DiffuSeg: Domain-Driven Diffusion for Medical Image Segmentation

Le Zhang, Fuping Wu, Kevin Bronik, and Bartlomiej W. Papiez

***Abstract*—In recent years, the deployment of supervised machine learning techniques for segmentation tasks has significantly increased. Nonetheless, the annotation process for extensive datasets remains costly, labor-intensive, and error-prone. While acquiring sufficiently large datasets to train deep learning models is feasible, these datasets often experience a distribution shift relative to the actual test data. This problem is particularly critical in the domain of medical imaging, where it adversely affects the efficacy of automatic segmentation models. In this work, we introduce DiffuSeg, a novel conditional diffusion model developed for medical image data, that exploits any labels to synthesize new images in the target domain. This allows a number of new research directions, including the segmentation task that motivates this work. Our method only requires label maps from any existing datasets and unlabelled images from the target domain for image diffusion. To learn the target domain knowledge, a feature factorization variational autoencoder is proposed to provide conditional information for the diffusion model. Consequently, the segmentation network can be trained with the given labels and the synthetic images, thus avoiding human annotations. Initially, we apply our method to the MNIST dataset and subsequently adapt it for use with medical image segmentation datasets, such as retinal fundus images for vessel segmentation and MRI images for heart segmentation. Our approach exhibits significant improvements over relevant baselines in both image generation and segmentation accuracy, especially in scenarios where annotations for the target dataset are unavailable during training. *An open-source implementation of our approach can be released after reviewing.*.**

***Index Terms*—Diffusion model, domain transfer, image generation, image segmentation.**

Received 24 July 2024; revised 26 November 2024; accepted 30 December 2024. Date of publication 7 January 2025; date of current version 7 May 2025. *(Corresponding author: Le Zhang.)*

Le Zhang is with the School of Engineering, College of Engineering and Physical Sciences, University of Birmingham, B15 2TT Birmingham, U.K., and also with the Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, OX3 7LF Oxford, U.K. (e-mail: l.zhang.16@bham.ac.uk).

Fuping Wu and Bartlomiej W. Papiez are with the Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, OX3 7LF Oxford, U.K..

Kevin Bronik is with the Department of Engineering Science, University of Oxford, OX3 7LF Oxford, U.K..

Digital Object Identifier 10.1109/JBHI.2025.3526806

I. Introduction

SUPERVISED semantic image segmentation is known to require large amounts of annotated data, generating in turn high costs due to the expertise required to annotate such data, especially in the medical imaging domain. Despite the availability of numerous datasets with curated labels, there remains significant inter-domain variability across these datasets, affecting the training and performance of downstream supervised machine learning models [1]. This problem is particularly pronounced in the medical domain, where collecting pixel-level labels for newly acquired data is often expensive, time-consuming, and sometimes impractical. Distribution shifts are commonly observed in this context. In multi-center studies, domain shifts commonly occur across different imaging centers as a result of variations in scanners, scanning protocols, and subject populations [2]. Accurate segmentation of vessels in fundus retinal images is challenging when machine learning models are pre-trained on data from different studies due to variability in data distribution. This issue is exacerbated by annotation biases present in the training data, leading to high prediction errors and significant performance degradation in the segmentation of anatomical structures in medical images [3]. For example, one study reported substantial inter-reader variability, with an average pairwise agreement between annotators of 0.78 for retinal vessel segmentation [4]. Consequently, despite over two decades of digitization resulting in an abundance of medical imaging data, accessing large imaging repositories with curated labels remains a significant challenge in healthcare. Additionally, the lack of general processing algorithms capable of handling data from multiple studies underscores the need for developing intelligent methods that can learn robustly across different datasets.

To reduce human annotations and mitigate domain variations for downstream supervised models (e.g., Fig. 1(a) and (b), recent works focused on designing realistic image simulation scenarios (e.g., Fig. 1(c)) in which ground truth annotations are readily available [5]. Image-to-image (I2I) translation has been proposed as a potential solution to enforce similar input data distributions across two domains [6]. More recently, generative techniques such as Generative Adversarial Networks (GANs) [7] and Variational Autoencoders (VAEs) [8] have achieved impressive results in generating high-quality images. These techniques have been utilized in conditional settings [5], [9] to address the I2I translation problem. For instance, Pix2Pix [5] is a prominent method that employs a conditional setting to learn I2I domain

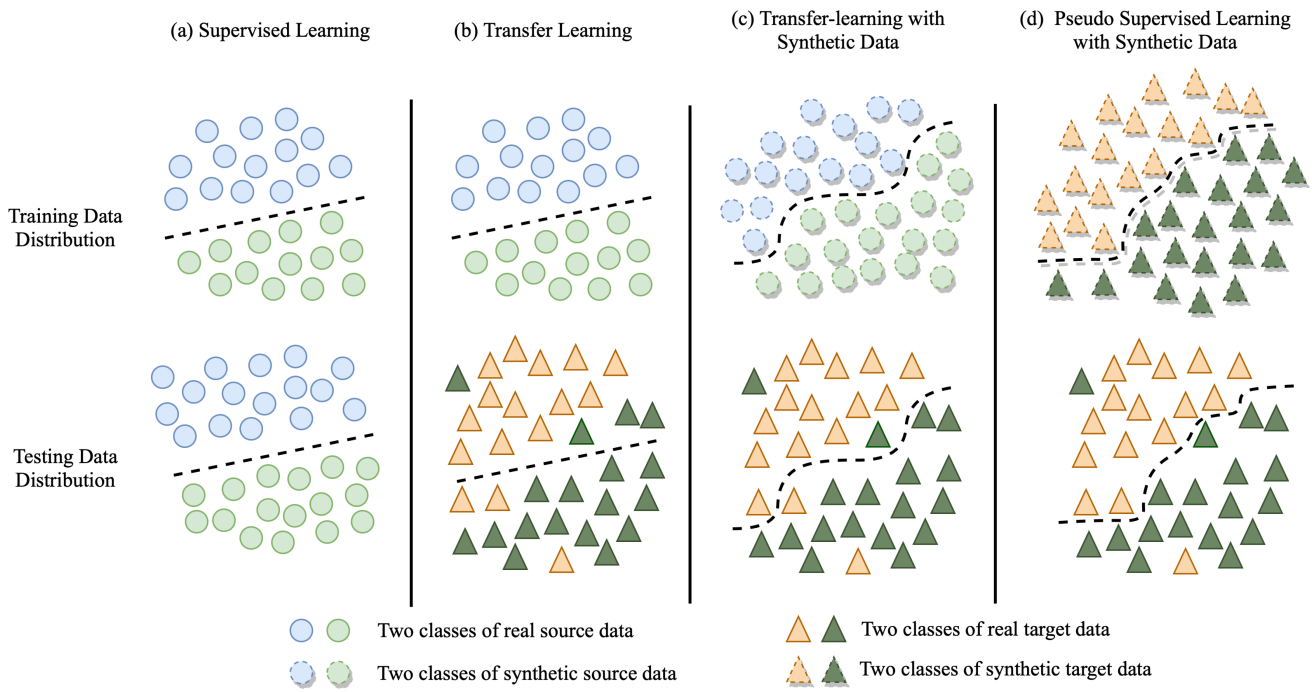

Fig. 1. Classification assumptions in different training data compositions. (a) Training and testing on source dataset; (b) Training on source dataset and fine-tuning with limited target data, testing on target dataset; (c) Training on a large number of synthetic source data and testing on target dataset; (d) Training on a large number of synthetic target data and testing on target dataset. The dashed lines indicate the classifiers that are trained under the corresponding approach.

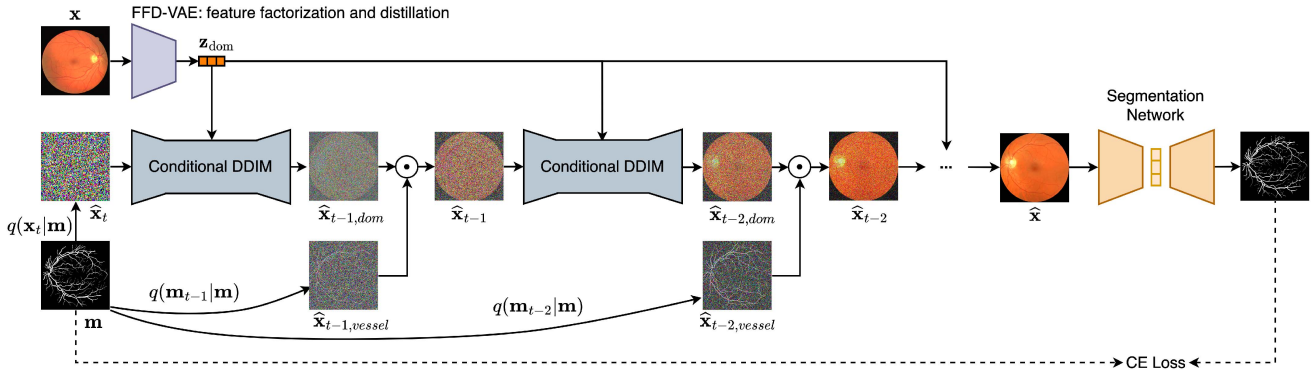

Fig. 2. The general architecture of our DiffuSeg model. Given a group of random images $\mathbf{x}$ from the target domain and a given mask $\mathbf{m}$, we utilize the diffusion process to generate a coherent image. The resulting image $\widehat{\mathbf{x}}$ aligns with the provided vessel description $\mathbf{m}$, while ensuring the complementary area closely resembles the original image $\mathbf{x}$. The obtained image $\widehat{\mathbf{x}}$ is then run through the segmentation network.

adaptation mapping and capture structural information. However, it necessitates paired cross-domain images for training, which are frequently challenging to acquire in practice. Furthermore, although GAN and VAE-based methods produce realistic visual results for several I2I translation tasks, they frequently exhibit implausible image artifacts in the translated images and sometimes fail to translate images with consistent structural and textural regularity (see in Fig. 5). This issue is especially problematic in domain adaptation scenarios, where maintaining strict fidelity of image content is crucial. More recently, score-based generative models such as Denoising Diffusion Implicit Model (DDIM) [10], have seen a significant increase in their application to image generation tasks, demonstrating the ability to encompass a wide range of visual semantics across various image domains. In particular, on the class-conditional ImageNet generation challenge, diffusion models have outperformed the state-of-the-art GAN baselines on Fréchet Inception Distance (FID) [11].

