無人航空機における画像検索のための深層学習に関する研究

理工学研究科後期課程第 21 期

電子情報工学系専攻　情報知能メディア学教育研究分野

ブイ ドク ヴェト

令和 6 年 3 月

論文審査委員主査	准教授	佐藤　浩
論文審査委員	教授	中村　康弘
論文審査委員	准教授	久保　正男
論文審査委員	准教授	岩切　宗利
論文審査委員	准教授	山下　倫央

研究成果の概要

近年，ドローン技術の急速な発達により，無人航空機（Unmanned Arial Vehicle，以下 UAV）に関する研究が大きく進展し，軍事や製造業だけでなく，一般社会への普及の兆しが見られる．特に，将来の利用者増を見据えて，タスクを自動で完了できる完全自律型無人機の技術開発に注目が集まっている．この為には，従来人間のオペレータが限られた観測情報を基に行っていた煩雑な作業（位置推定や実行するべきタスクの決定や実行可能なタスクの決定等）の重要な課題を限られた時間内で行うことが必要となる．例えば，スタジアムや都市等の広範囲な領域での監視と捜索等に自律型 UAV を運用するには現在位置の特定や障害物検出，監視対象の認識といった下位レベルのタスクを人を介さずに効率的に行う必要がある．

一方で，深層学習の登場により画像処理に関する研究は大きく進歩した．高度な深層学習アーキテクチャーが提案され，人間顔負けの機能を実現できるようになった．これを運用するには高い計算パワーが不可欠だが，近年のハードウェアの発展によって生み出された高性能な CPU と GPU を使えば航空機内においても，大規模な並列処理を短時間で実施可能になっている．現在では，より高度な深層学習を使った画像処理技術を運用できるプラットフォームも徐々に普及しており，これを発展させることによって自律化に向けて大きな寄与ができると考えた．

このようなことから，本論ではこれからの UAV に求められる画像処理技術の中でも次の画像検索に関わる課題に取り組んだ．画像検索とは，入力画像等の内容や特徴を解析し，それに基づいてデータベースから検索して関連する画像を取得するプロセスである．画像検索は，様々な分野で幅広く応用されており，特に，UAV では，「画像ベースの場所推定」及び「オブジェクト再同定」という二つの主な画像検索タスクがある．

画像ベースの場所推定とは，キャプチャされた画像を緯度経度といった地理情報が付帯した 2D もしくは 3D 画像データと照合することにより，その正確な場所を推定することである．特に UAV における画像ベースの場所推定では，主に斜めからの撮影された画像を使うクロスビュー場所推定が用いられる．この手法では，まず入力の UAV からの撮影画像に対し，事前に準備した環境の衛星画像データベースを検索し，最も類似している衛星画像を見つける．次に，出力された衛星画像には地理情報が付与されているため，これらの情報を基にして，現在

位置のさらなる推定を行う．オブジェクト再同定とは，異なるカメラビュー，異なる時間，あるいは異なる条件でキャプチャされた同じ物体や人物を識別して同定するタスクである．最近，このタスクを UAV 上で実装する研究の萌芽が見られ，UAV スワームによる監視及び追跡能力が大きく向上している．監視カメラシステムと異なり，オブジェクト再同定の能力を持つ UAV スワームは，徒歩や車両ではアクセスできない地域に飛行する能力を持っているため，様々な地域で偵察・探索性能を向上させることができる．これらのタスクは UAV の運用，特に将来の自律型 UAV システムにおいて非常に役に立つが，現状では以下の二つの課題がある．

一つ目の課題は, 精度の改善である．UAV における画像検索に関する研究は初期段階であるため，関連するデータセットが少なく，従来のアプローチでは簡単なアーキテクチャに留まる．画像検索のタスクは，異なる視点間で外観が大きく異なるため，同一のオブジェクト，または UAV ビュー画像と衛星ビュー画像をマッチングすることは簡単な問題ではない．また，これらのタスクを UAV で実施する場合には，視点だけではなく高度も異なるため，対象のオブジェクトも小さく見え，認識精度も落ちる傾向がある．

二つ目の課題は, 複数のタスクを実行できる相乗性を持った画像検索技術である．実際の UAV の運用では，多数のタスクを同時に処理することが必要とされるが，現在の UAV の画像検索システムでは，検索タスク毎に専用の画像検索システムが必要で，同一の画像検索システムを複数のタスクに活用することはかなり難しい．そのために，知識の共有が難しかった．特に，本研究で取り上げる UAV における画像検索のタスクについては，これまでの各タスクの対象ドメインに大きいな違いがあるため，これらのタスクを単純に統合するだけでは性能が悪化するのみである．

そこで，本研究では，無人航空機における画像検索の課題を中心として，深層学習の能力を利用することより，これらの課題を改善することを目的とする．まずはじめに，無人航空機で不可欠な画像検索タスクである，クロスビュー場所推定やオブジェクト再認定のタスクについて，それぞれを高精度に達成するモデルを提案する．次に，計算機リソースを有効に活用して，複数のタスクを同一の画像検索技術で遂行するマルチパーパス性を持った画像検索モデルを提案する．

本研究の成果には，以下の三つがある．まず，UAV におけるクロスビュー場所推定のために，二つの深層学習モデルを提案し，双方とも既存研究と比べて大幅に精度が向上したことを実験により確認した点．次に，UAV におけるオブジェクト再同定の精度向上のために，新しい損失関数を提案し，その有効性を実験により確認した点．最後に，UAV における三つの画像検索を解決できる単一の深層距離学習ベースのモデルを提案し，実験を通してその有用性を確認した点である．

以上の通り，本研究での提案モデルはこれまでの UAV における画像検索の問題を解決し，将来の UAV 開発に展開することができる有望な証左を与えた．

目次

図目次

表目次

第 1 章

序論

1.1 研究背景

1.1.1 背景

近年，ドローン技術の急速な発達により，無人航空機（Unmanned Arial Vehicle，以下 UAV）に関する研究は大きく進歩し，軍事や製造業だけでなく，一般社会への普及の兆しが見られる．そして，人間の介入なしにミッションを実行できるという完全自律型 UAV システムは様々な分野で注目されている [1]．完全自律型 UAV システムは，通常手段ではアクセスできない領域へのアクセスや探査能力を持ち，目的設定から実行，完了までを全て自律的に遂行できるため，災害救助の初期段階や地域偵察ミッションに対し特に活躍が期待されている．

UAV の完全な自律化には，限られた観測情報を基に，現在位置の推定，実行するべきタスクの決定及び実行可能なタスクの決定等の重要な課題を人間の介入なしに限られた時間内で実行することが必要となる．これは UAV の知覚と制御の根幹であり，高度な画像処理技術が非常に重要で，UAV の多様なタスクの性能向上の為のキーポイントである．例えば，優れた画像処理を導入できれば，自律型 UAV の運用に不可欠な下位レベルタスク（位置特定 [2]，地図作成 [3]，障害物検出 [4] や非常着陸検出 [5] 等）の精度を高められる．

画像処理技術の正確さ以外の側面での画像処理技術開発も自律化には不可欠である．例えば近年萌芽が見られる複数の自律型 UAV を群として運用する「UAV スワーム」では，精度面以外での画像処理技術の開発が重要であることが明らかになっている [6]．例えば自律型 UAV スワームにはスタジアムや都市等の広範囲な領域での監視や捜索を行って，屋外イベントで効果的にセキュリティを確保することが期待されている [7]．ところが現状では，画像関連の各タスクには異なるアルゴリズムやモデルが必要となるため， UAV の限られた計算リソースと電力エネルギを著しく消費してしまうことが明らかになってきた．このことは，将来の自律型 UAV システムを実現するには，正しく画像処理する能力だけではなく，関連する複数の画像

処理タスクを効率よく実装したり運用する能力も高める必要がある．

これを可能にする技術の一つは深層学習である．深層学習とは多層ニューラルネットワークとその学習技術で，これは従来考えられなかったほど高度な画像認識を可能にし，現在でも新しい技術とその応用が提案され続いている [8]．例えば，特徴学習ができる畳み込みニューラルネットワーク [9]（Convolutional Neural Network, 以下 CNN）や Vision Transformer [10] 等の深層学習モデルが，画像処理の多くのタスク（画像認識，セグメンテーション，物体検出等）において驚異的な成果を上げている．さらに，近年のハードウェアの急激な発展により, リアルタイム処理の能力が改善され，画像処理を用いるプラットフォームも徐々に多様化している [11]．特に，ロボット向けの GPU が発展し，GPU が搭載された UAV が登場した．このように，画像処理の深層学習技術は徐々に UAV に導入され，多様なアプリケーションやシナリオでの UAV の有用性を高めている．

そこで，本論ではこれからの無人航空機に求められる画像処理技術の中でも極めて重要な画像検索に関わる課題に取り組んだ．画像検索 (Image Retrieval)[12] は画像処理技術の 1 つで，入力画像の内容や特徴を解析し，それに見合った画像をデータベースやコレクションから取得するプロセスである．この技術は，様々な応用分野で幅広く活用されている．最も一般的な応用例の一つはインターネット上での画像検索である．ユーザーは検索エンジン (Google, Bing 等) に画像をアップロードして，関連する画像を探すことができる．検索エンジンは入力画像の特徴を活用して，類似した画像を提供する．また，オンラインショッピングプラットフォームでは，ユーザーが商品の写真をアップロードしたり，似た商品を探すために画像検索を使用し，ユーザーが欲しい商品を見つけやすくしている．

UAV や自動運転車等の自律型システムでは，この画像検索技術を自己位置推定や，周囲の環境を認識し，道路及び物体等を識別するために使用する [13]．特に，UAV においては次の二つの画像検索タスク — 画像ベースの場所推定とオブジェクト再同定 — が知られている．

画像ベースの場所推定: 画像ベースの場所推定 (Image-based Geo-Localization) [14][15][16][17] とは，撮影された画像を地理情報がある 2D か 3D 画像データと照合することにより，画像を撮った場所を推定することである．この問題は画像検索の問題として扱われている．画像ベースの場所推定は，地図学，地理空間解析，観光，拡張現実等の様々な分野で応用される技術であり，重要な役割を果たしている (Fig. 1.1).

画像ベースの場所推定には，大きく分けると 3 つの種類がある (Fig. 1.2)：

- Cross-time Geo-Localization(クロスタイム場所推定): 異なる時間帯（午前・午後等）で撮影された画像から同じ場所の画像を見つけることである．
- Cross-view Geo-Localization(クロスビュー場所推定): 異なるプラットフォーム（衛星・UAV・自動車等）で撮影された画像から同じ場所の画像を見つけることである．
- Cross-modal Geo-Localization(クロスモーダル場所推定): 異なるドメイン（2D マップ・

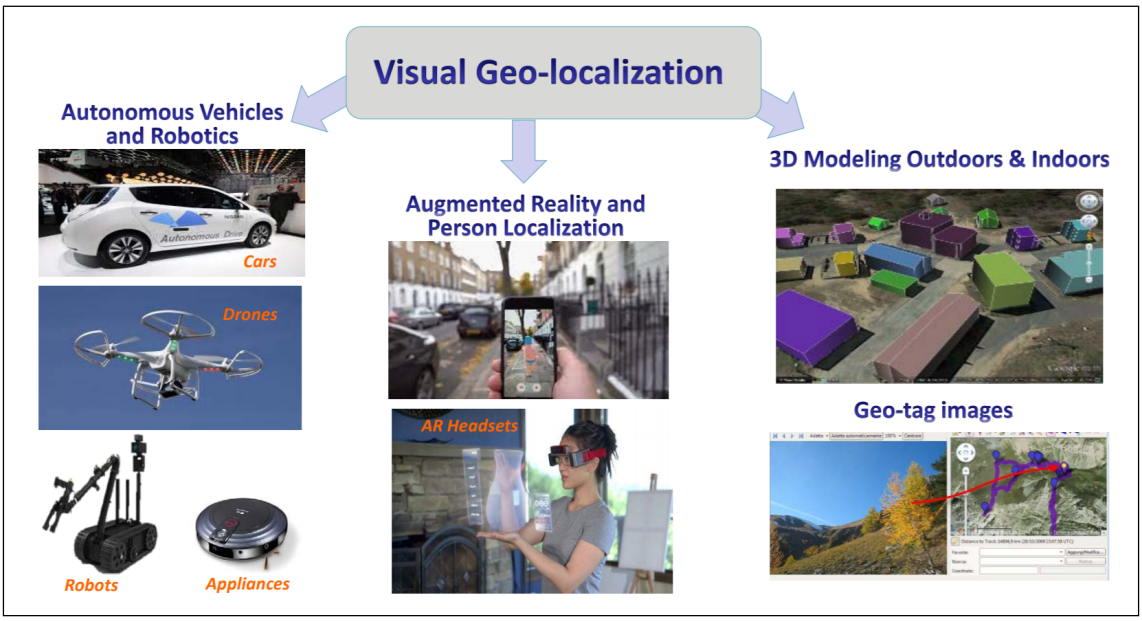

Fig.1.1　Applications of Image-based Geo-Localization (https://www.sri.com/computer-vision/cvpr-2021-tutorial-on-cross-view-and-cross-modal-visual-geo-localization)

3D マップ等）での画像から同じ場所の画像を見つけることである．

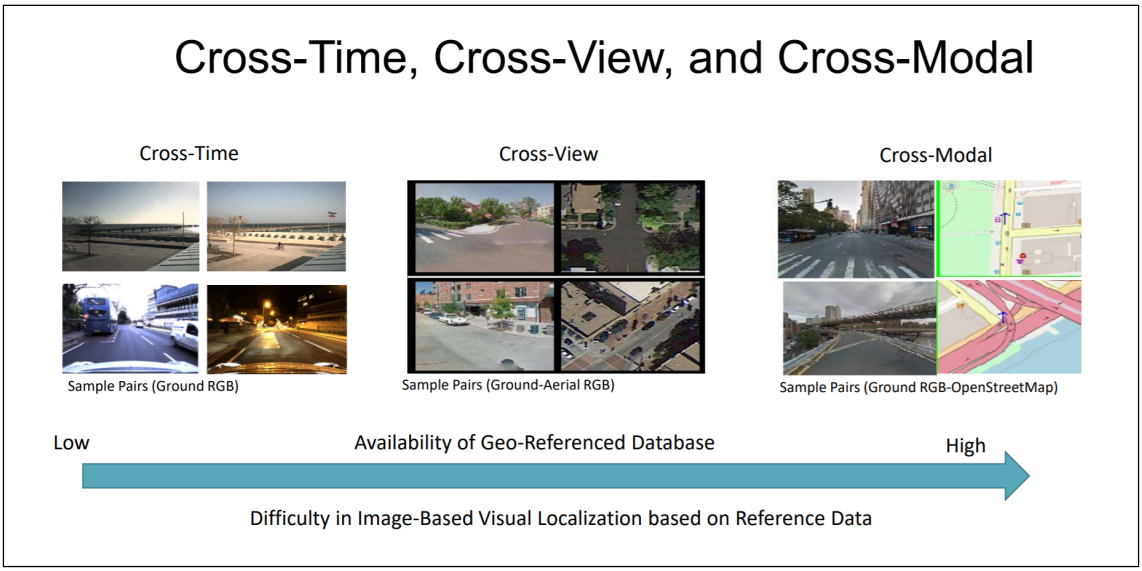

Fig.1.2　Three types of Image-based Geo-Localization (https://www.sri.com/computer-vision/cvpr-2021-tutorial-on-cross-view-and-cross-modal-visual-geo-localization)

現在，UAV における画像ベースの場所推定には，主にクロスビュー場所推定が応用されている [18]．具体的には，入力された UAV 画像に対し，事前に準備された衛星画像データベー

スを検索し，最も類似している衛星画像を見つける．出力された衛星画像には場所情報が付けられているため，入力画像を撮った場所を推定することが可能となる．これは GPS が信頼できない，または GPS にアクセスできない環境では非常に役立つ．また，これらの情報は UAV 上に搭載したセンサーの情報と合わせて，UAV の位置を推定することが可能になる [2]. 例えば，[19] では，クロスビュー場所推定から出力した複数の衛星画像により，UAV の位置を推定することが可能と主張した (Fig. 1.3)．Fig. 1.4 に示すように，クロスビュー場所推定 (Cross-view Geo-localization) により，UAV のグローバルポーズを推定することが可能になる．さらに，Visual Odometry [20][21][22]（画像のシーケンスを解析してカメラの位置と向きを決定するプロセス）と合わせることで，GPS データを用いることなしに， UAV の位置 (UAV pose output) を推定することができる．

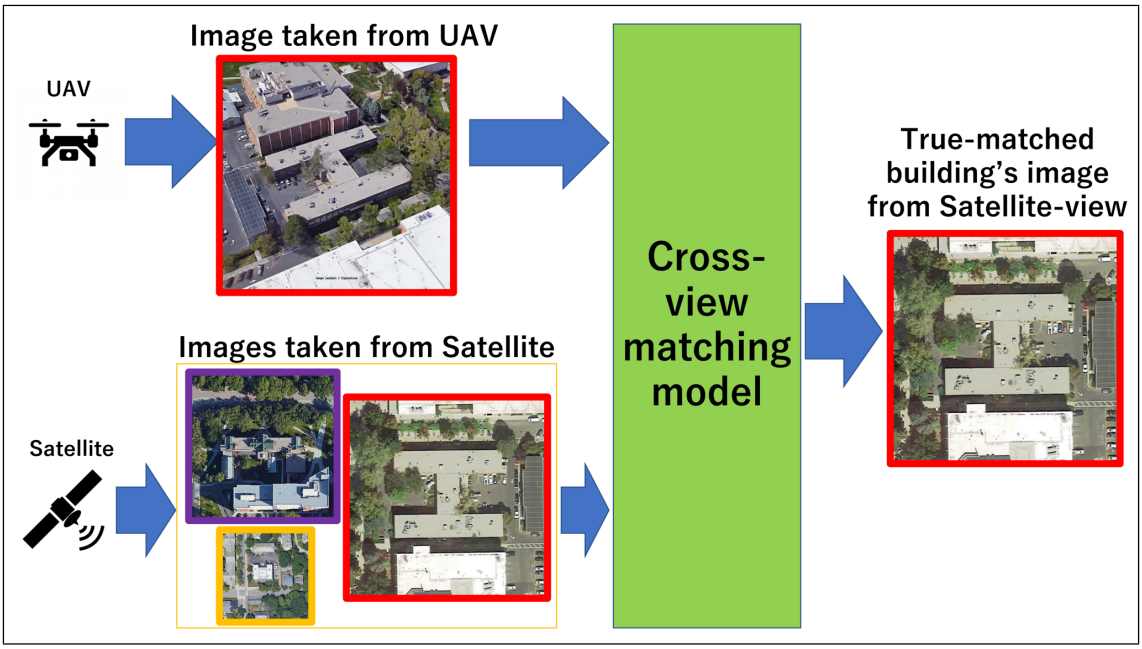

Fig.1.3 Example of Cross-view Geo-Localization for UAV

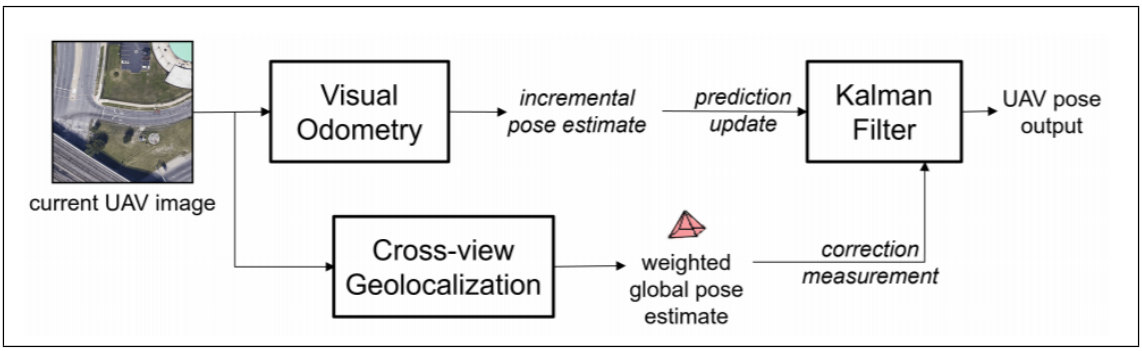

Fig.1.4 Example of Cross-view Geo-localization for estimating UAV pose

オブジェクト再同定:

オブジェクト再同定 [23][24]（Object Re-Identification, Object Re-ID） とは，異なるカメラビュー，異なる時間，あるいは異なる条件で撮影された同じ物体や人物を正確に識別して，同一物を同定することである．例えば，人物再同定 (Person Re-Identification，Person Re-ID [25]) や車両再同定 (Vehicles Re-Identification，Vehicle Re-ID [24]) といった画像検索タスクは，自動エリア監視システムのコア技術で，公共の安全とセキュリティにとって非常に重要である．Fig. 1.5 に人物再同定の例を示す．ここでは，カメラ 1 番 が撮った人物の画像に対し，他のカメラで取得した人物の画像の中に検索し，最も類似する画像を出力する．ただ，位置が固定された従来の監視システムではその盲点を突かれることも多く，近年，これらの業務を複数の UAV で代替しようという動きがある．監視カメラシステムと異なり，UAV スワームは広範な地域で運用できるので，人間がアクセスできない地域に飛行する能力を持っているため，現地での検索性能が向上する．ただ，これには他の UAV が撮影したオブジェクトの同一判定が行えることが不可欠で，またフェールセーフ上の観点からエッジでも行えることが望ましい．このようにオブジェクト再同定機能の UAV への実装は喫緊の課題となっている [26][27][28].

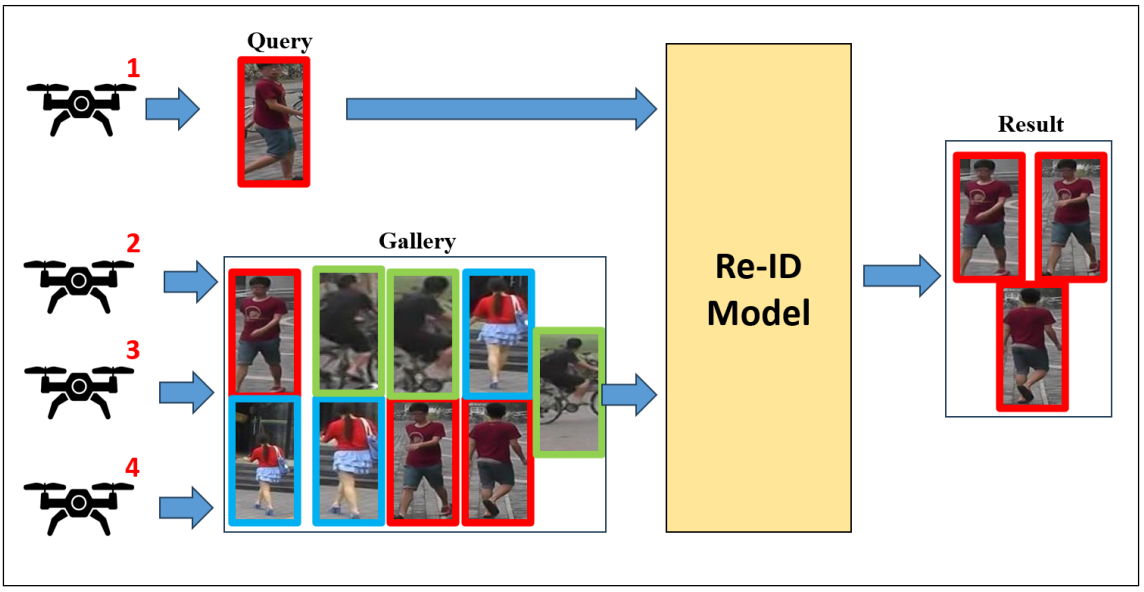

Fig.1.5　Example of Person Re-Identification for UAV

次に，UAV における場所推定及びオブジェクト再同定の応用をより詳しく理解するために，本研究の想定状況を定義する．

1.1.2　本論技術の自律型 UAV での意義

本節では無人航空機で上記の画像ベースの場所推定とオブジェクト再同定の利用例を説明し，これらの技術が無人航空機において重要であることを説明する．

無人航空機の設定

このシナリオで使用される無人航空機 (UAV) とは，中型以上のマルチロータードローンで，例えば一般的な 4 つのロータを持つクアッドコプターであれば，100m までの高さで飛行ができる．これらのドローンは安定性，操縦性及び静止能力で知られ，よく捜索・救助活動に使用されている．通常 30 分以上の長時間飛行が可能であり，詳細な画像とビデオを撮影できる高解像度のカメラを搭載している．このカメラは 4K 以上の解像度を持ち，捜索エリアの明瞭な視覚を撮影できる．必要に応じて追加のセンサーや装置を統合できるペイロード容量を持っている．例えば，低光条件や厳しい状況での捜索能力を向上させるために，サーマルカメラや赤外線センサーを追加することができる．また，この UAV にはリアルタイムで撮影された画像を処理できるために，高度なコンピュータビジョンアルゴリズムを実装できるハードウェアの搭載が必要となる．例えば，近年 NVIDIA 社が開発したモバイルデバイス用の GPU の Jetson を利用ができる．Jetson による組込みシステムは，高度な画像処理のアルゴリズムを実装することができ，一般的なドローンに搭載することができる．

想定する利用状況と意義

まず，本研究では市街地帯・山岳地帯・密林地帯等での捜索・救助活動を想定している．捜索の対象オブジェクトは，例えば都市部に災害した場合には行方不明の市民であり，山岳地帯や密林地帯の場合には行方不明のハイカーである．このような広い範囲での捜索・救助活動を実施する任務には多くの人員とリソースが必要であるが，将来，完全自律型 UAV システムを導入することで捜索や救助活動に関与する人間の負担を軽減することができる．UAV が機動性に優れており，空中からの視点を持ち，地上や他の交通手段がアクセスできないような困難な地形や環境でも迅速に移動することができる．山岳地帯や森林や水域等，人間には難しい場所に到達し，捜索範囲を拡大することができる．特に，完全自律型のスワームドローンは複数のドローンが連携して動作することができる．これにより，広範囲の地域を同時にカバーすることができ，大規模な捜索エリアの迅速な探査や効率的な情報収集が可能となる．また，自律性により，UAV はプログラムされたミッションに基づいて自律的に行動し，迅速で効率的な捜索や救助活動を実行できる．

また本論では，GPS を十分に利用できない状況を想定している．通常の UAV，特に完全自律型 UAV は GPS の情報を利用して自己位置推定を行っているが，民間用 GPS システムより軍事用 GPS システムの方が高性能であるため，軍事用 GPS システムを持っていない国々は高精度ミッションに UAV を利用することができない．そのため，高精度ミッションを行うには精度を高める何らかの追加機能が求められる．また，たとえ高精度な GPS を持っても，都市部の高い建物や構造物は，GPS 信号の反射や遮蔽を引き起こし，また建物から跳ね返された信号を受信することでマルチパスフェージングによって位置精度が低下することもある．また，

激しい雨や雪，濃霧等の悪天候条件によって，GPS 信号が減衰され，位置推定の精度を低下させることがある．このような場合，気象条件によっては，GPS に加えて UAV に搭載された他のセンサーの情報で自己位置推定を実施できることもある．

このような場合，複数のセンサの結果を統合する技術の一つは SLAM (Simultaneous Localization and Mapping) で，もう一つが本論で取り組んだ画像検索技術である．SLAM は GPS データ，加速度やジャイロ，カメラ画像，レーザ距離計（LIDAR），RGB-D センサ，ステレオカメラ等のセンサから得られた情報を利用して自己位置推定（Localization） と環境地図作成 (Mapping) を同時に実行して精度を高める方法である．通常，自己位置推定と地図作成は反復作業になり少なくない計算量が問題になる上，地図には 3D 環境モデルを使うことも多く，どの UAV でも利用できるわけではない [29]．もう一つの方法が提案手法が属する画像処理のアプローチである [30][31]．これは深層学習モデルをリアルタイム性を持って運用できる計算機資源さえあれば良いので，UAV におけるナビゲーションするための将来的に有望なアプローチと考えられる [32].

