マルチモーダル基盤モデルと拡散モデルに基づく対象物体の参照表現セグメンテーション

○今井悠人，飯岡雄偉，畑中駿平，九曜克之，杉浦孔明（慶應義塾大学）

生活支援ロボットが把持を行う場面では，正確な物体の位置や形状の特定が重要である．本研究では，複数の参照表現を含む命令文および画像から把持対象の物体を予測するタスクを扱う．既存研究では，対象となる物体の領域予測性能は不十分であった．そこで本論文では，2 段階のマルチモーダルセグメンテーションモデルを提案する．本手法の新規性は，画像の特徴抽出を並列に行うモジュールおよび拡散モデルによって抽出した画像特徴量と言語特徴量の交差注意機構の導入である．提案手法は本タスクと関連の深い参照表現セグメンテーションタスクの標準的な評価尺度において，ベースライン手法を上回った．

1. はじめに

高齢化が進行している現代社会において，日常生活における介助支援の重要性が一層高まっている．それに伴い，在宅介助者の不足が社会問題になっている．これを解決するために，被介助者に物理的な支援が可能な生活支援ロボットが注目されている．被介助者が自然言語による指示で生活支援ロボットの物体把持や移動に関する操作ができると便利である．しかし，生活支援ロボットの命令文理解性能には改善の余地がある．また，ロボットが物体を把持する場面では，その形状や位置を特定する必要がある．そのため，画素単位での把持物体の領域予測が有用である．

本研究では，自然言語による命令文の内容を解釈し，動作の対象となる物体を特定する Object Segmentation from Manipulation Instructions（OSMI）タスク [1] を扱う．OSMI タスクと関係が深いタスクとして，Referring Expression Segmentation（RES）タスク [2] がある．RES タスクは命令文ではなく参照表現を含む文から予測を行う．一方，本研究で扱う OSMI タスクでは，命令文に複数の参照表現が複数含まれる場合，動作対象と同じカテゴリに属する複数の物体が存在する場合があり，一般的な RES タスクよりも困難な問題である．

参照表現理解タスクに関する既存研究は数多く存在する [1,3,4]．しかし，これらの既存手法では，部分的なマスクしか生成されない場合や，複雑な参照表現を理解できず，対象と同カテゴリの異なる物体を予測してしまう場合があるため，ロボットの把持操作を目的とする OSMI タスクのモデルとしては不十分である．

本研究では，複雑な参照表現を理解するために，マルチモーダル特徴量を用いてマスクを洗練する手法を提案する．既存手法 [1,5] との主要な相違点は，拡散モデルの 1 つである DDPM [6] を拡張し，言語特徴量との交差注意機構を導入することで，マルチモーダルセグメンテーションを行う点である．

2. 関連研究

参照表現理解タスクは，画像および画像中の特定の領域を指すテキストをもとに，その領域を矩形領域や画素単位で予測するタスクであり，広く研究が行われている [7–9]．LAVT [3] は RES タスクを扱う代表的なモデルであり，言語特徴と視覚特徴の early fusion によって特徴量を抽出する．また，近年では対象物体を画素単位で予測するのではなく，ポリゴンの頂点群と矩形領域のシーケンス予測として参照表現理解タスクを解くアプローチも存在する [10,11]．一方，生活支援ロボットの参照表現理解を目的として，物体操作の指示文から把持物体を推定する問題に取り組んだ研究も存在する [12,13]．飯岡らは OSMI タスクに初めて取り組み，OSMI の評価のための SHIMRIE データセットを作成し，2 段階のセグメンテーションモデルである MDSM を提案している [1]．MDSM は，Encoder-Decoder モデルで生成したマスクに対し，DDPM [6] を用いて洗練を行っている．

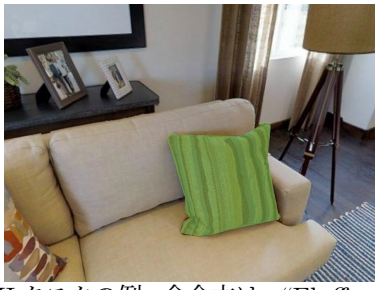

図 1: OSMI タスクの例．命令文は，“Fluff up the pillow with the brown yellow and blue stripes on the light colored sofa in the family room.”

