STEALING 3D MEDICAL SEGMENTATION MODELS VIA COLLABORATIVE DUAL-MODEL ARCHITECTURE

Yantao Li[1], *Kaimeng Yang*[1], *Song Ruan*[1,2], *Zhenyu Yang*[1,3], *Shaojiang Deng*[1]

[1] College of Computer Science, Chongqing University
[2] Department of Computer Science and Software Engineering, University of Western Australia
[3] D3 Center, Osaka University

ABSTRACT

Machine Learning as a Service (MLaaS) facilitates the deployment and accessibility of medical models, yet concurrently exposes proprietary models to potential adversaries. Attackers may exploit model stealing attacks (MSAs) to replicate these models illicitly, leading to loss of training investment and privacy vulnerabilities. While existing research has mainly focused on MSAs in the context of 2D natural image classification, this work presents the first investigation into stealing 3D medical segmentation models. We introduce collaborative dual-model 3D medical segmentation stealing (CDMSS-3D), which decomposes the model stealing objective into two complementary aspects: stealing accuracy and stealing robustness. With our adversarial proxy training, CDMSS-3D achieves superior model stealing performance. Furthermore, we incorporate a dual-model discrepancy sampling strategy, which enhances the fidelity of the substitute model by prioritizing uncertain samples. Extensive experiments on four 3D medical segmentation datasets demonstrate that CDMSS-3D consistently outperforms adapted baselines.

***Index Terms*—** 3D Medical Segmentation, Model Stealing Attacks, Collaborative Dual-model Architecture

1. INTRODUCTION

Machine Learning as a Service (MLaaS) has relieved resource-constrained hospitals from intensive computational burdens by offering cloud-based APIs for critical medical tasks such as 3D anatomical modeling and data-driven treatment planning [1, 2]. However, this convenience comes with security risks, as attackers may exploit APIs to perform model stealing attacks (MSAs) and steal proprietary models [3]. A successful theft not only results in the loss of high training costs but can also enable further attacks, such as membership inference [4] or model inversion [5], which threaten privacy. Since 3D medical segmentation, a common medical task, requires large amounts of rare, high-quality, privacy-sensitive 3D medical volumes for training, the consequences of model theft would be particularly severe. Against this backdrop, a critical question arises: ***Can 3D medical segmentation models be stolen?***

Most existing MSAs focus on stealing 2D natural image classification models. With minor modifications, these attacks could be adapted to steal 3D models. However, our experiments show that their performance is unsatisfactory. For instance, ActiveThief's entropy-based sampling [6] tends to over-select large-foreground 3D volumes, thereby reducing data diversity. Similarly, synthetic adversarial examples generated by Black-box Dissector [7] dilute the proportion of clean data, further degrading the effectiveness of these methods under strict query limitations common in medical applications. Such failure stems from three critical discrepancies: (1) *large and diverse natural images vs. limited medical images*, which challenge the basic assumptions; (2) *classification vs. segmentation tasks*, leading to poor algorithmic adaptability; and (3) *2D vs. 3D data*, which exhibit significant variation in data complexity.

In this paper, we introduce Collaborative Dual-Model Steal for 3D Medical Segmentation (CDMSS-3D), the first MSA method tailored for 3D medical image segmentation to the best of our knowledge. In contrast to existing approaches that primarily exploit the relationship between data and the target model, we advance the MSA paradigm by rethinking substitute training: we decouple the two objectives of model theft—segmentation accuracy and robustness—into two models within a collaborative dual-model architecture. Our approach achieves superior performance over baselines on 3D medical segmentation tasks. **Our key contributions are summarized as follows**: (1) We are the first to investigate model stealing attacks (MSAs) on 3D medical image segmentation models. (2) We introduce a novel framework, CDMSS-3D, that decouples stealing objectives via a collaborative dual-model architecture. (3) Extensive experiments validate the effectiveness and superiority of our method, demonstrating that 3D medical segmentation models are vulnerable to theft, highlighting the need for greater attention to their security.

