市街地での移動指示文に基づく目標領域予測

Target Region Prediction from Navigation Instructions in Urban Areas

畑中 駿平 *1 Shumpei Hatanaka　楊 巍 *1 Wei Yang　九曜 克之 *1 Katsuyuki Kuyo　細見 直希 *2 Naoki Hosomi　翠 輝久 *3 Teruhisa Misu　山田 健太郎 *2 Kentaro Yamada
杉浦 孔明 *1 Komei Sugiura

*1慶應義塾大学 Keio University　*2株式会社本田技術研究所 Honda R&D Co., Ltd.　*3Honda Research Institute USA, Inc.

In this work, we propose a Trimodal Navigable Region Segmentation Model (TNRSM), which can handle three modalities: image, navigation instruction, and semantic segmentation masks. We validated our model on the Talk2Car-Regseg dataset. The results demonstrated that our method outperformed the baseline method in all evaluation metrics in the Referring Navigable Regions task.

1. はじめに

本研究では，移動指示文を理解し目標領域のセグメンテーションマスクを生成するタスクを扱う．例えば，“木の横に駐車してください” という指示があった場合，木の横の領域を目標の駐車位置としてセグメンテーションマスクを生成することが望ましい．

先行研究 [Rufus 21] において，over IoU は約 31%しか達成できなかったという報告があるように，目標領域のセグメンテーションタスクは困難である．一方，自動運転分野においては，モビリティがさらにナビゲーションに対する経路計画を可能にするため，移動指示文から目標領域を予測することは非常に重要である．

既存手法の [Rufus 21] では，言語および画像を入力とするバイモーダルな transformer を提案している．しかし，[Rufus 21] のモデルでは，言語と画像に対して独立に事前学習された 2 つのユニモーダルエンコーダに基づいて学習されるため特徴量の融合の充実さに欠ける．さらに，視覚モダリティのエンコーダモジュールは，マスクに関連するセグメンテーションタスクである RNR のように，セマンティックセグメンテーションレベルで画像の特徴量を考慮することができない．一方，ユニモーダルやバイモーダルではなく，トリモーダルなエンコーダ・デコーダのアーキテクチャを RIS タスクなど様々なタスクに適用する研究が注目されている [Li 21, Bachmann 22]．

本論文では，言語，画像，セマンティックセグメンテーションマスクの 3 つのモダリティを扱うトリモーダルなエンコーダデコーダモデル TNRSM (Trimodal Navigable Region Segmentation Model) を提案する．これにより，画像全体とセマンティックセグメンテーションレベルで視覚モダリティを同時に考慮することにより，目標領域のセグメンテーションマスクタスクを解決することができる．

2. 関連研究

RNR タスクは Referring Expression Comprehension (REC) や RIS タスクと関連がある．REC および RIS タスクと RNR タスクの違いは，REC および RIS は参照表現されている対象物体の領域の予測を目的としているが，RNR タスクは移動指示文で参照されているモビリティの目的地の領域の予測を目的としている．

移動指示文: “pull in behind the blue van on the left”

図 1: RNR タスクの例

REC タスクは自然言語による参照表現から画像中の対象物体の矩形領域を予測することを目的としたタスクである．[Qiao 20] では，これらの REC 手法を視覚およびテキストのモダリティをエンコードする機構に応じていくつかのカテゴリに分類し，また REC タスクに関するデータセットを評論している．一方，RIS タスクは自然言語による参照表現から，対象物体のセグメンテーションマスクを生成することを目的としたタスクである [Hu 16]．VLT [Ding 21] は画像および言語情報から複数のクエリを生成し，transformer デコーダからの出力に基づいて最適なクエリを選択する RIS モデルである．MaIL [Li 21] はインスタンスセグメンテーションを新たな入力とするトリモーダルな RIS モデルである．MaIL は，トリモーダルな transformer エンコーダおよびインスタンスセグメンテーションマスクの出力から適切なマスクを選択するモジュールをデコーダの前に導入することで VLT を上回る性能を達成した．一方，LAVT [Yang 22] は既存の RIS モデルで用いられていた視覚およびテキストのそれぞれ独立したエンコーダで構成されていた問題を解消するために新たに Pixel-Word Attention Module（PWAM）を導入した RIS モデルである．MCN [Luo 20] は，REC および RIS を同時に扱うマルチタスク学習モデルである．[Magassouba 18] らは，曖昧な命令を入力として，Place 動作の対象領域の尤度を予測するモデルであ

