# Adaptively Distilled ControlNet: Accelerated Training and Superior Sampling for Medical Image Synthesis

Kunpeng Qiu[1,2], Zhiying Zhou[1,2], and Yongxin Guo[1,2,3(✉)]

[1] National University of Singapore, Singapore, Singapore
[2] National University of Singapore Suzhou Research Institute, Suzhou, China
[3] City University of Hong Kong, Hong Kong, China
kunpeng_qiu@u.nus.edu, elezzy@nus.edu.sg, yongxin.guo@cityu.edu.hk

**Abstract.** Medical image annotation is constrained by privacy concerns and labor-intensive labeling, significantly limiting the performance and generalization of segmentation models. While mask-controllable diffusion models excel in synthesis, they struggle with precise lesion-mask alignment. We propose **Adaptively Distilled ControlNet**, a task-agnostic framework that accelerates training and optimization through dual-model distillation. Specifically, during training, a teacher model, conditioned on mask-image pairs, regularizes a mask-only student model via predicted noise alignment in parameter space, further enhanced by adaptive regularization based on lesion-background ratios. During sampling, only the student model is used, enabling privacy-preserving medical image generation. Comprehensive evaluations on two distinct medical datasets demonstrate state-of-the-art performance: TransUNet improves mDice/mIoU by 2.4%/4.2% on KiTS19, while SANet achieves 2.6%/3.5% gains on Polyps, highlighting its effectiveness and superiority. Code is available at https://github.com/Qiukunpeng/ADC.

**Keywords:** Diffusion models · Medical Image Synthesis · Medical Image Segmentation.

## 1 Introduction

In medical image analysis, large, accurately annotated datasets are essential for high-performance segmentation [34,20]. Despite the rapid progress in deep learning [33,4,13,2], the high cost of acquiring annotated medical images, coupled with privacy and copyright constraints [23,6], hinders the full potential of segmentation models.

To mitigate data scarcity issue, diffusion models [28,10,21] have emerged as a leading paradigm for synthetic data generation, offering both training stability and high-fidelity image synthesis. Several existing approaches leverage lesion-free images [16,24] to synthesize abnormal samples; however, these methods fail to fully address privacy concerns. In contrast, mask-controllable synthesis eliminates the need for costly manual annotations and ethical constraints

while providing a more accessible and streamlined framework, making it a compelling alternative for broader adoption [6,23,5]. Regardless of the approach, precise lesion-mask alignment remains a notorious challenge in existing methods [19,35,15,6]. In this work, we advance the mask-controllable synthesis paradigm to generate high-quality synthetic medical images, specifically tackling lesion alignment limitations to enhance downstream segmentation performance.

To address this, studies [15,6] have embedded pretrained segmentation models within diffusion frameworks to provide iterative feedback, refining noise prediction. However, their reliance on pretrained segmentation models renders these methods task-specific and may introduce inherent biases into synthetic data. In a related effort, [6] introduces adaptive weighting to enhance lesion representation, yet the disproportionately low weight assigned to lesion-free regions impairs learning, leading to degraded image fidelity even after extensive training.

To overcome these limitations, we propose the **Adaptively Distilled ControlNet**, a novel field distillation framework [9,18,27]. Our approach leverages the regularization property of controllable diffusion models [3,11], where conditional inputs act as implicit regularizers to ensure stable optimization and enhanced image quality. Specifically, we adopt a teacher-student paradigm, where the teacher model—conditioned on mask-image pairs—regularizes the noise prediction of the student model, which is conditioned only on masks. A shared forward noise addition process enables a dual-diffusion decoder architecture. Furthermore, an adaptive weight distillation strategy reinforces lesion representation while preserving distributional fidelity. During sampling, the student model runs at ControlNet [35] speed while ensuring diversity and scalability without extra image conditions.

Our contributions are summarized as follows: (1) We introduce Adaptively Distilled ControlNet, which significantly accelerates training convergence and data fitting. Moreover, its task-agnostic nature allows seamless adaptation to diverse datasets and modalities without requiring modifications to the model architecture. (2) We propose Adaptive Distillation Loss, which substantially enhances lesion-mask alignment in synthetic images, generating high-quality training data for segmentation models. This ensures superior performance and generalization in downstream segmentation tasks. (3) Extensive experiments demonstrate that our method surpasses existing approaches in both image fidelity and segmentation accuracy. Specifically, TransUNet achieves 2.4% mDice and 4.2% mIoU improvements on the KiTS19 dataset, while SANet attains 2.6% mDice and 3.5% mIoU gains on Polyps, underscoring the efficacy of our approach.

