Siamese Diffusion と Segmentation Network の敵対的学習による腫瘍セグメンテーション精度向上

尾﨑 海里[1,a)]　堀田 一弘[1,b)]

概要

計算機支援診断（CAD）システムは，大腸内視鏡検査におけるポリープの見落とし低減にとってますます不可欠なものとなっている．近年のセグメンテーションモデルは画素レベルでの病変局在化を可能とするが，その性能は高品質なアノテーション付き医用データセットの不足によって根本的に制約されている．この課題に対し GAN や拡散モデルを用いたデータ拡張が検討されてきたが，これらはセグメンテーション精度を有意に改善する困難なサンプルではなく，典型的な症例を再現する傾向がある．本研究では，追加の大規模データを必要とすることなく生成サンプルの学習価値を最大化する敵対的フィードバックを提案する．提案手法は Noise-Consistent Siamese Diffusion（NCSD）を生成器として用い，凍結されたセグメンテーションネットワークによってセグメンテーション誤差に基づくフィードバックを与える．セグメンテーション誤差を生成過程へ逆伝播させることにより，曖昧な境界や微小病変といったタスク駆動型の困難例の合成を能動的に促進する．セグメンテーションの学習においては，生成された困難例を実画像と組み合わせて拡張学習セットを構成する．この戦略により，mIoU は SANet で 4.71%，Polyp-PVT で 1.50%，CTNet で 3.59%改善し，複数のアーキテクチャにわたって一貫した性能向上を示した．さらに従来の NCSD に基づく学習と比較しても，提案手法は追加的な性能改善を達成した．SANet を生成画像のみで学習させた場合においても，実画像のみによる学習に対して mIoU で 0.37%の改善を確認し，生成された困難例の有効性を実証した．

1. はじめに

大腸がんは世界的にがん関連死亡の主要な原因の一つであり続けており，早期発見を不可欠なものとしている．大腸内視鏡検査中に前がん性ポリープを同定することは，生存率を有意に改善する．しかしながら，経験豊富な臨床医であっても，ポリープの最大 20%が見落とされる可能性がある [1]．この限界が，正確な病変セグメンテーションのための深層学習に基づく CAD システムの開発を動機づけている．

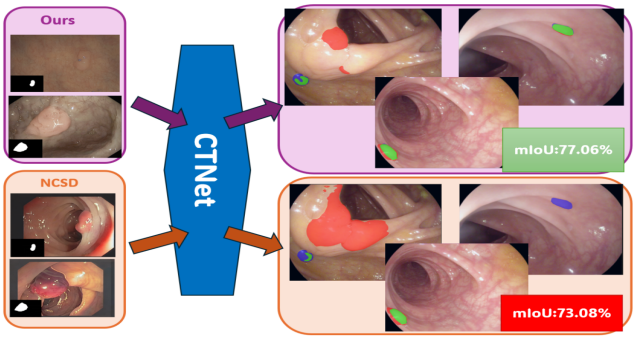

図 1: 合成画像による CTNet のセグメンテーション精度比較．色は TP（緑），FP（赤），FN（青）を示す．NCSD（橙）：腫瘍と背景のコントラストが明白．提案手法（紫）：境界が曖昧な難例（Hard examples）を生成し，FP および FN を抑制．

CAD に求められるのは，画像にポリープが写っているかを判定する分類ではなく，どの画素がポリープに対応するかを一画素ずつ判定するセマンティックセグメンテーションである．医師に病変を提示する以上，位置と範囲の情報が不可欠だからである．しかしこのタスクの精度は学習に用いる教師データに強く依存し，医用画像分野には教師データの確保を阻む二つの障壁がある．

第一は量の不足である．セグメンテーションの教師データを作るには，医師が画像一枚ごとに病変の輪郭を画素単位で描く必要があり，大変な時間とコストを要する．加えて医療画像は患者の個人情報そのものであり，収集と共有には倫理審査と厳格な管理が求められ，施設外への持ち出しが制度上不可能な場合も少なくない [2]．結果として，公開されている大腸内視鏡のセグメンテーション用データセットは，多くが数百枚から千数百枚の規模にとどまる．

