視覚的基盤モデルを用いたTrimodal Cross-Attentional Transformerに基づく再配置対象の検出

○西村喬行，松尾榛夏，杉浦孔明（慶應義塾大学）

本研究では，元の状態および現在の状態から再配置すべき物体を検出する Rearrangement Target Detection (RTD) を扱う．例えば，ロボットが机上の食器を棚に再配置する際に元の状態と現在の状態を比較することで再配置すべき物体の検出を行う．既存研究は，物体の移動を扱うことはできるが，ドア・引き出しの変化は扱わず，複雑な物体においてはセグメンテーション誤りがある．そこで，本研究では，RGBD 画像及びセグメンテーション画像を扱う Trimodal Cross-Attentional Encoder を提案する．RTD タスクにおいて深度情報及び物体のセグメンテーション情報を考慮することで，小物体の検出が可能になることが期待される．新たに作成したデータセットで検証した結果，提案手法はベースライン手法を mIoU 及び F_1-score において，それぞれ 14.5 ポイント及び 6.1 ポイント上回った．

1. はじめに

生活支援ロボットは，高齢化社会において予測される介助者不足の解決策となりうる．生活支援ロボットによって家庭環境内の rearrangement タスク [1] を行うことができれば，人間による指示がなくてもロボットが片付けをすることができ，介助者の負担を減らすことができる．rearrangement タスクでは元の状態と現在の状態を比較することにより再配置すべき物体の検出を行う必要がある．そのため元の状態および現在の状態から再配置すべき物体を検出する Rearrangement Target Detection (RTD) が重要である．

本研究では，目標状態および現在の状態から再配置すべき物体を検出する RTD タスクを扱う．例えば，目標状態では机上に置いてあった物体が，現在の状態で机上にない場合，その物体を検出しマスクした画像を出力することが望ましい．また，目標状態では閉じていたドアが，現在の状態では開いている場合，その扉を検出しマスクした画像を出力することが望ましい．

変化した物体を検出する方法として目標状態及び現在の状態における画像の組を画素値を用いて比較する手法がある．しかしながら、この手法では光や影により物体の色が僅かに変化した場合に誤りが発生する可能性がある．実際に，後述するように RTD タスクにおいて mean Intersection over Union (mIoU) は非常に低い．Scene Change Detection (SCD) を扱う既存研究 [2,3] は，物体の移動を扱うことはできるが，ドア・引き出しの変化は扱わず，複雑な物体においてはセグメンテーション誤りがある．また，[4] は，RTD タスクにおいて良好な結果が報告されているが小物体の検出において性能は十分でない．

本研究では，RGBD 画像及び SAM [5] で生成したセグメンテーション画像を扱う Trimodal Cross-Attentional を提案する．RTD タスクにおいて深度情報及び物体のセグメンテーション情報を考慮することで，小物体の検出が可能になることが期待される．既存手法と異なる点は，RGBD 画像及び SAM で生成したセグメンテーション画像を扱う Trimodal Cross-Attentional Encoder を導入している点である．depth 画像を活用することで，物体間の距離を把握することが可能となり，性能向上が期待される．また，SAM により生成したセグメンテーション画像を活用することで，物体の境界や領域を正確に認識し，性能の向上が見込まれる．提案手法の主要な新規性は以下である．

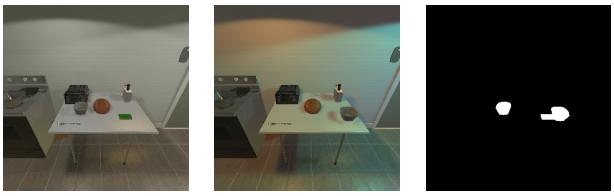

図 1 対象タスクの例．左から目標状態，現在の状態，正解マスク画像．

- RGBD 画像及び SAM で生成したセグメンテーション画像を扱う Trimodal Cross-Attentional Encoder を導入する．

2. 関連研究

SCD は，室内環境で得られた画像やストリートビュー画像を用いて，物体や街並みの変化を検出するタスクである [2,3,6]．CSCDNet [2] は，カメラの異なる視点の違いを扱うことができ，セマンティックなシーン変化検出タスクに取り組む手法である．C-3PO [6] は，変化情報を抽出し，時間的特徴を融合する．また，SCD に類似したタスクに本研究で扱う RTD がある．RTD を扱う [4] は，Cross-Attentional Transformer を導入し家庭環境において視覚情報から再配置対象を検出する．RoboCup@Home 競技会は，ロボットが物体の片付けを行うベンチマークである [7]．RoboCup@Home 競技会で行われているような片付けタスクにおいて，整頓済みの室内環境を目標状態とした場合，現在の状態との比較に RTD を適用できる．

