DeNAS-ViT: Data Efficient NAS-Optimized Vision Transformer for Ultrasound Image Segmentation

Renqi Chen[1], **Xinzhe Zheng**[2], **Haoyang Su**[3*], **Kehan Wu**[4]
[1]College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai, China
[2]School of Computing, National University of Singapore, Singapore
[3]Australian Institute for Machine Learning, The University of Adelaide, Adelaide, Australia
[4]Southern University of Science and Technology, Shenzhen, China
rqchen23@m.fudan.edu.cn, zxz.krypton@outlook.com, haoyang.su@adelaide.edu.au, 12010519@mail.sustech.edu.cn

Abstract

Accurate segmentation of ultrasound images is essential for reliable medical diagnoses but is challenged by poor image quality and scarce labeled data. Prior approaches have relied on manually designed, complex network architectures to improve multi-scale feature extraction. However, such handcrafted models offer limited gains when prior knowledge is inadequate and are prone to overfitting on small datasets. In this paper, we introduce DeNAS-ViT, a **D**ata **e**fficient **NAS**-optimized **Vi**sion **T**ransformer, the first method to leverage neural architecture search (NAS) for ultrasound image segmentation by automatically optimizing model architecture through token-level search. Specifically, we propose an efficient NAS module that performs multi-scale token search prior to the ViT's attention mechanism, effectively capturing both contextual and local features while minimizing computational costs. Given ultrasound's data scarcity and NAS's inherent data demands, we further develop a NAS-guided semi-supervised learning (SSL) framework. This approach integrates network independence and contrastive learning within a stage-wise optimization strategy, significantly enhancing model robustness under limited-data conditions. Extensive experiments on public datasets demonstrate that DeNAS-ViT achieves state-of-the-art performance, maintaining robustness with minimal labeled data. Moreover, we highlight DeNAS-ViT's generalization potential beyond ultrasound imaging, underscoring its broader applicability.

1 Introduction

Ultrasound is a critical medical imaging tool for diagnosing cardiac diseases, and precise segmentation of ultrasound images can greatly aid clinicians in thoroughly analyzing cardiac conditions (Leclerc et al. 2019). Nonetheless, the development of robust and efficient segmentation algorithms faces challenges such as the scarcity of annotated data, speckle noise in ultrasound images, and the difficulty of distinguishing adjacent anatomical structures (Zhou et al. 2023; Lin et al. 2023). These challenges highlight the need to **enhance feature extraction across multiple scales** (Lin et al. 2017; Su et al. 2020; Chen et al. 2024) and to **improve model robustness to limited data** (Tarvainen and Valpola 2017; Qiao et al. 2018; Xia et al. 2020).

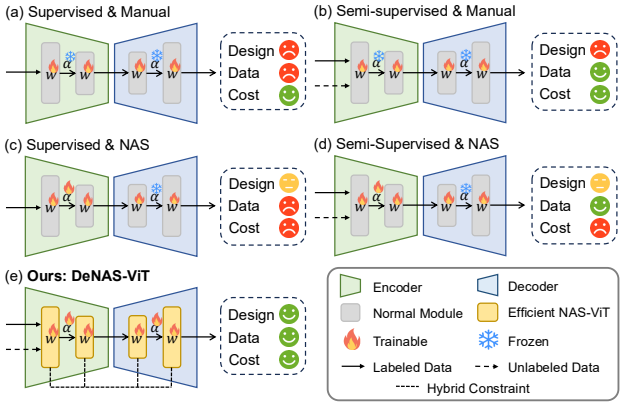

Figure 1: An illustration comparing DeNAS-ViT with existing baselines for segmentation task. Use smiley, neutral, and sad faces to present the performance. "Design" denotes the optimization of model architecture for multi-scale feature extraction, "Data" denotes the robustness to limited data, and "Cost" denotes the computational resource consumption. α and w present architecture and weights, respectively.

In recent years, deep learning has advanced medical image segmentation by improving multi-scale feature understanding. Ronneberger et al. (Ronneberger, Fischer, and Brox 2015) introduced U-Net, a U-shaped encoder-decoder architecture with exceptional feature extraction capabilities, inspiring variants like UNet++ (Zhou et al. 2019) that emphasize local feature perception. Concurrently, the Vision Transformer (ViT) (Dosovitskiy et al. 2020) emerged, leveraging self-attention to capture long-range pixel relationships and contextual information. This led to hybrid models such as TransUNet (Chen et al. 2021a), which integrates U-Net with ViT to balance local and global features, and EfficientViT (Cai et al. 2023), which employs multi-scale linear attention for efficient feature representation. However, manually designing such architectures demands significant expertise and struggles to optimize multi-scale feature extraction effectively. NAS has gained traction as an automated solution, optimizing network structures within a defined search space (Liu et al. 2019; Xu et al. 2019; Lu et al. 2022). While NAS has been applied to ultrasound image segmentation (Cao et al. 2022; Qian et al. 2022; Chen et al. 2024), ex-

*Corresponding author.

isting methods typically select operations at the module level (e.g., convolutions or Transformer blocks), neglecting finer-grained operations within these modules. This limitation restricts the representational capacity for multi-scale feature extraction. Additionally, module-level search strategies often suffer from low precision, resulting in inefficient computational resource utilization. In this work, we address both challenges by refining operation selection within modules.

Moreover, ultrasound datasets often lack sufficient labeled data, a challenge exacerbated by NAS's data-intensive nature. SSL offers a solution by leveraging unlabeled data, with methods like mean teacher (Tarvainen and Valpola 2017; Yu et al. 2019) and co-training (Chen et al. 2021b; Miao et al. 2023) enhancing robustness through constraints on unlabeled data. Combining NAS with SSL thus holds promise for addressing both limited labeled data and complex model design in ultrasound segmentation. However, prior efforts (Pauletto, Amini, and Winckler 2022; Chen et al. 2024) have largely adopted basic NAS-SSL integration without embedding additional constraints during NAS training, often leading to overfitting on small labeled datasets, especially with complex architectures (Huesmann et al. 2021; Song et al. 2023; Salehin et al. 2024).

To address these challenges, we propose DeNAS-ViT, a data-efficient, NAS-optimized ViT designed for ultrasound image segmentation. Fig. 1 highlights the key distinctions from existing methods. For architecture design, we employ NAS at multiple levels: at the cellular level, we introduce an efficient NAS-ViT module that integrates NAS with ViT to optimize multi-scale token representations while reducing computational overhead; at the module level, NAS is applied distinctly to the encoder and decoder, each with specialized search spaces to preserve their unique roles. To address ultrasound's data scarcity and NAS's data-intensive nature, we propose a NAS-based SSL framework with hybrid constraints, which includes: (1) a NAS-derived network independence loss to encourage complementary model representations; (2) a hierarchical NAS-based contrastive loss to maximize mutual information across views, boosting feature representation and generalization; and (3) a tailored stage-wise optimization strategy. In summary, the main contributions of this paper are:

- We propose a data efficient NAS-optimized Vision Transformer (DeNAS-ViT) for ultrasound segmentation, marking the first integration of NAS for token-level searching and multi-scale feature representation.
- We introduce a NAS-based constraint-driven SSL framework with a multi-stage optimization strategy, reducing overfitting with limited ultrasound image annotations and alleviating the data-intensive demands of NAS.
- Experiments on public datasets show that our method achieves state-of-the-art performance and shows potential for generalization beyond ultrasound segmentation.

