CM-010

深層学習を用いた魚種の自動判別における背景除去の影響
Effect of background removal on automatic identification of fish species using deep learning

長谷川 達人† 益本 英明† 瀬能 宏‡
Tatsuhito Hasegawa Hideaki Masumoto Hiroshi Senou

1. はじめに

資源管理とは，科学的根拠に基づく調査によって漁獲量を管理し，水産資源の維持回復を図る施策である．適切な資源管理により，水産資源の枯渇リスクを防ぎ，我々の食糧供給や漁業従事者の長期的な雇用につながるだけでなく，海洋生態系の保全にも貢献できる．適切な資源管理を行うには，正確な資源調査が必要となるが，国内における資源調査は現状水産試験場職員らの手作業で行われることが多く，即時性，正確性に課題が残っている．

資源調査の自動化に向けて，我々は画像認識による魚種の自動判別モデルを開発している．図 1 に示すように，コンベア等を流れる漁獲物を直上から撮影し，画像認識によって漁獲された魚種や魚体長，尾数を自動記録するシステムを目指している．先行研究[1]では，ごく少量の貼付け用画像データを準備し，画像合成により訓練データセットを自動生成して Mask R-CNN を訓練することで，魚体長自動計測手法を開発した．一方，日本近海に限定しても 4,700 を超える魚種が生息していると言われ，その姿も類似するものが多いことから，正確な魚種の判別は容易なタスクではない．

他方で，深層学習の発展は著しく，近年では基盤モデルと呼ばれる汎用モデルが登場している[2]．基盤モデルとは大量かつ多様なデータを用いて事前訓練された大規模モデルであり，様々なタスクに対して汎用的に動作したり，容易に適応できるモデルを総括する言葉である．例えば，OpenAI 社の CLIP[3]は，Web 上の大量の画像とテキストのペアデータを学習する事により，大規模な Vision-Language モデルを実現している． CLIP を用いることで，追加訓練を行わずに画像とテキストプロンプトの類似度を測ることが可能となった．一方，魚種判別のように CLIP の訓練データに多く含まれないであろう詳細なクラス分類はまだまだ推定精度が低いこともわかっている[4]．基盤モデルには CLIP の他にもいくつかのモデルが提案されている．特に画像のセグメンテーションの基盤モデルとして，Facebook Research 社の Detic[5]がある．Detic は Weakly-supervised object deteciton により物体検出用のデータセットのみでなく，画像分類用の大規模かつ多クラスなデータセットを併用してインスタンスセグメンテーションモデルを訓練している．この手法により，従来手法よりも高精度かつ多様なクラスのセグメンテーションを実現している．我々の先行研究[6]において，Detic が魚領域の検出に有効に働くことを検証しており，魚種の識別を行わなければ，高精度に魚領域のみを切り出すことが可能となっている．

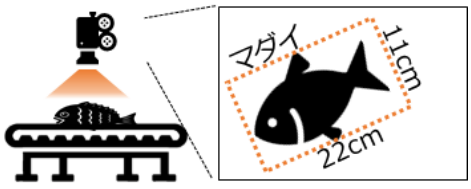

図 1　資源調査のための漁獲物自動認識

以上を踏まえ，本研究では深層学習を用いた魚種判別モデルに対して，前処理として基盤モデルを用いた背景除去を行う手法を提案する．我々が開発している魚種判別モデル[7]に対して，本手法を適用することにより，詳細魚種分類の推定精度を向上させることを目的とする．

本研究の貢献は以下の 2 点である．

- 魚分類モデルの訓練，検証データに対する前処理として，基盤モデルを用いて画像から魚領域のみを抽出する手法を新たに提案した．
- 詳細魚種分類問題において，背景除去がモデルの推定精度に与える影響を実験により明らかにした．

2. 関連研究

2.1 魚種分類に関する研究

魚種の画像認識は，古典的な手法も含めいくつか研究事例がある．Pornpanomchai らの研究[8]では，統制された環境で撮影された魚画像に対して，前処理とエッジに基づく特徴抽出を行い，Neural Network（NN）によって 30 魚種の認識を行っている．Convolutional Neural Network（CNN）を用いた魚種認識手法も多く提案されている．Alaba ら[9]は，訓練データの不均一性を調整する損失を導入したモデルを訓練する手法を開発している．Allken ら[10]は画像合成によりデータ拡張を行い，CNN を訓練する手法を提案している．ImageNet を用いて事前訓練した CNN モデルを転移学習する手法[11]等，魚種認識においても画像認識技術の応用が進みつつある．