3. 問題設定

本研究では，Object Segmentation from Manipulation Instructions (OSMI) タスク [1] を扱う．本タスクの目的は，複数の参照表現を含む物体操作に関する命令文から，動作の対象物体の領域を画素単位で予測することである．図 1 は OSMI タスクの例を示している．ここで，望ましい出力は図中の緑色で示されたセグメンテーションマスクである．

本稿において，入出力を以下とする．

- **入力:** 命令文と画像のペア
- **出力:** 対象物体の 2 値マスク画像

また，本論文で使用する用語を以下に定義する．

- **対象物体:** 与えられた命令文をロボットが実行する際に，動作の対象となる物体

本研究では，それぞれの入力における対象物体は 1 つしか存在しないことを前提とする．また，評価尺度は，RES タスクにおいて標準的である mean IoU (mIoU)，overall IoU (oIoU)，Precision@0.5 (P@0.5) を用いる．

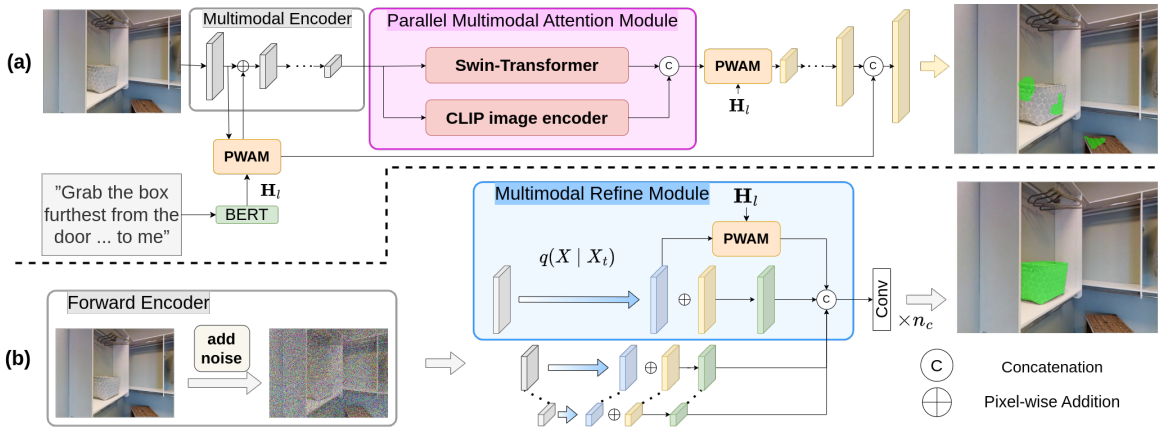

図 2: 提案手法のネットワーク構造．(a) intermediate training step (ITS)，(b) diffusion step (DS)

4. 提案手法

本研究では MDSM [1] を拡張する．この拡張は，拡散モデルを通じて得られた潜在表現と言語特徴量の関係をモデル化するアプローチであり，参照表現を用いたセグメンテーションタスク一般に応用可能な広範な手法であると考えられる．

本手法のネットワーク構造を図 2 に示す．図において，Conv は畳み込みを表す．提案手法は intermediate training step (ITS) および diffusion step (DS) から構成される．ITS で用いるモデルは，Multimodal Encoder，PMAM，PWAM [3] の 3 つの主要なモジュールを含む．DS で用いるモデルは Forward Encoder [1] および MRM の 2 つの主要なモジュールから成る．