## 2 Preliminary

Diffusion models [10,28] formalize data generation through two coupled chains: a destructive forward process that gradually corrupts data with Gaussian noise, and a learned reverse process that iteratively recovers the original signal. Following the standard variance-preserving formulation [10], the denoising network

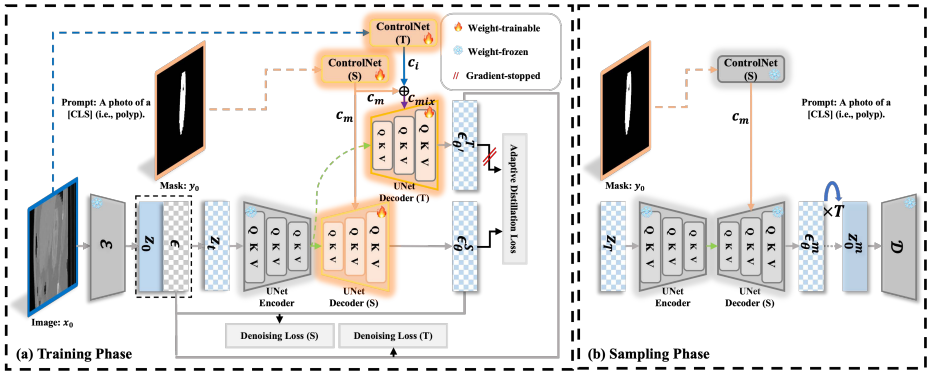


**Fig. 1.** (a) Illustration of our method during the training phase. (b) During sampling, only the student model is utilized with arbitrary masks.

$\epsilon_\theta(x_t, t)$ directly predicts the noise, reducing the training objective to:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_t,t,\epsilon}\left[\|\epsilon_\theta(x_t,t) - \epsilon\|_2^2\right], \tag{1}$$

where $t \sim \mathcal{U}\{1, T\}$ and $x_t$ is the noisy image.

Stable Diffusion [21] refines this framework through latent space optimization. A pretrained VAE [31] encoder $\mathcal{E}$ maps images $x_0$ into compact latent representations $z_0 = \mathcal{E}(x_0)$, facilitating diffusion in a reduced-dimensional space. Various extensions [19,35,15] of this model enable conditional generation via text prompts $c_t$ and task-specific control signals $c_f$, allowing for more precise content modulation. The generalized training objective is expressed as:

$$\mathcal{L}_{\text{cond}} = \mathbb{E}_{z_t,t,c_t,c_f,\epsilon}\left[\|\epsilon_\theta(z_t,t,c_t,c_f) - \epsilon\|_2^2\right]. \tag{2}$$

## 3 Methodology

### 3.1 Architecture of Adaptively Distilled ControlNet

Building upon the established ControlNet framework [35], we propose a distilled dual-branch diffusion architecture with shared latent projection, as illustrated in Fig. 1(a). The frozen VAE [31] encoder $\mathcal{E}$ establishes a deterministic mapping $\mathcal{E} : x_0 \mapsto z_0$ through latent space embedding, where $x_0$ denotes the input image and $z_0$ its latent representation. The student branch (S) ingests conditional masks through a dedicated ControlNet (S) module, generating encoded mask features $c_m$ that integrate with the student diffusion U-Net Decoder (S) through feature injection for noise prediction $\epsilon_\theta^S$.

The teacher branch (T) processes the paired image through a parallel ControlNet (T) to extract encoded image features $c_i$. These image features are fused with the corresponding mask features $c_m$ through element-wise summation:

