bioengineering

MDPI

Article

Verdiff-Net: A Conditional Diffusion Framework for Spinal Medical Image Segmentation

Zhiqing Zhang [1,2], Tianyong Liu [3], Guojia Fan [4], Yao Pu [5], Bin Li [1], Xingyu Chen [1], Qianjin Feng [6,*] and Shoujun Zhou [1,*]

[1] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; zq.zhang3@siat.ac.cn (Z.Z.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Institute of Unconventional Oil & Gas Research, Northeast Petroleum University, Daqing 163318, China
[4] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China
[5] Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China
[6] School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China
[*] Correspondence: fengqj99@smu.edu.cn (Q.F.); sj.zhou@siat.ac.cn (S.Z.)

Abstract: Spinal medical image segmentation is critical for diagnosing and treating spinal disorders. However, ambiguity in anatomical boundaries and interfering factors in medical images often cause segmentation errors. Current deep learning models cannot fully capture the intrinsic data properties, leading to unstable feature spaces. To tackle the above problems, we propose Verdiff-Net, a novel diffusion-based segmentation framework designed to improve segmentation accuracy and stability by learning the underlying data distribution. Verdiff-Net integrates a multi-scale fusion module (MSFM) for fine feature extraction and a noise semantic adapter (NSA) to refine segmentation masks. Validated across four multi-modality spinal datasets, Verdiff-Net achieves a high Dice coefficient of 93%, demonstrating its potential for clinical applications in precision spinal surgery.

Keywords: spinal segmentation; diffusion model; multi-modality

check for updates

Citation: Zhang, Z.; Liu, T.; Fan, G.; Pu, Y.; Li, B.; Chen, X.; Feng, Q.; Zhou, S. Verdiff-Net: A Conditional Diffusion Framework for Spinal Medical Image Segmentation. *Bioengineering* **2024**, *11*, 1031. https://doi.org/10.3390/bioengineering11101031

Academic Editor: Fabiano Bini

Received: 3 September 2024
Revised: 26 September 2024
Accepted: 10 October 2024
Published: 15 October 2024

1. Introduction

Spinal surgical conditions are prevalent and frequently result in high rates of disability, significantly diminishing patients' quality of life. They have gradually emerged as major health concerns, profoundly impacting individuals' daily lives [1–3]. An essential first step toward a thorough diagnosis and treatment of spinal illnesses is the automated extraction of the vertebral form from various vertebrae medical pictures. Specifically, the evaluation, diagnosis, surgery planning, and image-guided interventional processes of numerous vertebrae illnesses depend on the precise segmentation of vertebrae and intervertebral disks (IVDs) in vertebrae images. Extensive research efforts have been made by several researchers to enhance the segmentation performance of models for vertebrae images. Specifically, the methods in this field are mainly divided into two categories, such as single-modality and multi-modality. The single-modality approaches [4,5] are common and have drawbacks in clinical practice, and frequently necessitate an extended development cycle. A multi-modal approach [6] can leverage data from multiple imaging modalities at the same time, each of which offers distinct anatomical and pathologic details that contribute to a more thorough reveal of the spinal system and its diseases. Furthermore, by minimizing the dependence on a single imaging technology, the multi-modality approach lowers the possibility of misdiagnosis and enhances diagnostic accuracy and reliability through thorough analysis. However, investigations on multi-modality vertebrae image segmentation techniques in clinical settings are still very scarce. Meanwhile, researchers must overcome the following two primary obstacles in order to produce accurate and dependable multimodality vertebrae segmentation results because of the variety of imaging modalities used on spinal medical pictures as well as the distinctiveness of their anatomical structures:

(1) **Data heterogeneity.** The term "heterogeneity of data" describes the variety and diversity of the data, or the variations in the data across various dimensions. These variations could be caused by a variety of things, including the characteristics of the chiropractic data itself, the acquisition technique, the equipment used, the duration of the acquisition, and more. For instance, feature extraction varies throughout different types of imaging data (e.g., MRI, X-ray, CT, etc.), and variations in acquisition equipment can result in issues with data quality such as noise and distortion in vertebrae imaging. Figure 1a illustrates the blurred outlines caused by the physical properties of X-ray imaging. Second, interclass similarity is seen in spinal MR images [5]. This means that neighboring vertebrae (intervertebral disks) in the same subject (Figure 1b) exhibit a high degree of morphological resemblance, making it more challenging to distinguish between individual vertebrae. The surrounding tissues in Figure 1c have similar physics and tissue densities to the vertebrae in CT imaging, which leads to identical CT values that confuse the background features and result in erroneous detections. Analogously, variations in the duration of acquisition may cause data drift and variability. As a result, data heterogeneity must be taken into account and managed throughout data processing and analysis since it is one of the key elements influencing the outcomes of data analysis and mining.

(2) **Anatomical shape.** Vertebral pictures show features such as blurriness, uneven grayscale distribution, high noise levels, and low contrast because of the state of spinal medical imaging today. The vertebrae that make up the spinal structure also have a similar form but different types. Spinal illnesses like vertebral strain alter the anatomical form of the vertebral bodies, as shown in Figure 1d. Furthermore, the vertebrae are spatially displaced as a result of trauma, bad posture, muscular imbalance, and congenital deformities [7]. This causes aberrant modifications or misalignment of their locations in space, which severely distorts the morphology of the vertebral bodies. Particularly, in Figure 1e, the individual had fractures or breaks in the sacral vertebrae as a consequence of external pressures. This may lead to further deformation, which makes the connections between the lumbar and sacral vertebrae extremely tight. As such, defining the borders between these joint vertebrae based on pixel intensity is difficult. This frequently results in semantic segmentation of these linked vertebrae as a single object during vertebral segmentation, which causes misidentification as a single vertebra. Given the increased unpredictability in the contour forms and placements of the vertebral bodies, these features surely make vertebral segmentation tasks more challenging.

Most of the existing deep learning (DL)-based image segmentation methods perform prediction and discrimination by directly learning the probability of classifying image pixels, and these methods fall into two main categories: convolutional neural network (CNN)-based methods [8,9] and visual Transformer (ViT)-based methods [10–12]. These methods typically use the cross-entropy or Dice loss function to train a model that learns a mapping function from an input medical image to a segmentation mask. Although these methods have been effective in specific applications, they mainly focus on learning decision boundaries in the pixel feature space [13], ignoring the distributional properties of the underlying data [14] and failing to adequately capture the intrinsic class characteristics of the data. Additionally, the learned feature space of existing methods exhibits a rapid decline in performance in regions far from the decision boundary [15], posing challenges for handling fuzzy boundaries and fine objects. This instability limits the model's ability to recognize subtle variations in images, which is particularly significant in medical image segmentation where these variations can be crucial for diagnosis and treatment. Therefore, despite the progress made by CNN- and ViT-based methods in the field of image segmentation, there is still room for improvement in terms of deeply understanding data distribution and enhancing the stability of the feature space. Generative models [16–20] discover intricate probabilistic correlations between images and segmentation masks in the field of medical image segmentation. Instead of producing a fixed segmentation result, these algorithms generate a probabilistic distribution that directly models the underlying data distribution and predicts the segmentation mask conditional distribution. On the other hand, this method is more likely to capture image ambiguity [17,21] and produce segmentation results that are smoother and more reliable.