2.2 基盤モデルに関する研究

前述の通り，基盤モデルとは大量のデータを用いて訓練された汎用な大規模モデルである．CLIP や Detic の他にも，近年様々なモデルが提案されている．chatGPT でも話題が広がっている大規模言語モデルの GPT[12]や，自然言語のプロンプトから画像を生成する DALL-E[13]などが有名である．物体検出やセグメンテーションの基盤モデルもいくつか提案されている．Meta AI 社の Segmentation Anything Model（SAM）[14]は，独自に収集された 1100 万枚のライセンス画像と，10 億以上のマスクラベルで構成された大規模セグメンテーションデータセットを用いて訓練されており，かつて無い高精度なインスタンスセグメンテーションを実現している．追加訓練無しでパーツ単位のセグメンテーションが実現できるが，2023 年現在クラスラベルの予測手法は公開されていない．Grounding DINO[15]は物体検出モデル DINO[16]を Vision-Language モデルに拡張し，Zero-

† 福井大学大学院工学研究科
Graduate School of Engineering, University of Fukui
‡ 神奈川県立生命の星・地球博物館
Kanagawa Prefectural Museum of Natural History

shot で運用可能な Open-set 物体検出モデルを実現している．これにより，CLIP のように自然言語のプロンプトを与えるだけで，多様なクラスを Zero-shot で高精度に検出できるようになっている．

2.3 背景除去に関する研究

画像分類における背景の影響についてもこれまでいくつかの研究事例がある．Xiao らの研究[17]では，背景情報が最先端の物体検出モデルに与える影響を調査している．特に，検出対象の前景が正しく写っている場合であっても，敵対的に選択した背景画像を与えることで，前景の誤検出が増加することや，より高性能なモデルは背景に依存しない傾向が強いことなどを明らかにした．本研究のように背景除去を用いることで，画像認識精度向上を図った事例もいくつかある．KC らの研究[18]では，植物の葉の分類においてエッジとモルフォロジー変換に基づく葉領域のセグメンテーションを行い，葉以外の背景を除去することで，CNN の画像分類精度が向上する手法を提案している．Rajnoha らの研究[19]でも，人物の 2 値分類において余計な背景を除くことがモデルの収束に寄与することを示した．

2.4 詳細画像認識に関する研究

画像認識とは一般物体認識を意味することが多かったが，より詳細な種の分類を行うタスクとして詳細画像認識（Fine-Grained Image Recognition; FGIR）という研究分野がある．例えば，一般物体認識では犬や猫，机，飛行機といった形状や色等が大幅に異なる物体の識別を行うのに対し，FGIR では鳥の種別や車の車種といった外見の似通った物体の識別を対象とする[20]．本研究で対象とする魚種の分類も FGIR の一種と言える．2023 年現在，FGIR で高精度を挙げているモデルとして HERBS[21]がある．HERBS は，階層的な特徴を集約するモジュールと，背景を抑制するモジュールを併用することにより，高精細な画像認識モデルを実現している．IELT[22]は，重要な領域を注視するような投票機能により高精度化を図った Transformer ベースのモデルである．これらの事例のように，最先端の FGIR モデルは背景情報を明示的には用いてはいないものの，背景の影響を軽減する機能を暗黙的にモデルに導入する傾向がみられる．

2.5 本研究の立ち位置

以上を踏まえ，本研究の立ち位置は魚種認識という FGIR タスクにおいて，背景が分類に与える影響を明らかにすることである．従来研究では，植物の葉の分類[18]で背景除去の重要性は示されているものの，魚種分類における影響は明らかではない．また，背景除去の手法は古典的な画像処理に基づいており，基盤モデルを活用するものではない．最先端の FGIR 研究においても，背景を明示的に用いる研究事例は筆者らの調査の範囲では存在していない．以上より，基盤モデルを用いた魚領域の検出により背景除去を実装し，魚種分類モデルの推定精度向上を図る点が本研究の立ち位置である．