画像検索技術の導入方法

このような場合，画像検索技術の場所推定及びオブジェクト再同定は具体的に以下のように活用でき，UAV の運用向上が期待できる．

- **クロスビュー場所推定**：UAV を飛行する前に，捜索する予定の全体地域の衛星画像を UAV のメモリに保存する．この衛星画像には各場所の地理的情報が付けられる．このように，完全自律型 UAV が飛行している時，カメラで撮った画像を事前にメモリに保存された衛星画像の場所と比較し，クロスビュー場所推定によって UAV の現在の場所を確認することができる．
- **オブジェクト再同定**：UAV を飛行する前に，捜索の対象オブジェクト（人物・車両）UAV のメモリに保存する．このように，完全自律型 UAV システムを使用する際，UAV が取得したオブジェクトの画像と比較し，オブジェクト再同定によって対象オブジェクトを確認することができる．

1.2　UAV の画像検索に関する課題と本研究の目的

1.2.1　課題

上記で述べた画像検索のタスクは，UAV の運用，特に将来の自律型 UAV システムにおいて非常に役に立つが，これらに関する研究はまだ初期の段階である．次の二つの課題が明らかとなった．

一つ目の課題は, 精度の改善である．UAV における画像検索に関する研究はまだ初期の段階であり，関連するデータセットが少ないため，大量のデータが必要な深層学習は，この理由で困難になるとみられる．そして，異なる視点での外観は大きく異なるため，同一のオブジェクト（人物・車両）のマッチング及び UAV 画像と衛星画像のマッチングは簡単な問題ではない．また，これらのタスクを UAV で実施する場合には，視点だけではなく高度も異なり，オブジェクト（人物・車両）の画像も小さく見え，認識精度も落ちる傾向がある．しかし，従来の手法が複雑な特徴処理法を利用せずに，簡単なアーキテクチャを設計し，まだ高い精度を達成することができないため，これから改善の余地がある．

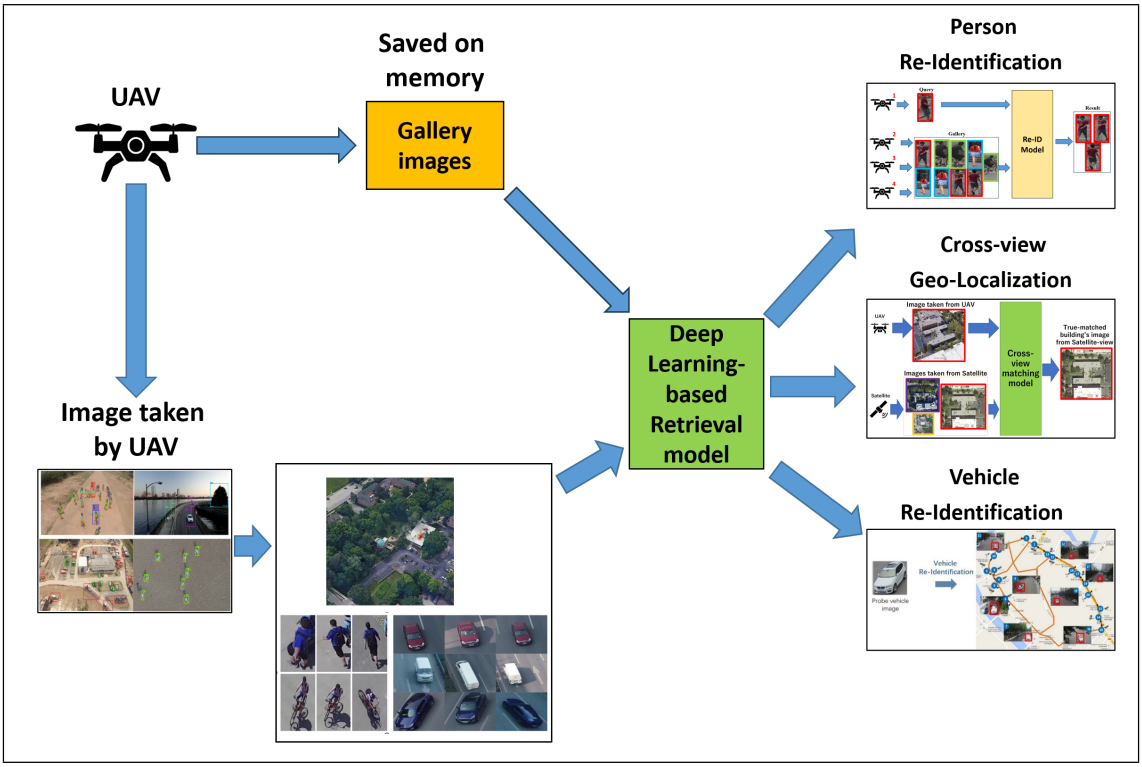

Fig.1.6　Example of Multitask Retrieval model for UAV

二つ目の課題は, 画像の情報だけで複数のタスクを実行できるモデルの開発である．上記で

述べたように，実際の UAV の運用では，目的の異なる多数のタスクを処理することが必要である．搭載されたセンサだけで可能な限り情報を取得し，多数のタスクを同一のモデルで実行することは UAV をはじめとするロボット研究における困難な課題である．従来，自律型 UAV システムでは，複数の画像検索タスクを解決できることが望まれてきたが，それらのタスクに個別に対応するモデルを作成する必要がある．この場合，タスクの増加に応じて計算するリソースも増加するため，リソースが限られた UAV に対しては困難になる．例えば，Fig. 1.6 に UAV におけるマルチパーパス[*1]の画像検索モデルの応用例を示す．想定する状況として，事前に撮影された各タスクの Gallery 画像と UAV がその場で撮影した画像を合わせて，上記のタスク（クロスビュー場所推定・人物再同定・車両再同定）を解決する場合を考える．人物や車両の捜索ミッション等に UAV を利用する場合，マルチパーパスな UAV なら，1 つのモデルで自律的に場所推定・人物再同定・車両再同定ができる．このように，監視・捜索用に個別の UAV スワームシステムを開発するためのリソースも削減できると考えられる．

しかし，複数のタスクを解決できるモデルの開発は，かなり難しいと考えられる．今回の UAV における画像検索のタスク（クロスビュー場所推定・人物再同定・車両再同定）については，各タスクの対象ドメインに大きいな違いがあるため，これらのタスクを統合することは非常に困難な課題と見られる．例えば，従来の監視システムによるオブジェクト再同定の研究では，2 つの異なる再同定タスク（人物再同定・車両再同定）に分割し，それぞれのタスクの独自の課題に対処できるネットワークを設計した研究であった．人物・車両を再同定するとそれらのオブジェクトの場所も推定できる深層学習モデルは，UAV システムの運用に効果的に貢献すると考えられるが，これまで複数の画像検索タスクの課題を取り込んだ研究は見られない．

1.2.2 本研究の目的

ここまでを踏まえ，本研究では無人航空機における画像検索の課題を中心として，深層学習の能力を利用することで，これらの課題を改善することを目的とする．この目的を達成するために，まずクロスビュー場所推定やオブジェクト再同定のタスク，それぞれを高精度に達成するモデルを提案する．加えて，計算機リソースを有効に活用して，多様で規模の大きな深層学習モデルの画像検索の利用を削減するために，マルチパーパスの画像検索モデルを提案する．

本研究の成果は，以下の三つである．一つ目は，UAV におけるクロスビュー場所推定のために，二つの深層学習モデルを提案し，双方とも既存研究と比べて大幅に精度の向上ができたことである．二つ目は，UAV におけるオブジェクト再同定のために，高精度学習のために新しい損失関数を提案し，ベンチマークデータセットで最先端技術を乗り越えた結果ができたこ

[*1] 一般的なマルチタスクには，タスクは並行して実行されるという前提があるが，本研究分野では，一つのモデルで複数のタスクをこなすことを「マルチタスク」と呼び，並行性は問わない．混乱をさけるために，以下では本研究分野の意味でのマルチタスクを「マルチパーパス」と呼称する．

とである．三つ目は，UAV における三つの画像検索（クロスビュー場所推定・人物再同定・車両再同定）を解決できる単一の深層距離学習ベースのモデルを提案したことである．

1.3　本論文の構成

本論文は以下の章により構成される．

第 1 章: 序論
本論文の概要を述べる．

第 2 章: 関連技術
本研究に関連する深層学習技術について紹介する．

第 3 章: 無人航空機のためのクロスビュー場所推定
無人航空機のためのクロスビュー場所推定について，精度改善のために二つの深層学習ベースモデルを提案する．

第 4 章: 無人航空機のためのオブジェクト再同定
無人航空機のためのオブジェクト再同定について，精度改善のために新しい損失関数を提案する．また，マルチパーパスの画像検索のための深層学習ベースアーキテクチャも提案する．

第 5 章: 結論と今後の展望
本研究の成果をまとめ， 今後の展望について述べる．

第 2 章

関連技術

本章では, 本研究に関連する技術について解説する．画像検索に関する技術を紹介した後, 画像検索のための主要な深層学習ベースのアプローチ（特徴学習や距離学習）を詳しく説明する．

2.1　画像検索について

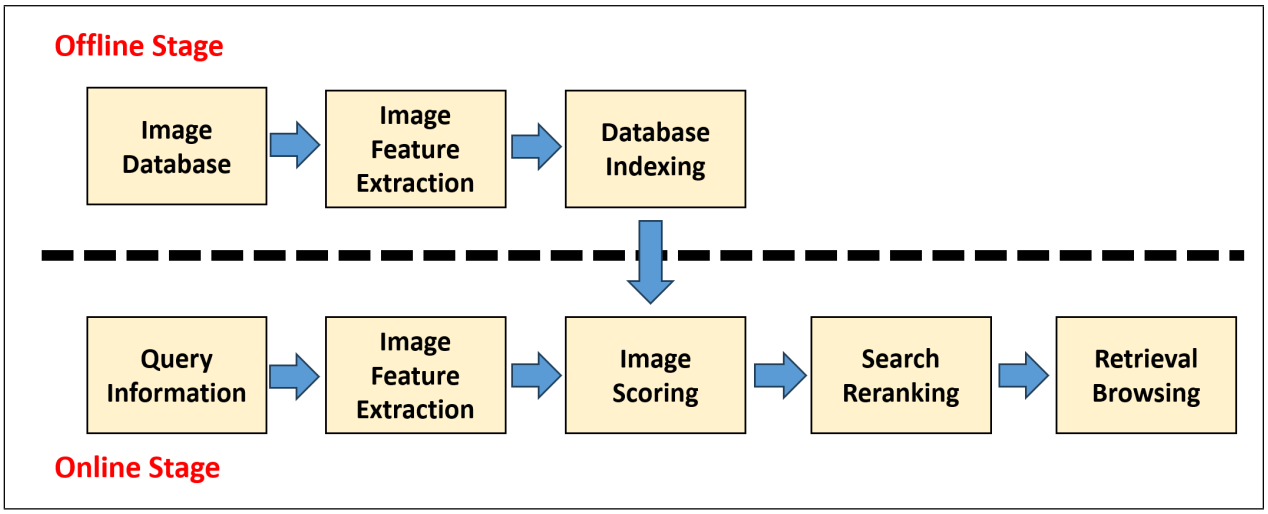

Fig.2.1　Flowchart of Image Retrieval

画像検索は，画像処理分野で 20 年以上にわたり注目されている課題である [33]．画像検索の一般的なフローチャートを Fig.2.1 に示す．基本的には，Offline Stage と Online Stage という 2 つのステージがある．Offline Stage では，データベース (Image Database) を構築し，各データベースの画像に対し特徴抽出し (Image Feature Extraction)，次にインデックス化する

(Database Indexing)．Online Stage では，いくつかの操作が必要である：Query 画像の特徴抽出 (Query Information・Feature Extraction)，類似度評価 (Similarity Measure) 及び検索のリランキング (Search Reranking) がある．画像特徴抽出モジュール (Image Feature Extraction) は，Offline Stage と Online Stage の両方で共使用されている．

画像検索システムを構築する場合，以下の要素に注目する：

- **画像の特徴抽出 (Image Feature Extraction)：** このフェーズでは，入力画像（以下，Query 画像）及びデータベースの画像（以下，Gallery 画像）を解析して，各画像から画像の特徴の情報が含まれる特徴ベクトルを抽出する．
- **類似度評価 (Similarity Measure)：** このフェーズでは，コサイン類似度 (Cosine Similarity) やユークリッド距離 (Euclidean distance) 等を利用して，Query 画像の特徴ベクトルと各 Gallery 画像の特徴ベクトルの間の距離（類似度）を計算する．ベクトル $\mathbf{p} = (p_1, p_2, \cdots, p_n)$ とベクトル $\mathbf{q} = (q_1, q_2, \cdots, q_n)$ のコサイン類似度 $\cos(\mathbf{p}, \mathbf{q})$ 及びユークリッド距離 $d(\mathbf{p}, \mathbf{q})$ は以下のように計算する：

$$\cos(\mathbf{p}, \mathbf{q}) = \frac{\sum_{i=1}^{N} p_i \cdot q_i}{\sqrt{\sum_{i=1}^{N} p_i^2} \sqrt{\sum_{i=1}^{N} q_i^2}} \tag{2.1}$$

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{N} (q_i - p_i)^2} \tag{2.2}$$

  ここでは，N はベクトルの次元数である．
- **リランキング (Reranking)：** このフェーズでは，求めた類似度に基づいて，ランキングリストを作成し，Query 画像と最も類似した画像を出力する．結果をさらに最適化したい場合，先行研究はいくつかのリランキングアルゴリズム [34][35][36] を利用している．

画像検索に関する研究は，主に「画像の特徴抽出」に注目しアルゴリズムやモデルを構築する．初期の研究では，手動により特徴量（例：色のヒストグラム，テクスチャの記述子，形状表現等）に焦点を当てた [33] が，実世界の画像の複雑さと変動性に対処するのが困難になる．そして，古典的な特徴抽出法 (SIFT [37], SURF[38] 等) が登場し，画像のための頑健な記述子が提供され，検索することが徐々に良くなったが，まだ精度の改善が必要である．その後，機械学習技術の導入が一般的になり，Bag-of-Words [39] や Support Vector Machine (SVM) [40] 等の使用が広まった．近年では，深層学習が画像検索を革新し，畳み込みニューラルネットワーク（CNN）や深層距離学習が生のピクセルデータから直接特徴を学習するデータ駆動の手法 [12][41] を提供している．これらのアプローチにより，よりコンテキストを考慮した意味のある画像検索システムの実現ができた．そのため，本研究は近年に人気になった深層学習の技

術を用いて，UAV における画像検索の問題を解決する．

次節では，深層学習の特徴学習アプローチによく利用される畳み込みニューラルネットワークや Vision Transformer を紹介する．

2.2　特徴学習

本節では，深層学習の最も利用されている特徴学習モデル，畳み込みニューラルネットワークについて紹介する．また，近年注目されている自己注意機構を持つ Vision Transformer について解説する．

2.2.1　畳み込みニューラルネットワーク

畳み込みニューラルネットワーク（Convolutional Neural Network, 以下 CNN）は，格子状のトポロジを持つデータの処理に使われる特殊なニューラルネットワークである [42]．格子状のトポロジーを有するデータの例としては，等時間間隔で取得したサンプルが 1 次元に配列された時系列データや，ピクセルが 2 次元に配列された画像データが考えられる． CNN は実用的なアプリケーションにおいて極めて成功している．畳み込みとは特殊な線形変換であり，CNN は，少なくともどこか一つの層で行列の掛け算の代わりに畳み込みを利用するニューラルネットワークのことである．CNN のネットワーク構造は，これまでのニューラルネットワークと同様に，多数のレイヤを積み重ねて作られる．CNN の場合，新たに「Convolution 層（畳み込み層）」と「Pooling 層」が登場する． 従来の CNN の構造は，基本的に「畳み込み層 - Pooling 層 - 全結合層」という流れである．Fig. 2.2 に CNN の構造の例を示す．

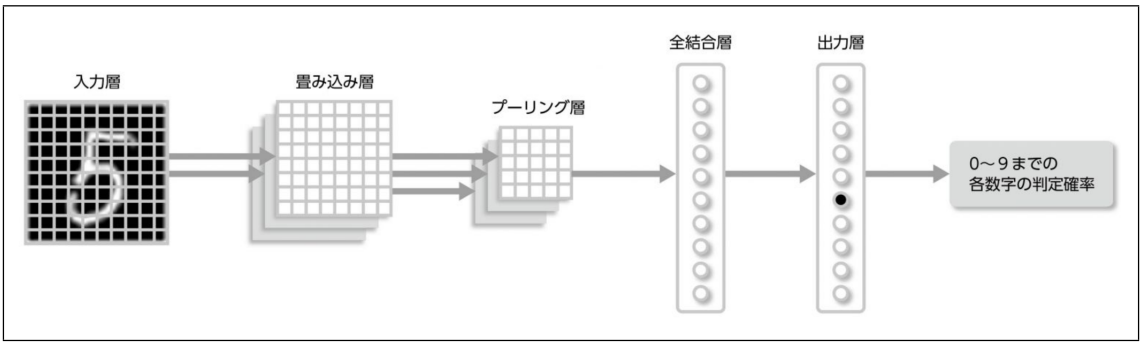

Fig.2.2　Example of CNN
(https://www.imagazine.co.jp/畳み込みネットワークの「基礎の基礎」を理解す/)

畳み込み層

畳み込み層で行う処理は「畳み込み計算」である [42]．畳み込み計算は，画像処理でいうところのボカシや細線化といった「フィルタ計算」に相当する．フィルタのパラメータが CNN のニューロン間の「重み」である．コンピュータ上で扱える画像の主な形式には，ベクタ画像形式とラスタ画像形式がある．ベクタ画像は，線や曲線といった図形集合の重ね合わせをデー

タとして保持した形式である．一方，ラスタ画像はピクセルと呼ばれる画素の 数値を格子状に並べたデータを保持した形式である．Fig. 2.3 に，入力データに対して畳み込み計算を適用した例を示す．入力データは縦・横の 2 次元のデータで，フィルタも同様に縦・横方向の次 元を持っている．フィルタのウィンドウを一定の間隔でスライドさせながら適用する．それぞれの場所で，フィルタの要素と入力の対応する要素を乗算し，その和を求める（積和演算と呼ばれる）．

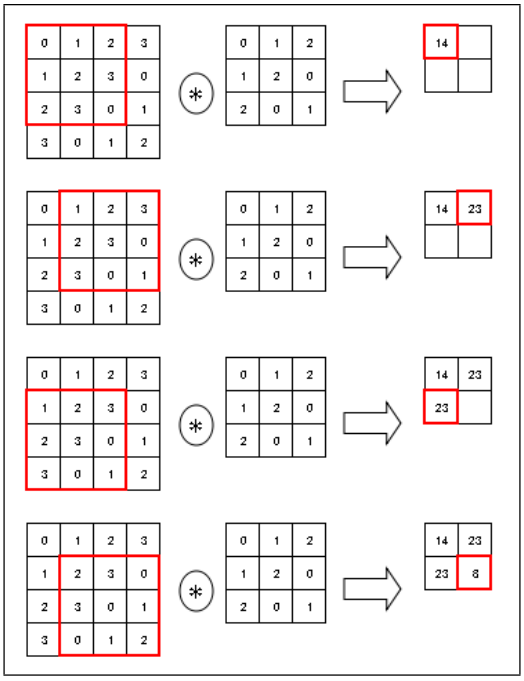

Fig.2.3　Example of a Convolutional process in CNN
(https://qiita.com/nvtomo1029/items/601af18f82d8ffab551e)

その後，結果を出力の対応する場所へ格納する．この操作を全ての場所で行うことで，畳み込み演算の出力を得ることができる． 積和演算は Fig. 2.4 に示すように，それぞれの場所でフィルタ要素と入力の対応する要素を乗算し，その和を求める．

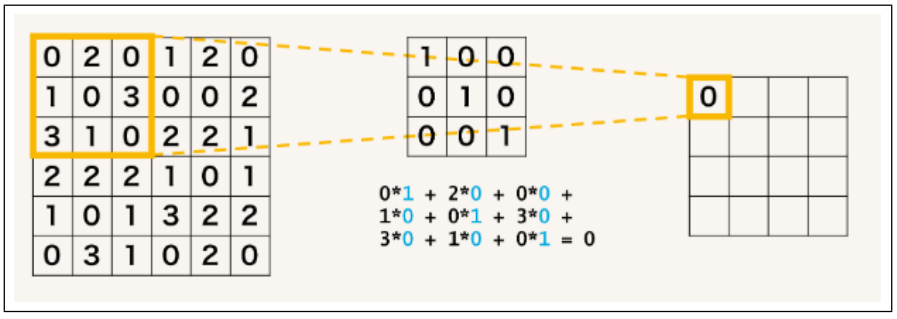

Fig.2.4　Example of a Convolutional calculation
(https://axa.biopapyrus.jp/deep-learning/cnn/convolution.html)

Pooling 層

Pooling 層はある場所でのネットワークの出力を，周辺の出力の要約統計量で置き換える [42]．例えば，よく使われる Max Pooling 処理では短形の近傍中で最大の出力を返す（Fig. 2.5 に示す）：その他の主要な Pooling 関数には，短形近傍の平均，中心ピクセルからの距離に基

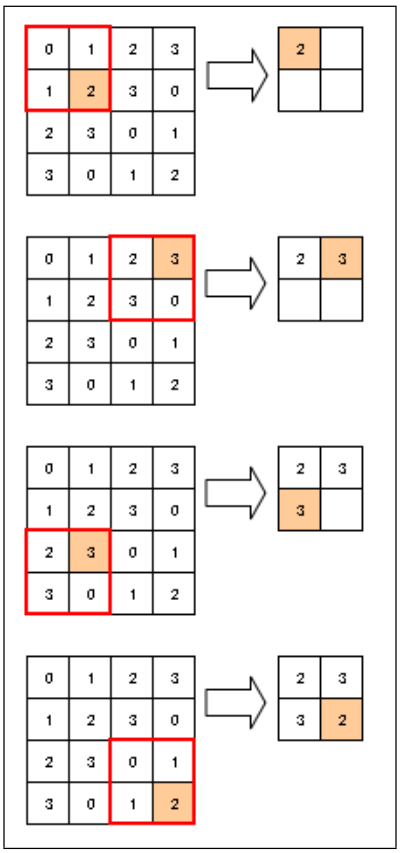

Fig.2.5　Example of a Pooling process in CNN
(https://qiita.com/nvtomo1029/items/601af18f82d8ffab551e)

づく重み付き平均がある．Pooling 層には，以下のような特徴がある：

- 学習するパラメータがない：Pooling 層は，畳み込み層と違って，学習するパラメータを持たない．例えば，Max Pooling は対象領域から最大値を取るだけの処理なので，学習すべきパラメータは存在しない．
- チャンネル数が変化しない：一般的に，画像データは RGB カラー画像であり，チャンネル数が 3 つがある．Pooling の演算によって，入力データと出力データのチャンネル数は変化しない．
- 微小な位置変化に対してロバストである：入力データの小さなズレに対して， Pooling は同じような結果を返す．そのため，入力データの微小なズレに対してロバストである

Pooling 層を利用する場合，以下のメリットが得られる：

- 次元削減: Pooling 層により，畳み込み層の出力や特徴マップの次元を削減できる．これにより，ネットワーク全体で処理するデータの量が減り，計算の効率が向上できる．
- 計算効率向上: Pooling 層はサブサンプリングを通じて情報を集約し，代表的な値を取り出すため，計算の冗長性が減り，処理速度が向上できる．
- 位置情報の保持: Pooling 層は局所的な領域から代表的な値を取得するが，その際に位置情報を一定程度保持するため，モデルは物体の位置や構造を把握しやすくなる．
- 過学習の防止：Pooling 層では，次元削減や情報の統合により，モデルはより汎化された特徴を学習しやすくなる．

2.2.2 Vision Transformer

Vision Transformer（ViT）[10] は，自然言語処理の分野で非常に成功した Transformer [43] の概念を画像処理の分野に適用したモデルである．ViT の構造を理解する前に，自然言語処理の Transformer について理解する必要があるため，以下には Transformer の仕組みを解説する．

Transformer とは

Vision Transformer を理解するために，Attention Mechanism (注意機構) 及び Transformer の概念を理解しなければならない．注意機構とは，自然言語処理の場合では文中の単語の意味を理解するのにどの単語に注目すれば良いのかをを示す機構であり，画像処理の分野では CNN が出力した特徴量マップのどの領域に注目すれば良いのかを表す機構である [44].

そして，Transformer は，2017 年に提案され [43]，自然言語処理に関するタスクの精度向上するための仕組みである．Transformer は，Encoder と Decoder という 2 つの部分で構成される (Fig. 2.6 に示す). Encoder の役割は，Transformer に入力されたデータを機械が処理できる形式に変換することである．例えば，言語翻訳のための学習では，英語等で書かれた文章が入力され Encoder によって数値のベクトルに変換される．Decoder の役割は，Encoder によって変換されたデータを受け取り，処理内容に応じて別の形式へ変換することである．例えば，英語から日本語への翻訳を行う際は，数値ベクトルに変換された英語の文章を日本語の文章へと変換する．Encoder 及び Decoder の主要な部分としては，自己注意機構 (Self-Attention Mechanism) を持つ Multi-Head Attention (いわゆる Multi-Head Self-Attention, MHSA) の部分である．本稿では，ViT が利用する自己注意機構を持つ Multi-Head Self-Attention (MHSA) ブロックを解説する．

まず，自己注意機構は文章の各単語を表す入力系列 $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ を同じ長さの系列 $\mathbf{y} = \{y_1, y_1, \cdots, y_n\}$ へと変換する関数である．系列中の各要素はベクトルである．はじめに，

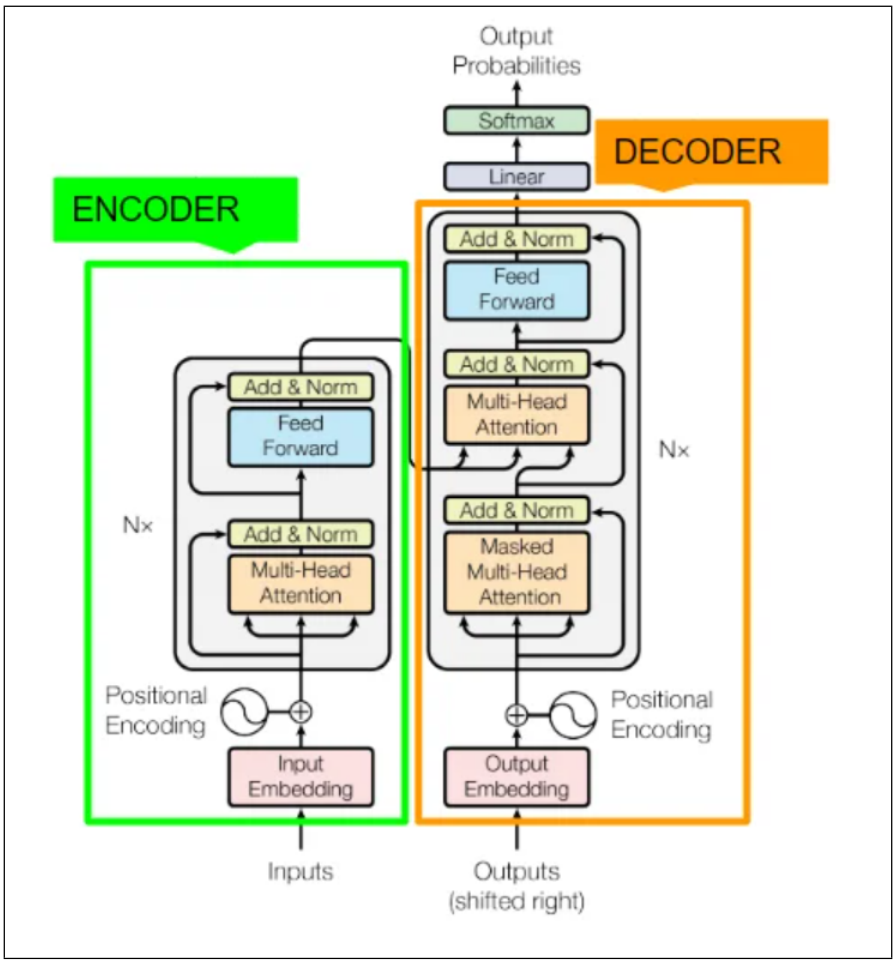

Fig.2.6　The architecture of Transformer

各要素 x_i 毎に要素を使って「Query」,「Key」,「Value」の 3 つのベクトルを計算する．この Query，Key，Value を行ごとに並べて作られた行列を $\mathbf{Q}, \mathbf{V}, \mathbf{K}$ とする．Query と Key の次元数は同じ d_k であり，Value の次元数は d_v とする．この際，スケール化内積である注意機構 (Scaled Dot-Product Attention) は次のように計算される：

$$\text{Attention}(\mathbf{Q}, \mathbf{V}, \mathbf{K}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}. \tag{2.3}$$

つまり，Query と Key の内積によりその関連度を計算し，その関連度を使って Value を加重和すると，注意度を適用した単語の特徴表現が手に入る．しかし，Transformer では，各単語に対して 1 組の Query，Key，Value を持たせるのではなく，比較的小さい Query，Key，Value を複数の読み込み *head* で用意し，それぞれの *head* で特徴表現を計算する．最終的にそれらを 1 つのベクトルに落とし込むことによって獲得された特徴表現をその単語の潜在表現とする．これらの複数の読み込み *head* を Multi-Head Self-Attention (MHSA) と呼び，以下のよう

に定義する：

$$\text{MultiHead}(\mathbf{Q},\mathbf{V},\mathbf{K}) = \text{Concat}(head_1, head_2, \cdots, head_n)\mathbf{W}^O \tag{2.4}$$

$$head_i = Attention(\mathbf{Q}\mathbf{W}_i^{\cdot}\mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V). \tag{2.5}$$

$\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V, \mathbf{W}^O$ は学習対象のパラメータ行列である．Transformer では k 個の $head$ を使って読み込んだ結果を最終的に統合して出力を決定する．

一般に，自己注意機構に位置ごとの非線形変換を組み合わせたものが 1 つの計算ブロックとして扱われる．例えば，Transformer では自己注意機構の後に 2 つの総結合層を用いる．また，スキップ接続や正規化，DropOut 等の正則化が適用される．

Vision Transformer とは

Vision Transformer (ViT) の構造を Fig. 2.7 に示す．自然言語処理の Transformer の場合には，各単語がベクトル表現となっている文を一気に入力する (Fig. 2.8) が，画像処理向けの ViT の場合には，画像をパッチに分けて各パッチを単語のように扱う (Fig. 2.9) .