4.1 Intermediate training step (ITS)

4.1.1 入力

ITS における入力は画像 $X \in \mathbb{R}^{H \times W \times C}$ 及び one-hot vector の形式で表現されている命令文 $L \in \{0,1\}^{v \times T}$ である．ここで，H，W，C，v，T はそれぞれ画像の幅，高さ，チャネル数，命令文の語彙サイズ，最大トークン数に対応する．

L に対しては，BERT [14] を用いて言語特徴量 $\boldsymbol{H}_l \in \mathbb{R}^{C_l \times T}$ を抽出する．ただし，C_l は各トークンにおける特徴量の次元数を表す．

4.1.2 Multimodal Encoder

Multimodal Encoder では，階層構造によって命令文と画像から特徴抽出を行う．以降，層の数を M とする．

第 i 層目の Multimodal Encoder では，マルチモーダル特徴量 $\boldsymbol{E}^{(i-1)} \in \mathbb{R}^{H^{(i-1)} \times W^{(i-1)} \times C^{(i-1)}}$ を入力として，$\boldsymbol{E}^{(i)}$ を出力する．ここで，$H^{(i)}$，$W^{(i)}$，$C^{(i)}$，はそれぞれ第 i 層での画像特徴量の高さ，幅，チャネル数に対応する．なお，$\boldsymbol{E}^{(0)} = X$ とする．以下，Swin Transformer による特徴抽出を $f_{\text{swin}}(\cdot)$ と表す．$\boldsymbol{E}^{(i)}$ は，後述する PWAM を用いて得られたマルチモーダル特徴量 $\boldsymbol{F}^{(i)} \in \mathbb{R}^{H^{(i)} \times W^{(i)} \times C^{(i)}}$ および Swin Transformer によって得られる画像特徴量 $V_{\text{swin}}^{(i)} = f_{\text{swin}}\left(\boldsymbol{E}^{(i-1)}\right) \in \mathbb{R}^{H^{(i)} \times W^{(i)} \times C^{(i)}}$ を用いて，以下のように計算できる．

$$\boldsymbol{E}^{(i)} = \boldsymbol{F}^{(i)} \odot \tanh\left(\omega_f\left(\boldsymbol{F}^{(i)}\right)\right) + V_{\text{swin}}^{(i)} \quad (1)$$

ここで，$\omega_{\cdot}$，$\odot$ はそれぞれ，1×1 の畳み込み，アダマール積を表す．

4.1.3 PMAM

PMAM では，特徴量抽出器を組み合わせることで，画像と言語の特徴抽出を行う．本モジュールにおける入出力を次のように定義する．入力は Multimodal Encoder の $\lambda - 1$ 層目で得られたマルチモーダル特徴量 $\boldsymbol{E}^{(\lambda-1)}$ とする．出力はマルチモーダル特徴量 $\boldsymbol{E}^{(\lambda)}$ とする．まず，$\boldsymbol{E}^{(\lambda-1)}$ をそれぞれ Swin Transformer および CLIP image encoder に入力する．このとき，前者の出力を $V_{\text{swin}}^{(\lambda)}$ とし，後者の出力を $V'_{\text{clp}} \in \mathbb{R}^{C_{\text{clp}}}$ とする．次に，V'_{clp} の次元数が $V_{\text{swin}}^{(\lambda)}$ と等しくなるように整形したものを V_{clp} とする．これらを用いて，出力は式 (1) と同様に以下の形で表せる．

$$\boldsymbol{E}^{(\lambda)} = \boldsymbol{F}^{(\lambda)} \odot \tanh\left(\omega_f\left(\boldsymbol{F}^{(\lambda)}\right)\right) + \text{Conv}\left([V_{\text{swin}}; V_{\text{clp}}]\right)$$