第二は，より深刻でありながら見過ごされやすい偏りの問題である．データセットとして整備される画像は，撮影条件が良好で病変が明瞭に写っているものに寄る．襞に隠

[1] 名城大学
[a)] 220442030@ccalumni.meijo-u.ac.jp
[b)] kazuhotta@meijo-u.ac.jp

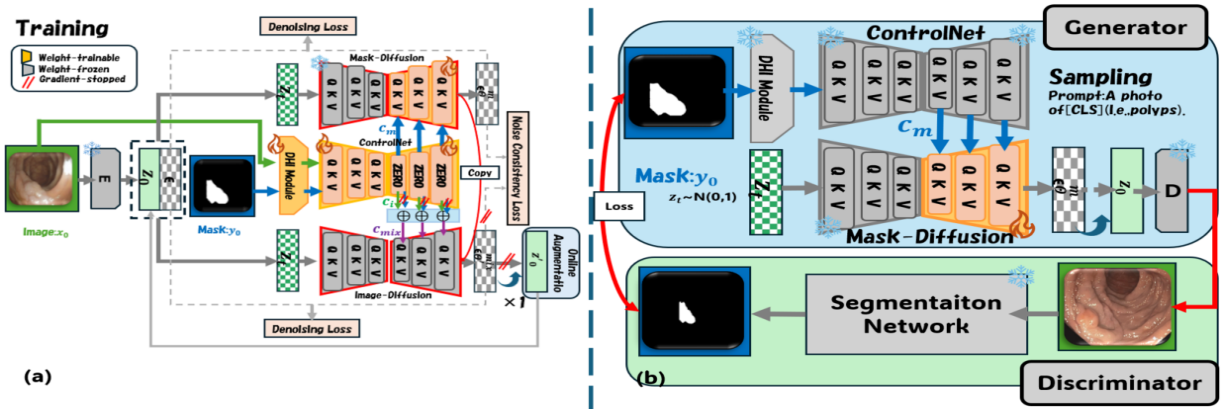

図 2: 提案手法の全体構造．(a) NCSD の学習構造：NCSD の良さを最大限活用するため NCSD の学習構造を採用する．(b) 提案手法: 推論時には，事前学習済みの Mask-Diffusion を生成器として用いる．生成画像に対して SANet からのフィードバックを導入することにより，セグメンテーション性能向上に寄与する困難例を生成する．

れた病変，反射に埋もれた病変，境界が判別できない扁平病変といった，臨床で実際に見落とされている種類の症例が，データセットの中で最も少なくなる．最も学習させたい症例が最も少ないという構造が，ベンチマーク上で高い平均精度を示すモデルが臨床現場で期待どおりに働かない理由の一端をなしている．

データ不足に対しては，生成モデルで合成画像を作り出して学習データを水増しするアプローチが研究されてきた．GAN [3] に始まり，近年では拡散モデル [4], [5], [6] が主流である．セグメンテーション向けには，病変マスクを条件として与える方式が有効である．マスクが生成条件であると同時に正解ラベルにもなり，画像とラベルの対を同時に得られるからである．SinGAN-Seg [7] や ArSDM [8] がこの系譜にあり，中でも NCSD [9] は，限られたデータからでも解剖学的に妥当かつ多様な画像生成を可能とする．

ここに，見過ごされてきた前提がある．本物らしく多様な画像を増やせば下流タスクの性能は上がる，という前提である．しかし生成モデルは学習データの分布を忠実に再現するよう最適化されている以上，その偏りもまた忠実に再現する．元のデータセットが明瞭な典型例に偏っているのだから，優秀な生成モデルほど典型例を大量に生産する．実際，図 1 に示すように，既存手法の生成画像はポリープと背景のコントラストが明瞭である．生成モデルとしては正しい振る舞いだが，データ拡張として見れば，既に解ける種類のサンプルを増やしているにすぎない．合成データの枚数を増やしても性能が頭打ちになる現象は，ここから説明できる．すなわち医用画像の課題はデータが足りないことではなく，データセットが覆う難易度の分布が臨床で問題になる側に届いていないことにある．

そこで本研究は，生成の目的関数を本物らしさから下流モデルの誤差へ置き換える．実画像のみで学習し重みを凍結したセグメンテーションネットワーク（SANet [10]）で生成画像を評価し，正解マスクとの誤差を最大化する方向に勾配を生成器へ逆伝播させる．凍結ネットワークは実画像で学習されているため，その苦手領域は実際のモデルの弱点を反映している．したがって生成される困難例は，恣意的に難しくした画像ではなくタスクにとって意味のある困難例となる．ただし誤差を最大化するだけでは画像は破綻する．病変を消し去れば誤差は最大になるが，それはマスクに対応する画像ではなく単なるラベルノイズである．この逸脱を防ぐため，更新対象をデコーダに限定して大域的な解剖構造を保持し，t-vMF Dice 損失 [11] で構造を，LPIPS [12] で視覚的自然さを拘束し，さらに識別器を凍結して更新しない．識別器を固定することで難化の目標が動かず，生成器と識別器が共進化して極端化する不安定性が構造的に回避される．

3 種類の構造が異なるセグメンテーションネットワークと 5 つの公開データセットで評価した結果，実画像のみのベースラインと比較して，SANet で mDice +4.28%，mIoU +4.71%，Polyp-PVT [13] で mDice +1.28%，mIoU +1.50%，CTNet [14] で mDice +3.28%，mIoU +3.59%の向上を得た．既存手法 NCSD と比較しても最大で mDice +1.69%，mIoU +2.10%の追加的な向上を達成した．

注目すべきは，改善がどこで生じたかである．5 つの評価データセットのうち ColonDB と ETIS は学習に一切用いておらず，撮影機器も施設も異なる分布外データである．改善はこの 2 つに集中し，CTNet では NCSD と比較して約 4%に達した．一方，学習画像の供給元である ClinicDB での改善は限定的であった．学習データと同じ分布のデータではモデルが既に十分な性能に達しており，追加の難例が与える情報は限界的だからである．この偏りは実応用の観点からむしろ望ましい．実運用で問題となるのは開発時に用いたデータでの性能ではなく，導入先の施設や機器で撮影されたデータでの性能だからである．さらに，実画像を 1 枚も用いず合成画像のみで学習した場合でも，実画像のみの学習を上回る mDice +0.44%，mIoU +0.37%を達成

