DiffAtlas: GenAI-fying Atlas Segmentation via Image-Mask Diffusion

Hantao Zhang [†1,2][0009−0008−9849−3507], Yuhe Liu [†3][0009−0006−2637−8839], Jiancheng Yang [‡1][0000−0003−0008−9036], Weidong Guo[2][0009−0002−7881−9510], Xinyuan Wang[3][0009−0005−7577−1648], and Pascal Fua[1][0000−0002−6702−9970]

[1] Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland
[2] University of Science and Technology of China (USTC), Hefei, China
[3] Beihang University, Beijing, China

Abstract. Accurate medical image segmentation is crucial for precise anatomical delineation. Deep learning models like U-Net have shown great success but depend heavily on large datasets and struggle with domain shifts, complex structures, and limited training samples. Recent studies have explored diffusion models for segmentation by iteratively refining masks. However, these methods still retain the conventional image-to-mask mapping, making them highly sensitive to input data, which hampers stability and generalization. In contrast, we introduce DiffAtlas, a novel generative framework that models both images and masks through diffusion during training, effectively "GenAI-fying" atlas-based segmentation. During testing, the model is guided to generate a specific target image-mask pair, from which the corresponding mask is obtained. DiffAtlas retains the robustness of the atlas paradigm while overcoming its scalability and domain-specific limitations. Extensive experiments on CT and MRI across same-domain, cross-modality, varying-domain, and different data-scale settings using the MMWHS and TotalSegmentator datasets demonstrate that our approach outperforms existing methods, particularly in limited-data and zero-shot modality segmentation. Code is available at https://github.com/M3DV/DiffAtlas.

Keywords: Atlas · GenAI · Diffusion · Few-Shot · Cross-Modality

1 Introduction

In recent years, deep learning-based segmentation methods have achieved remarkable success in the field of medical image analysis [1]. Standard segmentation architectures, such as the U-Net [20], are designed as feedforward neural networks and have demonstrated strong performance across a wide range of applications. The advantages of such feedforward models include their computational efficiency and ease of deployment. However, these methods also have notable limitations: they often struggle with fine-grained details, lack robustness

[†] Equal contribution. Conducted during the authors' research internships at EPFL.
[‡] Corresponding author: J. Yang (jiancheng.yang@epfl.ch), who led the project.

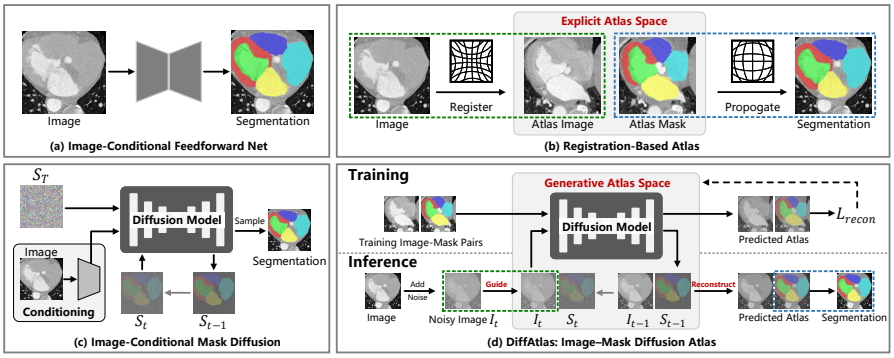

Fig. 1: **Illustration of different segmentation paradigms.** (a) Directly maps the input image to the segmentation mask; (b) Uses registration to align a labeled atlas with the target image and propagate atlas labels; (c) Conditioned on the input image, achieves image-to-segmentation mask mapping using diffusion; (d) Train: Parameterizes an atlas and simultaneously generates image-mask pairs. Test: Uses the noisy image as guidance to generate the corresponding atlas.

to domain shifts, and suffer from complex morphological variations and continuity issues when faced with challenging scenarios. To address these issues, many approaches have been proposed, such as incorporating attention mechanisms [23], multi-scale feature extraction [12], and adversarial training [16]. Despite these efforts, challenges persist, especially with complex anatomical structures under limited samples [7] and multi-modal cross-domain issues [6].