Our contribution: In this work, we present DiffuSeg, a novel diffusion model and neural network hybrid structure to readily segment images of any domains. Our DiffuSeg is trained on the synthetic images generated from the target domain by a diffusion model. To factorize the target domain knowledge as the conditional information for the diffusion model, we propose a novel feature factorization and distillation structure, which consists of paired VAE with "Teacher-Student" training strategy. Crucially, the synthetic images are sampled on the label maps from any datasets, which exposes the segmentation neural

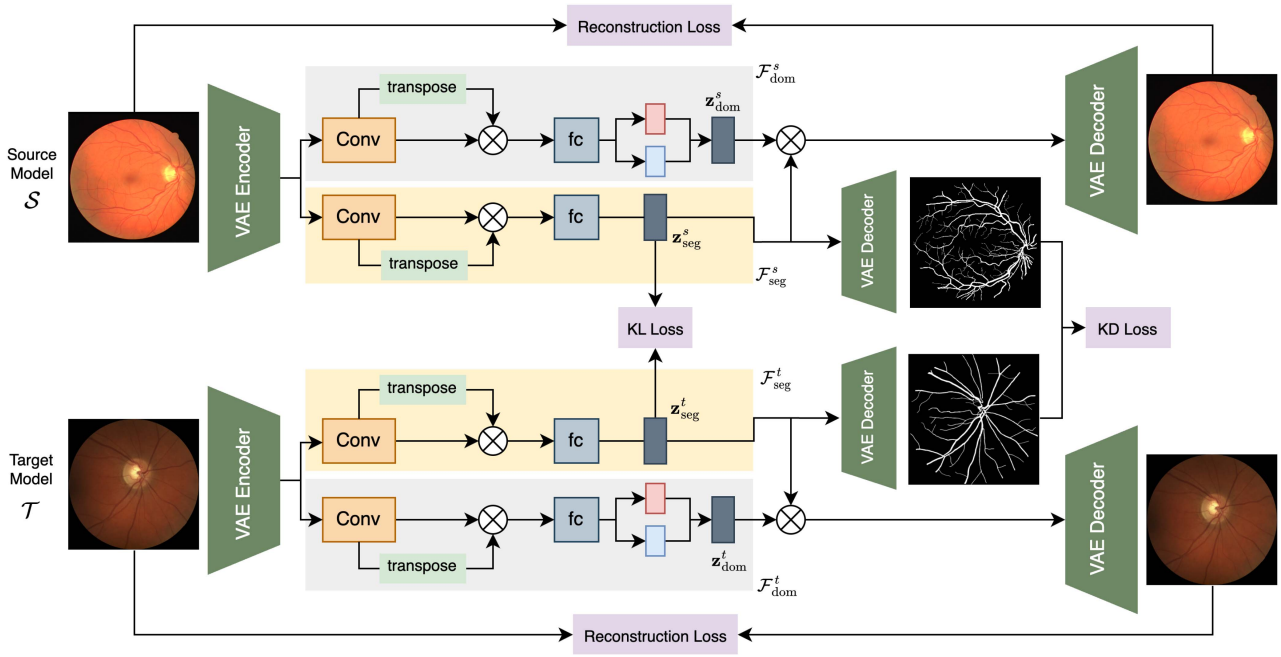

Fig. 3. Our proposed feature factorization and distillation (FFD-VAE) architecture.

network to ground truth annotations available, and thus avoids human annotations.

For evaluation, we first evaluate our method on the MNIST and publicly available retinal fundus image and heart MRI image datasets by performing domain-driven diffusion. Then we demonstrate the downstream utility of the diffusion results in real-world segmentation tasks. The results of the image generation process demonstrate that our method produces superior synthetic images in comparison to the current state of the art GAN and VAE-based frameworks, and the diffusion results are capable of boosting the cross-domain segmentation performance compared to the state-of-the-art transfer learning approaches.

II. Motivation

Modern image segmentation has made significant strides with the advent of deep neural networks. However, the robustness of these networks in domains different from the training data raises concerns over their generalization capabilities. We aims to address the issue of "domain-gap" or "distribution shift" [12] in deep learning networks, particularly its adverse effects on medical image segmentation. We explore the limitations of current approaches in handling unseen or limited datasets, emphasizing the challenges of generalizing across various modalities and resolutions in medical imaging [13], [14]. By understanding and mitigating these domain-generalization issues, we can enhance the reliability and accuracy of medical image segmentation techniques [15].

A. Domain Factorization for Image Diffusion

To synthesize domain-specific images, generative models are usually trained on the source dataset. However, these models may not generalize well to the target domain due to domain shift. Knowledge Distillation (KD) helps transfer domain-specific variations, creating a domain-invariant representation for better generalization to the target domain [16], but KD is often treated as a black-box approach, lacking interpretability. Knowledge Factorization (KF) [17], [18] provides an alternative method to identify and factorize domain-specific variations causing the distribution shift. It creates a better domain-invariant representation, leading to improved generalization. More specifically, we utilize domain factorization to enhance interpretability, as it clarifies features transferred from the source model to the target model. Inspired by KF, we propose a Feature Factorization and Distillation (FFD) architecture (see Fig. 3) for domain knowledge extraction, which is used for the condition of the target image diffusion process.

B. Consistency With Image Generation for Segmentation

Domain adaptation techniques are crucial for bridging the domain gap; however, they typically require a substantial amount of supervised data for training in both source and target domains. To mitigate this limitation, there is increasing interest in generating extensive target data. Recent approaches demonstrated competitive results in synthetic the target image data alone, optimizing the generation process through methods such as adversarial networks or semantic loss computation to reduce the "reality gap" [19]. However, incorporating the generated image data to train the downstream classification or segmentation models prior to real-world deployment [20], this strategy necessitates large training datasets that require human annotation to provide segmentation ground truth. In our paper, we propose an alternative solution—a diffusion model specifically developed for medical image data. This model allows the synthesis of

new images in the target domain without the need for human annotations. By sampling synthetic images based on label maps from any datasets, the segmentation neural network is exposed to available ground truth annotations, effectively circumventing the requirement for labour-intensive human annotations.

C. Proposal

Overall, although the recent image generation techniques such as GAN and VAE have achieved impressive results, these generated images show the random distribution of the anatomy (e.g., segmentation targets), therefore no ground truth label is available for training the downstream segmentation model. Meanwhile, implausible image artifacts are frequently observed in the translated images and they are sometimes unsuccessful in holistically translating images with consistent structural and textural regularity. This directly inspired us to focus on domain-driven image diffusion in our work that we want to design domain factorization and distillation structure to identify and factorize out the domain-specific variations or factors, which would lead to better generalization to the target domain. We present the VAE-based FFD structure in Section. III-B. Crucially, the synthetic images are sampled on the label maps from any datasets, which exposes the segmentation neural network to ground truth annotations available, and thus avoids human annotations. Therefore, we introduce our conditional blended diffusion model in Section III-C, which guarantees label preservation during image generation. The general pipeline of our model is presented in Fig. 2.

III. Method

A. Problem Formulation

Given an image $\mathbf{x}$ from the target dataset and a segmentation mask $\mathbf{m}$ from the source dataset, our target is to generate a modified image $\widehat{\mathbf{x}}$, s.t. the content $\widehat{\mathbf{x}}$ is consistent with the given mask $\mathbf{m}$, while the complementary area (e.g., background or image domain) closely resembles the target image $\mathbf{x}$, i.e., $\mathbf{x} \odot (1 - \widetilde{\mathbf{m}}) \approx \widehat{\mathbf{x}} \odot (1 - \mathbf{m})$, where $\widetilde{\mathbf{m}}$ is the segmentation map of the given image $\mathbf{x}$, $\odot$ is element-wise multiplication. Meanwhile, the transition between $\mathbf{x}$ and $\widehat{\mathbf{x}}$ should ideally appear seamless.

B. Feature Factorization and Distillation in VAE

Recently, Liu et al. [21] proposed to disentangle the input image to anatomy and modality factors to solve the domain shift problem. Inspired by the disentangled representations, we propose a Feature Factorization and Distillation (FFD) architecture for domain knowledge extraction, which is a process of subdividing the VAE into a network with two-factor branches. Each branch possesses distinctive knowledge to handle anatomy (e.g., blood vessels) and image domain (e.g., texture) feature distributions; then leveraging the knowledge from a source VAE to train the target VAE with "Teacher-Student" strategy. Given a source VAE model $\mathcal{S}$ with two-factor branches that are able to reconstruct the input image and handle segmentation simultaneously, $\mathbf{z}^s_{\text{seg}}$ could be estimated from the segmentation branch $\mathcal{F}^s_{\text{seg}}$, and reconstruction branch $\mathcal{F}^s_{\text{dom}}$ not only masters the source domain knowledge $\mathbf{z}^s_{\text{dom}}$ but also benefits from the segmentation feature $\mathbf{z}^s_{\text{seg}}$ to make the final image reconstruction. Fig. 3 shows our proposed FFD structure.