表 1: SANet，Polyp-PVT，および CTNet の性能比較．実画像のみで学習した各セグメンテーションネットワークおよび「+NCSD」に関するベースラインとその結果は，NVIDIA RTX 6000 Ada を用いたローカル環境において再評価した．「+」は実画像と合成画像を併用した学習を示し，「M」および「M+I」はそれぞれマスクのみ，およびマスクと画像を併用した事前条件を表す．

Methods	EndoScene		ClinicDB		Kvasir		ColonDB		ETIS		Overall	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
SANet	89.11	82.10	91.78	86.04	90.42	84.50	73.99	65.68	72.75	63.88	78.30	70.40
+SinGAN-Seg(M)[7]	88.30	81.60	90.90	85.30	91.00	85.80	77.30	69.40	73.70	65.40	80.00	72.60
+T2I-Adapter(M)[15]	86.50	78.90	90.60	85.00	89.90	84.10	77.70	69.40	78.60	69.70	81.10	73.20
+ArSDM(M)[8]	90.20	83.20	91.40	86.10	91.10	85.60	77.70	70.00	78.00	69.50	81.50	74.10
+ControlNet(M)[16]	90.20	83.70	91.60	85.40	90.60	85.00	77.00	69.00	75.80	66.50	80.50	72.80
+ControlNet(M+I)	89.90	83.30	91.40	86.20	91.20	85.50	78.20	70.50	75.40	66.60	81.00	73.60
+NCSD(M)	88.52	81.09	93.01	87.50	90.34	85.00	77.80	69.95	78.40	69.89	81.52	74.04
+Ours(M)	**90.51**	**83.97**	**93.45**	**88.40**	**91.48**	**86.82**	**79.38**	**71.06**	**78.48**	**70.08**	**82.58**	**75.11**
Polyp-PVT	90.07	83.20	92.32	87.81	91.84	86.92	80.32	72.30	79.35	71.38	83.19	75.93
+SinGAN-Seg(M)	87.00	79.70	91.70	87.00	**92.80**	**88.10**	76.90	69.00	74.20	66.70	80.10	73.00
+T2I-Adapter(M)	88.70	81.60	93.10	88.10	91.30	86.00	80.60	72.40	79.00	71.00	83.10	75.70
+ArSDM(M)	88.20	81.20	92.20	87.50	91.50	86.30	81.70	73.80	80.60	72.90	84.00	76.70
+ControlNet (M)	88.00	81.10	93.70	89.00	91.50	86.30	80.80	72.50	76.10	68.10	82.50	75.10
+ControlNet (M+I)	90.30	83.50	92.20	87.70	91.10	86.20	81.60	73.70	77.20	69.10	83.20	76.00
+NCSD(M)	89.82	81.95	**93.76**	**88.71**	91.35	86.32	80.46	72.63	**81.07**	73.28	83.64	76.47
+Ours(M)	**90.40**	**83.80**	93.65	88.66	92.15	87.21	**82.08**	**74.37**	80.90	**73.31**	**84.47**	**77.43**
CTNet	90.76	84.20	93.00	88.16	91.60	86.31	80.57	72.83	78.25	70.90	83.10	76.10
+SinGAN-Seg(M)	89.40	82.30	93.60	88.80	**92.90**	**88.60**	78.00	70.40	76.40	68.70	81.50	74.60
+T2I-Adapter(M)	90.80	84.30	93.70	88.70	91.90	86.90	81.70	74.10	80.40	72.50	84.30	77.20
+ArSDM(M)	89.80	82.90	92.90	87.80	90.50	85.20	81.90	74.00	80.90	73.50	84.20	77.00
+ControlNet (M)	90.20	84.10	92.30	87.50	91.70	86.90	81.40	73.50	79.50	71.70	83.70	76.60
+ControlNet (M+I)	90.70	84.20	93.20	88.60	92.00	87.20	81.10	73.60	80.90	73.10	84.10	77.10
+NCSD(M)	90.80	83.95	**93.84**	**88.80**	92.22	87.62	82.17	74.60	80.97	73.08	84.69	77.59
+Ours(M)	**91.64**	**85.08**	92.02	86.58	92.40	86.58	**84.18**	**77.26**	**84.16**	**77.06**	**86.38**	**79.69**

した．患者データを施設外へ持ち出せない環境でのモデル構築に道を開く結果である．

本研究の貢献は以下の三点である．

- **問題の再定式化**：優秀な生成モデルほど元データの偏りを再生産するという逆説を示し，課題をデータ量の不足ではなく難易度分布の偏りとして捉え直した．
- **手法の提案**：凍結セグメンテーションネットワークの予測誤差を生成器へ還元し，タスク依存の困難例を能動的に合成する枠組みを提案した．デコーダ限定更新，構造・知覚正則化，識別器凍結により，医学的妥当性を保ちながら難易度を連続的に制御可能とした．
- **実証と知見**：5 つのデータセットと 3 種のアーキテクチャで性能向上を確認し，特に分布外データで顕著な改善を得た．さらに，難化した画像は単独では性能を大きく低下させ，通常の生成画像と混合したときに初めて利得を生むことを見出した．有効なのは難易度を上げることではなく，難易度の幅を広げることである．

