HSRDiff: A Hierarchical Self-Regulation Diffusion Model for Stochastic Semantic Segmentation

Han Yang[1,2], **Chuanguang Yang**[1*], **Zhulin An**[1*], **Libo Huang**[1], **Yongjun Xu**[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{yanghan22s, yangchuanguang, anzhulin, huanglibo, xyj}@ict.ac.cn

Abstract

In safety-critical domains such as medical diagnostics and autonomous driving, single-image evidence is sometimes insufficient to reflect the inherent ambiguity of vision problems. Therefore, multiple plausible assumptions that match the image semantics may be needed to reflect the actual distribution of targets and support downstream tasks. However, balancing and improving the diversity and consistency of segmentation predictions under the high-dimensional output spaces and potential multimodal distributions is still challenging. This paper presents Hierarchical Self-Regulation Diffusion (HSRDiff), a unified framework that simulates joint probability distribution over entire labels. Our model self-regulates the balance between the two modes of predicting the label and noise in a novel "differentiation to unification" pipeline and dynamically fits the optimal path to model the aleatoric uncertainty rooted in observations. In addition, we preserve the high-fidelity reconstruction of the delicate structure in images by leveraging the hierarchical multi-scale condition priors. We validate HSRDiff in three different semantic scenarios. Experimental results show that HSRDiff is superior to the comparison method with a considerable performance gap.

Code — https://github.com/yanghan-yh/HSRDiff.git

Introduction

Due to the powerful representation ability (Yang et al. 2024a, 2023a) of deep neural networks, image semantic segmentation has made significant progress (Chen et al. 2017, 2018; Yang et al. 2022b; Zhao et al. 2017; Wang et al. 2020; Yang et al. 2022a). At present, most of the methods are deterministic models (Isensee et al. 2021; Yang et al. 2023b; Ronneberger, Fischer, and Brox 2015), that is, only one segmentation prediction with the highest probability and the best match is generated for a given image, which is practical in many scenarios. However, for some safety-critical domains, a single segmentation prediction is insufficient to reflect the inherent ambiguity of the natural visual world. For example, in medical diagnosis, different radiologists often make different diagnoses due to the ambiguity of lesions and organ margins (Yang et al. 2022b). Also, in road driving, different observation conditions may cause foreign judgments of objects by human eyes. Therefore, the goal of stochastic semantic segmentation is to model the aleatoric uncertainty of the label and condition to a given image, thereby providing multiple reasonable and effective segmentation assumptions to improve the decision confidence of downstream tasks.

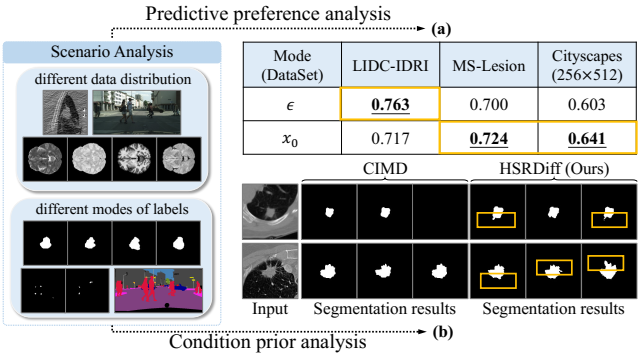


Figure 1: **The motivation of HSRDiff.** (a) mIoU of different datasets under different DDPM modes with global condition guidance; (b) Generated samples of CIMD (with the global condition) and HSRDiff (with the multi-scale condition).

Many VAE-based methods have been proposed to complete the stochastic semantic segmentation tasks (Kohl et al. 2018, 2019; Bian et al. 2020). Still, the diversity of their prediction results is limited because they rely on a strict condition of the axis-aligned Gaussian latent posterior distribution. In recent years, due to the powerful generative ability of Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020), some stochastic segmentation works based on DDPM have been proposed successively (Rahman et al. 2023; Zbinden et al. 2023; Wolleb et al. 2022). They predict the application noise or label map at each timestep and let the network model the underlying structure of the dataset in the reverse diffusion process with global guidance of condition information. However, there are still two problems: **(1) Global condition information is insufficient to guide the network to model the rich probability distribution of targets** (Kohl et al. 2019; Xu et al. 2024)**:** High-resolution features contain more fine-structure information, which is conducive to predicting small targets. Low-

*Corresponding Author.

resolution features include more contextual information and are more suitable for significant target prediction. As shown in Figure 1 (b), the global condition information ignores the multi-scale local variations in the target, which reduces the modeling ability of some fine details and makes the network unable to learn a richer probability distribution with more detailed information. **(2) Using only a single prediction mode of predicting noise or label map does not guarantee superior performance in various semantic scenarios, especially across domains:** As shown in Figure 1 (a), the data distributions of different semantic scenarios have considerable gaps, so they may have different preferences for being modeled by predicting noise with statistical properties and labels with multimodal details.

This paper considers the following two points: (1) How to introduce and effectively integrate hierarchical multi-scale structure information and (2) How to integrate the two prediction modes while maintaining their differentiation. Therefore, we propose a Hierarchical Self-Regulation Diffusion (HSRDiff) framework to solve the above problems. The proposed method adopts a novel "differentiation to unification" predicting pipeline. Firstly, in each denoising step, we adopt two modes of independently predicting the external noise and the label map simultaneously. Secondly, a trainable self-regulating parameter is used to unify the two modes dynamically to find the optimal path for reconstruction. This pipeline allows our proposed method to maintain superior performance across various semantic scenarios. Additionally, beyond the conventional global condition information, HSRDiff incorporates hierarchical multi-scale conditional priors to guide the decoder's reconstruction process. This enables HSRDiff to model richer probability distributions within the conditions, resulting in more refined segmentation. The fusion of hierarchical multi-scale conditional information and the dynamic self-regulation modeling process makes the distribution of predictions achieve maximum improvement and balance in diversity and consistency. The contributions of this paper are summarized as follows:

- We introduce a novel Self-Regulation Diffusion (SRDiff) method: SRDiff uses two independent modes to predict labels and noise separately, replacing the standard single-mode prediction. It then integrates information from both paths to self-regulate the application of each mode, choosing the optimal modeling path. This results in a "differentiation to unification" prediction pipeline.