With the emergence of generative AI (GenAI) [11], this novel paradigm has been introduced into many medical image analysis tasks [27], such as segmentation [24] and anomaly detection [13], to enhance performance. For instance, Wu et al. [24] incorporated image-based control into the diffusion to "GenAI-fy" segmentation. However, such diffusion-based methods [17] are highly sensitive to input images and fail to maintain robustness when crossing domains.

In this work, we revisit atlas-based segmentation [2,6]. This method, which involves image registration, label propagation, and refinement, was once the gold standard for medical image segmentation before the deep learning era [8]. It offers advantages such as interpretability, anatomical consistency, and robustness to small misalignments. However, atlas-based methods face key challenges: they are not easily scalable, struggle with fine-grained structures, and require labor-intensive, domain-specific atlases [28]. Despite these drawbacks, the robustness of atlas-based methods to misalignments remains a key advantage: even with misalignments, the segmentation remains meaningful [22]. This stands in stark contrast to many feedforward neural network-based methods [4], which often fail catastrophically when errors occur during segmentation.

In contrast, we propose a fundamentally different neural network segmentation approach based on the diffusion paradigm, parameterized for atlas-based segmentation, named **DiffAtlas**. Our method builds upon the principles of atlas-

based registration and label propagation, while leveraging modern generative AI techniques to GenAI-fy the process. Specifically, DiffAtlas parameterizes the atlas as a generative model, capturing the joint distribution of **image-mask pairs** during training. During inference, DiffAtlas performs segmentation through a guided generative process, where the noisy image progressively replaces the predicted image at each step in the image-mask pair. This process **guides the pair toward the target**, generating the corresponding mask. We conduct extensive experiments on two whole-heart segmentation datasets: MM-WHS [29] and TotalSegmentator [29], covering various settings including base segmentation, few-shot and extremely few-shot sample settings, as well as zero-shot cross-modality segmentation. Our results demonstrate that the proposed method achieves state-of-the-art performance in these scenarios.

2 Method

Atlas-based segmentation methods are well-known for their ability to enforce anatomical consistency and preserve global structural integrity in segmentation masks [18]. However, their reliance on predefined atlases and computationally intensive registration steps makes them inherently inflexible and difficult to scale to diverse anatomical variations [9].

To tackle these challenges, we propose DiffAtlas, a novel framework that combines atlas priors with the flexibility of generative diffusion models, as illustrated in Fig. 1(d). DiffAtlas constructs a generative atlas space through a learned diffusion process, which implicitly encodes diverse anatomical structures. This eliminates the need for explicit atlas registration while preserving the anatomical consistency central to atlas-based methods.

By modeling the joint prior of images and segmentation masks, DiffAtlas generates anatomically consistent and tightly aligned pairs. During inference, a noisy image (pre-noised from the input image at the corresponding time step t) replaces the predicted image, guiding the model to output the target pair and, consequently, the corresponding mask. Through this, DiffAtlas seamlessly integrates the target image's global anatomical priors into the inference process on any dataset, enabling segmentation without the need for additional training. At the same time, its reliance on a learned generative atlas space, rather than explicit registration, significantly reduces both time and spatial complexity, ensuring greater flexibility and scalability, and making it a robust solution.

2.1 Revisiting Previous Methods

Image-Conditional Feedforward Network. As illustrated in Fig. 1(a), feedforward neural networks, such as U-Net [20], are widely used for image segmentation tasks.These models learn a deterministic mapping from the input image $\mathbb{I}$ to the segmentation mask $\mathbb{S}$, denoted as $f_\theta : \mathbb{I} \to \mathbb{S}$. The training objective typically minimizes the cross-entropy loss, dice loss or their combination. While efficient, these methods entirely rely on the input image, making them highly sensitive to

input conditions. For instance, noise, artifacts, or incomplete information in the input image can significantly degrade segmentation accuracy. Furthermore, the lack of global anatomical priors limits the ability of these methods to enforce structural consistency, leading to poor robustness and generalization, particularly for complex anatomical regions or across diverse datasets [14].

