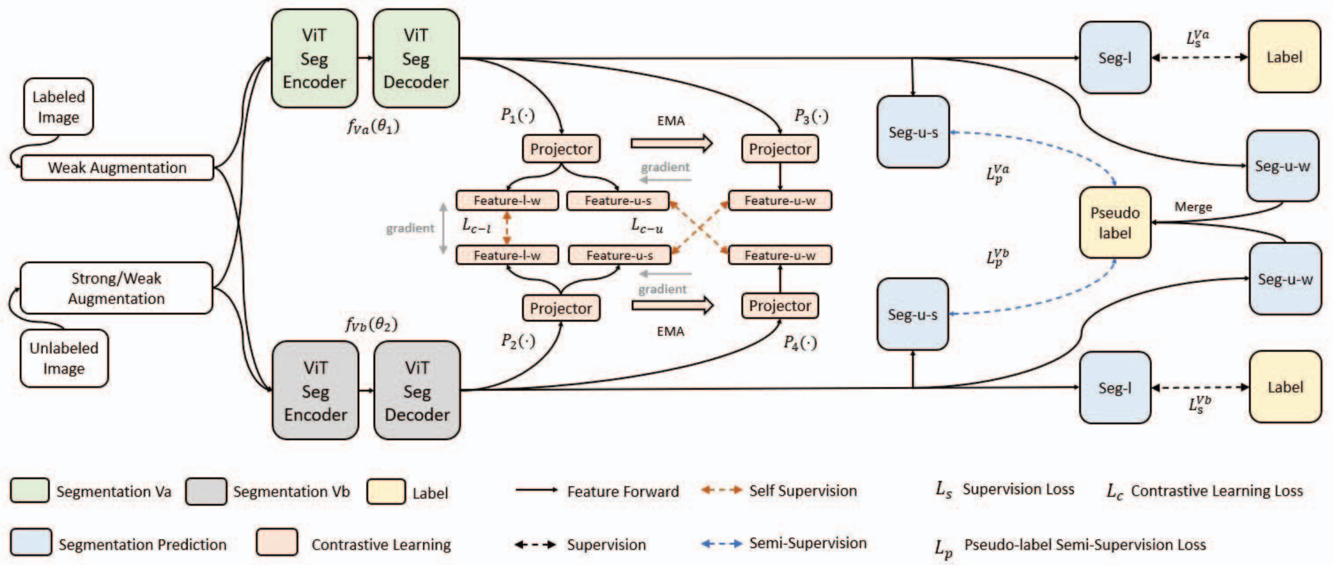

Figure 1. The Framework of the 3D-ViT for Semi-Supervised Medical Image Segmentation.

sistency concern with data perturbation (weak and strong image augmentation) and simultaneously network perturbation (two separate ViT) in a semi-supervised framework as dual-consistency learning.

3. Approach

The framework of 3D-ViT is briefly illustrated in Figure 1. In the semi-supervised study, $(\boldsymbol{X}_\mathrm{l}, \boldsymbol{Y}_\mathrm{l}) \in \mathbf{L}$, $(\boldsymbol{X}_\mathrm{u}) \in \mathbf{U}$, and $(\boldsymbol{X}_\mathrm{t}, \boldsymbol{Y}_\mathrm{t}) \in \mathbf{T}$ typically denote the labelled training dataset, unlabelled training dataset, and testing dataset, respectively. Here, $\boldsymbol{X} \in \mathbb{R}^{h\times w}$ represents a 2D grayscale image, and $\boldsymbol{Y}_\mathrm{l}, \boldsymbol{Y}_\mathrm{t} \in \mathbb{N}_4^{h\times w}$ represents a 4-class ground truth segmentation mask with pixel values ranging from 0 to 3. $\boldsymbol{p}_\mathrm{p}$, $\boldsymbol{p}_\mathrm{s}$ and $\boldsymbol{Y}_\mathrm{p}$ are the predicted probability distribution, feature with softmax activated, and the predicted segmentation mask by a segmentation network given $\boldsymbol{X}$ as $f(\theta) : \boldsymbol{X} \mapsto \boldsymbol{p}_\mathrm{p} \mapsto \boldsymbol{p}_\mathrm{s} \mapsto \boldsymbol{Y}_\mathrm{p}$ with the θ as parameters. Two segmentation networks f are based on ViT-based network but initialized separately as $f_{Va}(\theta_1)$ and $f_{Vb}(\theta_2)$. A pair of projectors $p(\cdot)$ is introduced to each network to extract representation features of $\boldsymbol{p}_\mathrm{p}$ for the labelled and unlabelled training sets, respectively, for contrastive learning purposes. The overall losses are categorized as supervision loss $\mathcal{L}_\mathrm{s}$, self-supervised contrastive loss $\mathcal{L}_\mathrm{c}$, and semi-supervised pseudo labelling loss $\mathcal{L}_\mathrm{p}$, and the final evaluation is conducted by calculating the difference between the pair of $(\boldsymbol{Y}_\mathrm{p}, \boldsymbol{Y}_\mathrm{t})$ on the test set. In the following sections, subscripts or superscripts $-l$ or $-u$ are used to denote variables on labelled or unlabelled training sets, and $-s$ and $-w$ are used to denote variables processed by strongly or weakly augumentation. The details of *dual-contrastive* learning about $p(\cdot)$, $\mathcal{L}_\mathrm{c}$, *dual-consistency* learning about $\mathcal{L}_\mathrm{p}$ are discussed in the following subsections, respectively.