3. 提案手法

本研究では，図 2 のように基盤モデルを用いて魚領域の抽出を前処理として行うことで，魚種分類モデルの判別精度を向上させる手法を提案する．FGIR モデルの多くが背景情報を分離するアイデアに基づいていること，近年の基盤モデルが魚領域を高精度に検出できることから着想を得て，本手法の提案に至った．

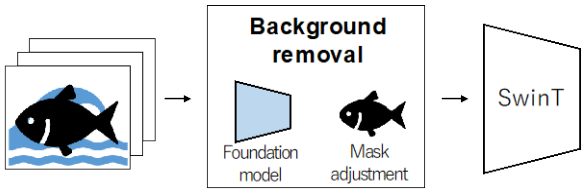

図 2. 背景除去を用いた魚種分類モデル

提案手法では，魚領域を検出する基盤モデルとして Grounding DINO[15]と SAM[14]を用いる．Grounding DINO のプロンプトを”Fish”に限定した上で魚領域の候補となる bbox を検出し，bbox を基準とした SAM により魚領域のインスタンスマスクを推定する．bbox を加工して SAM に複数候補のマスクを出力させ，後処理（Mask adjustment）により高精度なマスクを得る．後処理では複数のマスクから中央付近にある尤もらしいマスクを厳選し，モルフォロジー変換（オープニング・クロージング）によりマスクを整形する．最終的に得られたマスク$\boldsymbol{M}$と元画像$\boldsymbol{I}$，合成する背景画像$\boldsymbol{B}$のアダマール積 ⊙ 演算により，以下の式で新たな画像$\boldsymbol{I'}$を得る．

$$\boldsymbol{I'} = \boldsymbol{M} \odot \boldsymbol{I} + (\boldsymbol{1} - \boldsymbol{M}) \odot \boldsymbol{B}$$

4. 評価実験

4.1 データセット

本研究では，株式会社ズカンドットコムが運営している Web 魚図鑑[23]に登録されているデータから作成したデータセットを用いる．運営会社に許可を得た上で，2023 年 2 月頃にスクレイピングにより画像情報及び，魚種の情報を収集してデータセットを作成した．画像の解像度は統一されておらず，様々な背景の画像が含まれている．多くの画像が，釣り人により釣り上げられた魚であり，地上で 1 尾ずつ撮影されたものが比較的多い．収集した画像のうち，今回は画像分類の難易度が比較的高いサバ科の 17 魚種を対象に評価を行う．得られた画像データのうち，画像中に 1 尾のみ魚体の全体が写っている画像に目視で厳選した 601 枚の画像を用いる．前処理時の画像は 640px 程度の解像度で扱い，最終的には 224x224[px]にリサイズして画像分類に用いる．正解ラベルとして，同サービス上で提供されている種の和名を用いる．魚類学の専門家に伺ったところ，種の同定は概ね正確であるとの評価を得ている．

4.2 背景除去精度の精度評価

基盤モデルを用いた背景除去性能に関する評価実験を行う．601 枚のデータセットに対して 3 種類の背景除去手法を適用し，適切に魚領域のみを切り出せている画像数を評価する．領域検出の正解ラベルを持たないため，定性評価により概ね魚領域のみが切り出せているものを該当とした．今回，評価を行った手法は以下の 3 種である．

- Detic：先行研究[6]の知見を活かし，Detic[5]のクラスのプロンプトを”Fish”に限定した上で魚領域を検出し，魚領域以外の背景を除去する．

表 1. 種ごとの背景除去精度

	# of picts.	Detic	SAM-CLIP	GDINO-SAM
イソマグロ	19	84.20%	63.20%	94.70%
カツオ	41	90.20%	46.30%	100.00%
カマスサワラ	18	77.80%	94.40%	83.30%
グルクマ	22	90.90%	72.70%	100.00%
ゴマサバ	78	85.90%	75.60%	93.60%
マサバ	93	89.20%	71.00%	91.40%
サワラ	38	94.70%	89.50%	100.00%
ヨコシマサワラ	11	100.00%	90.90%	90.90%
スマ	43	90.70%	39.50%	100.00%
ヒラソウダ	47	97.90%	34.00%	100.00%
マルソウダ	77	94.80%	66.20%	96.10%
ニジョウサバ	9	88.90%	77.80%	88.90%
ハガツオ	28	100.00%	78.60%	96.40%
キハダ	35	85.70%	60.00%	91.40%
クロマグロ	29	93.10%	62.10%	100.00%
コシナガ	6	66.70%	66.70%	100.00%
メバチ	7	71.40%	100.00%	100.00%
総計	601	90.50%	65.90%	**95.70%**