2 Related Work

Neural Architecture Search. NAS is proposed to address the difficulty of network design and find the optimal network architecture within the search space (Ren et al. 2021; White et al. 2023). Differentiable architecture search (DARTS) (Liu, Simonyan, and Yang 2018) created a continuous relaxation algorithm to make gradient-based NAS efficiently trainable. Then, PC-DARTS (Xu et al. 2019) proposed partial channel connections and edge normalization, reducing GPU memory consumption. Edge normalization inspired the hierarchical NAS (HNAS) (Liu et al. 2019; Yu, Lee, and Chen 2023; Yang et al. 2023), enabling a multi-level architecture search. The Transformer is also used for context information. Yang et al. (2023) proposed DAST which incorporates the ViT layer as candidate operations for cell-level searching. Unlike these approaches, we delve into a more fine-grained search space to enhance multi-scale feature representation and reduce parameter costs.

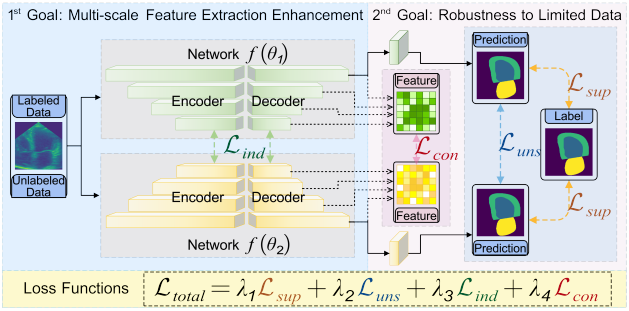

Figure 2: The pipeline of DeNAS-ViT. The hierarchical structure of NAS networks serves for the goal of multi-scale feature extraction enhancement, and multi-constraint SSL serves for the goal of robustness to limited data.

SSL for Medical Image Segmentation. SSL addresses the scarcity of labeled data in medical image segmentation, with approaches broadly classified into pseudo-labeling (Tarvainen and Valpola 2017; Bai et al. 2017; Wang et al. 2021) and consistency regularization (Qiao et al. 2018; Yu et al. 2019; Xia et al. 2020). Pseudo-labeling methods generate labels for unlabeled data during training, as seen in self-training (Bai et al. 2017). In contrast, consistency regularization enforces prediction alignment. Tarvainen et al. (Tarvainen and Valpola 2017) proposed the Mean Teacher (MT), averaging weights to create a teacher model that ensures consistency between predictions and targets. Beyond MT, Qiao et al. (Qiao et al. 2018) developed a deep co-training framework, training two networks on distinct views for complementary learning. In this work, we propose a hybrid constraint-driven SSL framework that not only mitigates the limited availability of medical image data but also reduces the sample dependency of NAS.

3 Methodology

An overview of DeNAS-ViT is presented in Fig. 2. To enhance multi-scale feature extraction, we introduce the Efficient NAS-ViT module (Sec.3.1) within the NAS backbone (Sec. 3.2). This module integrates NAS with a ViT to perform token-level search, optimizing multi-scale feature representation while alleviating the computational burden typically associated with NAS.

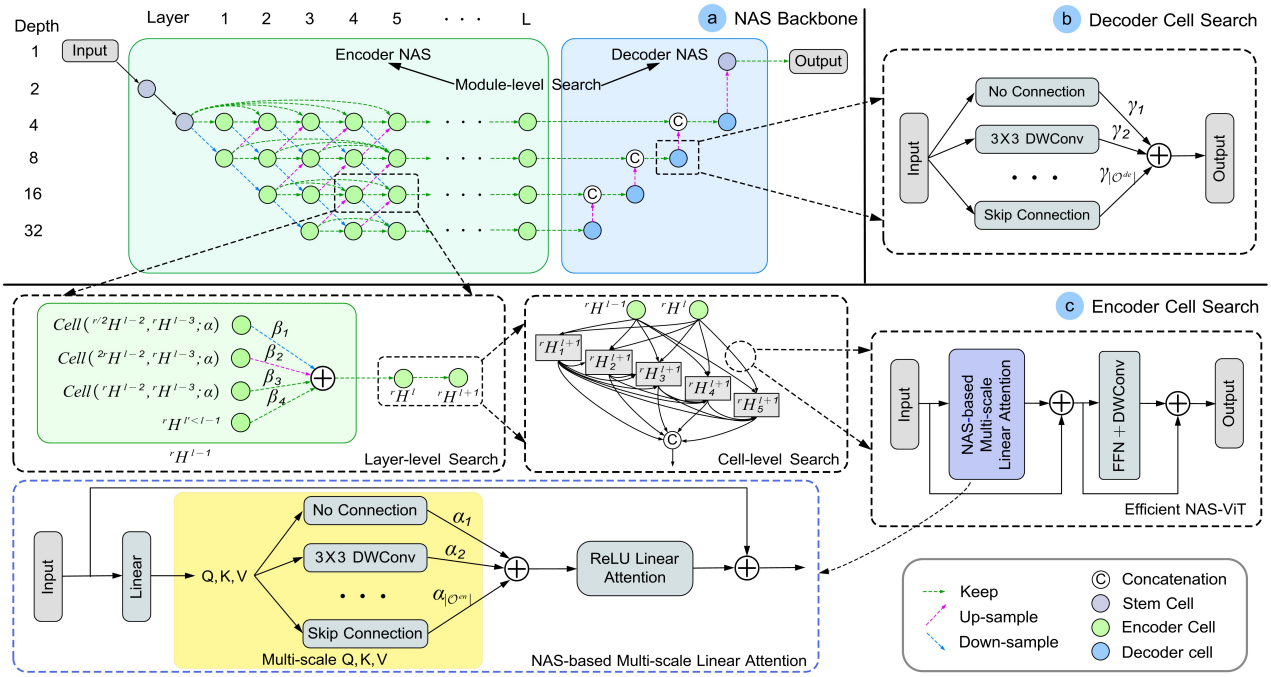

Figure 3: The proposed NAS backbone. (a) shows an overview of the NAS backbone, which consists of an encoder NAS and a decoder NAS, representing a module-level search. The input is passed through the encoder NAS to obtain multi-resolution feature maps, where a hierarchical encoder cell search is performed (shown in (c)). These outcomes are then processed by the decoder cells (shown in (b)), which concatenate features and complete the recovery process through further searching.