3.1. Training Objective

The training objective of the proposed 3D-ViT is to minimize the sum of the supervision loss $\mathcal{L}_\mathrm{s}$, self-supervision via contrastive learning loss $\mathcal{L}_\mathrm{c}$, and the semi-supervision via pseudo labelling loss $\mathcal{L}_\mathrm{p}$. The sum of loss is indicated in Eq. 1. A pair of optimizers then aim to minimize the total loss by updating the parameters of a pair of segmentation networks ($f_{Va}(\theta_1)$, $f_{Vb}(\theta_2)$) and the parameters of two pairs of projectors ($p_l^{Va}(\cdot)$, $p_u^{Va}(\cdot)$) and ($p_l^{Vb}(\cdot)$, $p_u^{Vb}(\cdot)$) for dual-ViT.

$$\mathcal{L} = \overbrace{\underbrace{\mathcal{L}_\mathrm{s}^{Va} + \mathcal{L}_\mathrm{s}^{Vb}}_{s}}^{labelled} + \underbrace{\lambda_1\mathcal{L}_{\mathrm{c}-1} + \lambda_1 (\lambda_2\mathcal{L}_{\mathrm{c}-\mathrm{u}}}_{self} + \overbrace{\underbrace{\mathcal{L}_\mathrm{p}^{Va} + \mathcal{L}_\mathrm{p}^{Vb}}_{semi})}^{unlabelled} \tag{1}$$

where $\mathcal{L}_\mathrm{s}$ is for labelled training set via supervised learning, $\mathcal{L}_\mathrm{p}$ is for unlabelled training set via semi-supervised learning, $\mathcal{L}_{\mathrm{c}-1}$ and $\mathcal{L}_{\mathrm{c}-\mathrm{u}}$ represent the contrastive learning loss for labelled and unlabelled training sets via self-supervised learning, respectively. λ_1 and λ_2 (initial values set as 1 and 0.1) are associated with a ramp-up function [34] for solely unlabelled training set enabling the whole 3D-ViT to be initialized from the labelled training set and gradually to be shifted to focusing on learning from the unlabelled training set. The supervision loss $\mathcal{L}_\mathrm{s}$ is the sum of the training losses $\mathcal{L}_\mathrm{s}^{Va}$ and $\mathcal{L}_\mathrm{s}^{Vb}$ of two networks on the labelled training set $(\boldsymbol{X}_\mathrm{l}, \boldsymbol{Y}_\mathrm{l}) \in \mathbf{L}$, illustrated as:

$$\begin{aligned}\mathcal{L}_{\mathrm{s}} &= \mathcal{L}_{\mathrm{s}}^{Va} + \mathcal{L}_{\mathrm{s}}^{Vb} \\ &= \mathrm{CE}(Y_{\mathrm{l}}, p_{\mathrm{p}}^{Va-l}) + \mathrm{Dice}(Y_{\mathrm{l}}, p_{\mathrm{s}}^{Va-l}) \\ &\quad + \mathrm{CE}(Y_{\mathrm{l}}, p_{\mathrm{p}}^{Vb-l}) + \mathrm{Dice}(Y_{\mathrm{l}}, p_{\mathrm{s}}^{Vb-l})\end{aligned} \tag{2}$$

where CE and $Dice$ represent the CrossEntropy-based and DiceCoefficient-based difference measures. The $\mathcal{L}_{\mathrm{c}}$ and $\mathcal{L}_{\mathrm{p}}$ are discussed in 3.2 and 3.3, respectively.

3.2. Dual-Consistency Learning

Inspired by the consistency-aware concept of two networks cross-teaching each other via pseudo labels [10, 26, 36], and for better integration with contrastive learning, we leverage the consistency learning via both image perturbation and network perturbation simultaneously.

Consistency under image perturbation is achieved by high-confidence pseudo-labelling. With the same weak and strong augmentation strategy for contrastive learning, two networks' predictions on a weakly augmented unlabelled image are merged and considered as a pseudo label to supervise the prediction of the same image which is strongly augmented. The augmentation implementation is detailed in Section 4.2. To generate a pseudo label, the softmax probability map p_{s} of each weakly augmented image is normalized and masked with a threshold to retain local predictions with high confidence; the procedure is denoted as $m(\cdot)$ and the threshold is set as 0.95. Then the masked probability maps from two networks are merged to generate a pseudo label. The generation of the pseudo label is illustrated in Eq. 3:

$$Y_{\mathrm{p}} = \mathrm{argmax}\big(m(p_{\mathrm{s}}^{Va-u}) + m(p_{\mathrm{s}}^{Vb-u})\big) \tag{3}$$

The pseudo labels are masked and merged to 1) handle uncertain predictions and stabilise the training procedure with only high confidence predictions utilized; 2) increase the consistency of the two networks by using an ensembled pseudo label to improve the performance of contrastive learning [14]. The above process enforces consistency between the network's predictions under different data augmentations, which helps the network learn more robust and discriminative features.

