This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418, IEEE Transactions on Emerging Topics in Computational Intelligence

A Novel Hierarchical Generative Model for Semi-Supervised Semantic Segmentation of Biomedical Images

Lu Chai, Zidong Wang, *Fellow, IEEE*, Yuheng Shao, and Qinyuan Liu

***Abstract*—In biomedical vision research, a significant challenge is the limited availability of pixel-wise labeled data. Data augmentation has been identified as a solution to this issue through generating labeled dummy data. While enhancing model efficacy, semi-supervised learning methodologies have emerged as a promising alternative that allows models to train on a mix of limited labeled and larger unlabeled data sets, potentially marking a significant advancement in biomedical vision research. Drawing from the semi-supervised learning strategy, in this paper, a novel medical image segmentation model is presented that features a hierarchical architecture with an attention mechanism. This model disentangles the synthesis process of biomedical images by employing a tail two-branch generator for semantic mask synthesis, thereby excelling in handling medical images with imbalanced class characteristics. During inference, the k-means clustering algorithm processes feature maps from the generator by using the clustering outcome as the segmentation mask. Experimental results show that this approach preserves biomedical image details more accurately than synthesized semantic masks. Experiments on various datasets, including those for vestibular schwannoma, kidney, and skin cancer, demonstrate the proposed method's superiority over other generative-adversarial-network-based and semi-supervised segmentation methods in both distribution fitting and semantic segmentation performance.**

***Index Terms*—Generative Adversarial Network, Semi-supervised Learning, Hierarchical Architecture, Attention Mechanism, Biomedical Image Segmentation**

I. Introduction

Medical experts commonly advocate for surgical resection and monitoring as primary treatments for serious disorders. A critical preoperative step involves accurately marking the lesion on computed tomography (CT) or magnetic resonance imaging (MRI) scans. However, due to the exponential increase in biomedical imaging data, manually annotating a vast quantity of medical images accurately and promptly has become nearly unfeasible. Consequently, there is growing research interest in automatic segmentation using computer-aided techniques, which offer high efficiency in precisely segmenting medical images.

Traditional segmentation methods predominantly utilize image edge detection algorithms to distinguish target object borders, and these techniques were initially prevalent in medical image segmentation [2], [41]. However, their effectiveness is limited in accurately delineating lesion boundaries in biomedical images, often hindered by intricate and similar pixel values in adjacent areas. As a response, convolutional neural network (CNN)-based approaches have risen to prominence in various computer vision fields. These methods excel at autonomously extracting critical features from deep layers of images.

Recently, a range of CNN-based segmentation networks has been developed, which automatically segment tumors or organs from diverse medical images. These networks capitalize on the proficiency of CNNs in extracting detailed features from biomedical images [1], [53]. While these advancements significantly improve upon traditional techniques, the majority of current models still depend heavily on large, high-quality labeled datasets. Such reliance poses a challenge as labeled biomedical data are often limited due to privacy issues and the scarcity of experts for image annotation. This shortage complicates the training of segmentation models that perform well on varied and unseen datasets, thus intensifying the challenges in automatic biomedical image segmentation.

To tackle the challenge of labeled biomedical data scarcity, two primary strategies have been proposed. The first involves the generation of synthetic labeled data pairs using advanced deep learning models to enhance the original dataset [28], [61], [63], [66], thereby effectively increasing the amount of data available for training. The second strategy is the application of various semi-supervised learning (SSL) methods according to specific application needs. SSL methodologies integrate a substantial volume of unlabeled data into the training process, allowing the model to utilize the full dataset more effectively, and this results in a more robust model compared to traditional supervised methods that rely solely on labeled data. The SSL strategy is particularly valuable in situations where obtaining labeled data is costly, and the model's performance may be adversely affected by a lack of diverse and extensive training data. Popular SSL techniques include pseudo-labeling [7], [27], [55] and consistent regularization [13], [14], [40], [56]. Among others, contrastive learning, which focuses on training powerful image feature extractors [20], [51], [60] using unsupervised contrast loss from image transformations, has shown promising SSL results.

While SSL methods facilitate model training with limited labeled data, they do not explicitly simulate the input data distribution, which can lead to overfitting during training. To address this, in [9], a semi-supervised segmentation technique has been introduced by using StyleGAN2 designed to fit the

Lu Chai, Yuheng Shao and Qinyuan Liu are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China. Email: liuqy@tongji.edu.cn.

Zidong Wang is with the Department of Computer Science, Brunel University London, Uxbridge, Middlesex, UB8 3PH, United Kingdom. Email: Zidong.Wang@brunel.ac.uk

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418,

actual dataset distribution. Despite its state-of-the-art performance on various datasets, this method faces challenges with biomedical images containing small lesions. The key limitation is the need for a highly consistent representation of the image and its semantic mask in the latent space, which is a condition often not met in cases of small lesions in biomedical images.

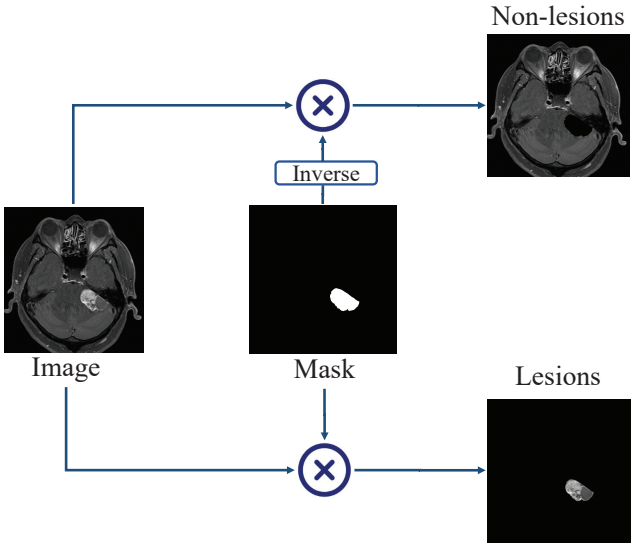

Fig. 1: Disentanglement of biomedical image according to the corresponding segmentation mask.

Addressing the issue of inadequate performance on small lesions in biomedical images, in this paper, we introduce a hierarchical architecture that integrates an attention mechanism into the generative model. Drawing from the principles of attention-GAN [12], [17], [58], [65], it's noted that an attention map, self-generated by the network, can significantly enhance the quality of generator synthesis. Building on this concept, we posit that in segmentation tasks, the semantic mask can function as a form of attention map. This approach is designed to focus the generative model more on lesion regions, which often occupy smaller areas in biomedical images. We achieve this by differentiating the biomedical image into lesion and non-lesion areas based on the semantic mask, as illustrated in Fig. 1.

Following the assumption that lesion textures and semantic masks possess analogous semantic representations, in [9], it has been suggested that images with similar semantics should correspond to similar representations in the latent space. Extending this idea, we introduce a tail two-branch generator derived from StyleGAN2 [46]. This generator is designed to produce both lesion textures and semantic masks with related semantics from the same latent representation. Concurrently, we utilize a standard StyleGAN2 [46] generator to create non-lesion textures from random noise. In this framework, the semantic mask is employed as an attention map. The final composite biomedical image is then formed by integrating the synthesized lesion texture, the semantic mask, and the non-lesion texture.

In the inference phase, we follow the method described in [9] by using an encoder with a ResNet [22] backbone to encode biomedical images into the targeted latent space. Subsequently, the corresponding semantic masks are generated using the trained Hierarchical Attention Generative Adversarial Network (HAGAN). However, due to the hierarchical design for image disentanglement, the grayscale mask from the tail two-branch generator often lacks detail compared to the desired segmentation mask. To address this, inspired by previous work [10], we apply the k-means clustering algorithm [44] to the feature maps from a specific layer of the tail two-branch generator during the inference stage. We use the results of this clustering as the final segmentation masks for the input images. Our experimental results show that these clustered feature maps retain more semantically meaningful details of the input images than the synthesized attention maps, leading to improved segmentation accuracy.

The proposed hierarchical attention generative model showcases significant potential in the accurate segmentation of biomedical images. Utilizing techniques in latent space mapping and attention map synthesis, the model achieves heightened accuracy in reconstructed images and their resultant segmentation masks. Further enhancement of accuracy and effectiveness is achieved through feature map extraction and clustering techniques. The primary contributions of this paper are summarized as follows.

1) Addressing the challenge of segmenting small lesions in medical images, we introduce a hierarchical architecture based on StyleGAN2 [46]. The integration of the attention mechanism in this approach aids in the semi-supervised segmentation of medical images with small lesions.
2) We develop a dual-branch generator within the generative model to simultaneously produce images with lesion textures and segmentation masks. This design enables more precise and controlled generation of medical images, specifically targeting lesion-specific characteristics.
3) To counter the potential loss of lesion boundary information in segmentation masks, K-means clustering is applied during the inference phase to refine the segmentation mask. Our experiments show that this method more effectively retains boundary details.
4) Experimental results indicate that our HAGAN surpasses existing semi-supervised segmentation methods, particularly in segmenting small lesions in medical images, thereby enhancing the accuracy of lesion boundary detection. By combining the data generation capability of the proposed method with consistency regularization-based semi-supervised segmentation methods, we achieve state-of-the-art performance in semi-supervised segmentation of biomedical images.