3. 問題設定

RTD では，ロボットに目標状態と現在の状態が与えられたときに状態が変化したことを検出する．本研究では，目標状態および現在の状態から再配置すべき物体を検出する RTD タスクを扱う．

本タスクでは，目標状態および現在の状態の画像から再配置すべき物体のマスク画像を生成することが期待される．図 1 に本対象タスクの代表例を示す．左か

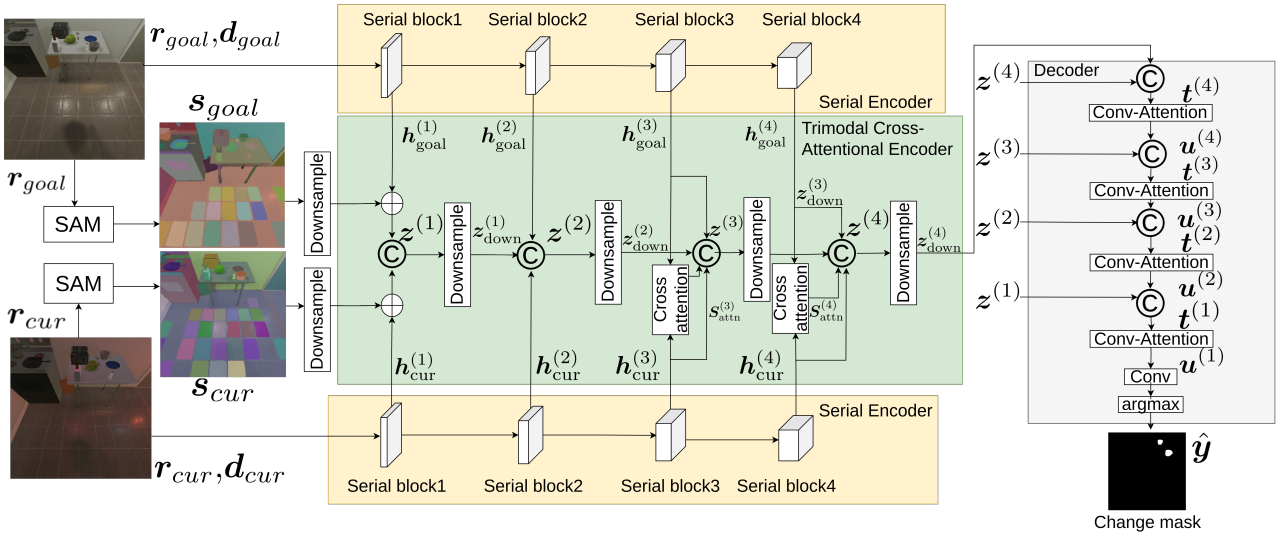

図 2　提案手法のネットワーク構造．ここで Ⓒ は concatenation である．

ら順に目標状態の画像，現在の状態の画像，再配置対象のマスク画像である．例では，器が移動し，緑のカードが消失しているため，これらの物体を再配置対象としてマスク画像が生成されることが望まれる．本タスクでは，RGBD 画像の組を入力とする．出力は，目標状態および現在の状態の画像から再配置すべき物体に対して，変化した物体の画素を 1，それ以外の画素を 0 としたマスク画像である．本論文で使用する用語を以下に定義する．

- **再配置対象**：目標状態と現在の状態の間で，位置および向きが変化する物体，開閉する引き出しおよび扉

本研究では，1 視点からの RTD タスクを前提とし，物体の変化距離が 30 cm 以下，扉の開閉が最大開閉角度の 60 %以下の物体を再配置対象としない．また，2022 AI2-THOR Rearrangement Challenge [8] の設定と合わせるため，1 画像中の再配置対象の個数は 5 個以下であることを前提とする．本実験では，正解マスク画像と予測マスク画像の mIoU と F_1-score を評価尺度として用いる．

モデルの学習訓練には大量のデータ及びアノテーションの付与が必要である．実機によるデータ収集では，人間が変化した物体を特定するため，多大な時間と労力を要する．そのため，大量のデータを短時間かつ自動で取得できるシミュレーション環境を用いる．