- SAM-CLIP：SAM[14]の mask_generator を用いてマスクを生成後， CLIP[3]を用いてラベル付けを行い，魚領域以外の背景を除去する．
- GDINO-SAM：提案手法により背景を除去する．

魚種ごとの各手法の背景除去精度を表 1 に示す．表より，全体の検出精度は GDINO-SAM が最も高く 95.7%となった．575/601 枚の画像で高精度に魚領域の検出が実現できており，前処理に利用可能な推定精度であると判断した．定性評価ではあるが，Detic は魚の検出は行えるが余計な領域も検出する事例が多く見られた．SAM-CLIP は魚の一部のみを検出する事例が多く見られた．また，いずれのモデルにおいてもヒレの一部は欠損しやすい傾向が見られたが，今回は一部のみ欠落していても検出できたものとして判定している．

4.3 分類精度評価

4.3.1 実験設定

前節の結果を踏まえ，本研究では背景除去モデルとして Grounding DINO と SAM を用いる．これにより，画像に対して魚領域の bbox とセグメンテーションマスクを得ることができる．前処理で魚が検出できなかった画像（ルールベースで該当画像を識別可能）に対してはマスクを適用しないこととする．検出が成功した画像に対しては，bbox に基づいて候補領域を切り出し（Crop），マスクに基づいて背景領域を単色で塗りつぶし（Mask）することで背景除去を行う．

推論モデルとして，先行研究[7]にて開発して魚種分類モデルを用いた評価を行う．モデル構造は Swin Transformer (Tiny)[24]であり，神奈川県立生命の星・地球博物館が提供する魚類写真資料データベースに含まれる 2826 魚種，176,819 枚の画像データセットで事前訓練されている．同論文内で提案されている 2 stage Masking Adaptation（2 stage MA）と，2 stage Transfer Learning（2 stage TL）の両方のケースにおける影響を評価する．評価指標は多クラス分類における micro-F1 score を用いる．

表 2. 種ごとの魚種分類精度（2 stage MA）

	元画像	Crop のみ	Mask のみ	Crop & Mask
イソマグロ	36.8%	52.6%	52.6%	52.6%
カツオ	92.7%	85.4%	90.2%	85.4%
カマスサワラ	100.0%	100.0%	100.0%	100.0%
グルクマ	63.6%	63.6%	36.4%	36.4%
ゴマサバ	76.9%	84.6%	76.9%	78.2%
マサバ	78.5%	72.0%	78.5%	79.6%
サワラ	84.2%	73.7%	73.7%	76.3%
ヨコシマサワラ	54.5%	72.7%	81.8%	72.7%
スマ	62.8%	58.1%	53.5%	53.5%
ヒラソウダ	85.1%	87.2%	83.0%	85.1%
マルソウダ	74.0%	84.4%	88.3%	92.2%
ニジョウサバ	66.7%	77.8%	77.8%	88.9%
ハガツオ	75.0%	71.4%	78.6%	75.0%
キハダ	57.1%	51.4%	48.6%	51.4%
クロマグロ	58.6%	65.5%	41.4%	41.4%
コシナガ	66.7%	66.7%	100.0%	83.3%
メバチ	0.0%	0.0%	0.0%	0.0%
micro-F1	73.2%	**74.0%**	72.7%	73.4%

表 3. 背景色の影響

	micro-F1
black	73.4%
blue	69.7%
green	69.9%
red	71.7%
white	70.2%

表 4. マスク膨張の影響

	micro-F1
n=0	73.4%
n=1	72.2%
n=3	68.6%
n=5	67.1%
n=10	69.4%
n=15	70.7%

4.3.2 推論時の背景除去の影響

推論時の背景除去が推定精度に与える影響を調査するため，先行研究[7]で事前訓練済みの Swin Transformer に対し，2 stage MA を用いて，追加訓練しない場合の推定精度を評価する．実験の結果を表 2 に示す．表 2 の実験では背景除去後の背景色は黒で統一している．

