視覚言語モデルを用いたパノプティックシーングラフ生成

許　敬斌[1,a)]　柳井　啓司[1,b)]

概要

既存の PSG One-stage モデルは、ロングテール問題で低頻度の関係の予測が困難のことがある。本研究では、視覚言語基礎モデルから高級特徴を抽出し、視覚と言語の両方から学習過程を強化することで、ロングテール問題を緩和する手法を提案する。結果として、低頻度の関係の精度表示する mRecall 指標で向上し、学習時間を短縮することも出来る。

1. はじめに

Panoptic Scene graph Generation (PSG) は、Scene graph Generation (SGG) から発展し、Panoptic segmentation を組み込むことで、画像のより豊かで詳細な表現を捉える。これには、オブジェクトには「物体」と「背景」クラスの両方が含まれる。PSG は視覚と言語の橋渡しとして、視覚質問応答、画像キャプション生成、視覚推論など多くの下流アプリケーションに利用される。また、エンボディッドナビゲーションやロボットの行動計画といった関連分野にも寄与することができる。

現在の PSG 手法 [17], [18], [22] は OpenPSG データセットにのみ基づいて学習しているため、関係カテゴリにおけるロングテール問題に直面している。例えば、最も頻繁に出現する上位 3 つの関係カテゴリが全サンプルの 50%以上を占めており、希少な関係は 1%未満しか出現しない。このため、これらの希少な関係を正確に予測することが困難である。現在、ロングテール分布に対応する主な方法としては、下位関係のデータを増やしたり、損失計算でそれぞれの関係の重みを割り当て直したりすることが挙げられる。以上の方法は事前調整済みのロスが必要か、ロングテール問題に特化したデータ拡張が必要である。我々は、基本モデルを使って関係性の知識を強化し、下位関係の学習を導く。

2. 関連研究

2.1 Scene graph Generation (SGG)

SGG のデータセット [4] における関係分布はばらつきが激しいため、ロングテール問題に特化したアプローチが多いである。一般的な対処策は、データの再サンプリング [7] や損失の再重み付け [3] を通じて希少な関係の予測精度を向上させる手法である。しかし、これらの手法のほぼは二段階方式であり、エンドツーエンドで学習することはできない。再サンプリング手法はデータの修正が必要であり、損失の再重み付け手法は大量の実験を要する。さらに、これらの方法は冗長な作業を必要とし、汎用性も失われる。

2.2 Panoptic Scene Graph Generation (PSG)

伝統的な手法　SGG とは対照的に、PSG タスク [18] は、より包括的なシーングラフを作成できる。[18] から PSG タスクを提案し、two-stage のベンチマーク [4], [9], [12], [16] と one-stage のベースライン法 [18] を提出した。CATQ[17] は PSG タスクに特化したパイプラインであり、3 つの異なるクエリを使用して三つ組みの各要素のデコーディングを行う。また、SOAG モジュールはオブジェクトの位置情報を事前知識として関係学習をガイドする。最後に、3 つのクエリの情報をコンテキストに統合する。この合理的な設計と高い効果により、本研究では CATQ をベースアーキテクチャとして採用する。

偽ラベルに基づく方法　HiLo [22] は、高頻度関係と低頻度関係の両方を学習する際に、異なるネットワーク分岐に特化することによって、ロングテール問題に取り組んでいる。この方法は、前処理で追加関係データや関係強化が必要である。

損失の最重み付けの方法　Pairnet[15] は、pair proposal ネットワークを用いて物体ペアをフィルタリングする新しいフレームワークである。さらに、SeeSaw[13] を関係として、関係分類に採用されている正と負のサンプルの勾配を動的に調整する。