Image-Conditional Mask Diffusion. Image-Conditional Mask Diffusion extends the image-conditional paradigm by modeling the conditional distribution $p(\mathbf{S}|\mathbf{I})$, where $\mathbf{S}$ is the segmentation mask and $\mathbf{I}$ is the input image. By introducing a probabilistic framework, Image-Conditional Mask Diffusion decomposes the generation of the segmentation mask into a step-by-step denoising process. The forward process $q(\mathbf{S}_t|\mathbf{S}_{t-1})$ progressively corrupts the ground-truth mask $\mathbf{S}$ by adding Gaussian noise. Specifically, the noisy mask $\mathbf{S}_t$ at step t is obtained as:

$$\mathbf{S}_t = \sqrt{\alpha_t}\mathbf{S}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}), \tag{1}$$

where α_t is a predefined noise schedule that controls the balance between the signal and noise, and $\mathcal{I}$ is the identity matrix.

The reverse process $p_\theta(\mathbf{S}_{t-1}|\mathbf{S}_t, \mathbf{I})$ iteratively refines the noisy mask by learning the conditional distribution:

$$p_\theta(\mathbf{S}_{t-1}|\mathbf{S}_t, \mathbf{I}) = \mathcal{N}(\mathbf{S}_{t-1}; \mu_\theta(\mathbf{S}_t, \mathbf{I}, t), \mathbf{\Sigma}_\theta(\mathbf{S}_t, \mathbf{I}, t)), \tag{2}$$

where μ_θ and $\mathbf{\Sigma}_\theta$ are the learned mean and covariance functions. The introduction of diffusion enables this paradigm to generate plausible masks through a step-by-step process. However, it remains fundamentally tied to the image-conditional paradigm, inheriting its limitations in the mapping from image to mask.

Registration-Based Atlas. Atlas-based methods address the lack of global priors by leveraging predefined labeled atlases $(\mathbf{A}_I, \mathbf{A}_S)$. These methods register the atlas image $\mathbf{A}_I$ to the target image $\mathbf{I}$, propagating the atlas labels $\mathbf{A}_S$ through the deformation field ϕ^*:

$$\mathbf{S} = \mathbf{A}_S \circ \phi^*, \tag{3}$$

where the deformation field ϕ^* is computed by minimizing an energy function:

$$\phi^* = \arg\min_\phi \mathcal{D}(\mathbf{I}, \mathbf{A}_I \circ \phi) + \lambda\mathcal{R}(\phi), \tag{4}$$

with $\mathcal{D}$ measuring similarity between $\mathbf{I}$ and the deformed atlas $\mathbf{A}_I \circ \phi$, and $\mathcal{R}$ being a regularization term. Although effective in enforcing anatomical consistency, these methods are computationally expensive and heavily depend on the quality of the predefined atlas. Their performance can degrade significantly when the atlas fails to represent diverse anatomical variations.

2.2 DiffAtlas: Image–Mask Diffusion Atlas

Joint Image-Mask Prior as Generative Atlas Space. DiffAtlas models the input image and segmentation mask as a pair $(\mathbf{I}, \mathbf{S})$ within a generative atlas space,

learned through a diffusion process. Unlike traditional methods that rely on predefined atlases and explicit registration, this learned space captures global anatomical consistency while remaining flexible to diverse anatomical variations. By training the model to reconstruct clean pairs $(\mathbf{I}, \mathbf{S})$ from noisy versions $(\mathbf{I}_t, \mathbf{S}_t)$, DiffAtlas learns the joint distribution $p_\theta(\mathbf{I}, \mathbf{S})$. This ensures that the segmentation mask $\mathbf{S}$ is both anatomically consistent and naturally aligned with its corresponding image $\mathbf{I}$.