The structure of this paper is organized as follows. Section II provides an overview of existing literature on biomedical image segmentation and synthesis. Section III delves into the specifics of the proposed generative network, detailing its architecture and mechanisms. It also introduces the semi-supervised segmentation process, explaining how it integrates with the generative network. In Section IV, the effectiveness

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418, IEEE Transactions on Emerging Topics in Computational Intelligence

of our method is substantiated through both qualitative and quantitative experiments conducted on three different datasets. Finally, Section V rounds off the paper with a discussion summarizing the key findings and contributions, followed by concluding remarks that highlight the implications and potential future directions arising from this research.

II. Related Work

A. Semi-supervised Semantic Segmentation

To offset the high labeling costs of biomedical images, semi-supervised methods have emerged as a solution, which include adversarial learning [3], [36], [56] and the use of weak annotations [16], [29] for semantic segmentation. Furthermore, strategies like consistency regularization [13], [14], [40], [56] and self-supervised learning [20], [51], [60] have been applied in medical segmentation and detection to leverage unlabeled data effectively. Unsupervised tasks [5], [31], [38], [45] are designed to extract meaningful features from this data. For example, a previous study [9] has implemented a two-branch GAN for semi-supervised semantic segmentation by reconstructing images and masks from latent vectors. However, this method has fallen short for images with small lesions or organs. To overcome this shortcoming, we introduce a hierarchical GAN that separates the synthesis process, allowing the generator to concentrate on smaller lesions or organs, thereby enhancing segmentation accuracy for such biomedical images. Recently, consistency regularization-based semi-supervised methods [67], [68] have achieved impressive results in image segmentation. However, the number of labels still significantly affects the model's performance. Using reconstructed images and pseudo-labels obtained from semi-supervised generative models for data augmentation helps mitigate the impact of label scarcity on consistency regularization methods.

B. GAN-based Image Disentanglement

Disentangling the image components can markedly improve the synthesized image quality in GAN-based approaches, especially for images with class imbalances. A hierarchical GAN, inspired by [25], [62], offers a promising strategy by segregating the input image into foreground and background layers through a masking technique. Furthermore, a three-stage generative model for biomedical image disentanglement has been introduced in [63] to further underscore the potential of image disentanglement in enhancing synthesis quality. Motivated by these advancements, our work introduces a hierarchical architecture that employs dual generators to independently synthesize lesion and non-lesion textures, addressing the semantic segmentation challenges in biomedical images. This method leverages disentanglement to generate high-quality, balanced representations of biomedical images, thereby improving the efficiency and effectiveness of semantic segmentation processes.

C. GAN-based Image Synthesis

Generative Adversarial Networks (GANs) have become a cornerstone in image synthesis, transforming random noise vectors into realistic images. The foundational structure of a standard GAN, as introduced in [18], consists of a generator that crafts images from noise, and a discriminator that differentiates between these synthesized images and actual ones. Various adaptations, such as [18], [30], [46], [52], [57], [59], have been developed to generate high-quality images from latent vectors tailored to specific synthesis goals. Innovatively, [61] merged StyleGAN2 [46] with classifiers to create images paired with semantic labels, while [9] introduced a two-branch GAN for direct image segmentation. Building on this, [28] developed a three-stage GAN aimed at generating extensive biomedical data pairs. Drawing from these two-branch and disentanglement concepts, our research introduces a Hierarchical Attention Generative Adversarial Network. This network is designed for the semi-supervised segmentation of biomedical images, leveraging hierarchical structures and attention mechanisms to enhance segmentation performance.

D. Attention Mechanism for Image Synthesis

To enhance the realism of synthesized images, attention maps that assign weights to each pixel are employed. This concept has seen various implementations, including an additional attention network ([58]), a self-attention strategy ([65]), and local sparse attention layers ([12]). A recent advancement by [15] involves the use of image synthesis loss to create attention masks. Expanding on the use of attention mechanisms in image synthesis, we dissect the input images into lesion and non-lesion components. Our methodology merges the synthesized non-lesion textures with the product of foreground textures and the generated attention map to form the final image. This technique allows for the nuanced weighting of pixels, yielding images that are not only more realistic but also visually richer.

III. Method

This section starts with an overview of our method, highlighting its key enhancements. We then detail the model's architecture, training regimen, and the segmentation process, focusing on the novel features that improve segmentation accuracy.

A. Overview

Neural network-based semantic segmentation methods aim to identify the optimal function $f : \mathcal{X} \rightarrow \mathcal{Y}$, which maps images $x \in \mathcal{X}$ to their segmentation masks $y \in \mathcal{Y}$. The conventional optimization objective is to maximize the conditional probability $p(y|x)$. However, this approach may lead to overfitting in datasets with limited annotations.

In contrast, generative adversarial network (GAN)-based semantic segmentation methods seek the optimal generator $G(z) : \mathcal{Z} \rightarrow (\mathcal{X}, \mathcal{Y})$, which fits the joint distribution $p(x, y)$ from a random noise vector z, adhering to a standard normal distribution $p(z) = \mathcal{N}(0, 1)$. By considering the latent vector z as an embedded semantic representation, the optimized $G(z)$ can concurrently generate an image and its corresponding segmentation mask. Subsequently, an encoder can map the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418,

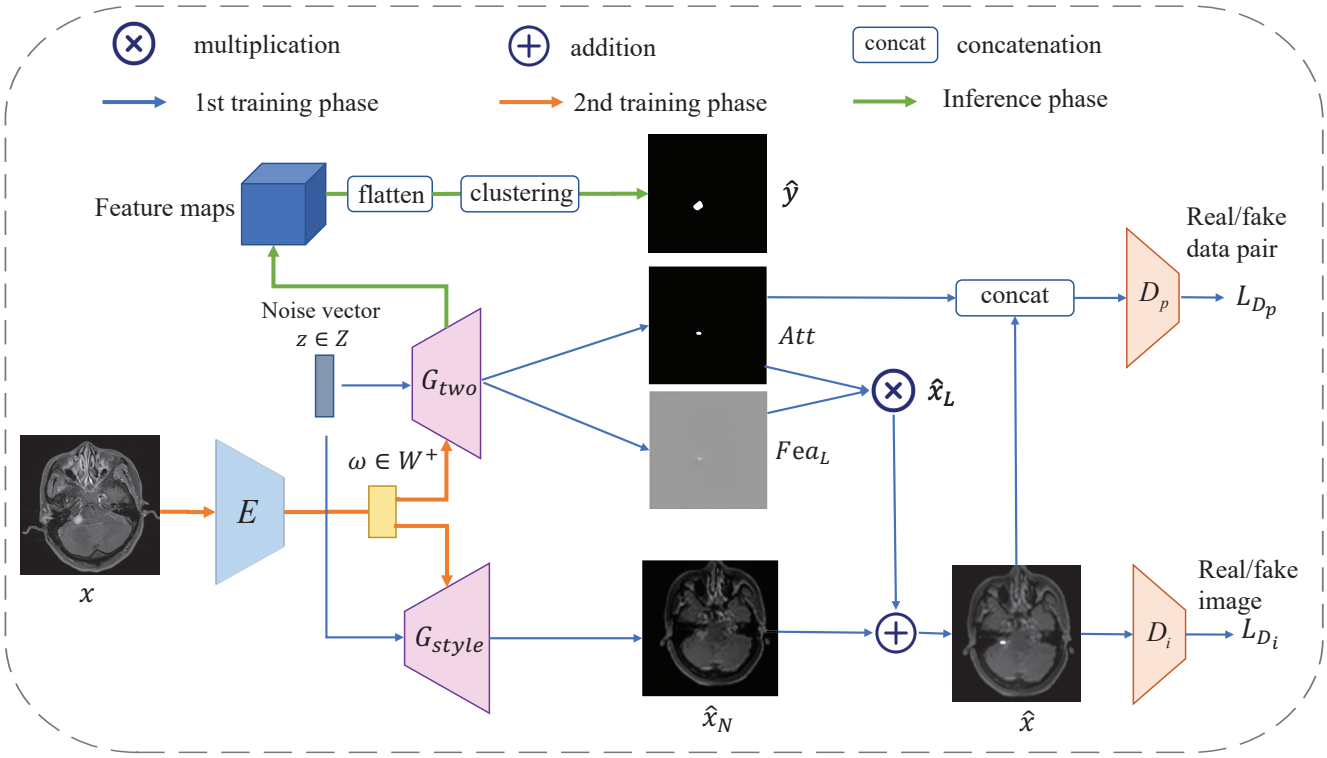

Fig. 2: **Model Structure of HAGAN**. The residual encoder maps image x into latent presentations ω in $\mathcal{W}^+$ space; Two generators reconstruct them into lesion textures Fea_L, attention maps Att and non-lesion textures $\hat{x}_N$; Feature maps from the seventh or ninth layer of G_{two} are clustered into 2 classes as final segmentation results.