Consistency under network perturbation is established via two networks cross teaching where the sum of semi-supervised consistency loss is illustrated in Eq. 4:

$$\begin{aligned}\mathcal{L}_{\mathrm{p}} &= \mathcal{L}_{\mathrm{p}}^{Va} + \mathcal{L}_{\mathrm{p}}^{Vb} \\ &= \mathrm{CE}(Y_{\mathrm{p}}, p_{\mathrm{p}}^{Va-u}) + \mathrm{Dice}(Y_{\mathrm{p}}, p_{\mathrm{s}}^{Va-u}) \\ &\quad + \mathrm{CE}(Y, p_{\mathrm{p}}^{Vb-u}) + \mathrm{Dice}(Y_{\mathrm{p}}, p_{\mathrm{s}}^{Vb-u})\end{aligned} \tag{4}$$

where $\mathcal{L}_{\mathrm{p}}^{Va}$ and $\mathcal{L}_{\mathrm{p}}^{Vb}$ represent the semi-supervision losses for f_{ViTa} and f_{ViTb} by the merged pseudo label, respectively.

3.3. Dual-Contrastive Learning

Contrastive learning has emerged as a prominent technique in numerous computer vision tasks, particularly when annotated data is scarce [41, 52]. The fundamental idea behind this approach is that an input image subjected to various augmentations should yield similar feature representations when processed by a neural network encoder while maintaining dissimilarity with the feature representations of distinct images [4]. To capture this notion, an appropriate contrastive loss function is formulated, and the projector is trained using raw data to minimize this loss, thereby improving the network's learning capabilities [8]. In our proposed dual-contrastive learning strategy, two pairs of projectors, ($p_l^{ViTa}(\cdot)$, $p_u^{ViTa}(\cdot)$) and ($p_l^{ViTb}(\cdot)$, $p_u^{ViTb}(\cdot)$), are added to two separate ViTs to extract image-level feature maps, which are then used to measure the similarity of two augmented images within a feature space. The InfoNCE loss [4] is used for this purpose, illustrated in Eq. 5.

$$\mathcal{L}_{\mathrm{c}} = -\log \frac{\exp(\mathbf{q}\cdot\mathbf{k}_{+}/\tau)}{\sum_{i=0}^{K}\exp(\mathbf{q}\cdot\mathbf{k}_{i}/\tau)} \tag{5}$$

Where either $\mathbf{q}$ and $\mathbf{k}$ are two batches of representation features generated by either pair of projectors ($p_l^{Va}(\cdot)$, $p_l^{Vb}(\cdot)$) or ($p_u^{Va}(\cdot)$, $p_u^{Vb}(\cdot)$); the similarity between them is measured by dot product. ($\mathbf{q}\cdot\mathbf{k}_{+}$) represents the similarity between a positive sample pair, and ($\mathbf{q}\cdot\mathbf{k}_{i}$) represents the similarity between a negative sample pair. Within a training batch, feature representations of the **same** image from different networks, with weak or strong augmentations are considered to be **positive** pairs, and feature representations for **different** images are **negative** pairs. We construct contrastive learning between and within the two networks. For labelled data, the comparison is between feature representations of weakly augmented images generated by separate networks; for unlabelled data, the comparison is between representations of weakly and strongly augmented images, generated from the same network while via different projectors, as shown in Figure 1. Two contrastive loss mechanisms are employed. For labelled data, the encoders generating both representations are updated end-to-end through back-propagation. For unlabelled data, inspired by the MoCo momentum contrast mechanism [14], the projector which generates the feature for weakly augmented images is updated as the exponential moving average (EMA) of the other projector for the same network, while only the projector generating the feature for strongly augmented images is updated through back-propagation. With a higher weight, the contrastive loss for labelled data $\mathcal{L}_{\mathrm{c-l}}$ is utilized to trigger contrastive training and guide the unlabelled data loss $\mathcal{L}_{\mathrm{c-u}}$. τ is a temperature parameter scaling the similarity [50] , set as 0.1, and K is the number of negative pairs. The architectures of the projectors are identical and are developed following VGG-style networks with the same structure of CNN and dense layers [25, 35].

4. Experiments

4.1. Datasets

The proposed 3D-ViT has been validated on the MRI Cardiac Segmentation(ACDC) dataset from the MICCAI Challenge 2017 [1]. It consists of 100 patient MRI scans divided into five different groups (normal cases, heart failure with infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle) with four annotated classes (right ventricle, myocardium, and left ventricle). This dataset provides a representative benchmark for evaluating the performance of our method in a challenging medical image segmentation task. All the images have been resized to 224 × 224 considering the ViT input fashion [23].

