H-050

Graph-based Zero-Shot 物体検出におけるエッジカット手法の提案
Edge Cutting Method for Graph-based Zero-Shot Object Detection

田場 クリスティアン[1)] 山崎 禎晃[1)] 伊東 聖矢[2)1)] 大原 剛三[1)]
Christian Taba Tomoaki Yamazaki Seiya Ito Kouzou Ohara

1 はじめに

画像中の物体の位置とクラスを自動的に検出する物体検出は，コンピュータビジョン分野での重要な課題の 1 つである．近年の物体検出では，豊富な注釈付きデータの下，深層学習を用いることで，学習したクラスに対して高精度な検出能力を達成している．しかし，実世界では，訓練時に与えられていない未知クラスに属する物体が出現する状況があり得る．未知クラスに対応する単純かつ有効な方法として，未知クラスの画像サンプルの収集と注釈付けが考えられるが，膨大な注釈コストが必要となる．そこで，未知クラスも含む各クラスの意味的特徴を外部情報として利用することで未知クラスの検出を試みるゼロショット物体検出（ZSD: Zero-shot Object Detection）[1, 9] が提案されている．ZSD は，ゼロショット学習（ZSL: Zero-shot Learning）[7] の概念を取り入れた手法であり，訓練時の画像サンプルから得られる視覚的特徴と意味的特徴を関連づけるようにしてモデルを学習する．これにより，訓練時に与えられていない未知クラスに属する物体の検出を試みる．特に，検出時に既知クラスと未知クラスの両方を検出することを目的とする場合は，一般化ゼロショット物体検出（GZSD: Generalizaed Zero-shot Object Detection）と呼ばれる．

ZSD と GZSD における基本的発想は，視覚的特徴空間と意味的特徴空間におけるクラス分布の整合性が取れていれば，ある物体の視覚的特徴はその物体を表す意味的特徴と紐づくため，その物体が既知であるか未知であるかに関わらず視覚的特徴に基づいてその物体のクラスを正しく検出できるというものである．そのため，ZSD と GZSD の多くの既存研究は，視覚的特徴と意味的特徴間の整合性の改善に注力している．具体的には，視覚的または意味的特徴空間で各特徴の整合を図る手法 [9, 4]，共有する特徴空間に視覚的および意味的特徴を射影することで整合を図る手法 [10, 3] が存在する．この中で Nie らが提案した手法 [3] は，検出された物体候補領域の視覚的特徴と意味的特徴を対応付ける際に，その領域だけを考えるのではなく，シーン中の他の物体候補領域，および，利用する物体クラスとの関係をも考慮することで，2 つの特徴空間の整合性を高めている．そのために，Nie らは，検出した物体候補領域の視覚的特徴を初期ノード特徴とする視覚関係グラフ，および，利用する物体クラスに対応する単語埋め込みを初期ノード表現とする意味関係グラフを構築し，それらを統合した異種混同グラフ全体に対して，Graph Align Network（GRAN）と呼ぶ手法を用いて各ノードの特徴を周辺ノードの特徴を利用して更新している．これにより，Nie らの手法は従来手法よりも高い検出精度を達成しているが，その値は十分高いものとは言えない．

このような背景の下，本研究では，GRAN で使われている異種混同グラフを構成するグラフがいずれも完全グラフであり，関連性の低いクラスに対するノードの特徴もノード特徴の更新に使われる場合があることに着目し，グラフ構造内の不要なエッジを事前に削除することで，視覚的特徴と意味的特徴の整合性の改善を試みる．具体的には，利用する物体クラスに対応する単語埋め込みのコサイン類似度を基に，完全グラフとして構成される意味関係グラフ中のエッジを削除するエッジカット手法を提案する．評価実験では，既知および未知クラスに対する検出精度に加え，既知クラスと未知クラスに対する検出精度の調和平均を用いてモデルの検出性能を評価する．また，提案手法の汎用性を検証するため，異なるドメインから獲得された 2 つの単語埋め込みを利用した結果を比較する．