4.　提案手法

本モデルの主要モジュールは，Serial Encoder 及び Trimodal Cross-Attentional Encoder の 2 つである．図 2 に Serial Encoder 及び Trimodal Cross-Attentional Transformer の構造を示す．また，図 2 に Decoder の構造を示す．ネットワークの入力を $\{\boldsymbol{r}_{\text{goal}} \in \mathbb{R}^{3\times300\times300}, \boldsymbol{r}_{\text{cur}} \in \mathbb{R}^{3\times300\times300}, \boldsymbol{d}_{\text{goal}} \in \mathbb{R}^{3\times300\times300}, \boldsymbol{d}_{\text{cur}} \in \mathbb{R}^{3\times300\times300}\}$ と定義する．ここで，順に目標状態の RGB 画像，現在の状態の RGB 画像，目標状態の depth 画像，現在の状態の depth 画像を表す．ただし，depth 画像は，surface normal [9] で変換したものである．$\boldsymbol{r}_{\text{goal}}$ に対して，SAM [5] の学習済みモデルを利用してセグメンテーション画像を取得し $\boldsymbol{s}_{\text{goal}}$ とした．次に，$\boldsymbol{r}_{\text{goal}}$ 及び $\boldsymbol{d}_{\text{goal}}$ をチャネル方向に結合し，$\boldsymbol{x}_{\text{goal}}$ とした．同様に，$\boldsymbol{s}_{\text{cur}}$ 及び $\boldsymbol{x}_{\text{cur}}$ を生成した．

4.1　Serial Encoder

Serial Encoder は M 個の serial block [10] を用いて画像特徴量を抽出する．i 番目の serial block での処理は以下である．まず，$\boldsymbol{x}_{\text{cur}}$ を入力として受け取り，パッチ埋め込み層を用いてダウンサンプリングを行い，$\boldsymbol{o}_{\text{cur}}^{(i)} \in \mathbb{R}^{H_i W_i \times C_i}$ を得る．ここで，H_i，W_i，C_i はそれぞれ i 番目の serial block から出力される画像特徴マップの高さ，幅，チャネル数を表す．次に，$\boldsymbol{v}_{\text{cur}}^{(i)} \in \mathbb{R}^{C_i}$ を平坦化して画像トークンとし，画像トークンに CLS トークン $\boldsymbol{v}_{\text{cur}}^{(i)} \in \mathbb{R}^{C_i}$ を結合した後，Conv-Attention Module を通じて，$\boldsymbol{v}_{\text{cur}}^{(i)}$ に Depthwise Convolution 及び Factorized Attention [10] を適用する．次に，画像トークンと CLS トークンを分離し，画像トークンに変形して $\boldsymbol{h}_{\text{cur}}^{(i)} \in \mathbb{R}^{H_i \times W_i \times C_i}$ を得る．$\{\boldsymbol{h}_{\text{cur}}^{(i)} \mid i = 1, \ldots, M\}$ がこのモジュールの出力である．同様に，入力 $\boldsymbol{x}_{\text{goal}}$ から，$\{\boldsymbol{h}_{\text{goal}}^{(i)} \in \mathbb{R}^{H_i \times W_i \times C_i} \mid i = 1, \ldots, M\}$ を得る．