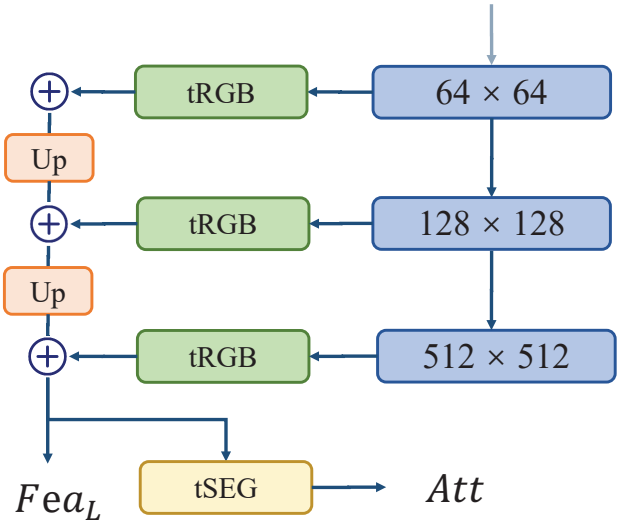

Fig. 3: Tail two-branch generator architecture

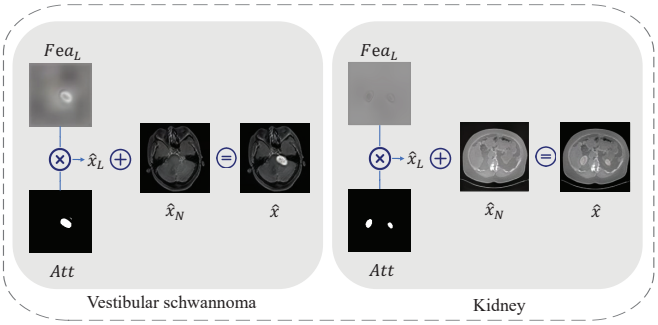

Fig. 4: Disentanglement of the biomedical images' synthesis process

image x to its latent space, allowing the generator to synthesize the segmentation masks $\hat{y}$.

Building on the above concept, a two-branch GAN has been introduced in [9] which, however, shows limited efficacy in class-imbalanced biomedical images. The fundamental premise of [9] is that the semantic mask and image representations in the low-dimensional latent space are similar, potentially leading to the omission of some semantic classes, especially those occupying smaller areas, during the synthesis process. To address this, especially for class-imbalanced images like those with small lesions, we segment the synthesis process into lesion image synthesis and non-lesion image synthesis. A tail two-branch generator $G_{two}(z) : \mathcal{Z} \rightarrow (Fea_L, Att)$ is devised to simultaneously synthesize the lesion textures Fea_L and the segmentation mask Att, with the latter serving as an attention map. The lesion images $\hat{x}_L$ are produced by the product of lesion textures and attention maps: $\hat{x}_L = Fea_L \times Att$. Concurrently, another generator $Gstyle(z) : \mathcal{Z} \rightarrow \mathcal{X}N$ generates the non-lesion image $\hat{x}N$. The final synthetic biomedical image $\hat{x}$ is a combination of the synthetic non-lesion image and the lesion image, derived from the product of the attention map and lesion texture: $\hat{x} = Fea_L \times Att + \hat{xN}$. This hierarchical approach aims

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418,

to fit the joint distribution of $p(x, y)$ through $Gtwo(z)$ and $G_{style}(z)$.

It is noted from experiments that the hierarchical architecture does not ideally separate biomedical images into lesions and non-lesions, often resulting in attention maps that depict lesions smaller than their actual size due to the loss of boundary detail. Due to the imbalance problem of lesions and non-lesions in biomedical images, our two branch generative model would lead to focus more on the generation of non-lesion areas during training, and the semantic mask generator G_{two} tend to generate smaller lesions. To improve this, we apply the k-means [44] clustering algorithm to the feature maps from the middle layer of G_{two}, using the clustered results as the final segmentation masks, which preserve more accurate boundary details than the synthesized attention maps.

The hierarchical attention GAN is less reliant on annotations compared to traditional CNN-based methods. Given that the loss functions for the generators and encoder include both supervised and unsupervised components, a minimal set of labeled images is adequate for HAGAN to perform semi-supervised segmentation of input images.

B. Model Structure

Our model leverages StyleGAN2 [46] to generate realistic images and their corresponding semantic masks, capitalizing on its capability to produce high-quality images from noise vectors. The generator accepts random noise vectors, adhering to a normal distribution, and initially maps them into a more complex space, often referred to as $\mathcal{W}$ space [49]. Through the application of the transformed noise vector to the generator's primary style layers, StyleGAN2 is able to produce realistic images.

Illustrated in Fig. 2, our model architecture introduces a hierarchical attention generative adversarial network tailored for biomedical imaging. A residual encoder maps the image x into latent representations ω within the $\mathcal{W}^+$ space. Two generators are employed to create biomedical images and their associated attention maps from the latent codes. The definitive segmentation masks are derived by clustering feature maps extracted from specific layers, either the seventh or ninth, of the tail two-branch generator G_{two}. The subsequent sections provide an in-depth explanation of the structure and functionality of each component.

Generators: Our model incorporates two generators. The first is a tail two-branch generator for lesion texture synthesis, and the second focuses on non-lesion synthesis. Drawing from StyleGAN2's residual skip-connection design, the tail two-branch generator (Fig. 3) ensures consistency between attention maps $\mathcal{A}$⊔⊔ and lesion images $\S_{\mathcal{L}}$ by sharing style convolutional layers. This generator, defined as $G_{two} : \mathcal{Z} \rightarrow (Fea_L, Att)$, takes noise vectors $z \in \mathcal{Z}$, following $p(z) = N(0, 1)$, to generate attention maps Att and lesion images $\hat{x}_L \in \hat{\mathcal{X}}_L$. We use convolutional layers named tRGB to reshape intermediate feature layers from $n * m * x$ to $n * m * t * 3$ where $x = t * 3$. By stacking tRGB results of different depths, the dual-branch generator G_{two} can extract lesion textures stored in RGB shape. Additionally, since the lesion textures are highly consistent with the lesion semantic mask, binarizing the stacked texture features through the tSEG layer can produce the corresponding semantic mask. The second generator, following StyleGAN2's design, synthesizes non-lesion images $\hat{x}_N \in \hat{\mathcal{X}}N$ from the same noise vectors and is defined as $Gstyle : \mathcal{Z} \rightarrow \hat{\mathcal{X}}_N$.

Discriminators: Two discriminators are employed: an image discriminator, $D_I : \mathcal{X} \rightarrow \mathbb{R}$, using a residual architecture to assess image realism, and a pair discriminator, $D_P : (\mathcal{X}, Att) \rightarrow \mathbb{R}$, evaluating the authenticity of image-mask pairs, thereby ensuring their consistency.

Encoders: An encoder, defined as $E : \mathcal{X} \rightarrow \mathcal{W}^+$, maps images into latent representations in $\mathcal{W}$ space, adopting a residual network design. This facilitates the use of a higher-dimensional space $\mathcal{W}^+$ for the StyleGAN-based generator, improving synthesis quality.

Cluster: The k-means clustering algorithm [44] is applied to feature maps from the tail two-branch generator's intermediate layer, using the clustering results as segmentation masks. This method, focusing on the feature maps from the seventh or ninth layer, preserves more accurate boundary information by avoiding detail loss during normalization and binarization.

Hierarchical Architecture: The hierarchical architecture enables focused synthesis on lesion and non-lesion areas. It addresses the limitations of single generators in synthesizing detailed segmentation masks and realistic images simultaneously, especially for small lesions. The synthesized image $\hat{x} = \hat{x}_L + \hat{x}_N$ combines the outputs of both generators, with discriminator scores guiding their training for improved quality. This architecture effectively disentangles the synthesis process, allowing for the concurrent generation of lesion images with semantic masks and their integration with non-lesion images to form complete biomedical images (Fig. 4).

C. Training Process

The training process is divided into two main phases: i) the first phase involves training the generators and discriminators to create biomedical images and attention maps from random noise vectors; and ii) in the second phase, the encoder is trained to map input images to latent representations in $\mathcal{W}^+$ space, using the previously trained generators to reconstruct the images and their semantic masks.

The dataset used for training, denoted as $F = \{S, U\}$, comprises both labeled data S and unlabeled data U. The labeled data, used for supervised learning, is represented as $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, whereas the unlabeled data is denoted as $\mathcal{U} = \{x_1, x_2, ..., x_n\}$, with $m \ll n$ indicating that the number of labeled instances is much smaller than the number of unlabeled instances.