[1] 電気通信大学
[a)] xu-j@mm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

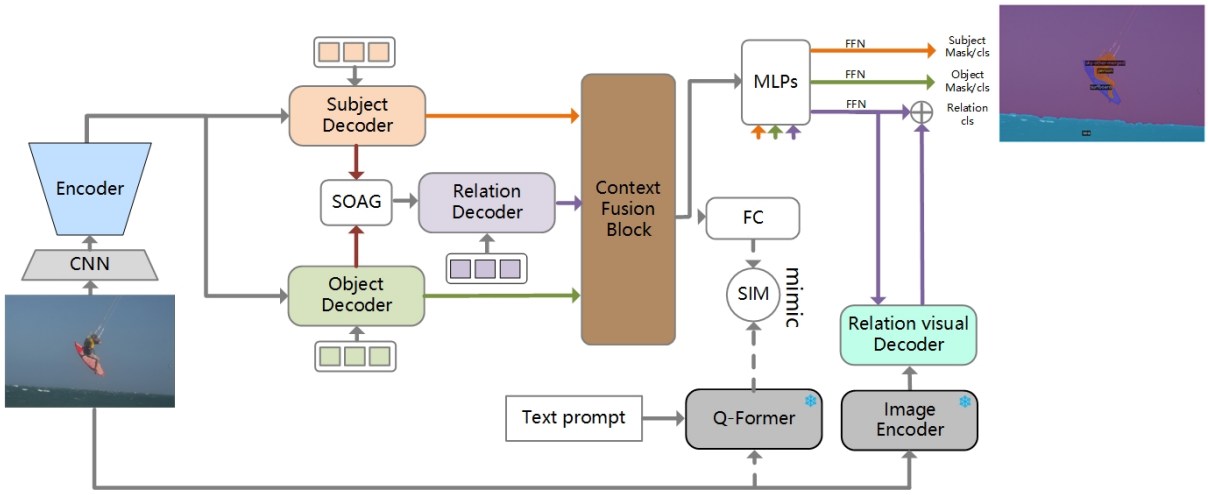

図 1 提案手法の全体図

2.3 大規模視覚言語モデル

最近、視覚言語モデル [5], [6], [11] は、物体検出 [19] や Human-Object Interaction（HOI）[10] などの視覚下流タスクにおいて驚異的な効果を発揮している。それにもかかわらず、ベースモデル上で効果的な迅速な学習を実施し、特に複雑なシナリオにおいて高次の関係を抽出することは、大規模なモデルの展開を成功させるために依然として重要である。したがって、大規模なモデルから効果的に特定の情報を抽出して利用の方法が最も重要である。

3. 手法

3.1 概要

既存の方法 [17] のフレームワークに従い、3 つのブランチを使用する手法では、3 つの異なるクエリをそれぞれトリプレット内の要素に対応する。さらに、優れた PSG 性能を達成するためには、高品質のパノプティックセグメンテーションが不可欠であるため、最新の Transformer ベースの Mask2former[2] に基づいてエンコーダとデコーダーを構築した。その上で、本研究の主要な焦点である物体ペアの空間的な位置プロンプトを用いて、視覚モデルからのターゲットプロンプト学習を行う。最後、実際のラベルを使用してコンテキスト情報の学習をサポートする。提案手法の概要図は図 1 に示す。

3.2 三つ組みアダプター

視覚特徴抽出と三つ組みでこデコーディングは CATQ に従う。画像 $X \in \mathbb{R}^{H\times W\times 3}$ を CNN バックボーンに入力し、画像特徴マップ $f \in \mathbb{R}^{K_f \times \frac{H}{32}\times\frac{W}{32}}$ を抽出する。その後、イメージエンコーダは、f を段階的にアップサンプリングして、マルチスケールの特徴量 $\tilde{f} = \left\{\tilde{f}_i\right\}_{i=1}^{4}$ を生成する。

エンコーダーから得られたグローバルな視覚特徴 V をオブジェクトデコーダーのソースインプットとして利用する。さらに、主語と目的語のブランチの埋め込み入力として、完全に独立した学習可能なクエリ $Q \in \mathbb{R}^{N_q\times D}$ を使用する。さらに、主語と目的語の位置に割り当てるために、位置誘導埋め込み $Q^p \in \mathbb{R}^{N_q\times D}$ が追加される。関係ブランチの埋め込み入力はオブジェクトブランチと同じであるが、ソースインプットが SOAG モジュールから得る。三つのブランチの出力 $Q^S, Q^O, Q^R \in \mathbb{R}^{N_q\times D}$ は、それぞれ主語、目的語、および関係の埋め込みに対応する。