2. 提案手法

2.1 全体構成

提案手法は二段階構成をとる．第 1 段階では，実画像のみを用いて生成器を通常の手順で学習し，第 2 段階で，凍結したセグメンテーションネットワークからの誤差フィードバックによって生成器を洗練する．段階を分けるのは，いきなり難化を目的とした最適化を行うと生成器が解剖構造そのものを学習し損ねるためである．基本構造の獲得と難化を分離することで，第 2 段階を既に妥当な画像をどの方向に難しくするかという限定された問題に還元できる．

図 2(a) に示す第 1 段階では，Stable Diffusion v1.5 を初期値とし，原論文に準拠したマスク条件付き拡散学習により NCSD を学習する．NCSD は，マスクのみを条件とする Mask-Diffusion と，画像とマスクの双方を条件とする Image-Diffusion からなる重み共有のシャム構造をとり，通常のノイズ除去損失に加えて両分岐のノイズ予測の一貫性を課すことで最適化される．この段階で，多様性と解剖学的一貫性を備えた生成事前分布が構築される．図 2(b) に示す第 2 段階では，学習済み Mask-Diffusion を生成器として用い，生成画像を凍結された SANet で評価して，その予測誤差を生成器へ還元する．

2.2 誤差フィードバックによる難化

生成はマスク条件付き逆拡散として行う．標準正規分布からサンプリングした潜在変数を出発点とし，各時刻で Mask-Diffusion がマスク条件のもとでノイズ成分を予測して潜在表現を更新する．これを反復して得た潜在符号を VAE デコーダで画像空間へ復号する．

生成された画像を凍結 SANet で評価し，正解マスクとの誤差を生成器へ逆伝播させる．ここで更新対象は，Mask-Diffusion のデコーダ（出力ブロックおよび最終層）に限定する．これは生成後の後処理的な更新であり，DDIM 軌道全体を通した不安定かつ高コストな逆伝播を回避する．符号化器と条件付け機構が凍結されているため大域的な解剖構造は厳密に保存され，デコーダのみの更新は境界や質感といった局所的な変動をもたらす．

セグメンテーションネットワークは凍結し，生成器のみを更新する．両者を同時に更新すると，生成器はより極端な改変へ，識別器はそれに適応へ，というエスカレーションが生じ，生成画像は医学的妥当性を失う．識別器を固定することで難化の目標が動かず，最適化は医学的に妥当な画像空間の内部に留まる．加えて，学習率を 10^{-7} と極めて小さく設定し，Dice スコアが 0.1 未満のサンプルについては勾配伝播を行わない．Dice が極端に低いサンプルは難例というより既に破綻した画像であり，これを除外することで意味のある学習信号のみが生成器に影響する．

2.3 損失関数設計

生成器は，解剖学的に妥当でありながらセグメンテーションにとって困難な画像を生成するよう学習される．難化と構造安定性を両立させるため，総損失を次式で定義する．

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \lambda_{tvMF}\mathcal{L}_{tvMF} + \lambda_{lpips}\mathcal{L}_{lpips}. \qquad (1)$$

難化損失は腫瘍領域のみに適用する．背景を含めて誤差を最大化すると，生成器は病変とは無関係な背景の改変に

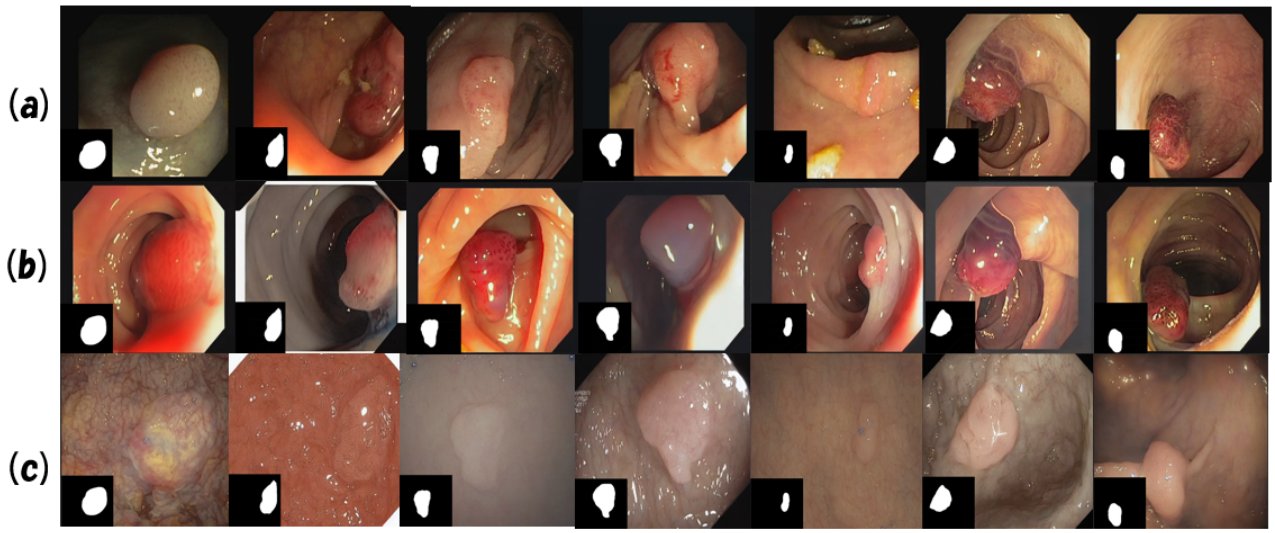

図 3: ポリープ画像の視覚的比較を示す．(a) 実画像のポリープ．(b) NCSD によって生成された合成画像．(c) 提案手法によって生成された合成画像．すべての合成画像は，マスク画像のみを条件とする同一条件下で生成されている．