ただし，Conv および $[\cdot;\cdot]$ は，それぞれ畳み込みおよびチャネル方向への結合を表す．

4.1.4 PWAM

Pixel-Word Attention Module (PWAM) [3] は，$\boldsymbol{E}^{(i)}$ および $\boldsymbol{H}_l$ を入力にとり，マルチモーダル特徴量 $\boldsymbol{F}^{(i)}$ を出力する．計算過程は以下で表せる．

$$\boldsymbol{G}'^{(i)} = \sqrt{C^{(i)}}\,\text{flatten}\left(\omega_{iq}\left(\boldsymbol{E}^{(i)}\right)\right)\left(\omega_{ik}\left(\boldsymbol{H}_l\right)\right)^{\top}$$

$$\boldsymbol{G}^{(i)} = \text{softmax}\left(\boldsymbol{G}'^{(i)}\right)\omega_{iv}\left(\boldsymbol{H}_l\right)$$

$$\boldsymbol{F}^{(i)} = \omega_{if}\left(\omega_{im}\left(\boldsymbol{E}^{(i)}\right)\right) \odot \omega_{iw}\left(\text{unflatten}\left(\boldsymbol{G}^{(i)\top}\right)\right)$$

また，この演算を $\boldsymbol{F}^{(i)} = \text{PWAM}\left(\boldsymbol{E}^{(i)}, \boldsymbol{H}_l\right)$ と表す．

4.1.5 Decoder

ITS では各 $\boldsymbol{F}^{(i)}$ を用いて最終的なセグメンテーションマスクを生成する．計算過程を以下に示す．

$$\boldsymbol{D}^{(i)} = \begin{cases} \boldsymbol{F}^{(i)} & i = M \\ \text{Conv}\left(\left[\text{upsample}\left(\boldsymbol{D}^{(i+1)}\right); \boldsymbol{F}^{(i)}\right]\right) & i \neq M \end{cases}$$

ここで，upsample は bilinear 補完を用いたアップサンプリングを表す．ITS におけるマスクの予測確率の推定値 $p\left(\hat{\boldsymbol{y}}_{\text{its}}\right)$ は，$p\left(\hat{\boldsymbol{y}}_{\text{its}}\right) = \text{softmax}\left(\text{Conv}\left(\boldsymbol{D}^{(1)}\right)\right)$ と計算できる．ITS の出力は，$p\left(\hat{\boldsymbol{y}}_{\text{its}}\right)$ を閾値 0.5 で二値化したマスク画像 $\hat{\boldsymbol{y}}_{\text{its}}$ である．

表 1: ベースライン手法との定量的比較

Method	Condition	λ	w/ PWAM	mIoU [%]	oIoU [%]	P@0.5 [%]
(i) LAVT [3]	-	-	-	24.27 ± 3.15	22.25 ± 2.85	21.27 ± 5.66
(ii) MDSM [1]	-	-	-	33.02 ± 5.51	30.25 ± 4.92	32.76 ± 5.28
(iii) Ours	(a-1)	1	√	30.79 ± 2.64	29.06 ± 2.03	29.20 ± 4.63
	(a-2)	2	√	31.37 ± 4.13	28.25 ± 4.17	30.72 ± 5.19
	(a-3)	3	√	31.91 ± 2.15	29.61 ± 1.41	31.38 ± 2.70
	(a-4)	4	√	**36.15 ± 5.95**	**33.18 ± 5.12**	**36.63 ± 6.92**
	(b)	4	-	33.02 ± 5.51	30.25 ± 4.92	32.76 ± 5.28

4.2 Diffusion step (DS)

DS の入力は X，$\boldsymbol{H}_l$，$p\left(\hat{\boldsymbol{y}}_{\text{its}}\right)$，$\boldsymbol{D}^{(i)}$ である．DS の構造は，主に Forward Encoder と MRM からなる．