- A novel hierarchical multi-scale framework Hierarchical Self-Regulation Diffusion (HSRDiff) for stochastic semantic segmentation is proposed based on SRDiff: HSRDiff uses hierarchical condition prior and SRDiff to effectively model label distribution and generate segmentation hypotheses that capture multi-scale uncertainty.

- We verify the proposed method on three semantic scenarios: We achieve state-of-the-art (SOTA) performance on three semantic scenarios of the LIDC-IDRI, MS-Lesion, and the multimodal Cityscapes dataset.

Related Works

Uncertainty in Deep Learning. The uncertainty in machine learning can be divided into epistemic uncertainty and aleatoric uncertainty (Hüllermeier and Waegeman 2021). Epistemic uncertainty refers to the uncertainty of model parameters, which measures whether the input exists in the data distribution that has already been seen. This uncertainty can be explained as long as there is enough data. The problem in this paper is aleatoric uncertainty, which refers to the inherent uncertainty in the observation. This uncertainty is not introduced by the model and cannot be resolved by increasing the amount of data. Well-calibrated quantification of aleatoric uncertainties is an essential step for further developing deep learning in safety-critical areas.

Stochastic Semantic Segmentation. In segmentation tasks, the most common paradigm is deterministic segmentation, which refers to learning a single label for each input image and predicting a single segmentation mask (Ronneberger, Fischer, and Brox 2015; Oktay et al. 2018; Isensee et al. 2021; Yang et al. 2022b, 2023b). However, in some safety-critical areas, it is necessary to quantify aleatoric uncertainties in the dataset and provide multiple segmentation assumptions. Initially, some works (Kendall and Gal 2017; Tanno et al. 2017) propose quantifying aleatoric uncertainty at the pixel level through Bayesian deep learning. However, these methods ignore the structural information and the joint distribution of labels. Later, Kohl *et al.* (Kohl et al. 2018) introduces a generative segmentation model Probabilistic U-Net (Prob. U-Net) based on the combination of U-Net and conditional variational autoencoder (c-VAE), which can effectively generate an infinite number of reasonable assumptions. Many works (Baumgartner et al. 2019; Badrinarayanan, Kendall, and Cipolla 2017; Kohl et al. 2019; Zhang et al. 2022) extend Prob. U-Net to a hierarchical version to improve the expressiveness of the model by fitting more complex distributions. There are also some concurrent works independent to c-VAE. Gao *et al.* (Gao et al. 2022) proposes a mixture of stochastic experts model, which simultaneously estimates different modes of aleatoric uncertainty through multiple expert networks, resulting in an effective two-level uncertainty representation.

Diffusion Model for Stochastic Segmentation. Diffusion models excel in various visual fields due to their ability to model complex distributions (Amit et al. 2021; Wu et al. 2022; Chung, Sim, and Ye 2022; Whang et al. 2022; Feng et al. 2024; Liu et al. 2023; Yang et al. 2024b). While many works (Monteiro et al. 2020; Chen, Zhang, and Hinton 2022; Rahman et al. 2023; Zbinden et al. 2023) use conditional diffusion models to simulate label distributions in stochastic segmentation, their reliance on global conditional information and single-mode denoising limits sample diversity and fine semantic expression. Benny *et al.* (Benny and Wolf 2022) addressed image quality issues during faster sampling by predicting noise and image simultaneously, which inspired us. However, their method has two key limitations: a) Shared parameters during decoding make it hard to distinguish between the two modes. b) The fusion parameter generation doesn't interact with the mode predictions, limiting the ability to achieve a good balance. Our work uses

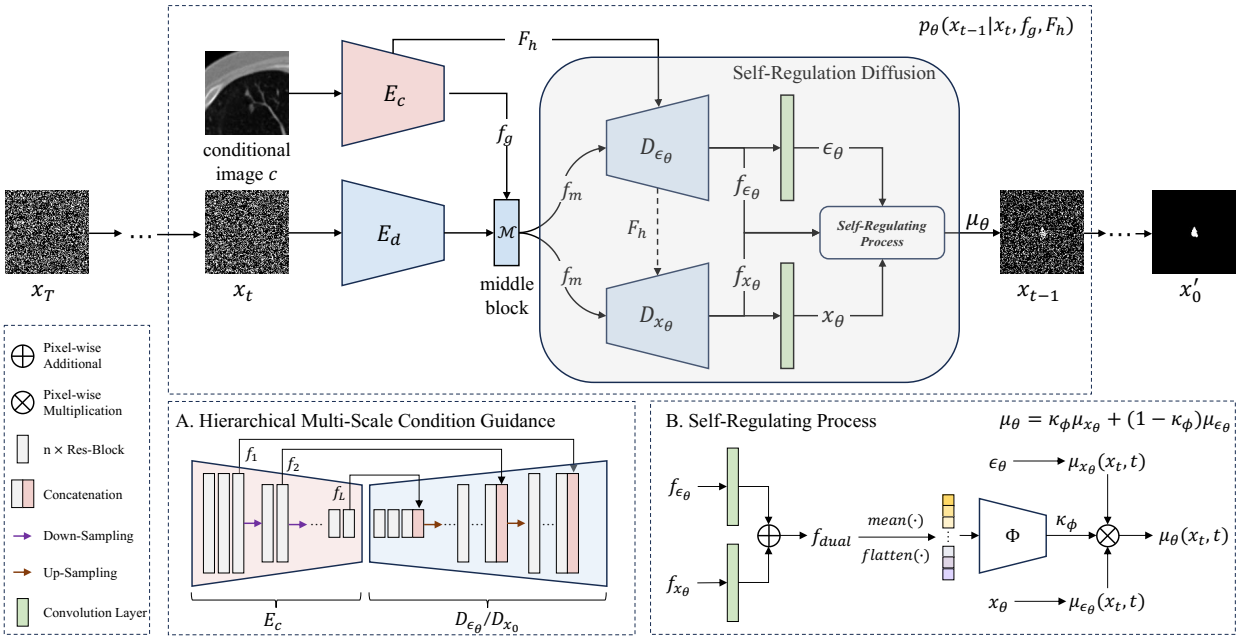


Figure 2: **The architecture of HSRDiff.** A) Hierarchical Multi-scale Condition Guidance. The feature of each step before down-sampling in E_C is composed of F_h, which is fused with the feature of corresponding resolution before each step of up-sampling in $D_{\epsilon_\theta/x_\theta}$. B) Self-Regulating Process. The mechanism takes f_{ϵ_θ}, f_{x_θ}, ϵ_θ, and x_θ as inputs. Through a series of transformations, a self-regulating parameter κ_ϕ is obtained from f_{ϵ_θ} and f_{x_θ}. We synthesize x_{t-1} in terms of κ_ϕ, ϵ_θ, and x_θ.

the hierarchical multi-scale condition guidance and the self-regulating dual-mode prediction pipeline of "differentiation to unification" to effectively improve the diversity of output space and the ability to reconstruct delicate structures.