1st Phase: The objective functions of generators $G = \{G_{two}, G_{style}\}$ and discriminators D_I, D_P are given as fol-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418,

lows.

$$\mathcal{L}_G = E_{(\hat{x},\cdot)=G(z)}[log(1 - D_I(\hat{x})] + E_{(\hat{x},Att)=G(z)}[log(1 - D_P(\hat{x}, Att)] \quad (1)$$

$$\mathcal{L}_{D_I} = E_{x\sim F}[log(D_I(x))] + E_{(\hat{x},\cdot)=G(z)}[log(1 - D_I(\hat{x})] \quad (2)$$

$$\mathcal{L}_{D_P} = E_{(x,y)\sim S}[log(D_P(x, y))] + E_{(\hat{x},Att)=G(z)}[log(1 - D_P(\hat{x}, Att)] \quad (3)$$

The learning objectives for the discriminators D_I and D_P involve maximizing $\mathcal{L}D_I$ and $\mathcal{L}D_P$, respectively. Concurrently, the generators G_{two} and G_{style} are collaboratively trained to minimize the combined loss $\mathcal{L}G$. Given that the synthesized image $\hat{x}$ consists of lesion parts $\hat{x}_L$ and non-lesion parts $\hat{x}N$, the first component of Eq. (3) ensures that D_I provides feedback to both G_{two} and G_{style}, encouraging the synthesis of more realistic textures. Additionally, the second component of Eq. (3) reflects the evaluation of how authentic and consistent the synthesized images and segmentation masks appear. Through the use of labeled data S, D_P incentivizes G_{two} to enhance its focus on lesion areas by assessing the congruence between the synthetic image and its attention map against the real image and its semantic mask. Consequently, the supervised loss component further encourages G_{two} to concentrate on lesion areas, while G_{style} is naturally directed towards synthesizing the non-lesion portions of the images. This approach ensures that the semantic details of lesions are preserved throughout the synthesis process, regardless of the size of the lesion areas.

2nd Phase: The encoder's objective function is a combination of supervised and unsupervised reconstruction losses formulated as:

$$L_E = L_S + L_U \quad (4)$$

where the reconstructed image and synthesized attention maps are given by $(\hat{x}, Att) = G(E(x))$. The supervised reconstruction loss of semantic masks L_s is defined as:

$$L_s = CE(y, Att) + DICE(y, Att) \quad (5)$$

with y being the semantic segmentation label for the input image x, $CE(\cdot)$ representing the pixel-wise cross-entropy, and $DICE(\cdot)$ denoting the dice loss as introduced in [11].

The unsupervised loss $\mathcal{L}_u$ is defined as:

$$L_u = LPIPS(x, \hat{x}) + \lambda\|x - \hat{x}\|^2 \quad (6)$$

Here, $LPIPS(\cdot)$ is the "Learned Perceptual Image Patch Similarity" metric [43], and the second term is a weighted L_2 norm loss.

After completing the second phase of training, the encoder is capable of mapping input images into latent representations ω in $\mathcal{W}^+$ space and reconstructing them into the corresponding realistic images $\hat{x}$ and attention maps Att.

Semi-supervised Learning: In our semi-supervised learning framework, we leverage both labeled and unlabeled biomedical images to provide gradient feedback to the generators G_{two} and G_{style}. The supervised loss component, represented as the second term in Eq. (3), encourages G_{two} to generate attention masks Att that closely resemble the distribution of real segmentation masks $\mathcal{Y}$. This process enables G_{two} to produce attention maps that accurately reflect the small lesion areas present in the real masks, based on the embedded latent representations ω.

The availability of a limited number of annotated data pairs biases G_{two} towards generating masks with smaller lesion areas. Conversely, the inclusion of unlabeled images compels G_{style} to focus on synthesizing the background regions by employing an unsupervised reconstruction loss, indicated as the first term in Eq. (3). Subsequently, by mapping the input image into latent semantic presentations within the $\mathcal{W}^+$ space using the residual encoder, G_{two} is capable of producing attention maps that align with the semantic content of the input images, thus facilitating effective semi-supervised learning.

D. Inference Phase

Given that segmentation masks are conceptualized as attention maps Att, they should not be directly employed as final semantic masks. Due to the imbalance problem of lesions and non-lesions in biomedical images, our two-branch generative model would lead to focus more on the generation of non-lesion areas during training, and the semantic mask generator G_{two} tends to generate smaller lesions. To alleviate this, we apply a cluster to the feature maps from the seventh or ninth layer from the middle layer of G_{two}. Compared to other clustering algorithms, k-means has a significant advantage in separating clusters with regular shapes and relatively symmetric convex forms, making it more suitable for lesion separation scenarios. The k-means clustering method, focusing on the feature maps from the seventh or ninth layer, preserves more accurate boundary information by avoiding detail loss during normalization and binarization. The objective function for the k-means algorithm, aimed at minimizing clustering error, is defined as:

$$J = \sum_{i=1}^{C}\sum_{j=1}^{N} r_{ij}\lambda||x_j - \mu_i||^2 \quad (7)$$

Here, x_j represents the data point from the flattened feature maps of size 256 or 512, and $r_{ij} \in 0, 1$ denotes the cluster assignment of data point x_j. The resultant clusters yield the final semantic segmentation masks, enriched with detailed information.

The segmentation workflow named Ours-seg in HAGAN's inference phase unfolds as follows:

1) *Step 1:* The residual encoder E transforms the input image into a latent representation ω within the $\mathcal{W}^+$ space.
2) *Step 2:* The tail two-branch generator G_{two} generates the attention map Att and lesion textures Fea_L, with feature maps being extracted from either the seventh or ninth layer during this process.
3) *Step 3:* The k-means clustering algorithm, utilizing two clusters, is applied to the aforementioned flattened feature maps. The clustering outcome serves as the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418, IEEE Transactions on Emerging Topics in Computational Intelligence

definitive semantic segmentation mask for the input image.

The data generation workflow named Ours-gen in HA-GAN's inference phase unfolds as follows:

1) *Step 1:* Randomly sample a noise vector z in the $\mathcal{Z}$ space.
2) *Step 2:* The generator module generates realistic image $\hat{x}$ and lesion textures Fea_L, with feature maps being extracted from either the seventh or ninth layer during this process.
3) *Step 3:* The k-means clustering algorithm, utilizing two clusters, is applied to the aforementioned flattened feature maps. The clustering outcome serves as the definitive semantic segmentation mask for the generated image $\hat{x}$.

IV. Experiments and Results

In this evaluation section, our semi-supervised segmentation approach is assessed across three biomedical datasets, employing varying quantities of annotated data during the training phase. Initially, the synthesis quality of our generative model is compared against numerous leading GANs to gauge its efficacy in replicating biomedical image distributions. Both qualitative and quantitative analyses indicate that our generative model adeptly conforms to the biomedical image distributions.

Subsequent to the synthesis evaluation, the segmentation capability of our model is benchmarked. Here, the performance is juxtaposed with that of Unet [35], along with other semi-supervised segmentation techniques. The outcomes of these experiments underscore the effectiveness of our method in the semi-supervised generation and segmentation domain of biomedical imagery. By utilizing the data generation workflow mentioned above, HAGAN can be used for data augmentation with the latest advanced semi-supervised segmentation models, reducing the impact of label scarcity and further achieving state-of-the-art semi-supervised segmentation results.

A. Experimental Settings

Dataset: In our experiments, we used three datasets to validate the effectiveness of our method: a vestibular schwannoma dataset (VS), a kidney dataset (Kits19 [33]), and a skin cancer dataset (ISIC2018 [39]), among which Kits19 [34] and ISIC2018 [39] are representative benchmark datasets for biomedical image segmentation. The vestibular schwannoma and kidney datasets consist of slices from 3D medical images. We retained 900 slices from vestibular schwannoma and 10,000 slices from the kidney dataset, discarding slices that did not contain valid lesion regions. The skin cancer dataset consists of 2D scanned images in PNG format, from which we randomly selected 30,000 images, including 2,500 with segmentation labels. All images across these datasets are resized to a uniform resolution of 128 × 128 for consistency in our experiments.

Settings: To assess the efficacy of our semi-supervised learning approach, we employ varying quantities of labeled data to create different experimental setups. The division of data into training, validation, and testing sets for each dataset adheres to a 7:1:2 ratio, ensuring a balanced distribution for comprehensive performance evaluation. Experiments were performed on one PC with a single RTX3080 GPU and another PC with two RTX3090 GPUs. As for 128 × 128 resolution images, the proposed method requires approximately 5.8 GFLOPs and contains about 30 million parameters.

Hyperparameters: During the training process, there are hyperparameters as shown in TABLE 1.

Hyperparameters	Value
Batchsize	4
Resolution	128 × 128
Optimizer	Adam
Learning rate	0.001
Epoch	120

TABLE 1: Hyperparameters.

Generative Models	Inception Score	Frechet Inception Distance
Pix2pixGAN [37]	1.18 ± 0.01	74.35
SPADE [47]	1.09 ± 0.02	88.72
NICE-GAN [42]	1.12 ± 0.01	159.54
DPGAN [28]	1.21 ± 0.01	69.42
StyleGAN2 [46]	$\mathbf{1.22 \pm 0.01}$	65.48
Ours	1.21 ± 0.01	65.42
Ours-Plus	$\mathbf{1.22 \pm 0.01}$	**62.38**

TABLE 2: Synthesis quality of several GANs using 315 labeled vestibular schwannoma data pairs to perform training. Ours-Plus also utilizes an additional 315 unlabeled images for training. IS (↑ better) and FID (↓ better).