連絡先: 畑中駿平，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，inovation@keio.jp

る MMC-GAN を提案している．

Vision-and-Language Navigation (VLN) の分野において，屋外での環境は一般的に屋内環境よりも複雑でかつより多くの物体が存在するため，多くの研究者は屋外でのナビゲーションタスクに関心がある [Wu 21]．TOUCHDOWN [Chen 19] は，目的地までのナビゲーションを追従し，目的地付近に存在するタッチダウンを見つけるタスクである．[Hermann 20] では，実世界の画像に対する強化学習に基づく対話的な屋外ナビゲーションである Street Learn タスクを提案している．Talk2Nav [Vasudevan 21] は，VLN タスクの大規模データセットである．これは都市レベルのナビゲーション経路を含み，都市内の位置をノードとし，それらをエッジとして結ぶことによって構築されている．VLN タスクと RNR タスクの違いは，VLN がエージェントが目的地に到達するための一連の行動を生成するのに対し，RNR は画像中の目的地となる領域のセグメンテーションマスクを予測するところである．

Talk2Car [Deruyttere 19] は nuScenes データセット [Caesar 20] の上に構築されたデータセットで，セマンティクスマップや GPS，LIDAR，RADAR，360 度カメラなどの広範なセンサモダリティを含んでいる．[Chen 19,Hermann 20,Vasudevan 21] と比較すると，Talk2Car ベンチマークは，指示文を持つランドマークの矩形領域を予測することに焦点を当てた REC タスクである．また，[Chan 22] では，各領域の層をまたいだマルチモーダル情報を融合する RSD (Region-Specific Dynamic) 注意機構を提案している．さらに，Talk2Car をデータ拡張し，正面図や上面図から移動指示文が指している目的地の領域に対してアノテーションを追加し，モビリティが向かうべき目的地を予測する手法を提案している [Rufus 21,Grujicic 22]．[畑中 22] では，目標領域ではなく，目標位置の回帰タスクとして 2 次元座標を予測する UNITER Regressor を提案している．

3. 問題設定

本研究では，移動指示文およびモビリティの前方カメラ画像から目標領域を予測する RNR（Referring Navigable Regions）タスクを扱う．本タスクでは，移動指示文および画像から対象領域のセグメンテーションマスクを生成することが望ましい．図 1 に本タスクの代表例を示す．図 1 の状況において，“pull in behind the blue van on the left” という指示文が与えられた場合に，緑色の領域である目標領域を予測することが求められる．

本研究では，以下の入出力を想定する．

- **入力**：移動指示文，画像
- **出力**：目標領域のセグメンテーションマスク

また，本論文で使用する用語を以下に定義する．

- **移動指示文**： モビリティが目的地へ移動するために自然言語で書かれた案内文．
- **目標領域**： 目的地としてマスクされた領域

本研究では，モデルは経路のウェイポイントの予測をするのではなく，目標領域のセグメンテーションマスクを生成することを前提としている．また，移動指示文に対応する軌道の予測も行わない．本論文では評価尺度としては mIoU，oIoU を使用する．

4. 提案手法

提案手法のネットワークは Trimodal Encoder および Trimodal Decoder の 2 つのモジュールから構成される．

4.1 入力

ネットワークの入力 $\boldsymbol{x}$ を以下のように定義する．

$$\boldsymbol{x} = \{\boldsymbol{x}_{\mathrm{img}}, \boldsymbol{x}_{\mathrm{inst}}\}$$