$$c_{\text{mix}} = c_i + c_m. \tag{3}$$

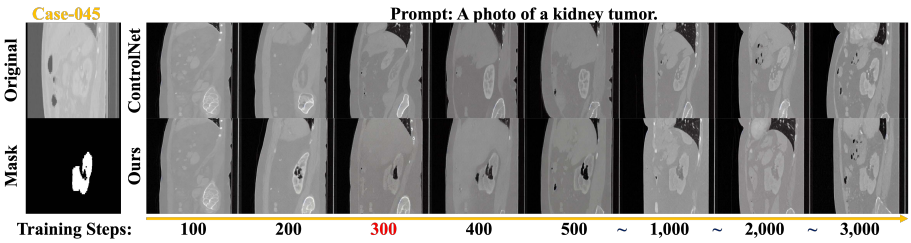


**Fig. 2.** Visualizing the difference between ControlNet and our method in training convergence and data fitting.

**Table 1.** Comparison of synthetic medical image quality generated by each method.

| Metrics | Polyps | | | | | KiTS19 | | |
|---|---|---|---|---|---|---|---|---|
| | SinGAN | ArSDM | T2I-Adapter | ControlNet | Ours | T2I-Adapter | ControlNet | Ours |
| FID (↓) | 103.142 | 98.085 | 150.546 | **65.609** | 66.587 | 92.717 | **69.240** | 70.786 |
| CLIP-I (↑) | 0.851 | 0.845 | 0.874 | 0.884 | **0.901** | 0.814 | 0.833 | **0.839** |

This fused representation $c_{\text{mix}}$ propagates through the teacher's diffusion U-Net decoder (T) to predict the noise $\epsilon^T_{\theta'}$. By sharing the forward process between the student and teacher branches, the architecture employs a unified latent space projection and diffusion U-Net encoder, significantly optimizing memory efficiency. The composite objective function integrates the following components:

$$\mathcal{L} = \underbrace{\mathcal{L}_S + \mathcal{L}_T}_{\text{Denoising Objectives}} + \underbrace{\mathcal{L}_{\text{Ada}}}_{\text{Distillation Regularizer}}, \tag{4}$$

with Denoising Objectives defined as:

$$\begin{aligned} \mathcal{L}_S &= \mathbb{E}_{z_t,t,c_t,c_m,\epsilon}\left[\|\epsilon_\theta(z_t,t,c_t,c_m)-\epsilon\|_2^2\right], \\ \mathcal{L}_T &= \mathbb{E}_{z_t,c_t,c_{mix},t,\epsilon}\left[\|\epsilon_{\theta'}(z_t,t,c_t,c_{\text{mix}})-\epsilon\|_2^2\right], \end{aligned} \tag{5}$$

where $\theta$ and $\theta'$, as in ControlNet [35], are both initialized with the parameters of a pretrained diffusion model, denote mutually independent parameters for each branch, and are optimized separately during training. Meanwhile, $\epsilon \sim \mathcal{N}(0, I)$ ensures stochastic consistency.

During sampling, as shown in Fig. 1(b), medical images are generated using the student branch with arbitrary masks at the same speed as ControlNet [35].

### 3.2 Adaptive Distillation Loss

The spatial alignment between synthesized lesion regions and their corresponding masks is critical for downstream segmentation tasks. However, the severe lesion-background imbalance in medical image synthesis often leads to the underrepresentation of lesion regions. To address this issue, we propose a spatially