表 2: 実画像および合成画像で学習した SANet の性能比較を示す．“Real” は，元の 1,450 枚の実画像のみを用いて学習した結果を示す．“Generated” は，実画像を一切使用せず，合計 4,350 枚の合成画像（NCSD により生成された 1,450 枚と，提案手法により難易度強化された 2,900 枚）を用いて学習した結果を示す．各評価指標における最良の結果は太字で示している．

	ClinicDB		Kvasir		ColonDB		ETIS		EndoScene		**Overall**	
Data	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice
Real	**86.04**	**91.78**	**84.50**	**90.42**	65.68	73.99	63.88	72.75	82.10	89.11	70.40	78.30
Generated	81.51	88.59	82.93	89.02	**67.10**	**75.03**	**64.31**	**74.03**	**83.76**	**90.26**	**70.77**	**78.74**

よって誤差を稼ぐ経路を見つけてしまい，境界の曖昧化という本来の目的が達成されないためである．具体的には，腫瘍領域における BCE 損失 [17] と Dice 損失 [18] を乗算することにより定義する [19].

$$\mathcal{L}_{adv} = -\alpha_{diff}\mathcal{L}_{BCE}^{tumor} \cdot \mathcal{L}_{Dice}^{tumor}. \tag{2}$$

ここで α_{diff} は難化の強度を制御する係数である．加算形式ではいずれか一方の誤差が大きいだけで損失が増大するが，乗算形式では各項が相互に重みとして作用するため，画素単位の誤差と構造的な誤差が同時に大きいサンプルが強調され，勾配更新が自動的にセグメンテーション上重要な領域へ集中する [19].

難化のみを最適化するとマスクの大域構造が歪むため，adaptive t-vMF Dice 損失 [11] を導入する．集中度パラメータ κ は，予測マスク $\mathbf{p}$ と正解マスク $\mathbf{g}$ の Dice 値に応じて動的に決定する．

$$\kappa = \alpha_\kappa \cdot \mathrm{Dice}(\mathbf{p}, \mathbf{g}), \quad \alpha_\kappa = 32. \tag{3}$$

t-vMF Dice 損失は次式で定義される．

$$\mathcal{L}_{tvMF} = \frac{1}{C}\sum_{i=1}^{C}\left(2 - \frac{1+\cos\theta_i}{1+\kappa(1-\cos\theta_i)}\right)^2. \tag{4}$$

この定式化はマスク間の方向的類似性を課し，難化に伴う摂動を許容しつつ大域構造を維持する．さらに，視覚的破綻を抑制するために LPIPS [12] による知覚的正則化を適用し，生成画像と参照画像の深層特徴の距離を制約する．以上により，解剖学的一貫性を維持しながらセグメンテーション難易度のみを高める．

3. 評価実験

本節では提案手法の性能を評価するための実験について述べる．3.1 節では実験条件や評価指標について説明する．3.2，3.3 節では実験結果を示す．

3.1 実験条件・評価指標

評価実験には，Kvasir [20]，CVC-ClinicDB [21]，CVC-ColonDB [22]，ETIS [23]，EndoScene [24] の 5 種の公開ポリープ画像データセットを使用した．学習には Polyps データセットを用い，Kvasir から 900 枚，CVC-ClinicDB から 550 枚の計 1,450 枚の実画像を組み合わせて使用した．NCSD および敵対的学習に用いるセグメンテーションモデル（SANet）は，これらの実画像のみを用いて事前学習した．評価は各データセットごとに個別に行い，EndoScene（60 枚），CVC-ClinicDB（62 枚），Kvasir（100 枚），CVC-ColonDB（380 枚），ETIS（196 枚）をテスト画像として用いた．全体性能は，各テストセットのサンプル数に基づく加重平均として算出した．また，汎化性能を評価するため，CNN ベースおよび Transformer ベースの複数のセグメンテーションネットワークを用いて比較実験を行った．生成モデルには NCSD を用いた．Stable Diffusion v1.5 [25] を初期値として学習した NCSD を事前学習モデルとして用い，その重みを初期値として敵対的学習を行った．敵

図 4: ガイダンススケール α を 0.1 から 0.5 まで変化させた場合に生成された合成ポリープ画像の定性的比較を示す．この結果は，提案する敵対的洗練機構がセグメンテーションの頑健性向上に有効な，より困難な学習サンプルを生成できることを示している．

対的学習段階では DDIM サンプリング [26] を 50 ステップで実行した．生成器の最適化には AdamW[27] を用い，学習率を 1×10^{-7}，ガイダンススケールを $\lambda = 1$ とした．損失関数に関する係数は，$\alpha = 0.4$，$\lambda_{\text{tvMF Dice}} = 0.7$，$\lambda_{\text{LPIPS}} = 1.0$ と設定した．評価指標として Dice 係数および Intersection over Union（IoU）を用いてその平均値である mean Dice（mDice）および mean IoU（mIoU）により性能評価を行った．