Method

We start by reviewing preliminary knowledge, followed by a description of the HSRDiff structure. Finally, we outline the training and inference procedures for the algorithm.

Review Process

The diffusion model trains parameterized Markov chains by variational inference, demonstrating better performance than Generative Adversarial Models, Variance Auto-Encoder, and Normalization Flow Model for many tasks.

In simple terms, the diffusion model is the process of gradually denoising data from pure noise through neural network learning, which contains two steps of fixed forward process q and learnable reverse process p_θ.

The forward process is an iterative process of gradually adding noise to a sample x_0 to generate x_t for $t \in [1, T]$. According to (Ho, Jain, and Abbeel 2020), this iterative process can be formalized as Equ.(1).

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}; \beta_t\mathbf{I}) \tag{1}$$

Here β_t is a variable that varies with time t, and its changing strategies include linear timetables, cosine timetables, and so on. Specific to each step of the calculation, $x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon$, where ϵ is a standard Gaussian distribution $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. By converting with $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$, we can use Equ.(2) to transform x_0 to x_t at any time.

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0; (1-\alpha_t)\mathbf{I}) \tag{2}$$

The reverse process is treated as $p_\theta(x_{t-1}|x_t)$, where θ represents the parameters of the neural network. Since the noise we added in the forward process is Gaussian noise, we assume that the reverse denoising process also removes Gaussian noise. The Gaussian distribution is determined by mean μ_θ and variance σ_t, so the reverse process can be represented by Equ.(3):

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I}) \tag{3}$$

σ_t is a constant determined by β_t. Referring to (Ho, Jain, and Abbeel 2020), the simplified optimization objective of the diffusion model can be represented as follows.

$$\mathcal{L} = \mathbb{E}_{t,x_0,\epsilon}||\epsilon - \epsilon_\theta(\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, t)||^2 \tag{4}$$

Hierarchical Self-Regulation Diffusion

In Fig.2, we present the architecture of the Hierarchical Self-Regulation Diffusion (HSRDiff). In the training phase, HSRDiff receives pairs of conditional image c with dimension $B \times C \times W \times H$ and ground truth x_0 with dimension $B \times M \times H \times W$ as inputs in the training dataset, where the M is the class number. Since each c corresponds to multiple annotations, we randomly select one annotation as x_0 in each iteration. With a sufficient number of iterations, this ensures learning the entire annotation set with equal probability. HSRDiff's prediction involves three steps: (1) *Forward Process and Encoding*; (2) *Hierarchical Multi-scale Condition Guidance*; (3) *Self-Regulation Diffusion (SRDiff)*.

Algorithm 1: HSRDiff Training Procedure

1. A label is **randomly** sampled from the annotation set as x_0 if multiple annotations exist; otherwise, the label is directly loaded as x_0.
2. Sample the timestep t of each training process according to the diffusion step T, and then compute x_t according to $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.
3. Input x_t and the conditional image c corresponding to x_0 into HSRDiff to generate ϵ_θ, x_θ and κ_ϕ.
4. Train the HSRDiff according to $\bigtriangledown_\theta(\lambda_1(||\epsilon - \epsilon_\theta||^2 + ||x_0 - x_\theta||^2) + \lambda_2||\tilde{\mu_t} - \kappa_\phi\mu_{x_\theta} - (1-\kappa_\phi)\mu_{\epsilon_\theta}||^2)$ until the specified maximum epoch number.

Algorithm 2: HSRDiff Inference Procedure

Require: Sample x_T from the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ and load the conditional image c to be segmented from the test dataset as inputs;
Ensure: Multiple prediction results $\{x_0^j\}_{j=1}^N$, where N represent the sample number.

```
for j = 1, ..., N do
    for t = T, ..., 1 do
        If t > 1, sample ε ~ N(0, I); Otherwise ε = 0;
        Input x_t and c into HSRDiff to generate ε_θ, x_θ and κ_φ.
        Obtain x_{t-1} iteratively according to x_{t-1} ← μ_θ + σ_t ε,
        where μ_θ = κ_φ(1/√(1-β_t) x_t − β_t/(√(1-α_t)√(1-β_t)) ε_θ) + (1 − κ_φ)(√(1-β_t)(1-α_{t-1})/(1-α_t) x_t + √(α_{t-1})β_t/(1-α_t) x_θ).
    end for
    return x_0^j
end for
return {x_0^j}_{j=1}^N
```

Forward Process and Encoding Firstly, we add Gaussian noise on the input x_0 to obtain x_t according to Equ.(2). Then, we feed x_t and c into an image encoder E_d and a condition encoder E_c, respectively. The condition encoder E_c generates a global condition f_g and a hierarchical multi-scale condition set F_h in *Hierarchical Multi-scale Condition Guidance* process. f_g and the feature extracted by the image encoder E_d are input into a middle block $\mathcal{M}$ to extracts the feature f_m, which is defined as $f_m = \mathcal{M}(E_d(x_t), f_g)$. Finally, the hierarchical multi-scale condition set F_h and f_m will be fed into *Self-Regulation Diffusion* to generate the predicted noise ϵ_θ, the predicted label map x_θ and the distribution mean μ_θ modeled by HSRDiff.