4.2.1 Forward Encoder

Forward Encoder では一般的な拡散モデル [6] で用いられるように，マルコフ過程に基づいてノイズを徐々に加算する．入力は X であり，出力は t 回ノイズが加算された画像 $X_t \in \mathbb{R}^{H\times W\times C}$ である．加算されるノイズは，正規分布 $q(X_t|X_0) = \mathcal{N}\left(X_t; \sqrt{\bar{\alpha}_t}X, \sqrt{1-\bar{\alpha}_t}I\right)$ に従う．ただし，I は単位行列を表す．したがって，X_t は以下のように表せる．

$$X_t = \sqrt{\bar{\alpha}_t}X + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

$$\alpha_t := 1-\beta_t, \quad \bar{\alpha}_t := \prod_{s=1}^{t}\alpha_s$$

ただし，β_t は t 回目に加算されるノイズの重みである．

4.2.2 MRM

MRM では，入力は X_t，$\boldsymbol{H}_l$，$p(\hat{\boldsymbol{y}}_{\text{its}})$，$\boldsymbol{D}^{(i)}$，出力は ITS 終了時と同様にマスク画像 $\hat{\boldsymbol{y}}_{\text{diff}}$ である．まず，X_t を用いて t 回目に加算されたノイズ $\epsilon_t \in \mathbb{R}^{H\times W\times C}$ を求める．予測ノイズ $\hat{\epsilon}_t \in \mathbb{R}^{H^{(n_b)}\times W^{(n_b)}\times C^{(n_b)}}$ は，事前学習された DDPM [6] に対して，X_t を入力して抽出する．ここで，n_b は U-Net 構造における各層のインデックスを表し，$H^{(n_b)}$，$W^{(n_b)}$，$C^{(n_b)}$ は，第 n_b 層の画像の高さ，幅，チャネル数に対応する．次に，以下の式に従い，第 n_b 層でのマルチモーダル特徴量である $\boldsymbol{H}'^{(n_b)}_{\text{seg}} \in \mathbb{R}^{H^{(n_b)}\times W^{(n_b)}\times C^{(n_b)}}$ を得る．

$$\hat{X}^{(n_b)}_{t-1} = X^{(n_b)}_t - \hat{\epsilon}_t$$

$$\boldsymbol{H}'^{(n_b)}_{\text{seg}} = \hat{X}^{(n_b)}_0 + \boldsymbol{D}^{(n_b)}$$

ここで，$\hat{X}^{(n_b)}_t$ は第 n_b 層においてノイズが t 回加算された画像 $X^{(n_b)}_t$ の予測値を表す．次に，$\boldsymbol{H}'^{(n_b)}_{\text{seg}}$ を bilinear 補完によって，$\boldsymbol{H}^{(n_b)}_{\text{seg}}$ に変形する．以下に，DS で得られる特徴量 $\boldsymbol{H}^{(i)}_{\text{diff}} \in \mathbb{R}^{H^{(i)}\times W^{(i)}\times C^{(i)}_{\text{diff}}}$ の計算過程を示す．

$$\boldsymbol{H}^{(i)}_{\text{diff}} = \begin{cases} \text{PWAM}\left(\hat{X}^{(1)}_0, \boldsymbol{H}_l\right) & i = 0 \\ \left[\boldsymbol{H}^{(i-1)}_{\text{diff}}; \boldsymbol{H}^{(i)}_{\text{seg}}\right] & i \neq 0 \end{cases}$$

ただし，$C^{(i)}_{\text{diff}}$ は特徴量の次元を表す．

最後に，DS におけるマスクの予測確率の推定値 $p(\hat{\boldsymbol{y}}_{\text{diff}})$ は $\boldsymbol{H}_{\text{diff}}$ と $p(\hat{\boldsymbol{y}}_{\text{its}})$ を用いて次のように表せる．

$$p\left(\hat{\boldsymbol{y}}_{\text{diff}}\right) = p\left(\hat{\boldsymbol{y}}_{\text{its}}\right) + f\left(\text{Conv}\left(\boldsymbol{H}^{(n_b)}_{\text{diff}}\right)\right)$$