B. Synthesis Results of GANs

Baselines. For synthesis quality, we compare our method with several state-of-the-art GANs including Pix2pixGAN [37], SPADE [47], NICE-GAN [42], StyleGAN2 [46] and DPGAN [28]. The implementation codes of these methods are cloned from their public repositories. The above generation methods do not have significant differences in model size and training mode. When setting the batch size to 4 the image resolution is 128x128, they all converge within 6-8 hours.

Evaluation Metrics. The quality of synthesized data pairs is evaluated by IS [48] and FID [32]. Specifically, IS measures the clarity and diversity of the data using Kullback-Leibler divergence, whereas FID focuses on the similarity between the synthesized and the real images. The higher the IS (or the lower the FID), the better the synthesis quality.

Generative Models	Inception Score	Frechet Inception Distance
Pix2pixGAN [37]	1.20 ± 0.01	55.47
SPADE [47]	1.17 ± 0.01	90.04
NICE-GAN [42]	1.14 ± 0.01	203.05
DPGAN [28]	1.19 ± 0.01	61.71
StyleGAN2 [46]	1.20 ± 0.01	65.74
Ours	1.21 ± 0.01	56.85
Ours-Plus	$\mathbf{1.22 \pm 0.01}$	**54.59**

TABLE 3: Synthesis quality of several GANs using 300 annotated KiTS19 data pairs to perform training. Ours-Plus also utilizes an additional 10000 unlabeled images for training. IS (↑ better) and FID (↓ better).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418, IEEE Transactions on Emerging Topics in Computational Intelligence

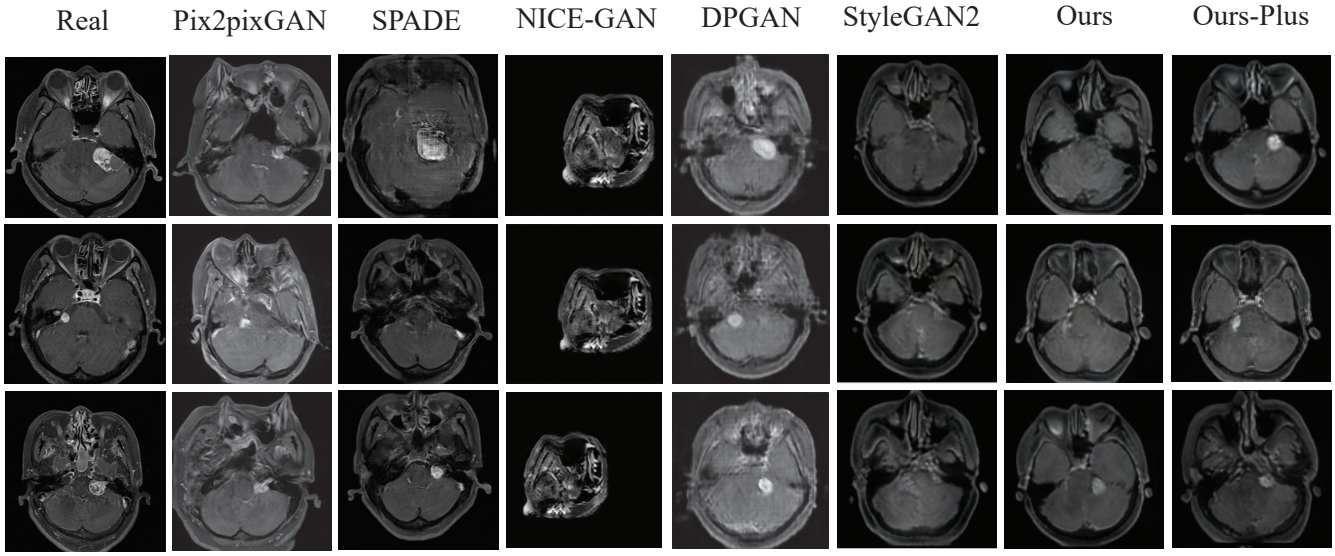

Fig. 5: Comparison of real and fake vestibular schwannoma images synthesized by different generative models. Ours-Plus also utilizes an additional 315 unlabeled images for training.

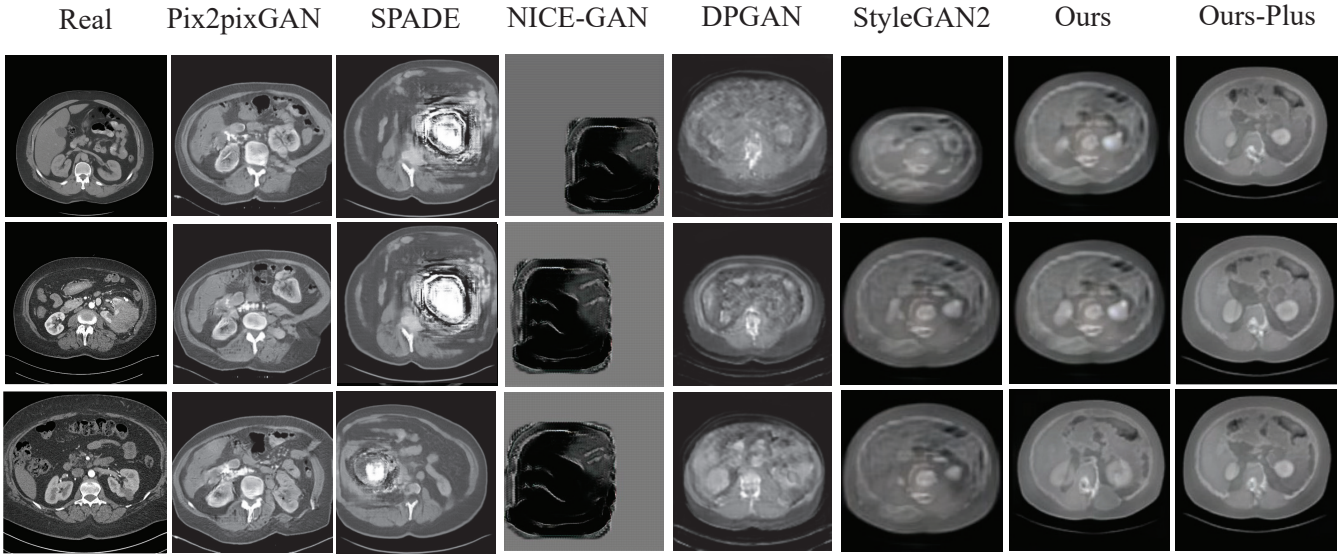

Fig. 6: Comparison of real and fake KiTS images synthesized by different generative models. Ours-Plus also utilizes an additional 5323 unlabeled images for training.

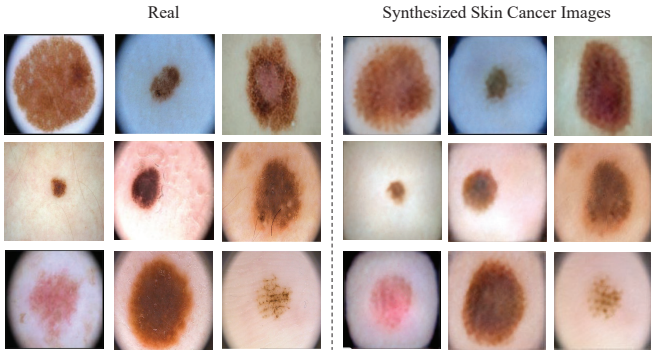

Fig. 7: Comparison of real and synthesized skin cancer images training with 200 annotated data pairs and 29800 unlabeled images.

As shown in TABLE 2 and TABLE 3, we utilize part of annotated data pairs to train the generative models. The quantitative results present that StyleGAN2 has better performance in the synthesis of biomedical images (Note: the proposed HAGAN is also constructed based on StyleGAN2). Further, in order to validate the semi-supervised learning strategy of the proposed method, we apply Ours-Plus which also applies amounts of unlabeled data to train the HAGAN. Combining the qualitative experimental results presented in Fig. 5 and Fig. 6, we find that Ours-Plus has a higher definition and more visible lesion areas than Ours. This further proves that HAGAN has the capacity to disentangle the biomedical image and extract expected features. Moreover, we also apply HAGAN to the ISIC2018 data set to evaluate HAGAN on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418, IEEE Transactions on Emerging Topics in Computational Intelligence

	20 labeled	80 labeled	315 labeled
Unet [35]	0.6216	0.7912	**0.8154**
MT [4]	0.5580	0.6369	0.6703
AdvSSL [54]	0.5735	0.6750	0.7059
GCT [64]	0.5817	0.7158	0.7249
SemanticGAN [9]	0.5042	0.6651	0.6912
Ours-seg	0.6252	0.7940	0.8098
UniMatch [67]	0.7052	0.8297	0.8564
ARCO [68]	0.8024	0.8580	0.8782
Ours-gen + UniMatch	0.8320	0.8324	0.8749
Ours-gen + ARCO	**0.8519**	**0.8641**	**0.8795**

TABLE 4: Segmentation results of vestibular schwannoma using different numbers of annotated data evaluated with the dice score.