3.3 クエリベースのプロンプト

BLIP2[5] は、画像特徴をビジュアルエンコーダでエンコードすることを提案しており、それをテキスト埋め込みと一緒に LLM に供給される視覚言語モデルである。本研究では、BLIP2 を用いて実験を行う。

3.3.1 視覚知識

まず、画像エンコーダーで画像 X から視覚特徴マップ $X' \in \mathbb{R}^{H\times W\times 3}$ に転換する。そして、BLIP2 の Q-former から高いレベルの視覚特徴 $V^Q \in \mathbb{R}^{N^Q\times D^Q}$ を抽出する。ここでは N^Q は Q-former のクエリ数、D^Q は Q-former の出力次元。画像エンコーダーは通常の視覚検出器よりも広範囲のデータでトレーニングされているため、画像の内容を理解する能力が高い。ここでは、視覚言語モデルの高次情報を物体ペア間の位置情報および複雑なセマンティック関係と関連付けるために、物体の位置情報と関係情報を含むプロンプトを使用する。Context Fusion Block(CFB)[17] からは、三つ組み情報を付加した関係クエリ $Q^R_{context}$ を生成し、この関係クエリが視覚プロンプトとする。さらに、関係の情報を基礎モデルの特徴トークンに関連付ける関係連結デコーダ Relation Visual Decoder ψ を設計し、視覚言語モデルの高次情報を三つ組み情報に接続する。この過程は公式 1 のように定義する。

$$Q^R_b = \Psi(V_Q, F(Q^R_{context})) \tag{1}$$

Transformer から BLIP 次元に合わせることで、線形層

Method	Backbone	*Recall@20*	*Recall@50*	*Recall@100*	*mRecall@20*	*mRecall@50*	*mRecall@100*	*PQ*
Two-stage								
IMP [16]	ResNet-50	16.5	18.2	18.6	6.5	7.1	7.2	40.2
MOTIFS [20]	ResNet-50	20.0	21.7	22.0	9.1	9.6	9.7	40.2
VCTree [12]	ResNet-50	20.6	22.1	22.5	9.7	10.2	10.2	40.2
GPSNet [9]	ResNet-50	17.8	19.6	20.1	7.0	7.5	7.7	40.2
One-stage								
PSGTR [18]	ResNet-50	28.4	34.4	36.3	16.6	20.8	22.1	13.9
PSGFormer [18]	ResNet-50	18.0	19.6	20.1	14.8	17.0	17.6	36.8
PairNet‡ [15]	ResNet-50	29.6	35.6	39.6	**24.7**	28.5	30.6	-
HiLo† [21]	ResNet-50	34.1	**40.7**	**43.0**	23.7	**30.3**	**33.1**	41.6
CATQ [17]	ResNet-50	**34.8**	39.7	40.3	20.9	24.9	25.2	35.9
CATQ+(提案手法)	ResNet-50	33.4	40.0	40.5	22.5(+1.6)	30.0(+5.1)	31.0(+5.8)	40.7

表 1 OpenPSG で PSG 手法の結果との比較。† マークは追加の関係データを使用する。‡ マークは損失の重み調整手法を用いる。

Method	*mRecall@20*	*mRecall@50*	*mRecall@100*
CATQ+	20.0	**31.1**	**34.1**
pairNet[15]	**24.7**	28.5	30.6

表 2 損失重み調整の実験

Method	*mRecall@20*	*mRecall@50*	*mRecall@100*
visual&text	**22.5**	**30.0**	**31.0**
text	20.4	24.1	24.4
w/o text	21.7	26.8	27.0