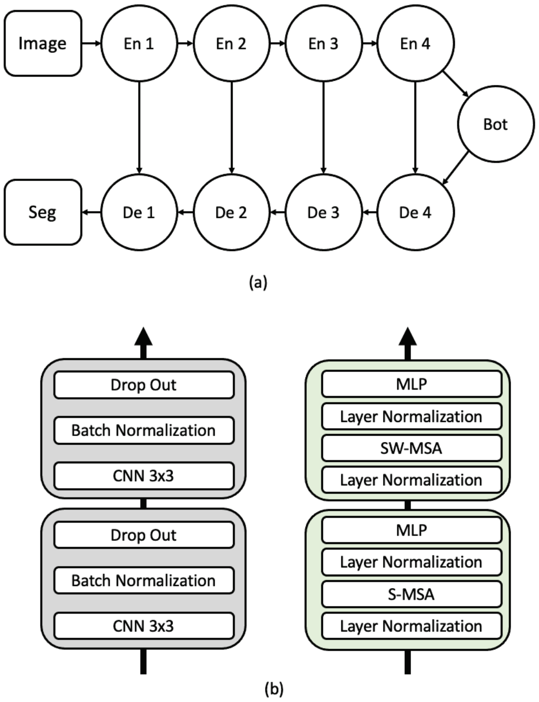

Figure 2. (a) The Encoder-Decoder Segmentation Network Architecture. (b) 2 Successive CNN or ViT Layers of Each Encoder or Decoder Network Block.

4.2. Augmentation Strategy

In 3D-ViT, two categories of data augmentation are employed: weak and strong augmentations, adapted from CTAugment [2]. These augmentations serve to 1) generate multi-view images for contrastive learning and consistency learning, as explained above; 2) generate additional training examples from the available labelled and unlabelled data, enhancing the generalization capabilities of the network. In this study, domain-adaptive augmentations are implemented on the dataset. The **Weak Augmentation** includes Rescale, Rotate, Affine, etc.; the **Strong Augmentation** includes Contrast, Blur, Sharpness, etc. Random times (within a certain range) of transformations of the above methods are implemented during either weak augmentation or strong augmentation. The random range is designed to both encourage larger feature space with deeper transformation and ensure consistency meanwhile with shallower transformations. For image segmentation tasks, the shape and location of objects in an image should be kept consistent between the two levels of augmentation, for either purpose of pseudo-labeling or contrastive learning. Hence, weak-augmented images are set as inputs for the strong augmentation, where no further geometric transformations (e.g. flips, affine) are involved. The strength and times of transformations are decaying during training.

4.3. Implementation

The experiments are conducted using PyTorch on 4 Nvidia GeForce RTX 3090 GPUs and an Intel Core i9-10900K CPU. The run times average around 5 hours for 30,000 iterations. The batch size is set to 24, and the optimizer is SGD [32], with a learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001. The networks are evaluated on the validation set every 200 iterations, and the network with the best performance is selected as the final network for testing. All the baseline methods and 3D-ViT are trained to employ the same hyperparameters, which include the segmentation network, optimizer, learning rate, batch size, and the number of iterations. The labeled training set, unlabeled training set, validation set, and test set are randomly selected, and the list of images for each set can be accessed.

4.4. CNN & ViT Segmentation Network

For the experiment, we utilize ViT and CNN layers within a unified Encoder-Decoder style segmentation network architecture, outlined in Figure 2, to facilitate a fair comparison [44]. For each level of the encoder or decoder, we employ two successive 3 × 3 convolutions, dropout, and batch normalization to construct the CNN-based network block. Alternatively, we use layer normalization, shift-window-based multi-head self-attention, residual connection, and multi-layer perception to construct the ViT-based network block. This results in pure CNN- or ViT-based UNet structures [3, 33].

4.5. Comparison with Baseline

Dice Coefficient(Dice), Intersection over Union(IoU), Accuracy(Acc), Precision(Pre), Sensitivity(Sen), Specificity(Spe), and Average Surface Distance(ASD) are uti-