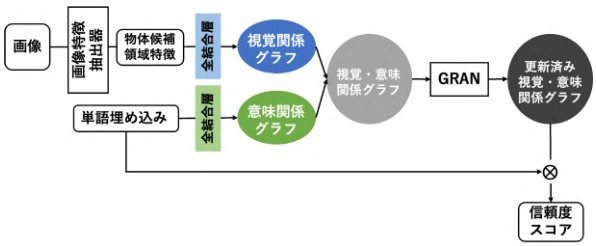


図 1: GRAN による物体検出時のクラス推定の概要

2 提案手法

提案手法は，Nie らの手法 [3] を基礎とし，単語埋め込みから獲得された物体クラスに対応する意味的特徴を初期ノード特徴としてもつ完全グラフ中のエッジを削除する．提案手法の適用対象となる GRAN による物体検出時のクラス推定の概要を図 1 に示す．本節では，GRAN [3] による物体検出の概要を説明したのち，提案するエッジカット手法について述べる．

2.1 GRAN による物体検出の概要

2.1.1 グラフの構築

GRAN [3] は，視覚関係グラフ（VRG: Visual Relational Graph）$\mathcal{G}^{vrg}$ と意味関係グラフ（SRG: Semantic Relational Graph）$\mathcal{G}^{srg}$ を結合した視覚的・意味関係グラフと呼ばれる異種混同グラフに対して，各ノードに付与された特徴表現（以下，ノード特徴）を周辺のノード特徴を用いて更新する．VRG は画像から検出された物体候補領域をノードとする完全グラフであり，各ノードには物体候補領域から抽出した視覚的特徴を学習可能な全結合層に入力して得られる特徴ベクトルが初期ノード特徴として与えられる．一方，SRG は利用する物体クラスをノードとする完全グラフであり，同様に，物体クラスに対する単語埋め込みを学習可能な全結合層に入力して得られる特徴ベクトルが初期ノード特徴として各ノードに与えら

1） 青山学院大学大学
2） 国立開発研究法人情報通信研究機構

れる．それぞれ以下のように定式化ができる．

$$\mathcal{G}^{vrg} = (X^{vrg}, A^{vrg}), \quad X^{vrg} = \{\boldsymbol{n}_i^{vrg} | 1 \le i \le N\} \quad (1)$$
$$\mathcal{G}^{srg} = (X^{srg}, A^{srg}), \quad X^{srg} = \{\boldsymbol{n}_i^{srg} | 1 \le i \le M\} \quad (2)$$

ここで，X^{vrg}，X^{srg} はそれぞれ，視覚関係グラフ，意味関係グラフにおけるノード特徴行列である．N，M はそれぞれ，画像から検出された物体候補領域数，考慮する既知・未知クラスの総数を表し，$A^{vrg} \in \mathbb{R}^{N \times N}$，$A^{srg} \in \mathbb{R}^{M \times M}$ は，視覚関係グラフ，意味関係グラフにおける隣接行列を表す．

異種混同グラフは，これら VRG と SRG の各ノード間にリンクを生成することで構築する．具体的には，VRG 中の各ノードに対応する物体候補領域に対しては，物体検出器による各既知クラスへの予測確率が得られているため，その値を重みとして VRG 中の各ノードから SRG 中のすべてのノードに向けてリンクを張り，SRG 中の各ノードから逆向きのリンクを張ることで，両グラフを双方向のリンクで結合する．なお，ここでの物体検出器は，通常の物体検出器を視覚的特徴と意味特徴を対応付けるように拡張したものである．

2.1.2 ノード特徴の更新

GRAN では，視覚・意味関係グラフのノード特徴の更新に Gated Recurrent Unit（GRU）[2] を利用する．具体的には，ある時刻 t において，各ノードは自身のノード特徴を用いて隣接ノードにメッセージ S を送信し，隣接ノードから受信したメッセージを集約した R と自身のノード特徴を GRU の入力として時刻 $t+1$ のノード特徴を生成する．送信メッセージ S と集約した受信メッセージ R はそれぞれ以下のように定義される．

$$S_i^{vrg} = \phi_{send}^{vrg}(\boldsymbol{n}_i^{vrg}) \quad (3)$$
$$S_i^{srg} = \phi_{send}^{srg}(\boldsymbol{n}_i^{srg}) \quad (4)$$
$$R_i^{vrg} = \psi^{vrg}(e_{ij}S_j^{vrg} + e_{ik}S_k^{srg}) \quad (5)$$
$$R_i^{srg} = \psi^{srg}(e_{ij}S_j^{vrg} + e_{ik}S_k^{srg}) \quad (6)$$