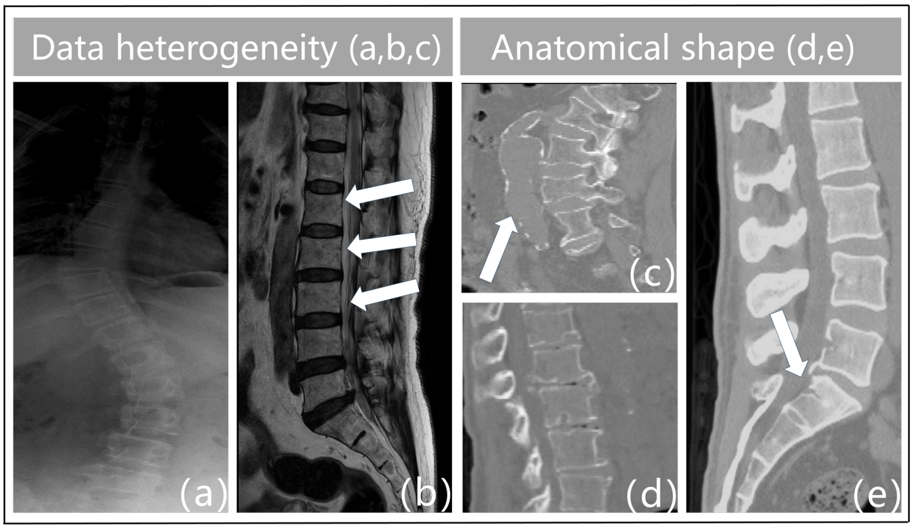

Figure 1. Challenges in resolving multi-modality vertebral images include: (**a**) boundary-blurring from X-ray imaging methods. (**b**) interclass similarity when vertebrae are present in MR. (**c**) The misdiagnosis in CT imaging of vertebrae and adjacent tissues with similar CT values. (**d**,**e**) weak vertebral contouring from the encroachment of vertebral degeneration/disease.

Diffusion probabilistic models are renowned as powerful generative models, possessing excellent capabilities for modeling (un)conditional data distributions. This brings significant advantages to the field of medical image segmentation. In practice, researchers have adopted two strategies: on the one hand, diffusion models are directly applied to treat image segmentation as a generative task, leveraging DDPM's ability to capture fine details in data distribution [20,22,23]. On the other hand, some studies aim to integrate the precision of discriminative methods with the creativity of generative methods [17,24,25], striving for superior performance in segmentation tasks. However, existing methods still have shortcomings in integrating the features of input data and segmentation masks. They often adopt a rather coarse concatenation approach when combining these two types of features for joint probabilistic modeling, leading to the loss of critical information. Moreover, although these methods are effective in general scenarios, they still require further customization and optimization in specific application contexts to meet unique demands and address corresponding challenges.

In this study, we built upon the foundation of diffusion models and cleverly integrated conditional models to provide more refined feature guidance. This approach achieves a deep combination of the advantages of existing discriminative segmentation models and generative diffusion probabilistic models. Our objective is to explore an efficient and high-quality method for precise segmentation of spinal multi-modal images. To this end, we propose a conditional diffusion segmentation framework named Verdiff-Net (as illustrated in Figure 2, which shows a conceptual diagram of our proposed segmentation framework). We observed that conventional diffusion segmentation models often adopt a simple concatenation approach when learning and modeling the joint probability of features from the original image and segmentation mask. To overcome this limitation, we introduced a multi-scale fusion module in the conditional U-Net to comprehensively capture the feature information of the original image. Additionally, by ingeniously applying the noise semantic adapter, we filtered out irrelevant information, allowing the model to focus more on the vertebral regions, thereby effectively enhancing the fineness of the segmentation masks and overall performance.

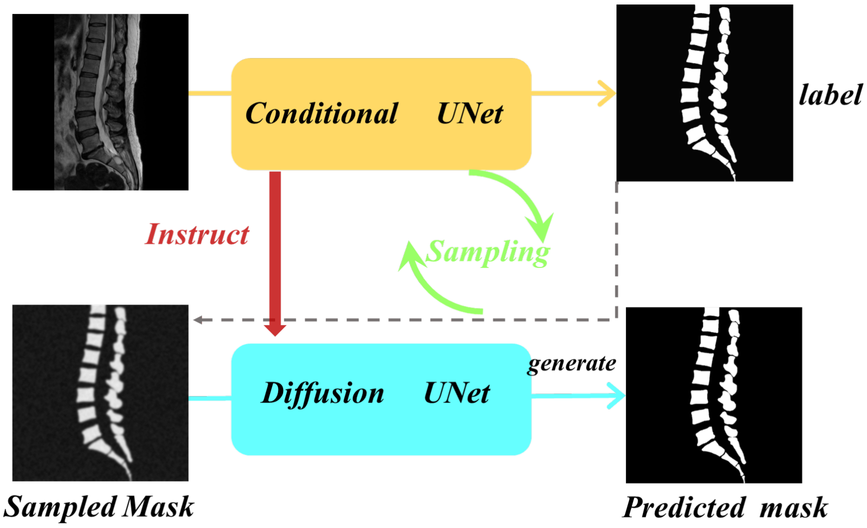

Figure 2. Modeling conceptual diagram. The diffusion model alternates between the noise addition phase and the sampling phase to train the network and ultimately generate the final prediction mask. The conditional U-Net model induces the diffusion model to synergize the advantages of the existing discriminative segmentation and generative diffusion models.

The innovations and contributions of this study can be summarized as follows:

(1) Synergistic combination of models: Verdiff-Net effectively blends the advantages of diffusion and conditional models, improving segmentation tasks' accuracy and stability, especially when modeling discrete targets.

(2) Multi-scale fusion module: Verdiff-Net integrates a multi-scale fusion module into the conditional U-Net to overcome the drawbacks of traditional diffusion models. By efficiently capturing and maintaining the underlying spinal feature information, this module lessens the loss that is commonly brought about by coarsely fusing image and mask features.

(3) Noise semantic adapter (NSA): During the diffusion process, the NSA serves as a selective filter, drawing the model's attention to important spinal aspects while rejecting unimportant data. This breakthrough increases inference efficiency as well as the model's accuracy in spinal segmentation.

(4) Comprehensive evaluation: The effectiveness and generalizability of the model are validated on four vertebrae medical datasets with three different modalities. To the best of our knowledge, this study is the first to thoroughly evaluate the model's robustness and generalization on a multi-modality vertebrae dataset.

2. Methodology

2.1. Framework Overview

Our proposed Verdiff framework consists of two key components: Conditional U-Net and diffusion U-Net. The core idea is to combine the discriminative power of existing segmentation models with the generative capabilities of diffusion models to enhance medical image segmentation. The conditional U-Net provides a prior segmentation mask for the diffusion model, while the diffusion U-Net iteratively refines the segmentation mask in an effective, efficient, and interactive manner.

In this section, we briefly introduce the proposed diffusion model framework. Unlike traditional medical image segmentation methods that directly input raw image data to predict the corresponding segmentation label map, the prior mask $f(b)$ generated by the conditional U-Net undergoes iterative refinement through the diffusion U-Net, ultimately producing the final segmentation mask. During the forward diffusion process, the prior mask is used for noise addition, and during the reverse diffusion process, it serves as the initial sampling point. The entire diffusion process can be represented as follows:

$$x_T \rightleftharpoons \cdots \rightleftharpoons x_t \underset{q(x_t|x_{t-1}, f(b))}{\overset{p_\theta(x_{t-1}|x_t, f(b)))}{\rightleftharpoons}} x_{t-1} \rightleftharpoons \cdots \rightleftharpoons x_0 \tag{1}$$

The diffusion model takes noisy segmentation labels $\mathbf{x}_t$ and the original image $\mathbf{b}$ as input and learns spinal features under the guidance of prior information from the original image. It predicts clear mask results through a reverse denoising process. Figure 3 illustrates the overall architecture of the proposed network, which consists of a forward diffusion stage and a reverse stage. Next, we provide a detailed explanation of the two key modules: the multi-scale fusion module (MSFM) and the noise semantic adapter (NSA).