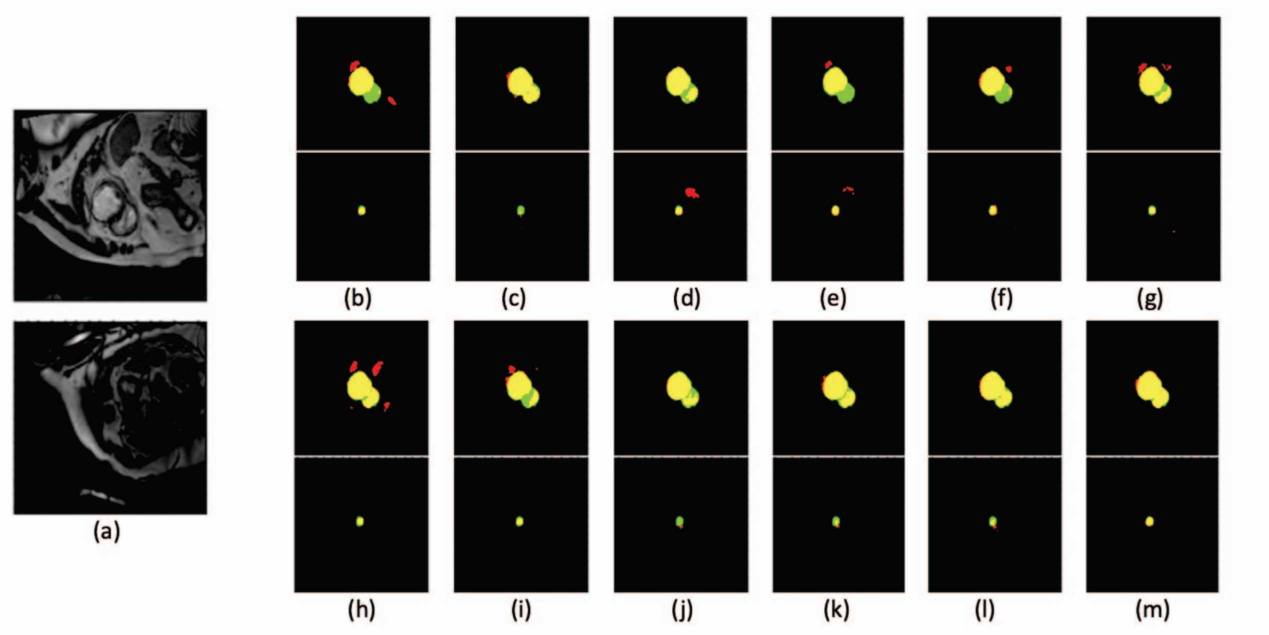

Figure 3. Example Raw Images with Corresponding Inferences Against Ground Truth of Each Method.

Table 1. The Performance of All Baseline Methods and 3D-ViT on MRI Cardiac Test Set When 20% of Training Set is Annotated.

Network	Dice↑	IoU↑	Acc↑	Pre↑	Sen↑	Spe↑	ASD↓
DAN [55]	0.8965	0.8124	0.9958	0.9052	0.8896	0.9735	1.9143
ADVENT [40]	0.9027	0.8227	0.9961	0.9155	0.8927	0.9740	1.8744
ICT [39]	0.9020	0.8215	0.9960	0.9103	0.8966	0.9749	2.2401
MT [37]	0.8949	0.8098	0.9958	0.9087	0.8837	0.9719	2.3086
UAMT [53]	0.9036	0.8242	0.9960	0.9098	0.8993	0.9756	2.1622
CPS [10]	0.9126	0.8392	0.9965	0.9179	0.9084	0.9790	1.6037
TVL [46]	0.9185	0.8493	0.9967	0.9152	0.9220	0.9842	1.8044
DCN [31]	0.8982	0.8152	0.9958	0.9100	0.8900	0.9724	2.2003
UAMTViT [47]	0.8921	0.8052	0.9958	0.8971	0.8881	0.9768	18.5518
CAAViT [44]	0.8956	0.8109	0.9959	0.9016	0.8903	0.9763	20.1265
CESSViT [45]	0.9107	0.8360	0.9964	0.9041	0.9178	0.9842	16.7005
3D-ViT	**0.9295**	**0.8722**	**0.9973**	**0.9274**	**0.9319**	**0.9878**	**1.1525**

lized as metrics. Our proposed method is compared with other semi-supervised frameworks, including Deep Adversarial Network (DAN) [55], Adversarial Entropy Minimization for Domain Adaptation(ADVENT) [40], Interpolation Consistency Training(ICT) [39], Mean Teachers (MT) [37], Uncertainty-Aware Mean Teachers (UAMT) [53], Cross Pseudo Supervision(CPS) [10], Deep Co-Training (DCN) [31], Triple-View Learning(TVL) [46], UAMTViT [47], CAAViT [44], and CESSViT [45]. For a fair comparison, the segmentation backbone networks for all baseline semi-supervised frameworks are based on CNN-based UNet [33] and Swin ViT-based UNet [3] without any modification. The detailed results are reported in Table 1, where all baseline methods are compared directly with 3D-ViT when 20% of training data is assumed as labelled data, and ↑, ↓ represent the higher/lower, the better. Different data situations with 20%, 30%, and 50% of training data as labelled data are further explored and reported in Table 2. The best performance is highlighted in **bold**, and the second-best performance of 3D-ViT is underlined. The quantitative results demonstrate that 3D-ViT outperforms all other baseline methods on various metrics under different data situations. To provide a more exhaustive assessment, we evaluated each prediction and present the distribution of IoU in Figure 4. The X-axis represents the IoU threshold, and the Y-axis signifies the number of predicted images corresponding to each threshold. The upper subfigure in Figure 4 displays the number of predictions where the corresponding

Table 2. The Performance of All Baseline Methods and 3D-ViT on MRI Cardiac Test Set Under Different Data Situations.