2. METHOD

2.1. Assumptions and Threat Model

The target model, denoted as f_{T}, refers to the victim model in MSAs. We consider a 3D medical segmentation target model deployed in an MLaaS setting. For any input $x \in \mathbb{R}^{c_{\mathrm{in}} \times w \times h \times d}$, f_{T} provides a feedback output $y = f_{\mathrm{T}}(x; \theta_T)$, where $y \in \mathbb{R}^{c_{\mathrm{out}} \times w \times h \times d}$, with c_{in} and c_{out} denoting the input and output channels respectively, and w, h, d denoting width, height, and depth of a 3D volume respectively. Each query to f_{T} incurs a certain cost, such as a monetary charge.

The attacker can query the f_{T} with data, but has no knowledge of its parameters θ_{T} or architecture. We assume the attacker possesses a seed set $\mathcal{X}$ to launch the attack, containing data similar to that used in training f_{T}. The attacker aims to train a substitute model f_{S} that approximates the functionality of f_{T} within a constrained query budget B. The stolen substitute model should achieve comparable segmentation accuracy and robustness, with its outputs exhibiting high-fidelity alignment to the responses of f_{T}:

$$\max_{f_{\mathrm{S}}} \mathbb{E}_{x \sim \mathcal{D}} \left[\mathcal{L}(f_{\mathrm{S}}(x), f_{\mathrm{T}}(x)) \right], \text{ s.t. } |\mathcal{D}| \leq B, \tag{1}$$