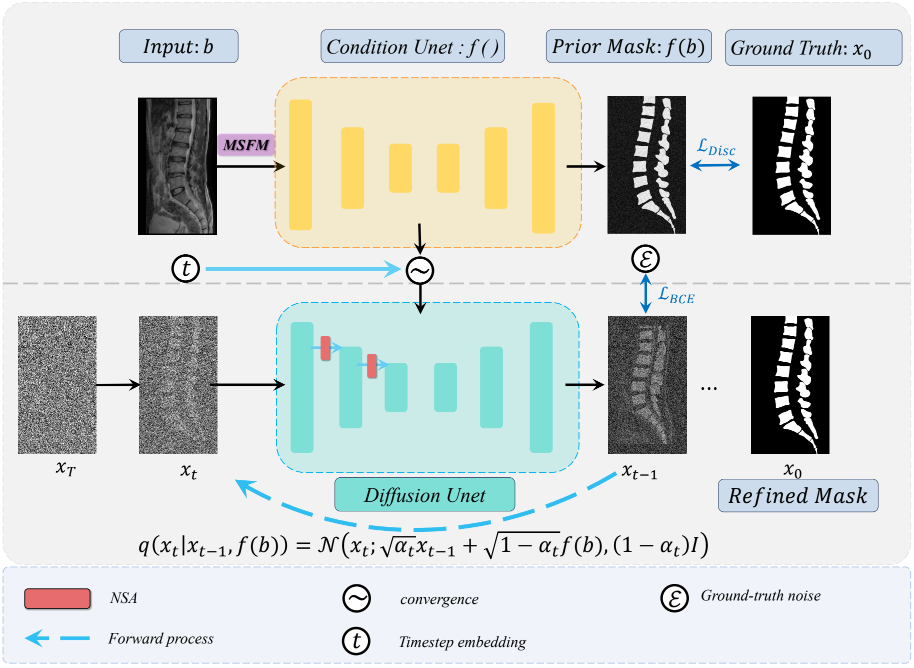

Figure 3. Overview of the Verdiff-Net framework. It contains two different U-Net backbone encoder–decoder structures.

2.2. Training Strategies for Diffusion Models

2.2.1. Diffusion Forward Process

During the training phase, variational inference is performed over a Markov process with T time steps. In the forward process at each time step $t \in T$, successive Gaussian noise is added to the label x_0 until the image becomes pure isotropic Gaussian noise x_T, thereby learning the data distribution. To better guide the diffusion process, we introduce an initial segmentation mask $f(x)$ provided by the conditional U-Net $f(b)$, which is used to correct the noise process. The forward noise process in the Markov chain at time step t in the forward diffusion process is represented as follows:

$$q(x_t \mid x_{t-1}, f(b)) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}f(b), (1-\alpha_t)\mathbf{I}), \tag{2}$$

where $\mathcal{N}(\cdot)$ denotes a normal distribution with a mean of $\alpha_t\mathbf{x}_{t-1} + (1-\alpha_t)f(\mathbf{x})$, and a variance of $(1-\alpha_t)\mathbf{I}$. Here, $f(\mathbf{x})$ represents the segmentation prior generated by the discriminative model, which is used to correct the noise distribution at each time step, allowing

Gaussian noise to better align with the segmentation prior. The parameter α_t controls the noise variance and determines the amount of noise added. $(x_0, x_1, \ldots, x_T)$ denotes the T steps in the Markov chain, and I is the identity matrix, representing isotropic noise.

After T steps, the data transform into a pure Gaussian noise image, represented by $\mathbf{x}_T$, and we can establish the relationship between $\mathbf{x}_0$ and $\mathbf{x}_t$ as follows:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}f(\mathbf{b}) + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{3}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative noise scheduling, which controls the degree of interpolation between $\mathbf{x}_0$ and the noise. The term $f(\mathbf{b})$ is the output of the conditional U-Net, which helps to guide the noise interpolation process. The term $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ represents standard Gaussian noise.

During the training process, after obtaining the noisy label $\mathbf{x}_t$ at time step t, the objective is to predict a clear label map $\mathbf{x}_0$ during the sampling phase. By combining the prior information from the conditional U-Net $f(\mathbf{b})$, the generative model can more efficiently recover the segmentation mask $\mathbf{x}_0$ from the noise, while further optimizing the synergy between the discriminative model and the generative model throughout the learning process.

2.2.2. Diffusion Reverse Process

Where $\bar{\alpha}$ is the cumulative product of α, used to simplify the relationship formula from the initial time step to the current time step. After obtaining the noisy label x_t at time step t, our goal is to predict the clean label map x_0 at the sampling stage. The holistic network Φ_θ uses the knowledge gained from the forward process to generate a sequence of incremental denoising operations in the reverse denoising process, which yields a clear image. The reverse distribution, as shown in Equation (4), reduces to a problem of minimizing the KL divergence between the forward and reverse distributions for all time steps.

$$p(x_{t-1} \mid x_t, f(b)) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, f(b)), \Sigma_\theta(x_t, t, f(b))) \tag{4}$$

where x_{t-1} follows a normal distribution (Gaussian distribution) with its mean and variance determined by the parameterized functions μ_θ and Σ_θ. μ_θ is the conditional mean function, and Σ_θ is the conditional variance (covariance) function. The optimization of the overall function requires sampling from the distribution $q(x_t \mid x_{t-1})$. Given x_0, the marginal distribution of x_t can be obtained from the intermediate latent variables.

$$q(x_t \mid x_0, f(b)) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}f(b), (1-\bar{\alpha}_t)\mathbf{I}) \tag{5}$$

By utilizing the posterior distribution $q(x_t \mid x_{t-1}, x_0)$, the KL divergence between the forward and reverse distributions is further simplified and reduced. Under the Markov assumption, the posterior distribution $q(x_{t-1} \mid x_t, x_0)$ is obtained through the iterative denoising process at each time step, yielding the estimate of x_{t-1}.

$$q(x_{t-1} \mid x_t, x_0, f(b)) = \mathcal{N}(x_{t-1}; \mu(x_t, x_0, f(b)), \sigma^2 I) \tag{6}$$

where

$$\mu(x_t, x_0, f(b)) = \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \sqrt{1-\bar{\alpha}_t}f(b)$$

and $\sigma^2 = \frac{(1-\bar{\alpha}_{t-1})(1-\alpha_t)}{1-\bar{\alpha}_t}$. Assuming that the covariance matrices of the two distributions $q(x_{t-1} \mid x_t, x_0)$ and $p(x_{t-1} \mid x_t)$ are the same, the corresponding expression of x_{t-1} as shown in Equation (7) is further obtained by the iterative inference process of the denoising model.

$$x_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\Phi_\theta(x_t, t, f(b))\right) + \sigma_t z \tag{7}$$

where $z \sim \mathcal{N}(0, I)$, $t = 1, \ldots, T$, σ is the inverse variance that can be learned, and z is the random sampling parameter. According to the above Equations (4)–(6), the simplified training objective is finally obtained by the following:

$$L_{\text{simple}} = \mathbb{E}_{x_0,\epsilon}\left[\|\epsilon - \Phi_\theta(\tilde{x}, t, f(b))\|_2^2\right] \tag{8}$$

where $\epsilon \sim \mathcal{N}(0, 1)$. We model the medical image segmentation task as a discrete data generation problem and directly predict the expected value of a label x_t with noise. Based on the variational upper bound of the negative log-likelihood of previous diffusion models, we use the Kullback–Leibler (KL) divergence and the binary cross entropy (BCE) mixed loss. Based on the combining Equations (6) and (10), we obtain the total hybrid loss [22] by the following:

$$L_{\text{BCE}} = -\mathbb{E}_{(\epsilon,\hat{\epsilon})} \sum_{i,j}^{H,W} \left[\epsilon_{i,j} \log \hat{\epsilon}_{i,j} + (1 - \epsilon_{i,j}) \log(1 - \hat{\epsilon}_{i,j})\right] \tag{9}$$

$$L_{\text{total}} = L_{\text{simple}} + L_{\text{BCE}} \tag{10}$$

By taking the average of different predicted segmentation masks generated by the network Φ_θ, a clear mask of significance can finally be obtained, providing a valuable reference for radiologists. The flow of the algorithm is shown in Algorithms 1 and 2.