ただし，n_c は畳み込みの回数を表す．f は Hardtanh 関数を用いた．DS の出力は，$p\left(\hat{\boldsymbol{y}}_{\text{diff}}\right)$ を閾値 0.5 で二値化した予測マスク画像 $\boldsymbol{y}_{\text{diff}}$ である．

損失関数として，ITS においては交差エントロピー誤差を，DS においては平均絶対誤差を用いた．

5. 実験設定

5.1 データセット

本研究では，SHIMRIE データセット [1] を使用した．ITS において，訓練集合はモデルの学習に，検証集合はハイパーパラメータを調整するために使用した．また，テスト集合はモデルの性能評価に使用した．DS においては，訓練集合内の一部をモデルの学習に，テスト集合のすべてをモデルの性能評価に使用した．

5.2 学習設定

ITS および DS における学習可能パラメータの総数は，それぞれ 1.23×10^8 と 2.68×10^8 であった．また，積和演算数の総数はそれぞれ 5.04×10^{11} と 1.06×10^{12} であった．学習はメモリ容量 24GB の Nvidia GeForce RTX 3090 および Intel Corei9 12900K，64GB の RAM を搭載した計算機上で行った．ITS および DS における学習時間はそれぞれ 1 時間程度および 3 分程度であった．また，1 サンプルあたりの推論時間は，約 0.53 秒であった．さらに，早期終了の条件は以下であった．

- ITS: 検証集合において mIoU が最高となるモデルをテスト集合における性能の算出に使用した．
- DS: 訓練集合において 4 エポック目以降，損失が最小となるモデルをテスト集合における性能の測定に用いた．また，損失が 51 イテレーション以上連続で減少しなかった場合，早期終了を行った．

6. 実験結果

6.1 定量的結果

表 1 に提案手法 (iii) とベースライン手法 (i)(ii) との定量的な比較を示す．各スコアは，5 回の実験における平均と標準偏差を表す．ベースライン手法として LAVT [3] 及び MDSM [1] を使用した．MDSM は OSMI タスクで良好な結果が得られている．また，LAVT は OSMI タスクと関連が深い RES タスクで高い性能が得られていることから，これらをベースラインとして選定した．

本実験における評価尺度には，mIoU，oIoU，および P@0.5 を用いた．我々は，OSMI タスクと関連が深い

(a)“Go to the bathroom on level 3 and bring me the picture frame that’s further into the room.”

(b)“Empty the tissue box in the bathroom on level one”

図 3: 提案手法の成功例．左列から，LAVT [3] での予測マスク，MDSM [1] での予測マスク，提案手法での予測マスク，正解マスク．

RES タスクにおいて標準的な尺度であるためこれらの評価指標を用いた．

表 1 より，主要尺度である mIoU において，ベースライン手法 (i)，(ii) と提案手法 (iii) は順に 24.27%，33.02%，36.15%であり，提案手法が各ベースラインに対しそれぞれ 11.88，3.13 ポイント上回った．oIoU，P@0.5 もおいても同様に，提案手法ではベースライン手法 (i)，(ii) を上回った．したがって，提案手法がすべての評価尺度において最も良好な性能であった．また，3 つの評価尺度における LAVT と提案手法の性能差は統計有意であった（$p<0.05$）．

図 3 に提案手法における成功例を示す．4 枚の画像は，左列から，LAVT での予測マスク，MDSM での予測マスク，提案手法での予測マスク，正解マスクである．ここで，予測マスクは緑，正解マスクは赤で示している．上段における命令文は “Go to the bathroom on level 3 and bring me the picture frame that’s further into the room” であり，対象物体は壁に掛けられた額縁である．ベースライン手法では，対象物体の領域のマスクが生成されていないのに対して，提案手法では対象物体の領域を正しく予測できている．また，下段における命令文は，“Empty the tissue box in the bathroom on level one” であり，対象物体はティッシュの箱である．ベースライン手法である LAVT および MDSM の場合，命令文内の “tissue” の部分しかマスクが生成されておらず，正解マスクに対して不十分な領域しか予測できていない．一方で，提案手法では “tissue box” という正しい対象物体のマスク画像を予測している．これは，画素単位で参照表現を利用する MRM が適切に機能しているからであると考えられる．