Hierarchical Multi-scale Condition Guidance Briefly, the denoising process for each timestep t of HSRDiff can be represented as $p_\theta(x_{t-1}|x_t, f_g, F_h)$. We use a global condition f_g and a hierarchical multi-scale condition set F_h to model the top-down dependence between condition information and label maps. Specifically, the highest level features extracted from E_c serve as the global condition f_g. Features before each down-sampling of E_c represent $f_k, k = [1, 2, ..., L]$, and L indicates the number of ResBlock in E_c. These features form the hierarchical multi-scale condition set $F_h, F_h = [f_1, f_2, ..., f_L]$. The fusion of F_h and the denoising process is shown in Figure 2 (A). Each feature f_k in F_h as the condition is connected with the feature of the corresponding scale in D_{ϵ_θ} and D_{x_θ} before the usual up-sampling operation. In this way, D_{ϵ_θ} and D_{x_θ} learn the semantic features of the conditions at different scales in each step t and can therefore perform high-fidelity reconstruction of the fine structure in images. Detailed structures of E_c, D_{ϵ_θ}, and D_{x_θ} are shown in *Appendix B.1*.

Self-Regulation Diffusion Because they are entirely different modeling logic for decoders to predict noise with statistical properties and label maps with delicate structures, SRDiff is designed in the following two steps: i) *Differentiation:* Two independent decoding paths are used to predict noise ϵ and ground truth x_0, respectively, to express the distinction of different prediction modes better. ii) *Unification:* The application probability of the two modes is self-regulated based on the posterior information from the differentiation decoding process, achieving the unification of both modes. We cover these two steps in detail below.

Step1: Differentiation. The two independent decoding paths have their own separate decoders which are represented as $D_{\epsilon_\theta}$and D_{x_θ}. D_{ϵ_θ} and D_{x_θ} take f_m and F_h as inputs to generate intermediate features f_{ϵ_θ} and f_{x_θ}. Then f_{ϵ_θ} and f_{x_θ} pass through a independent convolutional layer C and are supervised by ϵ and x_0 to obtain the preliminary prediction ϵ_θ and x_θ. This differentiation process can be formalized as:

$$\mathbb{P}_\epsilon : f_{\epsilon_\theta} = D_{\epsilon_\theta}(f_m, F_h), \epsilon_\theta = C_{\epsilon_\theta}(f_{\epsilon_\theta}) \tag{5}$$

$$\mathbb{P}_x : f_{x_\theta} = D_{x_\theta}(f_m, F_h), x_\theta = C_{x_\theta}(f_{x_\theta}) \tag{6}$$

Step2: Unification (Self-Regulating process). As shown in Figure 2 (B), f_{ϵ_θ} and f_{x_θ} are used as the posterior information of the differentiation path for further extraction and fusion. First, we pass f_{ϵ_θ} and f_{x_θ} through two independent convolution layers and add the extracted features to generate f_{dual}. Then, f_{dual} is compressed by a small and simple downsampling network Φ to obtain a two-dimensional encoding κ_ϕ, which is called the self-regulating parameter.

$$\kappa_\phi = softmax(\Phi(flatten(\underbrace{C_1(f_{\epsilon_\theta}) + C_2(f_{x_\theta})}_{f_{dual}}))) \tag{7}$$

Finally, we can obtain the final distribution mean μ_θ of HSRDiff by using κ_ϕ to weighted summation of the distribution means corresponding to ϵ_θ and x_θ.

$$\mu_\theta = \kappa_\phi\mu_{x_\theta} + (1-\kappa_\phi)\mu_{\epsilon_\theta} \tag{8}$$

The training procedure of HSRDiff is shown in Algorithm 1. The new loss function $\mathcal{L}$ of HSRDiff also consists of two parts. The first part, $\mathcal{L}_d$, is the independent optimization of the two decoding paths, representing the differentiation between the two modes. The second part, L_u, is the overall optimization of the network modeling distribution, promoting the unification of the two modes.

$$\begin{aligned}\mathcal{L} = &[\underbrace{\lambda_1(||x_0 - x_\theta||^2 + ||\epsilon - \epsilon_\theta||^2)}_{differentation\ loss\ \mathcal{L}_d} \\ &+ \underbrace{\lambda_2||\tilde{\mu_t} - \kappa_\phi\mu_{x_\theta} - (1-\kappa_\phi)\mu_{\epsilon_\theta}||^2}_{unification\ loss\ \mathcal{L}_u}]\end{aligned} \tag{9}$$

Where $\tilde{\mu_t} = \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}x_0 + \frac{\sqrt{\alpha_t}(1-\alpha_{t-1})}{1-\alpha_t}x_t$, denotes the mean of the true posterior $q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}, \tilde{\mu_t}, \tilde{\beta_t}\mathbf{I})$. λ_1 and λ_2 are used to regulate the relative degree of differentiation and unification.

Inference

According to (Ho, Jain, and Abbeel 2020), the sampling of HSRDiff can be expressed as:

$$p_\theta(x_{t-1}|x_t, f_g, F_h) = \mathcal{N}(x_t, \mu_\theta, \sigma_t^2\mathbf{I}) \tag{10}$$

Since we use the "differentiation to unification" prediction pipeline, we need to modify the calculation of μ_θ in standard DDPM. For standard DDPM, when predicting noise ϵ or label map x_0 alone, the formula of the mean of distribution p_θ can be calculated as:

$$\mu_{\epsilon_\theta} = \frac{1}{\sqrt{1-\beta_t}}x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}}\epsilon_\theta \tag{11}$$

$$\mu_{x_\theta} = \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}x_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}x_\theta \tag{12}$$

According to Equ.(8), we perform a weighted addition of μ_{ϵ_θ} and μ_{x_θ} by self-regulating parameter κ_ϕ.

$$\begin{aligned}\mu_\theta =& \kappa_\phi\left(\frac{1}{\sqrt{1-\beta_t}}x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}\sqrt{1-\beta_t}}\epsilon_\theta\right) + (1 \\ &- \kappa_\phi)\left(\frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}x_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}x_\theta\right)\end{aligned} \tag{13}$$

Then inference from Gaussian noise x_t according to $x_{t-1} \longleftarrow \mu_\theta + \sigma_t\epsilon$, and iteratively denoise at each timestep until x_0 is obtained. To sum up, we can simply express the inference process of HSRDiff as the Algorithm 2.