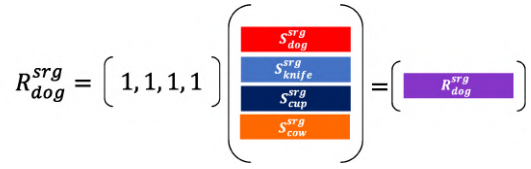
ここで，ϕ_{send}^{vrg}，ϕ_{send}^{srg} は，重み共有した学習可能な多層パーセプトロン，ψ^{vrg}，ψ^{srg} は学習可能な多層パーセプトロンである．また，e_{ij} はノード n_i，n_j 間のエッジ重みであり，両ノードがともに VRG もしくは SRG のいずれかに含まれる場合は 1 となる．このとき，GRU による時刻 $t+1$ におけるノード特徴 $\boldsymbol{n}_{t+1}$ の生成は以下のように定義される．

$$\boldsymbol{z}_t = \sigma(W_z R_t + U_z \boldsymbol{n}_t) \quad (7)$$
$$\boldsymbol{r}_t = \sigma(W_r R_t + U_r \boldsymbol{n}_t) \quad (8)$$
$$\boldsymbol{h}_t = \tanh(W_h R_t + U_h(\boldsymbol{r}_t \odot \boldsymbol{n}_t)) \quad (9)$$
$$\boldsymbol{n}_{t+1} = (1 - \boldsymbol{z}_t) \odot \boldsymbol{n}_t + \boldsymbol{z}_t \odot \boldsymbol{h}_t \quad (10)$$

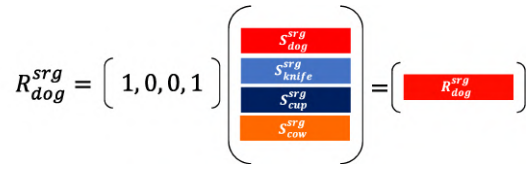
ここで W_z，W_r，W_h，U_z，U_r，U_h は学習可能な全結合層を表し，$\odot$ は要素ごとの乗算を意味する．ノード特徴は時刻 T まで T 回更新し，その結果得られたノード特徴から各物体候補領域の物体クラス予測値を算出する．

2.2 コサイン類似度に基づいたエッジカット

提案するエッジカット手法は，物体間の意味的な関連性に基づいてエッジを削除する．そのため，ともに完全グラフとして構成される視覚関係グラフ VRG と意味関係グラフ SRG のうち，SRG にのみ適用する．具体的には，SRG におけるエッジの削除は以下のように定義する．



(a) エッジカット非適用時（完全グラフ）



(b) エッジカット適用時

図 2: エッジカット適用例

$$e_{ik} = \begin{cases} 1 & \text{if} \quad \cos(\boldsymbol{d}_i, \boldsymbol{d}_k) > \theta \\ 0 & \text{otherwise} \end{cases}, \quad e_{ik} \in A^{srg} \quad (11)$$

ここで，$\boldsymbol{d}_i$，$\boldsymbol{d}_k$ は，物体クラス i，j に対する単語埋め込みを表し，$\cos(\boldsymbol{d}_i, \boldsymbol{d}_k)$ はそれらのコサイン類似度を表す．また，閾値 θ は提案手法におけるパラメータである．このように閾値 θ に基づいて SRG 中のエッジを削除することで，特定の物体クラスのノード特徴の更新に SRG 中の他のすべてノードからのメッセージを用いる Nie らの手法とは異なり，関係性の弱い物体クラスのノードから生成されたメッセージの影響を直接受けないノード特徴の更新を実現する．図 2 にエッジカット手法の適用例を示す．図 2a は SRG が完全グラフである時のノード特徴更新を示したものであり，dog に対応するノードは SRG 内のすべてのノード特徴から生成されたメッセージを集約している．これに対し，図 2b に示すエッジカット適用時のノード特徴更新では，dog と knife，cup に対応するノード間のエッジが削除されているため，それらが生成したメッセージ $S_{knife}^{srg}, S_{cup}^{srg}$ を除いて，受信メッセージを集約している．これによって，dog に対応するノード特徴は，関係性の弱い他の物体クラスに対応するノードが生成したメッセージを除いたノード特徴の更新が可能となる．