Labelled	20%			30%			50%		
Network	Dice↑	IoU↑	ASD↓	Dice↑	IoU↑	ASD↓	Dice↑	IoU↑	ASD↓
DAN [55]	0.8965	0.8124	1.9143	0.9093	0.8337	1.8898	0.9204	0.8525	1.3368
ADVENT [40]	0.9027	0.8227	1.8744	0.9177	0.8479	1.7516	0.9250	0.8605	1.3544
ICT [39]	0.9020	0.8215	2.2401	0.9153	0.8438	1.6228	0.9215	0.8544	1.5970
MT [37]	0.8949	0.8098	2.3086	0.9120	0.8382	1.7767	0.9121	0.8384	5.0554
UAMT [53]	0.9036	0.8242	2.1622	0.9145	0.8425	1.5430	0.9241	0.8589	1.2174
CPS [10]	0.9126	0.8392	1.6037	0.9172	0.8471	3.3960	0.9255	0.8613	**1.0584**
TVL [46]	0.9185	0.8493	1.8044	0.9186	0.8495	1.8636	0.9276	0.8650	2.3329
DCN [31]	0.8982	0.8152	2.2003	0.9119	0.8381	2.2896	0.9221	0.8555	2.9990
UAMTViT [47]	0.8921	0.8052	18.5518	0.9029	0.8230	17.5107	0.9173	0.8472	13.4582
CAAViT [44]	0.8956	0.8109	20.1265	0.9098	0.8345	1.8804	0.9252	0.8608	1.3828
CESSViT [45]	0.9107	0.8360	16.7005	0.9128	0.8396	13.2589	0.9210	0.8536	10.6794
3D-ViT	**0.9295**	**0.8722**	**1.1525**	**0.9316**	**0.8756**	**1.1263**	**0.9328**	**0.8777**	1.1215

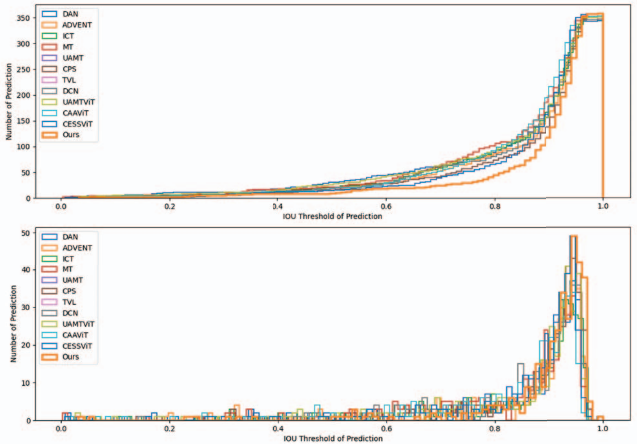

Figure 4. The IoU Distribution of Corresponding Inference Against Ground Truth of Each method.

IoU performance is equal to or lower than the IoU threshold. The bottom subfigure of Figure 4 illustrates the number of predictions whose IoU performance exactly matches the IoU threshold. These results underscore that our proposed 3D-ViT method is more inclined to generate predictions with high IoU values, thereby reinforcing its superior performance in comparison to other methods.

Some qualitative results, i.e., example images of inference against ground truth, are depicted in Figure 3, where yellow, red, green, and black indicate the true positive, false positive, false negative and true positive pixels; (a) represents input raw images, and (b-m) represents each baseline method and 3D-ViT following the same order of Table 1.

4.6. Ablation Study

To elucidate the individual and collective contributions of network structures and learning strategies in our proposed 3D-ViT, we undertake an ablation study on the case with 20% labelled data. For a fair comparison, we further investigate not only the ViT-based 3D-ViT, but also CNN-based network. Additionally, we scrutinize the impact of contrastive learning schemes, comparing scenarios in which they are utilized with or without labelled or unlabelled contrastive learning losses. The results of our in-

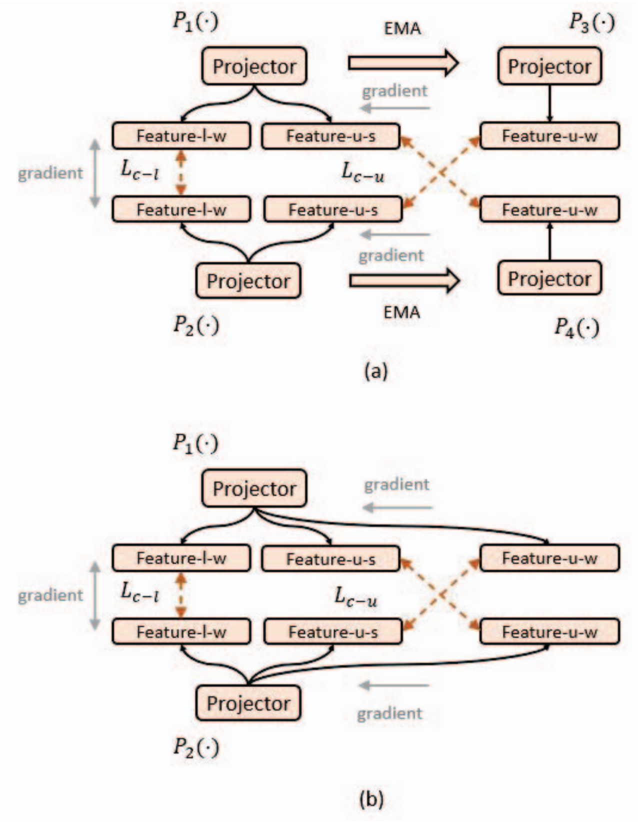

Figure 5. Illustrations of Two (a) and One (b) Pairs of Projectors.