表 2 より，従来手法（何もしない元画像）が 73.2%でサバ科 17 魚種を分類できるのに対し，Grounding DINO で検出した bbox で魚領域のみを切り取る Crop のみを適用した場合で精度が微増し 74.0%となった．一方で SAM のマスクで背景除去を行う場合精度が微減し 72.7%，Crop と Mask を併用すると 73.4%となった．魚種毎の特性を見ると，前処理によりイソマグロやマルソウダ，コシナガの再現率が向上しているが，カツオやグルクマ，サワラ，キハダ等は再現率が低下した．いずれの場合もメバチは他のマグロに誤認識される結果となった．Crop により注視領域を限定できたことで精度が向上したと考えられるが，一方で，マスク時にヒレの情報がまれに欠落したことが精度低下の要因になったと考えている．

背景色の影響を調査した結果を表 3 に示す．表より背景色は黒が最も高精度となった．事前訓練時には白背景のデータが多く含まれていたが，RGB 値で 0 を意味する黒が最も高精度を達成する結果となった．また，マスクの悪影響を低減できる可能性を考え，マスクの後処理として膨張処理を n 回適用したときの Crop&Mask の結果を表 4 に示す．

表より，余分にマスクを膨張するよりは，SAM で検出した通りに背景除去を行う方が高精度となった．適当な膨張によりヒレが除去されづらくはなるが，余分な情報が推定の悪影響につながった可能性がある．

4.3.3 訓練時及び推論時の背景除去の影響

続いて，訓練時及び推論時の背景除去が推定精度に与える影響を調査するため，先行研究[7]で事前訓練済みの Swin Transformer を転移学習した場合の推定精度を評価する．追加訓練は各魚種から K 枚のデータをランダムサンプリングした訓練データを用いて，新たに差し替えた出力層のみを Adam(lr=5e-4)で 100 エポック訓練する．

各魚種 K 枚をのぞいたデータを評価用とし，サンプリングを変えながら 10 試行した結果の平均精度を図 3 に示す．図の破線は元画像を用いた場合，点線は Mask のみ（Crop なし）を用いた場合である．また，線の色は先行研究の事前訓練モデル種別を意味する．None は事前訓練なし，imagenet は ImageNet で事前訓練されたモデル，KPM は ImageNet で訓練済みモデルを神奈川県立生命の星・地球博物館のデータセットでさらに事前訓練したモデルである．

図より，全体を通して転移学習時には Mask が有効に働いていることがわかる（点線＞破線）．特に魚種分類の事前訓練を行っていない場合（None，imagenet）で顕著な差が出ている．したがって，背景除去は追加訓練を前提とする場合において特に有効に働くことが明らかとなった．

5. おわりに

本研究では，資源調査の自動化に向けて，画像から正確に魚種認識を行う技術開発を行った．特に，認識が容易ではないサバ科の 19 魚種を対象に，基盤モデルを用いた背景除去が推定精度に与える影響を実験により調査した．実験の結果，Grounding DINO と SAM を併用する手法が，背景除去に最も適していることや，追加訓練を行わない場合背景除去よりも領域を Crop するだけの方が精度向上に寄与することを明らかにした．さらに，追加訓練を行う場合は背景除去が推定精度向上に寄与することも明らかにした．今後の課題として，検出したマスクをルールベースで背景除去に用いるのではなく，検出モデル内で動的に活用する方策を模索していきたい．

謝辞

本研究の一部は，JST ACT-X のグラント番号 JPMJAX20AJ の支援を受けたものである．また，本研究で用いたデータセットは神奈川県立生命の星・地球博物館及び，株式会社ズカンドットコムより提供を受けたものである．ここに謝意を表す．