3 評価実験

3.1 実験設定

本実験では，Nie らの手法 [3] を比較手法とした．データセットには，Nie らの研究をはじめとする多くの ZSD 研究で用いられている Microsoft COCO（MS-COCO）データセット [5] を用いた．MS-COCO データセットには，82,783 枚の訓練データ画像と 45,504 枚のテストデータ画像が含まれており，各画像には，画像に含まれる物体のクラスラベルと位置を示す Bounding Box の座標が正解ラベルとして付与されている．物体クラスは 80 種類あるが，ZSD タスクではこれらを既知および未知クラスに分割する必要がある．本研究では，Rahman ら [9] の研究と同様に，既知，未知クラスがそれぞれ 65

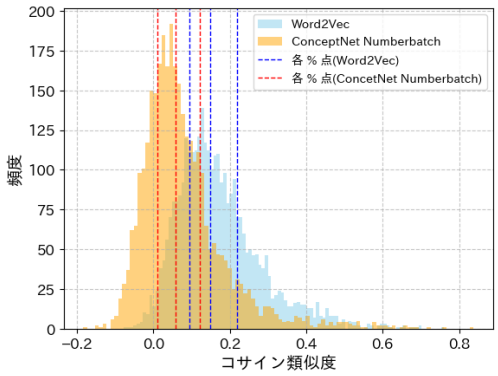


図 3: 物体クラス間のコサイン類似度分布

種類，15 種類となるように分割した．また，訓練時に未知クラスを背景として学習しないように，Bansal らの研究 [1] と同様に，訓練画像データから未知クラスに属する物体クラスが含まれる画像サンプルを削除した．単語埋め込みには，Nie らの研究でも利用されている Word2Vec [6]，および，提案手法の汎用性を検証するために ConceptNet Numberbatch [8] を用い，次元数はともに 300 とした．各単語埋め込みにおける物体クラス間のコサイン類似度分布を図 3 に示す．本実験では，提案手法のパラメータである閾値 θ として，各単語埋め込みにおける物体クラス間コサイン類似度の四分位数を用いた．図 3 では，各単語埋め込みにおける四分位数（25% 点，50% 点，75% 点）の位置を点線で示している．

3.2 評価指標

物体検出精度の評価には，GZSD 設定における物体クラスごとの適合率（AP: Average Precision）の平均である mean Average Precision（mAP），再現率（Recall），ならびに，既知および未知クラスに対する mAP と Recall それぞれの調和平均（HM: Harmonic Mean）を用いた．調和平均により，検出器が既知および未知クラスに対して偏りのない性能を有するかどうかを確認することができる．以下に，既知クラス（Seen）と未知クラス（Unseen）の評価指標 f に対する調和平均の定義式を示す．

$$HM_f = \frac{2 \times f_{seen} \times f_{unseen}}{f_{seen} + f_{unseen}} \tag{12}$$

ここで，f は，mAP または Recall であり，f_{seen} と f_{unseen} は，それぞれ既知クラスに対する評価指標値，未知クラスに対する評価指標値を表す．一方，これらの計算で用いる検出結果に対する真陽性，偽陽性は，正解ラベルの Bounding Box との重なり程度，および，既知クラスと未知クラスそれぞれに設定された最終信頼度スコアの閾値で判定され，Bounding Box の重なり具合の指標としては IoU（Intersection over Union）を使用する．2 つの領域 A，B が与えられたとき，IoU は次式のように定義される．

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} = \frac{\text{共通部分の面積}}{\text{合計面積}} \tag{13}$$

IoU の範囲は $[0, 1]$ であり，値が 1 に近いほど，2 つの領域の重なりが大きいことを意味する．本実験では，Nie らの研究 [3] 同様に，IoU，および，既知クラスと未知クラスそれぞれに設定された最終信頼度スコアの閾値はそれぞれ，0.5，0.20，0.05 とした．