表 3 テキストガイドの実験

F を用いる。Q_b^R は基層モデルの豊な関係視覚を持つ関係クエリとする。

3.3.2 テキストガイド

言語モデルを使用して、言語的な知識を視覚モデルに効果的に移すことができる。特に関係の曖昧な定義については、視覚的なプロンプトだけでは関係の意味を学びにくい。そのため、言語モデルを用いることで関係の意味理解を深めることが期待される。特に、データセット内でロングテール分布によって精度が低下している関係について、言語モデルによる強化が有効である。

しかし、言語知識だけではモデルを混乱させる可能性がある。同一の関係は異なる画像において全く異なる物体対のクラスや位置である可能性があるため、視覚言語モデルのマルチモーダル特徴を利用して言語知識を利用する。CFB からまとめたコンテキスト特徴 $Q_{context}$ を線形層で BLIP2[5] の次元 $D^{\mathcal{Q}}$ に合わせる。バッチ画像のグラウンドトゥルース三つ組みテキスト T_{tri} と画像 X を BLIP2[5] の Q-former に入力し、言語知識を付加した視覚情報クエリ $Q_\phi \in \mathbb{R}^{B\times N^{\mathcal{Q}}\times D^{\mathcal{Q}}}$ を出力する。ここの B はバッチサイズである。そして、コンテキスト三つ組み特徴 $Q_{context}$ と Q_ϕ をコサイン類似度行列を計算して最適マッチのインデックスからマッチした特徴量を取得。最後、予測とターゲット間の L_1 ロスを計算する。この過程は以下ように定義する。視覚情報の制約を持ち言語の特徴の抽出は既存の

Method	*mRecall@20*	*mRecall@50*	*mRecall@100*
Relation	**22.5**	**30.0**	**31.0**
Object pair (1)	20.9	26.52	28.44
Object pair (2)	20.0	24.1	25.8

表 4 視覚プロンプトの実験

λ	*mRecall@20*	*mRecall@50*	*mRecall@100*
1	20.4	24.1	24.4
10	20.7	27.2	28.1
15	22.5	**30.0**	**31.0**
20	**23.4**	29.0	29.5

表 5 テキストガイドのロスハイパーパラメータの実験

クエリの特徴を混乱させない。

$$Q_\phi = Q\text{-}Former(X, T_{tri}) \tag{2}$$

$$Sim = Cos(F(Q_{context}), Q_\phi) \tag{3}$$

$$j_i = \arg\max_j Sim \tag{4}$$

$$tgt_i = Q_{\phi_{ji}} \tag{5}$$

$$loss_{match} = L_1(tgt, Q^B_{context}) \tag{6}$$

4. 実験

4.1 実験設定

データセット すべての実験はデータセット OpenPSG[18] データセットで行われる。OpenPSG は 46,572 枚トレーニング画像と 2,177 枚テスト画像、総計 48,749 枚パノプティックセグメントとシーングラフアノテーション付いている画像がある。物体カテゴリは 133 のクラス (前景 80；背景 53)、関係カテゴリは 56 のクラスにある。

評価指標 画像が与えられた時、モデルはオブジェクトセグメンテーションとインスタンス間のペア単位の関係を同時に予測する。正しいマスクの iou の閾値は 0.5 に設定されており、正しいマッチングは 3 つの要素 subject,

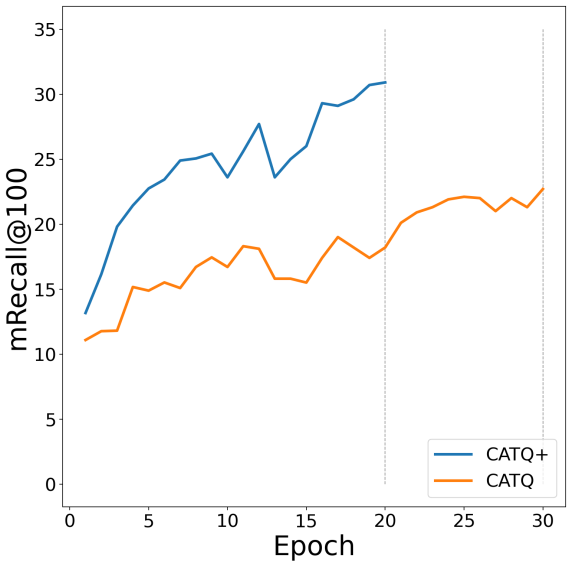

図 2 学習中の精度比較

relation, object 内のすべての要素が正しく分類されている。Recall@K[12] とm Recall@K[1] をシーングラフ生成の評価指標として、PQ はセグメンテーションの評価基準とする。