Noisy Image Guidance for Input-Conditioned Sampling. To ensure that the generated image-mask pair aligns with the input image during inference, DiffAtlas employs a noisy image replacement strategy. During the replacement process, the noisy version of the input image $\mathbf{I}_{\text{input}}$ is introduced at each timestep t to guide the sampling process. Specifically, the noisy image $\mathbf{I}_t$ is replaced with:

$$\mathbf{I}_t = \sqrt{\bar{\alpha}_t}\mathbf{I}_{\text{input}} + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{\epsilon}$ is Gaussian noise sampled from $\mathcal{N}(\mathbf{0}, \mathcal{I})$, and $\bar{\alpha}_t$ is the cumulative product of the noise schedule α_t up to timestep t. By conditioning the diffusion process on the input image, this guidance ensures that the sampled segmentation mask $\mathbf{S}$ aligns closely with the anatomical structures of $\mathbf{I}_{\text{input}}$, while preserving the global consistency enforced by the generative atlas space.

2.3 DiffAtlas Implementation

Training Process. During training, DiffAtlas optimizes a neural network to predict the added noise $\boldsymbol{\epsilon}$ on both the image $\mathbf{I}$ and the segmentation mask $\mathbf{S}$ in the forward diffusion process. The objective minimizes the mean squared error between the predicted noise $\boldsymbol{\epsilon}_\theta$ and the true noise $\boldsymbol{\epsilon}$:

$$\mathcal{L}_{\text{train}} = \mathbb{E}_{(\mathbf{I},\mathbf{S}),t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{I}_t, \mathbf{S}_t, t)\|^2\right]. \tag{6}$$

Modeling joint distributions of image-mask pairs has been explored in prior works, such as for representation learning or feature extraction [21]. We leverage this to construct a generative atlas space. Our primary focus is on guiding matching within the implicit atlas space for image segmentation. Furthermore, to enhance spatial continuity and structural coherence, DiffAtlas represents the mask $\mathbf{S}$ using a signed distance function (SDF), which encodes the distance to the mask boundary [3], improving the capture of fine anatomical details and ensuring smooth transitions between regions.

Inference Process. During inference, DiffAtlas starts with a randomly initialized noisy image-mask pair $(\mathbf{I}_T, \mathbf{S}_T)$ and refines it iteratively through the reverse diffusion process over T timesteps. At each timestep t, the noisy image $\mathbf{I}_t$ is replaced with the noisy version of the input image $\mathbf{I}_{\text{input}}$. This replacement gently anchors the generative process to the anatomical structures of the input, ensuring that the model remains guided by its features throughout the denoising process.

Table 1: **Segmentation under full training setting.**

Methods	Myo		LV		LA		RA		RV		Average	
	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD
Full *TS training set → TS test set*												
(a) ICF [10]	80.22	83.67	78.06	75.16	88.32	82.24	83.21	73.43	88.48	81.62	83.66	79.22
(b) RBA [6]	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
(c) ICMD [24]	77.80	84.30	74.20	73.34	87.44	81.64	82.72	76.25	86.53	79.68	81.74	79.04
(d) DA (Ours)	83.52	88.87	77.21	75.70	90.13	86.35	86.52	80.76	88.36	82.06	85.17	82.74
Full *MMWHS-CT training set → MMWHS-CT test set*												
(a) ICF [10]	57.88	47.18	60.05	22.41	73.31	42.62	49.68	25.35	69.24	43.25	62.03	36.16
(b) RBA [6]	51.13	40.56	65.28	30.66	60.04	31.13	66.96	39.54	55.38	29.47	59.76	34.27
(c) ICMD [24]	58.03	52.63	41.98	26.05	68.51	45.64	66.47	41.81	39.98	25.56	54.99	38.34
(d) DA (Ours)	78.02	73.75	87.27	74.74	86.68	74.41	80.25	65.87	84.02	66.79	83.25	71.11
Full *MMWHS-MRI training set → MMWHS-MRI test set*												
(a) ICF [10]	41.11	38.90	51.63	20.28	67.63	31.98	49.66	22.00	50.05	28.99	52.02	28.43
(b) RBA [6]	38.28	42.01	53.10	30.49	63.19	34.32	59.34	28.18	49.76	26.84	52.73	32.37
(c) ICMD [24]	53.14	55.40	32.97	21.98	65.51	40.97	61.84	37.48	39.87	27.10	50.67	36.59
(d) DA (Ours)	57.97	46.42	70.67	41.11	75.70	43.93	73.11	44.14	65.77	41.64	68.64	43.45