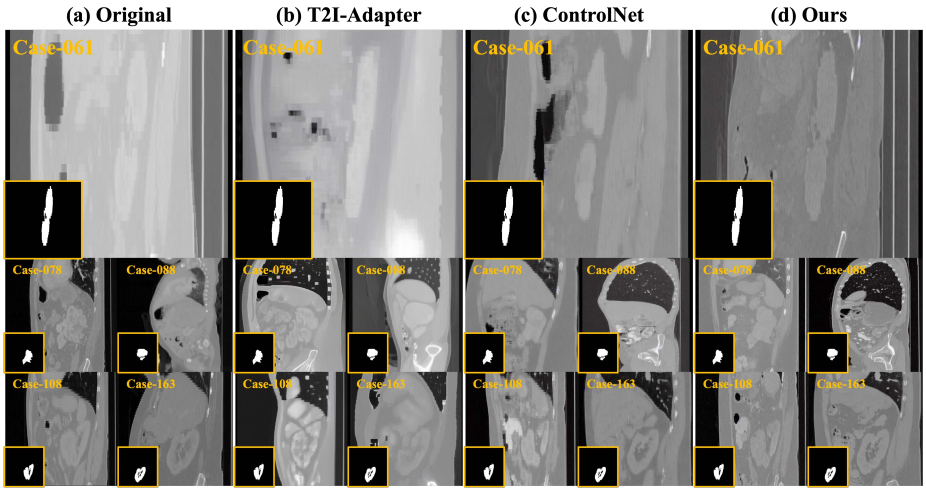


**Fig. 3.** Examples of real and synthetic kidney tumor images generated by each method.

**Table 2.** Comparisons of different methods applied on tumor segmentation baselines.

| Methods | TransUNet | | | | nnUNet | | | |
|---|---|---|---|---|---|---|---|---|
| | mDice | mIoU | Accuracy | Recall | mDice | mIoU | Accuracy | Recall |
| Real Dataset | 92.8 | 86.9 | 98.6 | 91.5 | 96.5 | 93.4 | 99.3 | 96.4 |
| +Copy-Paste | 93.3 | 87.7 | 98.7 | 91.5 | 96.5 | 93.6 | 99.3 | 96.0 |
| +T2I-Adapter | 94.5 | 89.9 | **99.0** | 92.6 | 96.3 | 93.6 | **99.8** | 95.8 |
| +ControlNet | 94.6 | 90.0 | **99.0** | **93.9** | 96.1 | 93.2 | **99.8** | 95.8 |
| +Ours | **95.2** | **91.1** | **99.0** | 93.8 | **97.9** | **96.0** | 99.6 | **97.8** |

adaptive distillation mechanism that enables the teacher model to dynamically modulate the regularization intensity for the student model, thereby emphasizing the learning of lesion-specific morphological features in the student model.

Unlike previous approaches that apply reweighting techniques to denoising losses [6], our method introduces lesion-aware attention through dual-stream gradient modulation, effectively addressing the lesion-background imbalance. The adaptive weight $w_{\text{Ada}}$ is derived from the mask statistics, with distinct weights assigned to lesion and lesion-free regions:

$$w_{\text{Ada}} = \begin{cases} \frac{N_{\text{lesion-free}}}{N_{\text{total}}}, & \text{for lesion regions} \\ \frac{N_{\text{lesion}}}{N_{\text{total}}}, & \text{otherwise} \end{cases} \tag{6}$$

where $N_{\text{lesion}}$ and $N_{\text{lesion-free}}$ denote pixel counts for respective regions, and $N_{\text{total}} = H \times W$ represents the total number of pixels in the image. These weights are normalized to form a spatially adaptive $W \times H$ weight matrix. The final adaptive distillation loss is formulated as:

$$\mathcal{L}_{\text{Ada}} = \mathbb{E}_{z_t,t}\left[w_{\text{Ada}} \cdot \|\epsilon_\theta^S - \text{sg}(\epsilon_{\theta'}^T)\|_2^2\right], \tag{7}$$

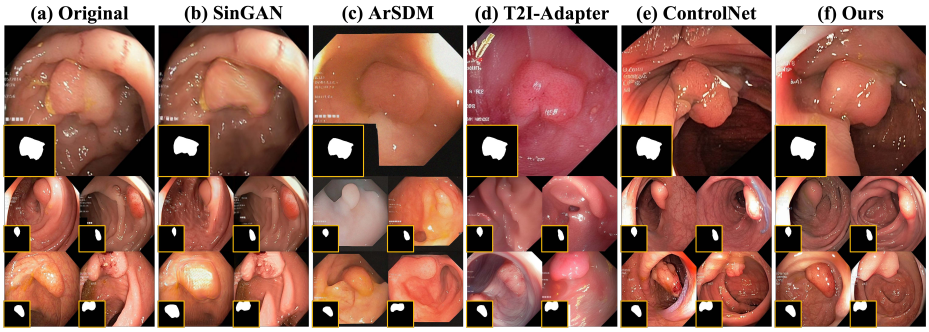


**Fig. 4.** Examples of real and synthetic polyp images generated by each method.

where sg(·) indicates stop-gradient operation.

## 4 Experiment

### 4.1 Dataset and Evaluation Metrics

We evaluate our method on two publicly available medical datasets: Polyps [14,1] (**RGB**) and KiTS19 [7] (**CT**, 2D slices), referred to as **Real Datasets**.

**Generative Model Training**: For Polyps, we use images from Kvasir [14] and CVC-ClinicDB [1]. For KiTS19 [7], 50 cases are randomly selected from 210 labeled cases, sliced into 2D, filtering out lesion-free slices.

**Generative Model Sampling and Evaluation**: Following [6], synthetic images are generated using masks from **Real Datasets**, referred to as **Synthetic Datasets**, and evaluated using FID [8] and CLIP-I [22].

**Segmentation Model Training**: **Synthetic Datasets** are combined with the **Real Datasets** as a new training set to train segmentation models.