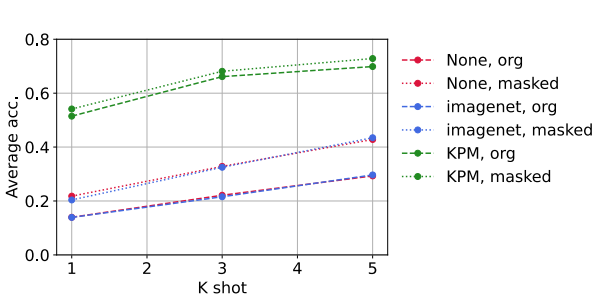

図 3. K shot 時の推定精度（2 stage TL）

参考文献

[1] 長谷川達人, 田中基貴, "水産資源管理に向けた Mask R-CNN による Few-shot 魚体長認識", 情報処理学会論文誌 コンシューマ・デバイス＆システム（CDS）, Vol. 12, No. 2, pp. 38-48, (2022).
[2] Rishi Bommasani, et al., “On the Opportunities and Risks of Foundation Models”, arXiv, 2108.07258 (2021).
[3] Alec Radford, et al., “Learning Transferable Visual Models From Natural Language Supervision”, In proceedings of the PMLR, pp. 8748-8763 (2021).
[4] 田中基貴，長谷川達人, “CLIP を用いた説明可能な Few-Shot 魚種分類手法”，第 85 回全国大会講演論文集，Vol. 2023, No. 1 (2023).
[5] Xingyi Zhou, et al., “Detecting Twenty-Thousand Classes Using Image-Level Supervision”, In proceedings of the ECCV, pp. 350-368 (2022).
[6] Tatsuhito Hasegawa, Motoki Tanaka, “Validation of the effectiveness of Detic as a zero-shot fish catch recognition system”, In proceedings of the ICIAE, pp. 1-5 (2023).
[7] 長谷川 達人, 近藤 圭, 瀬能 宏, “様々な産地市場に転用可能な魚種の自動判別モデル”, マルチメディア，分散，協調とモバイル(DICOMO2023)シンポジウム, (2023).
[8] Chomtip Pornpanomchai, et al., “Shape- and Texture-Based Fish Image Recognition System”, Agriculture and Natural Resources, Vol. 47, No. 4, pp. 624-634 (2013).
[9] Simegnew Yihunie Alaba, et al., “Class-Aware Fish Species Recognition Using Deep Learning for an Imbalanced Dataset”, Sensors, Vol. 22, No. 21, pp. 1-18 (2022).
[10] Vaneeda Allken, et al., “Fish species identification using a convolutional neural network trained on synthetic data”, ICES Journal of Marine Science, Vol. 76, Issue 1, pp. 342–349 (2019).
[11] Jaisakthi Seetharani Murugaiyan, et al., “Fish species recognition using transfer learning techniques”, International Journal of Advances in Intelligent Informatics, Vol. 7, No. 2. pp. 188-197 (2021).
[12] Alec Radford, et al. "Improving language understanding by generative pre-training." (2018). https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
[13] Aditya Ramesh, et al., “Zero-Shot Text-to-Image Generation”, In proceedings of the PMLR, pp. 8821-8831 (2021).
[14] Alexander Kirillov, et al., “Segment Anything”, arXiv, 2304.02643 (2023).
[15] Shilong Liu, et al., “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection”, arXiv, 2303.05499 (2023).
[16] Mathilde Caron, et al., “Emerging Properties in Self-Supervised Vision Transformers”, In proceedings of the ICCV, pp. 9650-9660 (2021).
[17] Kai Yuanging Xiao, et al., “Noise or Signal: The Role of Image Backgrounds in Object Recognition”, In proceedings of the ICLR, pp. 1-28 (2021)
[18] Kamal KC, et al., “Impacts of Background Removal on Convolutional Neural Networks for Plant Disease Classification In-Situ” Agriculture, Vol. 11, No. 9, pp. 1-16 (2021).
[19] Martin Rajnoha, et al., “Image Background Noise Impact on Convolutional Neural Network Training”, In proceedings of the ICUMT, pp. 1-4 (2018).
[20] Tsung-Yu Lin, et al., “Bilinear CNN Models for Fine-grained Visual Recognition”, In proceedings of the ICCV, pp. 1449-1457 (2015).
[21] Po-Yung Chou, et al., “Fine-grained Visual Classification with High-temperature Refinement and Background Suppression”, arXiv, 2303.06442 (2023).
[22] Qin Xu, et al., “Fine-Grained Visual Classification Via Internal Ensemble Learning Transformer”, IEEE Transactions on Multimedia, doi: 10.1109/TMM.2023.3244340 (2023). (Early Access)
[23] 株式会社ズカンドットコム，“Web 魚図鑑”, https://zukan.com/fish/ (accessed 2023/6/14).
[24] Ze Liu, et al., “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”, In proceedings of the ICCV, pp. 10012-10022 (2021).