$\boldsymbol{x}_{\mathrm{img}} \in \mathbb{R}^{H\times W\times 3}$ および $\boldsymbol{x}_{\mathrm{inst}} \in \mathbb{R}^{V\times L}$ はそれぞれ画像および移動指示文を表す．また，H, W, V，L はそれぞれ画像の高さ，画像の幅，指示文の語彙サイズ，最大トークン数を表す．本研究では，事前学習済みモデルである RoBERTa [Liu 19] を用いて $\boldsymbol{x}_{\mathrm{inst}}$ を言語特徴量 $\boldsymbol{h}_{\mathrm{inst}} \in \mathbb{R}^{C\times L}$ として埋め込む．ここで $C = 768$ は各トークンの特徴次元数を表す．

また，CitySpace データセット [Cordts 16] で学習された Mask2Former モデル [Cheng 22] を用いてセマンティクスセグメンテーションマスク $\boldsymbol{x}_{\mathrm{mask}}$ を生成する．具体的には，$\boldsymbol{x}_{\mathrm{img}}$ を入力とし，ゼロショット処理によりセマンティクスセグメンテーションマスクを得る．その後，CitySpace データセットに基づいて $\boldsymbol{x}_{\mathrm{mask}}$ を RGB 画像に変換する．最終的な $\boldsymbol{x}_{\mathrm{mask}}$ をの次元は $\boldsymbol{x}_{\mathrm{img}}$ と同じ次元である．

4.2 Trimodal Encoder

Trimodal Encoder は 4 つのブロックで構成されており，各ブロックは 3 つのサブブロックで構成される．具体的には Text-Image Encoder (TIE) Block および Text-Mask Encoder (TME) Block の 2 つのサブブロックから構成されている．

TIE Block は，Swin Transformer Block [Liu 21] および Pixel-Word Attention Module (PWAM) [Yang 22] から構成されている．i 番目の TIE Block は，$i-1$ 番目の TIE Block で得られた 2 つの特徴量 $\boldsymbol{h}_{\mathrm{img}}^{(i-1)}$ および $\boldsymbol{h}_{p,\mathrm{img}}^{(i-1)} \in \mathbb{R}^{H_{i-1}\times W_{i-1}c_{i-1}}$ を入力とし，$\boldsymbol{h}_{\mathrm{img}}^{(i)} \in \mathbb{R}^{H_i\times W_i\times c_i}$ を出力する．ここで，$i \in \{1,2,3,4\}$ 各ブロックのインデックスを表す．また，TME は TIE と同じ構造で構成されている．

PWAM では，$\boldsymbol{h}_{\mathrm{img}}^{(j)}$ および $\boldsymbol{h}_{\mathrm{inst}}$ を入力とし，マルチモーダル特徴量 $\boldsymbol{h}_{p,\mathrm{img}}^{(j)} \in \mathbb{R}^{H_i\times W_i\times c_i}$ を出力する．ここで，j は $j = i-1$ とする．PWAM は ViT [Dosovitskiy 21] と異なり，2 つの画像サイズの空間特徴マップ間のアテンションの重みを計算する必要がないため，メモリ使用量が少なく済む．具体的には，次の式の通りのように $\boldsymbol{h}_{p,\mathrm{img}}^{(j)}$ を得る．