**Segmentation Model Testing and Evaluation**: The Polyps test set includes images from five public datasets: EndoScene [32], CVC-ClinicDB [1], Kvasir [14], CVC-ColonDB [29], and ETIS [25]. For KiTS19 [7], 10 non-overlapping cases are selected from 210 labeled cases, sliced into 2D, filtering out slices without lesions. Evaluation metrics include mDice and mIoU for Polyps, and mDice, mIoU, Accuracy, and Recall for KiTS19.

### 4.2 Implementation Details

We detail the configuration of the generative and segmentation models as follows:

**Generative Model**: We use the pre-trained Stable Diffusion v1.5 [21]. The training setup is the same for both datasets: the AdamW [17] optimizer with a learning rate of $10^{-5}$ and weight decay of $10^{-2}$ is used for 3,000 iterations on 8×NVIDIA 4090 GPUs (global batch size of 32) with $384^2$ resolution inputs. A 5% probability for prompt dropout is applied. Sampling employs classifier-free guidance [11] (CFG=9) and deterministic DDIM [26] sampling ($\eta = 0$, 50

**Table 3.** Comparisons of different methods applied on polyp segmentation baselines.

| Methods | EndoScene | | ClinicDB | | Kvasir | | ColonDB | | ETIS | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mDice | mIoU | mDice | mIoU | mDice | mIoU | mDice | mIoU | mDice | mIoU | mDice | mIoU |
| nnUNet | 84.3 | 76.0 | **89.7** | **85.0** | 89.7 | 84.3 | 77.2 | 69.2 | 69.1 | 61.5 | 78.3 | 70.9 |
| +Copy-Paste | 85.0 | 76.8 | 89.5 | **85.0** | 89.8 | 84.3 | **77.7** | **70.2** | 69.4 | 61.8 | 78.7 | 71.5 |
| +SinGAN | 86.5 | 79.4 | 88.8 | 84.0 | 90.2 | 85.4 | 71.7 | 65.7 | 66.7 | 60.5 | 75.2 | 69.3 |
| +ArSDM | 86.2 | 79.1 | 89.3 | 84.5 | 90.2 | 84.8 | 75.3 | 68.0 | 73.2 | 65.7 | 78.6 | 71.7 |
| +T2I-Adapter | 83.9 | 76.6 | 87.9 | 82.9 | 91.1 | 85.5 | 75.5 | 68.9 | 69.2 | 61.7 | 78.0 | 70.9 |
| +ControlNet | 84.2 | 76.5 | 88.6 | 83.8 | 89.9 | 84.5 | 73.6 | 66.0 | 66.6 | 59.1 | 75.9 | 68.8 |
| +Ours | **87.7** | **79.8** | 88.9 | 84.0 | **91.3** | **85.9** | 76.2 | 68.8 | **74.3** | **67.8** | **79.5** | **72.7** |
| SANet | 88.8 | 81.5 | 91.6 | 85.9 | 90.4 | 84.7 | 75.3 | 67.0 | 75.0 | 65.4 | 79.4 | 71.4 |
| +Copy-Paste | 89.7 | 83.0 | 90.2 | 85.1 | 90.3 | 84.8 | 77.7 | 70.0 | 77.4 | 68.8 | 81.1 | 73.7 |
| +SinGAN | 88.3 | 81.6 | 90.9 | 85.3 | 91.0 | **85.8** | 77.3 | 69.4 | 73.7 | 65.4 | 80.0 | 72.6 |
| +ArSDM | **90.2** | **83.2** | 91.4 | 86.1 | 91.1 | 85.6 | 77.7 | 70.0 | 78.0 | 69.5 | 81.5 | 74.1 |
| +T2I-Adapter | 89.1 | 81.9 | 91.2 | 85.5 | 90.4 | 84.5 | 77.6 | 70.2 | 76.4 | 67.2 | 81.1 | 73.3 |
| +ControlNet | 89.3 | 82.1 | 91.1 | 85.8 | 90.8 | 85.2 | 76.2 | 68.2 | 75.7 | 65.8 | 80.0 | 72.2 |
| +Ours | 89.2 | 83.1 | **92.9** | **87.4** | **91.2** | 85.6 | **77.8** | **70.4** | **79.6** | **71.8** | **82.0** | **74.9** |
| Polyp-PVT | 90.0 | 83.3 | 93.7 | 88.9 | 91.7 | 86.4 | 80.8 | 72.7 | 78.7 | 70.6 | 83.3 | 76.0 |
| +Copy-Paste | 88.0 | 80.9 | 93.4 | 88.7 | 91.7 | 87.1 | 79.8 | 71.8 | 79.2 | 71.3 | 82.8 | 75.6 |
| +SinGAN | 87.0 | 79.7 | 91.7 | 87.0 | **92.8** | **88.1** | 76.9 | 69.0 | 74.2 | 66.7 | 80.1 | 73.0 |
| +ArSDM | 88.2 | 81.2 | 92.2 | 87.5 | 91.5 | 86.3 | 81.7 | 73.8 | 80.6 | 72.9 | 84.0 | 76.7 |
| +T2I-Adapter | 89.2 | 82.4 | **94.0** | **89.2** | 90.4 | 85.0 | 79.6 | 71.7 | 78.1 | 69.8 | 82.4 | 75.1 |
| +ControlNet | 86.1 | 78.8 | 91.3 | 85.9 | 91.1 | 86.2 | 79.7 | 71.4 | 78.7 | 70.2 | 82.3 | 74.6 |
| +Ours | **90.3** | **83.8** | 93.0 | 88.5 | 92.0 | 87.2 | **82.0** | **74.1** | **80.8** | **73.1** | **84.4** | **77.3** |