3.2 セグメンテーション性能

表 1 にポリープセグメンテーションの性能評価結果を示す．提案手法は SANet の性能を大幅に向上させ，mDice および mIoU はそれぞれ 4.28%，4.71%改善した．より高度なアーキテクチャである Polyp-PVT[13] および CTNet においても，mDice はそれぞれ 1.28%，3.28%，mIoU は 1.50%，3.59%向上した．従来の SoTA 手法である NCSD と比較しても，SANet，Polyp-PVT，CTNet において mDice で 1.06%，0.83%，1.69%，mIoU で 1.07%，0.96%，2.10%の追加的な性能向上を達成した．

改善は全データセットで一様ではない．CTNet は ColonDB と ETIS において NCSD と比較して約 4%の顕著な向上を示した一方，ClinicDB ではベースライン（mIoU 88.16）および NCSD（88.80）を下回る 86.58 となった．ColonDB と ETIS は学習に用いていない分布外データセットであり，ClinicDB は学習画像の供給元である．すなわち改善は分布外データに集中し，既に高い性能が達成されている内挿的なデータセットでは限定的か，場合によっては低下する．

3.3 合成画像のみによる学習

表 2 に，実画像のみで学習した場合と合成画像のみで学習した場合の SANet の性能比較を示す．合成画像のみで学習したモデルは，1,450 枚の実画像のみで学習したベースラインと比較して，Overall で mIoU 0.37%，mDice 0.44%の向上を達成した．内訳を見ると，実画像が直接学習に用いられた ClinicDB と Kvasir[20] ではベースラインをやや下回る一方，分布シフトの大きい ColonDB（mIoU 65.68 → 67.10）と ETIS（63.88 → 64.31），および EndoScene（82.10 → 83.76）ではベースラインを上回った．

この結果の含意は，平均値の改善幅が示すよりも大きい．実データを 1 枚も学習に用いることなく実データ学習と同等以上の性能に到達したということは，患者データを施設外へ持ち出せない状況下でもモデルを構築しうることを意味する．合成画像は難易度を制御可能な形で原理的に無制限に生成できるため，実データの収集に制約を受ける医用画像分野において，分布外データに対する汎化性能向上の有望な解決策となる．

表 3: 生成手法と難化係数 α_{diff} に関するアブレーション．識別器・評価モデルともに SANet を用い，5 つの公開データセットにおける総合性能（Overall）を示す．表は三部構成である．(1) NCSD または提案手法（$\alpha_{diff} = 0.4$）による 2,900 枚単独での比較．(2) NCSD 1,450 枚と提案手法 1,450 枚を組み合わせた設定における α_{diff} の感度分析．(3) NCSD 1,450 枚，提案手法（$\alpha_{diff} = 0.4$）1,450 枚，および α_{diff} を変化させた追加の 1,450 枚を加えた計 4,350 枚による複合的なデータ拡張設定．「N+O」は NCSD と提案手法の合成画像の組み合わせを表す．

Method	α_{diff}	Size	**mIoU (%)**	**mDice (%)**
NCSD	–	2900	68.90	76.77
Ours	–	2900	59.45	67.23
α_{diff} の感度分析（*N+O, 2,900* 枚）				
N+O	0.1	2900	67.93	75.88
N+O	0.2	2900	69.26	77.07
N+O	0.3	2900	68.87	76.86
N+O	0.4	2900	70.45	78.18
N+O	0.5	2900	68.72	76.39
難易度の混合（*N+O, 4,350* 枚）				
N+O	0.4+0.1	4350	70.45	78.32
N+O	0.4+0.2	4350	70.02	77.83
N+O	**0.4+0.3**	**4350**	**70.77**	**78.74**
N+O	0.4+0.5	4350	69.50	77.26

3.4 難化係数と難易度混合

表 3 に，生成手法と難化係数 α_{diff} に関するアブレーション結果を示す．まず，2,900 枚の生成画像のみを学習に用いた場合，提案手法（mIoU 59.45）は NCSD（68.90）を 9.45 ポイント下回った．本手法が意図的に低コントラ

スト領域と境界の曖昧さを含む困難なサンプルを生成していることの直接的な帰結である．この結果の解釈は本研究の中心的な論点であり，4 節で詳述する．

α_{diff} を 0.1 から 0.5 まで変化させたところ，$\alpha_{diff} = 0.4$ において最高性能（mIoU 70.45）が得られた．α_{diff} が小さすぎる場合には生成サンプルの難易度が不十分で頑健性の向上につながらず，逆に大きすぎる場合には形態的歪みが生じて学習信号が不安定になる．$\alpha_{diff} = 0.4$ は構造的一貫性を損なうことなく難易度を最大化する均衡点である．

さらに，$\alpha_{diff} = 0.4$ の合成画像に中程度の難易度である $\alpha_{diff} = 0.3$ のサンプルを追加した設定が，最良の総合性能（mIoU 70.77，mDice 78.74）を示した．単一の難易度水準のみを用いるよりも，複数の難易度分布を混合するほうが優れるという結果である．中程度の難易度のサンプルが基本的な識別能力を強化し，高難易度のサンプルが低コントラスト領域と境界の不確実性に対する頑健性を高めるという，相補的な役割分担が示唆される．

3.5 損失関数の各項の寄与

表 4 に損失関数の各項の寄与を示す．$\mathcal{L}_{tvMF}$ の導入により mDice は 80.61 から 81.91 へ，mIoU は 72.89 から 74.23 へ改善した．角度的な特徴整合が，合成画像とセグメンテーション表現との間の構造的一貫性を高めていると考えられる．