To improve robustness under limited labeled data, we propose a hybrid constraint-driven semi-supervised learning (SSL) framework (Sec.3.3), extending a co-training approach. This framework employs two networks, $f(\theta_1)$ and $f(\theta_2)$, which share the NAS backbone but operate with distinct parameters and independent optimizers. We implement a stage-wise optimization strategy (Sec. 3.4), incorporating an independence loss to promote complementary representations between the networks, alongside a contrastive loss for enhanced feature representation. This dual approach refines model updates, improving resilience to data scarcity.

3.1 Efficient NAS-ViT Module

In DAST (Yang et al. 2023), the ViT layer is treated as the candidate operation for multi-scale feature representation, leading to increased computation complexity. On the other hand, since the ViT is more complex than other operations, the model tends to select the ViT during backbone optimization, rendering the search meaningless. To solve these, we use the ViT as the basic unit, further implementing NAS before the attention calculation for searching multi-scale tokens, termed Efficient NAS-ViT. Compared to EfficientViT (Cai et al. 2023), where tokens are processed by fixed multi-scale convolution and sent into attention separately, our design can obtain a better multi-scale feature representation while reducing excessive parameter overhead. The effectiveness of this novel Transformer-based approach for NAS is proven through the empirical results (Sec. 4.5). An overview of the Efficient NAS-ViT is shown in the "Efficient NAS-ViT" of Fig. 3(c), used in the encoder NAS's cell search.

Given an input x, after a linear projection, the tokens are defined as $Q = xW_Q$, $K = xW_K$, and $V = xW_V$, where W_Q, W_K, and W_V represent the learnable linear projection matrices. Rather than using fixed operations to obtain multi-scale tokens, which constrains the token representation and affects attention calculation, we employ NAS to search for token representations, thereby enhancing the multi-scale learning capability. After incorporating partial channel connections to reduce memory overhead and reusing continuous relaxation for a differentiable search space, the multi-scale tokens $Q'/K'/V'$ are:

$$\{Q'/K'/V'\} = (1 - \mathbf{P}) \odot \{Q/K/V\} + \sum_{O_i \in \mathcal{O}^{en}} \frac{exp\{\alpha_i\}}{\sum_{j=1}^{|\mathcal{O}^{en}|} exp\{\alpha_j\}} \cdot O_i(\mathbf{P} \odot \{Q/K/V\}), \tag{1}$$

where $\mathbf{P}$ is the sampling mask for channel selection, $\odot$ denotes Hadamard product, O_i denotes the i-th operation selected from the set of encoder candidate operations $\mathcal{O}^{en}$ (encoder and decoder consider different sets of candidate operations due to their respective roles, please refer to Appendix A), and α is the cell architecture parameter, which measures the weight of the related candidate operation. Subsequently, tokens are computed by ReLU linear attention for contextual information, denoted as:

$$R_n = \sum_{i=1}^{S} \frac{ReLU(Q'_n)ReLU(K'_i)^T}{\sum_{j=1}^{S} ReLU(Q'_n)ReLU(K'_j)^T} V'_i, \tag{2}$$

where subscript n denotes the n-th row of R/Q, and S is the sequence length. Then, the output R is fed into the FFN+DWConv layer (the application of depthwise convolution on the FFN layer) for local information capture.

3.2 NAS Backbone

Our NAS backbone builds on HNAS framework (Liu et al. 2019; Fang et al. 2020; Chen et al. 2024), with enhancements for multi-scale optimization: (1) macro-level dual decoder NAS; (2) micro-level efficient NAS-ViT.

Encoder NAS We have cell- and layer-level searches for the multi-scale feature extraction. There are N intermediate nodes in cell search space, where the edge between nodes corresponds to the proposed Efficient NAS-ViT module, shown in the "Cell-level Search" of Fig. 3(c). After processing inputs by Efficient NAS-ViT ($f_{NAS-ViT}$), the output of the n-th node of the encoder cell in the l-th layer is ${}^{r}H_n^l = \sum_{{}^{r}H_i^l \in \mathcal{I}} f_{NAS-ViT}({}^{r}H_i^l; \alpha)$, where the input set $\mathcal{I}$ includes the previous cell's output and previous nodes' outputs in the current cell, and r denotes resolution.

To capture multi-resolution features, our model considers six values: $r = 1, 2, 4, 8, 16, 32$, where $r = 1$ corresponds to the original image size. Two stem cells down-sample the input from $r = 1$ to $r = 4$. When $r = 4$, the spatial size of the feature maps is $1/4$ of the case when $r = 1$. The final output tensor of the cell is the concatenation of outputs from all nodes. For simplicity, the cell-level search is written as $Cell_\alpha({}^{r}H^{l-1}, {}^{r}H^{l-2})$.

Following (Fang et al. 2020), the layer-level search aims at aggregating feature paths from different resolutions by employing relaxation and skip connections, shown in the "Layer-level Search" of Fig. 3(c). Path weights are referred to as architecture parameters β. There are L layers in the backbone, and the l-th layer level search is denoted as:

$$\begin{aligned} {}^{r}H^{l} = \beta_1 Cell({}^{\frac{r}{2}}H^{l-1}, {}^{r}H^{l-2}) + \beta_2 Cell({}^{2r}H^{l-1}, {}^{r}H^{l-2}) \\ + \beta_3 Cell({}^{r}H^{l-1}, {}^{r}H^{l-2}) + \beta_4 \{{}^{r}H^{l'} \in {}^{r}H | l' < l-1\}, \end{aligned} \tag{3}$$

where $\sum_i \beta_i = 1$, normalized and implemented as softmax.

Decoder NAS Rather than employing fixed convolution kernel sizes, we also introduce NAS for the decoder to enhance the capability. A U-shaped architecture is used as the backbone for the decoder NAS, as depicted in Fig. 3(a).

In the decoder NAS, the features from the $r = 32$ layer after the encoder NAS are fed into the decoder cell as the initial input. To mitigate the high computational complexity introduced by NAS, the decoder cell does not contain intermediate nodes, as shown in Fig. 3(b). The search process can be represented as:

$${}^{r}H^{de} = \sum_{O_i \in \mathcal{O}^{de}} \frac{exp\{\gamma_i\}}{\sum_{j=1}^{|\mathcal{O}^{de}|} exp\{\gamma_j\}} \cdot O_i({}^{r}H^{l=L}), \tag{4}$$

where γ are decoder architecture parameters, and ${}^{r}H^{de}$ is the final output tensor of the decoder cell in r-resolution.

After upsampling, the features ${}^{r}H^{de}$ are concatenated with ${}^{r/2}H^{L}$, and a convolution layer is applied to match the number of channels with the $r/2$ decoder NAS. This process is repeated until the $r = 4$ features are combined. Finally, an upsample layer is employed to recover the outputs to the full resolution, followed by a convolution layer to obtain the desired number of classes for target tasks.

3.3 NAS-based Constraint-driven SSL

To leverage the unlabeled data, the co-training framework of SSL is implemented, where the unsupervised loss is:

$$\mathcal{L}_{uns,a} = CE(p_{u,a}, \hat{y}_{u,b}), \tag{5}$$

where $a, b \in \{1, 2\}$ and $a \neq b$, corresponds to $f(\theta_a)$ and $f(\theta_b)$, respectively. $CE(\cdot)$ indicates the cross-entropy loss. $p_{u,a}$ is the predicted probability map generated by one network on the unlabeled data, and $\hat{y}_{u,b}$ is the corresponding one-hot pseudo label generated by another.