Table 3. Ablation Study on the 3D-ViT for Semi-Supervised Image Segmentation When 20% of Training Set is Annotated.

Dual Network		Contrastive Learning		Dual Projectors	mDice↑	mIOU↑
CNN	ViT	$\mathcal{L}_{\mathrm{c-l}}$	$\mathcal{L}_{\mathrm{c-u}}$			
✓× 2				-	0.9278	0.8695
✓× 2		✓			0.9242	0.8634
✓× 2		✓	✓		0.9260	0.8664
✓× 2		✓	✓	✓	0.9257	0.8657
✓	✓			-	0.9267	0.8676
✓	✓	✓			0.9260	0.8664
✓	✓	✓	✓		0.9263	0.8669
✓	✓	✓	✓	✓	0.9260	0.8664
	✓× 2			-	0.9283	0.8623
	✓× 2	✓			0.9288	0.8710
	✓× 2	✓	✓		0.9280	0.8697
	✓× 2	✓	✓	✓	**0.9295**	**0.8722**

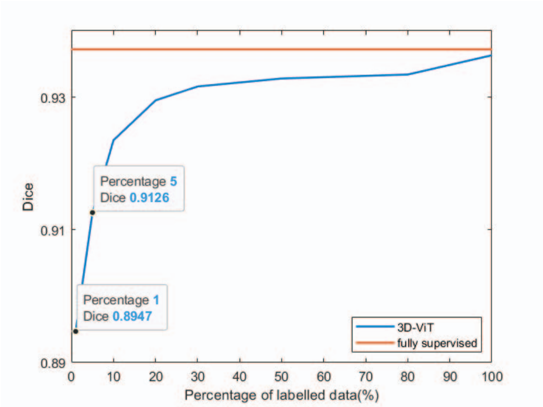

Figure 6. The Performance of 3D-ViT Compared to Fully Supervised under Various Data Situations with Dice.

quiry are illustrated in Table 3. A single checkmark (✓) within the Dual Network column signifies the mandatory employment of segmentation networks, encompassing hybrid CNN or ViT segmentation networks, along with varying contrastive learning schemes; double checkmarks (✓× 2) denote a framework constructed with two identical architecture backbone networks, i.e. 3D-ViT or an extended 3D-CNN. A checkmark (✓) within the Dual Projectors column indicates separate projectors for labelled and unlabelled set in a sub-network (Figure 5 (a)), as 3D-ViT. Otherwise the two sets share a single projector (Figure 5 (b)). The ablation study demonstrates insights into the individual and joint effects of proposed design and contributions. As shown in Table 3, overall, the performance of 3D-ViT is remarkably higher than all the baselines in Table 1 with the same amount (20 %) of labelled data, and higher or comparable to the performance of baselines with 50% labelled data, among which the architecture with two ViTs and separate projectors (3D-ViT) performs the best. The cases without contrastive learning (the first, fifth and ninth rows) prove the success of the consistency learning strategy. Furthermore, we illustrated the full range performance of 3D-ViT against the best performance it can attain (fully supervised), as shown in Figure 6. With only 1 % labelled data, the network achieves dice near 90 %, and with only 5 % labelled data, the network performs better than most of the baselines with 20 % labelled data. As shown in Figure 6, the network can achieve near-perfect performance with limited data. However, from Table 3, although the proposed learning strategies are robust across diverse constructions, only the setting with two ViTs can make good use of contrastive learning upon the confidence-aware pseudo labelling. Except the structures with two ViTs, the performance between one and two pairs of projectors are comparable, which is reasonable as the power of contrastive learning has not been fully utilized.

5. Conclusion

In this paper, we proposed 3D-ViT, a novel contrastive consistency segmentation transformer that effectively utilizes limited annotations for medical image segmentation. By incorporating a modified dual-contrastive learning strategy and a dual-consistency scheme, our method promotes the development of robust feature representations and ensures prediction consistency across perturbed input images. The dual-ViT is also explored within the proposed framework and validates its promising performance against other semi-supervised methods with various combinations of CNN- or ViT-based networks. The 3D-ViT achieves state-of-the-art performance to the best of our knowledge. We look forward to validating the 3D-ViT in other limited-annotation situations such as weakly-supervised learning.