$$\begin{aligned}
\boldsymbol{h}_{q,\mathrm{img}}^{(j)} &= \omega_{q,\mathrm{img}}^{(j)}(\boldsymbol{h}_{\mathrm{img}}^{(j)}) \\
\boldsymbol{l}_{k,\mathrm{img}}^{(j)} &= \omega_{k,\mathrm{img}}^{(j)}(\boldsymbol{h}_{\mathrm{inst}}) \\
\boldsymbol{l}_{v,\mathrm{img}}^{(j)} &= \omega_{v,\mathrm{img}}^{(j)}(\boldsymbol{h}_{\mathrm{inst}}) \\
\boldsymbol{g}_{\mathrm{img}}^{\prime(j)} &= \mathrm{softmax}\left(c_j^{-\frac{1}{2}}\,\mathrm{flatten}(\boldsymbol{h}_{q,\mathrm{img}}^{(j)})\boldsymbol{l}_{k,\mathrm{img}}^{(j)\top}\right)\boldsymbol{l}_{v,\mathrm{img}}^{(j)} \\
\boldsymbol{g}_{\mathrm{img}}^{(j)} &= \omega_{w,\mathrm{img}}^{(j)}\left(\mathrm{unflatten}(\boldsymbol{g}_{\mathrm{img}}^{\prime(j)\top})\right) \\
\boldsymbol{h}_{p,\mathrm{img}}^{(j)} &= \omega_{z,\mathrm{img}}^{(j)}\left(\left(\omega_{p,\mathrm{img}}^{(j)}(\boldsymbol{h}_{\mathrm{img}}^{(j)})\right)\odot \boldsymbol{g}_{\mathrm{img}}^{(j)}\right)
\end{aligned}$$

ここで，$\omega_{\cdot}$ は 1×1 の畳み込みを表す．また，$\boldsymbol{g}_{\mathrm{img}}^{\prime(j)}$ および $\boldsymbol{g}_{\mathrm{img}}^{(j)}$ はマルチモーダル特徴量を表す．このとき，flatten は 2 つの空間次元を 1 次元に展開する操作を表し，unflatten はその逆の操作を表す．同様に，$\boldsymbol{h}_{p,\mathrm{mask}}^{(j)}$ および $\boldsymbol{h}_{\mathrm{inst}}$ に PWAM を適用することによって，マルチモーダル特徴量 $\boldsymbol{h}_{\mathrm{mask}}^{(j)}$ を獲得する．

4.3 Trimodal Decoder

Trimodal Decoder では，以下の式のように $\boldsymbol{h}_{p,\mathrm{img}}^{(i)}$ および $\boldsymbol{h}_{p,\mathrm{mask}}^{(i)}$ を用いて最終的なセグメンテーションマスクを得る．

表 1: 定量的結果

Method [%]	Validation		Test	
	mIoU	oIoU	mIoU	oIoU
Baseline [Rufus 21]	17.11 ± 2.05	18.67 ± 2.29	32.28 ± 4.29	33.13 ± 4.45
Ours	**22.39 ± 1.24**	**24.63 ± 0.94**	**34.39 ± 1.43**	**37.14 ± 0.75**

$$\boldsymbol{H}^{(i)} = \begin{cases} \left[\boldsymbol{h}^{(i)}_{p,\mathrm{img}}; \boldsymbol{h}^{(i)}_{p,\mathrm{mask}}\right] & i = M \\ \mathrm{Conv}\left(\left[f_{\mathrm{up}}\left(\boldsymbol{H}^{(i+1)}\right); \boldsymbol{h}^{(i)}_{p,\mathrm{img}}; \boldsymbol{h}^{(i)}_{p,\mathrm{mask}}\right]\right) & i \neq M \end{cases}$$

M, $[\cdot;\cdot]$, f_{up} はそれぞれ Trimodal Encoder のブロック数，チャネル方向の結合，バイリニア補間によるアップサンプリングを表す．モデルの最終的な出力は，予測されたセグメンテーションマスクに対応する予測確率を 0.5 の閾値で二値化しすることで得られる $H \times W$ 次元のセグメンテーションマスク $\hat{\boldsymbol{y}}$ である．

4.4 損失関数

学習には，以下の損失関数 $\mathcal{L}$ を使用する．

$$\mathcal{L} = \lambda_{\mathrm{CE}}\mathcal{L}_{\mathrm{CE}}\left(p\left(\hat{\boldsymbol{y}}\right), \boldsymbol{y}\right) + \lambda_{\mathrm{focal}}\mathcal{L}_{\mathrm{focal}}\left(p\left(\hat{\boldsymbol{y}}\right), \boldsymbol{y}\right)$$