As the diffusion unfolds, the noisy pair $(\mathbf{I}_t, \mathbf{S}_t)$ progressively transforms into a clean and coherent pair $(\mathbf{I}, \mathbf{S})$. In this final pair, the generated image $\mathbf{I}$ naturally converges to closely resemble the input image $\mathbf{I}_{\text{input}}$, while the segmentation mask $\mathbf{S}$ emerges as an inherently compatible counterpart. Without any explicit enforcement, the mask $\mathbf{S}$ aligns seamlessly with the anatomical structures of $\mathbf{I}_{\text{input}}$, reflecting both global consistency and local precision. This effortless alignment arises naturally from the learned joint distribution in the generative atlas space, where the anatomical priors ensure harmony between the image and mask. By combining the implicit guidance of the generative prior with direct conditioning on the input image, DiffAtlas achieves segmentation results that are both precise and intuitively consistent.

3 Experiments

3.1 Setup

Dataset. This work uses two datasets: MM-WHS [29] and TotalSegmentator (TS) [29] dataset. For both datasets, we retained five labels for full heart segmentation: Myocardium of the LV (Myo), LV blood cavity (LV), left atrium (LA), right atrium (RA), and right ventricle (RV). We used the cardiac CT portion of the TS, which contains 746 cases, reserving 20% for testing. For MM-WHS, we used 20 available CT cases and 20 MRI cases.

The specific experimental settings are as follows: **Full setting:** For the TS, MMWHS-CT, and MMWHS-MRI datasets, we used 80% for training and 20% for testing. **Few-shot setting:** The test set remains the same as in the previous setting, comprising 20% of the total, with two configurations: 2-shot (2 training samples) and 4-shot (4 training samples). **Zero-shot cross-modality setting:** We used the same train-test split, dividing the settings into three groups: large

Table 2: **Segmentation under few-shot training setting.**

Methods	Myo		LV		LA		RA		RV		Average	
	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD
2-shot *MMWHS-CT training samples → MMWHS-CT test set*												
(a) ICF [10]	26.09	33.54	49.66	16.93	69.70	42.22	39.16	21.94	50.17	15.83	46.96	26.09
(b) RBA [6]	44.29	39.16	58.83	13.85	67.99	31.68	61.83	33.45	59.83	32.70	58.55	30.17
(c) ICMD [24]	38.26	39.59	38.31	24.22	49.05	33.38	50.91	25.70	29.31	19.08	41.17	28.39
(d) DA (Ours)	70.30	58.45	80.86	59.60	82.39	45.30	78.01	58.42	77.08	55.88	77.73	55.53
4-shot *MMWHS-CT training samples → MMWHS-CT test set*												
(a) ICF [10]	35.35	31.13	60.25	18.71	71.78	31.14	40.40	16.29	64.39	40.73	54.43	27.60
(b) RBA [6]	49.61	41.69	61.03	34.53	71.21	35.72	58.01	39.12	65.95	33.88	61.16	36.99
(c) ICMD [24]	43.32	40.74	38.49	24.55	62.80	38.74	51.05	25.59	28.65	19.34	44.86	29.79
(d) DA (Ours)	72.05	53.21	83.07	56.69	86.72	44.40	74.56	41.35	79.50	55.69	79.18	50.27
2-shot *MMWHS-MRI training samples → MMWHS-MRI test set*												
(a) ICF [10]	25.95	27.15	23.63	13.68	35.33	16.32	37.52	18.47	13.16	12.73	27.12	17.67
(b) RBA [6]	33.74	35.17	40.24	25.99	51.59	28.73	61.10	33.63	38.38	23.38	45.01	29.38
(c) ICMD [24]	45.39	44.72	7.90	6.97	61.35	36.69	56.13	33.71	22.43	17.91	38.64	28.00
(d) DA (Ours)	49.56	42.06	57.48	24.71	61.05	28.52	64.59	30.19	51.26	34.55	56.79	32.01
4-shot *MMWHS-MRI training samples → MMWHS-MRI test set*												
(a) ICF [10]	33.12	32.30	48.79	25.47	55.48	26.19	41.42	19.53	31.78	19.57	42.12	24.61
(b) RBA [6]	47.88	43.62	49.24	24.56	59.99	30.42	65.89	35.18	49.78	28.41	54.56	32.44
(c) ICMD [24]	49.19	44.88	20.64	16.48	65.81	35.22	55.19	33.74	9.44	17.09	40.06	29.48
(d) DA (Ours)	54.34	48.42	69.69	39.43	72.09	44.78	75.92	48.09	63.22	41.46	67.05	44.44