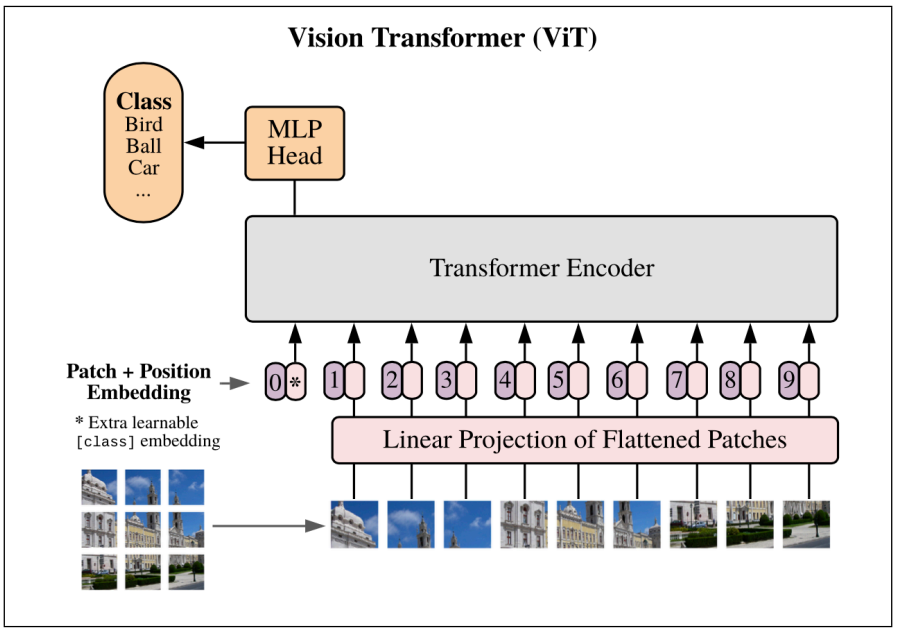

Fig.2.7 The architecture of Vision Transformer

そのため，入力画像を $\mathbf{x} \in \mathbb{R}^{H\times W\times S}$ とすると, 各パッチは以下のように定義できる:

$$\mathbf{x}_p^i | i = 1, 2, \cdots, N \tag{2.6}$$

パッチの数 N は次のように計算される:

$$N = \frac{HW}{K^2}. \tag{2.7}$$

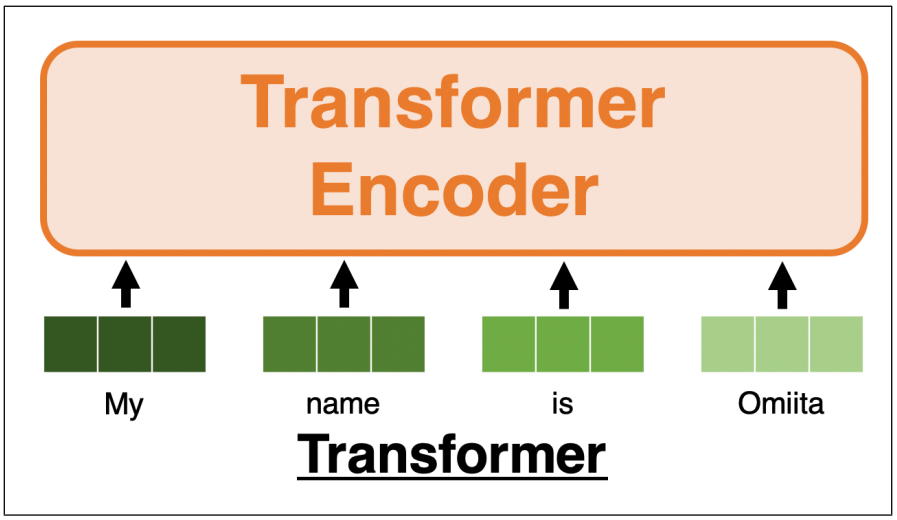

Fig.2.8 Example of input patches in Transformer

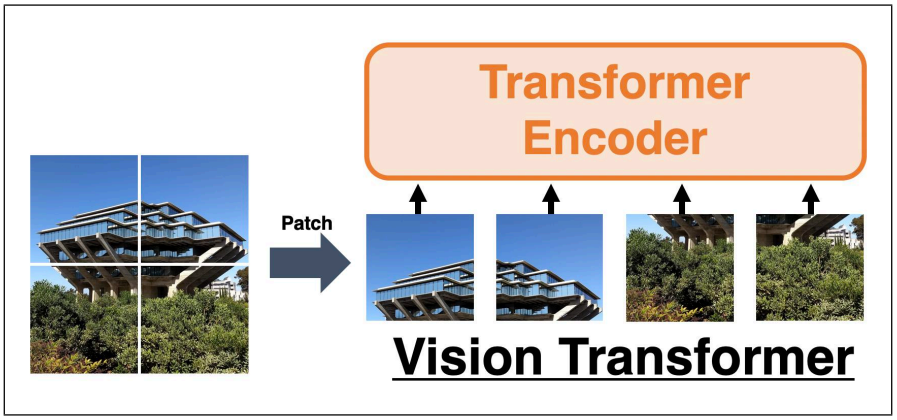

Fig.2.9 Example of nput patches in Vision Transformer

ここで， H と W は入力画像の高さと幅であり, K がカーネルストライドの高さと幅であり, S がカーネルストライドである. その後, Linear Projection Flatten Patches は，パッチ埋め込み関数 $\mathbf{E}$ を使用して，これらのパッチを D 次元に線形変換する.

$$\mathbf{E}(\mathbf{x}_p^i)|i = 1, 2, \cdots, N. \tag{2.8}$$

これらのパッチを Transformer Encoder に転送する前に，Transformer と同じく学習可能な埋め込みトークン (いわゆる分類トークン) $\mathbf{x}_{cls}$ を追加し，位置埋め込み P を融合する. N 個のパッチと $\mathbf{x}_{cls}$ で構成される最終のベクトル $\mathbf{z_0}$ を次のように定義する：

$$\mathbf{z_0} = [\mathbf{x}_{cls}; \mathbf{E}(\mathbf{x}_p^1); \mathbf{E}(\mathbf{x}_p^2); \ldots; \mathbf{E}(\mathbf{x}_p^N)] + P \tag{2.9}$$

ベクトル $\mathbf{z_0}$ は，複数の Transformer ブロックで構成される Transformer Encoder (MSA) に転送する：

$$\mathbf{z_l'} = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \quad l = 1 \cdots L. \tag{2.10}$$

最後に，出力された全てのトークン $\mathbf{z'}_l$ を Multi Layer Perceptron (MLP) に転送する：

$$\mathbf{z}_l = \mathrm{MLP}(\mathrm{LN}(\mathbf{z'}_l)), \quad l = 1 \cdots L, \tag{2.11}$$

$$\mathbf{y} = \mathrm{LN}(\mathbf{z}_L^0). \tag{2.12}$$

ここでは，LN が正規化レイヤー (Layer Normalization) であり，$\mathbf{z}_l$ は $\mathbf{x}_{cls}$ の最終出力である．

ViT は基本的に大規模なデータセットで事前学習され，その後，特定のタスクに対して微調整される. 大規模なデータセットでの事前学習により，モデルは豊富な特徴を学習し，少ないデータでの微調整においても高い性能を発揮できる．

2.3　距離学習

本節では，深層距離学習の概念を紹介する．その後，深層距離学習のための対照的アプローチ及び SoftMax ベースのアプローチについて解説する．

2.3.1　深層距離学習について

深層距離学習 (Deep Metric Learning) [45] とは，深層学習を使用してデータ間の距離や類似性を学習する手法である．距離学習の主要な目的としては, 同じクラスのサンプル（intra-class）間の距離を小さくしながら，異なるクラスのサンプル (inter-class) 間の距離を大きくすることとなる．

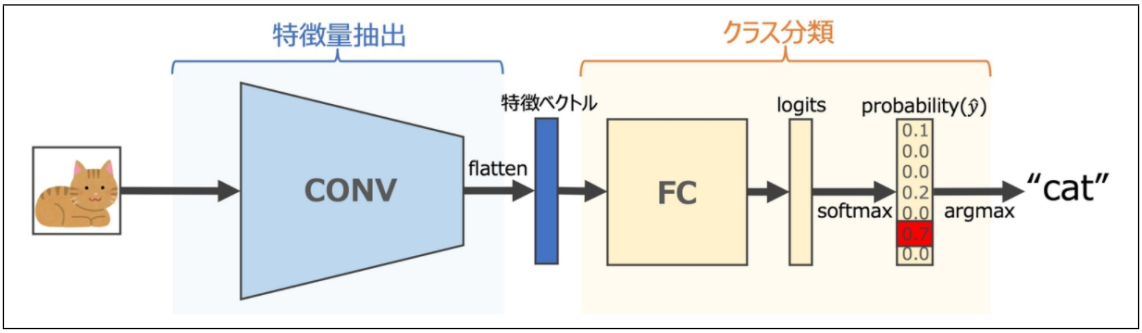

Fig.2.10　Example of a Basic Classifcation process
(https://tech-blog.optim.co.jp/entry/2021/10/01/100000)

深層距離学習の理解に向けて，まず深層学習における一般的なクラス分類（ここで，定番学習と呼ぶ）のメカニズムを理解する必要がある．定番学習の構造を Fig. 2.10 に示す．ここでは，特徴抽出及びクラス分類の部分に分割される．例えば，畳み込みニューラルネットワーク (CNN) の場合，特徴抽出の部分は複数の畳み込み層（CONV）と Pooling 層から構成され，最終的に一次元化される．この操作において，入力画像は特徴ベクトルに変換され，出力された特徴ベクトルは「埋め込み」(Embedding) と呼ばれる．そして，クラス分類の部分では，入力された特徴ベクトルを使用してクラスごとの所属確率を予測して出力する．通常の CNN では，この部分が一つまたは複数の全結合層 FC から構成され，出力された「Logits」は SoftMax 関数を介してクラス所属確率（サンプルが各クラスに属する確率）を計算する．最後に，argmax 関数を使用して最も確率の高いクラスを予測する．

定番学習と距離学習に基づくクラス分類の違いは，特徴抽出部分の学習方法である．定番学習の学習では，特徴抽出ネットワークを全結合層 FC を使用してトレーニングし，intra-class と inter-class のサンプル間の距離を考慮せず，線形的に分離可能な特徴量が抽出される．この

アプローチはトレーニングデータに各クラスのサンプルが十分に含まれている場合に高い精度を実現できるが，多くの場合にはトレーニングサンプルが不足しているクラスや未知のクラスが含まれる問題があるため，学習精度も落ちる．

一方，距離学習に基づく学習手法は，意図的に inter-class のサンプル間の距離を増加させ，intra-class のサンプル間の距離を縮小させるように学習を行い，識別的な特徴量が得られる．Fig. 2.11 に示すように，定番学習の学習と違って，距離学習では複数のサンプルの特徴量を抽出し，これらの類似度に基づいて学習を行う．

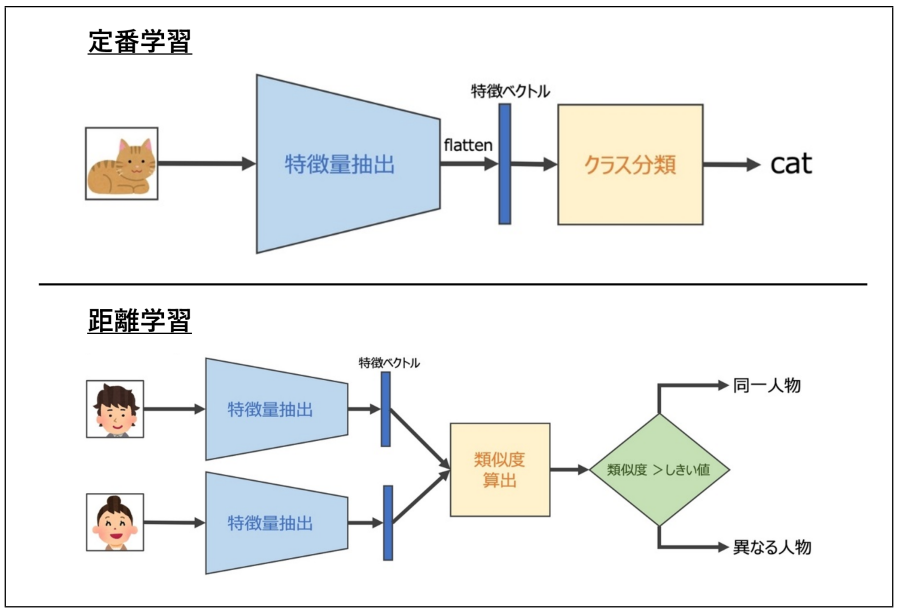

Fig.2.11　Differences between basic Feature Learning process and Metric Learning process (https://tech-blog.optim.co.jp/entry/2021/10/01/100000)

学習後の特徴量については，通常の学習手法と比較すると，intra-class のクラスターがよりコンパクトで，inter-class 間の距離がより離れているため，識別性の高い特徴量空間が形成される．このため，各クラスのサンプルが少ない場合や未知のクラスが存在する場合でも高い性能を発揮し，顔認識や異常検知などのタスクで広く使用されている．

深層距離学習の主要なアプローチには，対照的アプローチ及び SoftMax ベースのアプローチがある．

2.3.2　対照的アプローチ

対照的アプローチ (Contrastive Approach) [45][46] とは，学習データ内の異なるクラスやインスタンスの特徴を対比することに焦点を当てる手法である．対照的アプローチでは，直接的に intra-class のサンプルペアを引き寄せ，inter-class のサンプルペアを押しのけるようにデザインされた損失関数が使用される．このアプローチの重要な要素は，ネットワークの構造，損

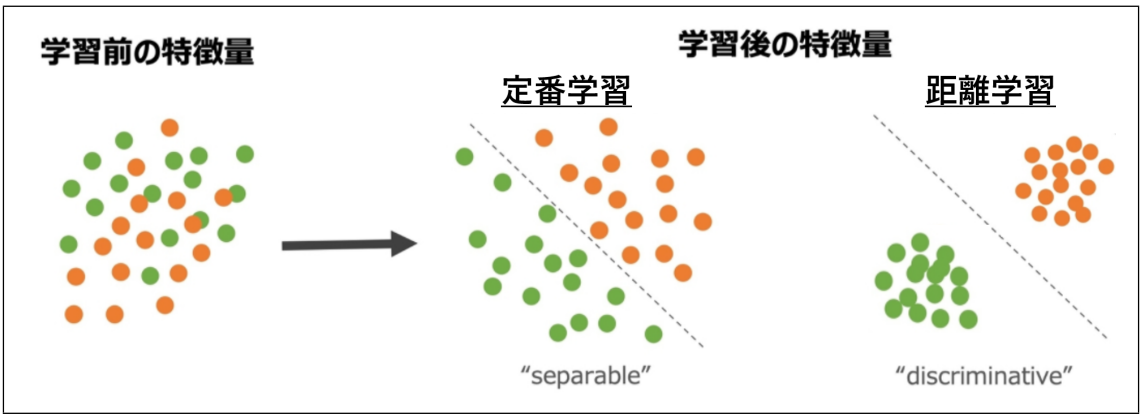

Fig.2.12　Differences between features of Metric Learning process and basic Feature Learning process
(https://tech-blog.optim.co.jp/entry/2021/10/01/100000)

失関数及びサンプルの選択である．

サンプル選択の重要性

対照的アプローチでは，サンプル間の距離に基づいて損失が直接計算されるため，サンプルの組み合わせの選択がモデルの訓練の成功と収束に大きな影響を与えるため，非常に大事である．また，全ての利用可能なサンプルの組み合わせでトレーニングを行うと，計算量と時間が非常にかかるため，効果的かつ効率的な学習を行うために有益なサンプルの組み合わせを選択することは非常に重要である．一般的なサンプル選択のアプローチを Fig. 2.13 に示す．

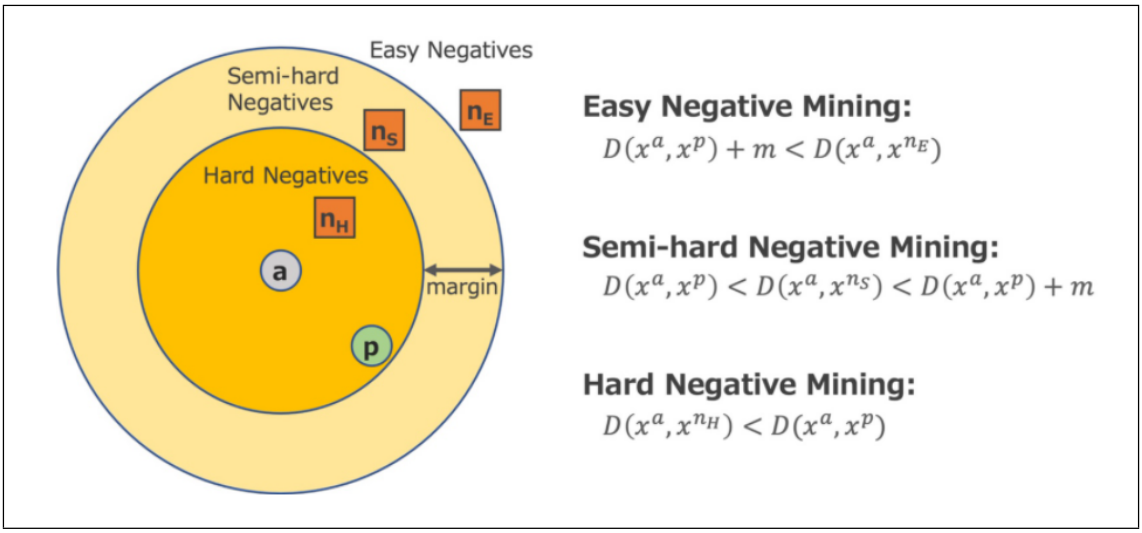

Fig.2.13　Example of common Sample Selection (Sample Mining) methods
(https://tech-blog.optim.co.jp/entry/2021/10/01/100000)

これらのアプローチは，サンプル間の距離及びあるマージン値 (margin) に基づいて学習の

ためのサンプルペアを生成する．ここでは，x_a が Anchor サンプル（基準のサンプル）であり，x_p は x_a と同じクラスの Positive サンプルであり，x_n は異なるクラスの Negative のサンプルである．$D(x_a, x_b)$ はサンプル x_a と x_a の間の距離を表し，マージン m は正の定数である．例えば，「Easy Negative Mining」では， intra-class サンプルの距離が inter-class サンプルの距離より短くて，intra-class のサンプルを引き寄せると inter-class のサンプルを押しのける作業が簡単になり，トレーニングに対して学習の効果が少なく，時間と計算リソースの無駄になる．そのため，多くの研究が「Semi-Hard Negative Mining」または「Hard Negative Mining」のアプローチを使用する．

これらのサンプル選択法は，深層距離学習の代表的な損失関数である「Contrastive Loss」及び「Triplet Loss」のためのものである．次に，Contrastive Loss を解説する．

Contrastive Loss

Contrastive Loss は，2006 年に次元削減の目的で提案された [47]．Contrastive Loss では，2 つのサンプルペアを入力して，共有重み (Shared weights) のネットワークで特徴量を抽出し，その距離 (D) によってネットワークを学習させる．ここで，intra-class のサンプルペア (いわゆる Positive ペア) を入力した時には距離を最小化するように学習し，inter-class のサンプルペア (いわゆる Negative ペア) を入力した時は，距離を最大化するようネットワークを学習させる．Contrastive Loss は，以下のように定義される：

$$L_{\text{ContrastiveLoss}} = (1 - Y)\frac{1}{2}(D)^2 + (Y)\frac{1}{2}max(0, m - D)^2, \tag{2.13}$$

$$D = \|f(x_i^a) - f(x_i^b)\|_2, \tag{2.14}$$

ここで，x_i^a と x_i^b は，i 番目の入力画像のペアを表す．Y はペアのラベルで，Positive ペアの場合はラベルとして 0 がつけられ，Negative ペアの場合はラベルとして 1 がつけられる．マージン m は，正の定数であるハイパーパラメータである．

また，上記の Contrastive Loss を実現するために，Siamese というネットワーク構造を利用する必要がある (Fig. 2.14 に示す)．Siamese ネットワークは，距離学習によく利用されており，類似度や距離を学習するためのニューラルネットワークの一種である．通常，Siamese ネットワークは 2 つの入力データポイントを受け取り，それらのデータポイントがどれだけ類似しているかを評価する役割を果たす．このアーキテクチャは，主に顔認識や物体追跡等の類似性を評価するタスクに使用される．

しかし，Contrastive Loss は以下のデメリットがある：

- Positive ペアと Negative ペアだけから学習しているため，情報が少ない推論時に正解ラベルが必要で，無標識データには対応できない．
- 境界線付近のサンプルからはほとんど学習できない．

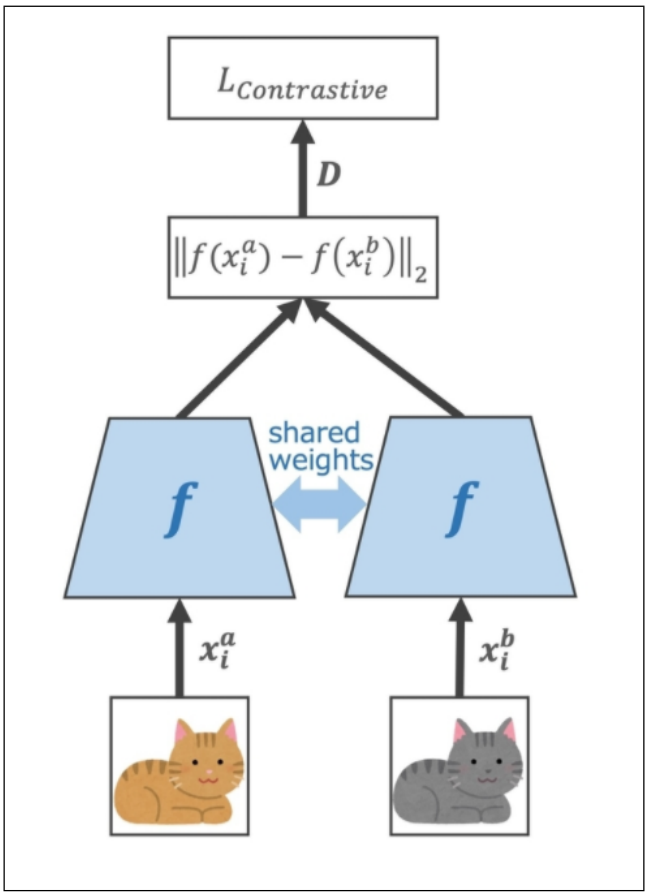

Fig.2.14　Example of Contrastive Loss
(https://tech-blog.optim.co.jp/entry/2021/10/01/100000)

- クラス内の乖離が大きいサンプルペアでは効果が低い.

Triplet Loss

Contrastive Loss の欠点を改善するために，2015 年に Triplet Loss が提案された [48]．Fig. 2.15 に示すように，Triplet Loss では，3 つの異なるサンプル (Triplet) x_a, x_p, x_n を活用してトレーニングが行われる.

ここで，x_a（Anchor サンプル）と x_p（Positive サンプル）は同じクラスのデータであり，x_n（Negative サンプル）は異なるクラスのデータである．Triplet Loss では， intra-class のペア (Positive ペア) をより近づけ，inter-class のペア (Negative ペア) をより遠ざける．つまり，x_a を x_p よりも x_n に近づける：

$$\|f(x_a) - f(x_p)\|_2^2 < \|f(x_a) - f(x_n)\|_2^2, \tag{2.15}$$

ここで，$f(x)$ は x の特徴埋め込みである．Triplet Loss は，異なるクラスのサンプルペア (inter-class) が少なくともあるマージン値 m だけで同じクラスのサンプルペア（intra-class）か

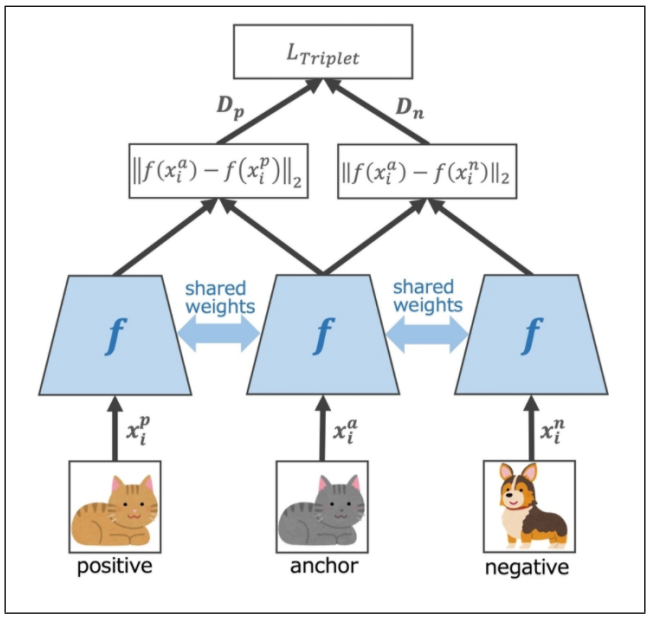

Fig.2.15　Example of Triplet Loss
(https://tech-blog.optim.co.jp/entry/2021/10/01/100000)

ら離れていることを確保するように設計されており，次の式で定義される：

$$L_{\text{TripletMarginLoss}} = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + m \right]_+, \tag{2.16}$$

ここで，$[z]_+ = max(z, 0)$ であり，m がハイパーパラメータである．

2.3.3　SoftMax ベースのアプローチ

SoftMax Loss

このアプローチはクラス分類用の SoftMax Loss に基づいた手法である [49]．SoftMax Loss は，以下のように定義される：

$$L_{\text{SoftMaxLoss}} = -\frac{1}{N} \sum_{i=1}^{N} log(\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}), \tag{2.17}$$

x_i は i 番目のサンプルの特徴量であり，y_i クラスに所属する．W_j と b_j は j 番目の列の重みとバイアスである．SoftMax Loss は，同クラスのサンプル間の類似度を高めながら他クラスの類似度を低くするように強制しないため，得られた特徴量でのクラス識別性能は低いという課題がある．それを改善するために，以下の Center Loss が提案された [50].

Center Loss

Triplet Loss は，d_p と d_n の差に焦点を当てているが，それらの絶対値を無視する傾向がある．例えば，$d_p = 0.3$，$d_p = 0.5$，$m = 0.1$ の場合，Triplet Loss は 0.1 となる．$d_p = 2.3$，$d_p = 2.5$，$m = 0.1$ の場合でも，Triplet Loss も 0.1 となる．そして，Triplet Loss は，ランダムに選択した 2 つのデータポイントによって決定するが，トレーニングデータセット全体で $dp < dn$ を確実にするのは困難である．この Triplet Loss の欠点を補完するために，2016 年に Center Loss は提案された [50]．この損失関数では，同クラスのサンプル間の距離を短縮するよう特徴量とそれに対応するクラスの中心位置との距離でペナルティを課すので，上記の対照的アプローチのような直接的にサンプル間の距離を比較する学習手法と違って，このアプローチではグローバルな情報 (クラスの中心位置) を利用して学習することが可能になり，さらに面倒なサンプル選択の工夫が不要になる．Center Loss は次のように定義される：

$$L_{\text{CenterLoss}} = \frac{1}{2}\sum_{i=1}^{N} \|f(x_i) - c_{y_i}\|_2^2, \tag{2.18}$$

ここでは，N が バッチサイズであり，$f(x)$ は x の特徴埋め込みとなり，c_{y_i} は i 番目のサンプルの正解クラス (y_i クラス) の中心を表す．Center Loss は，クラス内の変動を効果的に注目する．トレーニンの際，$L_{\text{SoftMaxLoss}}$ を加えて総合損失関数 (L_{total}) を計算する：

$$L_{\text{total}} = -\frac{1}{N}\sum_{i=1}^{N} log(\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}) + \frac{\lambda}{2}\sum_{i=1}^{N} \|f(x_i) - c_{y_i}\|_2^2. \tag{2.19}$$

ここでは，2 つの損失関数のバランスをとるためにパイパーパラメータ λ が導入される．

2.4　まとめ

本章では，画像検索の概要及び関連技術について関説明を行った．次章では，UAV におけるクロスビュー場所推定の問題に対し，これらの技術を用いてどのように解決するのかについて解説する．

第 3 章

無人航空機のためのクロスビュー場所推定

多くのクロスビュー場所推定の研究は，主に 2 つのタスクに焦点を当てている．一つ目は地上画像と衛星画像のマッチングである (CVUSA [51] 及び CVACT [52] データセットで実施). 二つ目は UAV 画像及び衛星画像のマッチングである (University-1652[53] データセットで実施)．これらの研究では，クロスビュー場所推定を画像検索のタスクとして扱う．ある画像を与えられると，提案モデルが異なるビューの Gallery から類似の画像を取得する．

初期の関連研究は，古典的画像処理技術の特徴抽出やマッピングのアルゴリズムである SIFT [37] または SURF[38] を使用して特徴を抽出し，その後両画像の類似度スコアを計算して画像を識別する．しかし，異なる視点間で外観が大きく異なると，多くの場合は検索に成功しない．その一方，近年は画像処理での深層学習の応用研究が多く見られる．高い学習能力を持つ CNN がクロスビュー場所推定に適用されており，優れた結果が報告されている [54][55][56][57]．ただし，クロスビュー場所推定を実現するために重要なことは，画像間の共有特徴を見つけ出し，画像の内容を完全に理解することである．CNN アーキテクチャは小さな識別特徴に注目するため, CNN にとって画像全体の大域的な特徴を求めるのは非常に困難である．これらの欠点を改善するために，注意機構 [58][59] に注目する研究がある．例えば，Vision Transformer 構造は一部のクロスビュー場所推定の関連研究 [60][61][62][63] に使われた．これらの手法は，Vision Transformer を使って，入力画像をピクセルレベルで処理 (SGM [62])， またはトークンレベルを処理しており (FSRA [63])，学習可能な埋め込み（分類トークン）の使用を十分に活用していないと見られる．

本章では, まず UAV におけるクロスビュー場所推定に関するデータセット (University-1652) 及び既存研究について紹介する. そして，本研究が提案する 2 つの深層学習モデル PAAN と TATN を説明し，実験結果及び考察を述べる．最後に，本章の成果についてまとめる．

3.1　データセット

本研究では，Zheng ら [53] により提供された University-1652 データセット (Table 3.1 に示す) を利用した．このデータセットには，世界中の 72 の大学の 1,652 の建物（1,652 の場所）が含まれている．各建物のデータは，衛星画像，UAV 画像及び地上画像の 3 つの異なる画像タイプで構成されている．全ての画像のサイズは 512×512 である．特にこのデータセットは，これまで UAV におけるクロスビュー場所推定の唯一のデータセットである．データセットの詳細内容を Table 3.1 に示す．

Table3.1　Details of Universities-1652

	Universities-1652			
	Views	Images	Classes	Universities
Training	UAV	37,854	701	33
	Satellite	701	701	
	Ground	2659	701	
Testing	Query (UAV)	37,855	701	39
	Query (Satellite)	701	701	
	Query (Ground)	2579	701	
	Gallery (UAV)	51,355	951	
	Gallery (Satellite)	951	951	
	Gallery (Ground)	2921	793	

Zheng らはまずインターネットにて，対象の大学の地理情報を収集し，その地理情報に基づいて，Google Map から衛星画像を収集し，データセットを作成した．これらの衛星画像は，UAV 画像と同一の縮尺であり，高い解像度を持っている．UAV 画像に関しては，UAV でのデータ収集にが高いコストがかかるため，本データセットでは全ての UAV 画像はシミュレーションで取得された．バーチャル地球儀システムである Google Earth において，各大学にある建物の 3D モデルとその周辺環境に対して，事前に設定された UAV の仮想的な飛行経路に従って撮影を行った (Fig. 3.1 に示す)．これらの画像は，シミュレートされた UAV ビューに基づいて取得されたものであるため，実際の UAV 画像に近いものである．University-1652 データセットでは，各建物とその周辺が「場所」を呼び，学習の時にはクラスとして扱う．ここで注意すべき点として，トレーニングとテストセットに重複するクラスはない．つまり，トレーニングで学習したクラスと同一のクラスはテストデータには含まれない．