As the algorithmic independence could facilitate the creation of distinct networks (Miao et al. 2023), especially in the co-training, complementary networks are capable of capturing diverse feature information and are less likely to overfit to a particular subset. Thus, we incorporate network independence loss and propose a stage-wise optimization strategy (Sec. 3.4) to fully utilize very few samples. The network independence loss is defined between two NAS backbones based on convolutional layers at the same position:

$$\mathcal{L}_{ind,a} = \frac{1}{L_{CNN}} \sum_{i=1}^{L_{CNN}} IND(\theta_{a,i}, \theta_{b,i}; G_{b,i}), \tag{6}$$

where L_{CNN} is the number of convolutional layers. $\theta_{a,i} \in \mathbb{R}^{C_{out} \times d} (d = K \times K \times C_{in})$ are the weights of the i-th convolutional layer in $f(\theta_a)$, reshaped into a matrix form, where K is kernel size. C_{in}, C_{out} are the number of input and output feature channels. $G_{b,i} \in \mathbb{R}^{C_{out} \times C_{out}}$ is the corresponding optimal coefficient matrix. Given matrices A, B, G_B, the independence loss function is defined as: $IND(A, B; G_B) = \frac{1}{C_{out}} \sum_{i=1}^{C_{out}} (\frac{\boldsymbol{v}_{A,i} \cdot \boldsymbol{q}_{B,i}}{|\boldsymbol{v}_{A,i}| \times |\boldsymbol{q}_{B,i}|})^2$, where $\boldsymbol{v}_{A,i}$ is the i-th row of A and $\boldsymbol{q}_{B,i} = (G_B \times B)_i$.

Since algorithmic independence essentially endows the network with the capability to observe the same image from different perspectives, we consider that contrastive loss can be utilized to maximize the mutual information across these views to enhance discriminative feature representation, which also helps the model to generalize to new samples. Benefiting from network independence, there is also no need to construct an asymmetric architecture for contrastive loss. In this work, an uncertainty-based contrastive loss (Huang et al. 2023) is calculated based on the hierarchical architecture of NAS, measured at different stages of the decoder NAS cell between two networks. The uncertainty estimation is defined using smoothed KL-divergence and considers features at various resolutions:

$$U_a^{r,h,w} = \sum_{c=0}^{C-1} ({}_a^{r}H^{de})^{c,h,w} \cdot \log \frac{({}_a^{r}H^{de})^{c,h,w} + \epsilon}{\overline{({}_a^{r}H^{de})^{c}} + \epsilon}, \tag{7}$$

where C is the channel dimension, ϵ is a small bias term, $r \in \{4, 8, 16, 32\}$ is used, and ${}_a^{r}H^{de}$ is the output tensor

of the decoder cell at the r-resolution. $\overline{(\cdot)}$ is the mean value across the channel dimension. A higher estimation value reflects a higher uncertainty, which can be used to compel the lower-quality features to align with their higher-quality counterparts (Huang et al. 2023). The positions of these features are estimated by $\mathcal{P}_a^{r,h,w} = \mathbf{1} \odot \{U_a^{r,h,w} > U_b^{r,h,w}\}$. Then, mean squared error is used in the loss function:

$$\mathcal{L}_{con,a} = \sum_{r\in\{4,8,16,32\}} \sum_{p\in\mathcal{P}_a^{r,h,w}} MSE({}_a^rH^{de})^p, ({}_b^rH^{de})^p). \tag{8}$$

The total loss function of our DeNAS-ViT is:

$$\mathcal{L}_{total,a} = \lambda_1\mathcal{L}_{sup,a} + \lambda_2\mathcal{L}_{uns,a} + \lambda_3\mathcal{L}_{ind,a} + \lambda_4\mathcal{L}_{con,a}, \tag{9}$$

where $\mathcal{L}_{sup,a} = \frac{1}{2}[CE(p_{l,a}, y_{l,a}) + DICE(p_{l,a}, y_{l,a})]$ is the supervised loss, which is the combination of the cross-entropy loss and Dice loss calculated on the labeled dataset; $\mathcal{L}_{uns,a}$ refers to the unsupervised loss in Eq. (5) calculated on the unlabeled dataset; $\mathcal{L}_{ind,a}$ refers to the network independence loss in Eq. (6) based on the network architecture; $\mathcal{L}_{con,a}$ refers to the contrastive loss in Eq. (8) based on the multi-resolution features. λ_1, λ_2, λ_3, and λ_4 are hyper-parameters to balance the relationship between losses.

3.4 Stage-wise Optimization Strategy

To adapt to the hybrid constraint-driven SSL, we propose a stage-wise optimization strategy, summarized in Algorithm 1. In each iteration, we first fix the network parameters and optimize the combination matrix for E_B epochs. In the second stage, we fix the combination matrix and network architecture parameters and then update the network weights by minimizing $\mathcal{L}_{total}$. In the third stage, we only update the architecture parameters by minimizing $\mathcal{L}_{total}$ after E_A epochs. During the optimization, continuous relaxation is implemented for the gradient descent algorithm.

Algorithm 1: Optimization Strategy

Input: Datasets $\mathcal{D}_l$, $\mathcal{D}_u$, weights w, architecture α, β, γ, combination matrices G_a, G_b, epochs E, E_A, and E_B.
Output: Searched $f(\theta_a^*), f(\theta_b^*)$.

```
for e = 1, ..., E do
    for f = 1, ..., E_B do
        Fix w, α, β, γ. Update G_a, G_b by maximizing L_ind,a and L_ind,b, respectively.
    Fix G_a, G_b, α, γ. Update w, β of f(θ_a) and f(θ_b) by minimizing L_total,a and L_total,b, respectively.
    if e > E_A then
        Fix G_a, G_b, w, β. Update α, γ of f(θ_a) and f(θ_b) by minimizing L_total,a and L_total,b, respectively.
```

4 Experiments

4.1 Datasets

Three public datasets are utilized for evaluation: (i) CAMUS dataset (Leclerc et al. 2019) is a large-scale 2D echocardiography dataset, comprising 2000 labeled images, and approximately 19000 unlabeled images. It includes four classes of labels: left ventricle endocardium (LV), left atrium, myocardium, and background. (ii) HMC-QU dataset (Kiranyaz et al. 2020) consists of 2D echocardiography videos. By splitting these sequences into individual images, a total of 4989 images are obtained, among which 2349 images are annotated with two classes of labels: left ventricle wall and background. (iii) CETUS dataset (Bernard et al. 2015) comprises 90 sequences of 3D ultrasound volumes. After randomly selecting 80 frames from each sequence, 7200 annotated images are obtained, with two classes of labels: LV and background. In our experiments, 3400 of these images are utilized as unlabeled data. More details on dataset preparation are provided in Appx. C.1.