Algorithm 1: Training

```
$(b, x_0) \sim q(b, x_0)$
$T \sim \text{Uniform}(\{1, 2, \ldots, T\})$
$\epsilon \sim \mathcal{N}(0, \mathbf{I})$
**Calculate Equation** (5)
**Take** gradient descent on $\nabla_\theta(L_{\text{total}})$
**until** converged
```

Algorithm 2: Sampling

```
$x_T \sim \mathcal{N}(0, \mathbf{I})$
**for** $t = T, \ldots, 1$ **do**
    $z \sim \mathcal{N}(0, 1)$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
    $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\Phi_\theta(x_t, t, f(b))\right) + \sigma_t z$
**end for**
**return** $x_0$
```

2.3. Multi-Scale Fusion Module (MSFM)

Conditional diffusion-based probabilistic models introduce noise into the original label and iteratively predict the segmented label map. However, the noise-added mask information at the current time step is often coarsely spliced and processed with the original image during fusion, which undoubtedly leads to feature loss and degrades the capability of subsequent network segmentation. To solve this problem, the proposed MSFM will play a role. As shown in Figure 4, before the input raw image b enters the conditional U-Net coding and decoding structure of the model, we first transform the raw data $b \in \mathbb{R}^{H \times W \times C}$ by three different convolutional kernel sizes of 3 × 3, 5 × 5, and 7 × 7, and fuse the results of the three branches by element summation to obtain $\tilde{b}$.

$$\tilde{b} = \tilde{A} + \tilde{B} + \tilde{C} \tag{11}$$

where $\tilde{b} \in \mathbb{R}^{H \times W \times C}$. After generating the channel information, the global information is embedded by using global average pooling to obtain $S \in \mathbb{R}^{C}$. The n-th element of S is obtained through contracting $\bar{b}$ by using the spatial dimension $H \times W$.

$$s_n = F_{gp}(\bar{b}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (\bar{b})_n(i,j) \tag{12}$$

Specifically, we use the ELU activation function in the fully connected layer along with the batch normalization technique for fine-tuning. To filter out useless dimensions to improve efficiency, we use the fully connected layer to create a compressed feature $z \in \mathbb{R}^{d}$ to achieve an accurate and adaptive selection guide.

$$z = Fc(\bar{b}) \tag{13}$$

where d is a hyperparameter that can be fine-tuned, it is used to control the dimensionality of the compressed feature z.

Then, cross-channel soft attention is used to adaptively select information at different spatial scales. Where a, b, c denote the soft attention vectors of $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, respectively, and a_n is the n-th element of a and $a_n + b_n + c_n = 1$, $A, B, C \in \mathbb{R}^{C \times d}$. Finally, the final feature map X is obtained from the attention weights on each convolutional kernel:

$$X_n = a_n \cdot \tilde{A}_n + b_n \cdot \tilde{B}_n + c_n \cdot \tilde{C}_n, \tag{14}$$

where $\mathbf{X} = [X_1, X_2, \ldots, X_n]$, $X_n \in \mathbb{R}^{H \times W}$,

$$\begin{aligned} a_n &= \frac{e^{A_n z}}{e^{A_n z} + e^{B_n z} + e^{C_n z}}, \\ b_n &= \frac{e^{B_n z}}{e^{A_n z} + e^{B_n z} + e^{C_n z}}, \\ c_n &= \frac{e^{C_n z}}{e^{A_n z} + e^{B_n z} + e^{C_n z}}. \end{aligned}$$

The output of X from this fusion module is then refined to extract the segmentation features of the original image, being fused with the information from the current step's noise-added mask x_t, and finally input to the U-Net coding structure of the diffusion model.

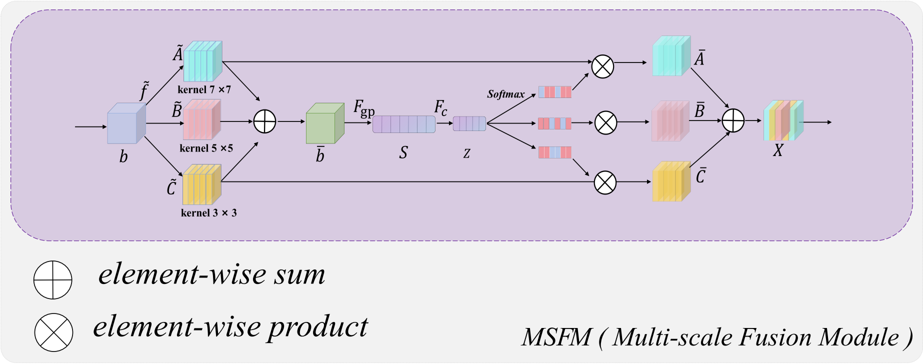

Figure 4. Schematic structure of the multi-scale fusion module (MSFM).

2.4. Noise Semantic Adapter (NSA)

Spinal pictures are used in anatomy to illustrate very intricate systems such as the spinal cord, IVD, and vertebrae. Not only are these structures connected, but they are also all extremely intricate. A continuous succession of structurally similar but individually unique vertebrae make up the vertebrae. The vertebrae are distinguished by their symmetry and continuity, which may not be as noticeable in other kinds of medical pictures. Vertebrae images from current medical imaging methods frequently have high aspect ratios, which makes processing image size and feature extraction difficult for popular deep learning networks. To further complicate the picture segmentation process, the key target regions of the vertebrae are typically more concentrated than in other medical images, and there is a certain semantic coherence between these regions. Relatively speaking, if researchers choose to compress the original vertebrae image into a square, or resize the original image to make it easier to input into the model, these operations will undoubtedly cause great damage to the spatial information features of the original image and lead to semantic loss. For example, in a single step in the diffusion process, t, the U-Net decoder of the diffusion model receives the fusion information from the noise mask, x_t, and the original image, and finally predicts the resultant clear mask. Along with input of different types of medical images, the lack of the model's ability (namely, to extract semantic information from a particular image) can lead to large discrepancies in the model's predictions.

To overcome this negative effect, we adopted the backbone structure of MedSegDiff [20]. The NSA was introduced into the standard U-Net of the conditional model to extract the spinal semantic information from the fused features. Specifically, we introduce the NSA into the U-Net encoder of the diffusion model. As shown in Figure 5, compared to the standard 2D convolution embedded in the standard U-Net, the NSA provides an operation to automatically select the convolution kernel. The input U-Net feature image information $X \in \mathbb{R}^{H \times W \times C}$ is first operated by the initial sliding window (kernel operation of the fine-grain convolution layer), and then by the spatial convolution layer of the 7 × 7 convolution kernel. The corresponding matrix obtained by these two operations is F and S, where $F, S \in \mathbb{R}^{H \times W \times C/2}$. The first fine-grained convolution layer can be selected with a 3 × 3 or 5 × 5 convolution layer, keeping the number of channels constant. Here, note that the specific selection strategy (for the convolution kernel geometry size) depends on the number of mask pixels around the center pixel. For fine particle extraction, 3 × 3 volume nodes are used when processing pixel data within the vertebrae, including vertebrae boundaries. For extracting large area background feature information, a 5 × 5 volume combination with a padding of 2 is used for coarse extraction, which improves the model's degree of matching to the target region features. In the next spatial convolution layer, a larger 7 × 7 convolution kernel with a dilation of 3 is employed. This approach increases the receptive field without adding more parameters, and each input channel is convolved with its own set of filters.

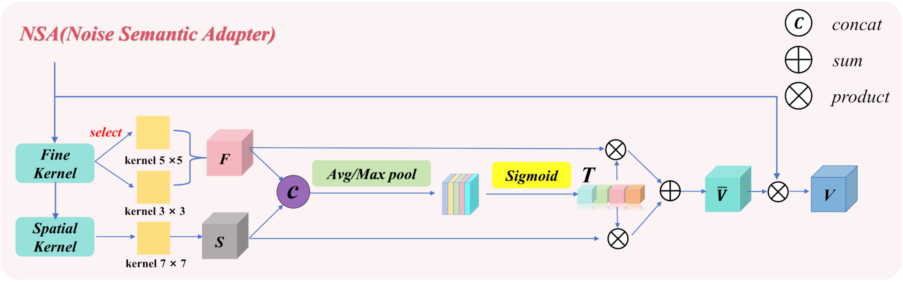

Figure 5. Noise semantic adapter (NSA) structure.