University-1652 データセットのサンプル画像は，Fig. 3.2 に示す．(a) が地上画像であり，

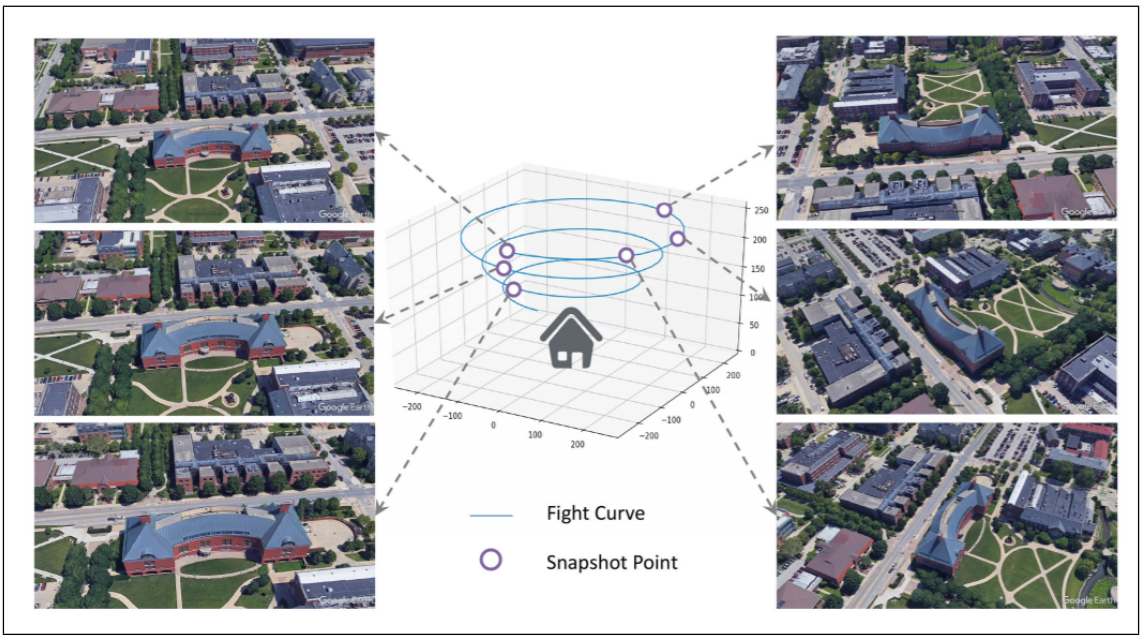

Fig.3.1　Example of the drone flight curve toward the target building

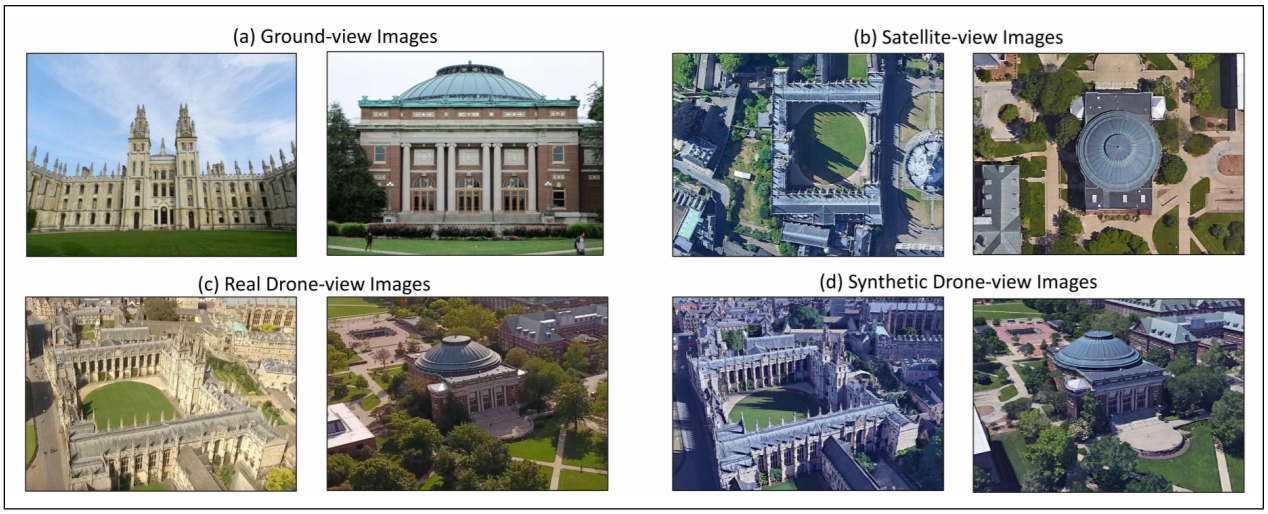

Fig.3.2　Sample images from University-1652: (a) Ground-view images, (b) Satellite-view images, (c) Real Drone-view images collected from public drone flights (d) Synthetic UAV-view images

(b) が衛星画像であり，(d) が UAV 画像である．University-1652 データセットは，実環境で飛行したドローンによる収集された 4K 解像度の画像も提供するため，(c) がこの 4K 解像度の画像の例を示す．これらの 4K 解像度の画像の数が少ないため，実環境の画像での検証実験に利用する．

University-1652 を用いた研究は，主に特徴学習アプローチを注目してモデルを構築する傾

向がある．既存手法は大きく分けると，畳み込みニューラルネットワークを用いるモデルと Vision Transformer を用いるモデルに分類できる．次節では，畳み込みニューラルネットワークベースの PAAN を解説する．

3.2 PAAN: Part-Aware Attention Network

本節では，まず畳み込みニューラルネットワークを用いた既存手法を紹介し，提案手法の PAAN を提案する．その後，提案手法を検証した実験の結果及び考察を述べる．

3.2.1 畳み込みニューラルネットワークを用いた手法

畳み込みニューラルネットワーク (CNN) を用いる既存手法がいくつかあるが，本稿では以下の 2 つの代表的なモデルを紹介する：

Baseline [53]: このモデルは，University-1652 を作成した Zheng らの提案モデルである．Fig. 3.3 に Baseline の構造を示す. Baseline は，Siamese ネットワークのような複数のブランチを持つアーキテクチャであるが，全てのブランチは 1 つの分類モジュール (Classifier Module) につながる．

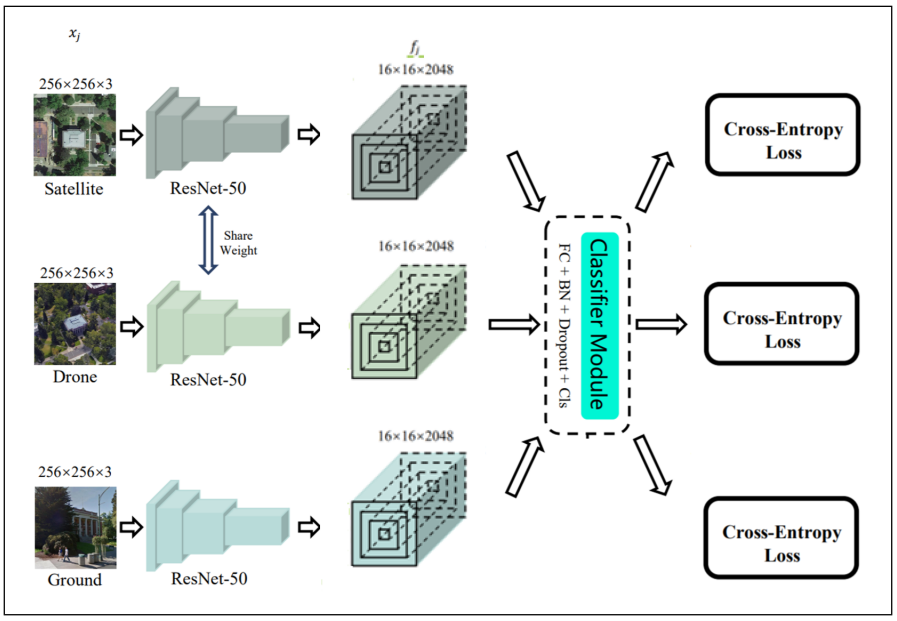

Fig.3.3 The architecture of Baseline model

Zheng らが提案した Baseline モデルを理解するために，まず通常の場所推定を理解する必要がある．Fig. 3.4 に示すように，各ドメインで画像で場所推定したいであれば，場所推定を分類問題として扱い，各ドメインに別々の分類モデルを採用して学習を行う．このように，各モデルの分類モジュール (Classifier Module) が対応するドメインの特徴を学習ができて，最高のパフォーマンスを発揮できる．

しかし，今回の問題は，クロスビュー場所推定であり，異なるドメインで同じ場所の画像を見つける問題であり，分類モジュールが異なるドメインの特徴を学習してマッピングできる能

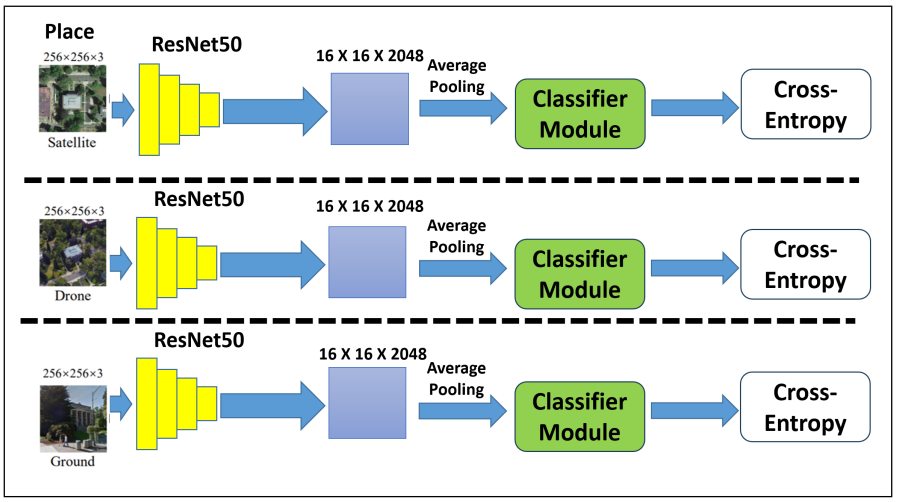

Fig.3.4　Approach of basic place classification

力が必要になる．そのため，Fig. 3.5 に示すように，Zheng らが提案した Baseline モデルでは全てのブランチが 1 つの分類モジュールを共有する．このように，異なるドメインの特徴を 1 つの共有特徴空間にマッピングできると期待される．

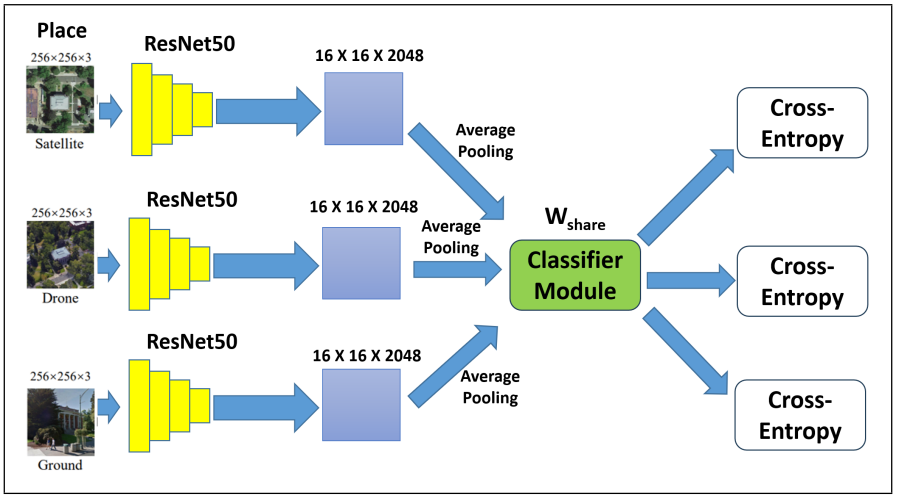

Fig.3.5　Approach of Baseline model

例えば，$\mathbf{x}_j$ が衛星・UAV・地上の入力画像として，各ブランチが出力した特徴ベクトル $\mathbf{g}_j$ は以下のように計算される:

$$\mathbf{f}_j = \mathbf{F}_{\text{backbone}}(\mathbf{x}_j), \quad j \in \{1, 2, 3\}, \tag{3.1}$$

$$\mathbf{g}_j = AvgPool(\mathbf{f}_j), \tag{3.2}$$

ここで，$\mathbf{F}_{\text{backbone}}$ は特徴抽出器，$AvgPool$ が Average Pooling 操作である．そして，学習時の

損失関数 L_{final} は以下のように計算される：

$$\mathbf{z}_j = \mathbf{F}_{\text{classifier}}(\mathbf{g}_j), \tag{3.3}$$

$$p(y|\mathbf{z}_j) = \frac{exp(\mathbf{z}_j(y))}{\sum_{c=1}^{C} exp(\mathbf{z}_j(c))}, \tag{3.4}$$

$$L_{\text{final}} = \sum_j -\log(p(y|\mathbf{z}_j)), \tag{3.5}$$

ここで，$\mathbf{F}_{\text{classifier}}$ は共有分類モジュールであり，$p(y|\mathbf{z}_j)$ が正解クラス y の予測確率である．

Local Pattern Network (LPN) [64]: LPN の構造を Fig. 3.6 に示す．

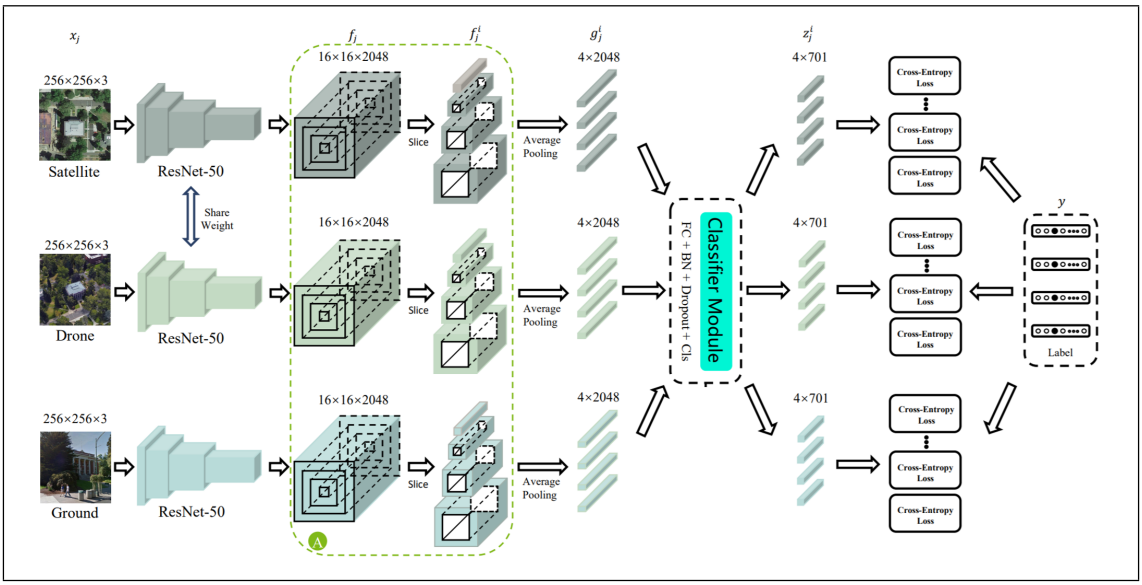

Fig.3.6 The architecture of LPN

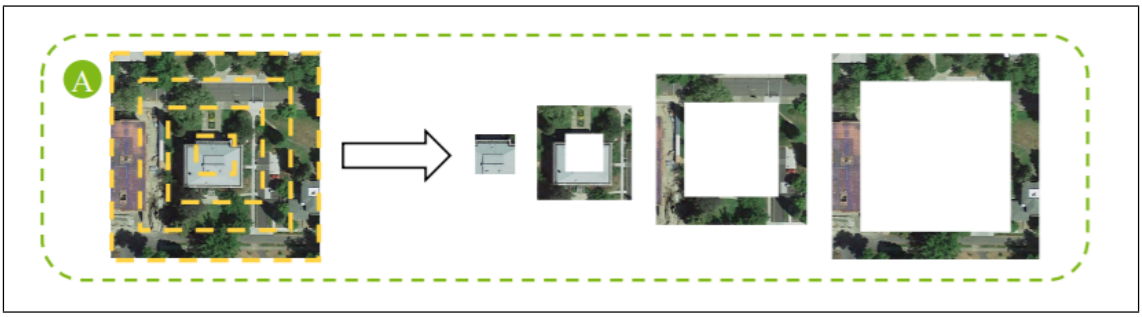

Fig.3.7 Feature Partition Strategy (LPN)

このモデルは，Zheng らの Baseline とほぼ同じ構造であるが，特徴分割法 (Feature Partition Strategy) (Fig. 3.7 に示す) という新しい特徴処理法を提案した．入力画像（UAV・衛星・地上）

を $\mathbf{x}_j$ として，ResNet-50 から出力された特徴ベクトルは以下のようになる：

$$\mathbf{f}_j = \mathbf{F}_{\text{backbone}}(\mathbf{x}_j), \quad j \in \{1, 2, 3\}, \tag{3.6}$$

ここでは，$\mathbf{F}_{\text{backbone}}$ が特徴抽出器とする. 特徴分割法は，特徴抽出器の ResNet-50 から抽出された特徴ベクトル $\mathbf{f}_j$ を 4 つの特徴パート $\mathbf{f}_j^i$ に分割して Average Pooling 層に入れて，特徴パートの $\mathbf{g}_j^i$ を出力する：

$$\mathbf{f}_j^i = \mathbf{F}_{\text{slice}}(\mathbf{f}_j, i), \quad i \in \{1, 2, 3, 4\}, \tag{3.7}$$

$$\mathbf{g}_j^i = AvgPool(\mathbf{f}_j^i), \tag{3.8}$$

ここで，$\mathbf{F}_{\text{slice}}$ は特徴分割法，*AvgPool* が Average Pooling 操作を表す. その後，全てのパートを共有分類モジュールに入れて，学習時の損失関数 L_{final} は以下のように計算される：

$$\mathbf{z}_j^i = \mathbf{F}_{\text{classifier}}(\mathbf{g}_j^i), \tag{3.9}$$

$$p(y|\mathbf{x}_j^i) = \frac{exp(\mathbf{z}_j^i y))}{\sum_{c=1}^{C} exp(\mathbf{z}_j^i(c))}, \tag{3.10}$$

$$L_{\text{final}} = \sum_{i,j} -\log(p(y|\mathbf{x}_j^i)), \tag{3.11}$$

ここで，$\mathbf{F}_{\text{classifier}}$ は共有分類モジュールであり，$p(y|x_j)$ が正解クラス y の予測確率である.

3.2.2　本研究のアプローチ

まず，上記の既存手法 (Baseline, LPN) から次の共通点を指摘する：

1. **ネットワーク構造:** 主に複数のブランチがある構造を利用する.
2. **損失関数:** 全ての研究は Cross Entropy Loss を利用する．その際，共有の分類モージュルを利用する.
3. **特徴処理:** CNN ベースのモデルは主に単なる ResNet50 モデルを特徴抽出器として利用する.

これらの点を踏まえて，本研究は以下の方針に基づいて提案モデルを構築する：

1. **ネットワーク構造:** 上記の研究は，複数のビューの画像から場所を推定したが，本研究は UAV・衛星の画像情報に基づいて場所を推定するため，提案手法のブランチ数は 3 つ（UAV・衛星・地上）ではなく，2 つ（UAV・衛星）とする.
2. **損失関数:** 既存研究はクロスビュー場所推定をクラス分類の問題として扱い，クラス分類のための分類モージュルや Cross-Entropy Loss を利用してドメイン違いの問題を解決するつもりであったが，まだ不十分だと考えられる．そのため，本研究では距離学習の分野によく利用される損失関数を導入すれば改善が期待できる.

3. **特徴処理:** CNN ベースのモデルについては，既存研究の結果により，ResNet-50 ベースのモデルが優れた精度に寄与したと考えられる．一方，深層学習にあるテックニック（例：注意機構の利用，Pooling 法の変更等）を利用すれば，モデルを画像の重要な部分に注目させることができ，精度の改善が期待できる．

次節では，本論が提案する ResNet-50 ベースの Part-Aware Attention Network (PAAN) を解説し，そのモデルの性能を示す．

3.2.3 提案手法

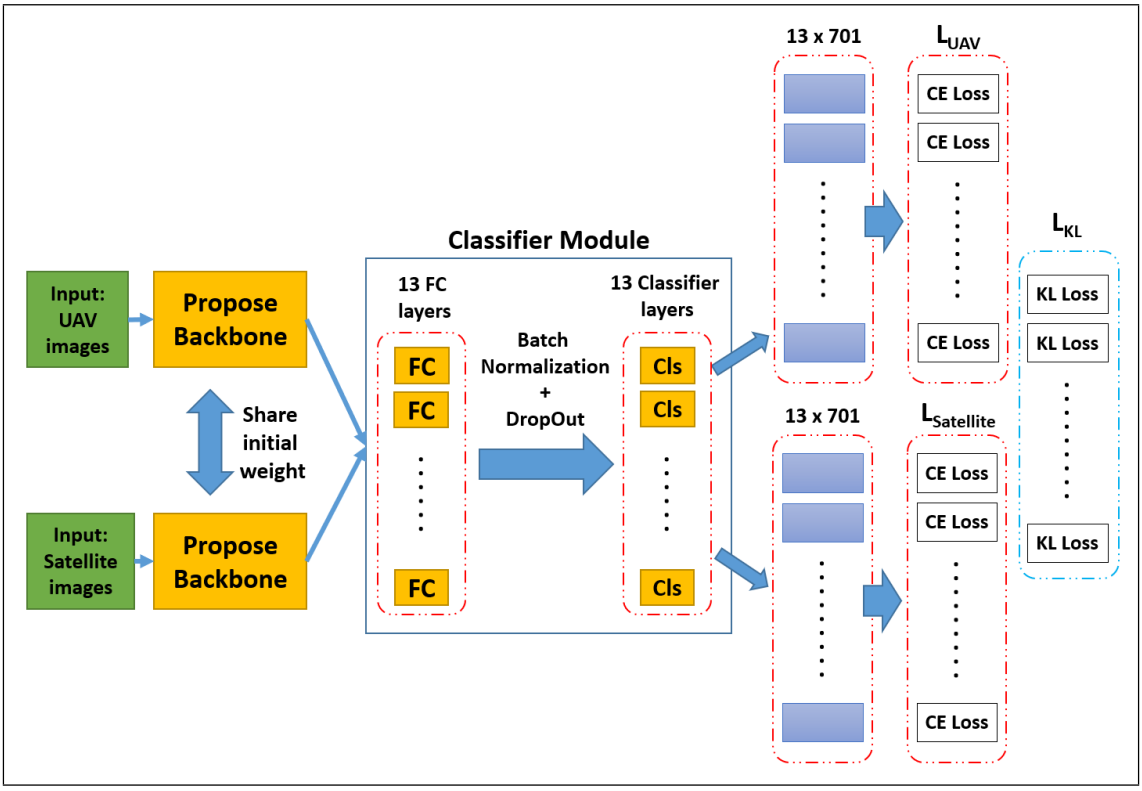

Fig.3.8 The proposed PAAN architecture

PAAN のアーキテクチャーの概要を Fig. 3.8 に示す．ネットワークは 2 つのブランチを持ち，各ブランチには本稿が提案する特徴抽出器（Proposed Backbone）を設定し，同じ初期重みを利用する．また，各ブランチで抽出した特徴は複数の完全接続層 (Fully Connected Layer - FC と記載する) と分類層 (Classifier Layer - Cls と記載する) で構成される分類モジュール (Classifier Module) に送信される．提案する特徴抽出器のアーキテクチャーを Fig. 3.9 に示す．

PAAN の主要な点は，以下のとおりである：

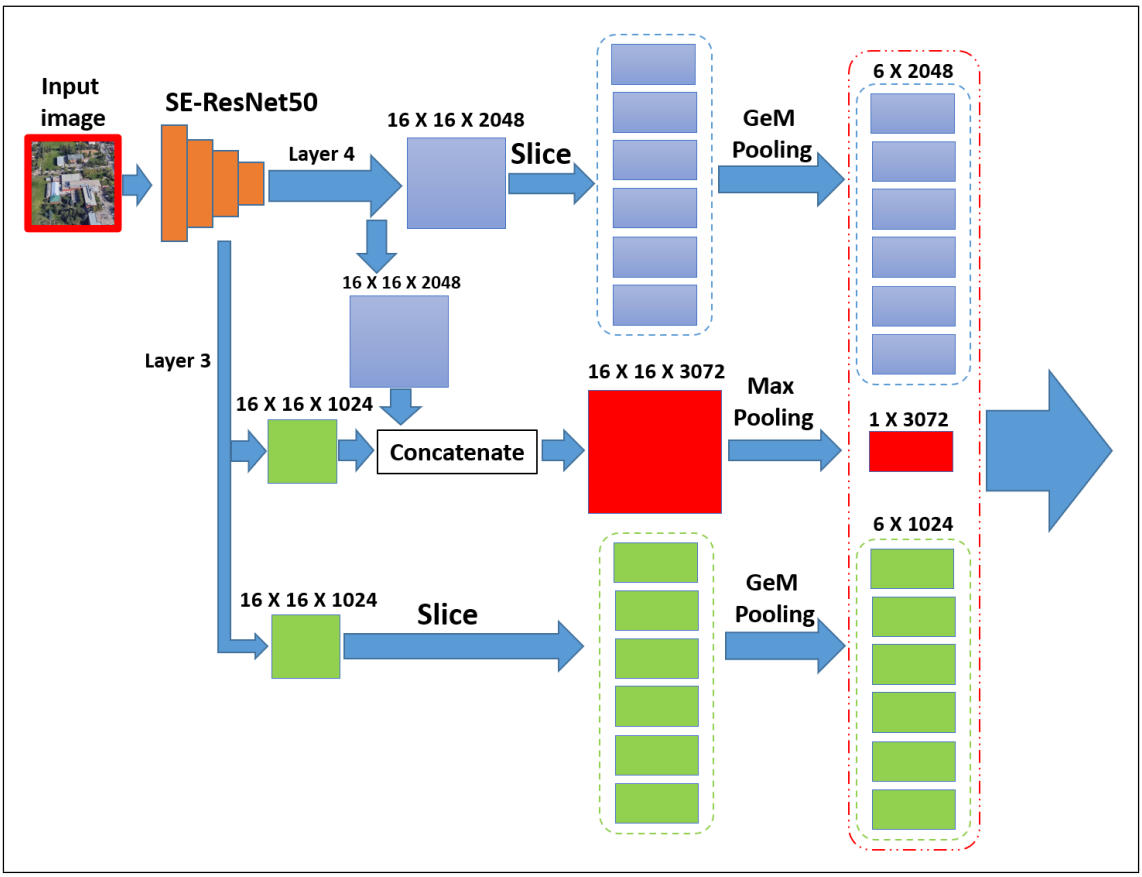

Fig.3.9　The proposed backbone (PAAN)

- 特徴抽出器（Proposed Backbone）：
  注意機構の利用： Squeeze-and-Excitation ブロック（SE ブロック）[65] と呼ばれる注意機構を持つ ResNet-50 (SE-ResNet50) の利用（Fig.3.9 のオレンジ色のボックス）.
  新しい特徴分割法： 6 個の特徴パーツに分割する方法（Fig.3.9 の赤色・緑色・青色のボックス）.
  新しい Pooling 法： GeM Pooling と Max Pooling の利用.
- 損失関数：Cross-Entropy Loss (CE Loss) 及び Kullback–Leibler Divergence Loss (KL Loss) の利用（Fig. 3.8 の CE Loss と KL Loss）.

次に，PAAN の各要素について解説する.

特徴抽出器

Squeeze-and-Excitation ブロック（SE ブロック）は，Hu ら [65] により提案された注意機構である．このブロックは，特定のニューラルネットワークのためのものではなく，ネットワーク中の追加経路として用いることができる注意機構である．SE ブロックでは，入力された

特徴マップが，全体平均 Pooling (Global Average Pooling - GAP) 及び 2 つの全結合層 (Fully Connected Layer - FC) により圧縮される．Fig. 3.10 にこのモジュールの構造を示す.

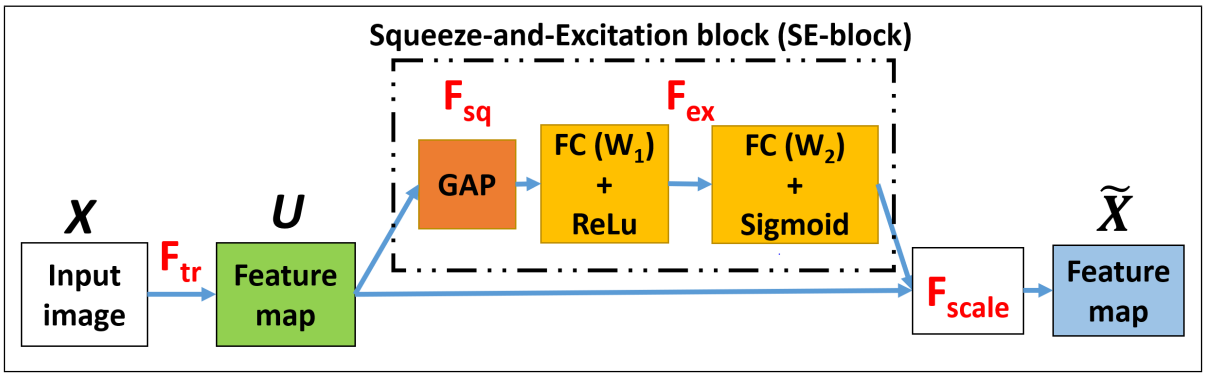

Fig.3.10 The architecture of SE-block module

まず，入力画像 $\mathbf{X} \in \mathbb{R}^{H' \times W' \times C'}$（$H'$ が高さ，W' が幅，C' がチャンネル数である）に対し，フィルタ $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N]$ で構成される畳み込み演算子 $\mathbf{F}_{tr}$ を適用する．この操作の後，特徴マップ $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$（$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_N]$）が得られ，$\mathbf{u}_c$ は以下のように計算される：

$$\mathbf{u}_c = \mathbf{v}_c * \mathbf{X} = \sum_{i=1}^{C'} \mathbf{v}_c^s * \mathbf{x}^s \tag{3.12}$$

ここで，“$*$” は畳み込み演算を示す（第 2.2.1 項に説明した）．

そして，SE ブロック内の Squeeze ステップ $\mathbf{F}_{\text{sq}}$ では，$H \times W$ での Global Average Pooling（GAP）を使用し，チャンネルごとの代表値として画素値平均 $\mathbf{z} \in \mathbb{R}^C$ を取る．ここで，$\mathbf{z}$ の c 番目の z_c は以下のように計算される：

$$z_c = \mathbf{F}_{\text{sq}}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{3.13}$$

その後，Excitation 操作 $\mathbf{F}_{\text{ex}}$ では，Squeeze ステップからの $\mathbf{z}$ を 2 つの FC 層で処理し，チャンネル間の非線形の関係を取る．ここでは，アクティベーション $\mathbf{s}$ を生成する：

$$\mathbf{s} = \mathbf{F}_{\text{ex}}(\mathbf{z}, W) = \sigma(\mathbf{W}_2 ReLU(\mathbf{W}_1 \mathbf{z})) \tag{3.14}$$

ここでは，$\mathbf{W}_1 \in R^{\frac{C}{r} \times C}$ 及び $\mathbf{W}_2 \in R^{C \times \frac{C}{r}}$ となる．最後に，アクティベーション $\mathbf{s}$ を使用し，元特徴マップ $\mathbf{U}$ を再スケーリングすることで，最終特徴マップ $\bar{\mathbf{X}} = [\bar{\mathbf{x}_1}, \bar{\mathbf{x}_2}, \cdots, \bar{\mathbf{x}_c}]$ が得られる：

$$\bar{\mathbf{x}_c} = \mathbf{F}_{\text{scale}}(\mathbf{u}_c, s_c) = s_c \mathbf{u}_c. \tag{3.15}$$

生成されたアクティベーション $\mathbf{s}$ を元の特徴マップ $\mathbf{U}$ にかけ合わせることで，最終特徴マップ $\bar{\mathbf{X}}$ にチャンネル間の相互作用を含めることができる．このように，価値の高いチャンネルを

強調することで表現の質を上げることを目指している．SE ブロックは複数のニューラルネットワークに簡単に適用できる追加モジュールとして設計された．本研究では，ResNet-50 モデルの各 Residual ブロックに SE ブロックを適用する．