	50 labeled	150 labeled	300 labeled
Unet [35]	0.3254	0.3562	0.4213
MT [4]	0.3616	0.4083	0.4846
AdvSSL [54]	0.3849	0.4260	0.5077
GCT [64]	0.4131	0.4686	0.5591
SemanticGAN [9]	0.3915	0.4526	0.5238
Ours-seg	0.4173	0.5087	0.6062
UniMatch [67]	0.6699	0.7330	0.7358
ARCO [68]	0.7424	0.7486	0.7536
Ours-gen + Unimatch [67]	0.7350	0.7494	0.7519
Ours-gen + ARCO [68]	**0.8071**	**0.8645**	**0.8433**

TABLE 5: Segmentation results of KiTS19 images using different scale annotated images evaluated with Dice.

biomedical data sets with big lesions. We use 200 annotated data pairs and 29800 unlabeled images, obtain IS = 2.04±0.01, and FID = 34.15. Samples in Fig. 7 indicate that HAGAN can synthesize detailed textures of skin cancer images.

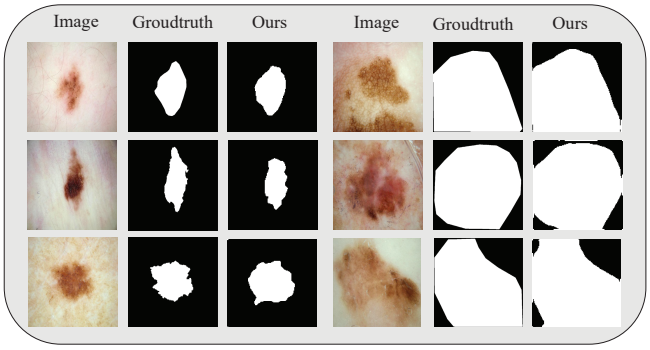

Fig. 8: **Segmentation samples of ISIC images**. We use 200 labeled data pairs and 29800 unlabeled images to perform training.

C. Semi-Supervised Segmentation Results

Baselines: For semi-supervised segmentation of biomedical images, we compare our method with both Unet [35] and several semi-supervised segmentation methods to evaluate the proposed strategy. Unet is trained only with annotated data while semi-supervised methods utilize unlabeled data also. As semi-supervised methods, we use the mean teacher model with transformation-consistency (MT), the adversarial training-based method (AdvSSL) [54], Guided Collaborative Training (GCT) [64], AN-based semi-supervised segmentation method (SemanticGAN) [9], and advanced consistency normalization based methods ARCO [68] and Unimatch [67]. Some of the implementations are adapted from PixelSSL[1] and their public code repositories. Among the above methods, when the batch size is set to 4 and the image resolution is 128x128, only Unet can achieve convergence within 2 hours, while the other methods require 6-8 hours to reach convergence.

	40 labeled	200 labeled	2000 labeled
Unet [35]	0.4935	0.6041	0.6469
MT [4]	0.5200	0.7052	0.7741
AdvSSL [54]	0.5016	0.6657	0.7388
GCT [64]	0.4759	0.6814	0.7887
SemanticGAN [9]	0.7144	0.7555	0.7890
Ours-seg	**0.7204**	**0.7631**	**0.8071**

TABLE 6: Segmentation results of ISIC skin lesion images using different scale annotated images evaluated with the Jaccard score.

Evaluation Metrics. The segmentation performance is evaluated with two metrics. Let n be a positive integer, given two image sets $P = \{p_i|\ 0 < i \leq n\}$ and $Q = \{q_i|\ 0 < i \leq n\}$. The Dice coefficient is defined as $\frac{1}{n}\sum_{i=1}^{n}\frac{2|p_i \cap q_i|}{|p_i|+|q_i|}$ to evaluate the similarity of sets, and the Jaccard index is $\frac{1}{n}\sum_{i=1}^{n}\frac{|p_i \cap q_i|}{|p_i|+|q_i|-|p_i \cap q_i|}$ to measure the similarity and diversity of sets.

Vestibular Schwannoma Segmentation. As shown in TABLE 4, we apply Unet to perform supervised segmentation and compare it with several semi-supervised networks. Considering that the method proposed in this paper has both segmentation and data generation capabilities, we compare the segmentation results obtained using the segmentation dataflow (Ours-seg) of HAGAN, with the results obtained using the generation dataflow (Ours-gen) for data augmentation in consistency normalization-based methods [67], [68]. The results show that using the generation dataflow for data augmentation yields the best semi-supervised segmentation results, further demonstrating the effectiveness of the proposed method in fitting distributions of biomedical images and corresponding masks.

Kidney Segmentation. Segmentation results in TABLE 5 and Fig 10 demonstrate that combining GAN-based semi-supervised generation with consistency regularization-based semi-supervised learning can achieve better segmentation results. Although the segmentation performance of Ours-seg is weaker than the consistency normalization-based methods Unimatch [67] and ARCO [68], when our method is applied as a generative method called Ours-gen and the generated medical images and corresponding pseudo-labels are used as augmented labeled data to ARCO and Unimatch, the segmentation metrics of both methods showed improvement across different dataset scales. This further demonstrates that the proposed method can better learn the distribution of medical images and their corresponding masks. This demonstrates that combining GAN-based semi-supervised learning methods and consistency normalization-based semi-supervised learning strategies could be a valuable future direction arising from this research.

[1] https://github.com/ZHKKKe/PixelSSL

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418,

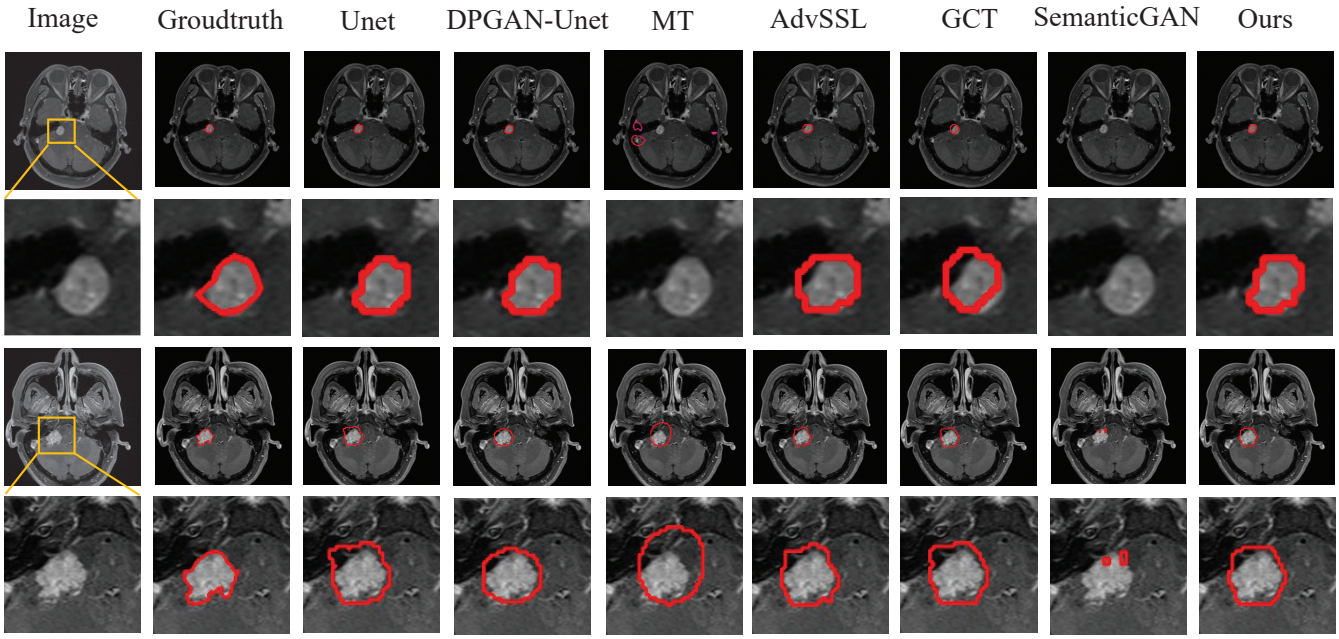

Fig. 9: **Segmentation samples of vestibular schwannoma images**. We use 80 labeled data pairs and 550 unlabeled images to perform training.