Anatomy Knowledge Distillation: Utilizing the approach of generalized knowledge distillation [22], we transfer the anatomy segmentation knowledge from the source VAE network $\mathcal{S}$ to the target VAE network $\mathcal{T}$ using the given segmentation dataset with soft labels $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$. These are computed as $\mathbf{y}_i = \sigma(f_s(\mathbf{x}_i))$ where σ is the softmax function, $f_s(\cdot)$ is the source feature extractor for segmenting image $\mathbf{x}_i$. The segmentation KD loss is defined as:

$$\mathcal{L}_{\text{kd}} = \sum_i \left[(1 - \text{Dist}\left(\tilde{\mathbf{y}}_i,\ \sigma\left(f_s\left(\mathbf{x}_i\right)\right)\right)) + \text{CE}\left(\mathbf{y}_i, \sigma\left(f_s\left(\mathbf{x}_i\right)\right)\right)\right] \tag{1}$$

where $\text{CE}(\cdot)$ is the cross-entropy measure for the segmentation accuracy of the source model $\mathcal{S}$, $\tilde{\mathbf{y}}_i$ is the segmentation prediction of the target model $\mathcal{T}$. The $\text{Dist}(\cdot)$ measures the similarity of the distribution of the two segmentation maps from $\mathcal{S}$ and $\mathcal{T}$. Furthermore, we utilize the Kullback-Leibler divergence as a loss function to compare the bottleneck representations between models $\mathcal{S}$ and $\mathcal{T}$. This approach ensures that the target model $\mathcal{T}$ effectively encodes the necessary information in its latent space for accurate segmentation of the input images:

$$\mathcal{L}_{latent}(\mathcal{T}_{\text{seg}}, \mathcal{S}_{\text{seg}}) = \sum_i \sum_j \mathcal{S}^i_{\text{seg}}(j) \log\left(\frac{\mathcal{S}^i_{\text{seg}}(j)}{\mathcal{T}^i_{\text{seg}}(j)}\right) \tag{2}$$

where $\mathcal{T}^i_{\text{seg}}$ and $\mathcal{S}^i_{\text{seg}}$ are the flattened and normalized vector of the segmentation bottleneck in target and source models. The complete objective function for segmentation knowledge distillation (SKD) is then:

$$\mathcal{L}_{skd} = \mathcal{L}_{kd} + \alpha \mathcal{L}_{latent} \tag{3}$$

where α balances the magnitude of the latent loss with respect to the KD loss.

Domain Feature Factorization via Image Reconstruction: The segmentation factor decoder is expected to inherit its knowledge and integrate with domain factor knowledge to reconstruct the image $\widehat{\mathbf{x}}$. θ and ϕ are the model parameters for the segmentation and image reconstruction branches, respectively. For each input sample $\mathbf{x}_i$, VEnc_θ is adopted to extract the segmentation feature $\mathbf{z}^i_{\text{seg}}$:

$$\mathbf{z}^i_{\text{seg}} = \text{VEnc}_\theta\left(\mathbf{x}_i; \theta_i\right). \tag{4}$$

On the contrary, VEnc_ϕ learns the domain-related knowledge $\mathbf{z}_{\text{dom}}$ from $\mathbf{x}$, which together with $\mathbf{z}^i_{\text{seg}}$ is processed by a decoder VDec_ϕ to make the image reconstruction:

$$\mathbf{z}_{\text{dom}} = \text{VEnc}_\phi\left(\mathbf{x}; \phi\right); \quad \widehat{\mathbf{x}} = \text{VDec}_\phi\left(\mathbf{z}^i_{\text{seg}},\ \mathbf{z}_{\text{dom}}; \phi\right). \tag{5}$$

which constrains domain factor encoder VEnc_ϕ to learn the domain feature and integrate it with the segmentation knowledge to handle the image reconstruction.

Intuitively, we expect that VDec_ϕ masters the image reconstruction task by using the domain knowledge $\mathbf{z}_{\text{dom}}$ and segmentation knowledge $\mathbf{z}^i_{\text{seg}}$. We accordingly define a feature factorization and distillation objective $\mathcal{L}_{\text{FFD}}$ to enforce the domain feature

factor branch of model $\mathcal{T}$ to imitate model $\mathcal{S}$ prediction while factorizing the domain knowledge via image reconstruction by minimizing the final loss:

$$\mathcal{L}_{\text{FFD}} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{skd} \tag{6}$$

where $\mathcal{L}_{\text{rec}}$ denotes the supervised loss for image reconstruction. In our work, $\mathcal{L}_{\text{rec}}$ takes the form of L2 norm for image reconstruction and λ is the weight coefficient. Notably, we may readily adopt various implementations for reconstruction here.

C. Conditional Blended Diffusion

The general idea of the diffusion model is that for an input image $\mathbf{x}$, we generate a series of noisy images $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T\}$ by a forward noising process that implicitly defines a sequence of image manifolds, each manifold consisting of progressively noisier images from 0 to T. During the reverse processing, we start from $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and predict $\mathbf{x}_{t-1}$ for $t \in \{T, \ldots, 1\}$. With this iterative denoising process, we can generate a fake image $\hat{\mathbf{x}}_0$. For more details about the image synthesis using the diffusion model please see the seminal paper [23]. In our work, to achieve a seamless result in which the image appearance aligns with the guiding domain feature $\mathbf{z}_{\text{dom}}$, while the target anatomy of the image is identical to the provided segmentation mask $\mathbf{m}$, the noisy images that are generated in a progressive manner by our conditional DDIM process are blended with the corresponding noisy versions of the segmentation maps. The key to this process is that although the intermediate blended noisy images may lack coherence at each step, the subsequent denoising diffusion step restores coherence by projecting the result onto the next manifold. The pipeline of the approach is depicted in Fig. 2.

Conditional DDIM: For each diffusion step, our conditional DDIM acts as an "encoder" $(\mathbf{x}_0 \rightarrow \mathbf{x}_T)$ and a "decoder" $((\mathbf{z}_{\text{dom}}, \mathbf{x}_T) \rightarrow \mathbf{x}_0)$. Here, $\mathbf{x}_T$ captures low-level stochastic variations from the random target image, $\mathbf{x}_T$ and $\mathbf{z}_{\text{dom}}$ together can be rendered back to the original image with high fidelity. In this work, we employ the publicly available latent diffusion model (LDM) known as Stable Diffusion [24] as our guidance model and give $\mathbf{z} = (\mathbf{z}_{\text{dom}}, \mathbf{x}_T)$ as the input to produce the synthetic image. The target image data is first passed through our FFD-VAE, which outputs its corresponding domain feature embedding $\mathbf{z}_{\text{dom}}$. Subsequently, we freeze the parameters of the domain knowledge learning model $\mathcal{T}$ and proceed to optimize the target domain embedding diffusion $\mathcal{D}(\mathbf{x}_t, t, \mathbf{z}_{\mathbf{dom}})$ using the objective function:

$$\mathcal{L}(\mathbf{x}_t, \mathbf{z}_{\text{dom}}) = \mathbb{E}_{\mathbf{x},\epsilon}\left[\|\boldsymbol{\epsilon} - \mathcal{D}(\mathbf{x}_t, t, \mathbf{z}_{\text{dom}})\|_2^2\right] \tag{7}$$

where $\mathbf{x}$ is the given segmentation map, $t \sim$ Uniform $[1, T]$, $\mathbf{x}_t$ is a noisy version of $\mathbf{m}$ obtained using $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t$ with $0 = \alpha_T < \alpha_{T-1} < \cdots < \alpha_1 < \alpha_0 = 1$ being the hyperparameters governing the diffusion process. This process yields a domain embedding that closely aligns with our target image. We perform this procedure for a limited number of steps to ensure proximity to the target image domain, resulting in $\widehat{\mathbf{x}}$. The maintenance of this proximity enables the implementation of meaningful linear interpolation within the embedding space. In the absence of this proximity, the embeddings exhibit a lack of linearity when they are more distant from one another.

Label Distribution Preserving: To make the generated image $\widehat{\mathbf{x}}$ identity to the provided segmentation map $\mathbf{m}$, we need to preserve the label distribution when diffusion. To achieve seamless results, Burt and Adelson [25] blend two images by smoothly combining each level of their Laplacian pyramids. Building on these promising results, we incorporate blending at various noise levels throughout the diffusion process. Our main concept involves projecting a noisy latent representation onto a medical image manifold corresponding to the noise level at each diffusion stage. The subsequent diffusion phase enhances coherence by projecting the result onto the next level manifold. In contrast, combining two images with the same noise level typically produces an outcome that lies outside the manifold.

During each step, we perform the conditional diffusion from a latent $\mathbf{x}_t$. This denoising operation is carried out in a direction that depends on the domain feature prompt, resulting in a new latent variable $\mathbf{x}_{t-1,\text{dom}}$. Concurrently, a noisy version of the segmentation mask $\mathbf{x}_{t-1,\text{seg}}$ is obtained by utilising the given segmentation mask using $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t$. Then we use the mask: $\widehat{\mathbf{x}}_{t-1} = \mathbf{x}_{t-1,\text{dom}} \odot (1 - \mathbf{m}) + \mathbf{x}_{t-1,\text{seg}} \odot \mathbf{m}$, to blend the above two latents, and the process is repeated. In the last stage, the whole region outside the segmentation map is modified by the corresponding background knowledge learned from the given image, which ensures that the given mask is strictly preserved in the target image domain.