4.2 Experimental Settings

Network Architecture. In the encoder NAS, each cell has $N = 5$ intermediate nodes. For the resolution $r = 4$, the channel numbers are fixed as 8 for each node. When r doubles, the number of channels doubles accordingly. For partial channel connections, $\frac{1}{4}$ of the channels are allowed for searching. The default number of layers L is set to 8 and the default proportion of labeled data utilized is 50%.

Training Setup. The hyper-parameters $\lambda_1 = 1$, $\lambda_2 = \lambda_4 = 5exp(-5(1 - \frac{min(i, I_{ramp})}{I_{ramp}})^2)$ are adopted following (Yu et al. 2019; Huang et al. 2023) at the i-th epoch, where $I_{ramp} = 50$, and $\lambda_3 = 2$. The linear coefficient matrices G are optimized by Adam with a fixed learning rate of 0.001. The network weights w and architecture β are optimized using SGD with an initial learning rate of 0.001, a momentum of 0.9, and a weight decay of 0.0003. For the architecture α and γ, the Adam optimizer is applied with a learning rate of 0.003 and weight decay of 0.001. The total number of epochs is set to $E = 40$ and the architecture optimization begins at $E_A = 10$. In each epoch, G is updated $E_B = 6$ times. Experiments are conducted on an Nvidia A100 GPU and are repeated with 5 random seeds, where mean value and standard deviation are reported in $mean_{(std)}$ format.

Metrics. Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and 95% Hausdorff Distance (95HD). Wilcoxon signed-rank test for statistical significance.

4.3 Main Results

Comparison with SOTA Methods. We evaluate DeNAS-ViT against other SOTAs, including supervised handcrafts: UNet++, nnU-Net (Isensee et al. 2021), Transfuse (Zhang, Liu, and Hu 2021); supervised NAS: Auto-DeepLab, M^{3}NAS (Lu et al. 2022), GeNAS (Jeong et al. 2023); semi-supervised handcrafts: URPC (Luo et al. 2021), CPS, CnT-B (Huang et al. 2023), ARCO-SG (You et al. 2024); and semi-supervised NAS: Se^{2}NAS (Pauletto, Amini, and Winckler 2022), SSHNN (Chen et al. 2024), implemented by their officially released code and pre-trained model.

Table 1 and Appx. Fig. A1 provide quantitative and qualitative comparisons across HMC-QU, CAMUS, and CETUS datasets. Supervised methods were trained on 100% annotations, while semi-supervised learning (SSL) methods used 50% annotated data, excluding unlabeled data. DeNAS-ViT consistently outperforms all methods across all metrics. On

Method	HMC-QU			CAMUS			CETUS		
	DSC↑	IoU↑	95HD↓	DSC↑	IoU↑	95HD↓	DSC↑	IoU↑	95HD↓
UNet++	$0.899_{(0.005)}*$	$0.898_{(0.004)}*$	$4.860_{(0.102)}*$	$0.919_{(0.006)}*$	$0.855_{(0.009)}*$	$6.584_{(0.578)}*$	$0.952_{(0.002)}*$	$0.968_{(0.001)}*$	$2.386_{(0.091)}*$
nnU-Net	$0.908_{(0.005)}*$	$0.907_{(0.005)}*$	$3.843_{(0.310)}*$	$0.922_{(0.003)}*$	$0.860_{(0.004)}*$	$6.075_{(0.412)}*$	$0.958_{(0.001)}*$	$0.970_{(0.001)}*$	$2.099_{(0.048)}*$
Transfuse	$0.903_{(0.004)}*$	$0.903_{(0.004)}*$	$4.304_{(0.228)}*$	$0.923_{(0.004)}*$	$0.861_{(0.006)}*$	$5.853_{(0.496)}$	$0.957_{(0.002)}*$	$0.971_{(0.001)}*$	$2.152_{(0.102)}*$
Auto-DeepLab	$0.908_{(0.004)}*$	$0.907_{(0.004)}*$	$3.857_{(0.164)}*$	$0.918_{(0.003)}*$	$0.851_{(0.006)}*$	$6.641_{(0.473)}*$	$0.954_{(0.004)}*$	$0.968_{(0.001)}*$	$2.278_{(0.133)}*$
M^{3}NAS	$0.910_{(0.007)}*$	$0.909_{(0.006)}*$	$3.709_{(0.288)}*$	$0.920_{(0.004)}*$	$0.856_{(0.004)}*$	$6.405_{(0.556)}*$	$0.956_{(0.005)}*$	$0.969_{(0.004)}$	$2.175_{(0.174)}*$
GeNAS	$0.913_{(0.004)}*$	$0.908_{(0.005)}*$	$3.748_{(0.219)}*$	$0.917_{(0.003)}*$	$0.852_{(0.003)}*$	$6.782_{(0.425)}*$	$0.949_{(0.002)}*$	$0.966_{(0.001)}*$	$3.267_{(0.119)}*$
URPC†	$0.892_{(0.004)}*$	$0.891_{(0.004)}*$	$5.355_{(0.125)}*$	$0.912_{(0.002)}*$	$0.842_{(0.002)}*$	$6.708_{(0.152)}*$	$0.941_{(0.004)}*$	$0.963_{(0.003)}*$	$3.374_{(0.185)}*$
CPS†	$0.878_{(0.005)}*$	$0.877_{(0.005)}*$	$6.908_{(0.296)}*$	$0.901_{(0.004)}*$	$0.827_{(0.005)}*$	$8.864_{(0.477)}*$	$0.923_{(0.005)}*$	$0.954_{(0.003)}*$	$4.306_{(0.282)}*$
CnT-B†	$0.911_{(0.006)}*$	$0.908_{(0.005)}*$	$4.102_{(0.496)}*$	$0.917_{(0.003)}*$	$0.847_{(0.002)}*$	$6.846_{(0.421)}*$	$0.943_{(0.004)}*$	$0.959_{(0.003)}*$	$3.488_{(0.341)}*$
ARCO-SG†	$0.908_{(0.003)}*$	$0.905_{(0.004)}*$	$3.912_{(0.203)}*$	$0.916_{(0.004)}*$	$0.850_{(0.003)}*$	$6.410_{(0.494)}*$	$0.948_{(0.005)}*$	$0.964_{(0.004)}*$	$3.125_{(0.282)}*$
Se^{2}NAS†	$0.907_{(0.004)}*$	$0.907_{(0.004)}*$	$3.941_{(0.303)}*$	$0.920_{(0.002)}*$	$0.856_{(0.003)}*$	$6.410_{(0.311)}*$	$0.955_{(0.003)}*$	$0.967_{(0.003)}*$	$2.235_{(0.111)}*$
SSHNN†	$0.906_{(0.002)}*$	$0.904_{(0.002)}*$	$4.011_{(0.145)}*$	$0.922_{(0.002)}*$	$0.859_{(0.003)}*$	$6.116_{(0.278)}*$	$0.949_{(0.002)}*$	$0.961_{(0.001)}*$	$3.053_{(0.087)}*$
DeNAS-ViT†	$\mathbf{0.933_{(0.002)}}$	$\mathbf{0.931_{(0.002)}}$	$\mathbf{2.480_{(0.161)}}$	$\mathbf{0.937_{(0.002)}}$	$\mathbf{0.884_{(0.003)}}$	$\mathbf{5.042_{(0.168)}}$	$\mathbf{0.972_{(0.001)}}$	$\mathbf{0.978_{(0.001)}}$	$\mathbf{1.620_{(0.036)}}$

Table 1: Comparison with SOTAs on public datasets. † : Note that SSL methods are tested under 50% annotations. The best and second-best results are highlighted in **bold** and underlined, respectively. * : $p < 0.01$ comparing against our method in each metric. Moreover, we test the generalization capability of DeNAS-ViT on additional datasets of other fields, shown in Sec. 4.4.