References

[1] Olivier Bernard et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE TMI*, 2018. 1, 2, 5

[2] David Berthelot et al. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 5

[3] Hu Cao et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV*. Springer, 2023. 1, 2, 5, 6

[4] Krishna Chaitanya et al. Contrastive learning of global and local features for medical image segmentation with limited annotations. *NIPS*, 2020. 1, 2, 4

[5] Jieneng Chen et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 1, 2

[6] Liang-Chieh Chen et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017. 1, 2

[7] Shuai Chen et al. Multi-task attention-based semi-supervised learning for medical image segmentation. In *MICCA*. Springer, 2019. 1

[8] Ting Chen et al. Big self-supervised models are strong semi-supervised learners. *NIPS*, 2020. 4

[9] Ting Chen et al. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[10] X Chen et al. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021. 2, 4, 6, 7

[11] Cicek et al. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, 2016. 2

[12] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[13] Kaiming He et al. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[14] Kaiming He et al. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 4

[15] Xinrong Hu et al. Semi-supervised contrastive learning for label-efficient medical image segmentation. In *MICCAI*. Springer, 2021. 1, 2

[16] Gao Huang et al. Densely connected convolutional networks. In *CVPR*, 2017. 2

[17] Guoliang Kang et al. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, pages 4893–4902, 2019. 2

[18] Zhanghan Ke et al. Guided collaborative training for pixel-wise semi-supervised learning. In *ECCV*. Springer, 2020. 1, 2

[19] Byoungjip Kim et al. Selfmatch: Combining contrastive self-supervision and consistency for semi-supervised learning. *arXiv preprint arXiv:2101.06480*, 2021. 1

[20] S Laine et al. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 1, 2

[21] Xiaomeng Li et al. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *BMVC*, 2018. 1

[22] Tsung-Yi Lin et al. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[23] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, 2021. 1, 2, 5

[24] J Long et al. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 2

[25] Ange Lou et al. Min-max similarity: A contrastive semi-supervised deep learning network for surgical tools segmentation. *IEEE TMI*, 2023. 2, 5

[26] Xiangde Luo et al. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *MIDL*, 2022. 2, 4

[27] Fausto Milletari et al. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *IEEE 3DV*. IEEE, 2016. 1, 2

[28] Pim Moeskops et al. Deep learning for multi-task medical image segmentation in multiple modalities. In *MICCAI*. Springer, 2016. 1

[29] Ozan Oktay et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 2

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[31] Siyuan Qiao et al. Deep co-training for semi-supervised image recognition. In *ECCV*, 2018. 1, 2, 6, 7

[32] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, 1951. 5

[33] Olaf Ronneberger and et al. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1, 2, 5, 6

[34] Laine Samuli and Timo Aila. Temporal ensembling for semi-supervised learning. *ICLR*, 2016. 3

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 5

[36] Kihyuk Sohn et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NIPS*, 2020. 2, 4

[37] A Tarvainen et al. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017. 1, 2, 6, 7

[38] Ashish Vaswani et al. Attention is all you need. *NIPS*, 2017. 1, 2

[39] V Verma et al. Interpolation consistency training for semi-supervised learning. In *IJCAI*, 2019. 2, 6, 7

[40] T H Vu et al. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 2, 6, 7

[41] Kaiping Wang et al. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *MedIA*, 2022. 4

[42] Xiaolong Wang et al. Non-local neural networks. In *CVPR*, 2018. 1

[43] Ziyang Wang et al. Quadruple augmented pyramid network for multi-class covid-19 segmentation via ct. In *EMBC*, 2021. 2

[44] Ziyang Wang et al. Adversarial vision transformer for medical image semantic segmentation with limited annotations. In *BMVC*, 2022. 2, 5, 6, 7

[45] Ziyang Wang et al. Computationally-efficient vision transformer for medical image semantic segmentation via dual pseudo-label supervision. In *ICIP*. IEEE, 2022. 2, 6, 7

[46] Z Wang et al. Triple-view feature learning for medical image segmentation. In *MICCAI*, 2022. 2, 6, 7

[47] Z Wang et al. Uncertainty-aware transformer for MRI cardiac segmentation via mean teachers. *MIUA*, 2022. 2, 6, 7

[48] Ziyang Wang et al. When cnn meet with vit: Towards semi-supervised learning for multi-class medical image semantic segmentation. In *ECCV*. Springer, 2023. 2

[49] Huisi Wu et al. Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In *CVPR*, 2022. 2

[50] Zhirong Wu et al. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 4

[51] Enze Xie et al. Segformer: Simple and efficient design for semantic segmentation with transformers. *NIPS*, 2021. 1, 2

[52] Chenyu You et al. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE TMI*, 2022. 2, 4

[53] L Yu et al. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *MICCAI*, 2019. 2, 6, 7

[54] Dan Zhang and Fangfang Zhou. Self-supervised image denoising for real-world images with context-aware transformer. *IEEE Access*, 2023. 1

[55] Y Zhang et al. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *MICCAI*, 2017. 2, 6, 7

[56] Fangfang Zhou, Zhengming Fu, and Dan Zhang. High dynamic range imaging with context-aware transformer. In *IJCNN*. IEEE, 2023. 1