ここに λ_{CE} および λ_{focal} はそれぞれパイパーパラメータを表す．また，$\mathcal{L}_{\mathrm{CE}}$, $\mathcal{L}_{\mathrm{focal}}$, $\boldsymbol{y}$ はそれぞれクロスエントロピー損失関数，Focal 損失関数 [Lin 17]，正解のセグメンテーションマスク領域を表す．

5. 実験

5.1 実験設定

本実験では，モデルの評価に Talk2Car-RegSeg のデータセット [Rufus 21] を使用した．Talk2Car-RegSeg は，nuScenes [Caesar 20] および Talk2Car [Deruyttere 19] データセットに基づいて構築されたデータセットである．Talk2Car-RegSeg データセットは，Talk2Car データセットを元に，対象領域のセグメンテーションマスクを新たにアノテーションしている．まず，アノテーターが移動指示文やランドマーク情報に基づき適切な領域を対象領域としてマスクした．次に，3 名のアノテーションレビュアーがアノテーションされた領域が適切かどうかを確認を行い，過半数のレビュアーが同意するまでアノテーションされた領域を修正を行っている．本研究では，RNR タスクに適用可能な移動指示文および目標領域を含む標準的なデータセットであるため Talk2Car-RegSeg を用いた．

Talk2Car-RegSeg データセットのサンプル数は 10,019 である．各サンプルは，1,600 × 900 サイズの画像，移動指示文，および対象領域のセグメンテーションマスクで構成される．本データセットに含まれる移動指示文の語彙数，総語数，平均文長はそれぞれ 1,746, 116,501, 11.63 である．本研究では，解像度 1,600 × 900 の画像を 224 × 224 にリサイズした．

表 2: ハイパーパラメータ設定

Epoch	30
Batch size	16
Learning Rate	5.0×10^{-5}
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)
Loss Weights	$(\lambda_{\mathrm{CE}}, \lambda_{\mathrm{focal}}) = (1, 1)$
Channel size (encoder block-1)	96
Swin Transformer layer size	$\{2, 2, 6, 2\}$
# of tokens (max)	30

“pull over by that green trashcan on the right.”

(a) 正解　(b) ベースライン手法　(c) 提案手法

図 2: 提案手法の成功例．(a): 正解．(b): ベースライン手法．(c): 提案手法．

また，[Rufus 21] にしたがって訓練集合，検証集合，テスト集合に分割した．訓練集合，検証集合，テスト集合のサンプル数はそれぞれ 8,349，1,163，508 である．提案手法では各エポックごとに検証集合を用いて評価を行った．最終的な性能評価には検証集合の mIoU が最も高いときのモデルを利用した．最後に，テスト集合でモデルの評価を行った．

表 2 に提案手法のハイパーパラメータ設定を示す．提案手法のモデルのパラメータ数および積和演算数はそれぞれ 140M および 118G であった．モデルの学習には，24GB のメモリを搭載した TITAN RTX および 64GB のメモリを搭載した Intel Core i9-9820X を使用し，学習は約 4 時間 3 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 480 ミリ秒程度であった．

5.2 定量的結果

定量的結果を表 1 に示す．これらの結果は，5 回の実験から得られた平均値と標準偏差である．ベースライン手法としては RNR タスクで良好な結果が報告されている [Rufus 21] を選択した．

評価指標として，mIoU，oIoU を使用した．これらは RIS および RNR のタスクで標準的な指標である．mIoU は，大小の目標領域を同等に扱う性質を持つため，本実験の主要尺度とした．IoU($\hat{\boldsymbol{y}}$, $\boldsymbol{y}$) は 2 つのマスク間の Intersection over Union (IoU) を表す．