表 1: Word2Vec を用いてエッジカットを適用した場合の検出精度

閾値	mAP			Recall		
	Seen	Unseen	HM	Seen	Unseen	HM
無し	31.8	**13.9**	**19.4**	59.19	61.31	60.23
25%点	**33.2**	13.2	18.9	63.84	64.06	63.95
50%点	33.0	13 .3	19.0	**64.77**	**64.67**	**64.72**
75%点	31.0	13.4	18.7	55.52	59.32	57.36

表 2: ConceptNet Numberbatch を用いてエッジカットを適用した場合の検出精度

閾値	mAP			Recall		
	Seen	Unseen	HM	Seen	Unseen	HM
無し	29.8	**14.3**	19.4	60.78	66.15	63.35
25%点	**33.2**	14.0	**19.5**	**64.29**	**69.62**	**66.85**
50%点	29.2	13.8	18.7	61.17	65.20	63.12
75%点	30.2	13.6	18.8	62.72	67.45	65.00

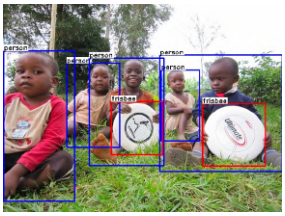
3.3 学習設定

学習設定は，比較手法である Nie らの研究に準じた．具体的には，物体検出器として Faster R-CNN を利用し，最適化アルゴリズムは確率的勾配降下法（SGD: Stochastic Gradient Descent），モーメンタム，重み減衰係数はそれぞれ 0.9，0.0001 とした．学習率および学習エポック数はそれぞれ 0.01，12 とし，8，11 エポック開始時に学習率に 0.1 を乗じた．

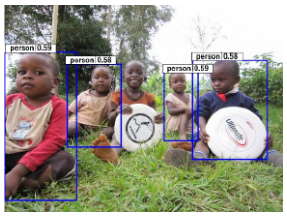
3.4 定量評価

単語埋め込みとして Word2Vec を利用した場合，ConceptNet Numberbatch を利用した場合，それぞれにおいて各閾値でエッジカット手法を適用した結果を表 1 と表 2 に示す．各表における閾値無しの結果は，完全グラフのままノード特徴を更新した場合の結果であり，そのため，表 1 における閾値なしは，比較手法とする Nie らの手法の結果に相当する．また，Seen および Unseen はそれぞれ既知クラスと未知クラスを指す．

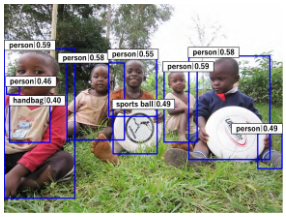
表 1 より，閾値が 50% 点のとき，比較手法に該当する閾値無しと比較して，既知および未知クラスの mAP とそれらの調和平均は僅かに低下したが，Recall がそれぞれ，5.5 ポイント，2.7 ポイント，3.7 ポイント上回る結果となったことが確認できる．表 2 からは，閾値が 25% 点のとき，未知クラスの mAP が僅かに低下したが，完全グラフである閾値無しと比較して，mAP では既知クラスと調和平均がそれぞれ 2.4 ポイント，0.1 ポイント改善し，Recall では，既知および未知クラスとそれらの調和平均が約 3.5 ポイント上回る結果となったことが確認できる．このことから，意味関係グラフ内でのエッジカットが検出精度の改善に寄与すると考えられる．ただし，単語語埋め込みの種類を問わず，エッジカット手法を適用することで，未知クラスの mAP が低下し，Recall が向上した．これは，エッジカット手法によって，未知クラスの最終信頼度スコアが閾値 0.05 を上回る予測が増加し，正しい物体クラスの検出と同時に，正解ラベルの Bounding Box とは重なりが小さい領域に対する誤検出が増加したと考えられる．また，利用する単語埋め込みによって最も精度が改善する閾値が異なることから，利用する単語埋め込みに応じて適切な閾値を設定する必