The two convolution layers are used to halve the dimension of the channel, and the obtained results are concatenated by aggregating the average and maximum attention. The sigmoid function is then applied to smoothly activate the weighted attention, attempting to expand the channel dimension back to its original size to retain the most relevant information and obtain the detailed features T of each channel.

$$T = \text{Sigmoid} \cdot f(F * S) \tag{15}$$

where $*$ represents the product, $T \in \mathbb{R}^{H_1 \times W_1 \times C}$.

The two convolution layers are used to halve the dimension of the channel, and the obtained results are concatenated by aggregating the average and maximum attention, and the sigmoid function is applied to smoothly activate the weighted attention, trying to expand the channel dimension back to its original size to retain the most relevant information and obtain the detailed features T of each channel.

After multiplying and superimposing the matrices F and S obtained from two scale convolution layers, $\overline{V} \in \mathbb{R}^{H \times W \times C}$ is then obtained. The final output feature $V \in \mathbb{R}^{H \times W \times C}$ results from multiplying with the residual connection of the original input information X. The V can be expressed as follows:

$$V = \overline{V} * X, \overline{V} = F * T + S * T \tag{16}$$

3. Experiments and Results

3.1. Dataset Introduction

To validate the generalizability and robustness of the proposed model, we comprehensively evaluated our proposed model on four vertebrae imaging dataset, including two sets of CT datasets, respectively, from publicly and privately available ones, a single set of MR ones, and a single set of X-ray ones.

Public CT dataset. We used the 42 sets of 3D spine1K data in the "verse" folder of the large-scale spine CT dataset called CTSpine1K. Based on this dataset, we conducted vertebrae segmentation experiments to select slices and contacted radiologists to manually select images one by one, and finally selected 2508 images with clear layers and anatomical information with masks for training and sampling. All the image sizes are standardized to 166 × 369.

Private CT dataset. Sagittal CT images of the vertebrae were collected from 630 volunteers (396 females, 234 males; mean age 26 ± 3 years, range 19 to 36 years). Three radiologists and one vertebral surgeon labeled these vertebral image regions as the ground truth for the vertebrae segmentation task and checked the labeled regions against each other to ensure reliability. Thus, each subject had a T2-weighted MR image and a corresponding mask as the initial ground truth, where each vertebra was assigned a unique label. We chose images of clean lumbar vertebral regions containing the caudal vertebrae and standardized the dimensions to 534 × 768 as inputs.

Private MR dataset. The dataset, collected from a local hospital, consists of T2-weighted MR volumetric images from 215 subjects. Among the 215 subjects, there were 6 normal subjects, 177 patients with vertebrae degeneration (VD), 204 patients with intervertebral disk degeneration (IDD), 21 patients with lumbar spondylolisthesis (LS), 91 patients with spinal canal stenosis (SCS), 22 patients with Schmorl's node (SN), and 53 patients with vertebral endplate osteochondritis (VEO). The delineated mask was corrected by the senior expert using the ITK-SNAP1 to be the ground truth of vertebrae parsing [26]. The average pixel spacing within the plane of the MR image was 0.35 mm, while the average slice thickness was 4.42 mm. By focusing solely on segmenting the vertebrae and excluding the spinal fluid portion between them, we first removed the spinal fluid from the labels of these data. We then smoothly extracted the 3D raw data and their corresponding mask data from the sagittal plane, ultimately obtaining pure 880 × 880 paired images and mask.

Public X-ray dataset. The dataset consists of 609 spinal anterior–posterior x-ray images [27]. The landmarks were provided by two professional doctors at the London Health Sciences Center. Each vertebra was located by four landmarks with respect to four corners. All the image sizes were 250 × 750. In addition, since the noise artifacts in this dataset were extremely severe, and some of the data were mixed with medical clinical instruments on the vertebrae, in order to further test the segmentation ability of the model in this particular condition, we ignored the detailed features of the vertebrae contour lines and performed a binary transform only for the vertebral trunk regions in order to obtain the original annotations.

3.2. Data Preprocessing

We used dynamic data augmentation during training to avoid network overfitting and to increase the robustness of our model. Three types of data augmentation are included, as follows: (1) randomly rotating the image from -30° to 30° to simulate the rotation variance; (2) randomly shifting the image by 1–5% to simulate the shift variance; (3) random elastic deformation and random contrast adjustment to improve the generalization of the model. To ensure all the methods used the same augmented training data, we used the same random state (seed = 35) for all methods when conducting data augmentation.

3.3. Training/Validation Setup and Evaluation Metrics

All the networks were built using PyTorch on a Linux 20.04 system, and the code ran on a server equipped with an A6000 GPU unless otherwise specified. The training process had a maximum iteration limit of 9000 steps. In our implementation, we trained our networks from scratch without relying on any pre-trained models. We employed the Adam optimization algorithm to minimize the loss function as described in Equation (9). Given our computational resources, a batch size of 1 was chosen. Batch normalization was utilized to facilitate a higher learning rate, resulting in relatively shorter training times. The learning rate was set to 10^{-4}. The stopping criterion was determined by the point where the validation loss stopped decreasing.

The Dice similarity coefficient (DSC) and intersection over union (IoU) serve as quantitative metrics for assessing segmentation performance. The DSC quantifies the overlap between the ground truth segmentation (Y) and the predicted segmentation results (X). The IoU describes the ratio of the overlapping area to the union area between the predicted and annotated segmentation regions. Specifically, the IoU is calculated as the area of intersection divided by the area of union between the predicted and annotated regions.

$$\text{Dice} = \frac{2|X \cap Y|}{|X| + |Y|} \tag{17}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{18}$$

Here, TP is the number of true positive pixels, which are correctly identified as part of the segment of interest. FP is the number of false positive pixels, representing the count of pixels incorrectly classified as part of the segment. FN is the number of false negative pixels, which are the pixels that are part of the segment in the ground truth but were missed by the prediction.

3.4. Vertebrae Segmentation Results

To quantitatively evaluate Verdiff-Net, we used Dice and IoU to measure the difference between the predicted results and the ground truth. During the evaluation, the ground truth labels for the vertebrae were obtained by manually drawing lines along the vertebral contours and were confirmed by certified radiologists. Verdiff-Net achieved average Dice scores of 94.37 ± 4.75%, 93.84 ± 1.56%, 93.86 ± 1.98%, and 88.74 ± 21.6% for vertebral structure segmentation on four different spinal multi-modality datasets. These quantitative results indicate that Verdiff-Net effectively balances over-segmentation and under-segmentation. As shown in Figure 6, we visualized the segmentation results of a single sagittal slice of the lumbar vertebrae. The high similarity between the model-predicted masks and the ground truth masks demonstrates that the model successfully achieves precise 2D spinal image segmentation across four multi-modality datasets.