Experiments

Datasets

Lung Nodule Segmentation (LIDC-IDRI) The LIDC-IDRI dataset is a typical abnormality dataset representing the CT image ambiguity. It contains 1018 lung 3D CT scans from 1010 lung cancer patients, where four anonymous experts annotate scans for each volume. The four experts first annotated all the scans independently and then adjusted their masks according to the annotations of others, so up to three masks per image can be empty sets. We obtain 2026 slices after preprocessing, and each slice is cut to a 128×128 patch. Finally, the training set consists of 1620 images, and the test set consists of 406 images.

Multiple Sclerosis Lesion Segmentation (MS-Lesion) The dataset includes 84 longitudinal MRI scans from 5 subjects (Carass et al. 2017). Two experts, with 4 and 10 years of experience, annotated the MS-related white matter lesions. Each MRI slice is converted into a 2D image with corresponding segmentation masks, featuring four scan types: PD, Flair, MP RAGE, and T2. The training set contains 2300 slices, and the test set includes 531 slices.

Multimodal Semantic Segmentation (Cityscapes) Cityscapes is a multi-class semantic segmentation dataset. It consists of street view images taken from cars with corresponding semantically segmented maps, where each image is labeled using 19 possible classes, and each image corresponds to one annotation. The official training dataset consists of 2975 images, and the validation dataset contains 500 images. We set up two experiments on Cityscapes. *Experimental Group 1:* Following (Kohl et al. 2018, 2019; Gao et al. 2022), we randomly flipped the five original semantic classes into five new classes with a certain probability for stochastic semantic segmentation. *Experimental Group 2:* We following (Zbinden et al. 2023) provided results in deterministic segmentation without injecting ambiguity. The detailed operation can be found in the *Appendix C*.

Experiment Setup

Metrics For the LIDC-IDRI, we use *Generalized Energy Distance* (GED), *Hungarian-Matched Intersection-over-Union* (HM-IoU), and *Maximum Dice Matching* (MDM) to measure model performance. Subscripts indicate the number of samples. For the Cityscapes dataset, we adopt the GED and HM-IoU in Experimental Group 1, and *Mean Intersection-over-Union* (mIoU) in Experimental Group 2. In Experimental Group 2, we fuse the results of multiple samples using the average predicted probability.

Implementation Details HSRDiff is implemented using Pytorch. For LIDC-IDRI, we crop images to 128×128 resolution and train for 500 epochs with a batch size of 40. MS-Lesion dataset uses the same settings as LIDC-IDRI but we resize the slice to 128×128. For Cityscapes, our two experimental groups followed the settings of (Kohl et al. 2018) and (Zbinden et al. 2023) respectively, training for 800 epochs with a batch size of 4. Across all experiments, we use 250 time steps with a linear noise schedule and the AdamW optimizer with a learning rate of 10^{-4}. Both λ_1 and λ_2 are set to 1. Based on the different sizes of C1 and C2, we define two model scales: HSRDiff and HSRDiff-light. Specific parameters can be found in the *Appendix B.2*.

LIDC: Typical Stochastic Segmentation

We quantitatively compare HSRDiff with the previous methods under sample number $N = \{16, 32\}$. It is worth noting in Table 1 that HSRDiff performs best on all metrics of GED, HM-IoU, and MDM, and HSRDiff-light performs second best. This proves that HSRDiff can achieve stable and excellent performance improvement in both large and small samples, and HSRDiff-light with a small parameter scale can also achieve better performance than existing works. Additionally, we obtain a deterministic segmentation result by averaging all generated samples ($N = 16$) for each method and reported their mIoU scores in Table 3. It shows that HSRDiff still achieves the best performance. Figure 3 shows the generated samples of MoSE, CIMD, and HSRDiff. Our results provide better consistency with ground truth and accurately captures fine structures like spiculations. To verify HSRDiff captures uncertainty, we provide entropy maps and more visualizations in the *Appendix D.1*.

MS-Lesion: Multiple Sclerosis Segmentation

Table 2 and Figure 4 present the comparison results of HSRDiff with previous works on the MS-Lesion dataset. As seen in Table 2, HSRDiff and HSRDiff-light achieve the best performance on most metrics. In some cases, HSRDiff-light even outperforms HSRDiff, indicating that the effectiveness of HSRDiff comes not from increased model parameters, but from the hierarchical multi-scale conditional guidance and

Method	$GED_{16}(\downarrow)$	$HM\text{-}IoU_{16}(\uparrow)$	$MDM_{16}(\uparrow)$	$GED_{32}(\downarrow)$	$HM\text{-}IoU_{32}(\uparrow)$	$MDM_{32}(\uparrow)$	#Params.
Prob. Unet (Kohl et al. 2018)	$0.354_{\pm 0.05}$	$0.518_{\pm 0.01}$	$0.687_{\pm 0.02}$	$0.336_{\pm 0.01}$	$0.523_{\pm 0.03}$	$0.681_{\pm 0.02}$	74.82M
HProb. Unet (Kohl et al. 2019)	$0.270_{\pm 0.01}$ †	$0.530_{\pm 0.01}$ †	-	-	-	-	-
PhiSeg (Baumgartner et al. 2019)	$0.262_{\pm 0.00}$ †	$0.586_{\pm 0.00}$ †	-	$0.247_{\pm 0.00}$ †	$0.595_{\pm 0.00}$ †	-	74.82M
MoSE (Gao et al. 2022)	$0.259_{\pm 0.003}$	$0.598_{\pm 0.002}$	$0.768_{\pm 0.004}$	$0.239_{\pm 0.004}$	$0.603_{\pm 0.002}$	$0.767_{\pm 0.002}$	41.63M
AB (Chen, Zhang, and Hinton 2022)	$0.221_{\pm 0.001}$	$0.621_{\pm 0.001}$	$0.787_{\pm 0.003}$	$0.201_{\pm 0.002}$	$0.630_{\pm 0.001}$	$0.792_{\pm 0.002}$	51M
CIMD (Rahman et al. 2023)	$0.241_{\pm 0.004}$	$0.577_{\pm 0.002}$	$0.805_{\pm 0.005}$	$0.228_{\pm 0.002}$	$0.580_{\pm 0.001}$	$0.814_{\pm 0.003}$	85.6M
CCDM (Zbinden et al. 2023)	$0.194_{\pm 0.001}$	$0.664_{\pm 0.002}$	$0.793_{\pm 0.001}$	$0.183_{\pm 0.001}$	$0.670_{\pm 0.001}$	$0.790_{\pm 0.003}$	41M
HSRDiff (Ours)	$\mathbf{0.181}_{\pm 0.002}$	$\mathbf{0.697}_{\pm 0.001}$	$\mathbf{0.891}_{\pm 0.003}$	$\mathbf{0.175}_{\pm 0.003}$	$\mathbf{0.702}_{\pm 0.002}$	$\mathbf{0.893}_{\pm 0.003}$	96.14M
HSRDiff-light (Ours)	$\underline{0.188}_{\pm 0.001}$	$\underline{0.677}_{\pm 0.001}$	$\underline{0.889}_{\pm 0.002}$	$\underline{0.180}_{\pm 0.002}$	$\underline{0.683}_{\pm 0.001}$	$\underline{0.891}_{\pm 0.002}$	12.89M