6.2 Ablation studies

Ablation study として，以下の条件を定めた．

(a) PMAM の位置の変更

(b) MRM における PWAM の有無

表 1 より，条件 (a-1)，(a-2)，(a-3) における mIoU は条件 (a-4) と比較して，それぞれ 2.24，1.66，1.04 ポイント低かった．これより，PMAM はより深い層であるほど性能向上に寄与するといえる．

また，表 1 より，条件 (b) における mIoU は，条件 (a-4) よりも 3.13 ポイント低かった．また，oIoU，P@0.5 においても，条件 (b) は (a-4) よりもそれぞれ 2.93 ポイント，3.86 ポイント低かった．以上から，マスクの洗練において PWAM によるマルチモーダル特徴量が有効に機能しているといえる．

7. 結論

本研究では，実環境における室内の画像および物体操作に関する命令文から対象物体のマスク画像を生成する OSMI タスクを扱った．

本研究の貢献を以下に示す．

- Multimodal Encoder の最終層に並列クロスモーダル特徴抽出機構 (Parallel Multimodal Attention Module）を導入した．
- MDSM [1] を拡張し，マルチモーダル入力によるマスク改善を行う Multimodal Refine Module を導入した．
- 標準的な評価尺度である mIoU，oIoU，P@0.5 に関して，提案手法はベースライン手法を上回った．

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] Y. Iioka, Y. Yoshida, Y. Wada, S. Hatanaka, et al., “Multimodal diffusion segmentation model for object segmentation from manipulation instructions,” IROS, 2023.

[2] R. Hu, M. Rohrbach, and T. Darrell, “Segmentation from Natural Language Expressions,” ECCV, pp.108–124, 2016.

[3] Z. Yang, J. Wang, et al., “LAVT: Language-Aware Vision Transformer for Referring Image Segmentation,” CVPR, pp.18155–18165, 2022.

[4] P. Wang, et al., “OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework,” ICML, pp.23318–23340, 2022.

[5] D. Baranchuk, A. Voynov, I. Rubachev, V. Khrulkov, and A. Babenko, “Label-Efficient Semantic Segmentation with Diffusion Models,” ICLR, 2022.

[6] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” NeurIPS, vol.33, pp.6840–6851, 2020.

[7] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, et al., “MAttNet: Modular Attention Network for Referring Expression Comprehension,” CVPR, 2018.

[8] L. Ye, M. Rochan, Z. Liu, and Y. Wang, “Cross-Modal Self-Attention Network for Referring Image Segmentation,” CVPR, 2019.

[9] G. Luo, Y. Zhou, X. Sun, L. Cao, et al., “Multi-Task Collaborative Network for Joint Referring Expression Comprehension and Segmentation,” CVPR, 2020.

[10] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, et al., “SeqTR: A Simple yet Universal Network for Visual Grounding,” ECCV, pp.598–615, 2022.

[11] J. Liu, H. Ding, Z. Cai, Y. Zhang, R.K. Satzoda, et al., “PolyFormer: Referring Image Segmentation as Sequential Polygon Generation,” CVPR, pp.18653–18663, 2023.

[12] A. Magassouba, et al., “Understanding Natural Language Instructions for Fetching Daily Objects Using GAN-Based Multimodal Target-Source Classification,” IEEE RA-L, vol.4, no.4, pp.3884–3891, 2019.

[13] S. Ishikawa, et al., “Target-dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots,” IEEE RA-L, vol.6, no.4, pp.8401–8408, 2021.

[14] J.D.M.-W.C. Kenton and L.K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of NAACL-HLT, pp.4171–4186, 2019.