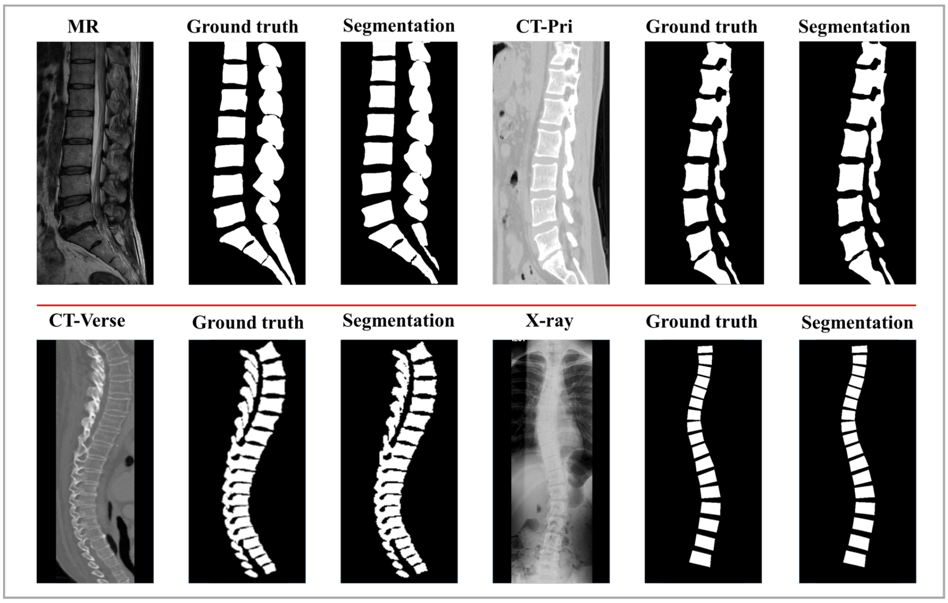

Figure 6. Verdiff-Net achieves reliable spinal segmentation performance. Above are the midsagittal sections of the vertebrae of each of the four subjects. The four subjects were imaged using different vertebrae imaging devices.

3.5. Comparison of the Results

In a comparative evaluation across four multi-modality datasets, we compared our proposed segmentation method with state-of-the-art deep learning approaches, as detailed in Table 1. We evaluate our approach against a number of cutting-edge segmentation approaches, such as U-Net [28], FCN [29], DeepLabv3 [30], nnU-Net [31], and Swin-U-Net [12]. For a thorough comparison, we additionally take into account the most recent state-of-the-art methods, SAM [32], and MedSegDiff [20].

Table 1. Metrics unique to each experiment for the four datasets.

	CT-Verse		CT-Pri		X-ray		MR	
	Dice (%)	**IoU (%)**	**Dice (%)**	**IoU (%)**	**Dice (%)**	**IoU (%)**	**Dice (%)**	**IoU (%)**
U-Net [28]	83.25 ± 4.08	71.35 ± 3.23	93.22 ± 1.83	87.34 ± 3.19	60.78 ± 10.23	43.75 ± 9.82	91.23 ± 2.76	84.00 ± 4.59
FCN [29]	82.22 ± 4.34	69.88 ± 3.61	93.24 ± 1.79	87.36 ± 3.07	65.30 ± 9.53	52.60 ± 8.17	91.50 ± 2.75	84.44 ± 4.57
DeepLabv3[30]	80.41 ± 4.50	67.29 ± 3.27	93.06 ± 1.88	86.94 ± 4.05	72.44 ± 8.59	63.91 ± 6.72	91.06 ± 2.93	83.70 ± 5.15
nnU-Net [31]	92.98 ± 3.86	86.97 ± 2.93	93.33 ± 1.69	88.04 ± 2.97	**89.77 ± 6.69**	**81.69 ± 5.26**	92.16 ± 2.73	84.97 ± 4.53
Swin-U-Net [12]	81.65 ± 4.19	69.27 ± 3.44	91.27 ± 2.15	84.02 ± 5.64	68.93 ± 9.06	53.13 ± 8.08	85.86 ± 7.94	75.39 ± 16.48
SAM [32]	79.78 ± 4.23	66.89 ± 3.50	80.36 ± 3.44	76.35 ± 8.21	67.89 ± 9.22	52.56 ± 13.32	81.00 ± 9.27	70.92 ± 17.57
MedSegDiff [20]	92.44 ± 3.75	88.27 ± 2.91	91.07 ± 1.69	84.51 ± 2.86	87.59 ± 7.24	76.68 ± 5.78	92.96 ± 2.87	86.53 ± 3.71
Verdiff-Net(ours)	**94.37 ± 3.73**	**90.89 ± 2.87**	**93.84 ± 1.56**	**91.87 ± 2.79**	88.74 ± 7.04	77.04 ± 5.75	**93.86 ± 1.98**	**88.58 ± 3.83**

All experiments were conducted on the same datasets, and statistical significance was noted in the results. The results show that Verdiff-Net performs better than its rivals, attaining the highest Dice similarity in the datasets CT-Pri, CT-Verse, and MR. More specifically, on the MR public dataset, Verdiff-Net demonstrated a considerable Dice improvement of roughly 1.7% over nnU-Net, and improvements of 0.51% and 1.4% on CT-Pri and CT-Verse, respectively. Verdiff-Net demonstrated its supremacy by routinely delivering Dice scores above 93% on CT-Pri, CT-Verse, and MR. Verdiff-Net reduces false positives and improves Dice and IoU scores by incorporating the suggested module and taking advantage of diffusion models' pixel-level semantic segmentation capabilities. Verdiff-Net generates smoother and more accurate vertebral mask edges with a considerable reduction in pixel shifts and misclassifications as compared to other DDPM-based techniques such as MedSegDiff [20]. Figure 7 makes clear that alternative deep learning techniques are not able to reliably produce accurate spinal segmentation in all datasets, especially when it comes to X-ray data, where Dice coefficients for U-Net, FCN, Swin-U-Net, and SAM were less than 70%. In regions above the implant, FCN and DeepLabv3 are unable to segment the vertebrae; in contrast, nnU-Net, MedSegDiff, and our suggested model can recognize and find the vertebrae.

In conclusion, Verdiff-Net outperforms other techniques in terms of Dice similarity and robustness to alterations in input modalities, demonstrating its strength in delivering accurate spinal segmentation, even in difficult settings.

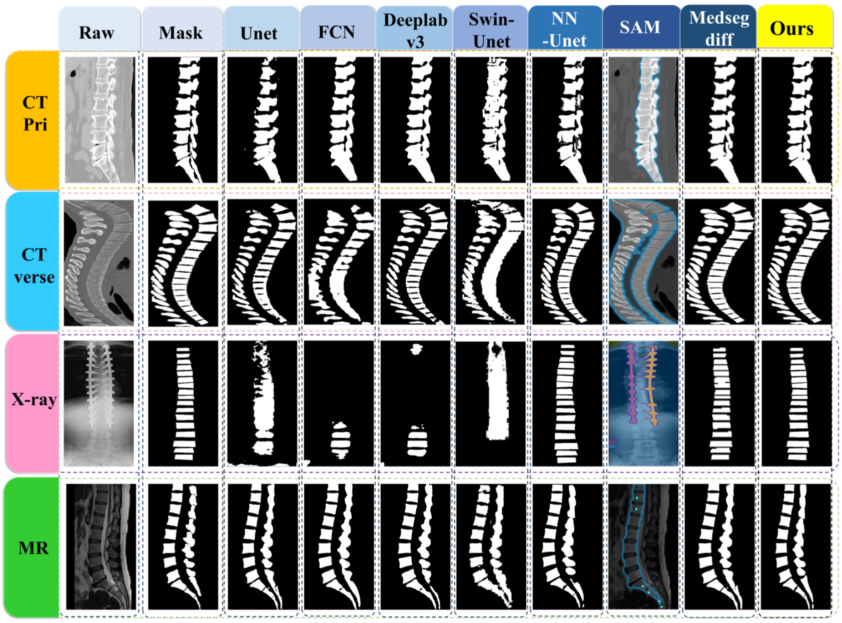

Figure 7. The comparison of Verdiff-Net with several SOTA segmentation methods.

3.6. Ablation Study

We conducted a thorough ablation experiment on the MR dataset to confirm the efficacy of our proposed module. To assess how well the three activities were performed, we used dice scores (%). The segmentation results of the underlying denoising diffusion probabilistic model (DDPM) on the MR dataset are greatly improved by both the Noise semantic adapter (NSA) alone and the multi-scale fusion module (MSFM) module, as shown in Table 2. In particular, the model's Dice coefficient is improved by 3.28% just by using the suggested multi-scale fusion module (MSFM) in place of the earlier technique.