Table 1: Quantitative results on the LIDC-IDRI dataset. **Bold** and underline indicate the best and second-best of each metric. The score of (†) comes from (Zbinden et al. 2023), and the rest of the results are ours. Our results are verified over 3 times.

Method	$GED_{16}(\downarrow)$	$HM\text{-}IoU_{16}(\uparrow)$	$MDM_{16}(\uparrow)$	$GED_{32}(\downarrow)$	$HM\text{-}IoU_{32}(\uparrow)$	$MDM_{32}(\uparrow)$
Prob. Unet (Kohl et al. 2018)	$0.572_{\pm 0.002}$	$0.640_{\pm 0.003}$	$0.754_{\pm 0.002}$	$0.571_{\pm 0.001}$	$0.640_{\pm 0.002}$	$0.754_{\pm 0.003}$
MoSE (Gao et al. 2022)	$0.340_{\pm 0.002}$	$0.675_{\pm 0.001}$	$0.768_{\pm 0.003}$	$0.333_{\pm 0.002}$	$0.675_{\pm 0.002}$	$0.767_{\pm 0.001}$
AB (Chen, Zhang, and Hinton 2022)	$0.336_{\pm 0.003}$	$0.679_{\pm 0.001}$	$0.779_{\pm 0.002}$	$0.330_{\pm 0.001}$	$0.680_{\pm 0.001}$	$0.781_{\pm 0.001}$
CIMD (Rahman et al. 2023)	$0.311_{\pm 0.002}$	$0.699_{\pm 0.001}$	$0.831_{\pm 0.003}$	$0.307_{\pm 0.003}$	$0.698_{\pm 0.002}$	$0.833_{\pm 0.001}$
CCDM (Zbinden et al. 2023)	$\mathbf{0.303}_{\pm 0.001}$	$\underline{0.701}_{\pm 0.002}$	$0.816_{\pm 0.002}$	$\mathbf{0.301}_{\pm 0.002}$	$0.702_{\pm 0.001}$	$0.824_{\pm 0.001}$
HSRDiff (Ours)	$\underline{0.306}_{\pm 0.002}$	$\mathbf{0.702}_{\pm 0.001}$	$\underline{0.833}_{\pm 0.002}$	$\underline{0.304}_{\pm 0.001}$	$\underline{0.703}_{\pm 0.002}$	$\underline{0.834}_{\pm 0.001}$
HSRDiff-light (Ours)	$0.329_{\pm 0.002}$	$\mathbf{0.702}_{\pm 0.002}$	$\mathbf{0.835}_{\pm 0.001}$	$0.323_{\pm 0.002}$	$\mathbf{0.704}_{\pm 0.001}$	$\mathbf{0.837}_{\pm 0.001}$

Table 2: Quantitative results on the MS-Lesion dataset.

Method	LIDC-IDRI	MS-Lesion
Prob. Unet (Kohl et al. 2018)	0.611	0.640
AB (Chen, Zhang, and Hinton 2022)	0.689	0.701
MoSE (Gao et al. 2022)	0.708	0.738
CIMD (Rahman et al. 2023)	0.725	<u>0.776</u>
CCDM (Zbinden et al. 2023)	0.769	0.731
HSRDiff (Ours)	<u>0.803</u>	**0.778**
HSRDiff-light (Ours)	**0.825**	<u>0.776</u>

Table 3: mIoU of MS-Lesion and LIDC-IDRI dataset.

Method	GED_{16}	$HM\text{-}IoU_{16}$
Prob. Unet (Kohl et al. 2018)	0.206	0.512
MoSE (Gao et al. 2022)	0.203	0.580
AB (Chen, Zhang, and Hinton 2022)	0.194	0.605
CIMD (Rahman et al. 2023)	0.186	0.619
CCDM (Zbinden et al. 2023)	0.171	0.628
HSRDiff (Ours)	**0.161**	**0.643**
HSRDiff-light (Ours)	<u>0.165</u>	<u>0.637</u>

Table 4: Quantitative results on the Cityscapes dataset with stochastic semantic segmentation methods.

dynamic self-regulation diffusion foundation. Additionally, Figure 4 further demonstrates HSRDiff's superior ability to capture fine structural changes compared to other methods.

Method	mIoU		
	128×256	256×512	512×1024
DeepLabv3 (Chen et al. 2017)	0.433	0.565	0.571
DeepLabv3+ (Chen et al. 2018)	0.450	0.565	0.679
PSPNet (Zhao et al. 2017)	0.428	0.546	0.666
HRNet (Wang et al. 2020)	0.501	0.639	0.701
UNet++ (Zhou et al. 2019)	0.572	0.653	0.689
Prob. Unet (sample=16)	0.497	0.552	0.687
CCDM (sample=16)	0.569	0.643	0.711
HSRDiff (sample=1)	0.609	0.673	0.761
HSRDiff (sample=16)	<u>0.651</u>	**0.701**	**0.779**
HSRDiff-light (sample=1)	0.631	0.667	0.751
HSRDiff-light (sample=16)	**0.662**	<u>0.698</u>	<u>0.768</u>

Table 5: Quantitative results on the Cityscapes dataset with classic semantic segmentation methods.