4.2　Trimodal Cross-Attentional Encoder

Trimodal Cross-Attentional Encoder は 2 つの目標状態と現在の状態の関係性を考慮した画像特徴量を得る．Serial Encoder から得た $\{\boldsymbol{h}_{\text{goal}}^{(i)}, \boldsymbol{h}_{\text{cur}}^{(i)} \mid i = 1, \ldots, M\}$ 及び $\{\boldsymbol{s}_{\text{goal}}, \boldsymbol{s}_{\text{cur}}\}$ を入力とする．ここで，$i = 1$ のときのみ，$\boldsymbol{h}_{\text{goal}}^{(1)}$ に $\boldsymbol{s}_{\text{goal}}$ を加算し，同様に $\boldsymbol{h}_{\text{cur}}^{(1)}$ に $\boldsymbol{s}_{\text{cur}}$ を加算する．まず，$\boldsymbol{h}_{\text{goal}}^{(1)}$ と $\boldsymbol{h}_{\text{cur}}^{(1)}$ を連結して $\boldsymbol{z}^{(1)} \in \mathbb{R}^{(C_1\times2)\times H_1 \times W_1}$ を得る．次に，$\boldsymbol{z}^{(1)}$ に対してダウンサンプリングを行い $\boldsymbol{z}_{\text{down}}^{(1)} \in \mathbb{R}^{C_2 \times H_2 \times W_2}$ を得た後，$\boldsymbol{z}_{\text{down}}^{(1)}$，$\boldsymbol{h}_{\text{goal}}^{(2)}$，$\boldsymbol{h}_{\text{cur}}^{(2)}$ を連結して $\boldsymbol{z}^{(2)} \in \mathbb{R}^{(C_1+C_2\times2)\times H_2 \times W_2}$ を得る．さらに，$\boldsymbol{z}^{(2)}$ に対してダウンサンプリングを行い $\boldsymbol{z}_{\text{down}}^{(2)} \in \mathbb{R}^{C_2 \times H_3 \times W_3}$ を得る．任意の行列 $\boldsymbol{X}_{\text{A}}$ 及び

表 1　RTDD データセットにおける定量的結果

手法		$\boldsymbol{d}'$	$\boldsymbol{s}$	mIoU [%]	F_1-score [%]
画素値比較		-	-	1.7	11.7
baseline [4]		-	-	59.0±0.5	85.2±0.3
Ours	(i)	✓		73.4±0.6	91.3±0.2
	(ii)		✓	58.3±0.7	84.9±0.3
	(full)	✓	✓	**73.5±0.3**	**91.3±0.1**

$\boldsymbol{X}_{\mathrm{B}}$ に対する cross-attention の式を以下に定義する．

$$\boldsymbol{f}_{\mathrm{attn}}^{(j)}(\boldsymbol{X}_{\mathrm{A}}, \boldsymbol{X}_{\mathrm{B}}) = \mathrm{softmax}(\frac{(W_q^{(j)}\boldsymbol{X}_{\mathrm{A}})(W_k^{(j)}\boldsymbol{X}_{\mathrm{B}})^\top}{\sqrt{d}})(W_v^{(j)}\boldsymbol{X}_{\mathrm{B}})$$

ここで，W_q，W_k，W_v は学習可能な重み，d はスケーリングファクターである．次に，$i = 3, \ldots, M-1$ のとき，$\boldsymbol{h}_{\mathrm{goal}}^{(i)}$ および $\boldsymbol{h}_{\mathrm{cur}}^{(i)}$ に対して cross-attention を適用して，Attention スコア $\boldsymbol{S}_{\mathrm{attn}}^{(i)}$ を算出する．

$$\boldsymbol{S}_{\mathrm{attn}}^{(M)} = \{\boldsymbol{f}_{\mathrm{attn}}^{(j)}(\boldsymbol{h}_{\mathrm{goal}}^{(i)}, \boldsymbol{h}_{\mathrm{cur}}^{(i)}) \mid j = 1, \ldots, A\}$$

ここで，Attention の Head 数を A とする．次に，$\boldsymbol{z}_{\mathrm{down}}^{(i-1)}$，$\boldsymbol{S}_{\mathrm{attn}}^{(i)}$，$\boldsymbol{h}_{\mathrm{goal}}^{(i)}$ および $\boldsymbol{h}_{\mathrm{cur}}^{(i)}$ を連結して $\boldsymbol{z}^{(i)} \in \mathbb{R}^{(C_{i-1}+C_i\times 3)\times H_i \times W_i}$ を得る．その後，$\boldsymbol{z}^{(i)}$ に対してダウンサンプリングを行い $\boldsymbol{z}_{\mathrm{down}}^{(i)} \in \mathbb{R}^{C_i \times H_{i+1} \times W_{i+1}}$ を得る．同様に $\boldsymbol{S}_{\mathrm{attn}}^{(M)}$，$\boldsymbol{z}^{(M)} \in \mathbb{R}^{(C_{M-1}+C_M\times 3)\times H_M \times W_M}$ を得る．ここで，$\boldsymbol{z}^{(M)}$ に畳み込み層を適用して $\boldsymbol{z}_{\mathrm{down}}^{(M)} \in \mathbb{R}^{C_M \times H_M \times W_M}$ を得る．したがって，$\{\boldsymbol{z}^{(i)} \mid i = 1, \ldots, M\}$ および $\boldsymbol{z}_{\mathrm{down}}^{(M)}$ が本モジュールの出力である．