$\mathcal{L}_{lpips}$ 単独の効果は限定的である．LPIPS は生成画像と実画像の高次意味特徴を整合させることでテクスチャ統計を正則化するが，幾何的一貫性を直接強制するものではないため，単独の寄与は $\mathcal{L}_{tvMF}$ に比べて小さい．一方，三項を同時に最適化した場合に最良の性能（mDice 82.58，mIoU 75.11）が得られており，各項が相補的に働いていることが分かる．$\mathcal{L}_{tvMF}$ による構造的制約が特徴幾何を安定させ，$\mathcal{L}_{lpips}$ が知覚的な自然さを保つ．効果的な難例合成には，幾何的一貫性と知覚的整合性の双方が必要である．

3.6 識別器の選択

表 5 に，識別器として用いるセグメンテーションモデルを変えた場合の比較を示す．識別器として SANet を用いた場合に，どのセグメンテーションモデルを学習させた場合でも最高性能が得られた．CNN ベースのアーキテクチャである SANet が低コントラスト領域や曖昧な境界に対して相対的に頑健でないため，凍結識別器として用いると生成器がその弱点を突く方向に強く誘導され，結果としてより効果的な学習信号が得られたことが一因と推測される．ただしこの説明は現時点で直接検証しておらず，仮説にとどまる．

表 4: 損失関数の各項に関するアブレーション．$\mathcal{L}_{base}$ は CrossEntropy 損失と Dice 損失を組み合わせた難化損失，$\mathcal{L}_{tvMF}$ は adaptive t-vMF Dice 損失，$\mathcal{L}_{lpips}$ は LPIPS 知覚損失を表す．

Settings	$\mathcal{L}_{base}$	$\mathcal{L}_{tvMF}$	$\mathcal{L}_{lpips}$	**mDice**	**mIoU**
1	✓			80.61	72.89
2	✓	✓		81.91	74.23
3	✓		✓	80.59	73.05
Ours	✓	✓	✓	**82.58**	**75.11**

表 5: 判別器として用いる異なるセグメンテーションモデルの比較を示す．性能は，3 種類のバックボーンアーキテクチャ（SANet，Polyp-PVT，CTNet）における mDice（%）および mIoU（%）を用いて評価した．

	SANet		Polyp-PVT		CTNet	
Discriminator	mIoU	mDice	mIoU	mDice	mIoU	mDice
SANet (Ours)	**75.11**	**82.58**	**77.43**	**84.47**	**79.69**	**86.38**
Polyp-PVT	73.89	81.49	77.02	83.86	78.53	85.51
CTNet	73.97	81.30	77.36	84.29	79.21	85.85

3.7 生成画像の定性評価

図 3 にポリープ画像の定性的比較を示す．NCSD による生成画像はポリープの形状や質感を忠実に再現しており，境界も明瞭に保たれている．生成モデルとしては優れた性質だが，セグメンテーションモデルにとっては既に解ける種類のサンプルである．一方，提案手法による生成画像では，ポリープと周囲粘膜との境界がより曖昧になっている．特に低コントラスト領域や光反射を含む領域において，境界の識別が困難なサンプルが多く生成されている．同時に病変の位置・大きさ・大域的な形状は保持されており，形態的整合性を維持したまま難易度のみが上昇していることが視覚的に確認できる．

図 4 は α_{diff} を変化させたときの生成画像を示している．α_{diff} の増加に伴って病変と背景のコントラストが単調に低下する一方，病変の形態的一貫性は維持されており，難易度が連続的に制御可能であることが確認できる．

4. 考察

4.1 難例は単独では劣化し，混合において効く

本研究の実験結果のうち最も解釈を要するのは，提案手法が生成した 2,900 枚の難例のみで学習したモデルが，mIoU 59.45 と NCSD 単独の 68.90 を 9.45 ポイント下回った点である．手法の有効性を主張するうえで，一見不都合な結果である．

しかしこれは手法の欠陥ではなく，難例というものの性質を示していると考えられる．難例はモデルの決定境界の

近傍に位置するサンプルである．境界近傍にしかサンプルが存在しない学習データからは，境界がどこにあるべきかを規定するのに必要な，境界から十分に離れた領域の情報が得られない．分類問題の比喩でいえば，サポートベクターだけを与えられても，それがどちらのクラスに属するかを規定する広い領域の情報がなければ安定した決定面は定まらない．

この解釈は，$\alpha_{diff} = 0.4$ の高難易度サンプルに $\alpha_{diff} = 0.3$ の中難易度サンプルを加えた設定が最良であったという結果と整合する．もし難しければ難しいほど良いのであれば，最も難しい設定が最良となるはずだが，そうはならない．有効なのは難易度の絶対的な高さではなく，学習データが覆う難易度の範囲の広さである．

したがって本手法の貢献は，難しい画像を作る手法としてではなく，学習データの難易度分布を制御可能な形で拡張する手法として理解されるべきである．従来の生成型拡張は難易度分布の容易側にサンプルを追加していた．本手法は困難側に，制御されたかたちでサンプルを追加する．両者を組み合わせることで分布全体が広がる．表 2 において合成画像のみでの学習が成立したのも，そこに NCSD による 1,450 枚（容易側）と提案手法による 2,900 枚（困難側）の両方が含まれていたためであり，この解釈と符合する．