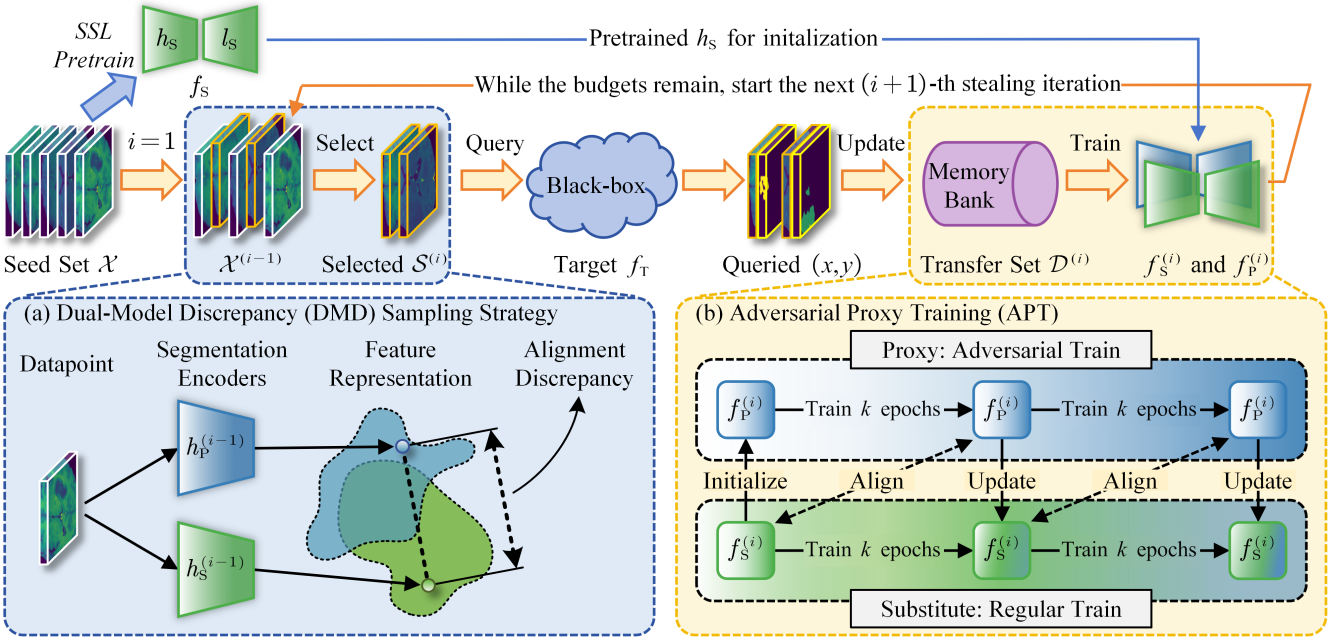

Fig. 1. Overview of CDMSS-3D. Starting from a self-supervised learning (SSL) pretrained model, we progressively refine the substitute model f_S iteratively until the query budget B is exhausted. (a) Illustration of the dual-model discrepancy (DMD) strategy used in the AL-driven sampling phase, where samples with high alignment discrepancy are selected. (b) Illustration of the adversarial proxy training (APT) used in the substitute training phase, where a proxy model trained with adversarial examples helps the substitute model to steal the target model.

where $\mathcal{D}$ denotes the transfer set constructed by the attacker, $\mathcal{L}$ quantifies the functional similarity between f_S and f_T.

2.2. Adversarial Proxy Training

Previous studies suggest that aligning the decision boundary of the substitute model f_S and target model f_T is key to successful MSA, and they focus on uncovering this boundary from the data perspective. However, in data-scarce medical domains, data options are limited, and synthetic samples may dilute the proportion of high-quality medical supervision, increasing the risk of overfitting [3, 8, 9] and potentially reducing theft success. We introduce adversarial proxy training (APT) to enhance substitute training phase through a dual-model architecture, under which a substitute model f_S focuses on capturing segmentation accuracy, while a proxy model f_P targets robustness. These models are trained collaboratively, with knowledge periodically fused to refine the overall stolen model, as illustrated in Fig. 1(b). By following three mechanisms, APT enables more effective MSAs:

(1) *Boundary Shaping*: f_S is trained exclusively on clean samples x and their corresponding pseudo-labels y queried from f_T, directly aligning decision boundary of f_S to f_T.

(2) *Robustness Exploration*: f_P is trained on dynamically generated [10] $x_{adv} = \text{AdvAttack}(x, f_P)$ [11], while retaining the original clean queried y, thereby expanding the decision boundary beyond the capability of f_S.

(3) *Knowledge Integration*: Every k epochs, inject robustness from f_P to f_S via $\theta_S = \alpha\theta_S + (1-\alpha)\theta_P$, where α is a weight parameter, we set $\alpha = 0.5$. Note that the parameter fusion is unidirectional, and the parameters of f_P remain unchanged. After each stealing iteration, f_P is reset to explore new robustness directions in subsequent iterations.

Further, we propose the Barlow Twins loss [12] to prevent semantic conflicts during parameter fusion:

$$\mathcal{L}_{\text{align}} \overset{\text{def}}{=} \frac{1}{n}\sum_{n}\left(\sum_{i}(1-C_{ii})^2 + \beta\sum_{i}\sum_{i\neq j}C_{ij}^2\right). \quad (2)$$

This loss is adapted to the segmentation task, where $n = w' \times h' \times d'$ denotes voxel count, with w', h', and d' denote the width, height, and depth of the volume features, respectively. C is the cross-correlation matrix between features $h_S(x)$ and $h_P(x_{adv})$, where $h_S : \mathbb{R}^{c_{in}\times w\times h\times d} \to \mathbb{R}^{c_{hidden}\times w'\times h'\times d'}$ is the encoder of f_S, with its decoder denoted as $l_S : \mathbb{R}^{c_{hidden}\times w'\times h'\times d'} \to \mathbb{R}^{c_{out}\times w\times h\times d}$, such that $f_S = l_S \circ h_S$ and similarly, $f_P = l_P \circ h_P$. β is a weight parameter; we set $\beta = 0.05$.