特徴分割法

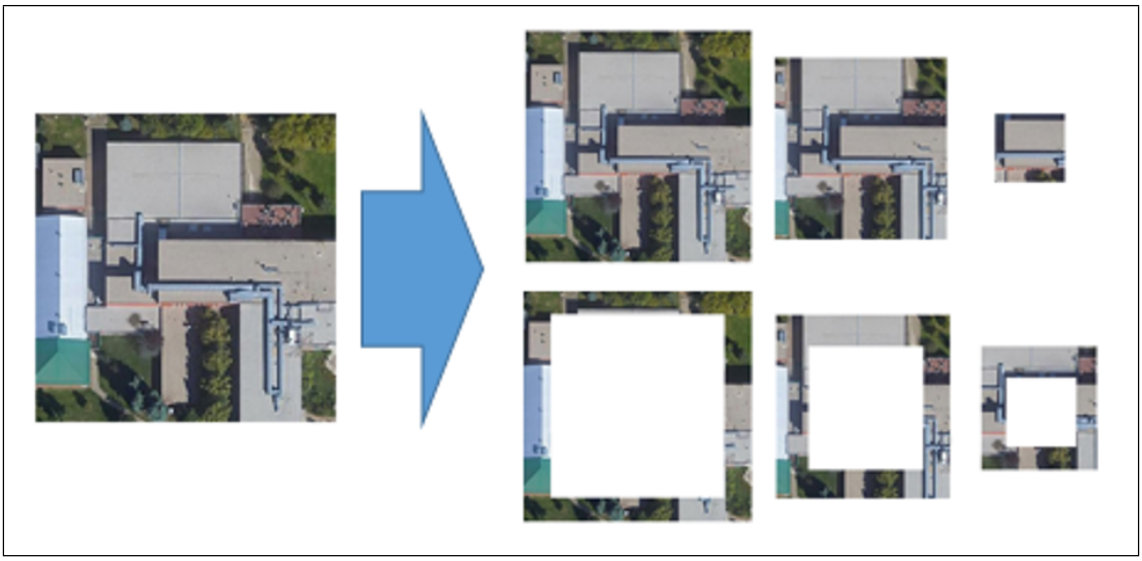

Fig.3.11　The proposed Feature Partition Strategy (PAAN)

本研究が利用した特徴分割法は 6 個のパーツに分割する (Fig. 3.11 に示す)．そして，既存研究 LPN では，ResNet-50 の第 4 番目レイヤーの特徴に対して特徴分割法を適用しているが，PAAN モデルでは，他のレイヤーからの特徴も抽出し，それらの特徴に対し特徴分割法を適用する．Fig. 3.9 に示すように，第 4 番目レイヤーの特徴だけではなく，第 3 番目レイヤーの特徴にも特徴分割法を適応した．さらに，多レベルの特徴の統合が最終結果に大きく貢献すると考えられるため，第 3 番目レイヤー及び第 4 番目レイヤーの特徴を結合したグローバル特徴マップも作成する（Fig. 3.9 の赤色ボックス）．

※第 3 番目レイヤーと第 4 番目レイヤーの特徴サイズが異なるため，これらの特徴を結合するためには，第 4 番目レイヤーの最終的なダウンサンプリングレイヤーのストライドは 2 から 1 に調整する必要がある．

Pooling 層

これまでの関連研究では，ほとんどの CNN アーキテクチャが Average Pooling や Max Pooling 等の一般的な Pooling 手法を適用している．ところが，従来の Pooling 手法は，バックプロパゲーションプロセスを通じて学習できず，空間分解能の低下によってすべての空間情報をうまく保存しないため，入力の最終的なグローバル記述に影響を与える可能性がある．例

えば，Max Pooling は Pooling 領域からの最大要素のみを考慮し，他の要素を無視するため，情報の損失によって悪い結果につながる可能性がある．上記に述べたように，クロスビュー場所推定の分野では，画像の全体的な背景の情報を理解することが重要である．

そこで，関連研究とは異なり，本研究では Average Pooling の代わりに，Generalized Mean Pooling (GeM Pooling) [66] を導入した．GeM Pooling は画像検索用の Pooling 方法として多くの検索システムで広く応用され，有望な性能を収めている．GeM Pooling は次のように定義される：

$$f^{(g)} = [f_1^{(g)} \cdots f_k^{(g)} \cdots f_K^{(g)}]^T, f_k^{(g)} = \left(\frac{1}{|\mathbf{X}_k|} \sum_{x \in X_k} x^{p_k} \right)^{\frac{1}{p_k}} \tag{3.16}$$

ここで，$\mathbf{X}_k$ は特徴マップを表し，k がチャンネルの数であり，p_k が Pooling パラメータである．この Pooling パラメータは手動で設定するか，学習プロセスを通じて変更できる．ちなみに，Average Pooling 及び Max Pooling が GeM Pooling の特別な場合である．p_k が 1 に近づくと，GeM Pooling 関数は Average Pooling になり，p_k が ∞ に近づくと，GeM Pooling は Max Pooling になる．

損失関数

上記の既存手法では，各場所を 1 つのクラスとして扱い，このモデルを分類モデルとしてトレーニングする．学習の際，共有分類モジュールにより，異なるソースの特徴を 1 つの共有特徴空間にマッピングし，ネットワーク全体の損失は Cross-Entropy Loss で計算する．本研究では，機械学習・深層学習の分野にある Kullback–Leibler Divergence Loss (KL Loss) を導入する．

機械学習の分野では，Kullback–Leibler Divergence は確率分布間の類似度を測定する指標である．本モデルに KL Loss を導入する目的は，異なるドメインからの出力間の関係を学習し，類似したインスタンス間の距離を狭めることである．ここでは，正規化された確率スコアを取得するためにソフトマックス関数を使用し，その後 KL Loss を計算する：

$$L_{\mathrm{KL}(p_2 \| p_1)} = \sum_{i=1}^{N} p_2^i \log(\frac{p_2^i}{p_1^i}), \tag{3.17}$$

ここで，p_1 と p_2 は各ブランチの予測結果である．

分類モジュール

分類器モジュール は Fig. 3.12 の Classifier Module に該当する．ここでは，提案した特徴抽出器が出力した特徴に対応する FC レイヤー：6 個の FC レイヤー（入力サイズと出力サイズ：2048 と 512），6 個の FC レイヤー（入力サイズと出力サイズ：1024 と 512），1 個の FC レイヤー（入力サイズと出力サイズ：3072 と 512）を用意する．また，バッチ正規化；DropOut レ

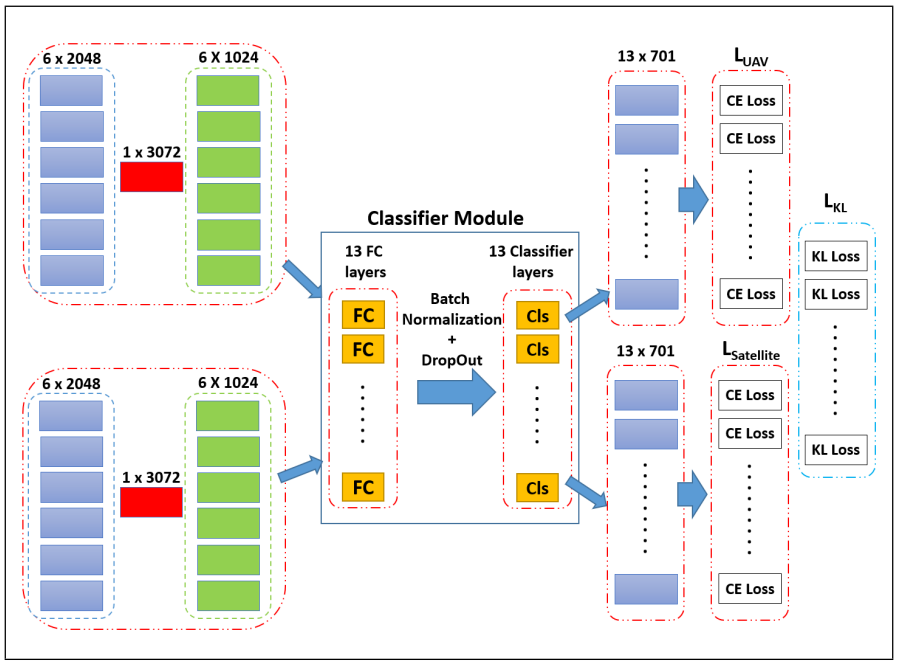

Fig.3.12　The proposed Classifier Module (PAAN)

イヤー；13 個 の分類レイヤー（入力サイズと出力サイズ：512 と 701）を用意する. 総合の損失関数 L_{final} は次のように計算される：

$$L_{\text{final}} = \sum_{i=0}^{N} L^{i}_{\text{UAV}} + L^{i}_{\text{Satellite}} + L^{i}_{\text{KL}} \tag{3.18}$$

ここで，L_{UAV} と $L_{\text{Satellite}}$ は各ビューで計算された損失であり，N は特徴ベクトルの数 (以下では 13 となる) である.

3.2.4　実験設定

トレーニングフェーズ

先行研究では，256 × 256 の画像サイズで実験を行っており，本論でもこれに従った画像サイズで実験を行った．トレーニングでは，入力データに対しデータ拡張 (Cropping, Rotation) を使用した．オプティマイザーには，運動量 (Momentum) の 0.9 の確率的勾配降下法 (Stochastic Gradient Descent - SGD) を採用した．初期学習率は 0.001 であり，モデルの訓練期間を 120 エポックとした．DropOut 率は 0.75 とした．全てのプログラムは Pytorch フレームワークで構成し，NVIDIA Titan XP で実行した.

テストフェーズ

テストの際，分類器モジュールの最終的な分類器レイヤーを削除し，それぞれから出力された特徴ベクトルを連結した最終の 1 つの特徴マップを生成する．ユークリッド距離 (Euclidean Distance) を用いて Query 画像及び Gallery 画像の距離を計算し，Gallery の中で最も Query 画像を類似している画像を探索する (Fig. 3.13)

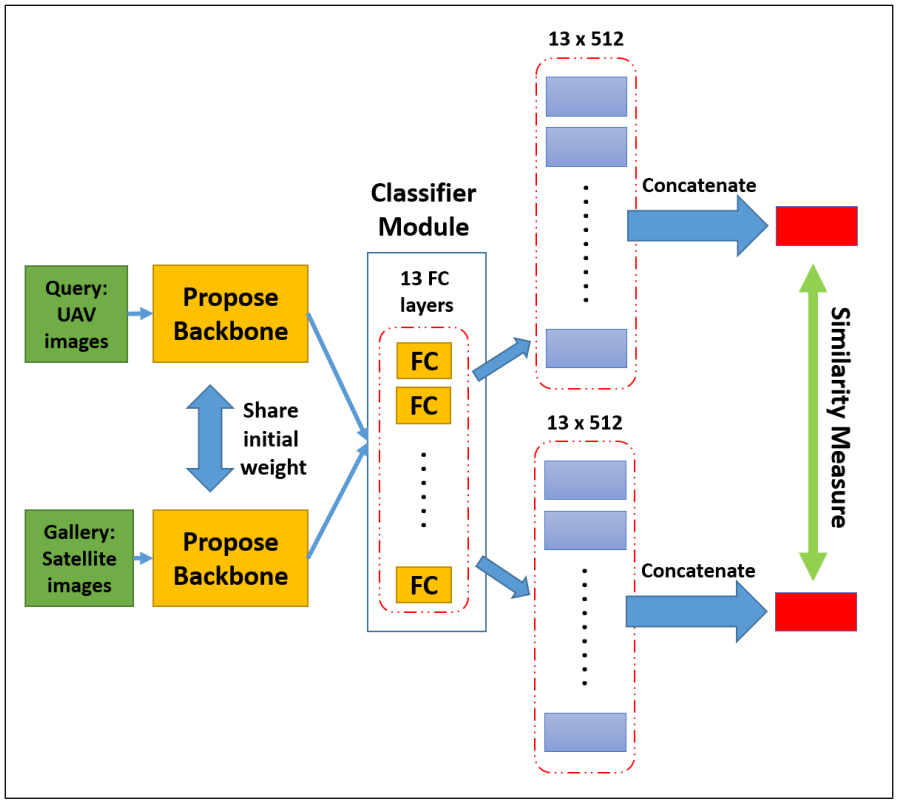

Fig.3.13 Testing phase (PAAN)

その探索作業を評価するために， 先行研究では Recall @K と Average Precision (AP) という 2 つの評価指標を使って評価する：

- Recall@K（K 件での再現率）は，画像検索やレコメンデーションの評価指標の一つで，上位 K 番目の検索結果または検索したアイテムのうち，正解（真陽性）の割合を示す．Recall@K は，特に上位 K 件での再現率を評価するために使われる．

$$\text{Recall@K} = \frac{|a \cap p_K|}{|a|} \tag{3.19}$$

ここで，K は考慮する上位ランキングの数で，a は正解の集合，p_K はランキングリストの上位 K 番目である．本論では，Recall@1 を利用する．

- Average Precision (AP) とは，推薦システムや検索システム等の性能を表す評価指標である．この値は適合率・再現率 (Precision・Recall) 曲線の下の面積 (Area Under Curve - AUC) であり，各適合率の変化に対する平均的な適合率を示す．AP が大きいほど，検索パフォーマンスが向上していることを示す．AP は，以下のように計算する：

$$\text{Precision@K} = \frac{|a \cap p_K|}{K} \tag{3.20}$$

$$y_K = \begin{cases} 1 : \text{上位 K 番目が適合結果} \\ 0 : \text{それ以外} \end{cases} \tag{3.21}$$

$$\text{AP} = \sum_{k=1}^{N} \frac{\text{Precision@K} \cdot y_K}{\sum_{i=1}^{k} y_i} \tag{3.22}$$

3.2.5　実験結果

Table 3.2 は，提案したモデルを他の関連研究と比較した結果である．本研究で提案した PAAN モデルは，UAV → Satellite タスクにおいて 84.51% の Recall@1 精度と 86.78% の AP を達成し，Satellite → UAV タスクにおいては 91.01% の Recall@1 精度と 82.28% の AP を達成した．PAAN は，全ての既存の CNN ベースモデルを大幅に上回ったことがわかる．

Table3.2　Comparisons with state-of-the-art methods on University-1652. The best accuracy is highlighted in **bold**

Method	Backbone	Resolution	Testing inference time	Task			
				UAV → Satellite		Satellite → UAV	
				Recall@1	AP	Recall@1	AP
Baseline [67]	ResNet-50	256 × 256	-	58.49	63.13	71.18	58.74
LPN [64]	ResNet-50	256 × 256	1.00×	74.16	77.39	85.16	73.68
LPN [64]	ResNet-101	256 × 256	1.51×	76.13	79.29	85.45	75.45
PAAN	SE-ResNet50	256 × 256	1.17×	**84.51**	**86.78**	**91.01**	**82.28**

3.2.6　考察

本節では，提案手法の各要素を評価するために，いくつかの比較実験を実施し，その結果を考察する．また，テストデータで検証した場合における提案手法が検索した画像を可視化する．

各要素の検証実験

今回の比較実験については，提案手法の各要素を評価する：SE ブロックの利用 (SE block usage)，ブランチ数 (Number of branches)，Pooling 法 (Type of Pooling)，特徴分割法の利用 (Feature Partition Strategy)，特徴を利用する層 (Feature from layer)，KLLoss の利用 (KLLoss usage) を比較項目として，以下の 6 つの実験を実施した (Table 3.3 に示す)：

- Exp.1, Exp. 2, Exp.3: 異なる特徴抽出器 (ResNet50・SE-ResNet50) 及び Pooling 法 (Average Pooling・GeM Pooling) で構成されたモデルで実験を行い，GeM Pooling 及び SE ブロックの効果を検証する．
- Exp.4, Exp.5, Exp.6: 特徴抽出器 (SE-ResNet50) 及び GeM Pooling があるモデルで特徴抽分割方法 (4 パーツ・6 パーツ)，特徴を利用する層 (第 3 番目のレイヤー・第 3 番目 + 第 4 番目のレイヤー)，KLLoss の効果を検証する．

Table3.3 Ablation studies on University-1652.

Method	SE block usage	Number of branches	Type of Pooling	Feature Partition Strategy	Feature from layer	KL Loss usage
Baseline	×	3	Avg	×	4	×
Baseline	×	2	Avg	×	4	×
Exp. 1	×	2	GeM	×	4	×
Exp. 2	✓	2	Avg	×	4	×
Exp. 3	✓	2	GeM	×	4	×
LPN	×	3	Avg	4	4	×
Exp. 4	✓	2	GeM	4	4	×
Exp. 5	✓	2	GeM	6	3+4	×
Exp. 6	✓	2	GeM	6	3+4	✓

Table 3.4 に示すように，GeM Pooling 及び SE ブロックは，両方とも結果を向上させることが分かる (Exp.1 から Exp.3 までの結果)．ResNet50 の各レイヤーで SE ブロックを使用することで，両方のタスクで Recall@1 の精度が約 4 % 向上した．そして，GeM Pooling の利用により，提案モデルの性能は大幅に向上した．

GeM Pooling 及び SE ブロックの効果をさらに理解するために，各特徴抽出器（ResNet50・SE-ResNet50）からの出力特徴マップを取得し，ヒートマップの形式で可視化した．Fig.3.14

Table3.4　Ablation studies on University-1652. The best accuracy is highlighted in **bold**

	SE block usage	Number of branches	Type of Pooling	Feature Partition Strategy	Feature from layer	KL Loss usage	UAV → Satellite		Satellite → UAV	
							R@1	AP	R@1	AP
Baseline	×	3	Avg	×	4	×	58.49	63.13	71.18	82.31
Baseline	×	2	Avg	×	4	×	58.23	62.91	74.47	59.45
Exp. 1	×	2	GeM	×	4	×	59.45	64.00	73.89	59.44
Exp. 2	✓	2	Avg	×	4	×	63.69	68.22	77.03	63.61
Exp. 3	✓	2	GeM	×	4	×	66.81	70.87	77.6	59.44
LPN	×	3	Avg	4	4	×	75.93	79.14	86.45	74.79
Exp. 4	✓	2	GeM	4	4	×	78.95	81.79	89.59	77.92
Exp. 5	✓	2	GeM	6	3+4	×	82.85	85.27	90.16	**82.51**
Exp. 6	✓	2	GeM	6	3+4	✓	**84.51**	**86.78**	**91.01**	82.28

では，画像のクラス固有の領域を強調し，画像を分類する際にニューラルネットワークがどの部分に焦点を当てているかを理解するのに役立つ．例えば，暖かい色（赤，黄色）はより活性化された領域を表し，涼しい色（緑，青）はあまりに活性化されない領域を表す．Fig.3.14 では，異なる Backbone と Pooling 方法の違いによる注視点の違いをヒートマップで示す．これを見ると，(a) 及び (b) の ResNet50 ベースの方法は，画像の中央の建物に焦点を当てているだけである．一方，(c) 及び (d) の SE-ResNet50 ベースの方法は，中心だけではなく小さな近隣地域も活性化させており，ネットワークが入力画像の様々なグローバル情報に焦点を当てている事がわかった．特に，提案したモデル（d）は，SE-ResNet50 と GeM Pooling の組み合わせを使用すると，他の部分（道路，周辺の建物等）を強調し，画像上でより大きな活性領域を持っていることが分かった．

そして，Table 3.4 の Exp.4 及び Exp.5 の結果も特徴分割法の性能を示した．LPN の特徴分割法を利用した Exp.4 と比べて，新しい特徴分割法を利用した Exp.5 は両方のタスクで Recall@1 の精度が約 4 % 向上した．ここでは，深層学習の深層とは浅層の特徴の効果をより深く理解するために，いくつかの追加実験を行った（Table 3.5 の Exp.7, Exp.8, Exp.9, Exp.10）．全ての実験は，SE-ResNet50 と GeM Pooling を利用したが，特徴分割法では異なるレイヤーからの特徴を利用した．ここでは，1・2・3・4 の数字は，どの層の特徴に特徴分割法をを適用したかを表す．例えば，Table 3.5 の (3 + 4) は，SE-ResNet50 の第 3 及び第 4 番目レイヤーからの特徴に特徴分割法を適用したことを意味する．

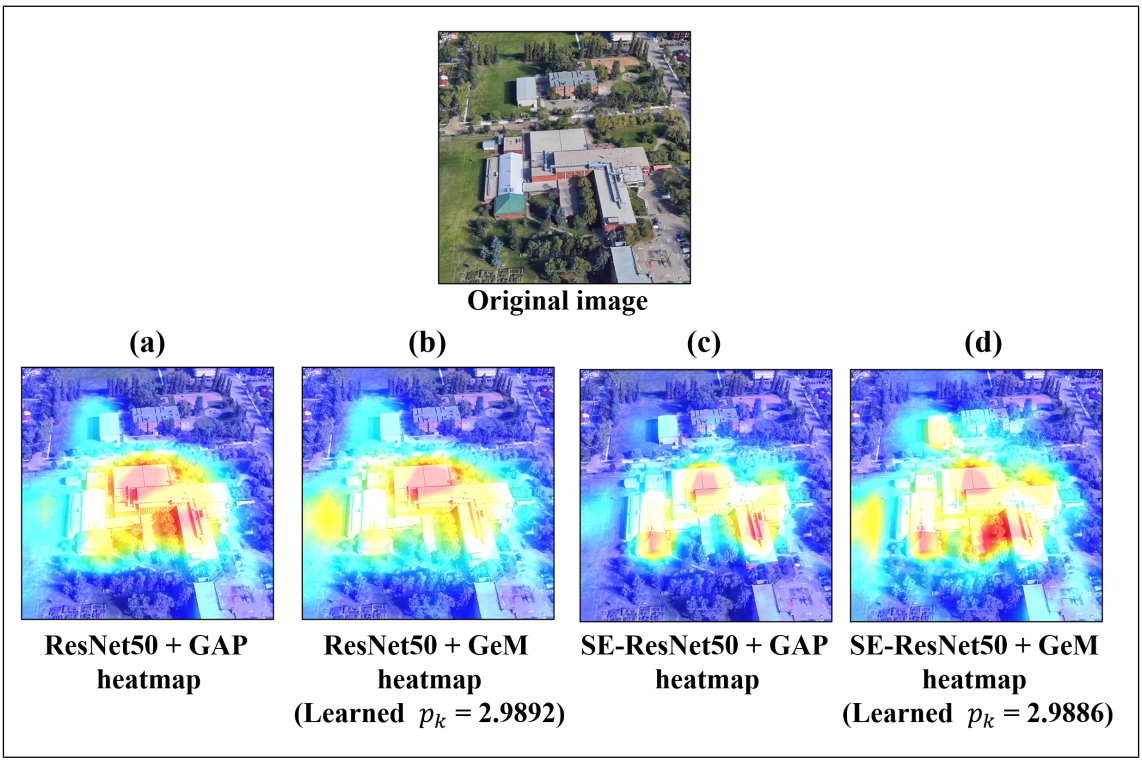

Fig.3.14 Collected heatmap on different models

Table3.5 Ablation studies on diferent feature partition strategy. The best accuracy is highlighted in **bold**

	Feature from layer	University-1652			
		UAV → Satellite		Satellite → UAV	
		Recall@1	AP	Recall@1	AP
Exp. 7	4	82.87	85.13	90.87	82.06
Exp. 8	3 + 4	**84.51**	**86.78**	**91.01**	**82.28**
Exp. 9	2 + 3 + 4	82.49	84.93	88.30	78.95
Exp. 10	1 + 2 + 3 + 4	77.70	80.67	85.02	75.32

Table 3.5 の結果を見ると，第 1 及び 第 2 番目レイヤーからの特徴を使用すると，Recall@1 及び AP の結果が低下し，第 3 及び第 4 番目レイヤーからの特徴を組み合わせると最高のパフォーマンスが得られたことがわかる．これらの結果により，第 1 及び第 2 番目レイヤーからの特徴は，浅い初期のレイヤーの特徴であり，画像全体のグローバル情報を表すことができないため，それらの特徴を最後の特徴表現に使用するとモデルの動作が悪化する可能性があると考えられる．

また，これまでの既存手法は，共有分類モジュールが優れていることから，異なるソースの

特徴を 1 つの共有特徴空間にマッピングするという方向を利用したが，Exp. 6 の優れた結果から見ると，KL Loss の利用により，さらに異なるソースの特徴マッピング能力を向上するの可能性が指摘できる．

検索結果の可視化

University-1652 データセットには，実環境で飛行したドローンによる収集された 4K 解像度の画像が限定的であるが提供されている．実際のミッションに対する提案手法の信頼性を確認するために，これらの 4K 解像度画像に対して検索を行う（Real UAV 画像 → 衛星画像）．正解の画像は黄色のボックス，不正解の画像が青いボックスに表示される．Fig. 3.15 に示すように，衛星 Gallery には各場所につき 1 つの画像しか含まれていないため，正しい答えを見つけるのは難しいが，提案されたモデルは正解の画像を見つけることができた．この結果により，提案手法の PAAN がシミュレーション画像で訓練されたとしても実環境の画像で優れたパフォーマンスを発揮ができ，実際の UAV ミッションに適用可能と考えられる．

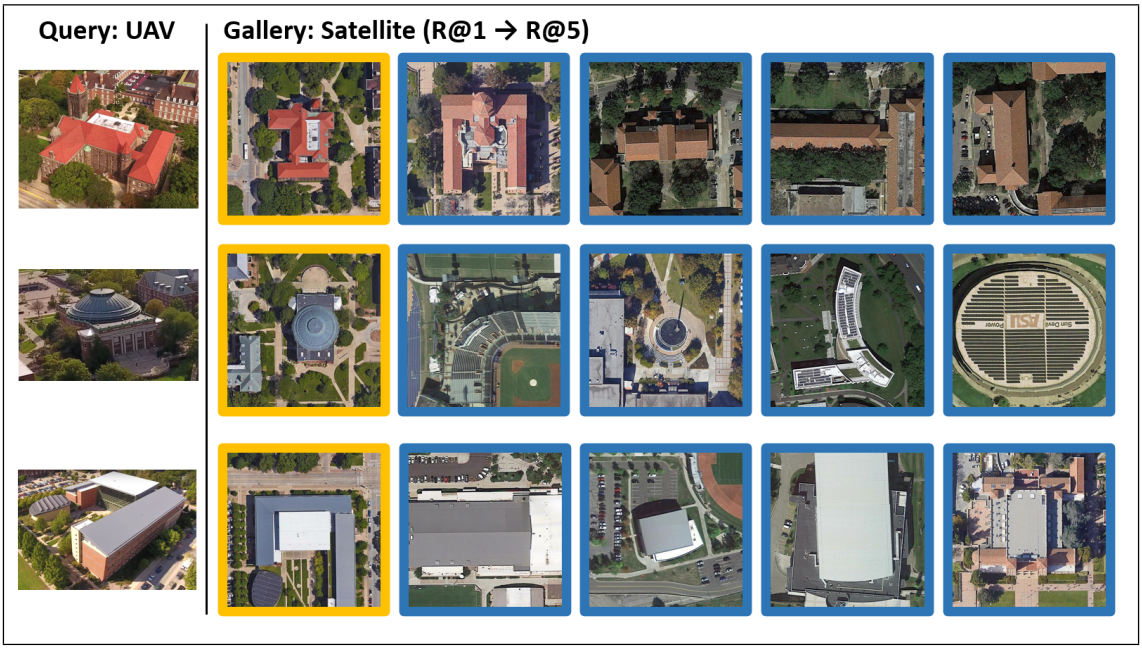

Fig.3.15　Visualization on 4K-image UAV data

3.2.7　結論

本節では，ResNet50 ベースの既存手法のアイデアに基づいて，適切な改善を行い，新しい ResNet50 ベースの PAAN モデルを提案した．このモデルでは，深層学習の注意機構及び新しい特徴処理法を利用することで，既存手法と比べて大幅に精度の向上ができた．追加実験の結果により提案モデルの各要素を検証した．

次節では，ViT ベースの既存手法を紹介する．その後，本稿が提案した ViT ベースの Token-Aware Transformer Network (TATN) を紹介し，そのモデルの性能を示す.