表 1 から，テスト集合におけるベースライン手法および提案手法の mIoU はそれぞれ 32.28%および 34.39% あった．したがって，ベースライン手法と比較して提案手法では 2.11 ポイント向上した．また，提案手法は oIoU においてもより良好な結果を得ることができた（33.13% vs. 37.14%）．

5.3 定性的結果

図 2 に成功例の定性的な結果を示す．図中の (a)，(b)，(c) は，それぞれ正解，ベースライン手法の予測結果，提案手法の予測結果を示す．画像に対する移動指示文は，“pull over by that green trashcan on the right.” であった．このとき，ベースライン手法では，対象領域に加えて，左側のゴミ箱付近も誤ってマスクしていたのに対し，提案手法では，対象領域を正しく予測することができた．

表 3: 失敗例の分類

エラー ID	詳細	# Errors
(a)	セグメンテーションマスクとして生成されない.	40
(b)	移動指示文と違う領域に予測.	31
(c)	$\hat{\boldsymbol{y}}$ と $\boldsymbol{y}$ の重なりが小さい.	17
(d)	曖昧な移動指示文による失敗.	5
(e)	アノテーションエラー.	7
Total	-	100

5.4 エラー分析

テスト集合において $\mathrm{IoU}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i) = 0$ であったサンプル数は 84 であった．表 3 に失敗例の分類を示す．テスト集合において $\mathrm{IoU}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i) = 0$ を含む IoU の値が小さい上位 100 サンプルについて，失敗の原因を分析した．失敗の原因としては大きく 5 つに分類される．

(a) セグメンテーションマスクとして生成されない: このカテゴリは，モデルが対象領域のセグメンテーションマスクを生成しなかった場合を指す.

(b) 移動指示文と違う領域に予測: このカテゴリは，移動指示文の内容やランドマークの向きに対して，モデルが誤った領域を目標領域として予測した場合を指す.

(c) $\hat{\boldsymbol{y}}$ と $\boldsymbol{y}$ の重なりが小さい: このカテゴリは，$\hat{\boldsymbol{y}}$ と $\boldsymbol{y}$ の重なりが小さい場合を指す.

(d) 曖昧な移動指示文による失敗: このカテゴリは，与えられた移動指示文に曖昧な文が含まれていたために目標領域の予測に失敗した場合を指す.

(e) アノテーションエラー: このカテゴリは，Talk2Car-RegSeg データセットにおける正解のセグメンテーションマスクまたは移動指示文のアノテーションエラーの場合を指す.

表 3 から，主なボトルネックは，モデルが目標領域のセグメンテーションマスクを生成しないことであった．これは，セグメンテーションマスクが閾値よりも下回って予測されたためと考えられる．解決策として，セグメンテーションマスクと矩形領域を同時に予測するためのデコーダモジュールを導入し，さらに追加の損失関数として対象領域に対する回帰関数を導入することが考えられる．

6. おわりに

本研究では，移動指示文およびモビリティの正面カメラ画像が与えられた場合に，目標領域に対するセグメンテーションマスクを生成する RNR タスクに取り組んだ．将来研究としては，RNR タスクのマルチタスクとして，目標領域の矩形領域を予測するデコーダモジュールを新たに導入することが考えられる．

参考文献

[Bachmann 22] Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A.: MultiMAE: Multi-modal Multi-task Masked Autoencoders, in *ECCV*, pp. 348–367 (2022)

[Caesar 20] Caesar, H., Bankiti, V., et al.: nuScenes: A multimodal dataset for autonomous driving, in *CVPR*, pp. 11621–11631 (2020)

[Chan 22] Chan, H., et al.: Grounding Commands for Autonomous Vehicles via Layer Fusion with Region-specific Dynamic Layer Attention, in *IROS*, pp. 12464–12470 (2022)

[Chen 19] Chen, H., Suhr, A., et al.: TOUCHDOWN: Natural Language Navigation and Spatial Reasoning in Visual Street Environments, in *CVPR*, pp. 12538–12547 (2019)