Final optimization objective integrates all components, including boundary shaping, robustness exploration, and Eq. (2)), i.e., representation alignment:

$$\min_{\theta_S,\theta_P} \underbrace{\mathcal{L}_{\text{seg}}(f_S(x;\theta_S), y)}_{\text{boundary shaping}} + \underbrace{\mathcal{L}_{\text{seg}}(f_P(x_{adv};\theta_P), y)}_{\text{robustness exploration}} + \lambda \underbrace{\mathcal{L}_{\text{align}}(h_S(x;\theta_S), h_P(x_{adv};\theta_P))}_{\text{representation alignment}}, \quad (3)$$

where $\lambda = 1$ is a trade-off weight. $\mathcal{L}_{\text{seg}}$ denotes the segmentation loss function, specifically, the Dice loss [13] in our approach.

2.3. Dual-Model Discrepancy Sampling

Previous studies have demonstrated the effectiveness of AL-driven sampling [14], which we retain in our framework. However, in-

Seed Set	Method	UNETR				MedNeXt-Base			
		ET	TC	WT	Mean	ET	TC	WT	Mean
Full overlap	Random	79.50 (90.47)	84.21 (90.80)	89.47 (94.61)	84.39 (91.96)	79.19 (84.14)	83.66 (85.48)	89.10 (91.47)	83.98 (87.03)
	JBDA	76.58 (85.46)	79.17 (84.23)	86.87 (91.25)	80.87 (86.98)	74.98 (77.89)	77.23 (78.46)	86.33 (88.28)	79.51 (81.54)
	ActiveThief	80.07 (**91.39**)	85.07 (**91.85**)	89.64 (94.84)	84.92 (**92.69**)	80.60 (85.22)	84.99 (86.86)	89.76 (92.13)	85.12 (88.07)
	Black-box Dissector	79.44 (89.95)	83.94 (90.23)	88.94 (94.08)	84.11 (91.42)	77.79 (81.22)	81.56 (82.97)	87.75 (89.85)	82.36 (84.68)
	CDMSS-3D (Ours)	**80.43** (91.06)	**86.07** (91.79)	**90.17** (**94.84**)	**85.56** (92.56)	**81.40** (**87.22**)	**87.03** (**89.16**)	**90.87** (**93.50**)	**86.43** (**89.96**)
Partial overlap	Random	77.00 (86.15)	80.29 (85.86)	87.83 (92.41)	81.71 (88.14)	76.23 (80.07)	80.00 (81.27)	86.74 (88.97)	80.99 (83.44)
	JBDA	75.18 (82.37)	75.97 (80.10)	84.59 (88.79)	78.58 (83.76)	72.59 (74.92)	73.61 (74.47)	83.32 (85.07)	76.51 (78.16)
	ActiveThief	78.07 (88.46)	82.52 (88.58)	88.58 (93.27)	83.06 (90.10)	76.63 (80.87)	79.86 (81.32)	87.22 (89.34)	81.24 (83.84)
	Black-box Dissector	77.17 (85.85)	80.26 (85.62)	87.18 (91.71)	81.54 (87.73)	73.94 (77.11)	77.13 (78.19)	86.04 (88.05)	79.04 (81.11)
	CDMSS-3D (Ours)	**78.89** (**88.98**)	**83.96** (**89.43**)	**89.25** (**93.69**)	**84.03** (**90.70**)	**78.04** (**82.49**)	**83.17** (**84.99**)	**89.08** (**91.51**)	**83.43** (**86.33**)
Disjoint	Random	77.91 (88.09)	81.37 (87.46)	88.19 (93.02)	82.49 (89.52)	76.69 (81.88)	80.72 (82.33)	87.52 (89.83)	81.64 (84.68)
	JBDA	74.83 (82.45)	74.72 (78.88)	84.96 (89.04)	78.17 (83.46)	67.95 (70.16)	66.25 (67.26)	82.00 (84.07)	72.07 (73.83)
	ActiveThief	77.39 (87.58)	81.03 (86.60)	88.14 (92.80)	82.19 (88.99)	75.96 (80.54)	79.51 (81.06)	87.46 (89.75)	80.98 (83.78)
	Black-box Dissector	76.04 (85.11)	79.50 (84.72)	87.42 (91.84)	80.99 (87.22)	73.52 (76.56)	76.01 (77.24)	85.55 (87.68)	78.36 (80.49)
	CDMSS-3D (Ours)	**78.88** (**89.59**)	**84.08** (**89.77**)	**89.29** (**93.97**)	**84.08** (**91.11**)	**77.21** (**83.08**)	**82.40** (**84.53**)	**88.67** (**91.11**)	**82.76** (**86.24**)