3.3　TATN: Token-Aware Attention Network

本節では，まず Transformer ベースの既存手法を紹介し，提案手法の TATN を提案する．その後，提案手法を検証した実験の結果及び考察を述べる．

3.3.1　Transformer を用いた既存手法

Transformer を用いた既存手法については，これまで以下の 2 つのモデルが知られる：

Feature Segmentation and Region Alignment (FSRA) [63]: このモデルは，Fig. 3.16 に示す．Baseline 及び LPN と違い，FSRA モデルの構造は 1 つのブランチのみで構成されており，特徴抽出器は ViT である．

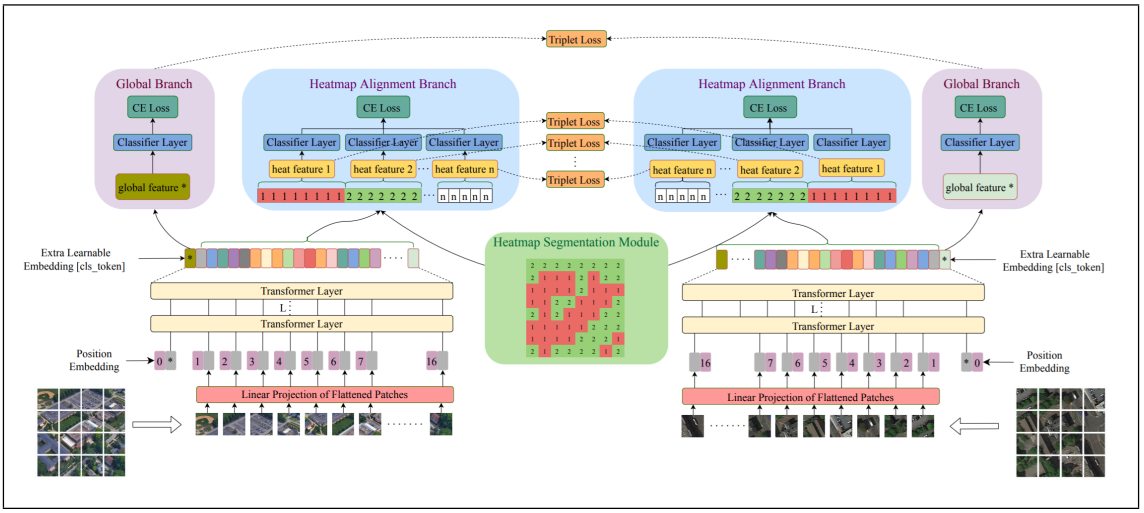

Fig.3.16　The architecture of FSRA

FSRA では，まず特徴抽出器の ViT を通じて，学習可能な埋め込み $\mathbf{x}_c ls$ を除いて全てのトークン $L \in \mathbb{R}^{B\times N\times S}$ (B はバッチサイズを表し，N はトークンサイズを表し，S は各トークに対応する特徴ベクトルの長さを表す) を取得する：

$$L = [F(\mathbf{x}_p^1); F(\mathbf{x}_p^2); \ldots; F(\mathbf{x}_p^N)]. \tag{3.23}$$

各トークンの熱値 (Thermal value) は次のように定義される：

$$P^c = \frac{1}{S}\sum_{i=1}^{S} M^i \quad c = 1, 2, \ldots, N. \tag{3.24}$$

ここで，P^c は c 番目のトークンの発熱量 (Heat Value) を表す．M_i は，特徴ベクトルの i 番目の値に対応する c 番目のトークンを表す．次に，P^{1-N} の値を降順に並べ替えて n 個の領域に

応じてトークンを均等に分割する．各領域に対応するトークンの数 N は以下のようになる：

$$N^i = \begin{cases} \lfloor \frac{N}{n} \rfloor & i = 1, 2, \cdots, n-1 \\ N - (n-1) \times \lfloor \frac{N}{n} \rfloor & i = n \end{cases} \tag{3.25}$$

ここで，N_i は i 番目の領域のトークンの数を表し，$\lfloor \cdot \rfloor$ は床関数である．その後，L を n 個の部分に分割する．Fig. 3.17 にこの操作の例を示す．ここでは，$n = 3$ であるため，特徴マップ (Feature Maps) から 3 つの特徴ベクトル (Feature Vectors) に分ける．最後に，これらの特徴ベクトルを利用して学習を行う．

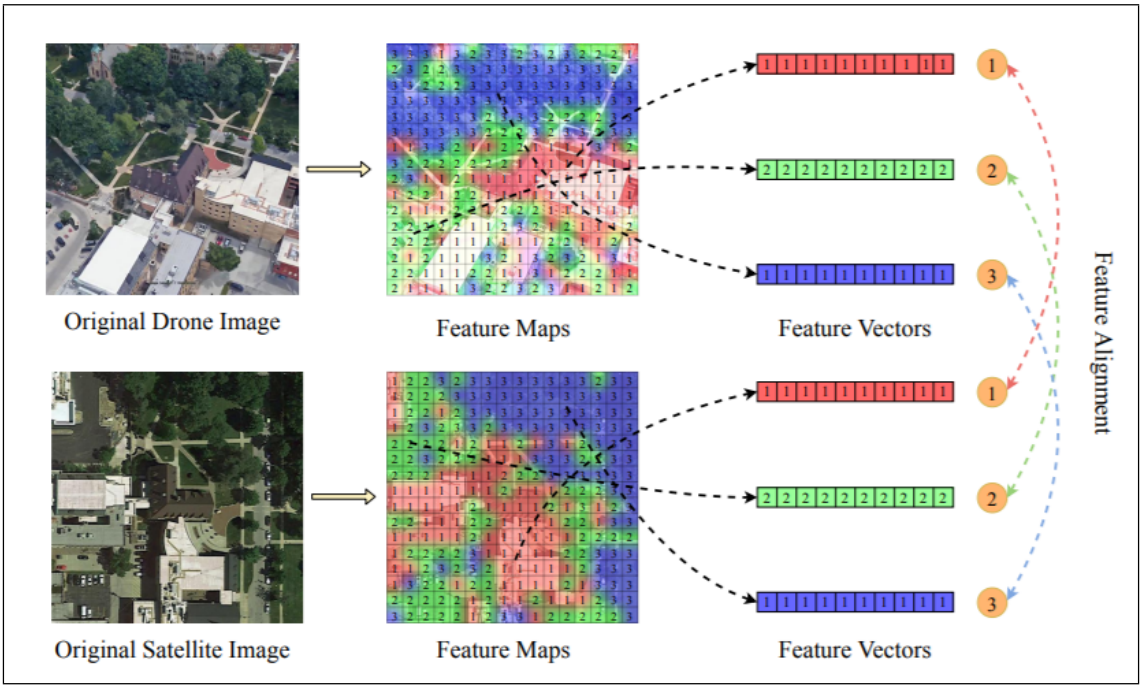

Fig.3.17 Feature Alignment (FSRA)

Semantic Guidance Module (SGM) [62]: SGM のモデルを Fig. 3.18 に示す．Baseline 及び LPN と違い，SGM モデルの構造は，UAV ビューと衛星ビューの 2 つのブランチを持ち，特徴抽出器は Transformer ベースの Swin-Transformer [68] とする．特徴抽出器を通じて 64×768 のサイズの特徴マップ M_i^j が取得される．
SGM モジュールはチャネル方向で M_i^j を合計し，次の M_i が計算される：

$$M_i = \sum_{j=0}^{768} M_i^j \quad i \in [0, 63]. \tag{3.26}$$

ここで，M_i のサイズは 64×1 となる．M_i は次のように正規化される：

$$M_{i'} = \frac{M_i - \ Minimum(M_i)}{Maximum(M_i) - \ Minimum(M_i)} \tag{3.27}$$

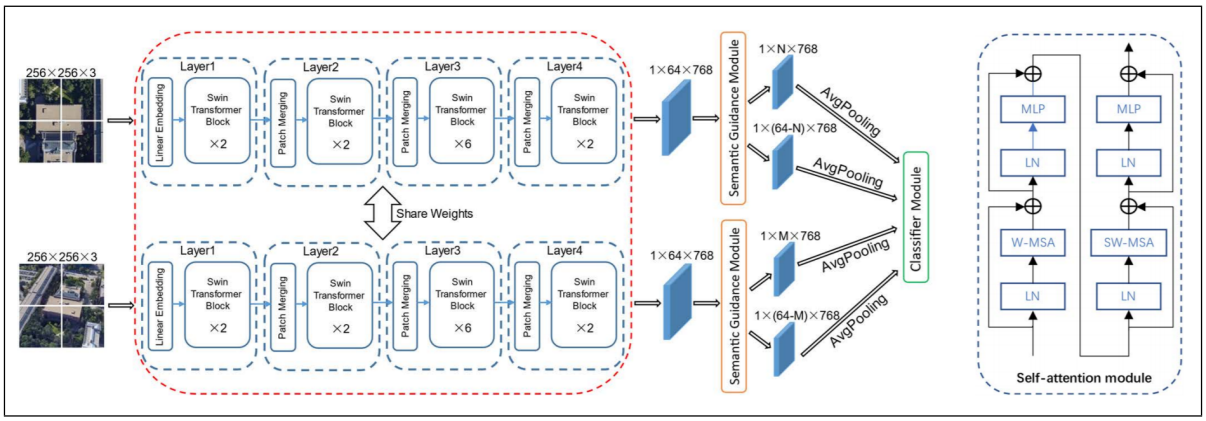

Fig.3.18　The architecture of SGM

ここで，$M_{i'}$ は正規化された M_i であり，$Maximum$ と $Minimum$ がそれぞれ M_i の最大値と最小値を取得することを表す. 次に，SGM は特徴マップの発熱量を計算し，計算結果に基づいて特徴マップを異なる領域に分割する. Fig. 3.19 はこの操作の例を示す．Fig. 3.19 (a) の赤い矢印は，勾配が大きい位置を示す．SGM は，この位置を基準位置とみなし，特徴マップを黄色と緑色の部分 (Fig. 3.19 (b) に示す) に分割する．

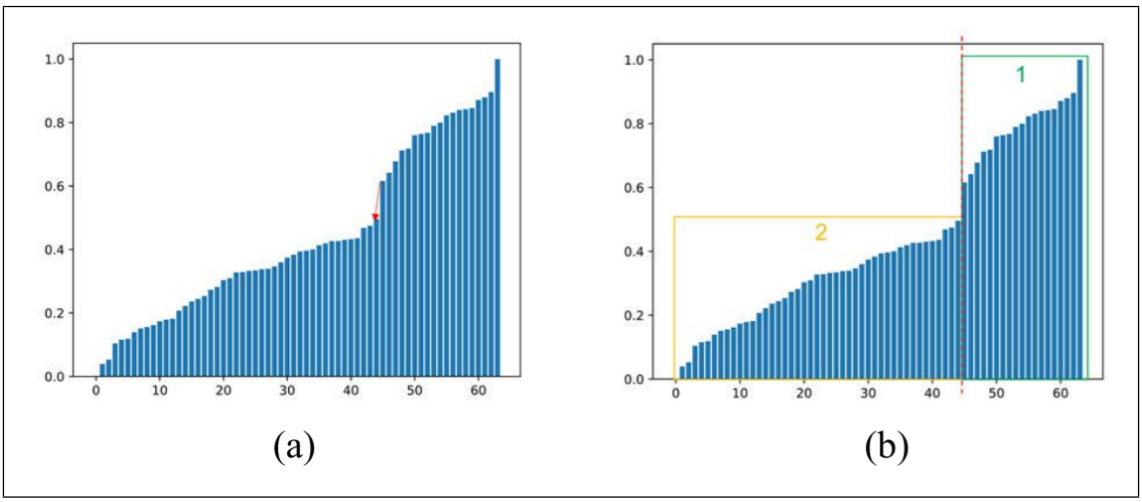

Fig.3.19　Semantic Guidance Module (SGM)

基準位置 i'_{position} は，以下の式で決められる：

$$i'_{\text{position}} = argmax_{i'}\left(\frac{M_{i'+1} - M_{i'}}{M_{i'}}\right) \tag{3.28}$$

SGM により，特徴マップは明らかに建物 (前景) と周りの環境 (背景) の 2 つのパーツに分割される．この結果は，画像のコンテキスト情報を抽出するための優れた基盤となり，オフセットとスケールに対して堅牢であると考えられる．最後に，各パーツに対して Average Pooling 操作を実行し，分類モジュールに送信して学習を行う．

3.3.2 本研究のアプローチ

まず，上記の既存手法 (FSRA, SGM) から次の共通点を指摘する：

1. **特徴抽出器:** Transformer ベースのモデル (Vision Transformer, Swin Transformer) を特徴抽出器として利用する．
2. **特徴処理法:** 入力画像をピクセルレベルで発熱量を注目して処理する．

これらの点を踏まえて，TATN は以下の方針に基づいて提案モデルを構築する：

1. **ネットワーク構造:** TATN の構造は 2 つのブランチで（UAV・衛星）構成される．
2. **特徴抽出器:** ViT を特徴抽出器として利用する．
3. **特徴処理法:** 既存の Transformer ベースの研究は入力画像をピクセルレベルで注目して各トークンの配置を調整したが，ViT から出力された学習可能な埋め込 $\mathbf{x}_{cls}$ の最終出力を十分に活用していないと見られる．本論が提案した手法では，グローバルトークン（Transformer Encoder に埋め込みトークンを入力することで収集されたトークン）とローカルトークン（Transformer Encoder に画像パッチを入力することで収集されたトークン）の組み合わせる方法を導入し，その効果を確かめる．

3.3.3 提案手法

Fig. 3.20 に提案モデル (TATN) のアーキテクチャを示す．このネットワークは 2 つのブランチがあり，それぞれのブランチを UAV 画像のブランチと衛星画像のブランチとする．各ブランチに事前学習された Vision Transformer (ViT)（初期値の重みを共有する）を特徴抽出器として設置する．また，各ブランチの抽出器の後にトークン強化法という特徴処理戦略を導入し，最後に，両ブランチの特徴マップを共有の分類モジュールに転送する．

TATN の主要な点は，以下である：

- 特徴抽出器：自己注意機構を持つ ViT の利用．
- トークン強化法 (Token Enhancement)：ViT が出力したトークンを組み合わせて再配置戦略を採用する．
- 損失関数：Cross-Entropy Loss の利用．

特徴抽出器

Fig. 3.21 に 特徴抽出器の操作を示す．

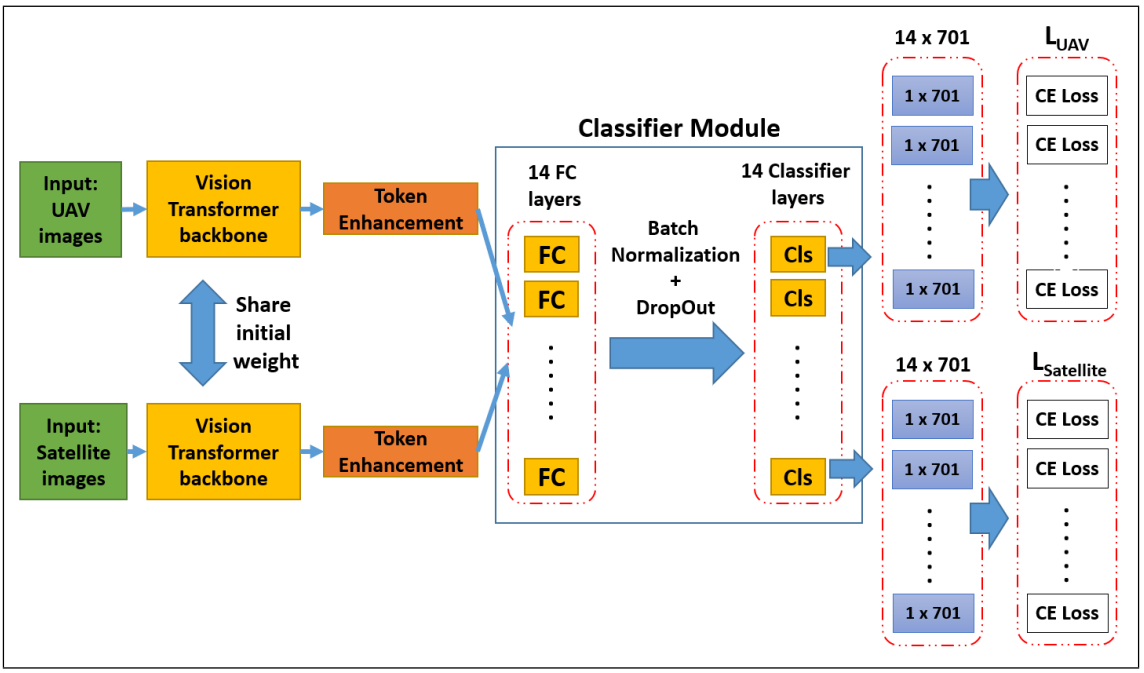

Fig.3.20　The proposed architecture of TATN

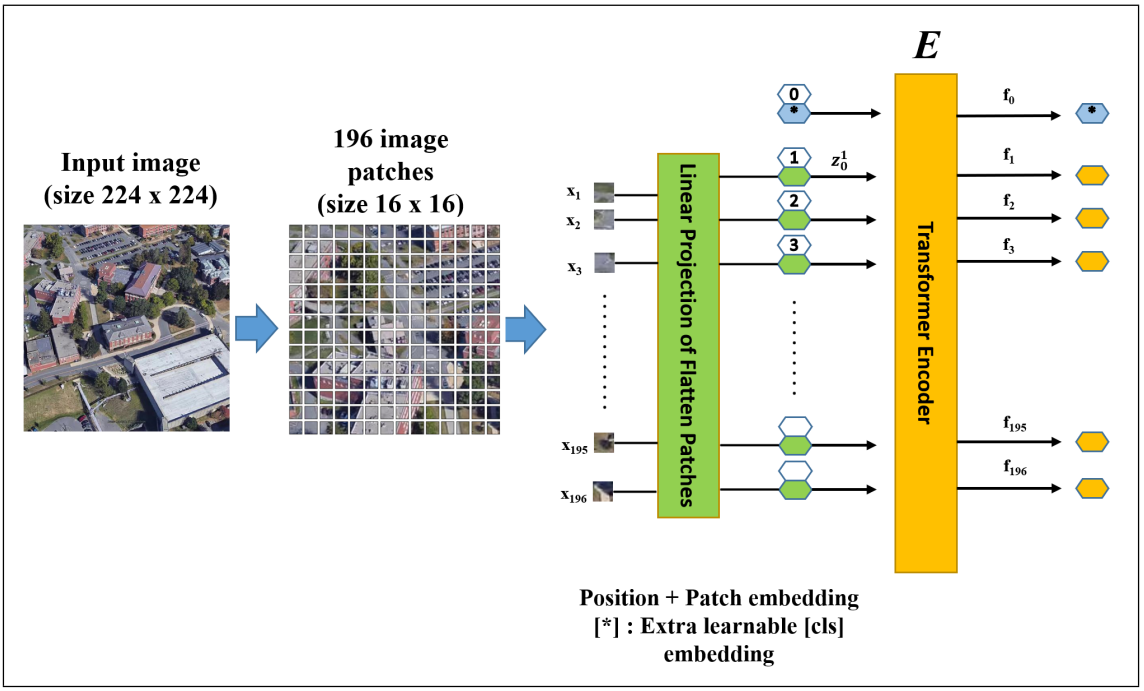

Fig.3.21　The Vision Transformer Backbone.

まず，入力画像を $\mathbf{x} \in \mathbb{R}^{H \times W \times S}$ として, 画像を複数のパッチに分ける:

$$\mathbf{x}_p^i | i = 1, 2, \cdots, N \tag{3.29}$$

パッチの数 N は次のように計算する:

$$N = \frac{HW}{K^2}. \tag{3.30}$$

ここで，H と W は入力画像の高さと幅であり, K_H と K_W は，カーネルストライドの高さと幅であり, S はカーネルストライドである. 今回の TATN は，入力画像サイズを 224 × 224，パッチサイズを 16 × 16 として扱った. カーネルストライドの高さと幅 K_H, K_W 及び カーネルストライド S はパッチの大きさと等しく, したがってパッチの数 N は 196 である．パッチを分けた後, Linear Projection Flatten Patches は，パッチ埋め込み関数 E を使用して，これらのパッチを D 次元の線形に変換する．

$$\mathbf{E}(\mathbf{x}_p^i)|i = 1, 2, \cdots, N. \tag{3.31}$$

これらのパッチを Transformer Encoder に転送する前に，学習可能な埋め込みトークン (いわゆる分類トークン)，x_{cls} を追加し，位置埋め込み (Position) P を融合し，最終のベクトル z_0 を次のように定義する：

$$\mathbf{z}_0 = [\mathbf{x}_{cls}; \mathbf{E}(\mathbf{x}_p^1); \mathbf{E}(\mathbf{x}_p^2); \ldots; \mathbf{E}(\mathbf{x}_p^N)] + P \tag{3.32}$$

ベクトル $\mathbf{z}_0$ は，複数の Transformer ブロックで構成される Transformer Encoder F に転送される. 出力を $O \in \mathbb{R}^{B \times S \times C}$ とする．ここで，B はバッチサイズであり，S はパッチサイズであり，C はパッチの特徴ベクトルの長さである. この出力には，$N+1$ 個の特徴ベクトルが含まれ，次のように定義する：

$$F(\mathbf{z}_0) = [f_0, f_1, f_2, \cdots, f_N] \tag{3.33}$$

次節では，トークン強化法の詳細について説明する.

トークン強化法

ViT が抽出したトークンの強化は，入力画像から包括的な情報を抽出する上での困難さに対処するために重要である．ViT はパッチベースで画像を処理するため，トークンが広範なコンテキストを理解するのに制約が生じる可能性がある．この制約を改善するため，追加の処理法がトークンの埋め込みを補完し，画像のグローバルなコンテキストを理解できる．そのため，ViT が提案された後，ほとんどの Transformer ベースの研究は Multi-head Self-Attention (MHSA) ブロック（第 2.2.2 項に説明した）の強化に焦点を当てていた [68][69][70] が，一部の研究者は生成されたトークン間の相関に注目した．Beal ら [71] は，物体検出用の空間特徴マップを作成するために初めてローカルトークン（Transformer Encoder に画像パッチを入力することで収集されたトークン）を組み合わせた．Jiang ら [72] は，トークンのラベル付け方法を提案した．この手法は，すべてのトークンを利用するために，各トークンの属性に基づいてトークンにラベルを付ける新しいトレーニング法である．Yuan らの Tokens-to-Tokens

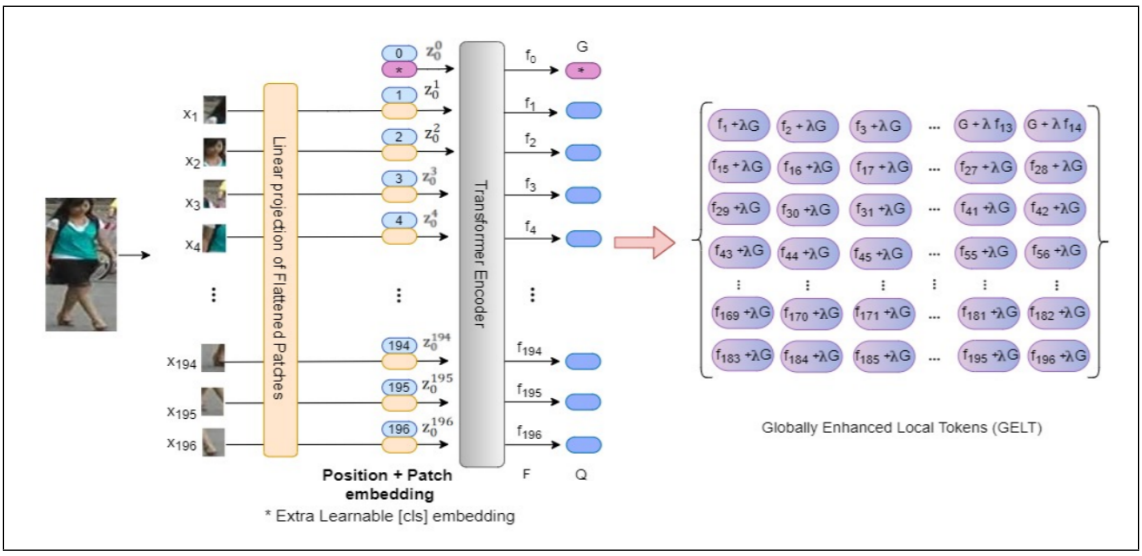

Fig.3.22　Token Enhancement (LA Transformer)

Vision Transformer (T2T-ViT) [73] は，隣合うトークンを 1 つのトークンに融合する手法で，徐々に入力画像から構造化された表現を抽出し，同時に計算量を軽減する．

本研究のアプローチは，人物再同定の分野に応用された LA-Transformer [74] のアイデアを導入する．Fig. 3.22 に示すように，[74] では，グローバルトークン $G = \mathbf{f}_0$（Transformer Encoder に埋め込みトークンを入力することで収集されたトークン）をローカルトークン $\mathbf{f}_i (i = 1, 2, \cdots, 196)$（Transformer Encoder に画像パッチを入力することで収集されたトークン）と組み合わせて，出力の特徴マップを 2D グリッドの形式で全てのトークンを再配置することで，Globally Enhanced Local Tokens (GELT) という特徴マップを作成した．最後に，この GELT を用いて学習を行う．

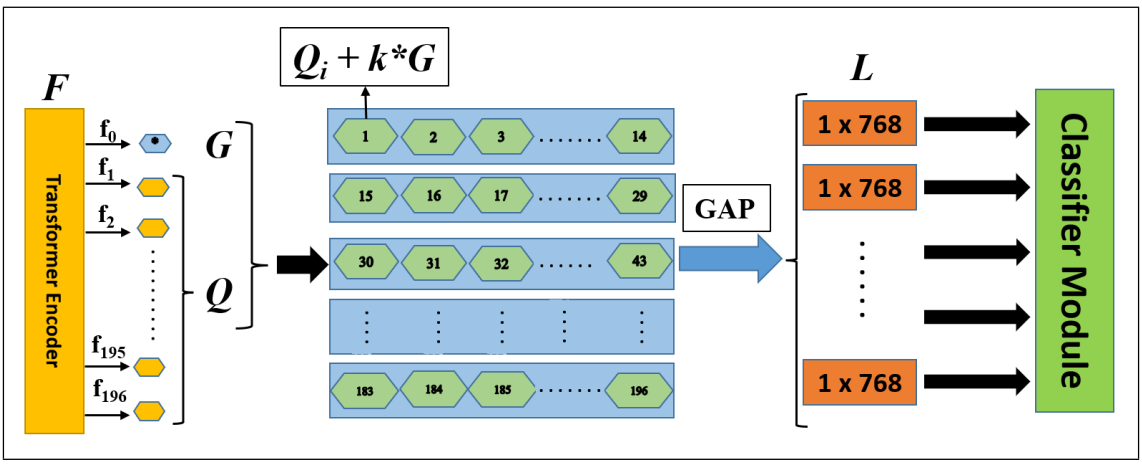

Fig.3.23　The proposed Token Enhancement Strategy (TATN)

Fig. 3.23 に，本研究が利用したトークン強化法を示す．Transformer Encoder は $(N+1)$ トー

クン $[\mathrm{f}_0, \mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_N]$ を出力する．ここでは，グローバルトークン $G = \mathbf{f}_0$ 及びローカルトークン $Q = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_N]$ とする．トークン強化法では，各ローカルトークン $Q_j, (j = 1, 2, \cdots, N)$ とグローバルトークン G を組み合わせて，組み合わせたトークン $(Q_j + k * G)$ を 2D グリッドの形式で再配置する．その後，Global Average Pooling（以下：GAP) を実行し，最後にトークンの 2D グリッドを $\sqrt{N} = 14$ 個の特徴ベクトルに分割する．GAP プロセスの後に得られる最終の各特徴ベクトル L_i を次のように定義する：

$$L_i = \frac{1}{N_R} \sum_{j=iN_R+1}^{(i+1)N_R} \frac{Q_j + kG}{1+k} \qquad i = 0, 1, \cdots, N_R - 1 \tag{3.34}$$

ここで，Q_i と G はローカルトークンとグローバルトークンである．N_R と N_C は，それぞれ 2D グリッドの行と列のパッチの数 $(N_R = N_C = \sqrt{N})$ である．k は強化されたトークンにおけるグローバルトークン G の重要性を示すハイパーパラメータである．最後に，L を分類モジュール (Classifier Module) に転送する

分類モジュール

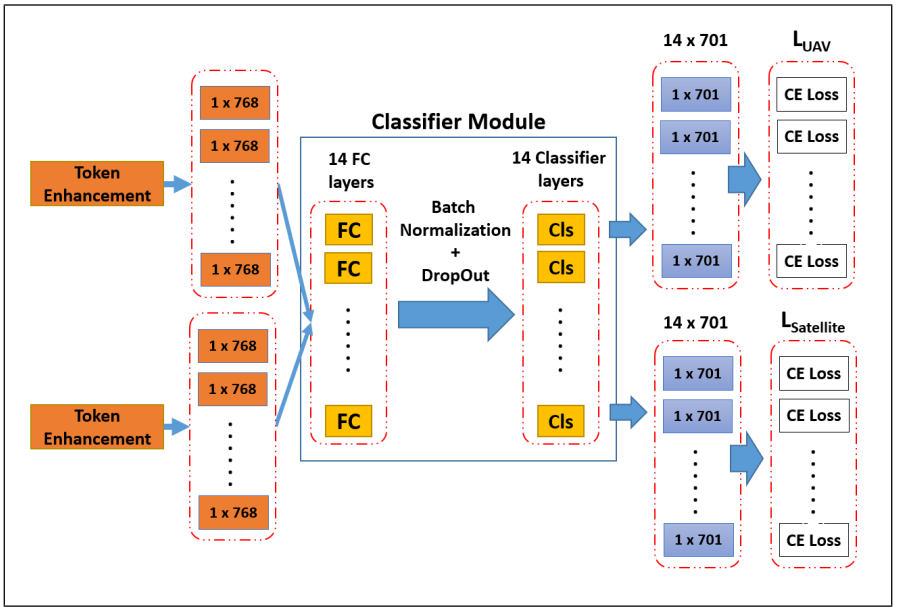

Fig.3.24 The proposed Classifier Module (TATN)

分類器モジュール (Fig. 3.24 の Classifier Module) は次のように構成される：14 個の FC レイヤー（入力サイズと出力サイズはそれぞれ 768 と 512 である)，バッチ正規化，DropOut レイヤー，14 個の分類レイヤー（入力サイズと出力サイズが 512 と 701 である)．総合の損失関数 L_{final} は次のように計算される：

$$L_{\text{final}} = \sum_{i=0}^{N} L^i_{\text{UAV}} + L^i_{\text{Satellite}} \tag{3.35}$$

ここで，L_{UAV} と $L_{\mathrm{Satellite}}$ は各ビューで計算された損失であり，N は分類モジュールが出力した特徴ベクトルの数 (以下では 13 となる) である．

3.3.4　実験設定

トレーニングフェーズ

トレーニングでは，入力データに対しデータ拡張 (Cropping, Rotation) を使用した．オプティマイザーには，運動量 (Momentum) 0.9 の確率的勾配降下法 (Stochastic Gradient Descent - SGD) を採用した．初期学習率は 10^{-4} とし，40 番目と 70 番目のエポック後に 10 分の 1 に減少させた．モデルの訓練期間を 120 エポックとした．DropOut 率は 0.75 とした．トークン強化法のハイパーパラメータ k は 8 とした．テストの際，分類器モジュールの最終的な分類器レイヤーを削除し，出力の特徴が連結されて最終の 1 つの特徴マップを生成する．全てのプログラムは Pytorch フレームワークで構成し，NVIDIA Titan XP で実行した．

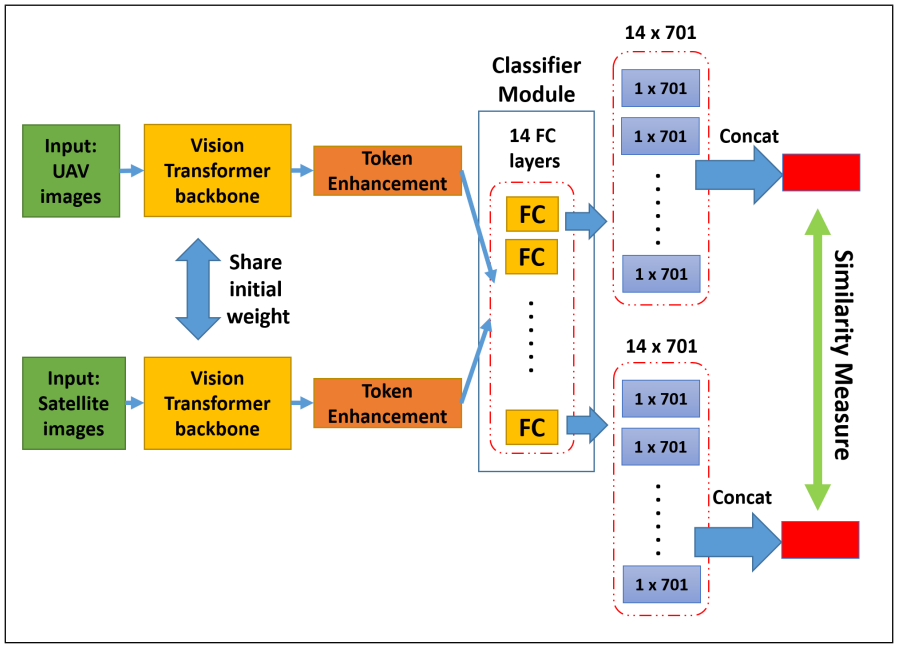

Fig.3.25　Testing phase (TATN)

テストフェーズ

テストの際 (Fig. 3.25)，分類器モジュールの最終的な分類器レイヤーを削除し，出力の特徴を連結して一つにまとめた最終の 1 つの特徴マップを生成する．ユークリッド距離を用いて Query 画像及び Gallery 画像の距離を計算し，Gallery の中で最も Query 画像を類似している画像を探索する．評価指標に関しては，Recall@1 及び AP を利用する．

3.3.5 実験結果

Table3.6 Comparisons of the proposed method TATN with state-of-the-art methods on University-1652.