During each step, we conduct conditional diffusion starting from a latent $\mathbf{x}_t$, denoising it in a direction influenced by the domain feature prompt, resulting in a latent $\mathbf{x}t - 1, \text{dom}$. Concurrently, a noised version of the segmentation mask $\mathbf{x}_{t-1,\text{seg}}$ is derived from the given segmentation mask using $\mathbf{x}t = \sqrt{\alpha_t}\mathbf{x}0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}t$. Subsequently, the above two generated latents are blended using the mask. $\widehat{\mathbf{x}}_{t-1} = \mathbf{x}_{t-1,\text{dom}} \odot (1 - \mathbf{m}) + \mathbf{x}_{t-1,\text{seg}} \odot \mathbf{m}$, and the process is then repeated. In the end, the whole image except the segmentation mask region is modified with the corresponding background knowledge generated from the given image, thereby faithfully preserving the given mask in the target image domain.

IV. Experiments and Analysis

A. Datasets

MNIST Segmentation Dataset: MNIST [26] $\rightarrow$ MNIST-M [27]. The MNIST dataset comprises 60,000 training examples and 10,000 testing examples. All the provided images are 28-by-28in grayscale with digits from 0 to 9. We threshold the intensity values of the image at 0.5 to derive all the segmentation labels. The MNIST-M dataset was created by blending the original MNIST images with patches randomly extracted from colour photos from BSDS500 dataset [28], therefore generating a quite distinct domain. In our work, we distil the segmentation knowledge from MNIST in order to factorise the domain feature in MNIST-M.

Retinal Vessel Segmentation Dataset: The IOSTAR retinal vessel segmentation dataset [32] comprises 30 images with a

resolution of 1024 × 1024pixels. The dataset comprises annotations of all vessels by a group of experts specialising in retinal image analysis. We applied data augmentation, such as rotation, scaling, flipping, translating, to increase the limited number of images. DRIVE [33] is another retinal vessel segmentation dataset. Each image was captured using 8 bits per colour plane at 768 × 584pixels. For this database, the images have been cropped around the field of view (FOV). The set has 40 images containing manual segmentation for each image. In our work, we use this dataset to evaluate the segmentation performance of the transfer learning approach.

Cardiac Ventricles Segmentation Datasets: Multi-sequence Cardiac Magnetic Resonance Segmentation (MSCMRseg) Challenge dataset [34], [35] includes 45 image subjects with cardiomyopathy. In our work, we use late gadolinium enhancement (LGE) CMR images to evaluate our model. Each LGE CMR sequence includes 10 to 18 slices that cover the main body of the ventricles, with an acquisition matrix of 512 × 512, providing an in-plane resolution of 0.75 × 0.75 mm. Manual annotations of the left ventricular (LV), right ventricular (RV), and left ventricular myocardium (Myo) were performed by three observers. The Automated Cardiac Diagnosis Challenge (ACDC) [36] provides a dataset specifically for the segmentation of the LV, RV, and Myo in cardiac MRI scans. The study's target population consists of 150 patients. A series of short-axis slices spans the LV from base to apex, with spatial resolution ranges from 1.37 to 1.68 mm^2/pixel, and the number of images covering the cardiac cycle ranges from 28 to 40 varying by patient.

Whole Heart Segmentation Datasets: The Multi-Modality Whole Heart Segmentation Datasets (MM-WHS) [37] is a collection of 120 multi-modal whole heart images gathered from various sources, including 60 cardiac CT/CTA scans and 60 cardiac MRI scans in 3D. These images encompass the entire heart and its substructures with axial plane slices acquired. The in-plane resolution is approximately 0.78 × 0.78 mm. The MRI data were acquired using 3D balanced steady-state free precession (b-SSFP) sequences in the axial view and cover the heart from the upper abdomen to the aortic arch. The dataset includes seven cardiac structures: LV, RV, LA, RA, Myo, ascending aorta (Ao), and pulmonary artery (PA). Each of these substructures was manually labeled in the images to create an atlas label map, which serves as the gold standard for validation. The dataset was divided into two subsets, with 20 CT and 20 MRI images used for training, and 40 CT and 40 MRI images reserved for testing.

B. Implementation

The proposed DiffuSeg model is implemented in Keras-TensorFlow GPU deep learning library and using AdamW for optimization. We split each dataset into 70%, 10%, 20% as the training, validation, and test sets, respectively. The learning rate is initialized to 1×10^{-4} with maximum 500 epochs. In our model, an early stopping routine was implemented that allowed us to stop training once the model performance on a holdout validation dataset stopped improving. The network is trained on a Quadro-RTX8000 GPU with 48GB of memory. The number of model parameters is roughly 26 M.

C. Comparison Methods and Evaluation Metrics

We evaluate our model from two different tasks: image generation and the downstream segmentation task. For image generation, we compare our method against multiple generative models. In particular, we consider five state-of-the-art (SOTA) I2I translation methods: cGAN [5], CVAE [29], which are trained with paired image data. CoCosNet [30] and CoCosNet V2 [31], which are trained with non-paired data. For image segmentation, we train a segmentation model on different ways of mixing data. Note that we adopt a UNet architecture [38] for the segmentation tasks as it is the most widely used in medical image segmentation, but the network architecture is not a focus of this work, and it could be replaced with any other segmentation network.

Quantitative evaluation of generative models is widely recognized as challenging, we adopt FID [11] to measure the quality of the synthesized results with the real one that these criteria values should be lower if the generated images are more realistic. To evaluate the background (domain) quality of the synthesized images, we employ the mutual background similarity (MBS) metric [40]. This metric assesses the consistency of the background between generated and real images that are expected to have identical backgrounds. A lower MBS score signifies a higher degree of background consistency between the paired images. For segmentation evaluation metrics, we use mean Intersection over Union (mIoU) between estimated segmentation $\widehat{\mathbf{m}}$ and expert consensus label $\mathbf{m}_{GT}$. $\text{mIoU}_c = \sum_i \frac{|\widehat{\mathbf{m}} \cdot \mathbf{m}_{GT} \cdot \mathbf{U}_c|}{|\widehat{\mathbf{m}} \cdot \mathbf{U}_c| + |\mathbf{m}_{GT} \cdot \mathbf{U}_c| - |\widehat{\mathbf{m}} \cdot \mathbf{m}_{GT} \cdot \mathbf{U}_c|}$ where $\mathbf{U}_c$ means the one-hot vector for class c, $\mathbf{U}_c = (U_1, \ldots, U_N), U_i = \begin{cases} 0 (i \neq c) \\ 1 (i = c) \end{cases}, c = 1, 2, \ldots, N$. Dice score coefficient $\left(DSC = \frac{2\widehat{\mathbf{m}} \cap \mathbf{m}_{GT}}{\widehat{\mathbf{m}} + \mathbf{m}_{GT}} \right)$, a region-based metric, is used to evaluate the region overlap. Jaccard index ($JI = \frac{\widehat{\mathbf{m}} \cap \mathbf{m}_{GT}}{\widehat{\mathbf{m}} \cup \mathbf{m}_{GT}}$)), also known as the Jaccard similarity coefficient, is a metric used to measure the similarity between two segmentations. It is defined as the size of the intersection of two segmentations divided by the size of their union. Hausdorff distance is a mathematical concept used to measure the dissimilarity between two sets of points. In the image segmentation task, Hausdorff distance ($HD = \max_{a \in \widehat{\mathbf{m}}} (\min_{b \in \mathbf{m}_{GT}} d(a, b))$) is employed to compare the similarity between two masks.

D. Performance on Image Generation

To evaluate the image generation performance, we utilize our approach and the selected comparison methods to synthesize the same number of images for each dataset. Fig. 4 shows the label → image translation performance on MNIST dataset using our conditional blended diffusion model. We can see our method successfully generate realistic images with the target domain and also preserve the given label maps. Meanwhile, We show the qualitative comparison and quantitative evaluation with the competitors in Fig. 5 and Table. I for retinal images. cGAN [5] and CVAE [29] leads to reasonable but noisy and corrupted results. This is because cGAN and CVAE usually need a large number of training data to achieve good results. CoCosNet [30]

Fig. 4. The diffusion results (MNIST → MNIST-M) using our method. For each group, *left* is MNIST digit mask, *middle* is our prediction, *right* is the corresponding MNIST-M image.