Table 1. The comparison of Clean Segmentation Accuracy CA (with corresponding Fidelity) (%) between our proposed method and baseline approaches. The ET, TC, and WT represent the enhancing tumor, tumor core, and whole tumor sub-regions, respectively.

stead of directly adopting techniques designed for 2D classification tasks—many of which perform poorly even compared to random sampling in 3D medical segmentation—we base our AL-driven sampling on dual-model discrepancy (DMD), as shown in Fig. 1(a). This stems from a straightforward insight: the feature divergence between two models directly reflects data uncertainty [15], which typically lies around the decision boundary [8].

Specifically, at the end of i-th stealing iteration, both the substitute model $f_{\mathrm{S}}^{(i)} = l_{\mathrm{S}}^{(i)} \circ h_{\mathrm{S}}^{(i)}$ and the proxy model $f_{\mathrm{P}}^{(i)} = l_{\mathrm{P}}^{(i)} \circ h_{\mathrm{P}}^{(i)}$ are trained on current transfer set $\mathcal{D}^{(i)}$, where h and l represent the encoder and decoder modules, respectively. For each sample $x \in \mathcal{X}^{(i)}$, we compute the feature distance discrepancy between $h_{\mathrm{S}}^{(i)}(x)$ and $h_{\mathrm{P}}^{(i)}(x)$ using the alignment loss $\mathcal{L}_{\mathrm{align}}$ mentioned above. Then, we sample a subset $\mathcal{S}^{(i+1)} \subset \mathcal{X}^{(i)}$ containing $b^{(i+1)}$ samples, where $b^{(i+1)}$ is determined by the labeling budget B:

$$\begin{aligned} &\underset{\mathcal{S}^{(i+1)} \subset \mathcal{X}^{(i)}}{\arg\max} \ \mathbb{E}_{x \in \mathcal{S}^{(i+1)}} \left[\mathcal{L}_{\mathrm{align}} \left(h_{\mathrm{S}}^{(i)}(x), h_{\mathrm{P}}^{(i)}(x) \right) \right], \\ &\text{s.t. } |\mathcal{S}^{(i+1)}| = b^{(i+1)}. \end{aligned} \tag{4}$$

2.4. Pretraining for Representation Enhancement

Knowledge integration in our dual-model architecture involves fusing feature representations, which are initially weak and insufficient during early training. Existing studies indicate that rich and diverse features enhance substitute training [14], and pretraining can provide such informative features [3]. Our approach is likely to benefit more from pretraining. Inspired by this, we introduce a self-supervised pretraining phase using the full seed set $\mathcal{X}$ via the masked autoencoder [16] before launching attack queries. For each sample, 75% of the 3D patches are randomly masked, and the model is trained to reconstruct them.

3. EXPERIMENTS AND DISCUSSIONS

3.1. Datasets and Experimental Setup

We conducted experiments on four public 3D brain tumor segmentation datasets: BraTS2018, BraTS2019, BraTS2020, and BraTS2021 [17, 18, 19]. The experimental settings adopt the configuration detailed in Table 3, following [20, 7].