4.2 分布外データにおける改善が主たる利得である

データセット別の結果には明確な偏りがある．学習に用いていない ColonDB と ETIS において CTNet は NCSD と比較して約 4%の改善を示した一方，学習画像の供給元である ClinicDB や Kvasir における改善は限定的か，場合によっては低下している．

この偏りは本手法の作用機序から説明できる．学習データと同じ分布に属するデータについてはモデルが既に十分な性能を達成しており，追加の難例が与える情報は限界的である．むしろ既に学習された表現に対する摂動として作用し，わずかな性能低下を招く場合すらある．他方，撮影機器や施設が異なる分布外データに対しては，モデルは低コントラストや境界の曖昧さといった条件で脆弱性を露呈する．本手法が生成する難例は，まさにその脆弱性に対応した学習信号を供給する．

なお，Overall は各テストセットのサンプル数に基づく加重平均であり，全 798 枚のうち ColonDB（380 枚）と ETIS（196 枚）が計 576 枚，約 72%を占める．したがって Overall の改善は主として分布外汎化性能の改善を反映している．これは実応用の観点からむしろ望ましい性質である．実運用で問題となるのは開発時に用いたデータでの性能ではなく，導入先の施設・機器で撮影されたデータでの性能だからである．

4.3 アーキテクチャによる改善幅の差

CTNet が ColonDB と ETIS で特に大きな改善を示した理由は，このアーキテクチャの構造的特性と本手法が生成する難例の特性との適合にあると考えられる．CTNet は深層の広い受容野から得られる文脈情報と高解像度の局所境界情報とを段階的に融合する構造を有する．本手法による難例は境界の曖昧さと低コントラスト領域を強調するため，CTNet はこれらのサンプルから境界に関する特徴を抽出しやすい．

対照的に，SANet は画像中の高コントラスト領域への応答に依存する傾向があり，低コントラストなポリープに対する感度が構造的に制限される．また Polyp-PVT は強力な大域表現能力を持つが，ETIS のように分布シフトが大きく規模の小さいデータセットでは性能が飽和しやすく，追加された合成データから得られる勾配の寄与が相対的に小さくなったと考えられる．これらの説明はアーキテクチャの既知の特性に基づく推測であり，直接の検証は今後の課題である．

5. おわりに

本稿では，画像合成とセグメンテーション最適化を統合する敵対的学習を提案した．固定したセグメンテーションモデルの予測誤差を生成器へ還元することにより，より困難で学習効果の高い画像を生成し，学習の高度化を実現した．

参考文献

[1] J. C. van Rijn *et al.*, “Polyp miss rate determined by tandem colonoscopy: A systematic review,” *The American Journal of Gastroenterology*, vol. 101, no. 2, pp. 343–350, 2006.

[2] W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen, “DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1206–1217, 2023.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[4] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[5] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2256–2265, PMLR, 2015.

[6] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” *arXiv preprint arXiv:2011.13456*, 2020.

[7] V. Thambawita, P. Salehi, S. A. Sheshkal, S. A. Hicks, H. L. Hammer, S. Parasa, T. de Lange, P. Halvorsen,

and M. A. Riegler, "SinGAN-Seg: Synthetic training data generation for medical image segmentation," *PLOS ONE*, vol. 17, no. 5, p. e0267976, 2022.

[8] Y. Du, Y. Jiang, S. Tan, X. Wu, Q. Dou, Z. Li, G. Li, and X. Wan, "ArSDM: Colonoscopy image synthesis with adaptive refinement semantic diffusion models," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 416–426, Springer, 2023.

[9] K. Qiu, Z. Gao, Z. Zhou, M. Sun, and Y. Guo, "Noise-Consistent Siamese-Diffusion for medical image synthesis and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15672–15681, 2025.

[10] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 699–708, Springer, 2021.

[11] S. Kato and K. Hotta, "Adaptive t-vmf dice loss: An effective expansion of dice loss for medical image segmentation," *Computers in Biology and Medicine*, vol. 168, p. 107695, 2024.

[12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

[13] B.-C. Hu, G.-P. Ji, D. Shao, and D.-P. Fan, "PraNet-V2: Dual-supervised reverse attention for medical image segmentation," *arXiv preprint arXiv:2504.10986*, 2025.

[14] B. Xiao, J. Hu, W. Li, C.-M. Pun, and X. Bi, "CT-Net: Contrastive transformer network for polyp segmentation," *IEEE Transactions on Cybernetics*, vol. 54, no. 9, pp. 5040–5053, 2024.

[15] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 4296–4304, 2024.

[16] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[18] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, 2016.

[19] Y. Yokoi and K. Hotta, "Multiplicative loss for enhancing semantic segmentation in medical and cellular images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1273–1281, 2025.

[20] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *Proceedings of the International Conference on Multimedia Modeling*, pp. 451–462, Springer, 2019.

[21] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[22] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2015.

[23] Z. Chen, K. Wang, and Y. Liu, "Efficient polyp segmentation via integrity learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1826–1830, IEEE, 2024.

[24] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of Healthcare Engineering*, vol. 2017, no. 1, p. 4037190, 2017.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

[26] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.