Type	Method	Usage Variants	DSC↑	IoU↑	95HD↓	Params (M)↓	FLOPs (G)↓
Manual	TransUNet	Employ Transformer as Encoder	$0.906_{(0.003)}*$	$0.905_{(0.004)}*$	$4.032_{(0.207)}*$	96.07	**48.34**
Manual	EfficientViT-L2	Multi-scale Tokens	$0.916_{(0.005)}*$	$0.915_{(0.004)}*$	$3.356_{(0.260)}*$	52.12	91.45
Manual	ViT-H Med-SAM	Employ Transformer as Encoder	$0.920_{(0.001)}*$	$0.918_{(0.001)}*$	$3.093_{(0.107)}*$	636.00	246.20
NAS	Auto-DeepLab	No Transformer is Applied	$0.908_{(0.004)}*$	$0.907_{(0.004)}*$	$3.857_{(0.164)}*$	44.42	347.52
NAS	SSHNN†	Treat Transformer as Additional Branch	$0.906_{(0.002)}*$	$0.904_{(0.002)}*$	$4.011_{(0.145)}*$	38.82	52.78
NAS	DAST	Treat Transformer as Candidate Operation	$0.915_{(0.002)}*$	$0.913_{(0.002)}*$	$3.438_{(0.119)}*$	192.44	110.36
NAS	**DeNAS-ViT-E†**	Treat EfficientViT as Candidate Operation	$0.923_{(0.001)}*$	$0.922_{(0.002)}*$	$2.884_{(0.123)}$	**38.48**	55.47
NAS	**DeNAS-ViT†**	Efficient NAS-ViT	$\mathbf{0.933_{(0.002)}}$	$\mathbf{0.931_{(0.002)}}$	$\mathbf{2.480_{(0.161)}}$	41.20	67.50

Table 2: Discussion of various Transformer usage variants. DeNAS-ViT achieves the highest accuracy while maintaining no increase in parameter size (Params) or computation complexity (FLOPs). † : SSL methods are tested under 50% annotations.

HMC-QU, it surpasses the previous SOTA (CnT-B) by 2.2% in DSC, 2.3% in IoU, and 1.622 in 95HD, while also demonstrating superior results on CAMUS and CETUS.

Comparison with ViT Usage Variants. Table 2 provides an evaluation of SOTA models utilizing different Transformer application strategies on the HMC-QU dataset. TransUNet and Med-SAM (Ma et al. 2024) utilize ViTs as their encoders. EfficientViT-L2 (Cai et al. 2023) incorporates fixed multi-scale tokens within the ViT architecture. Auto-DeepLab, a NAS model, does not leverage Transformers. SSHNN integrates ViTs as an auxiliary branch, while DAST employs ViTs as a candidate operation at the feature scale. In our DeNAS-ViT-E, we also explored a variant where EfficientViT was used as a candidate operation instead of the Efficient NAS-ViT. Our approach achieves the highest accuracy with a smaller model size and reduced FLOPs. Furthermore, the DeNAS-ViT-E variant demonstrates that the choice of Transformer application significantly influences performance, providing deeper insights into our design.

Search Cost. Fig. 4 compares the architecture search time of DeNAS-ViT with Transfuse (a high-performing handcrafted model, see Table 1) and other NAS frameworks on the HMC-QU dataset. For Transfuse, the search cost reflects its training time, while for NAS models, it encompasses both search and training durations, measured on an Nvidia A100 GPU. DeNAS-ViT requires 0.7 GPU days, with its search process being notably efficient and competitive compared to handcrafted models like Transfuse.

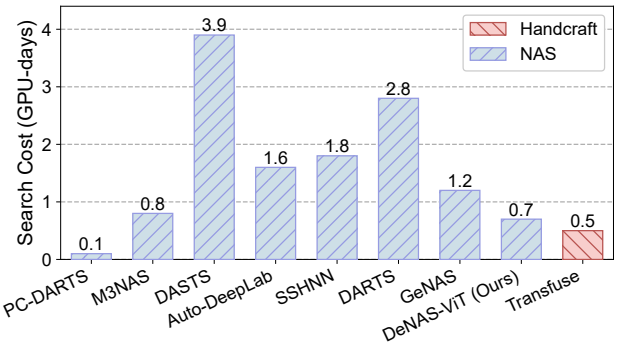

Figure 4: Search cost comparison with SOTA methods.

4.4 Robustness and Generalization

Robustness on Ultrasound Images. Fig. 5 illustrates the results of representative SSL networks and DeNAS-ViT with varying proportions of annotations on the CAMUS dataset. It can be observed that DeNAS-ViT performs more stably than other advanced SSL methods. When the annotation ratio descends from 50% to 5%, the IoU of CPS, CnT-B,

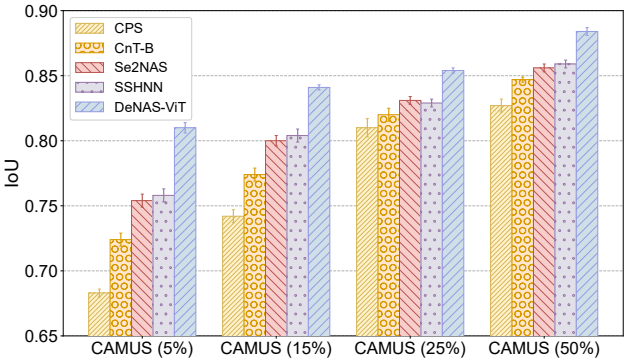

Figure 5: The impact of annotation proportions on SOTAs is evaluated using the CAMUS dataset, where DeNAS-ViT exhibits both robustness and effectiveness.

Se^{2}NAS, and SSHNN drops by 14.4%, 12.3%, 10.2%, and 10.1%, respectively. In contrast, DeNAS-ViT's IoU drops by only 7.4%, highlighting its robustness. This stability is attributed to its hybrid constraint-driven SSL, which improves data utilization.