data with different source CT to MRI, small data with same-source CT to MRI, and MRI to CT, as shown in Tab. 3.

Method Comparison. We compared our method with publicly available state-of-the-art methods from each category shown in Fig.1, including **(a) ICF** (Image-Conditional Feedforward): nnU-Net[10]; **(b) RBA** (Registration-Based Atlas): CMMAS [6]; **(c) ICMD** (Image-Conditional Mask Diffusion): MedSegDiffv2 [24]; **(d) DA (Ours)**: DiffAtlas. For fairness, all models use the same data augmentation, with no additional augmentation for nnU-Net.

3.2 Performance Evaluation

For all experiments, we followed Metrics Reloaded [15] and related works [6,24] and employed the Dice similarity coefficient (Dice) and normalized surface distance (NSD) [5] as metrics. We use red to highlight the best.

Full Training. As shown in Tab. 1, under the same-dataset setting, our proposed (d) DA (DiffAtlas) outperforms state-of-the-art methods, including (a) ICF (nnU-Net[10]), on both the large-scale CT TS dataset and the smaller MMWHS CT and MRI datasets, demonstrating its effectiveness and scalability. However, (b) RBA (Registration-Based Atlas): CMMAS [6] is limited by the high time and space complexity of the atlas paradigm, restricting its scalability. As a result, it could not be executed on the large-scale TS dataset, with results marked as 'NaN'.

Table 3: **Segmentation under zero-shot cross-domain evaluation setting.**

Methods	Myo		LV		LA		RA		RV		Average	
	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD
Full TS training set ONLY → MMWHS-MRI dataset												
(a) ICF [10]	28.36	34.71	45.57	21.13	51.66	22.89	50.75	24.71	45.25	22.80	44.32	25.25
(b) RBA [6]	24.13	29.63	40.79	24.28	46.09	23.59	39.48	22.98	42.03	24.02	38.51	24.90
(c) ICMD [24]	29.26	24.01	75.53	62.06	57.86	30.02	66.18	43.26	47.74	37.07	55.31	39.29
(d) DA (Ours)	62.97	52.31	81.49	66.65	83.25	37.85	80.13	57.95	74.37	53.89	76.44	53.73
Full MMWHS-CT dataset ONLY → MMWHS-MRI dataset												
(a) ICF [10]	26.89	28.38	45.95	19.39	41.96	19.34	31.33	13.69	30.18	16.34	35.26	19.42
(b) RBA [6]	31.70	34.53	42.01	24.88	50.71	27.71	43.05	29.92	39.72	23.55	41.44	28.12
(c) ICMD [24]	14.09	14.56	48.60	37.04	41.35	22.60	38.93	27.47	16.55	16.82	31.90	23.70
(d) DA (Ours)	46.43	24.80	61.59	30.55	70.42	38.13	65.88	33.14	51.79	21.08	59.22	29.54
Full MMWHS-MRI dataset ONLY → MMWHS-CT dataset												
(a) ICF [10]	25.03	30.29	47.54	20.01	56.74	24.12	52.59	21.85	48.46	18.22	46.07	22.90
(b) RBA [6]	32.44	32.61	51.43	23.30	52.76	23.18	51.26	26.25	40.43	23.06	45.67	25.68
(c) ICMD [24]	46.52	20.84	41.53	11.84	67.41	18.54	66.62	18.21	16.25	4.91	47.67	14.87
(d) DA (Ours)	54.24	33.72	72.22	36.97	61.90	21.58	69.14	32.06	61.49	27.78	63.80	30.42