実装詳細 BLIP2 の画像エンコーダーは ViT-L/14 を用いる。Transformer の潜在次元数 $D = 256$、クエリ数 $N^q = 100$。実験には PSGTR[18] のトレーニング設定に従い、AdamW オプティマイザーで学習率は $1e-4$（バックボーンは $1e-5$)。イメージエンコーエンコーダー、三つのデコーダーは COCO[8] で事前学習済みの Mask2former で初期化する。8 枚の NVIDIA A6000 GPU でバッチサイズ 8(GPU あたりに 1 枚画像)、20 エポック (15 エポックで学習率減衰) の設定にする。

学習と推論 学習は図 1 のパイプライン通り行う。学習ロスは

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{cls} + \lambda \mathcal{L}_{match}$$

この中で $\mathcal{L}_{seg}$、$\mathcal{L}_{cls}$ はそれぞれセグメンテーションと分類ロスである。$\mathcal{L}_{match}$ はテキストガイドロス。λ はテキストガイドロスのハイパーパラメータである。各実三つ組みは、言語モデル処理のための文に変換される。例えば、「subject, relation, object」は $The < subject > is < relation > the < object >$ に変える。推論はテキストガイドの手順はなく、視覚知識だけ使用する。

4.2 ベース手法との比較

表 1 では、OpenPSG データセットにおいて、提案手法と最近の SOTA 手法を比較した。CATQ より Recall@50 と Recall@100 指標では向上したが、mRecall の 3 つのメトリクスでは 7.7%、20.5%、23.0%と上回る。

4.3 アブレーション実験

視覚プロンプト実験 本手法には画像エンコーダーのプロンプトとしてコンテキスト情報を含む関係クエリを使用した。物体ペア情報だけ使って画像エンコーダーを提示することも試した。表 4 の示すよう、2 つのプロンプト方法がある。Object pair (1) は公式 7 のようにコンテキスト情報を持った物体ペアクエリに足して 2 で割る。Object pair (2) は公式 8 のようにコンテキスト情報を持った物体ペアクエリに concat 操作する。結果的には視覚言語モデルは関係情報も理解できることを示す。

$$Q_{prompt} = (Q^S_{context} + Q^O_{context})/2 \tag{7}$$

$$Q_{prompt} = concat(Q^S_{context}, Q^O_{context}) \tag{8}$$

損失重み調整 PairNet[15] は、Seesaw Loss[14] を関係損失に応用し、正例と負例の勾配を動的に調整する。PairNet と比べるため、提案手法の関係損失にも SeeSaw Loss を適用した結果は表 2 に示す。mRecall@50 および mRecall@100 において大幅な精度向上が見られ、これらの指標において現行の SOTA である HiLo をそれぞれ 2.6%および 3.0%上回る。

テキストガイド実験 表 3 では、マルチモーダル特徴（visual&text）を用いた結果とテキストのみ (text) を使用した結果を比較すると、視覚的な情報が提供されることで、言語的な情報の有効性が高まり、これにより本来の CATQ より精度が向上する。さらに、テキスト情報のみを用いた場合、元の学習パフォーマンスは向上せず、逆に学習効果が低下することが証明された。また、テキストの情報なし (w/o text) で視覚情報のみを利用した場合、視覚情報とテキスト情報を併用した場合に比べて若干の精度低下が見られるが、テキスト情報のみを用いる場合よりも高い精度が得られることが示された。表 5 では、テキストガイドの損失ハイパーパラメーの重みを調整する結果を示す。