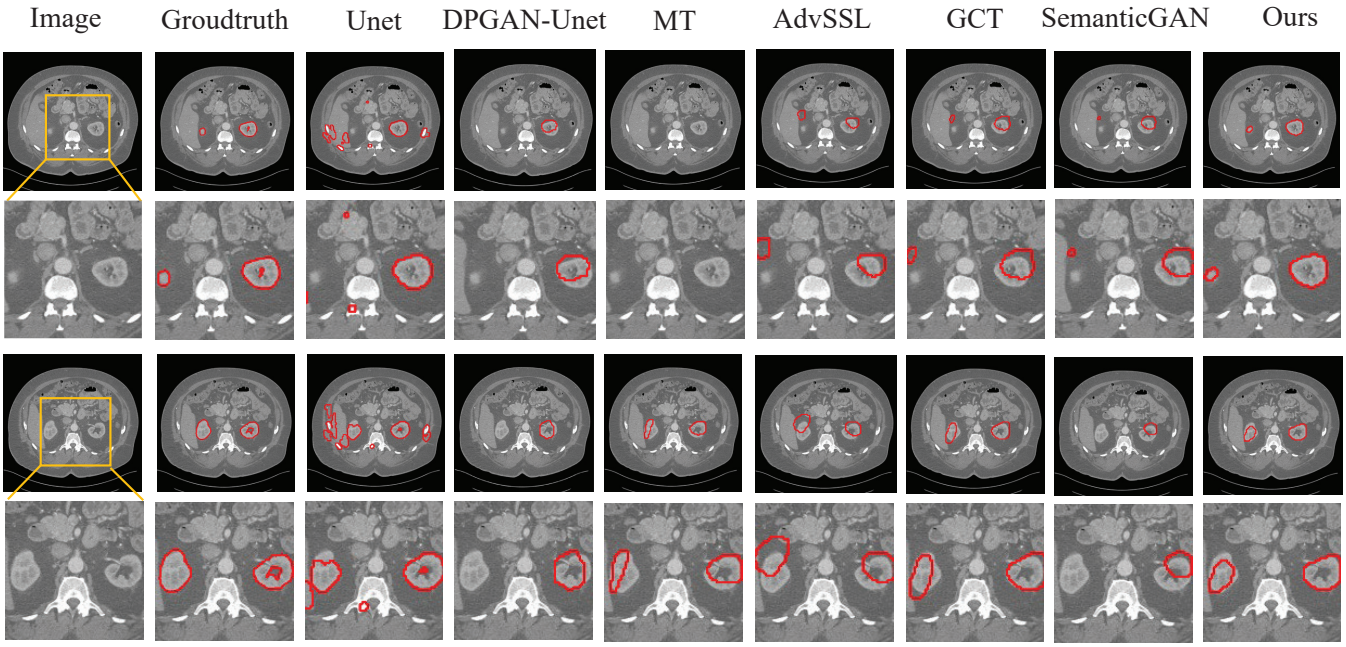

Fig. 10: **Segmentation samples of KiTS19 images**. We use 300 labeled data pairs and 10000 unlabeled images to perform training.

Skin Lesion Segmentation. To evaluate the proposed method on biomedical datasets with large lesions, we also apply our method to ISIC2018 [39] skin lesion data set. Results shown in TABLE 6 and Fig 8 show that the hierarchical structure can slightly improve the segmentation performance of images with large lesion areas. The final segmentation semantic masks are obtained by applying the two-category k-means clustering algorithm on the ninth layer feature maps of G_{two}.

	Vestibular Schwannoma	KiTS19	ISIC2018
Ours–no clustering	0.6094	0.6047	0.7263
Ours	**0.6252**	**0.6062**	**0.7631**

TABLE 7: Segmentation score of using attention maps or clustered feature maps as the final semantic segmentation masks.

D. Ablation Analysis

In the inference phase, we apply two k-means clusters to the feature maps extracted from certain layers of G_{two}. Comparing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418,

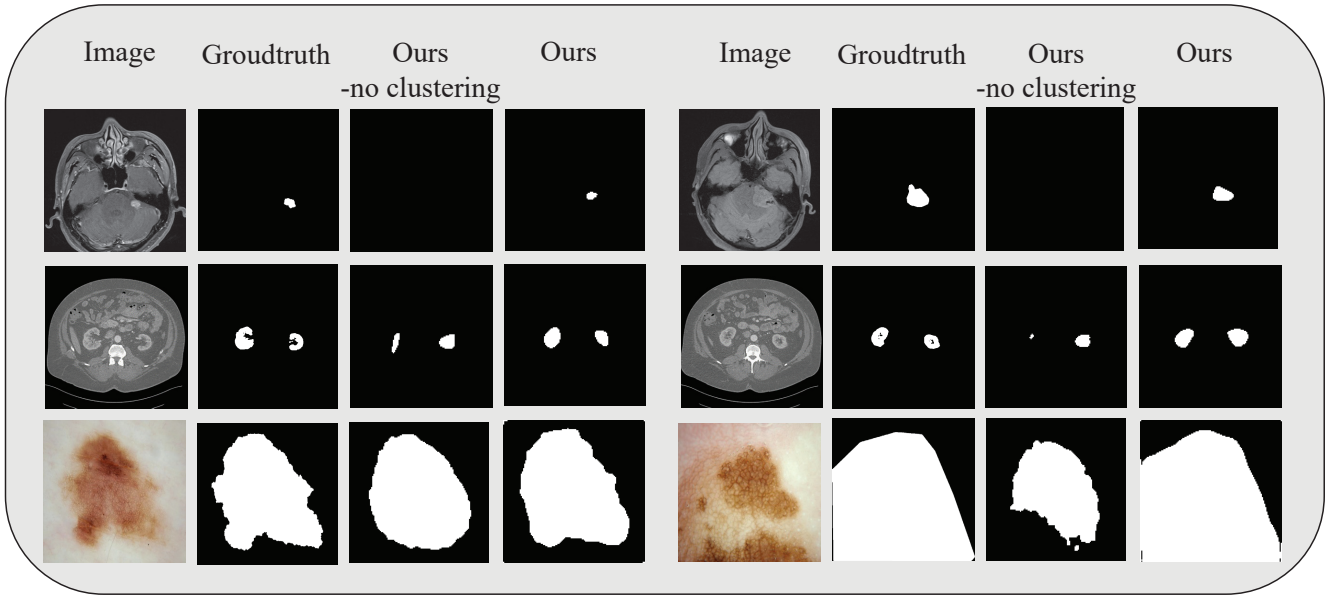

Fig. 11: Segmentation samples of using attention maps or clustered feature maps as the final semantic segmentation masks.

with the no-clustering results presented in TABLE 7 and Fig. 11, we find that the clustered results preserve more details of the lesion boundaries than attention maps and obtain a better segmentation score. One of the main reasons is that attention maps synthesized by G_{two} are not strictly 0 or 1 masks. Some minor changes in the values of adjacent pixels that may contain boundary information are lost during the normalization and binarization. Instead, the k-means algorithm preserves more accurate boundary information by clustering the flattened feature maps.

V. Discussion and Conclusion

In this study, a novel hierarchical attention GAN has been developed for semi-supervised semantic segmentation of biomedical images, particularly those with small lesions. Our approach effectively separates the synthesis process into lesion and non-lesion components, enhancing the clarity and visibility of lesion areas. A key feature of our model is the tail two-branch generator, inspired by StyleGAN2 [46], engineered to concurrently generate semantic masks and lesion textures from identical latent codes. These generated semantic masks function as attention maps throughout the training process. One challenge with this method is that the attention maps produced by the tail two-branch generator are not strictly binary, which means subtle variations in adjacent pixel values, which are potentially crucial for delineating boundaries, might be overlooked during the normalization and binarization processes. To address this, we have employed the k-means clustering algorithm on the flattened feature maps from specific layers of the tail two-branch generator during the inference stage. This technique aids in preserving detailed boundary information within the final segmentation masks. The performance of our proposed method has been rigorously assessed against leading generative models and GAN-based semi-supervised segmentation techniques through both qualitative and quantitative evaluations, showcasing its effectiveness and potential in enhancing segmentation accuracy for biomedical images with small lesions. However, this method also has some limitations. GAN-based semi-supervised segmentation methods learn the mapping from the noise space to the medical image space to acquire the underlying semantic information of the image. Since the adversarial loss primarily focuses on the validity of the generated images, the model tends to focus more on shallow image representations (such as global texture distribution) and neglects the deeper features that truly affect the image semantics. In contrast, by extracting the common features of the same image under different noise influences, adversarial consistency methods are more effective in extracting deep features that impact image semantics. Considering that GAN-based methods can generate high-quality images and pseudo-labels, Combining GAN-based semi-supervised learning methods and consistency normalization-based semi-supervised learning strategies could be a potential future direction arising from this research.

References

[1] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. Grimson, and A. Willsky, A shape-based approach to the segmentation of medical imagery using level sets,*IEEE Transactions on Medical Imaging*, vol. 22, no. 2, pp. 137–154, 2003.

[2] A. Yezzi, S. Kichenassamy, A. Kumar, P. Olver, and A. Tannenbaum, A geometric snake model for segmentation of medical imagery,*IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 199–209, 1997.

[3] A. Kumar Mondal, J. Dolz, and C. Desrosiers. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning.*arXiv preprint*, arXiv:1810.12241, 2018.

[4] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, pp. 1195–1204, 2017.

[5] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In Proceedings of the 2015

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI:10.1109/TETCI.2025.3540418, IEEE Transactions on Emerging Topics in Computational Intelligence

IEEE International Conference on Computer Vision (ICCV), page 1422–1430, USA, 2015. IEEE Computer Society.