4.3　Decoder

Decoder では，Trimodal Cross-Attentional Encoder から得た $\{\boldsymbol{z}^{(i)} \mid i = 1, \ldots, M\}$ および $\boldsymbol{z}_{\mathrm{down}}^{(M)}$ を入力とし，M 個の Conv-Attention Module を用いて，モデルの入力と同じサイズの予測マスク画像を得る．まず，パッチ埋め込み層を，$\{\boldsymbol{z}_{\mathrm{down}}^{(M)}, \boldsymbol{z}^{(M)}\}$ に適用し $\boldsymbol{t}^{(M)} \in \mathbb{R}^{H_M \times H_M \times W_{M-1}}$ を得る．次に，Conv-Attention Module 及びアップサンプリング層を $\boldsymbol{t}^{(M)}$ に適用して $\boldsymbol{u}^{(M)} \in \mathbb{R}^{H_{M-1} \times H_{M-1} \times W_{M-1}}$ を得る．さらに，$i = M-1, \ldots, 2$ のとき，$\{\boldsymbol{u}^{(i+1)}, \boldsymbol{z}^{(i)}\}$ にパッチ埋め込み層を適用する．その後，Conv-Attention Module 及びアップサンプリング層を $\boldsymbol{t}^{(i)} \in \mathbb{R}^{H_i \times W_i \times C_{i-1}}$ に適応し $\boldsymbol{u}^{(i)} \in \mathbb{R}^{C_{i-1} \times H_{i-1} \times W_{i-1}}$ を得る．最後に，$\{\boldsymbol{u}^{(2)}, \boldsymbol{z}^{(1)}\}$ から $\boldsymbol{u}^{(1)} \in \mathbb{R}^{C_{i-1} \times H_{i-1} \times W_{i-1}}$ を得る．$\boldsymbol{u}^{(1)}$ にカーネルサイズ 1 の畳み込み層を適用した後，2 値化し，モデル全体の最終的な出力である 256×256 の 2 値マスク画像 $\hat{\boldsymbol{y}}$ を得る．損失関数 L は [4] に示す式と同じ式を用いる．ここで，K，N，$\boldsymbol{y}_{i,j}$ 及び $\hat{\boldsymbol{y}}_{i,j}$ は順にクラス数，全画素数，正解マスク画像の i 番目のクラスの j 番目の画素値，予測マスク画像の i 番目のクラスの j 番目の画素値を示す．ϵ は，0 除算を避けるために使用される小さな正の値で 1×10^{-7} に設定される．

5.　実験設定

5.1　データセット

本論文では，シミュレーション環境として AI2-THOR [8] を用いて新たに RTDD データセットを作成した．本データセットは以下の手順で作成された．まず，環境の目標状態において，ロボットのヘッドカメラから取得された RGBD 画像を保存した．その後，部屋にある物体をランダムに移動または扉・引き出しを開閉し，再びロボットのヘッドカメラから取得された RGBD 画像を保存した．各状態の RGBD 画像を保存する際，再配置対象の ID，目標状態および現在の状態におけるシミュレーション環境上での位置，エージェントのヘッドカメラから取得された画像上での位置のそれぞれを記録した．ただし，RTDD データセットでは，環境上物体の変化距離が 30 cm 以下，扉・引き出しにおいては再配置前後の角度差が最大開閉角度の 60%未満の物体を再配置対象としない．

変化検出タスクにおいては，PCD [11] や ChangeSim [12] などが標準データセットとして提案されている．しかし，これらのデータセットは rearrangement タスクが行われる家庭環境で作成されたものではなく，RTD には適さないため，新たに RTDD データセットを作成した．depth 画像を法線マップに基づいて 256 階調の RGB 画像とした [9]．RTDD データセットのサンプル数は，12000 である．1 サンプルは，目標状態及び現在の状態の RGBD 画像の組，および変化マスク画像からなる．本研究のデータセットは，訓練集合，検証集合，テスト集合のサンプル数はそれぞれ 10000，1000，1000 である．また，それぞれの集合が収集された環境に重複は無く，テスト集合は未知の環境である．本論文では，訓練集合および検証集合をそれぞれパラメータの更新およびハイパーパラメータの選択に使用した．また，テスト集合を性能の評価に使用した．