Method	Backbone	Resolution	Testing inference time	Task			
				UAV → Satellite		Satellite → UAV	
				Recall@1	AP	Recall@1	AP
Baseline [67]	ResNet-50	256 × 256	1.00×	58.49	63.13	71.18	58.74
LPN [64]	ResNet-50	256 × 256	1.00×	74.16	77.39	85.16	73.68
LPN [64]	ResNet-101	256 × 256	1.51×	76.13	79.29	85.45	75.45
PAAN	ResNet-50	256 × 256	1.51×	84.51	86.78	91.01	82.28
SGM [62]	Swin-Tiny	224 × 224	-	79.59	82.50	87.73	79.59
SGM [62]	Swin-Tiny	256 × 256	1.04×	82.14	84.72	88.16	81.81
FSRA [63]	Vit-S	224 × 224	1.21×	80.81	83.65	87.73	80.02
FSRA [63]	Vit-S	256 × 256	1.21×	84.51	86.71	88.45	83.37
TATN	Vit-S	224 × 224	0.89×	83.88	86.25	90.87	83.65
TATN	Vit-B	224 × 224	1.28×	87.33	89.28	90.16	86.93
TATN	Vit-L	224 × 224	1.50×	**88.18**	**89.99**	**91.30**	**87.44**

提案されたモデルと先行研究の結果との比較を Table 3.6 に示す． Table 3.6 より， UAV → Satellite のタスクで は，提案モデルが R@1 精度の 86.00% と AP 精度の 88.12% に達した．そして， Satellite → UAV のタスクでは， R@1 精度の 91.44%， AP 精度の 86.31% を達成した. この結果は全ての ResNet-50 ベースモデルを上回った．これにより， Transformer ベースのモデルは CNN ベースのモデルで同じパフォーマンスを達成でき，さらに優れたパフォーマンスを発揮できることが確認された．また，他の Transformer ベースの SGM [62] 及び FSRA [63] と比較して，提案モデルはそれぞれのモデルに対し約 4% と 3% 上回った．特に，大きなサイズの ViT（ViT-L）を利用した時，提案モデルは最大の精度を達成した： R@1 精度の 91.30%， AP 精度の 86.31%（UAV → Satellite タスク）．この結果は，全ての関連手法に対しほぼ 4% 上回った．これらの結果から、 ViT の入力画像の全体情報を含むグローバルトークンは最終的な特徴表現に大きな影響を与える可能性があると言える．また， 適切なトークン強化法をを利用ことで FSRA [63] と SGM [62] よりも効果的であることを確認した．

3.3.6　考察

本節では，いくつかのアブレーションスタディを実施し，その結果で提案手法の各要素を評価する．

トークン強化法の性能評価

Table3.7　Ablation study on the influence of global and local tokens with different backbones. The best accuracy is highlighted in **bold**

Backbone	Tokens	Task			
		UAV → Satellite		Satellite → UAV	
		Recall@1	AP	Recall@1	AP
ViT-S	Local	77.61	80.77	84.74	77.98
Vit-S	Local + Global	83.88	86.25	90.87	83.65
ViT-B	Local	81.19	83.99	88.30	81.72
ViT-B	Local + Global	87.33	89.28	90.16	86.93
ViT-L	Local	85.11	87.29	90.30	85.11
Vit-L	Local + Global	**88.18**	**89.99**	**91.30**	**87.44**
Swin-S	Local	79.09	82.06	85.59	78.27
Swin-S	Local + Global	79.52	82.45	87.02	79.09
Swin-B	Local	83.97	86.41	88.45	83.61
Swin-B	Local + Global	83.88	86.25	90.87	83.65
Swin-L	Local	84.72	87.04	90.30	84.74
Swin-L	Local + Global	84.38	86.68	90.30	83.81

Transformer ベースのモデルにおけるグローバルトークンとローカルトークンの影響を詳しく理解するために，Transformer ベースの特徴抽出器とトークンの組み合わせでモデルのパフォーマンスを検証した．ここでは，特徴抽出器は Transformer ブロック数（8，12，24）を持つ ViT (ViT-S, ViT-B, ViT-L) 及び ImageNet チャレンジで ViT を上回った Transformer ベースの Swin Transformer の Swin-S, Swin-B, Swin-L を利用し，合計 6 種から選んだ．

Table 3.7 に示すように，トークン強化法は，UAV → Satellite タスクと Satellite → UAV タスクの両方において ViT ベースのモデルに対して優れた効果が出た．ViT-S 及び ViT-B ベースのモデルでは，約 5% の精度向上があり，ViT-L ベースのモデルでは約 1 〜 3% の精度向上ができた．これらの結果から，ViT のグローバルトークンは画像の特徴表現にも大きな影響

を与え，FSRA や SGM のトークンの再配置よりも効果的であると考えられる．一方，Swin ベースのモデルにトークン強化を適用した場合，Swin-S ベースのモデルを除いて，ほとんどの Swin ベースのモデルは UAV → Satellite タスクで Recall@1 と AP の精度が低下し，Satellite → UAV タスクでは約 1% の精度向上を達成した．この結果により，トークン強化法は Swin ベースのモデルには良い影響を与えなかったと考えられる．

そして，Swin ベースのモデルにトークン強化法を適用した場合，Swin-S ベースのモデルを除いて，UAV → Satellite タスクにおいて Recall@1 及び AP の精度が低下し，Satellite → UAV タスクでは約 1% の精度向上があった．この結果から，トークンの強化法は Swin ベースのモデルには肯定的な影響を与えなかったと言える．この問題については，Swin Transformer モデルが自己注意機構を適用する方法が原因だと考えられる．Swin Transformer (Fig. 3.26) では，Transformer ブロック内の標準の Multi-head Self-Attention（MHSA）に対しシフトウィンドウ (Shifted Window) に基づいた Swin Transformer ブロックに置き換えて構築された．これらのシフトウィンドウは，階層的な特徴マップ（層ごとにマージされた特徴マップ）を構築し，効果的に特徴マップの空間次元を 1 つの層から別の層に減少させる．Fig. 3.27 に示すように，これらの操作から生成された特徴マップは ViT の特徴マップと異なり，元の入力パッチからの空間情報を保持していないため，提案された方法のトークン強化法に大きく影響を与えると考えられる．

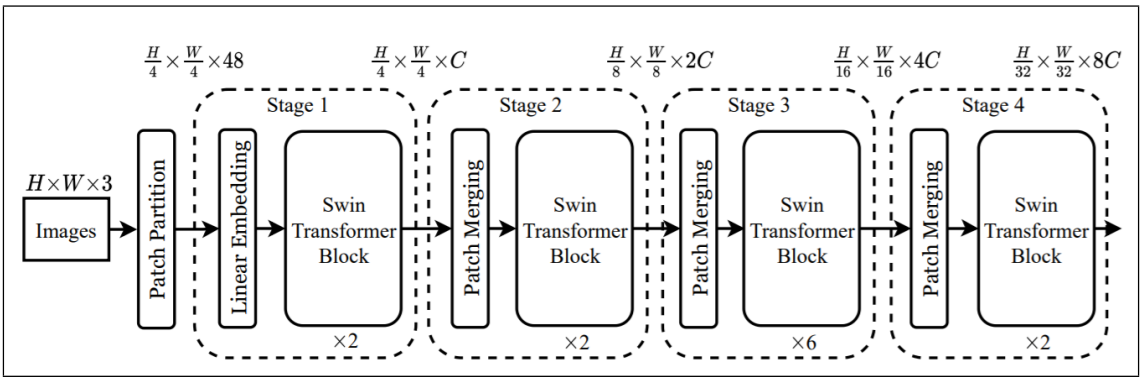

Fig.3.26 Architecture of the Swin Transformer

グローバルトークンの性能評価

トークン強化法では，ハイパーパラメータ k はグローバルトークンの重要性を示す要素である (式 3.34 の k である)．デフォルトでは，$k = 8$ を設定する．k の精度への影響を検証するために，異なる k の値（0.5 から 20 までの範囲）で実験を行った（Fig. 3.28）．結果として，UAV → Satellite タスクでは，k の値が増加すると，Recall@1 と AP の精度が上昇し，$k = 12$ のときに最高の精度（87 %と 89 %）に達し，その後急激に低下する傾向がある．一方，Satellite → UAV タスクでは，k が 6 未満の場合，Recall@1 と AP の精度が向上し，k が増加すると非

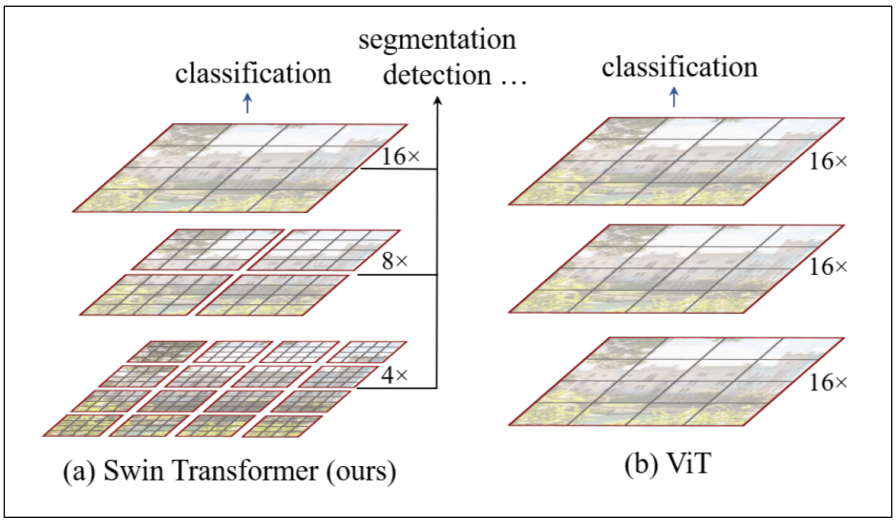

Fig.3.27　Difference between feature map of (a) Swin Transformer and (b) Vision Transformer

常に不安定になる．これらの結果により，k が大きくなると，グローバル分類トークンの情報がローカルトークンの情報を圧倒する可能性があり，それによって各パッチから抽出される特徴表現に影響を与えると考えられる．そのため，トレーニングフェーズでの k の設定は非常に重要であり，この現象については，今後の調査が必要と考えられる．

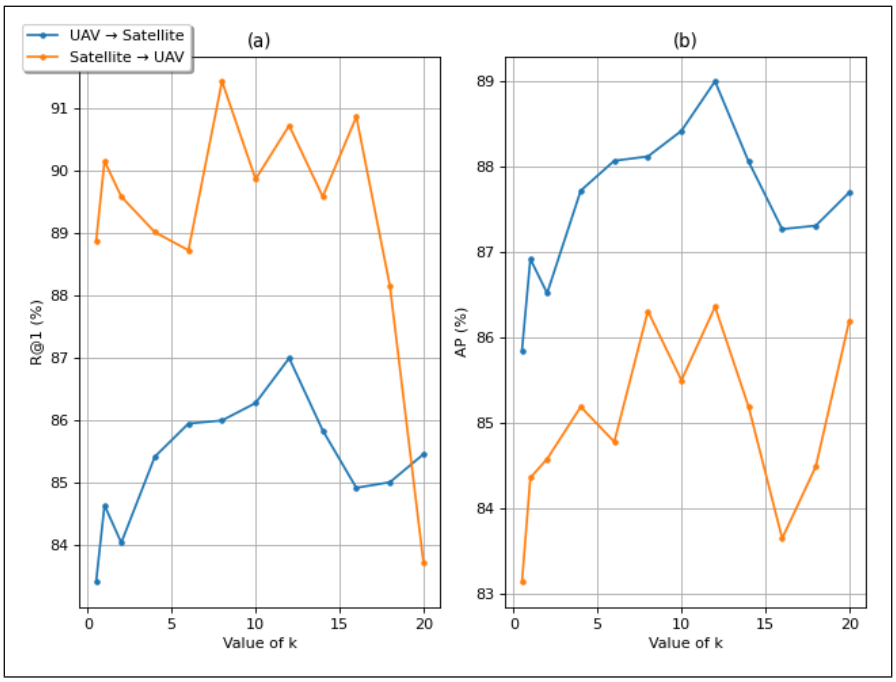

Fig.3.28　Compare the effect of the hyperparameter k on two task: UAV → Satellite (blue line) and Satellite → UAV (orange line). (a) Show the effect of hyperparameter k on the accuracy of Recall@1. (b) Show the effect of hyperparameter k on the accuracy of AP.

検索結果の可視化

実際のミッションに対する提案手法の信頼性を確認するために，今回も University-1652 の 4K 解像度画像に対して検索を行う（Real UAV 画像 → 衛星画像）．正解の画像は黄色のボックス，不正解の画像が青いボックスに表示される．Fig. 3.29 に示すように，衛星 Gallery には各場所につき 1 つの画像しか含まれていないため，正しい答えを見つけるのは難しいが，提案されたモデルは正解の画像を見つけることができた．この結果により，提案手法の TATN がシミュレーション画像で訓練されたとしても実環境の画像で優れたパフォーマンスを発揮ができ，実際の UAV ミッションに適用できることが可能と考えられる．

Fig.3.29 Visualization on 4K-image UAV data (TATN)

3.3.7 結論

本節では，Transformer ベースの既存手法のアイデアに基づいて，新しい ViT ベースの TATN モデルを提案した．具体的には，このモデルは PAAN のような特徴分割法や新たな Pooling 法を導入することではなく，ViT が出力した全てのトークンをトークン強化法で処理して再調整することで，既存手法と比べて大幅に精度の向上ができた．

3.4　第 3 章における結論

UAV 技術の発展に伴い，UAV の自律制御，特に GPS 信号を使用せずに UAV をナビゲートする必要性が急速に高まっている．画像ベースの場所推定は，この問題の重要な解決策の一つである．本章では，UAV における画像ベースのクロスビュー場所推定の改善に焦点を当てた．

本章では，まず UAV におけるクロスビュー場所推定において，既存研究の CNN 及び ViT ベースのモデルを紹介した．これらの手法の欠点を明らかにし，2 つの深層学習ベースの PAAN と TATN を開発した．CNN ベースの PAAN は，複数の深層学習のテクニックを利用し，ViT ベースの TATN は新たなトークン強化法を導入した．両モデルは，ベンチマークデータセットから，既存の深層学習ベースのモデルと比べ，より優れた性能を見せた．特に，実環境のデータで検証したところ，提案モデルも正解の画像を見つけることができた．この結果より，提案モデルは実際の UAV に応用する可能性が高いと考えられる．

次章では，UAV におけるオブジェクト再同定タスクに取り組み，改善方法を提案する．

第 4 章

無人航空機のためのオブジェクト再同定

オブジェクト再同定は，時間と空間が超えて異なるカメラの視点で捉えられた同じオブジェクトまたは人物の画像を一致させることを目指す．近年，このタスクは，画像処理における基本的でありながら今も難しい課題であり，ビデオ監視，自動運転，スマートシティ等の幅広い発展が期待されている．

初期の研究は，異なるカメラの視点での変化に対して不変で，頑健で識別的な特徴を抽出することに焦点を当てている．色のヒストグラム [75] や SURF 特徴 [76] 等の低レベルの特徴が使用された．また，最大安定色領域 [77] や再帰パッチモデル [78] 等の中間レベルの特徴を利用した研究によって進歩がもたらされた．近年では，深層学習手法が広く使用され，オブジェクトの画像から高レベルの意味的な特徴を抽出することが普及した．特に，既存研究 [79][80][81] は識別能力が高い距離学習を用いて，オブジェクト再同定のベンチマークデータセットに対し優れた結果を達成した．

しかし，現在の手法を用いてもなお，再同定のタスクでは照明，視点，姿勢，遮蔽の様々な理由で精度が落ちる傾向がある．そして，人物再同定タスクでは衣服の変化や持ち物の状態や非剛体変形等の様々な課題 [25] が，車両再同定タスクでは汚れの蓄積や損傷や照明の変動等 [82] があり，検索が困難になっている．特に，UAV でこれらのタスクを実施する時，高度に依存して取得した画像の画質やオブジェクトの認識が不十分になる可能性が高く，望ましい結果を得ることが難しくなる．また，複数のオブジェクト再同定のモデルを実装することで計算量が増えて，UAV の限られたリソースで対応できない状態が起こることも想定される．そこで，複数のタスクを 1 つのモデルで解決することが望ましいと考えた．

本章では，まずオブジェクト再同定に関する既存研究について紹介する．次に，オブジェクト再同定タスクの精度向上のため，新しい損失関数「Centroid Tuplet Loss」を提案する．また，複数の画像検索タスクを解決できるマルチパーパスに対応した深層学習ベースモデルを提

案する．最後に，本章の成果についてまとめる．

4.1 既存研究

オブジェクト再同定の主要な目的は，異なるポーズや視点等の様々な実環境の条件下でオブジェクトを認識し，一致させるアルゴリズムを開発することである．この分野の既存研究は，通常，特徴抽出（特徴学習アプローチ）または異なる距離学習のテクニック（距離学習アプローチ）に焦点を当てていた．本節では，特徴学習アプローチと距離学習アプローチを詳しく解説し，現在の課題と本研究のアプローチを紹介する．

4.1.1 特徴学習アプローチ

まず，特徴抽出の改善のアプローチでは，学習能力を持つ畳み込みニューラルネットワーク（CNN）アーキテクチャが多く利用されている．その中で，最も利用されるアプローチとしては，グローバル特徴表現の学習 (Global Feature Representation Learning) 及びロカール特徴表現の学習 (Local Feature Representation Learning) が知られている．

- **グローバル特徴表現の学習 [83]：** グローバル特徴表現学習は，各人物画像に対してグローバルな特徴ベクトルを抽出する．グローバル特徴学習は元々画像分類に適用されていたため，オブジェクト再同定の初期の段階ではグローバル特徴学習を応用することが最適なアプローチであった．有名な CNN の VGG16 [84] や ResNet50 [85] がよく利用されており，優れた結果を達成した [86][87]．Qian ら [88] は，グローバル特徴表現の様々なスケールで識別特徴を取得するためのマルチスケール深層表現学習モデルを開発した．IDE モデル [89] は，各人物の ID を個別のクラスとして扱うことにより，トレーニングプロセスをマルチクラス分類問題として扱う．そして，注意機構も表現学習を向上させるために広く研究されている [90]．例えば，ピクセルレベルの注意 [91]，チャネルごとの特徴の重み付け [92]，[93]，[94]，または背景の抑制 [95] が提案された．
- **ローカル特徴表現の学習 [83]：** ローカル特徴表現の学習は，オブジェクトのパーツや領域の集約された特徴を学習し，位置ずれに対して堅牢になる．これらのパーツは，人間による解析・姿勢推定によって自動生成される，またはおおよその水平分割によって生成される．例えば，2018 年に Part-based Convolutional Baseline (PCB) [96] は，特徴分割法を導入して，モデルが画像にある特別部分に注目して認識する能力を向上させている．Fig. 4.1 に示すように，PCB は CNN が出力した特徴ベクトルに対して特徴分割法を適応し，各パーツで学習を行う．

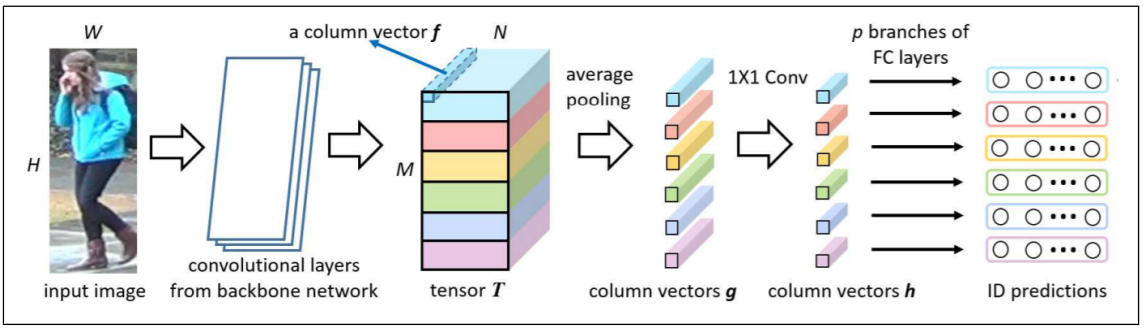

Fig.4.1　The architecture of PCB

4.1.2　距離学習アプローチ

近年の距離学習アプローチでは，CNN モデルの表現学習能力を向上させる損失関数を注目している．例えば，Re-ID の分野では Identity Loss (以降 ID Loss) という損失関数をよく採用している．ID Loss の本質は通常画像分類の問題で使用される一般的な Cross-Entropy Loss である．Re-ID の問題では， ID Loss はトレーニングプロセスを画像分類の問題として扱い，各アイデンティティが異なるクラスとして扱われる．テストの際，Pooling 法，または埋め込み層の出力が特徴抽出器として使用される．入力画像 x_i とラベル y_i が与えられた場合，ID Loss は次のように計算される：

$$L_{\text{IDLoss}} = -\frac{1}{N}\sum_{i=1}^{N}\log(p(y_i|x_i)), \tag{4.1}$$

ここで，$p(y_i|x_i)$ は，x_i がクラス y_i として認識される確率を表し，SoftMax 関数で計算される．N は各バッチ内のトレーニングサンプルの数を表す．ID Loss は，既存の研究で広く使用される．

そして，いくつかの研究では，SoftMax Loss とその他の SoftMax Loss ベースの損失関数 (Sphere Loss [97], AM Loss [98]) ，または対照学習での代表的な損失関数 (Contrastive Loss, Triplet Loss 等) もよく利用されている．対照学習の成功と収束性はサンプル選択に大きく依存するため，慎重なサンプル選択が必要になることが多く，Triplet Loss のために様々な有益なサンプル選択法が設計された．例えば，重み制約がある Positive Sample Mining というサンプル選択法が [99] に提案されており，Triplet Loss をで直接最適化した．Hermans ら [100] は，各トレーニングバッチ内で Hardest Positive Mining (距離が大きい Anchor - Positive ペアでトレーニングする) 及び Hardest Negative Mining (距離が小さい Anchor - Negative ペアでトレーニングする) というサンプル選択法が Re-ID モデルの識別能力に貢献できることを検証した．

また，複数の損失関数の組み合わせ (例：ID Loss，Triplet Loss, Center Loss 等) は多くの Re-ID の研究に利用され，優れた性能を示した [80][101].

4.1.3 既存研究の課題

Triplet Loss の問題

Re-ID に関する研究の中では，最も利用される損失関数が Triplet Loss であるが，いくつかの研究で Triplet Loss [8][9] の欠点を指摘している．その主要な問題の一つは計算コストの高さである．Triplet Loss は，ミニバッチ内の全てのサンプルペアの距離を比較することため，これはトレーニングサンプルの数が増加するにつれて，学習可能なペアと必要な計算の数も増加することになる．Hardest Positive Mining (距離が遠い Anchor - Positive ペアでトレーニングする) 及び Hardest Negative Mining (距離が近い Anchor - Negative ペアでトレーニングする) 等のサンプル選択法が提案され，トレーニングフェーズの収束を速めることができたが，グローバルではなくローカル情報で学習することになり，悪い局所最小値を引き起こす可能性があり，モデルが最高のパフォーマンスを発揮するのを妨げる可能性がある．

UAV におけるオブジェクト再同定

最近，UAV スワーム技術はハードウェアやソフトウェアのコンポーネントにおいて大きな進歩を遂げ，これらの自律システムの能力と汎用性を新たなレベルに引き上げている [102][7]. 特に，機械学習と深層学習の技術は UAV スワームの意思決定能力に重要な役割を果たし，ルートの最適化 [103][104]，障害物の回避 [105]，複数のタスクの同時実行 [106][107][108][109] 等を可能になりつつある．オブジェクト再同定に関しては，特定の場所に固定された監視カメラとは異なり，UAV スワームシステムは様々な地形を自由に移動し，柔軟的にデータを収集できる能力を持っている．これは，多様な領域での監視にとって大きなメリットである．

近年，一部の研究では，UAV 上での Re-ID タスクを実施し始めた [26][27][28]．しかし，UAV が上昇するにつれて，Re-ID の対象物はサイズが小さくなり，認識精度も落ちる傾向がある．また，UAV に関するオブジェクト再同定のデータセットが少ないため，良いモデルを作成することが難しい．

4.1.4 本研究のアプローチ

既存研究の課題を解決するために，次の 2 つの手法を提案する．

- Centroid Tuplet Loss: Triplet Loss の計算コストを削減するために，[110] では，新しい Fast Approximated Triplet Loss (FAT Loss) を提案した．この損失関数では，同じクラスに属する全ての画像をクラスターとして扱い，Anchor，クラスが同じクラスターの中心，及びクラスが異なるクラスターの中心をサンプルとして利用する．すなわち，FAT Loss の主なアイデアは，Triplet のポイントとポイント間の距離の計算 (point-to-point)

をポイントとクラスター間の距離の計算 (point-to-cluster) に置き換える (Fig. 4.2)．FAT Loss は，人物再同定のベンチマークデータセットで Triplet Loss を上回った．

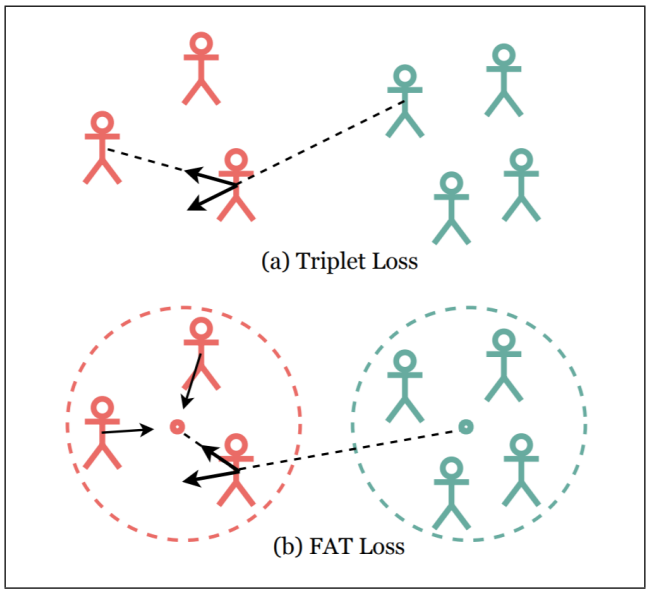

Fig.4.2　(a) Triplet Loss: Point-to-point strategy and (b) FAT Loss: Point-to-cluster strategy

本研究では，FAT Loss のアイデアを参考して，オブジェクト再同定の精度向上のために距離学習に利用できる Centroid Tuplet Loss を提案する．この損失関数は Triplet Loss 理論に基づいて構成される損失関数であり，FAT Loss や Center Loss のクラス中心の概念を活用している．

- マルチパーパスな画像検索のモデル：本論では，クロスビュー場所推定，人物再同定及び車両再同定のタスクを画像検索の問題と見なし，この 3 つのタスクを統合したマルチパーパス画像検索モデルを提案する．このモデルは，これらのタスクを同時に処理する最初に提案されたフレームワークである．提案モデルをトレーニングするために，特徴抽出器の SE-ResNet50 及び新しい Centroid Tuplet Loss を採用する．

次節では，本研究が提案する Centroid Tuplet Loss の詳細内容及びその性能を示す．

4.2 Centroid Tuplet Loss

本節では，まず提案手法の Centroid Tuplet Loss を詳しく解説する．その後，提案手法を評価するためのデータセットを紹介し，実験の設定及びその結果を示す．最後に，考察を述べる．

4.2.1 提案手法

まず，Triplet Loss の使用を一般化するために，次のように Tuplet の定義を導入する：

$$t = (x_a, x_p, x_{n_i}, ..., x_{n_{k-1}}), \forall i = 1, \cdots, k-1, \tag{4.2}$$

ここで，k がミニバッチ内のクラス数であり， x_{n_i} が異なるクラスのネガティブサンプルを示す. この定義により，各 Tuplet $t = (x_a, x_p, x_{n_i}, ..., x_{n_{k-1}})$ には共通の Positive ペア (x_a, x_p) を共有する $(k-1)$ 個の Triplet が含まれていることがわかる：

$$(x_a, x_p, x_{n_i}), \forall i = 1, \cdots, k-1. \tag{4.3}$$

Tuplet の概念を利用することで，Triplet Loss は次の Tuplet Loss のように一般化できる；

$$L_{\text{TupletLoss}} = log\Big(1 + \sum_{i=1}^{k-1} e^{d(x_a,x_p)-d(x_a,x_{n_i})}\Big) \tag{4.4}$$

ここで， $d(\cdot,\cdot)$ は距離計算の操作を示す. 例えば： $d(x_a, x_p) = \|f(x_a) - f(x_n)\|_2^2$.