Seed Set	Method	UNETR		MedNeXt-Base	
		RA(Fidelity)	ASR	RA(Fidelity)	ASR
Full overlap	Random	60.49 (75.37)	40.15	89.47 (94.61)	12.19
	JBDA	59.62 (70.46)	27.98	58.22 (61.63)	10.46
	ActiveThief	60.77 (74.23)	36.40	60.04 (63.88)	11.59
	Black-box Dissector	63.11 (76.96)	35.09	62.19 (66.62)	13.92
	CDMSS-3D (Ours)	**71.96** (**87.22**)	**40.87**	**72.61** (**79.06**)	**15.37**
Partial overlap	Random	57.97 (70.24)	33.67	57.20 (60.63)	9.77
	JBDA	58.28 (67.61)	25.81	52.76 (55.37)	8.31
	ActiveThief	60.78 (73.44)	31.17	57.23 (60.50)	8.79
	Black-box Dissector	60.44 (72.34)	31.47	57.46 (60.74)	9.36
	CDMSS-3D (Ours)	**69.95** (**83.96**)	**36.17**	**68.23** (**73.68**)	**12.65**
Disjoint	Random	61.85 (75.52)	36.34	58.65 (62.31)	9.01
	JBDA	57.39 (67.26)	25.69	47.30 (49.03)	6.71
	ActiveThief	60.67 (72.78)	30.80	59.19 (62.96)	10.29
	Black-box Dissector	62.19 (74.17)	30.77	56.88 (60.17)	9.32
	CDMSS-3D (Ours)	**71.16** (**85.11**)	**36.91**	**67.01** (**72.62**)	**11.95**

Table 2. The comparison of Robust Accuracy (with corresponding Fidelity) (%) and ASR (%) under SegPGD3, evaluated between our proposed method and the baseline approaches.

We considered two target model architectures: UNETR [21] and MedNeXt-Base [22], and they were well-trained on BraTS2020 and BraTS2021 using adversarial training. The substitute model was chosen as UNETR, thereby covering both cases with the same and different architectures.

For baselines, data-free approaches [9, 23] were excluded due to the complexity of generating 3D medical segmentation data. The baselines included three representative approaches: (1) *JBDA [24]*, (2) *ActiveThief [6]*, and (3) *Black-box Dissector [7]*. All methods were adapted to support the segmentation tasks. Additionally, we included a naive baseline, (4) *Random*, which randomly samples the training data to train the substitute model.

We evaluated the substitute model in terms of segmentation accuracy, robustness, and attack transferability using the following four metrics: (1) *Clean Accuracy (CA)* measures segmentation accuracy on clean inputs using the Dice score: $\mathrm{CA} = \mathrm{Dice}(f_{\mathrm{S}}(x), \mathrm{GT})$, where GT denotes the ground-truth segmentation label of clean

Seed Set	Defination	Overlap	Query Sequence
Full-overlap	BraTS20 ∪ BraTS21	100%	{100, 200, 300, 400, 500}
Partial-overlap	BraTS19 ∪ BraTS20	22.76%	{40, 80, 120, 160, 200}
Disjoint	BraTS18 ∪ BraTS19	0%	{40, 80, 120, 160, 200}

Table 3. Seed set configuration. Overlap denotes the portion of data that is shared with the target model's training set.

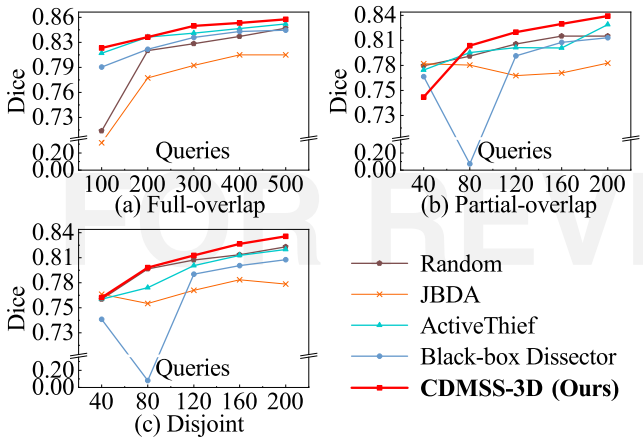

Fig. 2. The mean CA of our method and baseline approaches, with the number of queries consumed, over three seed set scenarios, when attacking UNETR.