5.2　パラメータ

最適化関数は Adam($\beta_1 = 0.5$,β_2=0.999) を使用し，学習率は 0.001，バッチサイズは 16，損失関数の重みは $\gamma_{ce} = 1$，$\gamma_{sDice} = 1$ とした．エポック数の 5 回平均は 49 であった．Serial Encoder の serial block 及び Decoder における Conv-Attentional Module の層数は順に [2, 2, 2, 2], [1, 1, 1, 1] と設定した．次に、$H_1 = W_1 = C_1 = 64$ および $(H_{i+1}, W_{i+1}, C_{i+1}) = (H_i/2, W_i/2, C_i \times 2)$ とする．ただし，$C_3 = 320$ および $C_4 = 512$ である．Serial Encoder，Trimodal Cross-Attentional Encoder，Decoder のそれぞれにおける Attention の Head 数はすべて 8 とした．

提案手法における訓練可能パラメータ数は約 2600 万である．積和演算数は 1.06×10^{11} である．学習は GeForce RTX 3090（メモリ 24GB）および Intel Corei9 12900K を搭載した計算機上で行った．学習は，3 時間程度で完了した．また，1 サンプルあたりの推論に要した時間は 16ms 程度であった．

6.　実験結果

6.1　定量的結果

定量的結果を表 1 に示す．表 1 の行は手法，列は mIoU 及び F_1-score を表している．なお，実験はそれぞれ 5 回行い，その平均及び標準偏差を示す．RTD において，良好な結果が報告されている [4] をベースラインとして選択した．評価尺度は mIoU と F_1-score を用いた．mIoU を，本実験の主要尺度とした．mIoU は，

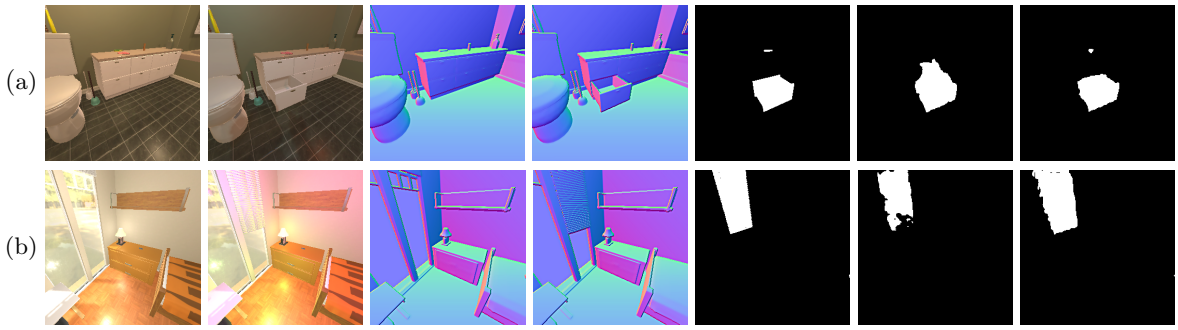

図 3 定性的結果（成功例）．左列から順番に $\boldsymbol{r}_{\text{goal}}$，$\boldsymbol{r}_{\text{cur}}$，$\boldsymbol{d}'_{\text{goal}}$，$\boldsymbol{d}'_{\text{cur}}$，$\boldsymbol{y}$，baseline 手法によって得られた $\hat{\boldsymbol{y}}$，及び提案手法によって得られた $\hat{\boldsymbol{y}}$.

以下の式で定義される．

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \text{IoU}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i)$$

ここで，N はサンプル数，$\hat{\boldsymbol{y}}_i$ ，$\boldsymbol{y}_i$ は i 番目のサンプルにおける予測及び正解マスク，IoU は 2 つのマスク間の Intersection over Union (IoU) を示す．mIoU および F_1-score は，RTD と関連の深い scene chnage detection において標準的な尺度として用いられる．したがって，本研究でも評価尺度として用いた．表より，提案手法及びベースラインの mIoU はそれぞれ 73.5%及び 59.0%であり，また，F_1-score はそれぞれ 91.3%及び 85.2%であった．すなわち，提案手法はベースラインを，mIoU 及び F1-score でそれぞれ 14.5 ポイント及び 6.1 ポイント上回った．この性能差は統計有意であった ($p < 0.05$)．