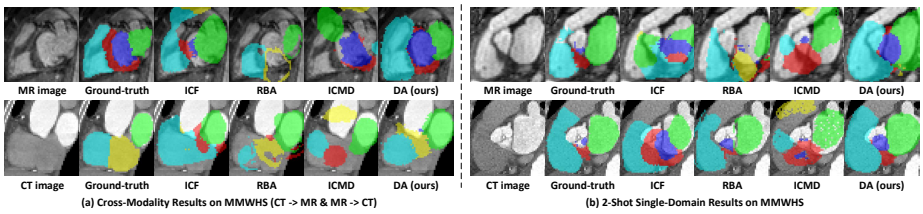

Fig. 2: **Visualization of cross-domain and few-shot (2-shot) results.**

Few-shot Training. As illustrated in Tab. 2, atlas-based methods ((b) and (d)) demonstrate greater robustness in small-data learning than image-mask mapping methods ((a) and (c)). DiffAtlas (d) achieves the best overall performance, followed by (b), leveraging the atlas-based registration paradigm. By parameterizing the atlas and modeling both image and mask, it utilizes diffusion's ability to enhance learning with limited samples.

Zero-shot Cross-domain Evaluation. Tab. 2 presents results across three settings with varying pretraining sample sizes, domains, and sources. Similar to the small-sample experiments, atlas-based methods ((b) and (d)) exhibit superior performance and robustness. DiffAtlas consistently achieves the best results, largely due to its guided inference process, which enhances target image matching and mitigates semantic gaps from domain shifts.

3.3 Visual Analysis

Fig. 2 displays the segmentation results with Myo (red), LV (green), LA (dark blue), RA (yellow), and RV (light blue). In cross-domain scenarios, the first row of Fig.2(a) demonstrates zero-shot segmentation from CT pretraining to MRI, while the second row shows MRI to CT. The results indicate that atlas-based methods (RBA and DA (ours)) maintain basic anatomical integrity across

domains, avoiding significant misclassifications, such as the mis-segmentation of Myo (red) and RA (yellow) shown in ICMD. In 2-shot scenario, our method ensures the most accurate segmentation by preserving contour integrity.

4 Conclusion

In conclusion, we present DiffAtlas, which jointly models image-mask pairs during training and guides the model to generate target image-mask pairs, ultimately producing the corresponding masks during testing. This approach alleviates the limitations of the atlas paradigm, such as scalability challenges and struggles with fine-grained structures and domain-specific atlases, while preserving its advantages of anatomical consistency and robustness to small misalignments. We thoroughly explore its effectiveness across various settings. In the future, we plan to further enhance the models by integrating implicit shape priors [26,25] and latent space diffusion [19].

Acknowledgment. This work was supported in part by the Swiss National Science Foundation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
2. Bai, W., Shi, W., de Marvao, A., Dawes, T.J., O'Regan, D.P., Cook, S.A., Rueckert, D.: A bi-ventricular cardiac atlas built from 1000+ high resolution mr images of healthy subjects and an analysis of shape and motion. Medical image analysis **26**(1), 133–145 (2015)
3. Bogensperger, L., Narnhofer, D., Falk, A., Schindler, K., Pock, T.: Flowsdf: Flow matching for medical image segmentation using distance transforms. arXiv preprint arXiv:2405.18087 (2024)
4. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis **97**, 103280 (2024)
5. DeepMind: Surface-distance (2018), https://github.com/google-deepmind/surface-distance
6. Ding, W., Li, L., Zhuang, X., Huang, L.: Cross-modality multi-atlas segmentation via deep registration and label fusion. IEEE Journal of Biomedical and Health Informatics **26**(7), 3104–3115 (2022)