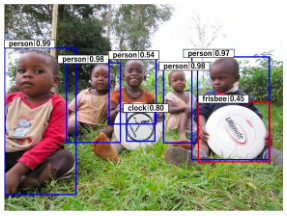
(a) 正解ラベル



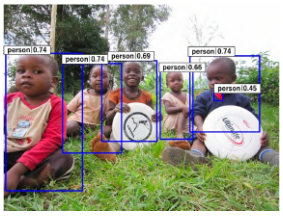
(b) Word2Vec（閾値なし）



(c) Word2Vec（閾値 50% 点）



(d) Numberbatch（閾値なし）



(e) Numberbatch（閾値 25% 点）

図 4: 各単語埋め込みでエッジカット適用時の検出例

要があることがわかる．

3.5 定性評価

次に，同一画像における検出結果を定性的に評価した．単語埋め込みとして Word2Vec，ConceptNet Numberbatch を利用してエッジカット手法を適用したときの検出結果例を図 4 に示す．各画像の青枠と赤枠は，それぞれ既知クラス，未知クラスの Bounding Box を示す．図 4b は Word2Vec を単語埋め込みとして利用し，エッジカット手法を適用しなかった場合，すなわち比較手法の検出例である．図 4b から，中央を除く person の検出に成功していることが確認できる．一方，Word2Vec を用い，閾値 50% 点のエッジカットを適用した図 4c では，handbag，sports ball，person が誤検出されているが，すべての person の検出に成功している．この結果から，エッジカット手法の適用により，モデルの真陽性と偽陽性がともに増加していることが確認できる．図 4d は ConceptNet Numberbatch を単語埋め込みとして利用し，エッジカットを適用しなかった場合の検出例である．すべての person，未知クラスである frisbee の 2 つのうちの一方を検出できていることが確認できる．しかし，中央の frisbee に対しては clock として誤検出している．一方，ConceptNet Numberbatch を用い，閾値 25% 点のエッジカットを適用した図 4e では，中央の frisbee に対する誤検出がなくなっているが，画像右側の少年の胸元における誤検出が新たに生じている．この結果から，ConceptNet Numberbatch を利用した場合においても，エッジカット手法の適用によって，モデルの偽陽性が増加してしまう場合があることが確認できる．

4 おわりに

本研究では，GRAN を用いた ZSD において，関係性の弱い物体クラス間でのノード特徴更新を抑制するために，物体クラスに対応する単語埋め込み間のコサイン類似度に基づいて意味関係グラフのエッジを削除する手法を提案した．提案するエッジカット手法を適用することで，既知および未知クラスに属する物体クラスの再現率が改善することを実験を通して定量的に確認した．また，異なるドメインから獲得された単語埋め込みを利用した場合でも精度が改善することを確認した．今後は，視覚関係グラフ内および視覚関係グラフと意味関係グラフ間のエッジも含め，各エッジが物体検出結果に与える影響をより詳細に調査し，検出結果に悪影響を与えるエッジの選択的な削除方法を検討する必要がある．

参考文献

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the 15th European Conference on Computer Vision (ECCV 2018)*, pages 384–400, 2018.

[2] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1724–1734, 2014.

[3] Nie Hui, Wang Ruiping, and Chen Xilin. From node to graph: Joint reasoning on visual-semantic relational graph for zero-shot detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2022)*, pages 1109–1118, 2022.

[4] Zhang Licheng, Wang Xianzhi, Yao Lina, Wu Lin, and Zheng Feng. Zero-shot object detection via learning an embedding from semantic space to visual space. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020)*, pages 906–912, 2020.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piootr Dollàr. Microsoft coco: Common objects in context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV 2014)*, pages 740–755, 2014.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th Neural Information Processing Systems (NeurIPS 2013)*, volume 2, pages 3111–3119, 2013.

[7] Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In *Proceedings of the 22nd Neural Information Processing Systems (NeurIPS 2009)*, pages 1410–1418, 2009.

[8] Speer Robyn, Chin Joshua, and Havasi Catherine. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st Conference Association for the Advancement of Artificial Intelligence (AAAI 2017)*, pages 4444–4451, 2017.

[9] Rahman Shafin, Khan Salman, and Porikli Fatih. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Proceedings of the 14th Asian Conference on Computer Vision (ACCV 2018)*, pages 547–563, 2018.

[10] Caixia Yan, Xiaojun Chang, Minnan Luo, Huan Liu, Xiaoqin Zhang, and Qinghua Zheng. Semantics-guided contrastive network for zero-shot object detection. *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 46(3):1530–1544, 2021.