6.2 定性的結果

図 3 に提案手法の成功例の定性的結果を示す．左列から順番に $\boldsymbol{r}_{\text{goal}}$，$\boldsymbol{r}_{\text{cur}}$，$\boldsymbol{d}'_{\text{goal}}$，$\boldsymbol{d}'_{\text{cur}}$，$\boldsymbol{y}$，baseline 手法によって得られた $\hat{\boldsymbol{y}}$，及び提案手法によって得られた $\hat{\boldsymbol{y}}$ である．図 3(a) は，棚上に置かれたスポンジの移動及び引き出しの開閉がある例である．ベースライン手法は，移動したスポンジを再配置対象として検出できていない．また，開いた引き出しの領域を正確に予測できていない．一方，提案手法は，移動したスポンジの検出及び開いた引き出しの領域予測に成功した．図 3(b) は，目標状態では開いていたブラインドが現在の状態では閉じている例である．ベースライン手法は，ブラインドの輪郭及び内部を正確に予測できなかった．一方，提案手法は，ブラインドの輪郭及び内部をより正確に予測できた．

6.3 Ablation Study

ablation 条件として，以下の 2 条件を定めた．

(i) SAM [5] 画像を使わない場合：$\boldsymbol{s}_{\text{goal}}$ 及び $\boldsymbol{s}_{\text{cur}}$ を用いないことによる提案手法への影響を調査した．

(ii) depth 画像を使わない場合：$\boldsymbol{d}'_{\text{goal}}$ 及び $\boldsymbol{d}'_{\text{cur}}$ を用いないことによる提案手法への影響を調査した．

表 1 に ablation study の定量的結果を示す．表 1 より，(i) では，mIoU 及び F_1-score はそれぞれ 73.4 及び 91.3 であった．同様に (ii) では，それぞれ 58.3 及び 84.9 であった．これより，depth 画像は，性能向上に大きく寄与していることがわかった．また，SAM [5] で作成したセグメンテーション画像は僅かながら性能向上に寄与したことが示唆された．

7. 結論

本研究では，目標状態および現在の状態から再配置すべき物体を検出する RTD タスクを扱った．提案手法の貢献は以下である．

- RGBD 画像及び SAM [5] で生成したセグメンテーション画像を扱う Trimodal Cross-Attentional Encoder を導入した．
- RTDD データセットにおいて，提案手法がベースライン手法を上回る性能を得た．

謝辞

本研究の一部は，JSPS 科研費 23H03478，JST ムーンショット，NEDO の助成を受けて実施されたものである．

参 考 文 献

[1] L. Weihs, M. Deitke, A. Kembhavi, et al., “Visual Room Rearrangement,” CVPR, pp.5922–5931, 2021.
[2] K. Sakurada, M. Shibuya, and W. Wang, “Weakly Supervised Silhouette-based Semantic Scene Change Detection,” ICRA, pp.6861–6867, 2020.
[3] S. Chen, K. Yang, and R. Stiefelhagen, “DR-TANet: Dynamic Receptive Temporal Attention Network for Street Scene Change Detection,” IV, pp.502–509, 2021.
[4] 松尾榛夏，石川慎太郎，杉浦孔明，“物体再配置タスクのための Co-Scale Cross-Attentional Transformer,” JSAI，2023.
[5] A. Kirillov, E. Mintun, N. Ravi, H. Mao, et al., “Segment Anything,” arXiv preprint arXiv:2304.02643, 2023.
[6] G.-H. Wang, et al., “How to Reduce Change Detection to Semantic Segmentation,” Pattern Recognition, 2023.
[7] L. Iocchi, et al., “RoboCup@ Home: Analysis and results of evolving competitions for domestic and service robots,” Artificial Intelligence, vol.229, pp.258–281, 2015.
[8] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, et al., “AI2-THOR: An Interactive 3D Environment for Visual AI,” arXiv preprint arXiv:1712.05474, 2017.
[9] A. Aakerberg, K. Nasrollahi, et al., “Depth Value Pre-Processing for Accurate Transfer Learning based RGB-D Object Recognition,” IJCCI, pp.121–128, 2017.
[10] W. Xu, Y. Xu, et al., “Co-Scale Conv-Attentional Image Transformers,” ICCV, pp.9981–9990, 2021.
[11] K. Sakurada and T. Okatani, “Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation,” BMVC, 2015.
[12] J.-M. Park, J.-H. Jang, et al., “ChangeSim: Towards End-to-End Online Scene Change Detection in Industrial Indoor Environments,” IROS, pp.8578–8585, 2021.