inputs. (2) *Robust Accuracy (RA)* measures robustness against adversarial inputs: $\mathrm{RA} = \mathrm{Dice}(f_{\mathrm{S}}(x_{adv}), \mathrm{GT})$. (3) *Fidelity* quantifies the prediction similarity between f_{S} and f_{T} on the same inputs x:$\mathrm{Fidelity} = \mathrm{Dice}(f_{\mathrm{S}}(x), f_{\mathrm{T}}(x))$. (4) *Attack Success Rate (ASR)* is defined as the relative performance degradation of f_{T} when attacked by transferred adversarial examples generated by f_{S}:

$$\mathrm{ASR} = \frac{\mathrm{Dice}\left(f_{\mathrm{T}}(x), \mathrm{GT}\right) - \mathrm{Dice}\left(f_{\mathrm{T}}(x_{\mathrm{adv}}^{f_{\mathrm{S}}}), \mathrm{GT}\right)}{\mathrm{Dice}\left(f_{\mathrm{T}}(x), \mathrm{GT}\right) - \mathrm{Dice}\left(f_{\mathrm{T}}(x_{\mathrm{adv}}^{f_{\mathrm{T}}}), \mathrm{GT}\right)}, \quad (5)$$

where $x_{\mathrm{adv}}^{f_{\mathrm{S}}} = \mathrm{AdvAttack}(x; f_{\mathrm{S}})$ and $x_{\mathrm{adv}}^{f_{\mathrm{T}}}$ denotes the white-box adversarial example.

3.2. Comparison with Baselines

Table 1 shows the CA and corresponding fidelity. Our method consistently outperforms the baselines in CA. In terms of fidelity, it performs better in most cases, except slightly underperforming ActiveThief when attacking UNETR under full-overlap configuration. Meanwhile, it can be observed that as the amount of data and overlapping samples decreases, some baseline methods gradually degrade, even performing worse than Random baseline. In contrast, our method consistently outperforms the Random baseline. Table 2 shows a similar conclusion: our method outperforms others in RA and attack transferability in most cases. The above comparison confirms the effectiveness and advantages of our method. By decoupling the accuracy and robustness in model stealing through the dual-model architecture, we achieve better substitute model performance. This advantage also extends to efficiency: as shown in Fig. 2, our method consistently achieves better performance under the same query budget.

MAE	DMD	APT	Partial-overlap		
			CA	RA	ASR
✗	✗	✗	81.67	60.31	31.92
✗	✗	✓	80.30	67.79	35.77
✗	✓	✓	78.69	54.18	36.00
✓	✗	✗	83.43	58.11	27.17
✓	✗	✓	83.41	69.87	34.57
✓	✓	✓	**84.03**	**69.95**	**36.17**

Table 4. Ablation study of the proposed components, evaluated under the partial-overlap seed set scenario when attacking UNETR. Segmentation accuracy, as well as robust accuracy and ASR under SegPGD3, are reported, with the best values highlighted in bold.

k	CA	SegPGD3		SegPGD8	
		RA	ASR	RA	ASR
5	82.88	68.85	32.04	53.51	38.19
10	83.43	69.71	33.47	53.92	40.36
20	83.72	**70.20**	33.73	**54.58**	40.41
25	84.03	69.95	36.17	54.50	**43.73**
30	**84.28**	67.66	**36.36**	49.03	42.68

Table 5. Sensitivity of fusion interval k, evaluated under the partial-overlap seed set scenario when attacking UNETR. Segmentation Accuracy and corresponding Fidelity are reported and the best values are highlighted in bold.

3.3. Ablation Study

When no modules are enabled, the ablation baseline uses random sampling combined with standard adversarial training. The DMD sampling strategy relies on the dual-model setup, so DMD cannot be used without enabling APT. Table 4 shows that enabling APT substantially improves RA and ASR. However, combining APT with the DMD sampling strategy without MAE pre-training reduces performance, which aligns with our previous finding that feature alignment is important for DMD sampling. Enabling MAE, APT, and DMD together achieves the best performance.