[6] D. Kingma and M. Welling, Auto-encoding variational bayes, *arXiv preprint*, arXiv: 1312.6114, 2013.

[7] D. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 2013.

[8] D. Kingma, D. Rezende, S. Mohamed, and M. Welling. Semi-supervised learning with deep generative models. NIPS'14, page 3581–3589, Cambridge, MA, USA, 2014. MIT Press.

[9] D. Li, J. Yang, K. Kreis, A. Torralba and S. Fidler. Semantic segmentation with generative models: semi-supervised learning and strong out-of-domain generalization, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8296–8307, doi: 10.1109/CVPR46437.2021.00820, 2021.

[10] D. Pakhomov, S. Hira, N. Wagle, Kemar E. Green, N. Navab. Segmentation in style: unsupervised semantic image segmentation with stylegan and CLIP, *arXiv preprint*, arXiv: 2107.12518, 2021.

[11] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, Paul F. Jaeger, S. Kohl, J. Wasserthal, G.Koehler, T. Norajitra, S. Wirkert, and K. Maier-Hein, nnu-net: Self-adapting framework for unet-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[12] G. Daras, A. Odena, H. Zhang, A. Dimakis. Your local GAN: designing two dimensional local attention mechanisms for generative models, *arXiv preprint*, arXiv: 1911.12287, 2019.

[13] H. Basak, R. Bhattacharya, R. Hussain and A. Chatterjee, An exceedingly simple consistency regularization method for semi-supervised medical image segmentation, *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1-4, doi: 10.1109/ISBI52829.2022.9761602, 2022.

[14] H. Seo, L. Yu, H. Ren, X. Li, L. Shen and L. Xing, Deep neural network with consistency regularization of multi-output channels for improved tumor detection and delineation, *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3369–3378, doi: 10.1109/TMI.2021.3084748, 2021.

[15] H. Tang, D. Xu, N. Sebe and Y. Yan, Attention-guided generative adversarial networks for unsupervised image-to-image translation, *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, doi: 10.1109/IJCNN.2019.8851881, 2019.

[16] H. Wu, H. Wang, H. He, Z. He and G. Wang, A novel weakly supervised framework based on noisy-label learning for medical image segmentation, *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1768–1772, doi: 10.1109/ISBI48211.2021.9433883, 2021.

[17] H. Emami, M. M. Aliabadi, M. Dong and R. B. Chinnam, SPA-GAN: Spatial attention gan for image-to-image translation, *in IEEE Transactions on Multimedia*, vol. 23, pp. 391–401, doi: 10.1109/TMM.2020.2975961, 2021.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks, *arXiv preprint*, arXiv: 1406.2661, 2014.

[19] J. Zhu, T. Park, P. Isola, and A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

[20] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning. *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*, 2020.

[21] J. Kim, M. Kim, H. Kang and K. H. Lee, U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation, *International Conference on Learning Representations (ICLR)*, 2020.

[22] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[23] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville and R. Salakhutdinov, Show, attend and tell: neural image caption generation with visual attention, *Computer Science*, pp. 2048–2057, 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[25] K. Singh, U. Ojha, and Y. Lee, Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6483–6492, 2019.

[26] L. Beyer, X. Zhai, A. Oliver, and A. Kolesnikov. S4l: Selfsupervised semi-supervised learning. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 1476–1485, 2019.

[27] L. Beyer, X. Zhai, A. Oliver, and A. Kolesnikov. S4l: Selfsupervised semi-supervised learning. *In 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1476–1485, 2019.

[28] L. Chai, Z. Wang, J. Chen, G. Zhang, Fawaz E. Alsaadi, Fuad E. Alsaadi, Q. Liu, Synthetic augmentation for semantic segmentation of class imbalanced biomedical images: A data pair generative adversarial network approach, *Computers in Biology and Medicine*, Volume 150, art. no. 105985, 2022, DOI: 10.1016/j.compbiomed.2022.105985.

[29] L. Zhang, V. Gopalakrishnan, L. Lu, R. M. Summers, J. Moss and J. Yao, Self-learning to detect and segment cysts in lung CT images without manual annotation, *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 1100–1103, doi: 10.1109/ISBI.2018.8363763, 2018.

[30] M. Arjovsky, S. Chintala, and L. Bottou, Wasserstein generative adversarial networks,*Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 214–223, 2017.

[31] M. Noroozi, P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, Computer Vision, ECCV 2016, pages 69–84, Cham, 2016. Springer International Publishing.

[32] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6629–6640, 2017.

[33] N. Heller, N. Sathianathen, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, M. Oestreich et al., The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes, *arXiv preprint*, arXiv:1904.00445, 2019.

[34] N. Heller, N. Sathianathen, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, M. Oestreich et al., The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes, *arXiv preprint*, arXiv:1904.00445, 2019.

[35] O. Ronneberger, P.Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, pp. 234–241, 2015.

[36] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint*, arXiv:1611.08408, 2016.

[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, Image-to-image translation with conditional adversarial networks, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.

[38] P. Goyal, D. Mahajan, A. Gupta, and I. Misra. Scaling and benchmarking self-supervised visual representation learning. In ICCV, 2019.

[39] P. Tschandl, C. Rosendahl, and H. Kittler, The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, *Sci. Data 5, 180161*, DOI: 10.1038/sdata.2018.161, 2018

[40] Q. -Q. Chen, Z. -H. Sun, C. -F. Wei, E. Q. Wu and D. Ming, Semi-supervised 3d medical image segmentation based on dual-task consistent joint learning and task-level regularization, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, doi: 10.1109/TCBB.2022.3144428, 2022.

[41] R. Pohle and K. D. Toennies, Segmentation of medical images using adaptive region growing, *Medical Imaging 2001: Image Processing*, vol. 4322, pp. 1337–1346, 2001.

[42] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, Reusing discriminators for encoding: Towards unsupervised image-to-image translation, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8165–8174, 2020.

[43] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

[44] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[45] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In International Conference on Learning Representations, 2018.

[46] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019.

[47] T. Park, M. Liu, T. Wang, and J. Zhu, Semantic image synthesis with spatially-adaptive normalization, *2019 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pp. 2332–2341, 2019.

[48] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, Improved techniques for training gans, *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2234–2242, 2016.

[49] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

[50] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[51] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. *In NeurIPS*, 2020.

[52] T. Karras, S. Laine, and T. Aila, A style-based generator architecture for generative adversarial networks, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019.

[53] V. Grau, A. Mewes, M. Alcaniz, R. Kikinis, and S. Warfield, Improved watershed transform for medical image segmentation using prior information, *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 447–458, 2004.

[54] W. Hung, Y. Tsai, Y. Liou, Y.Lin, and M. Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018.

[55] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert. Semisupervised learning for network-based cardiac mr image segmentation. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 253–260. Springer, 2017.

[56] X. Li, L. Yu, H. Chen, C. -W. Fu, L. Xing and P. -A. Heng, Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523–534, doi: 10.1109/TNNLS.2020.2995319, 2021.

[57] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, Least squares generative adversarial networks, *2017 IEEE International Conference on Computer Vision (ICCV)* pp. 2813–2821, 2017.

[58] X. Chen, C. Xu, X. Yang, D. Tao, Attention-gan for object transfiguration in wild images, *arXiv preprint*, arXiv: 1803.06798, 2018.

[59] Y. Wang, P. Wang, B. Sun, K. He, and L. Huang, Iinfogan: Improved information maximizing generative adversarial networks, *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, pp. 1487–1490, 2020.

[60] Y. Liu, Z. Li, S. Pan, C. Gong, C. Zhou and G. Karypis, Anomaly detection on attributed networks via contrastive self-supervised learning, *in IEEE Transactions on Neural Networks and Learning Systems*, doi: 10.1109/TNNLS.2021.3068344, 2021.

[61] Y. Zhang et al., DatasetGAN: Efficient labeled data factory with minimal human effort, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10140–10150, 2021, DOI: 10.1109/CVPR46437.2021.01001.

[62] Y. Li, K. K. Singh, U. Ojha, and Y. J. Lee, Mixnmatch: Multifactor disentanglement and encoding for conditional image generation, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8036–8045, 2020.

[63] Y. Zou, Z. Yu, B. Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. *In Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.

[64] Z. Ke, Di Qiu, K. Li, Q. Yan, and R. Lau. Guided collaborative training for pixel-wise semi-supervised learning. *arXiv preprint arXiv:2008.05258*, 2020.

[65] Z. Han, G. Ian, Metaxas, D. Odena, Augustus, Self-attention generative adversarial networks, *arXiv preprint*, arXiv: 1805.08318, 2018.

[66] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim. Image-to-image translation for domain adaptation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 4500–4509, 2018.

[67] L. Yang, L. Qi, L. Feng, W. Zhang, Y. Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation,*In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023.

[68] C. You, W. Dai, Y. Min, F. Liu, D. Clifton, S. Zhou, L. Staib, J. Duncan. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective, *Advances in Neural Information Processing Systems*, 2023.