steps), as described in [35]. T2I-Adapter [19] and ControlNet [35] share the same configuration as our method, while SinGAN [30] and ArSDM [6] use their default settings. Notably, for ControlNet [35], unlocking the weights of Stable Diffusion is more effective for medical image synthesis.

**Segmentation Model**: Both CNN-based and Transformer-based models are utilized with default configurations. Specifically, nnUNet [13] is trained for 200 epochs with five-fold cross-validation, and the final results are obtained by ensembling five models, followed by postprocessing.

### 4.3 Qualitative Comparison

Fig. 2 demonstrates that the teacher model's adaptive regularization accelerates the student model's data fitting within approximately 300 steps, mitigating the sudden convergence phenomenon in ControlNet [35].

Fig. 3 and Fig. 4 present kidney tumor and polyp images generated by various methods. SinGAN [30], although designed for the Polyps dataset, often introduces artifacts and lacks diversity. ArSDM [6] suffers from texture degradation in polyps and fails to generalize to KiTS19 due to its task-specific nature. T2I-Adapter [19] generates unrealistic textures in RGB data and underperforms on CT data. ControlNet [35] struggles with mask-lesion alignment. In contrast, our

**Table 4.** Comparison of the impact of $\mathcal{L}_{\text{Ada}}$ on kidney tumor image segmentation.

| Settings | TransUNet | | | | nnUNet | | | |
|---|---|---|---|---|---|---|---|---|
| | mDice | mIoU | Accuracy | Recall | mDice | mIoU | Accuracy | Recall |
| w/o | 94.6 | 90.0 | **99.0** | **93.9** | 96.1 | 93.2 | **99.8** | 95.8 |
| w/(Standard) | 94.9 | 90.6 | **99.0** | 93.5 | 97.4 | 95.3 | 99.6 | 97.6 |
| w/(Adaptive) | **95.2** | **91.1** | **99.0** | 93.8 | **97.9** | **96.0** | 99.6 | **97.8** |

model excels in both mask-lesion alignment and morphological features, clearly outperforming the others.

### 4.4 Quantitative Comparisons

Table 1 shows FID [8] and CLIP-I [22] results. Notably, more precise mask-lesion alignment does not significantly lower the FID score, with our method's FID score slightly higher than ControlNet [35]. We attribute this to the inherent limitations of FID [12], which overfits with limited data. Nevertheless, CLIP-I [22] confirms our method achieves higher semantic similarity.

Table 2 and Table 3 highlight the enhancement of segmentation models using synthetic data from various generative models. We establish a new baseline by retraining models on a duplicated dataset (*i.e.*, "Copy-Paste"). Our method significantly outperforms others. On KiTS19 [7], it improves mDice by 2.4%, mIoU by 4.2%, and Recall by 2.3% over TransUNet [2], and mDice by 1.4%, mIoU by 2.6%, and Recall by 1.4% over nnUNet [13]. On Polyps, our method outperforms nnUNet [13] by 1.2% in mDice and 1.8% in mIoU, SANet [33] by 2.6% in mDice and 3.5% in mIoU, and Polyp-PVT [4] by 1.1% in mDice and 1.3% in mIoU. Interestingly, in comparison to ArSDM [6] and ControlNet [35], we observe that there is no consistency between image quality and segmentation performance, indirectly highlighting that our method's superior mask-lesion alignment is key to improvements across diverse segmentation models.