Table 2. Ablation study results on the MR dataset.

MSFM	NSA	MR (Dice)	MR (Iou)
		0.8868	0.8396
✓		0.9196	0.8682
	✓	0.9027	0.8521
✓	✓	0.9386	0.8858

This shows that by helping to retrieve the original vertebrae image's finely segmented features, which can then be used to include semantic requirements in the diffusion U-Net model that follows, multi-scale fusion module (MSFM) can optimize the DDPM-based model. By learning an architecture based on the content-switching convolutional kernel of the original vertebrae images, the noise semantic adapter (NSA) improves the interaction between the noise mask and the anatomical semantic features of the images. This greatly improves the segmentation results of the underlying diffusion model, leading to an improvement of the Dice coefficients by 1.59%, respectively. In this MR dataset, our suggested two modules enhance the base denoising diffusion probabilistic model by over 5%, yielding cutting-edge segmentation outcomes. The ablation study's findings confirm the efficacy of the suggested modules by demonstrating that the semantic feature extractor is responsible for the notable increase in vertebrae parsing performance.

4. Discussion

4.1. Segmentation Effects

We analyzed the segmentation effects across four disparate datasets. The comparative trials revealed significant variations in the models' performance, as depicted in Figure 8. Specifically, on the CT-Verse and CT-Pri datasets, the models exhibited superior segmentation capabilities, accompanied by a notably smaller total standard deviation. In contrast, the X-ray dataset presented a challenge, with the segmentation metrics Dice and IoU recording lower values and the overall standard deviation being slightly elevated. To elucidate these discrepancies, we undertook an exhaustive review of the dataset composition. Less overall variability of the pictures in the dataset was caused by the subject's mid-sagittal vertebrae segment data in the CT dataset being smoother and having higher resolution. On the other hand, the X-ray dataset's poor segmentation can be attributed to two key factors. Firstly, there is a lot of noise and artifacts in the X-ray collection, and the vertebrae shape differs a lot from patient to patient, making segmentation more challenging. Second, with no discernible edge enhancement or description, the target vertebrae region is remarkably similar to the background pixels. Many of the most advanced segmentation methods struggle to handle this high degree of similarity, making it simple to confuse the vertebrae region with the backdrop. Furthermore, our observations indicate that the standard deviation of the SAM model's segmentation outcomes on the MR dataset is significantly higher when compared to other models, including Swin-U-Net. The spinous processes in the MR images exhibit substantially lower resolution and brightness compared to the vertebral trunk, often blending in with the background. Additionally, the manually annotated masks for these areas are less regular, contrasting with the more uniform imaging conditions typical of CT scans. In this context, the absence of pre-training hinders the SAM model's ability to effectively segment the vertebral feature region from a semantic standpoint, thereby reducing the overall segmentation accuracy of the dataset.

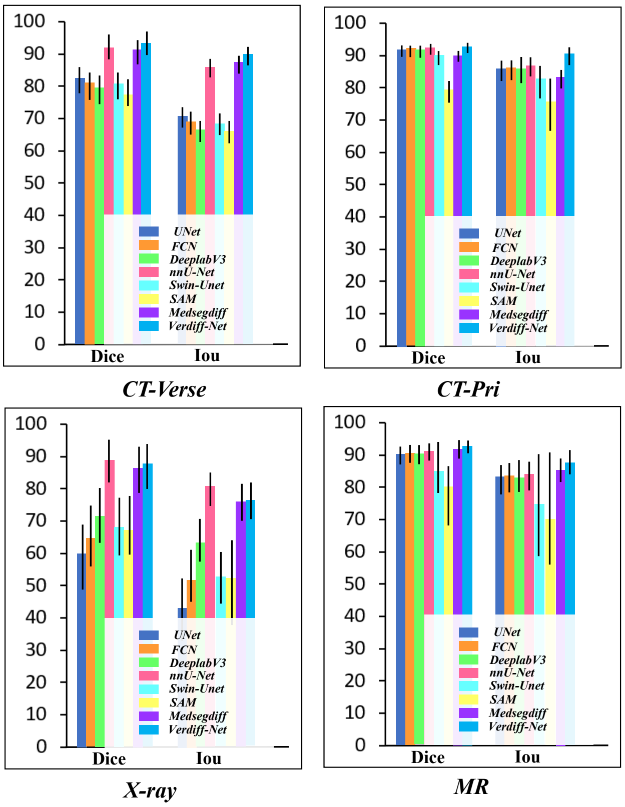

Figure 8. Vertebrae segmentation metrics (Dice vs. IoU) on the four datasets. In the majority of vertebrae datasets, Verdiff-Net achieved the highest average Dice (%) results.

Due to Swin-U-Net's relatively slow global sensitivity and convergence rate during actual training, its segmentation performance is further compromised by the system's limited capability to extract meaningful features from the vertebral region. The performance of these models on the MR dataset also highlights their limitations in dealing with low-resolution images of vertebrae that have a background highly similar to the foreground, underscoring the need for more sophisticated models that can better handle such challenging imaging conditions.

Furthermore, we examined the instances in which certain models failed the comparison test, as illustrated in Figure 7. Of the models tested on the X-ray dataset, only nnU-Net, MedSegDiff, and Verdiff-Net were able to accurately predict the specific regions that belong to the vertebrae in the X-ray images when bone nails were present, and they also maintained a high degree of similarity with the mask that the qualified physicians had labeled. Based on our analysis, MedSegDiff and Verdiff-Net's segmentation success can be attributed to distinct factors. Specifically, the DDPM algorithm, which emphasizes learning only the vertebrae region's specific features during the noise addition process, disregards the features of the other parts as noise. By classifying all non-vertebrae pixels on the vertebrae layer as noise during the reverse denoising stage, these models successfully segment the vertebrae area. On the other hand, nnU-Net's integrated data preparation module improves the vertebrae region's semantic information and sensory field, which helps the network better collect vertebrae properties.

4.2. Limitation

From an application perspective, we acknowledge that we only considered the segmentation of spinal images and did not evaluate its generalizability to other medical images. Additionally, due to the difficulty in obtaining and annotating data, there are differences in the spinal regions (such as thoracic and lumbar spine) among different datasets, and the annotation methods vary. The imaging techniques for different spinal regions also lead to variations in imaging directions, resulting in different morphological representations of the spine. This not only increases the difficulty of data annotation but also complicates the model's ability to recognize vertebral regions.

On the one hand, our initial intention was to explore the generalization performance of the diffusion model on multi-modal spinal images and the stability of segmentation performance across different modality datasets. This led us to overlook the model's multi-class segmentation performance, simplifying the spinal segmentation task to a binary segmentation task. On the other hand, the segmentation task for the generative model requires significant CUDA memory, thereby increasing computational demands. Segmenting multiple types of spinal regions ($\geq$9 categories) inevitably leads to high CUDA usage, potentially causing errors.

Despite the algorithm's superior segmentation performance, there are still shortcomings. First, the DPM-based algorithm takes a long time to train. Due to the necessity of performing addition and denoising operations over several iterations, the training process is time-consuming. Secondly, this approach heavily relies on GPU resources. The training phase requires handling a large number of computational operations, typically necessitating high-performance GPU support, which limits its applicability to some extent. It is recommended that future research focus on optimizing the algorithm, enhancing training efficiency, and reducing dependence on hardware resources to further promote the application of this technology in clinical settings.

5. Conclusions

This study proposes a novel medical image segmentation method called Verdiff-Net, which combines the advantages of discriminative segmentation models and generative diffusion probabilistic models to achieve accurate segmentation of multi-modal spinal images. The contributions of this study are summarized as follows:

Feature extraction mechanism: A feature extraction mechanism incorporating multi-scale convolutional layers is introduced into the conditional U-Net. This mechanism reduces the loss of low-level features, effectively models the underlying data distribution, and maximizes the model's ability to learn multi-scale spatial features from the original spinal images.