Cityscapes: Multimodal Semantic Segmentation

Table 4 (Experiment Group 1) shows the results for stochastic semantic segmentation, evaluated on 16 samples. Table 5 (Experiment Group 2) presents the results for classic semantic segmentation, where we average the sampling results ($N = 16$) following (Zbinden et al. 2023) and evaluate at three resolutions: 128×256, 256×512, and 512×1024. Our method achieves state-of-the-art performance on GED, HM-IoU, and mIoU scores across both settings, demonstrating its effectiveness. Notably, HSRDiff-light outperforms HSRDiff at lower resolutions in classic segmentation, suggesting that self-regulating parameters derived from a large receptive field help capture fine structures in low-resolution

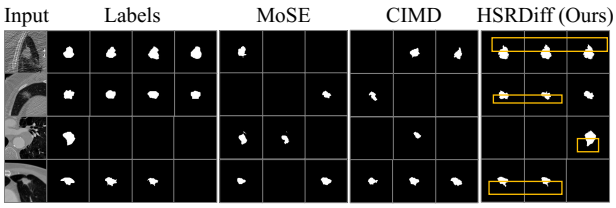


Figure 3: Qualitative analysis on the LIDC-IDRI dataset.

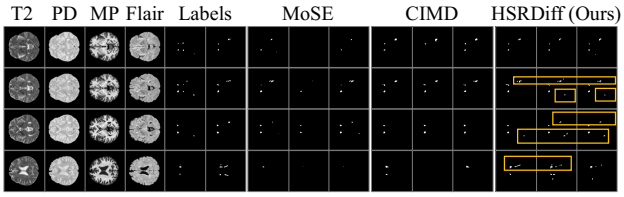


Figure 4: Qualitative analysis on the MS-Lesion dataset.

ϵ	x_0	f_g	F_h	dual	SRDiff	GED_{16}	HM-IoU_{16}	MDM_{16}
✓						0.425	0.429	0.594
✓		✓				0.276	0.626	0.881
✓		✓	✓			0.219	0.684	0.887
	✓	✓	✓			0.230	0.674	0.888
✓	✓	✓	✓	✓		0.205	0.685	0.886
✓	✓	✓	✓		✓	**0.181**	**0.697**	**0.891**

Table 6: Ablation studies for stochastic segmentation on LIDC-IDRI. Results of the first behavior standard DDPM.

images. As shown in Figure 5, our model predicts multi-modal outputs with uncertainty while achieving fine-grained semantic segmentations.

Ablation Study

We constructed five ablation settings using ϵ prediction as the baseline. x_0 and ϵ are the targets of the model. f_g and F_h represent global condition guidance and hierarchical multi-scale condition guidance, respectively. "dual" refers to the independent prediction of x_0 and ϵ using the method from (Benny and Wolf 2022). SRDiff is our proposed "differentiation to unification" prediction pipeline. Table 6 shows the ablation results of HSRDiff on the LIDC-IDRI dataset. First, introducing condition information significantly improves all metrics (Line 1 vs. Line 2). Second, adding hierarchical multi-scale condition priors further boosts the metrics (Line 2 vs. Line 3). Finally, using SRDiff achieves the best performance across all metrics (Line 6), demonstrating the effectiveness of our proposed methods. Additionally, we compared with the multi-objective training method from (Benny and Wolf 2022) (Line 5 vs. Line 6). As expected, our "differentiation to unification" pipeline shows better results.

Discussion of Self-Regulating Parameter κ_ϕ

Self-regulating parameter across datasets. Figure 6(a) shows notable variations in self-regulation parameters across datasets. For LIDC-IDRI, as denoising step t increases, the model shifts from a slight preference for ϵ prediction to a neutral stance. For Cityscape, the model transitions from relying primarily on x_0 to almost exclusively using x_0. For MS-Lesion, the dispersed and small lesions cause greater parameter fluctuations, though x_0 predictions remain favored overall. These differences reflect dataset-specific preferences for prediction modes and demonstrate HSRDiff's ability to dynamically self-regulate its modeling path, ensuring high performance across domains.

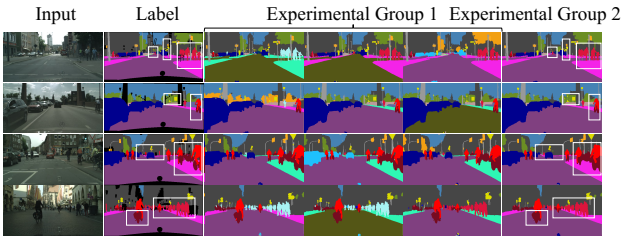
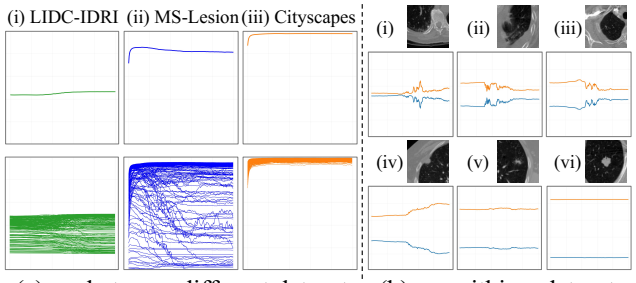


Figure 5: The visualization of HSRDiff on the Cityscapes.



(a) κ_ϕ between different dataset (b) κ_ϕ within a dataset

Figure 6: The trend of the self-regulating parameter κ_ϕ. The horizontal axis of all figures represents the diffusion step t (from left to right, 0-250), and the vertical axis represents the self-regulation parameter κ_ϕ (from bottom to top, 0-1).

Self-regulating parameters within a dataset. Figure 6(b) shows that as lung nodule uncertainty decreases from (i) to (vi), the fluctuations in self-regulation parameters also stabilize. This visualizes HSRDiff's self-regulation process.

In summary, self-regulation parameters can help the model find optimal prediction paths across datasets and generate the best predictions for various cases within a dataset.