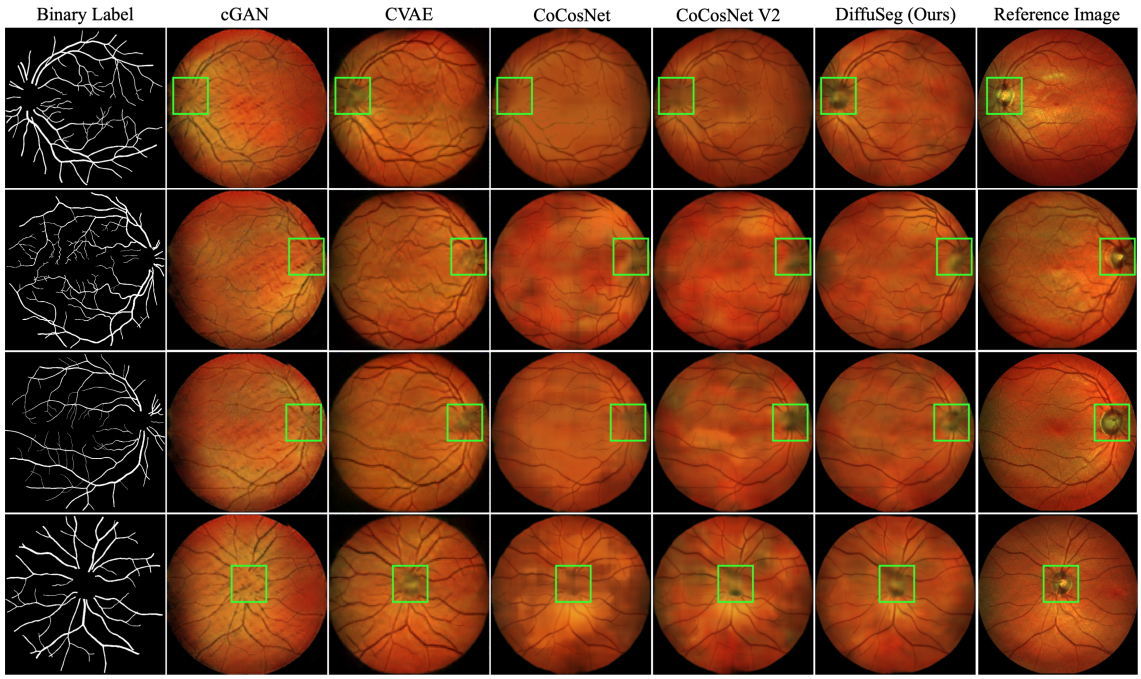

Fig. 5. The image generation results on IOSTAR datasets using different SOTA methods (zoom in to visualize the details).

TABLE I
QUANTITATIVE EVALUATION ON IOSTAR RETINAL DATASET: OUR MODEL VS. DIFFERENT IMAGE GENERATION MODELS (FID & MBS, MEAN ± STANDARD DEVIATION)

Methods	cGAN [5]	CVAE [31]	CoCosNet [32]	CoCosNet V2 [33]	DiffuSeg (Ours)
FID (↓)	90.21 ± 0.62	82.79 ± 0.52	66.59 ± 0.35	42.38 ± 0.28	**26.33 ± 0.43**
MBS (↓)	93.17 ± 0.46	80.22 ± 0.34	56.41 ± 0.78	39.19 ± 0.24	**23.82 ± 0.54**

Numbers in bold indicate the best method that statistically ($p < 0.01$) better than other methods by computing the p-values of paired t-tests on FID and MBS.

and CoCosNet V2 [31] produce better quality but blurry results, especially the missing information around the optic disc cup area. It can be clearly seen that our method generates the most visually appealing results and the least visible artefacts. We find that the distinctive patterns of the vessel in the exemplars have been remarkably well preserved in the output. On the other hand, our output depicts subtle details of the optic disc cup (see green box in each image in Fig. 5) that are of particular importance to a high-quality image and disease analysis, demonstrating the advantage of our diffusion model. In addition, we present the image generation results on another two medical datasets, which focus on cardiac ventricles (Fig. 6) and the whole heart structures (Fig. 7) segmentation. Because of sufficient training images from the two datasets, the generated images of each method have significant improvements in terms of the anatomy and domain character details. However, when overlapping the given contours of the segmentation structures, DiffuSeg shows comparable alignment performance of the image and contours among the different synthetic methods. We believe the reason is that the blending technique we utilized in the diffusion process. Therefore, we can further figure out from Fig. 6 and Fig. 7 that our model conditioned with the domain features and the preservation of the given labels generated high-quality images compared to other methods.

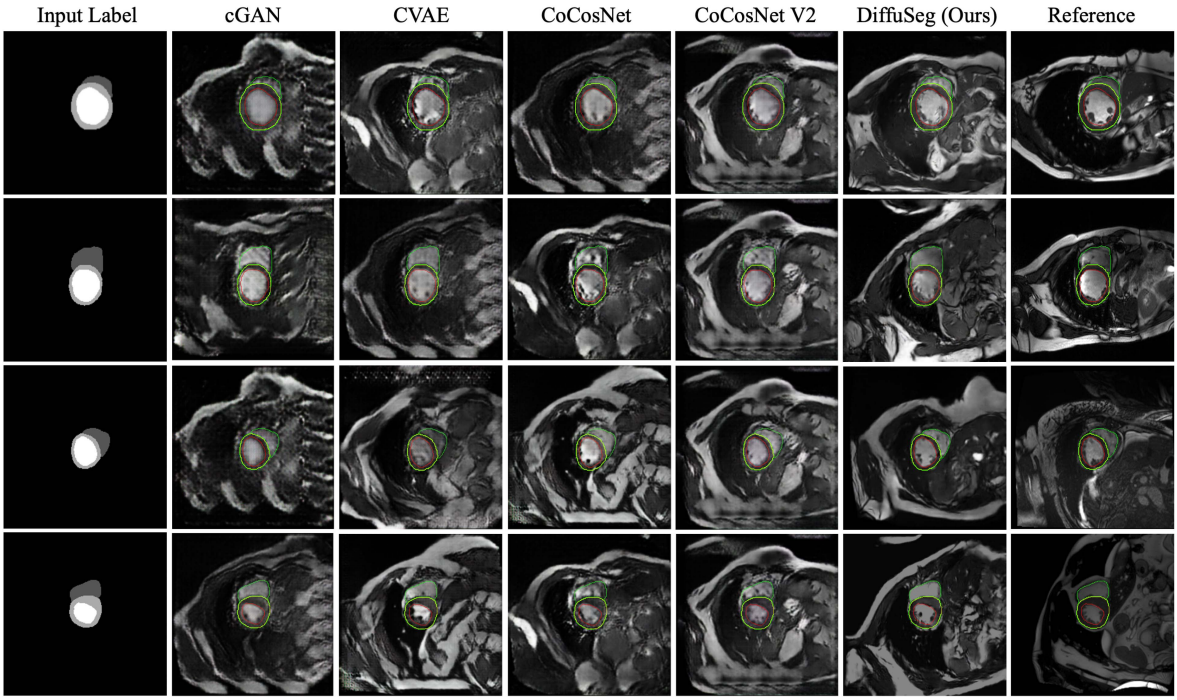

Fig. 6. The image generation results on MSCMR dataset using different SOTA methods. The contours of the input ventricle labels are overlapped on each image. The alignment performance between the image and the segmentation label can indicate the quality of the generated image.

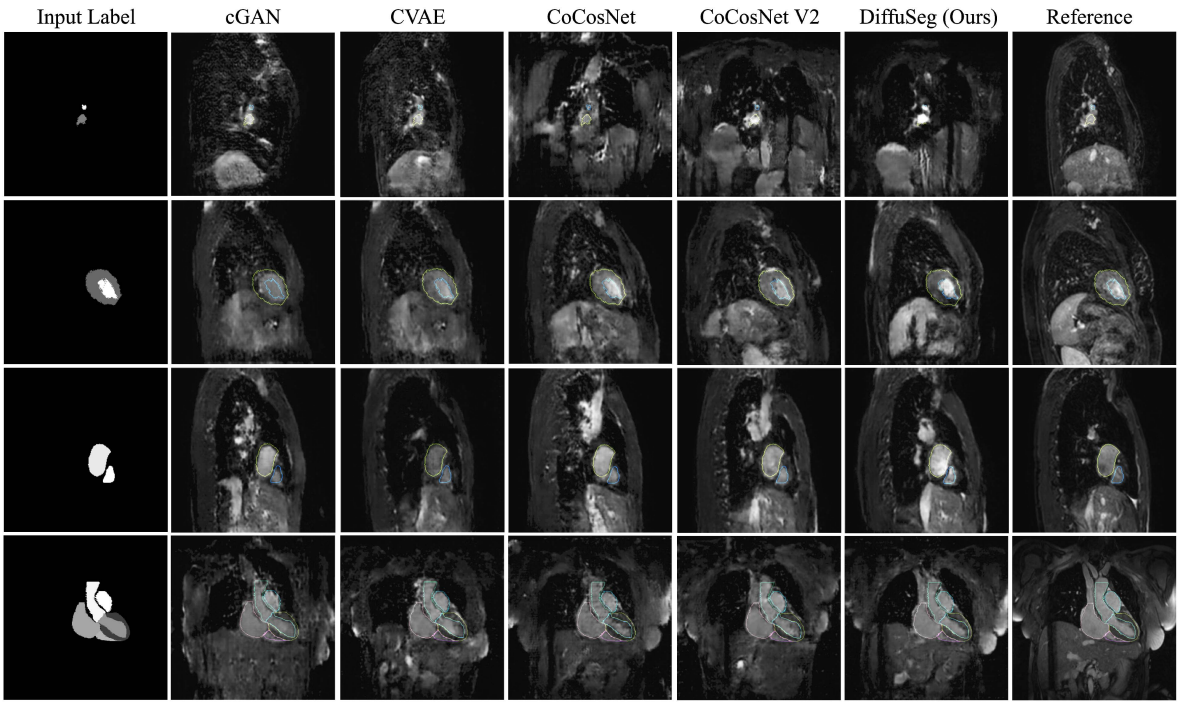

Fig. 7. The image generation results on WHS dataset using different SOTA methods. The contours of the input structures are overlapped on each image. The alignment performance between the image and the segmentation label can indicate the quality of the generated image.