Noise semantic adapter (NSA): Given the characteristic large aspect ratio of spinal medical images, NSA is proposed. NSA filters the fused features input into the diffusion model and further adjusts the model to focus on attention feature responses in the spinal target area, thus accommodating the unique morphology and structural factors of the spine.

Validation and generalization capability: The effectiveness and generalization capability of the model were thoroughly evaluated on four medical datasets covering three different imaging modalities. The results demonstrate that Verdiff-Net exhibits outstanding performance in the domain of spinal medical image segmentation, effectively identifying and segmenting low-level spinal features while showing strong potential for application across different datasets. This study is the first to comprehensively evaluate the robustness and generalization capability of the model on multi-modal spinal datasets, providing new perspectives and references for research in this field, and promoting further exploration and development in medical image segmentation technology.

Author Contributions: Data curation, T.L. and Y.P.; formal analysis, G.F.; funding acquisition, Q.F.; investigation, Y.P.; methodology, Z.Z.; resources, X.C.; software, T.L.; supervision, S.Z.; writing—original draft, Z.Z.; writing—review and editing, Z.Z., B.L., Q.F. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key R&D Project of China (2018YFA0704102, 2018YFA0704104), in part by the Natural Science Foundation of Guangdong Province (No. 2023A15150 10673), in part by the Shenzhen Technology Innovation Commission (JSGG20220831110400001, CJGJZD20230724093303007), and in part by the Shenzhen Engineering Laboratory for Diagnosis & Treatment Key Technologies of Interventional Surgical Robots (XMHT20220104009).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: To access the MR dataset, please contact Dr. Pang (pangshumao@126.com) and sign a confidentiality agreement. Public CT dataset is available at https://github.com/MIRACLE-Center/CTSpine1K. The X-ray dataset is available at https://www.dropbox.com/scl/fi/80hduycgyrbse281sdtjf/scoliosis-xray-Single-View.zip?rlkey=opye7f5de0avh3isc1br5gwjh&e=2&dl=0, accessed on 1 Septermber 2024. The other data that support the findings of this study are available from the corresponding author, upon reasonable request.

Acknowledgments: The authors thank "The Key Laboratory of Biomedical Imaging Science and System, Chinese Academy of Sciences."

Conflicts of Interest: The authors declare that there are no conflicts of interest related to this article.

References

1. Smith, M.W.; Reed, J.; Facco, R.; Hlaing, T.; McGee, A.; Hicks, B.M.; Aaland, M. The reliability of nonreconstructed computerized tomographic scans of the abdomen and pelvis in detecting thoracolumbar spine injuries in blunt trauma patients with altered mental status. *JBJS* **2009**, *91*, 2342–2349. [CrossRef] [PubMed]
2. Yao, J.; Burns, J.E.; Munoz, H.; Summers, R.M. Detection of vertebral body fractures based on cortical shell unwrapping. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2012: 15th International Conference, Nice, France, 1–5 October 2012; Proceedings, Part III 15; Springer: Berlin/Heidelberg, Germany, 2012; pp. 509–516.
3. Huang, M.; Zhou, S.; Chen, X.; Lai, H.; Feng, Q. Semi-supervised hybrid spine network for segmentation of spine MR images. *Comput. Med. Imaging Graph.* **2023**, *107*, 102245. [CrossRef]
4. Han, Z.; Wei, B.; Mercado, A.; Leung, S.; Li, S. Spine-GAN: Semantic segmentation of multiple spinal structures. *Med. Image Anal.* **2018**, *50*, 23–35. [CrossRef] [PubMed]
5. Pang, S.; Pang, C.; Zhao, L.; Chen, Y.; Su, Z.; Zhou, Y.; Huang, M.; Yang, W.; Lu, H.; Feng, Q. SpineParseNet: Spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation. *IEEE Trans. Med. Imaging* **2020**, *40*, 262–273. [CrossRef]

6. Zhao, S.; Wang, J.; Wang, X.; Wang, Y.; Zheng, H.; Chen, B.; Zeng, A.; Wei, F.; Al-Kindi, S.; Li, S. Attractive deep morphology-aware active contour network for vertebral body contour extraction with extensions to heterogeneous and semi-supervised scenarios. *Med. Image Anal.* **2023**, *89*, 102906. [CrossRef] [PubMed]
7. Zhang, D.; Chen, B.; Li, S. Sequential conditional reinforcement learning for simultaneous vertebral body detection and segmentation with modeling the spine anatomy. *Med. Image Anal.* **2021**, *67*, 101861. [CrossRef] [PubMed]
8. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.
9. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [CrossRef] [PubMed]
10. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
11. Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H.R.; Xu, D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In Proceedings of the International MICCAI Brainlesion Workshop, Virtual Event, 27 September 2021; pp. 272–284.
12. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
13. Liu, W.; Wen, Y.; Yu, Z.; Yang, M. Large-margin softmax loss for convolutional neural networks. *arXiv* **2016**, arXiv:1612.02295.
14. Bernardo, J.; Bayarri, M.; Berger, J.; Dawid, A.; Heckerman, D.; Smith, A.; West, M. Generative or discriminative? Getting the best of both worlds. *Bayesian Stat.* **2007**, *8*, 3–24.
15. Ardizzone, L.; Mackowiak, R.; Rother, C.; Köthe, U. Training normalizing flows with the information bottleneck for competitive generative classification. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 7828–7840.
16. Liang, C.; Wang, W.; Miao, J.; Yang, Y. Gmmseg: Gaussian mixture based generative semantic segmentation models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 31360–31375.
17. Chen, T.; Wang, C.; Chen, Z.; Lei, Y.; Shan, H. HiDiff: Hybrid Diffusion Framework for Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2024**. [CrossRef] [PubMed]
18. 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, Vancouver, BC, Canada, 17–24 June 2023; pp. 11536–11546.
19. Xing, Z.; Wan, L.; Fu, H.; Yang, G.; Zhu, L. Diff-unet: A diffusion embedded network for volumetric segmentation. *arXiv* **2023**, arXiv:2303.10326.
20. Wu, J.; Fu, R.; Fang, H.; Zhang, Y.; Yang, Y.; Xiong, H.; Liu, H.; Xu, Y. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In Proceedings of the Medical Imaging with Deep Learning, Paris, France, 3–5 July 2024; pp. 1623–1639.
21. Ng, A.; Jordan, M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Adv. Neural Inf. Process. Syst.* **2001**, *14*.
22. Wolleb, J.; Sandkühler, R.; Bieder, F.; Valmaggia, P.; Cattin, P.C. Diffusion models for implicit image segmentation ensembles. In Proceedings of the International Conference on Medical Imaging with Deep Learning, Zurich, Switzerland, 6–8 July 2022; pp. 1336–1348.
23. Amit, T.; Shaharbany, T.; Nachmani, E.; Wolf, L. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv* **2021**, arXiv:2112.00390.
24. Guo, X.; Yang, Y.; Ye, C.; Lu, S.; Peng, B.; Huang, H.; Xiang, Y.; Ma, T. Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation. In Proceedings of the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia, 18–21 April 2023; pp. 1–5.
25. 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, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 6030–6038.
26. Yushkevich, P.A.; Pashchinskiy, A.; Oguz, I.; Mohan, S.; Schmitt, J.E.; Stein, J.M.; Zukić, D.; Vicory, J.; McCormick, M.; Yushkevich, N.; et al. User-guided segmentation of multi-modality medical imaging datasets with ITK-SNAP. *Neuroinformatics* **2019**, *17*, 83–102. [CrossRef] [PubMed]
27. Wu, H.; Bailey, C.; Rasoulinejad, P.; Li, S. Automatic landmark estimation for adolescent idiopathic scoliosis assessment using BoostNet. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, 11–13 September 2017; Proceedings, Part I 20; Springer: Berlin/Heidelberg, Germany, 2017; pp. 127–135.
28. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
29. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

30. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
31. 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. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef] [PubMed]
32. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4015–4026.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.