Method	ISIC (10%)		ISIC (15%)	
	DSC↑	95HD↓	DSC↑	95HD↓
CPS	0.746	15.722	0.779	12.910
CnT-B	0.790 (+0.044)	11.879 (-3.843)	0.814 (+0.035)	9.583 (-3.327)
ARCO-SG	0.798 (+0.052)	11.352 (-4.370)	0.816 (+0.037)	9.600 (-3.310)
SSHNN	0.788 (+0.042)	12.064 (-3.658)	0.809 (+0.030)	10.119 (-2.791)
DeNAS-ViT	**0.817** (+0.071)	**9.024** (-6.698)	**0.840** (+0.061)	**7.391** (-5.519)

Table 3: Comparison with SOTA methods on the ISIC under varying annotation ratios to assess the generalizability.

Generalization Across Image Domains. To test the generalization capability of DeNAS-ViT, we conduct experiments on the International Skin Imaging Collaboration (ISIC) (Codella et al. 2018) dataset. As shown in Table 3, we evaluate our method against other SOTAs, including CPS, CnT-B, ARCO-SG, and SSHNN. Our method achieves the best performance in all semi-supervised settings (10% and 15%), with the improvements of DSC: 1.9%, 2.4%, and 95HD: -2.328mm, -2.209mm over the runner-up. Extensive results demonstrate the generalization capability of our proposed model. More experiments are shown in Sec. B.

4.5 Ablation Studies

Effects of Sub-modules. To evaluate the impact of sub-modules, ablation studies are conducted as shown in Tab. 4. The baseline model (No. 1) is the NAS backbone (Sec. 3.2) without the Efficient NAS-ViT module, trained on 50% labeled data. Adding the Efficient NAS-ViT module (No. 2) improves segmentation performance, increasing the DSC from 0.905 to 0.912, demonstrating enhanced context extraction capability. Introducing the co-training SSL framework (No. 6) further boosts performance by 0.7%, attributed to the additional information gained from unlabeled data. Incorporating network independence loss (No. 7) results in a 9.5% improvement in 95HD over No. 6. The effectiveness

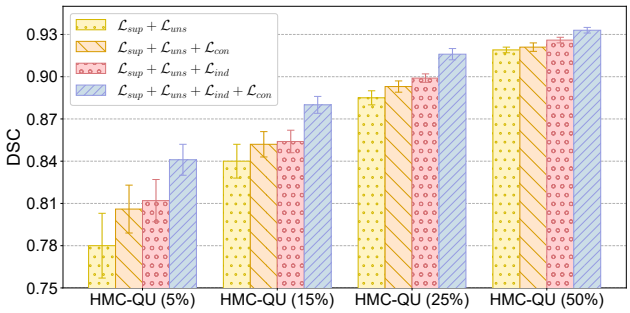

Figure 6: Sensitivity analysis of hybrid constraints. The case considering additional constraints shows better robustness.

of contrastive learning is validated in No. 5, which achieves a 0.7% higher DSC compared to using alone $\mathcal{L}_{uns}$ in No. 3.

No.	NAS	$\mathcal{L}_{uns}$	$\mathcal{L}_{ind}$	$\mathcal{L}_{con}$	DSC↑	IoU↑	95HD↓
1	✗	✗	✗	✗	$0.905_{(0.002)}$	$0.906_{(0.003)}$	$4.310_{(0.204)}$
2	✓	✗	✗	✗	$0.912_{(0.001)}$	$0.911_{(0.002)}$	$3.623_{(0.131)}$
3	✗	✓	✗	✗	$0.910_{(0.001)}$	$0.910_{(0.001)}$	$3.674_{(0.100)}$
4	✗	✓	✓	✗	$0.919_{(0.001)}$	$0.920_{(0.002)}$	$3.058_{(0.093)}$
5	✗	✓	✗	✓	$0.917_{(0.003)}$	$0.915_{(0.002)}$	$3.356_{(0.297)}$
6	✓	✓	✗	✗	$0.919_{(0.002)}$	$0.918_{(0.002)}$	$3.041_{(0.165)}$
7	✓	✓	✓	✗	$0.926_{(0.002)}$	$0.925_{(0.001)}$	$2.752_{(0.102)}$
8	✓	✓	✓	✓	$\mathbf{0.933}_{(0.002)}$	$\mathbf{0.931}_{(0.002)}$	$\mathbf{2.480}_{(0.161)}$

Table 4: Ablation studies of each component in the DeNAS-ViT structure on the HMC-QU dataset. "NAS" denotes our proposed Efficient NAS-ViT. More experiments on the other two datasets are shown in Appx. D.1.

Analysis of Hybrid Constraints. We investigate whether algorithmic independence loss $\mathcal{L}_{ind}$ or contrastive loss $\mathcal{L}_{con}$ would impact its performance under varying annotations on the HMC-QU, presented in Fig. 6. When both $\mathcal{L}_{ind}$ and $\mathcal{L}_{con}$ are not considered, there is a 13.9% drop in DSC as the annotation ratio decreases from 50% to 5%, which is larger than the 11.4% and 11.5% DSC decreases observed when $\mathcal{L}_{con}$ or $\mathcal{L}_{ind}$ is excluded, respectively. Notably, the DSC of DeNAS-ViT decreases by 9.2%, representing the smallest drop among these cases. Results show that (1) $\mathcal{L}_{ind}$ encourages the creation of complementary networks that assist in training by providing multi-view information from the same data; (2) $\mathcal{L}_{con}$ considers feature-level uncertainty where regions with lower uncertainty are filtered out, which is helpful for stable training under limited annotations.

5 Conclusion

In this paper, we propose a data efficient NAS-optimized ViT (DeNAS-ViT) to address two key challenges in ultrasound segmentation: multi-scale feature extraction and model robustness to limited data. For the first issue, DeNAS-ViT implements a three-level search, coupled with the Efficient NAS-ViT to enhance context extraction. For the second issue, we propose a NAS-based constraint-driven SSL for limited annotations. While experiments show DeNAS-ViT's effectiveness, our future direction is to enhance its zero-shot generalization capability across modalities.

References

Bai, W.; Oktay, O.; Sinclair, M.; Suzuki, H.; Rajchl, M.; Tarroni, G.; Glocker, B.; King, A.; Matthews, P. M.; and Rueckert, D. 2017. Semi-supervised learning for network-based cardiac MR image segmentation. In *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*, 253–260. Springer.

Bernard, O.; Bosch, J. G.; Heyde, B.; Alessandrini, M.; Barbosa, D.; Camarasu-Pop, S.; Cervenansky, F.; Valette, S.; Mirea, O.; Bernier, M.; et al. 2015. Standardized evaluation system for left ventricular segmentation algorithms in 3D echocardiography. *IEEE transactions on medical imaging*, 35(4): 967–977.

Cai, H.; Li, J.; Hu, M.; Gan, C.; and Han, S. 2023. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17302–17313.