## 5 Ablation Study

We conducted an ablation study to evaluate the importance of the Adaptive Distillation Loss ($\mathcal{L}_{\text{Ada}}$). Table 4 presents the results on KiTS19 [7]. The findings show that regularizing the student model with Distillation Loss (Standard) improves segmentation performance, while $\mathcal{L}_{\text{Ada}}$ (Adaptive) further enhances the baseline model's accuracy, highlighting its crucial role in mask-lesion alignment.

## 6 Conclusion

We present Adaptively Distilled ControlNet, a novel image synthesis method. During training, a teacher model with image-conditioned inputs adaptively regularizes the student model. During sampling, only the enhanced student model

is used, maintaining ControlNet's [35] sampling speed. We generate high-quality medical images with accurate mask-lesion alignment and rich morphological features using arbitrary masks. Extensive experiments across two modalities demonstrate the robustness, effectiveness, and superiority of our approach.

**Acknowledgments.** This work was supported in part by the Startup Grant for Professor (SGP) — CityU SGP, City University of Hong Kong under Grant 9380170.

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

## References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. COMPUT MED IMAG GRAP **43**, 99–111 (2015)
2. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis (2024)
3. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
4. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. CAAI AIR **2**, 9150015 (2021)
5. Dorjsembe, Z., Pao, H.K., Xiao, F.: Polyp-ddpm: Diffusion-based semantic polyp synthesis for enhanced segmentation. In: EMBC (2024)
6. Du, Y., Jiang, Y., Tan, S., Wu, X., Dou, Q., Li, Z., Li, G., Wan, X.: Arsdm: colonoscopy images synthesis with adaptive refinement semantic diffusion models. In: MICCAI (2023)
7. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
9. Hinton, G.: Distilling the knowledge in a neural network. In: NeurIPS Workshop (2015)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS Workshop (2022)
12. Hu, T., Zhang, J., Yi, R., Du, Y., Chen, X., Liu, L., Wang, Y., Wang, C.: Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In: AAAI (2024)
13. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods (2021)

14. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MMM (2020)
15. Li, M., Yang, T., Kuang, H., Wu, J., Wang, Z., Xiao, X., Chen, C.: Controlnet_plus_plus: Improving conditional controls with efficient consistency feedback. In: ECCV (2024)
16. Liu, S., Chen, Z., Yang, Q., Yu, W., Dong, D., Hu, J., Yuan, Y.: Polyp-gen: Realistic and diverse polyp image generation for endoscopic dataset expansion (2025)
17. Loshchilov, I.: Decoupled weight decay regularization. In: ICLR (2017)
18. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: CVPR (2023)
19. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: AAAI (2024)
20. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In: NeurIPS (2024)
21. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
22. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)
23. Shao, S., Yuan, X., Huang, Z., Qiu, Z., Wang, S., Zhou, K.: Diffuseexpand: Expanding dataset for 2d medical image segmentation using diffusion models. In: IJCAI Workshop (2024)
24. Sharma, V., Kumar, A., Jha, D., Bhuyan, M.K., Das, P.K., Bagci, U.: Controlpolypnet: towards controlled colon polyp synthesis for improved polyp segmentation. In: CVPR Workshop (2024)
25. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. INT J COMPUT ASS RAD **9**, 283–293 (2014)
26. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2020)
27. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: ICML (2023)
28. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2020)
29. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE TMI **35**(2), 630–644 (2015)
30. Thambawita, V., Salehi, P., Sheshkal, S.A., Hicks, S.A., Hammer, H.L., Parasa, S., Lange, T.d., Halvorsen, P., Riegler, M.A.: Singan-seg: Synthetic training data generation for medical image segmentation. PloS one **17**(5) (2022)
31. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: NeurIPS (2017)
32. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. J HEALTHC ENG **2017**(1), 4037190 (2017)
33. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S., Cui, S.: Shallow attention network for polyp segmentation. In: MICCAI (2021)
34. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In: ICCV (2023)

35. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)