The fusion interval k is a critical hyperparameter in our dual-model architecture. As shown in Table 5, the performance improves as k increases up to 20. At $k = 25$, ASR shows a significant jump compared to $k = 20$. Further increasing k beyond 25 results in an increase in CA but a degradation in RA. Overall, $k = 25$ achieves the best trade-off and is selected for our experiments.

4. CONCLUSION

This paper presents the first systematic study of model extraction attacks against 3D medical image segmentation models, demonstrating that high-fidelity model stealing is achievable even under stringent query budget constraints. Unlike prior work focused on exposing decision boundaries with data, our method, CDMSS-3D, a novel collaborative dual-model architecture based on adversarial proxy training, enables the simultaneous steal of both segmentation accuracy and robustness from the target. The proposed dual-model discrepancy sampling and pretraining phase further enhances attack performance. Experiments demonstrate the effectiveness and superiority of our approach.

5. REFERENCES

[1] Mauro Ribeiro, Katarina Grolinger, and Miriam A.M. Capretz, "Mlaas: Machine learning as a service," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015, pp. 896–902.

[2] Ahmad P Tafti, Eric LaRose, Jonathan C Badger, Ross Kleiman, and Peggy Peissig, "Machine learning-as-a-service and its application to medical informatics," in *International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2017, pp. 206–219.

[3] Jiacheng Liang, Ren Pang, Changjiang Li, and Ting Wang, "Model extraction attacks revisited," in *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, 2024, pp. 1231–1245.

[4] Lakshmi Prasanna Pedarla, Xinyue Zhang, Liang Zhao, and Hafiz Khan, "Evaluation of query-based membership inference attack on the medical data," in *Proceedings of the 2023 ACM Southeast Conference*, 2023, pp. 191–195.

[5] Shagufta Mehnaz, Sayanton V Dibbo, Ehsanul Kabir, Ninghui Li, and Elisa Bertino, "Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4579–4596.

[6] Soham Pal, Yash Gupta, Aditya Shukla, Aditya Kanade, Shirish Shevade, and Vinod Ganapathy, "Activethief: Model extraction using active learning and unannotated public data," in *Proceedings of the AAAI conference on artificial intelligence*, 2020.

[7] Yixu Wang, Jie Li, Hong Liu, Yan Wang, Yongjian Wu, Feiyue Huang, and Rongrong Ji, "Black-box dissector: Towards erasing-based hard-label model stealing attack," in *European conference on computer vision*, 2022.

[8] Guanlin Li, Guowen Xu, Shangwei Guo, Han Qiu, Jiwei Li, and Tianwei Zhang, "Extracting robust models with uncertain examples," in *The Eleventh International Conference on Learning Representations*, 2023.

[9] Xiaojian Yuan, Kejiang Chen, Wen Huang, Jie Zhang, Weiming Zhang, and Nenghai Yu, "Data-free hard-label robustness stealing attack," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 6853–6861.

[10] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[11] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr, "Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness," in *European Conference on Computer Vision*, 2022.

[12] Dong Kyu Cho, Inwoo Hwang, and Sanghack Lee, "Peer pressure: Model-to-model regularization for single source domain generalization," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 15360–15370.

[13] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[14] Lijun Gao, Wenjun Liu, Kai Liu, and Jiehong Wu, "Augsteal: Advancing model steal with data augmentation in active learning frameworks," *IEEE Transactions on Information Forensics and Security*, 2024.

[15] Han Liu, Hao Li, Xing Yao, Yubo Fan, Dewei Hu, Benoit M Dawant, Vishwesh Nath, Zhoubing Xu, and Ipek Oguz, "Colossal: A benchmark for cold-start active learning for 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 25–34.

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16000–16009.

[17] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[18] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.

[19] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al., "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[20] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz, "Knockoff nets: Stealing functionality of black-box models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4954–4963.

[21] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[22] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein, "Mednext: transformer-driven scaling of convnets for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023.

[23] Mingwen Shao, Lingzhuang Meng, Yuanjian Qiao, Lixu Zhang, and Wangmeng Zuo, "Latent code augmentation based on stable diffusion for data-free substitute attacks," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[24] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017.