学習コストの軽減 提案手法が学習コストを軽減することを示すために、ベースモデルである CATQ との学習プロセスを可視化し、比較した。図 2 に示すように、提案手法は CATQ より早く収束し、最終的には性能が向上することを示した。

5. おわりに

この論文では、大規模視覚言語モデルを用いて、既存の Transformer ベース手法 CATQ を改良した手法を提案する。既存モデルの知識をプロンプトとして視覚言語モデルから三つ組み関連付ける事前知識を抽出し、三つ組み視覚情報を強化する。そして、マルチモーダル知識から関係デコーディングをガイドする学習手法を提案する。結果には、ベース手法より mRecall 指標に大幅の精度向上が達成した。

参考文献

[1] Chen, T., Yu, W., Chen, R. and Lin, L.: Knowledge-embedded routing network for scene graph generation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6163–6171 (2019).

[2] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. and Girdhar, R.: Masked-attention Mask Transformer for Universal Image Segmentation, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).

[3] Kang, H. and Yoo, C. D.: Skew class-balanced re-weighting for unbiased scene graph generation, *Machine Learning and Knowledge Extraction*, Vol. 5, No. 1, pp. 287–303 (2023).

[4] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A. et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations, *International journal of computer vision*, Vol. 123, pp. 32–73 (2017).

[5] Li, J., Li, D., Savarese, S. and Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, *International conference on machine learning*, PMLR, pp. 19730–19742 (2023).

[6] Li, J., Li, D., Xiong, C. and Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, *International conference on machine learning*, PMLR, pp. 12888–12900 (2022).

[7] Li, R., Zhang, S., Wan, B. and He, X.: Bipartite graph network with adaptive message passing for unbiased scene graph generation, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11109–11119 (2021).

[8] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: *Microsoft COCO: Common Objects in Context*, p. 740–755 (2014).

[9] Lin, X., Ding, C., Zeng, J. and Tao, D.: Gps-net: Graph property sensing network for scene graph generation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3746–3753 (2020).

[10] Ning, S., Qiu, L., Liu, Y. and He, X.: Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23507–23517 (2023).

[11] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al.: Learning transferable visual models from natural language supervision, *International conference on machine learning*, PMLR, pp. 8748–8763 (2021).

[12] Tang, K., Zhang, H., Wu, B., Luo, W. and Liu, W.: Learning to compose dynamic tree structures for visual contexts, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6619–6628 (2019).

[13] Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C. C. and Lin, D.: Seesaw Loss for Long-Tailed Instance Segmentation, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021).

[14] Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C. C. and Lin, D.: Seesaw loss for long-tailed instance segmentation, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9695–9704 (2021).

[15] Wang, J., Wen, Z., Li, X., Guo, Z., Yang, J. and Liu, Z.: Pair then Relation: Pair-Net for Panoptic Scene Graph Generation.

[16] Xu, D., Zhu, Y., Choy, C. B. and Fei-Fei, L.: Scene graph generation by iterative message passing, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5410–5419 (2017).

[17] Xu, J., Chen, J. and Yanai, K.: Contextual Associated Triplet Queries for Panoptic Scene Graph Generation, *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, pp. 1–5 (2023).

[18] Yang, J., Ang, Y. Z., Guo, Z., Zhou, K., Zhang, W. and Liu, Z.: Panoptic scene graph generation, *European Conference on Computer Vision*, Springer, pp. 178–196 (2022).

[19] Zang, Y., Li, W., Zhou, K., Huang, C. and Loy, C. C.: Open-vocabulary detr with conditional matching, *European Conference on Computer Vision*, Springer, pp. 106–122 (2022).

[20] Zellers, R., Yatskar, M., Thomson, S. and Choi, Y.: Neural motifs: Scene graph parsing with global context, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5831–5840 (2018).

[21] Zhou, Z., Shi, M. and Caesar, H.: HiLo: Exploiting High Low Frequency Relations for Unbiased Panoptic Scene Graph Generation.

[22] Zhou, Z., Shi, M. and Caesar, H.: Hilo: Exploiting high low frequency relations for unbiased panoptic scene graph generation, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21637–21648 (2023).