E. Performance on Downstream Segmentation Tasks

We additionally evaluate the effectiveness of our diffusion results on the segmentation tasks. In Table. II we quantitatively compared the retinal vessel segmentation performance on different training strategies. When replacing all training data from DRIVE dataset with IOSTAR dataset, the segmentation performance shows an approximate 6% increment. We believe this is a reasonable gap considering the challenges of exploiting domain adaptation among different datasets. When applying

TABLE II

COMPARISONS OF SEGMENTATION RESULTS USING DIFFERENT TRAINING APPROACHES. THE TERM "TOTAL" SHOWS THE NUMBER OF TRAINING IMAGES, "W/ IOSTAR [39]" SHOWS THE NUMBER OF TRAINING IMAGES FROM IOSTAR DATASET, "W/ DRIVE [33]" SHOWS THE NUMBER OF TRAINING IMAGES FROM DRIVE DATASET, AND "W/ DIFFUSEG" SHOWS THE NUMBER OF SYNTHETIC IMAGES FROM OUR MODEL. ALL METHODS ARE TRAINED ON THE UNET SEGMENTATION NETWORK

Training Strategies	w/ IOSTAR [41]	w/ DRIVE [35]	w/ DiffuSeg	total	mIoU (%)
Transfer-learning	–	15	–	15	42.5 ± 0.4
Self-learning	15	–	–	15	48.2 ± 0.2
Semi-self-learning	15	–	15	30	53.7 ± 0.2
Ours (DiffuSeg)	–	–	50	50	56.1 ± 0.3

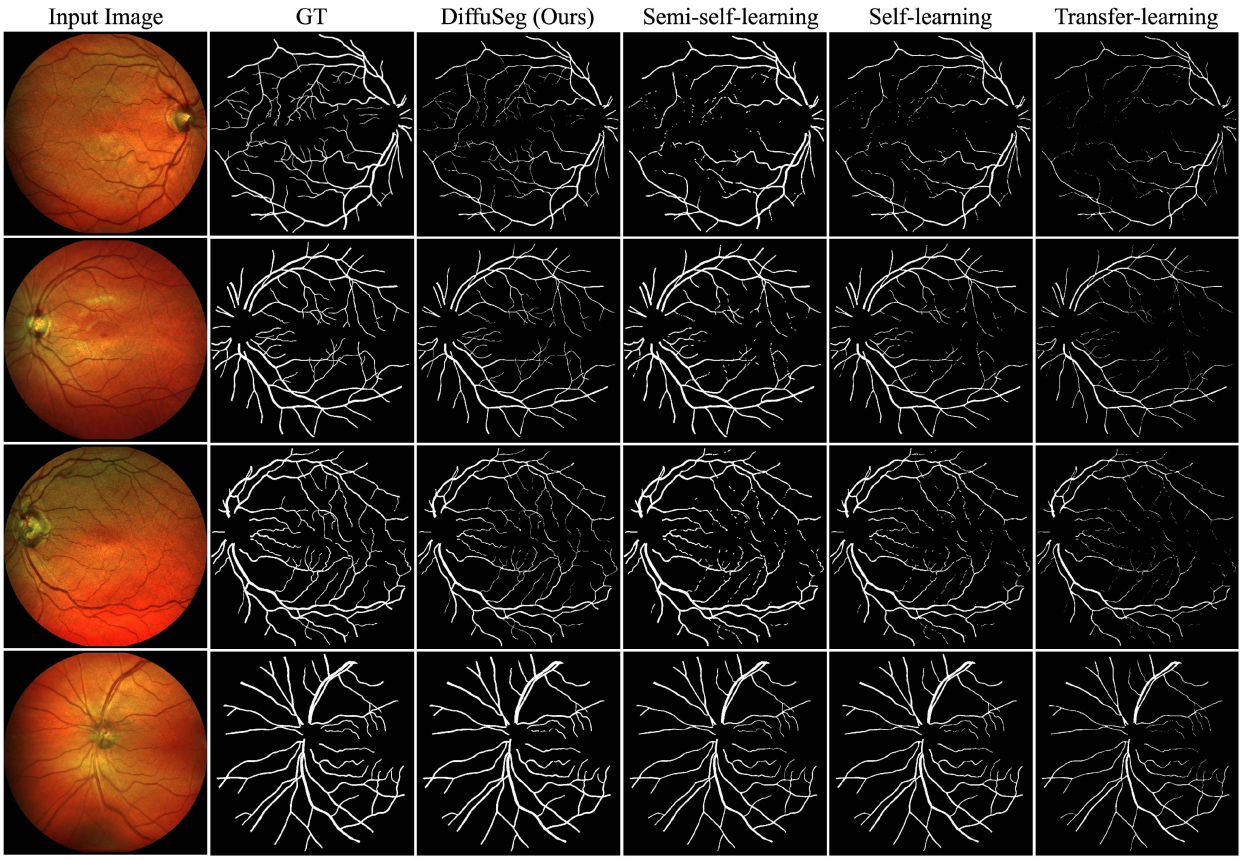

Fig. 8. The vessel segmentation results on IOSTAR dataset using different learning approaches.

extra 15 synthetic images of our diffusion model, the segmentation has an upgradation of about 5 points on the supervised-learning approach. When only using 50 of our synthetic images, we achieved about 8% increments. These results imply that our DiffuSeg method generated reliable images with the target domain and can be a novel data augmentation approach for the segmentation task. Meanwhile, the segmentation results on IOSTAR with different training approaches are visualized in Fig. 8, we can see that the segmentation model training on our DiffuSeg generated images show the best performance. This is not only because of the larger amount of training data that DiffuSeg provided, but also because the generated images show more diversities compared with the existing datasets with limited annotated images. In Fig. 9 and Fig. 10, we present the segmentation performance of our model on cardiac ventricles and heart imaging data, respectively. Notably, our model demonstrates exceptional capabilities in accurately delineating structures in both datasets. The utilization of DiffuSeg-generated images for training contributes significantly to the robust performance, leveraging a substantial increase in the training dataset size. This augmented dataset not only enhances the model's capacity to generalize but also introduces a rich diversity of anatomical variations that might be underrepresented in conventional datasets with limited annotations. The superior segmentation results underscore the efficacy of our approach, showcasing its adaptability and strength across different medical imaging domains. The model's proficiency in handling cardiac and heart segmentation tasks is a testament to its potential for broader applicability in diverse medical image analysis scenarios, highlighting its versatility

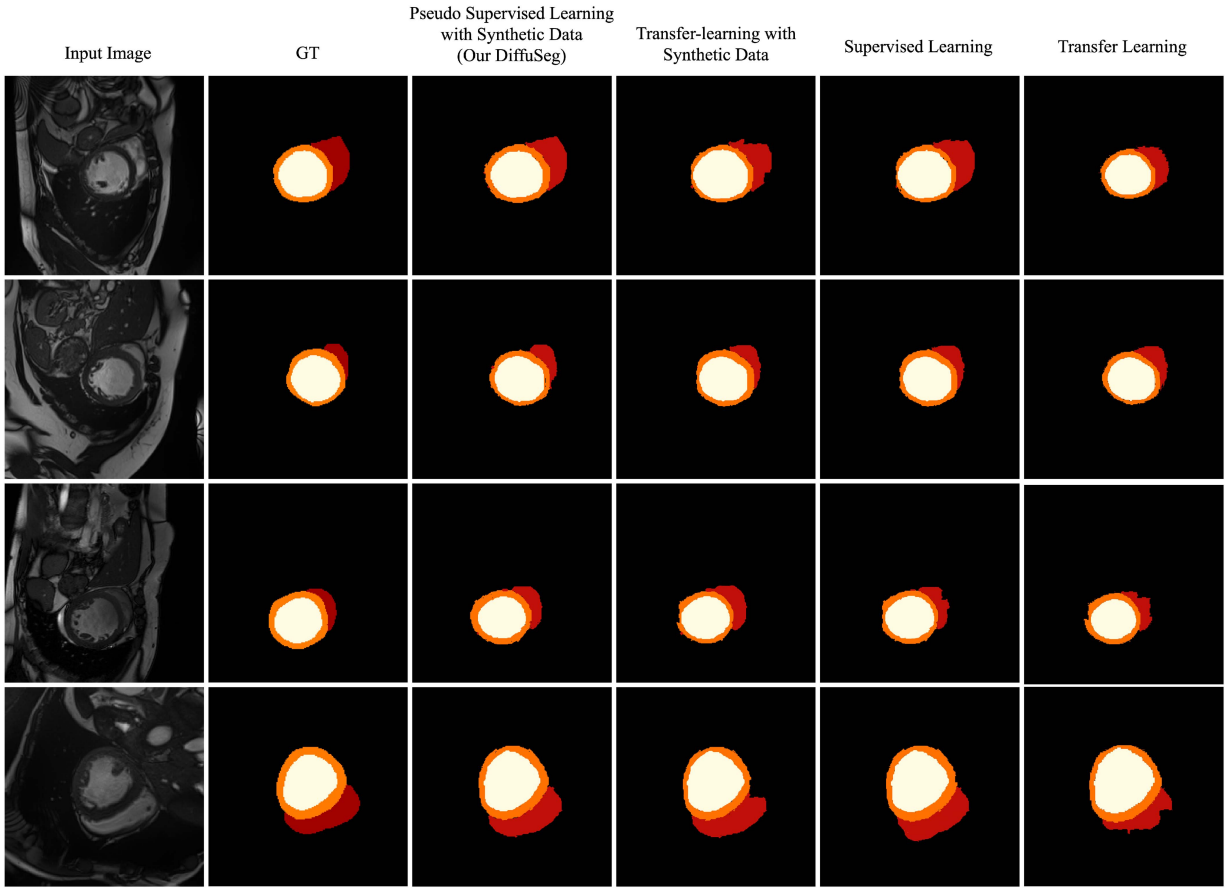

Fig. 9. The cardiac ventricle segmentation results on MSCMRseg dataset using different training approaches.