7. Feng, R., Zheng, X., Gao, T., Chen, J., Wang, W., Chen, D.Z., Wu, J.: Interactive few-shot learning: Limited supervision, better medical image segmentation. IEEE Transactions on Medical Imaging **40**(10), 2575–2588 (2021)
8. Feuerstein, M., Glocker, B., Kitasaka, T., Nakamura, Y., Iwano, S., Mori, K.: Mediastinal atlas creation from 3-d chest computed tomography images: application to automated detection and station mapping of lymph nodes. Medical image analysis **16**(1), 63–74 (2012)
9. Gibbons, E., Hoffmann, M., Westhuyzen, J., Hodgson, A., Chick, B., Last, A.: Clinical evaluation of deep learning and atlas-based auto-segmentation for critical organs at risk in radiation therapy. Journal of Medical Radiation Sciences **70**, 15–25 (2023)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
11. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Medical image analysis **88**, 102846 (2023)
12. Kushnure, D.T., Talbar, S.N.: Ms-unet: A multi-scale unet with feature recalibration approach for automatic liver and tumor segmentation in ct images. Computerized Medical Imaging and Graphics **89**, 101885 (2021)
13. Liu, J., Ma, Z., Wang, Z., Liu, Y., Wang, Z., Sun, P., Song, L., Hu, B., Boukerche, A., Leung, V.: A survey on diffusion models for anomaly detection. arXiv preprint arXiv:2501.11430 (2025)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
15. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al.: Metrics reloaded: recommendations for image analysis validation. Nature methods **21**(2), 195–212 (2024)
16. Nie, D., Shen, D.: Adversarial confidence learning for medical image segmentation and synthesis. International journal of computer vision **128**(10), 2494–2513 (2020)
17. Rahman, A., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M.: Ambiguous medical image segmentation using diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11536–11546 (2023)
18. Ranem, A., González, C., dos Santos, D.P., Bucher, A.M., Othman, A.E., Mukhopadhyay, A.: Continual atlas-based segmentation of prostate mri. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 7563–7572 (2024)
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
21. Sauvalle, B., Salzmann, M.: Hybrid diffusion models: combining supervised and generative pretraining for label-efficient fine-tuning of segmentation models. arXiv preprint arXiv:2408.03433 (2024)
22. Vakalopoulou, M., Chassagnon, G., Bus, N., Marini, R., Zacharaki, E.I., Revel, M.P., Paragios, N.: Atlasnet: Multi-atlas non-linear deep networks for med-

ical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11. pp. 658–666. Springer (2018)
23. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2441–2449 (2022)
24. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 6030–6038 (2024)
25. Yang, J., Sedykh, E., Adhinarta, J.K., Le, H., Fua, P.: Generating anatomically accurate heart structures via neural implicit fields. In: Conference on Medical Image Computing and Computer Assisted Intervention. pp. 264–274. Springer (2024)
26. Yang, J., Wickramasinghe, U., Ni, B., Fua, P.: Implicitatlas: learning deformable shape templates in medical imaging. In: Conference on Computer Vision and Pattern Recognition. pp. 15861–15871 (2022)
27. Zhang, H., Yang, J., Wan, S., Fua, P.: Lefusion: Synthesizing myocardial pathology on cardiac mri via lesion-focus diffusion models. arXiv e-prints pp. arXiv–2403 (2024)
28. Zhuang, X., Bai, W., Song, J., Zhan, S., Qian, X., Shi, W., Lian, Y., Rueckert, D.: Multiatlas whole heart segmentation of ct data using conditional entropy for atlas ranking and selection. Medical physics **42**(7), 3822–3833 (2015)
29. Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Medical image analysis **58**, 101537 (2019)