Cao, X.; Chen, H.; Li, Y.; Peng, Y.; Zhou, Y.; Cheng, L.; Liu, T.; and Shen, D. 2022. Auto-DenseUNet: Searchable neural network architecture for mass segmentation in 3D automated breast ultrasound. *Medical image analysis*, 82: 102589.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021a. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, R.; Luo, J.; Nian, F.; Cen, Y.; Peng, Y.; and Yu, Z. 2024. SSHNN: Semi-Supervised Hybrid NAS Network for Echocardiographic Image Segmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1541–1545. IEEE.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021b. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2613–2622.

Codella, N. C.; Gutman, D.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Dusza, S. W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, 168–172. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, J.; Sun, Y.; Zhang, Q.; Li, Y.; Liu, W.; and Wang, X. 2020. Densely connected search space for more flexible neural architecture search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10628–10637.

Huang, H.; Huang, Y.; Xie, S.; Lin, L.; Ruofeng, T.; Chen, Y.-w.; Li, Y.; and Zheng, Y. 2023. Semi-Supervised Convolutional Vision Transformer with Bi-Level Uncertainty Estimation for Medical Image Segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5214–5222.

Huesmann, K.; Rodriguez, L. G.; Linsen, L.; and Risse, B. 2021. The impact of activation sparsity on overfitting in convolutional neural networks. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*, 130–145. Springer.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Jeong, J.; Yu, J.; Park, G.; Han, D.; and Yoo, Y. 2023. GeNAS: neural architecture search with better generalization. *arXiv preprint arXiv:2305.08611*.

Kiranyaz, S.; Degerli, A.; Hamid, T.; Mazhar, R.; Ahmed, R. E. F.; Abouhasera, R.; Zabihi, M.; Malik, J.; Hamila, R.; and Gabbouj, M. 2020. Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. *IEEE Access*, 8: 210301–210317.

Leclerc, S.; Smistad, E.; Pedrosa, J.; Østvik, A.; Cervenansky, F.; Espinosa, F.; Espeland, T.; Berg, E. A. R.; Jodoin, P.-M.; Grenier, T.; et al. 2019. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE transactions on medical imaging*, 38(9): 2198–2210.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, X.; Xiang, Y.; Zhang, L.; Yang, X.; Yan, Z.; and Yu, L. 2023. SAMUS: Adapting Segment Anything Model for Clinically-Friendly and Generalizable Ultrasound Image Segmentation. *arXiv preprint arXiv:2309.06824*.

Liu, C.; Chen, L.-C.; Schroff, F.; Adam, H.; Hua, W.; Yuille, A. L.; and Fei-Fei, L. 2019. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 82–92.

Liu, H.; Simonyan, K.; and Yang, Y. 2018. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*.

Lu, Z.; Xia, W.; Huang, Y.; Hou, M.; Chen, H.; Zhou, J.; Shan, H.; and Zhang, Y. 2022. M 3 NAS: Multi-scale and multi-level memory-efficient neural architecture search for low-dose CT denoising. *IEEE Transactions on Medical Imaging*, 42(3): 850–863.

Luo, X.; Liao, W.; Chen, J.; Song, T.; Chen, Y.; Zhang, S.; Chen, N.; Wang, G.; and Zhang, S. 2021. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, 318–329. Springer.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Miao, J.; Chen, C.; Liu, F.; Wei, H.; and Heng, P.-A. 2023. CauSSL: Causality-inspired Semi-supervised Learning for Medical Image Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21426–21437.

Pauletto, L.; Amini, M.-R.; and Winckler, N. 2022. Se 2 NAS: Self-Semi-Supervised architecture optimization for Semantic Segmentation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 54–60. IEEE.

Qian, J.; Li, R.; Yang, X.; Huang, Y.; Luo, M.; Lin, Z.; Hong, W.; Huang, R.; Fan, H.; Ni, D.; et al. 2022. Hasa: hybrid architecture search with aggregation strategy for echinococcosis classification and ovary segmentation in ultrasound images. *Expert Systems with Applications*, 202: 117242.

Qiao, S.; Shen, W.; Zhang, Z.; Wang, B.; and Yuille, A. 2018. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, 135–152.

Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Chen, X.; and Wang, X. 2021. A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys (CSUR)*, 54(4): 1–34.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Salehin, I.; Islam, M. S.; Saha, P.; Noman, S.; Tuni, A.; Hasan, M. M.; and Baten, M. A. 2024. AutoML: A systematic review on automated machine learning with neural architecture search. *Journal of Information and Intelligence*, 2(1): 52–81.

Song, X.; Xie, X.; Lv, Z.; Yen, G. G.; Ding, W.; Lv, J.; and Sun, Y. 2023. Efficient evaluation methods for neural architecture search: A survey. *arXiv preprint arXiv:2301.05919*.

Su, P.; Wang, K.; Zeng, X.; Tang, S.; Chen, D.; Qiu, D.; and Wang, X. 2020. Adapting object detectors with conditional domain normalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 403–419. Springer.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Wang, G.; Zhai, S.; Lasio, G.; Zhang, B.; Yi, B.; Chen, S.; Macvittie, T. J.; Metaxas, D.; Zhou, J.; and Zhang, S. 2021. Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung CT scans with multi-scale guided dense attention. *IEEE transactions on medical imaging*, 41(3): 531–542.

White, C.; Safari, M.; Sukthanker, R.; Ru, B.; Elsken, T.; Zela, A.; Dey, D.; and Hutter, F. 2023. Neural architecture search: Insights from 1000 papers. *arXiv preprint arXiv:2301.08727*.

Xia, Y.; Yang, D.; Yu, Z.; Liu, F.; Cai, J.; Yu, L.; Zhu, Z.; Xu, D.; Yuille, A.; and Roth, H. 2020. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical image analysis*, 65: 101766.

Xu, Y.; Xie, L.; Zhang, X.; Chen, X.; Qi, G.-J.; Tian, Q.; and Xiong, H. 2019. Pc-darts: Partial channel connections for memory-efficient architecture search. *arXiv preprint arXiv:1907.05737*.

Yang, D.; Xu, Z.; He, Y.; Nath, V.; Li, W.; Myronenko, A.; Hatamizadeh, A.; Zhao, C.; Roth, H. R.; and Xu, D. 2023. DAST: Differentiable Architecture Search with Transformer for 3D Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 747–756. Springer.

You, C.; Dai, W.; Min, Y.; Liu, F.; Clifton, D.; Zhou, S. K.; Staib, L.; and Duncan, J. 2024. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *Advances in neural information processing systems*, 36.

Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; and Heng, P.-A. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 605–613. Springer.

Yu, Z.; Lee, F.; and Chen, Q. 2023. HCT-net: hybrid CNN-transformer model based on a neural architecture search network for medical image segmentation. *Applied Intelligence*, 1–17.

Zhang, Y.; Liu, H.; and Hu, Q. 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, 14–24. Springer.

Zhou, G.-Q.; Zhang, W.-B.; Shi, Z.-Q.; Qi, Z.-R.; Wang, K.-N.; Song, H.; Yao, J.; and Chen, Y. 2023. DSANet: Dual-branch Shape-Aware Network for Echocardiography Segmentation in Apical Views. *IEEE Journal of Biomedical and Health Informatics*.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6): 1856–1867.