Conclusion

This paper introduces a Hierarchical Self-Regulation Diffusion (HSRDiff) framework, which learns the probability distribution of targets within a label set and generates multiple segmentation hypotheses. HSRDiff includes two key components: Self-Regulation Diffusion (SRDiff) and Hierarchical Multi-Scale Condition Guidance. SRDiff dynamically determines the optimal path for modeling target distribution using a novel "differentiation to unification" prediction pipeline. Hierarchical Multi-Scale Condition Guidance ensures high-fidelity segmentation of delicate structures in images. HSRDiff demonstrates significant performance improvements across three different semantic scenarios.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No.62476264 and No.62406312), China National Postdoctoral Program for Innovative Talents (No.BX20240385) funded by China Postdoctoral Science Foundation, Beijing Natural Science Foundation (No.4244098), and Science Foundation of the Chinese Academy of Sciences.

References

Amit, T.; Shaharbany, T.; Nachmani, E.; and Wolf, L. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12): 2481–2495.

Baumgartner, C. F.; Tezcan, K. C.; Chaitanya, K.; Hötker, A. M.; Muehlematter, U. J.; Schawkat, K.; Becker, A. S.; Donati, O.; and Konukoglu, E. 2019. Phiseg: Capturing uncertainty in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 119–127. Springer.

Benny, Y.; and Wolf, L. 2022. Dynamic dual-output diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11482–11491.

Bian, C.; Yuan, C.; Wang, J.; Li, M.; Yang, X.; Yu, S.; Ma, K.; Yuan, J.; and Zheng, Y. 2020. Uncertainty-aware domain alignment for anatomical structure segmentation. *Medical Image Analysis*, 64: 101732.

Carass, A.; Roy, S.; Jog, A.; Cuzzocreo, J. L.; Magrath, E.; Gherman, A.; Button, J.; Nguyen, J.; et al. 2017. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage*, 148: 77–102.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, T.; Zhang, R.; and Hinton, G. 2022. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*.

Chung, H.; Sim, B.; and Ye, J. C. 2022. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12413–12422.

Feng, W.; Yang, C.; An, Z.; Huang, L.; Diao, B.; Wang, F.; and Xu, Y. 2024. Relational diffusion distillation for efficient image generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 205–213.

Gao, Z.; Chen, Y.; Zhang, C.; and He, X. 2022. Modeling Multimodal Aleatoric Uncertainty in Segmentation with Mixture of Stochastic Expert. *arXiv preprint arXiv:2212.07328*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hüllermeier, E.; and Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110.

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

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Kohl, S.; Romera-Paredes, B.; Meyer, C.; De Fauw, J.; Ledsam, J. R.; Maier-Hein, K.; Eslami, S.; Jimenez Rezende, D.; and Ronneberger, O. 2018. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31.

Kohl, S. A.; Romera-Paredes, B.; Maier-Hein, K. H.; Rezende, D. J.; Eslami, S.; Kohli, P.; Zisserman, A.; and Ronneberger, O. 2019. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*.

Liu, Z.; Zhang, F.; He, J.; Wang, J.; Wang, Z.; and Cheng, L. 2023. Text-guided mask-free local image retouching. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2783–2788. IEEE.

Monteiro, M.; Le Folgoc, L.; Coelho de Castro, D.; Pawlowski, N.; Marques, B.; Kamnitsas, K.; van der Wilk, M.; and Glocker, B. 2020. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in neural information processing systems*, 33: 12756–12767.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Rahman, A.; Valanarasu, J. M. J.; Hacihaliloglu, I.; and Patel, V. M. 2023. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11536–11546.

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

Tanno, R.; Worrall, D. E.; Ghosh, A.; Kaden, E.; Sotiropoulos, S. N.; Criminisi, A.; and Alexander, D. C. 2017. Bayesian image quality transfer with CNNs: exploring uncertainty in dMRI super-resolution. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017:*

20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20, 611–619. Springer.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Whang, J.; Delbracio, M.; Talebi, H.; Saharia, C.; Dimakis, A. G.; and Milanfar, P. 2022. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16293–16303.

Wolleb, J.; Sandkühler, R.; Bieder, F.; Valmaggia, P.; and Cattin, P. C. 2022. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, 1336–1348. PMLR.

Wu, J.; Fu, R.; Fang, H.; Zhang, Y.; Yang, Y.; Xiong, H.; Liu, H.; and Xu, Y. 2022. Medsegdiff: Medical image segmentation with diffusion probabilistic model. *arXiv preprint arXiv:2211.00611*.

Xu, Z.; Qin, J.; Li, C.; Bu, D.; and Zhao, Y. 2024. MiHATP:A Multi-hybrid Attention Super-Resolution Network for Pathological Image Based on Transformation Pool Contrastive Learning. In Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; and Schnabel, J. A., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 488–497. Cham: Springer Nature Switzerland.

Yang, C.; An, Z.; Huang, L.; Bi, J.; Yu, X.; Yang, H.; Diao, B.; and Xu, Y. 2024a. CLIP-KD: An Empirical Study of CLIP Model Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15952–15962.

Yang, C.; An, Z.; Zhou, H.; Zhuang, F.; Xu, Y.; and Zhang, Q. 2023a. Online knowledge distillation via mutual contrastive learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10212–10227.

Yang, C.; Zhou, H.; An, Z.; Jiang, X.; Xu, Y.; and Zhang, Q. 2022a. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12319–12328.

Yang, H.; Shen, L.; Zhang, M.; and Wang, Q. 2022b. Uncertainty-guided lung nodule segmentation with feature-aware attention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 44–54. Springer.

Yang, H.; Wang, Q.; Zhang, Y.; An, Z.; Chen, L.; Zhang, X.; and Zhou, S. K. 2023b. Lung Nodule Segmentation and Uncertain Region Prediction with an Uncertainty-Aware Attention Mechanism. *IEEE Transactions on Medical Imaging*.

Yang, Y.; Cheng, D.; Fang, C.; Wang, Y.; Jiao, C.; Cheng, L.; and Wang, N. 2024b. Diffusion-based Layer-wise Semantic Reconstruction for Unsupervised Out-of-Distribution Detection.

Zbinden, L.; Doorenbos, L.; Pissas, T.; Huber, A. T.; Sznitman, R.; and Márquez-Neila, P. 2023. Stochastic segmentation with conditional categorical diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1119–1129.

Zhang, W.; Zhang, X.; Huang, S.; Lu, Y.; and Wang, K. 2022. A probabilistic model for controlling diversity and accuracy of ambiguous medical image segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4751–4759.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

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