TABLE III
DICE COMPARISON WITH TTA [41], CM-TRANCAF [42], DEY-NET [43] AND DDSP [44] METHODS FOR SEGMENTATION TASKS ON IOSTAR, MSCMR AND MM-WHS DATASETS. DATA USED FOR TRAINING: L_S: SOURCE LABEL, L_T: TARGET LABEL, I_T: TARGET IMAGE; METRICS USED: DICE ↑

Method	Data Access			IOSTAR	MSCMR	MM-WHS
	L_S	L_T	I_T			
TTA [43]	✓	×	×	48.3	66.2	69.4
CM-TranCaF [44]	✓	×	✓	53.8	71.8	74.2
DeY-Net [45]	✓	×	✓	51.7	68.6	69.6
DDSP [46]	✓	×	✓	54.5	71.2	76.2
Ours	✓	×	✓	**56.1**	**77.2**	**79.8**

The best values are highlighted in bold.

and reliability in capturing intricate structures within complex anatomical regions.

Furthermore, we compared our method with the State-of-the-art domain transfer segmentation methods in [41], [42], [43], [44] and evaluate them on all three datasets. The results presented in Table III clearly demonstrate that the our method outperforms existing state-of-the-art approaches, achieving the highest Dice scores across all datasets. The improvement is particularly significant on the MSCMR and MM-WHS datasets, where the proposed approach leverages target domain information more effectively than its competitors. These results validate the efficacy of our diffusion-based model in generating high-quality segmentations and its potential as a robust tool for medical image analysis across multiple domains. In addition, Table.IV provides a comprehensive comparison of whole-heart segmentation performance across different methods on the MM-WHS testing dataset. Our proposed method consistently outperforms the competing approaches across most metrics: 1) Our method achieves the highest Dice scores across all anatomical structures, particularly excelling in segmenting the LV (81.6%), RV (80.3%), and Ao (75.7%), demonstrating superior segmentation quality; 2) Our approach also records the highest Jaccard indices, indicating better overlap with the ground truth, especially for RV (0.89), LA (0.86), and Ao (0.83); 3) In terms of Hausdorff Distance, which measures boundary accuracy, the proposed method shows the lowest values across all regions, notably improving precision in boundary delineation for LV (7.51 mm), RV (9.38 mm), and Ao (5.83 mm). Therefore, our method demonstrates substantial improvements in both segmentation accuracy and boundary delineation compared to other state-of-the-art methods. Its robust performance across all key metrics and anatomical structures highlights its effectiveness for whole-heart segmentation, making it a promising approach for clinical applications. The results validate its ability to accurately capture complex anatomical

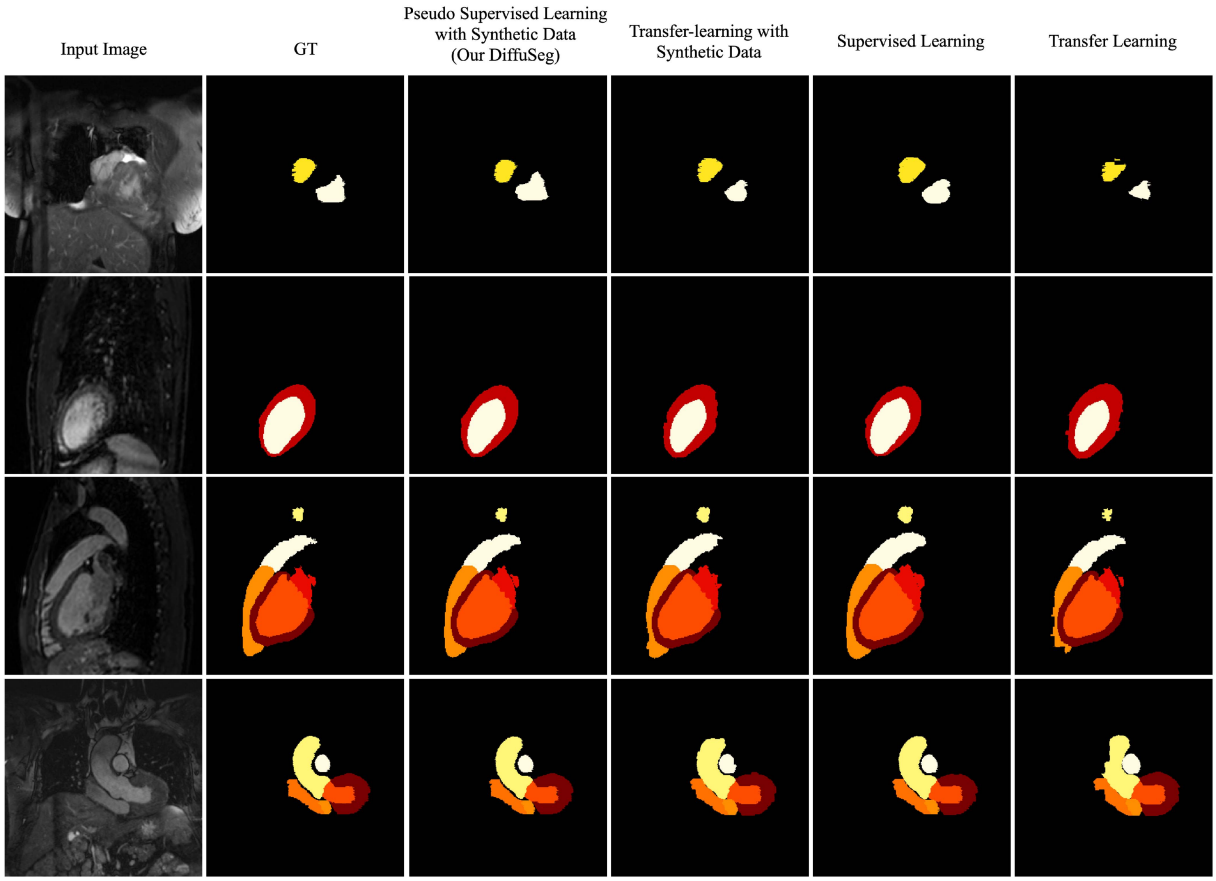

Fig. 10. The heart segmentation results on MM-WHS dataset using different training approaches.

TABLE IV
RESULTS OF DIFFERENT WHOLE-HEART SEGMENTATION METHODS ON MM-WHS TESTING DATASET

	Methods	LV	RV	LA	RA	Myo	Ao	PA	WHS
Dice	TTA [43]	74.1	76.3	71.2	65.6	61.4	74.6	63.8	69.4 ± 0.3
	CM-TranCaF [44]	75.5	74.7	76.4	73.2	66.3	71.6	61.7	74.2 ± 0.2
	DeY-Net [45]	71.8	72.9	71.3	67.1	64.9	69.3	60.2	69.6 ± 0.4
	DDSP [46]	78.6	76.8	77.1	75.9	65.9	73.7	65.3	76.2 ± 0.2
	Ours	**81.6**	**80.3**	**79.2**	**77.1**	**68.3**	**75.7**	**67.6**	**79.8 ± 0.2**
Jaccard	TTA [43]	0.81	0.79	0.81	0.77	0.77	0.81	0.72	0.77 ± 0.03
	CM-TranCaF [44]	0.82	0.79	0.77	0.75	0.74	0.80	0.71	0.80 ± 0.04
	DeY-Net [45]	0.86	0.83	0.78	0.75	0.73	0.77	0.78	0.78 ± 0.06
	DDSP [46]	0.81	0.80	0.77	0.78	0.75	0.81	0.76	0.79 ± 0.11
	Ours	**0.88**	**0.89**	**0.86**	**0.83**	**0.84**	**0.83**	**0.80**	**0.84 ± 0.04**
HD (mm)	TTA [43]	10.12	12.34	13.53	15.92	14.91	8.96	10.26	20.26 ± 8.81
	CM-TranCaF [44]	9.88	10.80	9.46	14.19	13.28	6.89	8.28	17.42 ± 9.25
	DeY-Net [45]	8.34	11.31	13.24	14.56	14.78	7.88	9.41	18.84 ± 6.87
	DDSP [46]	8.10	10.2	10.91	13.53	12.78	6.32	8.12	17.43 ± 8.55
	Ours	**7.51**	**9.38**	**8.11**	**12.32**	**10.65**	**5.83**	**6.67**	**15.39 ± 6.12**

features and reduce boundary discrepancies, crucial for precise medical image analysis.

V. DISCUSSION

A. Insights Into DiffuSeg

In medical image analysis, diffusion models have emerged as powerful tools for image generation and facilitating various clinical applications. Our DiffuSeg achieved superior performance within this landscape are the integration of conditional diffusion model and strategy of blending techniques, which offer valuable insights that hold great promise for advancing medical image generation and the downstream tasks.

Our proposed DiffuSeg method leverages domain-driven diffusion and feature factorization techniques, which are specifically designed to preserve clinical anatomy and ensure the clinical quality of the synthesized images. These mechanisms

are crucial in maintaining the visual fidelity and clinical relevance of the generated images, directly addressing the demands of medical settings. DiffuSeg employs a conditional diffusion model that is guided by domain-specific features extracted from real medical images. This ensures that the generated images closely mimic the appearance, texture, and structural consistency of clinically relevant data. By embedding domain-specific knowledge into the diffusion process, DiffuSeg can generate images that maintain the existing anatomical structures, which is critical for clinical applications where accurate representation of anatomy is essential.