[Cheng 22] Cheng, B., Misra, I., Schwing, A., Kirillov, A., and Girdhar, R.: Masked-attention Mask Transformer for Universal Image Segmentation, in *CVPR*, pp. 1290–1299 (2022)

[Cordts 16] Cordts, M., Omran, M., et al.: The Cityscapes Dataset for Semantic Urban Scene Understanding, in *CVPR*, pp. 3213–3223 (2016)

[Deruyttere 19] Deruyttere, T., Vandenhende, S., Grujicic, D., Gool, L., and Moens, M.: Talk2Car: Taking Control of Your Self-Driving Car, in *EMNLP-IJCNLP*, pp. 2088–2098 (2019)

[Ding 21] Ding, H., Liu, C., Wang, S., and Jiang, X.: Vision-Language Transformer and Query Generation for Referring Segmentation, in *ICCV*, pp. 16321–16330 (2021)

[Dosovitskiy 21] Dosovitskiy, A., Beyer, L., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *ICLR* (2021)

[Grujicic 22] Grujicic, D., Deruyttere, T., et al.: Predicting Physical World Destinations for Commands Given to Self-Driving Cars, in *AAAI*, Vol. 36, pp. 715–725 (2022)

[Hermann 20] Hermann, K., Malinowski, M., Mirowski, P., Horvath, A., et al.: Learning to Follow Directions in Street View, in *AAAI*, Vol. 34, pp. 11773–11781 (2020)

[Hu 16] Hu, R., et al.: Segmentation from Natural Language Expressions, in *ECCV*, pp. 108–124 (2016)

[Li 21] Li, Z., Wang, M., Mei, J., and Liu, Y.: MaIL: A unified mask-image-language trimodal network for referring image segmentation, *arXiv preprint arXiv:2111.10747* (2021)

[Lin 17] Lin, T., Goyal, P., et al.: Focal Loss for Dense Object Detection, in *ICCV*, pp. 2980–2988 (2017)

[Liu 19] Liu, Y., Ott, M., et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv preprint arXiv:1907.11692* (2019)

[Liu 21] Liu, Z., Lin, Y., et al.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in *ICCV*, pp. 10012–10022 (2021)

[Luo 20] Luo, G., Zhou, Y., et al.: Multi-task Collaborative Network for Joint Referring Expression Comprehension and Segmentation, in *CVPR*, pp. 10034–10043 (2020)

[Magassouba 18] Magassouba, A., Sugiura, K., and Kawai, H.: A Multimodal Classifier Generative Adversarial Network for Carry and Place Tasks from Ambiguous Language Instructions, *RAL*, Vol. 3, No. 4, pp. 3113–3120 (2018)

[Qiao 20] Qiao, Y., Deng, C., and Wu, Q.: Referring Expression Comprehension: A Survey of Methods and Datasets, *IEEE-TMM*, Vol. 23, pp. 4426–4440 (2020)

[Rufus 21] Rufus, N., Jain, K., Nair, U., Gandhi, V., and Krishna, K.: Grounding Linguistic Commands to Navigable Regions, in *IROS*, pp. 8593–8600 (2021)

[Vasudevan 21] Vasudevan, A., et al.: Talk2Nav: Long-Range Vision-and-Language Navigation with Dual Attention and Spatial Memory, *IJCV*, Vol. 129, pp. 246–266 (2021)

[Wu 21] Wu, W., Chang, T., and Li, X.: Visual-and-Language Navigation: A Survey and Taxonomy, *arXiv preprint arXiv:2108.11544* (2021)

[Yang 22] Yang, Z., Wang, J., et al.: LAVT: Language-Aware Vision Transformer for Referring Image Segmentation, in *CVPR*, pp. 18155–18165 (2022)

[畑中 22] 畑中駿平, Wei, Y., 細見直希, 翠輝久, 山田健太郎, 杉浦孔明:市街地での移動指示理解タスクにおける UNITER Regressor による目標位置予測, SIG-KBS 126 回 (2022/08), pp. 34–39 (2022)