In the future, we would like to develop DiffuSeg's ability on the potential to aid in disease-specific image analysis. By allowing clinicians to input pertinent clinical information, our DiffuSeg can generate images tailored to a particular pathology or clinical scenario. This customization enhances diagnostic accuracy and provides insights that align with clinical requirements.

Furthermore, the application of blending techniques contributes to the creation of images that are visually coherent and clinically accurate. Our DiffuSeg integrates the blending techniques within the diffusion process, which is a crucial step toward ensuring clinically meaningful outcomes. In medical image generation, maintaining the fidelity of anatomical structures and preserving the integrity of relevant features are paramount. By employing strategies such as spatially varying noise modulation, the blending process mitigates discrepancies between enhanced regions and the surrounding anatomy, yielding images that are both diagnostically informative and visually consistent. The blending technique ensures that the boundaries of segmented regions are smooth and integrated with neighbouring structures. This is imperative for tasks like tumour delineation, and organ segmentation over time, which is the future work to investigate.

While diffusion models are known for their computational intensity, we adopted several strategies to enhance efficiency. First, we utilized GPU-accelerated training (Quadro RTX 8000) and incorporated the AdamW optimizer, which allows faster convergence with reduced computational overhead. Additionally, our early stopping routine terminated training when validation performance plateaued, thereby preventing unnecessary computations. By employing batch processing, our model effectively handled multiple inputs in parallel during both training and inference, significantly reducing runtime. These optimizations allowed us to maintain a high level of performance, achieving competitive segmentation results with state-of-the-art methods, while mitigating computational demands. Future work will focus on further optimizing the architecture, such as using model pruning or lightweight diffusion variants, to facilitate real-time deployment scenarios.

B. Strengths and Challenges

Our DiffuSeg utilizes the diffusion model for generating images in the context of image segmentation, which offers a novel approach to address the challenges of data scarcity and annotation costs. First, our diffusion model aims to denoise images by iteratively refining noise, which can lead to diversity in the generated images. Therefore, the synthetic data could fully capture the variations present in real-world data. Segmentation models trained on such data could generalize well to real-world scenarios, impacting their performance on complex images. Second, DiffuSeg generates images with the given labels, which exposes the segmentation neural network to ground truth annotations available, and thus avoids human annotations. This introduces potential consistencies between the generated images and their corresponding annotations, which leads to suboptimal training and segmentation results.

The evaluation of DiffuSeg on diverse datasets such as MNIST, retinal fundus images, and MRI heart images demonstrates its potential to generalize across various types of medical image data. This selection of datasets covers a broad spectrum of image characteristics. Together, these evaluations indicate that DiffuSeg has a broad adaptability to various image types and conditions. Its foundation on domain-driven diffusion and feature factorization allows it to adapt to different image modalities without extensive retraining. While specific evaluations on CT scans, X-rays, or histopathological images were not included, the inherent flexibility of DiffuSeg's architecture suggests it can generalize effectively to other imaging modalities, including those with different noise levels, resolutions, and structures. Future work can further validate DiffuSeg's generalization ability by testing on additional medical datasets like CT scans, X-rays, and histopathology. However, the current results strongly indicate that DiffuSeg's design is not limited to a single image type, making it a versatile solution for a wide range of medical imaging challenges.

However, our approach is not without its limitations and challenges, which must be carefully considered for effective integration into image segmentation pipelines. On one hand, diffusion models focus on enhancing visual quality, but they might not always preserve semantic accuracy. The generated images could contain artefacts, anomalies, or subtle shifts in object boundaries that might affect the alignment between image content and segmentation labels. Training on such images could lead to confusion for segmentation models, impacting their ability to accurately delineate objects. On the other hand, the quality and characteristics of the generated images are highly dependent on the configuration of the diffusion model, including hyperparameters and training settings. Improper tuning of these parameters could lead to undesirable artifacts or unnatural visual effects, which could subsequently affect the quality of the generated dataset and the performance of segmentation models.

VI. Conclusion

In this paper, we presented DiffuSeg, a two-stage framework, for medical image segmentation. Specifically, in the first stage, to address the issue of domain shift among different datasets, we introduced a diffusion model condition on the domain knowledge extracted by our proposed FFD-VAE to synthesize the images with the target domain. In the second stage, we trained a segmentation model using the given label maps and the corresponding synthetic images. DiffuSeg obtains competitive

image generation ability to SOTA I2I translation methods while contributing the data augmentation for the segmentation tasks. Furthermore, our proposed method also shows the potential ability to synthesize disease-related image textures, which could be our future work by exploring these synthetic data for disease analysis.

ACKNOWLEDGMENT

The authors would like to thank Tobias Højgaard Dovmark and Sile Hu for reviewing the paper, and acknowledge the funding from Novo Nordisk to support this work.

REFERENCES

[1] L. Zhang et al., "Learning from multiple annotators for medical image segmentation," *Pattern Recognit.*, vol. 138, 2023, Art. no. 109400.
[2] Y. Liu, S. Zhang, Y. Li, and J. Yang, "Learning to adapt via latent domains for adaptive semantic segmentation," in *Proc. 32nd Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 1167–1178, 2021.
[3] S. Kohl et al., "A probabilistic U-Net for segmentation of ambiguous images," *Adv. Neural Inf. Process. Syst.*, 2018, pp. 6965–6975.
[4] M. Pilch et al., "Automated segmentation of retinal blood vessels in spectral domain optical coherence tomography scans," *Biomed. Opt. Exp.*, vol. 3, no. 7, pp. 1478–1491, 2012.
[5] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.
[6] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.
[7] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[8] T. Cemgil, S. Ghaisas, K. Dvijotham, S. Gowal, and P. Kohli, "The autoencoding variational autoencoder," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 15077–15087, 2020.
[9] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, and C. Zhang, "Marginal contrastive correspondence for guided image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10663–10672.
[10] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10619–10629.
[11] C. Saharia et al., "Palette: Image-to-image diffusion models," in *Proc. ACM SIGGRAPH Conf.*, 2022, pp. 1–10.
[12] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. Artif. Neural Netw. Mach. Learn.*, Springer, 2018, pp. 270–279.
[13] K. Kushibar et al., "Supervised domain adaptation for automatic subcortical brain structure segmentation with minimal user interaction," *Sci. Rep.*, vol. 9, no. 1, pp. 6742, 2019.
[14] B. Billot et al., "Synthseg: Segmentation of brain MRI scans of any contrast and resolution without retraining," *Med. Image Anal.*, vol. 86, 2023, Art. no. 102789.
[15] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1521–1528.
[16] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 268–284.
[17] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[18] Y. Yang, Z. Feng, M. Song, and X. Wang, "Factorizable graph convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 20286–20296.
[19] N. Jakobi, P. Husbands, and I. Harvey, "Noise and the reality gap: The use of simulation in evolutionary robotics," in *Proc. Adv. Artif. Life: 3rd Eur. Conf. Artif. Life Granada*, 1995, vol. 3, pp. 704–720.
[20] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. K. Carla, "An open urban driving simulator," in *Proc. Conf. Robot Learn.*, 2017, pp. 1–16.
[21] X. Liu, P. Sanchez, S. Thermos, A. Q. O'Neil, and S. A. Tsaftaris, "Learning disentangled representations in the imaging domain," *Med. Image Anal.*, vol. 80, 2022, Art. no. 102516.
[22] M. Hu et al., "Knowledge distillation from multi-modal to mono-modal segmentation networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2020, pp. 772–781.
[23] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 8780–8794.
[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.
[25] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," in *Readings in Computer Vision*. Amsterdam, Netherlands: Elsevier, 1987, pp. 671–679.
[26] Y. LeCun et al., "Gradient-based learning applied to document recognition," *Proc. IEEE Proc. IRE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[27] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.
[28] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.
[29] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Proc. 28th Adv. Neural Inf. Process. Syst.*, 2015, pp. 3483–3491.
[30] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5143–5153.
[31] X. Zhou et al., "Cocosnet v2: Full-resolution correspondence learning for image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11465–11475.
[32] J. Zhang et al., "Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores," *IEEE Trans. Med. Imag.*, vol. 35, no. 12, pp. 2631–2644, Dec. 2016.
[33] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. V. Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.
[34] X. Zhuang, "Multivariate mixture model for myocardial segmentation combining multi-source images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2933–2946, Dec. 2019.
[35] K. Zhang and X. Zhuang, "Cyclemix: A holistic strategy for medical image segmentation from scribble supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Los Alamitos, CA, USA, Jun. 2022, pp. 11646–11655.
[36] O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.
[37] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Med. Image Anal.*, vol. 31, pp. 77–87, 2016.
[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Springer, 2015, pp. 234–241.
[39] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 203–210, Mar. 2000.
[40] Y. Xue, Y. Li, K. K. Singh, and Y. J. Lee, "GIRAFFE HD: A high-resolution 3D-aware generative model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18440–18449.
[41] H. Basak and Z. Yin, "Quest for clone: Test-time domain adaptation for medical image segmentation by searching the closest clone in latent space," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Springer, 2024, pp. 555–566.
[42] Y. Ding et al., "A cascaded framework with cross-modality transfer learning for whole heart segmentation," *Pattern Recognit.*, vol. 147, 2024, Art. no. 110088.
[43] R. Wen, H. Yuan, D. Ni, W. Xiao, and Y. Wu, "From denoising training to test-time adaptation: Enhancing domain generalization for medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 464–474.
[44] B. Zheng et al., "Dual domain distribution disruption with semantics preservation: Unsupervised domain adaptation for medical image segmentation," *Med. Image Anal.*, vol. 97, 2024, Art. no. 103275.