しかし，Positive と Negative ペアの間のマージンは，特徴埋め込みのノルム $\|f(x)\|_2$ 及び特徴埋め込みの方向の両方から影響を受けると考えられる [111]． このマージンはまた特徴埋め込みのノルムにより上に制限される [112]．さらに，トレーニング中には，簡単なサンプルの特徴埋め込みのノルムを増加させることで，損失関数を最小化させることが簡単になり，難しいサンプルを無視する傾向がある [113]．これらの点を改善するために，Yu ら [112] は新しい Tuplet Margin Loss という Tuplet ベースの損失関数を提案した．この Tuplet Margin Loss では，特徴埋め込みの方向を保持し，特徴埋め込みのノルムを制御するためのスケールファクター s を使用した．Tuplet Margin Loss の式は，以下のように定義される：

$$L_{\text{TupletMarginLoss}} = log\Big(1 + \sum_{i=1}^{k-1} e^{s\left(cos\theta_{an_i} - cos(\theta_{ap}-\beta)\right)}\Big), \tag{4.5}$$

ここで，θ_{an_i} は $f(x_a)$ と $f(x_{n_i})$ の間の角度であり，θ_{ap} は $f(x_a)$ と $f(x_p)$ の間の角度である． スラックマージン $\beta \geq 0$ は，Tuplet Margin Loss の性能を向上させる手段として導入された．検証実験 [112] では，Tuplet Margin Loss は Triplet Loss より大幅に精度の向上ができた．

本論では，Tuplet Margin Loss のアイデアに基づいて，Center Loss や FAT Loss のクラス中心 (本論では Centroid と呼ぶ) の概念を適用し，新しい損失関数「Centroid Tuplet Loss」を提案する．具体的には，上記の Triplet Loss 及び Tuplet Margin Loss は各クラスから Positive のサンプル x_p と Negative サンプル x_n を利用したが， Centroid Tuplet Loss が各クラスのサンプルの代わりに，各クラスの中心 Centroid を利用する．本研究が提案した Centroid Tuplet Loss は以下のように定義される：

- まず，トレーニングフェーズでは，各ミニバッチごとに P 個のクラスがあり，各クラスに M 個のサンプル ($P \times M$ バッチサイズ) が使用され，訓練バッチが作成される．ここで，$\mathbf{T}_k = x_1, x_2, \cdots, x_M$ はトレーニングのミニバッチ内のクラス k のサンプルセットを表し，$x_i \in R^D$ (D が特徴埋め込みのサイズ) がクラス k のサンプルセット内の第 i サンプルの特徴埋め込みを表す．$\mathbf{T}_k$ の各サンプルはクエリ q_k として使用され，残りの $(M-1)$ サンプルはそのクラスのクラスター特徴の平均（いわゆるクラスの中心 Centroid）を計算して，Centroid c_k は以下のように表される:

$$c_k = \frac{1}{|\mathbf{T}_k \backslash \{q_k\}|} \sum_{x_i \in T_k \backslash \{q_k\}} f(x_i), \tag{4.6}$$

  ここで，f は画像を T_k から D 次元の特徴埋め込み空間にエンコードする深層学習モデルである．
- 本研究が提案した Centroid Tuplet Loss は，Anchor x_a と Positive クラスの Centroid c_p, Negative クラスの Centroid c_{n_i} の距離を計算する:

$$L_{\text{CentroidTupletLoss}} = log\Big(1 + \sum_{i=1}^{k-1} e^{s\left(cos\theta_{ac_{n_i}} - cos(\theta_{ac_p} - \beta)\right)}\Big), \tag{4.7}$$

  ここで，θ_{a_p} は，$f(x_a)$ と $f(c_p)$ の間の角度を表し，$\theta_{ac_{n_i}}$ は，$f(x_a)$ と $f(c_{n_i})$ の間の角度を表す．c_p は， Anchor と同じクラスの中心であり，c_{n_i} は，異なるクラスの中心である．

4.2.2　データセット

上記の提案手法を評価するために，4 つのベンチマークデータセット： Market-1501 [114], CUHK03 [115], PRAI [26] 及び VRU [28] を利用して実験を行った．オブジェクト再同定に関するデータセットの詳細内容を Table. 4.2.2 に示す．

- **Market-1501:** 2015 年に提案された Market-1501 は，清華大学で 6 つの監視カメラでキャプチャされたデータセットであり，合計 1,501 の人物（クラス）の 32,668 個の確認がある．トレーニングセットには 751 のクラスが含まれ，テストセットには 750 のクラスが含まれる．Fig.4.3 に Market-1501 のサンプル画像を示す．

Table4.1 Details of PRAI and VRU dataset

	Training set		Testing set	
	Images	Classes	Images	Classes
Market-1501	19,523	751	80,532	750
CUHK03	7,365	767	6,732	700
PRAI	19,523	782	19,938	799
VRU	80,532	7085	91,595	8000

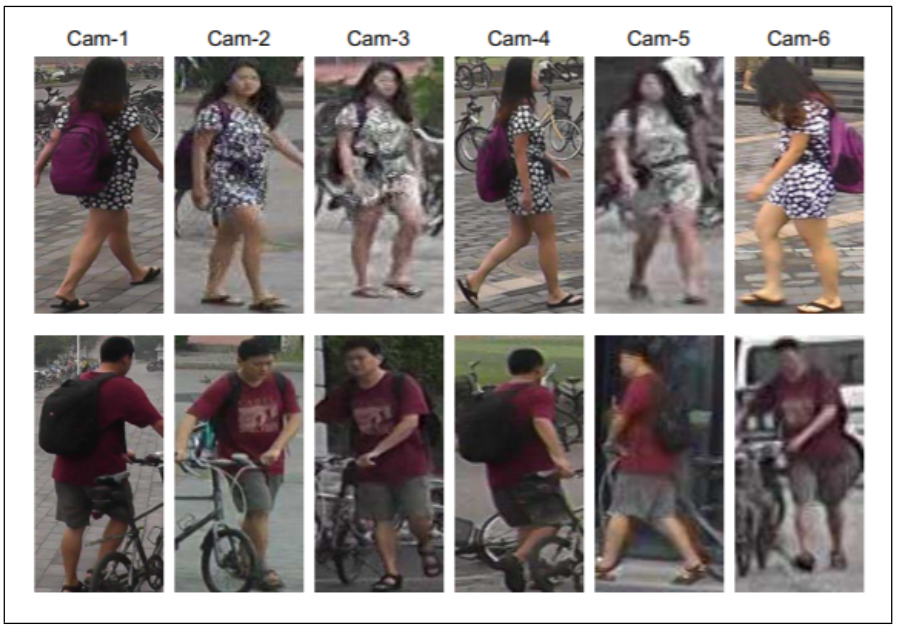

Fig.4.3 Sample images from Market-1501

- **CUHK03:** 2014 年に提案された CUHK03 は，6 つの監視カメラでキャプチャされたデータセットであり，合計 1,467 の人物（クラス）の 14,097 個の確認がある．このデータセットは，767 のクラスがトレーニングに使用され，700 のクラスがテストに使用される．CUHK03 データセットには 2 つのバージョンがあり，「label」バージョンはオブジェクトのボックスが人間によってラベル付けされていることを意味し，「detected」バージョンはオブジェクトのボックスが人物検出器によってラベル付けされていることを意味する．Fig.4.4 に CUHK03 のサンプル画像を示す．
- **PRAI-1581:** 2020 年提案された PRAI は，実用的な UAV 監視シナリオを含む最初の Person Re-ID データセットである．PRAI-1581 には，高さ 20 から 60 メートルの屋外環境で 2 つの UAV によって 39,461 個のボックスを撮影を行い，合計のクラス数は 1,581 である．Fig.4.5 に PRAI-1581 のサンプル画像を示す．
- **VRU:** これまで，UAV による車両再同定に関するデータセットは複数ある（例：VRAI

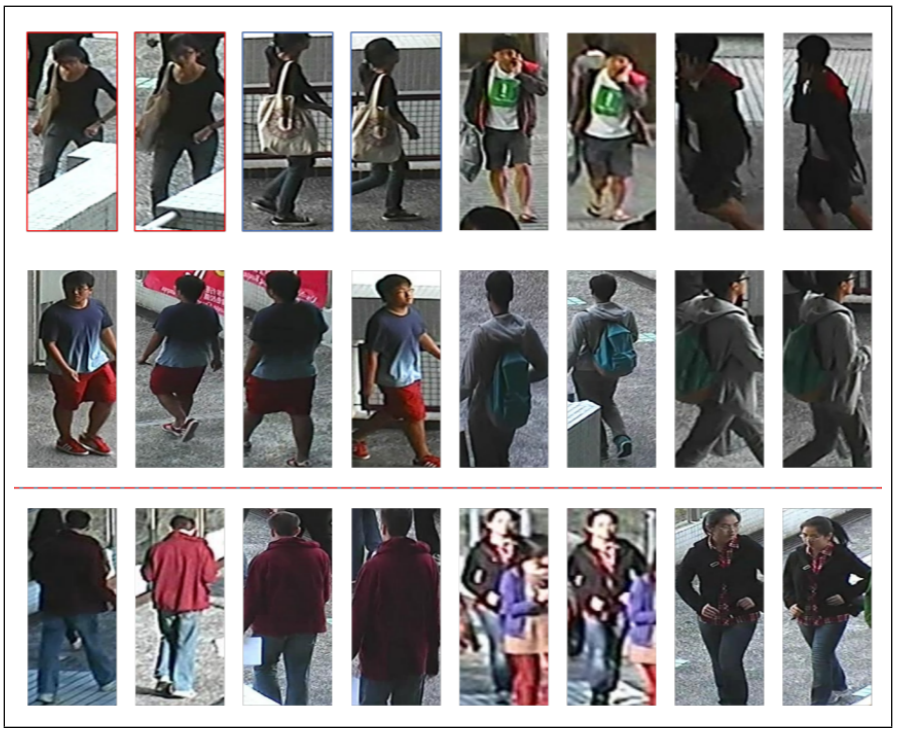

Fig.4.4　Sample images from CUHK03

Fig.4.5　Sample images from PRAI

[116]，UAV-ReID [117])．VRU データセットは異なるシナリオで異なる撮影戦略があり，車両画像の多様性を持つ最初のデータセットである．VRU の作成には，5 台の DJI Mavic 2 Pro を利用し，異なる高度（15 から 60 メートルの間）で道路上の車両を撮影した．Fig.4.6 に VRU のサンプル画像を示す．

Fig.4.6 Sample images from VRU

4.2.3 実験設定と評価指標

実験設定

本実験の目的は，オブジェクト再同定タスクに提案した Centroid Tuplet Loss の性能を検証することである．そのため，各オブジェクト再同定のデータセットに提案モデルを学習させ，既存研究のモデルと比較する．

評価実験では，オブジェクト再同定の分野にある Baseline：Bag-of-trick Baseline [80] に Centroid Tuplet Loss を導入する．Fig. 4.7 に示すように， Bag-of-trick Baseline は単なる ResNet50 が特徴抽出器として利用し，ID Loss や Triplet Loss や Center Loss の組み合わせで学習を行う．

そして，本実験が導入する設定を Fig. 4.8 に示す．ここでは，特徴抽出器 (ResNet50 + Global Average Pooling) が抽出した特徴に対し，Bag-of-trick Baseline のように Triplet Loss 及び Center Loss を計算する．Centroid Tuplet Loss の計算については，各クラスの Centroid を計算し，これらの Centroid で Centroid Tuplet Loss を計算する．最後に，それらの特徴を Fully Connected Layer を通して ID Loss を計算する．総合の損失関数 L_{final} は，次のように計算する：

$$L_{\text{final}} = L_{\text{IDLoss}} + L_{\text{TripletLoss}} + \alpha L_{\text{CenterLoss}} + L_{\text{CentroidTupletLoss}}. \tag{4.8}$$

実験では，Center Loss の重み α は 5×10^{-4} とした．

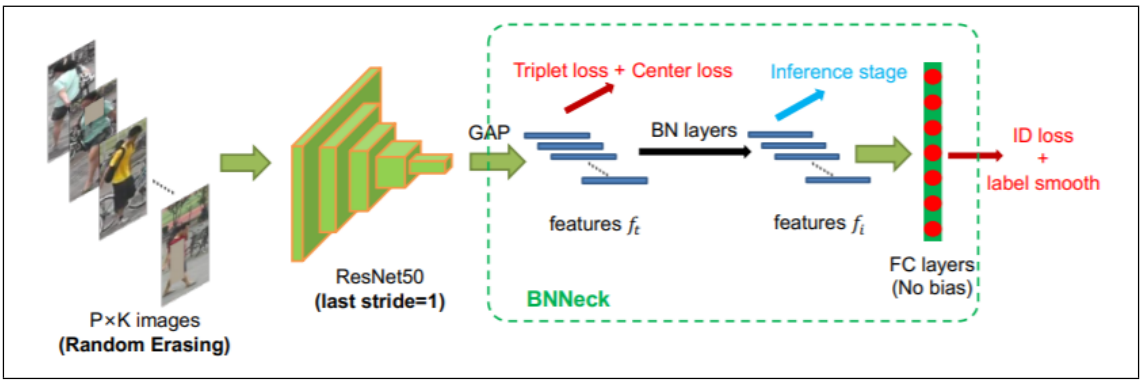

Fig.4.7　Architecture of Bag-of-trick Baseline

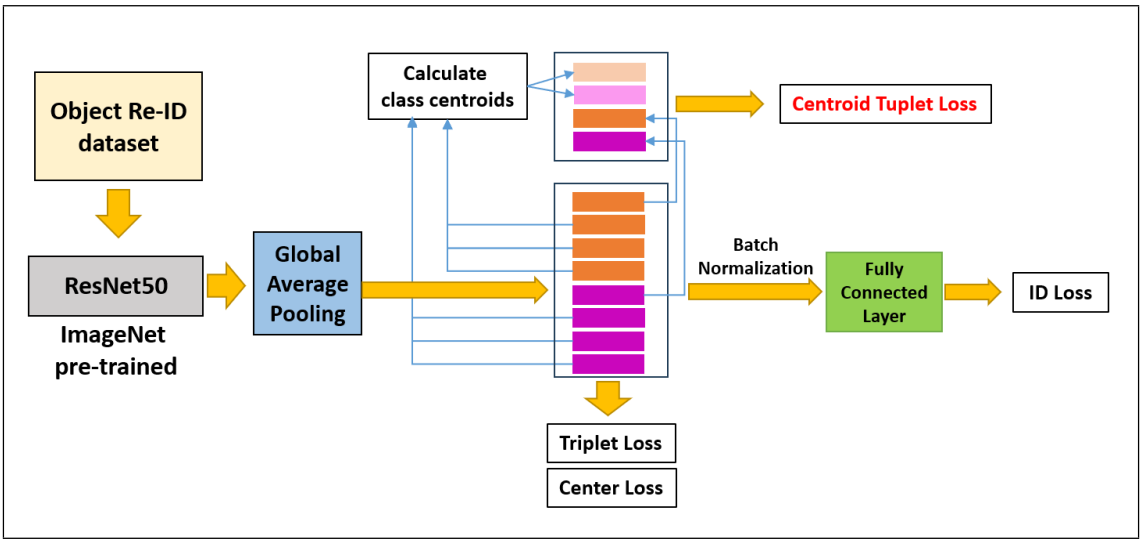

Fig.4.8　Experiment settings for training phase (Centroid Tuplet Loss)

本実験の実装詳細は，以下のようになる：

- **トレーニングフェーズ**：トレーニングの際，全ての入力画像サイズを（256，128）に変更した．また，入力データに対しデータ拡張 (Cropping, Rotation) を使用した．実験では，ResNet50 は ImageNet データセットで事前にトレーニングされた．Centroid Tupet Loss のハイパーパラメータは，Tuplet Margin Loss のハイパーパラメータの値 ($s = 64$, $\beta = 0.10$rad) を利用した．バッチサイズは 16，エポック数は 120 とした．オプティマイザーには，Adam オプティマイザを適用し，初期学習率は 1×10^{-4} で，40 番目と 70 番目のエポック後に 10 分の 1 に減少させた．全てのプログラムは Pytorch フレームワークで構成され，NVIDIA Titan XP で実行した．
- **テストフェーズ**：テストの際，Query 画像が入力として利用し，Fully Connected Layer が削除され，提案モデルが特徴ベクトルを出力する．ここで，事前に Gallery の各クラスの Centroid を用意し，提案モデルが出力した特徴ベクトルと各クラスの Centroid の

ユークリッド距離を計算し，その距離に基づいて検索を行う (Table) Gallery セットの

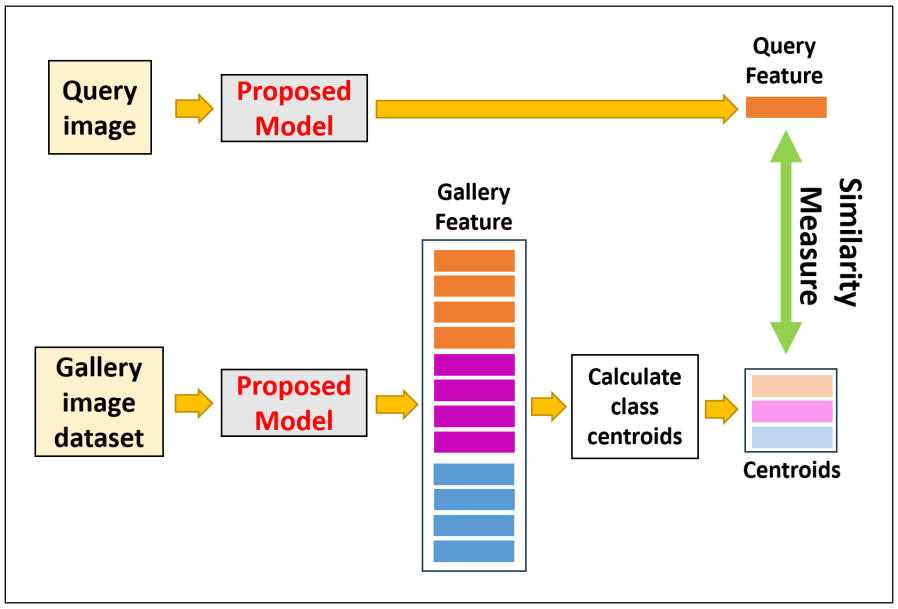

Fig.4.9 Experiment settings for testing phase (Centroid Tuplet Loss)

各クラスの Centroid の c_k は，次のように計算される：

$$c_k = \frac{1}{|\mathbf{G}_k|} \sum_{x_i \in \mathbf{G}_k} f(x_i). \tag{4.9}$$

ここで，x_i はクラスのデータで，$f()$ は提案モデルである．

評価指標

評価指標に関しては，オブジェクト再同定の分野によく利用される 2 つの指標を利用する：

- Cumulative Matching Characteristic-K (CMC-K, いわゆる Rank@K) は，Query ごとに上位 K 位で正確な一致を見つける確率を表す．Rank@K は次のように計算する：

$$Rank@K = \frac{\text{上位 } K \text{ 位での正確な一致の数}}{K} \times 100\%. \tag{4.10}$$

- Mean Average Precision (mAP) は複数の Query やクラスに対する AP の平均である．mAP は次のように計算する：

$$Precision@K = \frac{|a \cap p_K|}{K} \tag{4.11}$$

$$y_K = \begin{cases} 1：\text{上位 } K \text{ 番目が適合結果} \\ 0：\text{それ以外} \end{cases} \tag{4.12}$$

$$AP(u) = \sum_{k=1}^{N} \frac{Precision@K \cdot y_K}{\sum_{i=1}^{k} y_i} \tag{4.13}$$

$$mAP = \frac{1}{U} \sum_{u \in U} AP(u) \times 100\%. \tag{4.14}$$

ここで，U は Query の集合である．

4.2.4　実験結果

Table4.2　Centroid Tuplet Loss in comparisons with state-of-the-art methods on Market-1501 and CUHK03 dataset. The best accuracy is highlighted in **bold**

Method	Market-1501		CUHK03 (labeled)		CUHK03 (detected)	
	Rank@1	mAP	Rank@1	mAP	Rank@1	mAP
FAT Loss [110]	91.40	76.40	-	-	-	-
Circle Loss [118]	94.20	84.90	-	-	-	-
BoT Baseline [80]	95.43	94.24	-	-	-	-
Pyramid [79]	95.70	88.20	78.90	76.90	78.90	74.80
FPB [119]	96.10	90.60	85.90	83.80	-	-
LightMBN [81]	96.30	91.50	87.20	85.10	84.90	82.40
DiP [120]	95.80	90.80	87.00	85.70	85.40	83.10
Top-DB-Net + RK [86]	95.50	94.10	86.70	88.50	85.70	86.90
SOLIDER [121]	96.70	95.60	-	-	-	-
Viewpoint-Aware Loss [122]	96.79	95.43	-	-	-	-
Centroid Tuplet Loss	**98.30**	**98.57**	**92.30**	**94.38**	**91.10**	**93.39**

Table 4.2 に実験の結果を示す．Market-1501 データセットでは，同じ実験設定のもとで，提案された Centroid Tuplet Loss は Rank@1 と mAP で BoT Baseline をそれぞれ 2.87% と 4.33% 上回り，最先端技術の Viewpoint-Aware Loss をそれぞれ 1.51% と 2.97% 上回った．CUHK03 データセットでは，提案した手法がデータセットの両バージョンで既存の方法を約 6% 上回ったことがわかった．そして，UAV のデータセットの PRAI や VRU に対しても提案手法は優れた結果を達成した (Table 4.3)．PRAI に関する最先端技術の既存手法と比較して，提案手法は Rank@1 と mAP で約 30% 上回り，大幅に精度の向上ができた．VRU にも Rank@1 と mAP で最先端技術の DpA をそれぞれ 5.42% と 1.19% 上回った．

Table4.3 Centroid Tuplet Loss in comparisons with state-of-the-art methods on PRAI and VRU dataset. The best accuracy is highlighted in **bold**

Method	PRAI		VRU	
	Rank@1	mAP	Rank@1	mAP
PCB [96]	47.47	37.15	-	-
MGN [119]	49.64	40.86	66.25	71.53
OSNet [81]	54.40	42.10	-	-
PVT [120]	59.18	51.45	-	-
SCAN [86]	-	-	86.70	88.50
GASNet [121]	-	-	90.29	93.93
DpA [121]	-	-	92.30	95.29
Centroid Tuplet Loss	**92.16**	**88.90**	**97.72**	**96.30**

ここで，4 つのオブジェクト再同定のデータセット (Market-1501・CUHK03・PRAI・VRU) を用いた検証結果より，提案した Centroid Tuplet Loss は全てのプラットフォーム（監視カメラ・UAV）のデータで優れた結果を達成したことがわかった．このように，人物再同定の問題において提案した Centroid Tuplet Loss の効果を確認できた．[112] で説明したように，Tuplet Margin Loss は難しいサンプルに高い重みを割り当て，簡単なサンプルには低い重みを割り当てる．そのため，Tuplet Margin Loss ベースの提案モデルは一般化能力を大幅に向上させ，難しいサンプルでより良いパフォーマンスを発揮させ，簡単なサンプルに過度に影響されないことがわかる．さらに，今回ではクラスの代表的な Centroid にモデルを注目させ，異なる Re-ID の状況でより良いパフォーマンスを実現できる．

4.2.5 考察

本節では，Centroid Tuplet Loss の特徴を理解するために，いくつかのアブレーションスタディを実施し，その結果で Centroid Tuplet Loss の特徴を分析する．

バッチサイズの影響の検証実験

Centroid Tuplet Loss の Centroid の計算方法としては，バッチ内の各クラスにあるサンプルを利用する．バッチサイズが大きくなると，サンプルの数が増えて，各クラスの Centroid が変化し，学習に影響する可能性があると考えられる．この仮説を検証するために，本論ではトレーニングフェーズのバッチサイズを変えて，PRAI データセットで検証した．Table 4.4 にこの検証実験の結果を示す．学習のバッチが大きくなると，検索精度が大きくなる傾向があると

見られる．考えられる理由としては，学習のバッチが大きくなればなるほど，各クラスのサンプル数も増えるので，計算した Centroid はそのクラス全体の表現をより一般化できるようになり，モデルもよく学習できたと考えられる．ちなみに，一般的な Triplet Loss もバッチサイズの影響を受ける．バッチサイズの変更により，作成できる Triplet の数も変わるので，学習の効果に影響を与える．本論が提案した Centroid Tuplet Loss は，Triplet Loss ベースの損失関数であり，Triplet Loss の特徴を受け継いだことを本実験で確認できた．

Table4.4　Evaluation of Centroid Tuplet Loss on different training batch size

Batch size	R@1	R@5	R@10	R@20	mAP
4	85.91	95.81	97.34	98.04	90.08
8	87.45	95.96	97.42	98.23	91.14
16	88.90	96.82	97.64	98.25	92.16
32	89.12	96.73	97.77	98.32	92.37

検索速度の評価実験

テストの際，提案手法は 全ての Gallery 画像の特徴と Query 画像の特徴の距離で検証することではなく，Gallery の各クラスの Centroid と Query 画像の特徴の距離で検証を行った．既存手法と比べて，提案手法は検索時間を減らすことができると考えられる．この仮説を検証するために，本論では各データセットに検証する時間を測定した (Table 4.5)．ここでは，Instance が既存手法となり，Centroid が本論の提案手法となる．Table 4.5 に示すように，本論の提案手法は既存手法より早く検索することができた．特に，クラス数や画像の枚数が非常に大きい場合 (VRU データセット)，提案手法は既存手法より 3 分の 1 ほどの時間に減少した．理由としては，通常 1 つのクラスあたり複数の画像があるため，クラスの全ての画像を Centroid で表現することにより，検索に必要な特徴の数を減らすことができ，モデルが似ている画像を早く見つけることができた．

ハイパーパラメータの影響の検証実験

スケール s とマージン β は，Tuplet Margin Loss [112] 及び Centroid Tuplet Loss の大事なハイパーパラメータである．Tuplet Margin Loss では，$s = 64$ 及び $\beta = 0.10$rad が最適だったが，これらのハイパーパラメータがどのように Centroid Tuplet Loss の学習能力に影響するのか確認するために，本論は s と β の値を変更して PRAI データセットで検証実験を行った．Table 4.6 及び Table 4.7 に示すように，s と β が学習の結果に少し影響を与えたと見られる．そして，$s = 64$ と $\beta = 0.10$rad の時，提案手法は最大の結果を達成した．このように，Centroid Tuplet Loss は元の Tuplet Margin Loss と同じハイパーパラメータを利用することができると

Table4.5 Comparison of inference time on Re-ID dataset

Dataset	Method	Number of query images	Number of gallery images	Total eval time (min)
Market-1501	Instance	3,368	15,913	3min 02s
	Centroid			2min 12s
CUHK03	Instance	1,400	5,328	31s
	Centroid			29s
PRAI	Instance	4,680	15,258	3min 24s
	Centroid			1min 24s
VRU	Instance	8,000	83,595	27min 21s
	Centroid			9min 34s

見られる．検索の問題によって s と β を再度調整する必要があると考えられるが，他の深層距離学習の課題においてこれらのハイパーパラメータを最適する方法は今後の課題とする．

Table4.6 Comparison of different s on PRAI dataset ($\beta = 0$)

s	R@1	R@5	R@10	R@20	mAP
1	88.10	96.31	97.64	98.25	91.64
8	88.32	96.21	97.75	98.11	91.85
16	88.33	96.12	97.62	98.36	91.83
32	88.70	96.08	97.64	98.22	92.11
64	**88.87**	**96.34**	**97.84**	**98.40**	**92.13**
128	88.30	96.64	97.51	98.32	91.85

Table4.7 Comparison of different β on PRAI dataset ($s = 1$)

$\beta(rad)$	R@1	R@5	R@10	R@20	mAP
0	88.10	96.30	97.65	98.23	91.64
0.05	88.71	96.73	97.63	98.10	92.13
0.10	**88.85**	**96.74**	**97.81**	**98.40**	**92.14**
0.15	88.70	96.50	97.61	98.20	92.01
0.20	88.80	96.43	97.66	98.35	92.09

4.2.6　結論

本節では，オブジェクト再同定に関する研究を紹介し，既存手法によく利用される Triplet Loss の欠点を明らかにした．その上，Tuplet ベースの Tuplet Margin Loss を基にして，Center Loss の中心の概念である「Centroid」を導入し，新しい損失関数「Centroid Tuplet Loss」を提案した．オブジェクト再同定のベンチマークデータセットで検証したところ，提案した損失関数が全ての既存手法を上回った．また，追加実験の結果により，Centroid Tuplet Loss の特徴を詳しく分析した．これらの結果より，今後のオブジェクト再同定の分野に利用ができると考える．特に，Centroid Tuplet Loss の展開性が高いと考えられており，オブジェクト再同定タスクだけでなく，他の深層距離学習のタスクに応用することも可能と考えられる．

次節では，提案した「Centroid Tuplet Loss」を利用したマルチパーパス画画像検索モデルを紹介する．

4.3　マルチパーパス画像検索モデル

本節ではまず，マルチパーパスに対応した画像検索のためのモデルを解説する．その後，提案手法を評価するためのデータセットを紹介し，実験の設定及びその結果を示す．最後に，考察を述べる．

4.3.1　提案手法

Fig. 4.10 に提案手法の概要を示す．本研究では，UAV の視点からキャプチャした 3 つの異なるドメイン（人物の再識別・車両の再識別・場所推定）のデータを単一の混合データセットに組み合わせた．具体的には，人物再同定タスクの PRAI データセット，車両再同定タスクの VRU データセット及び クロスビュー場所推定の University-1652 である．University-1652 では，各クラス（いわゆる場所）に UAV 画像や衛星画像が含まれる．混合データセットの詳細内容を Table 4.8 に示す．

提案モデルでは特徴抽出器は注意機構の SE ブロックを持つ SE-ResNet50 を利用する．損失関数は，ID Loss, Triplet Loss, Center Loss 及び Centroid Tuplet Loss の組み合わせとなる．

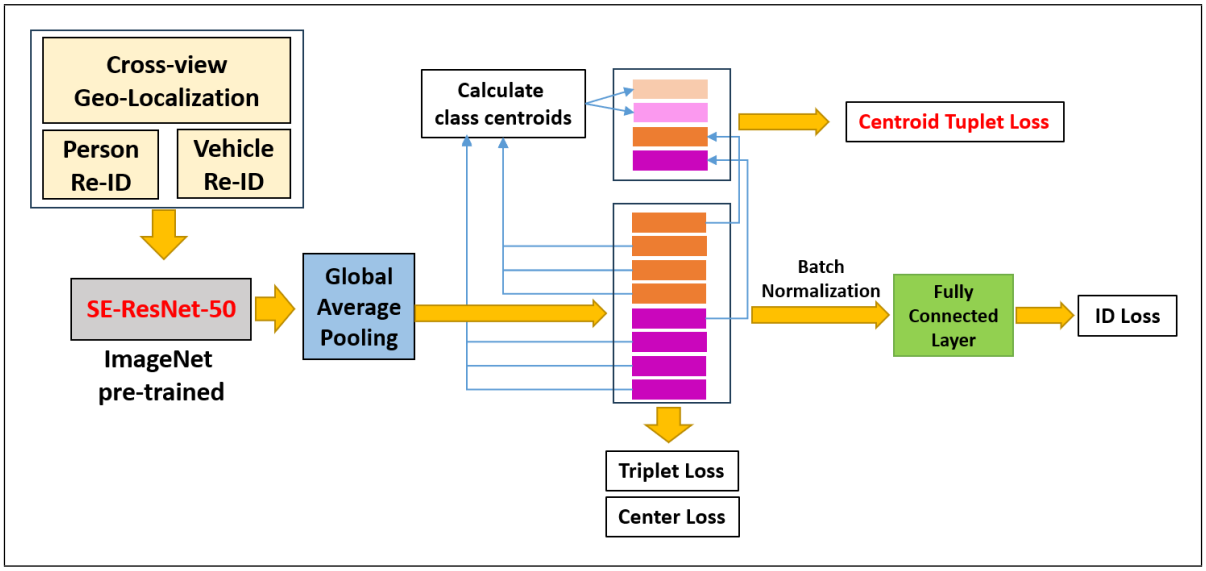

Fig.4.10　Details of the proposed method

4.3.2　実験設定

本実験の目的は，各タスク（人物の再識別・車両の再識別・場所推定）に対して提案したマルチパーパス画像検索モデルの性能を評価することである．そのため，混合データセットで提

Table4.8　Details of our dataset: PRAI + VRU + University-1652

	Training set	
	Images	Classes
PRAI	19,523	782
VRU	80,532	7085
University-1652	40,513	701
PRAI + VRU + University-1652	140,568	8568

案モデルを学習させ，テストの際，各データセットの Gallery 画像で精度を検証し，既存研究のモデルと比較する．

評価実験では，節 4.2 の評価実験と同じ設定で実装する．特徴抽出器 (SE-ResNet50 + Global Average Pooling) が抽出した特徴に対し，Triplet Loss 及び Center Loss を計算する．Centroid Tuplet Loss の計算については，各クラスの Centroid を計算し，これらの Centroid で Centroid Tuplet Loss を計算する．最後に，それらの特徴を Fully Connected Layer を通して ID Loss を計算する．総合の損失関数を，次のように計算する：

$$L_{\text{final}} = L_{IDLoss} + L_{\text{TripletLoss}} + \alpha L_{\text{CenterLoss}} + L_{\text{CentroidTupletLoss}}. \tag{4.15}$$

実験では，Center Loss の重み α は 5×10^{-4} とした．

本実験の実装詳細は，以下のようになる：

- **トレーニングフェーズ**：トレーニングの際，全ての入力画像サイズを（256，128）に変更した．また，入力データに対しデータ拡張 (Cropping, Rotation) を使用した．実験では，ResNet50 は ImageNet データセットで事前にトレーニングされた．Centroid Tupet Loss のハイパーパラメータは，Tuplet Margin Loss のハイパーパラメータの値 ($s = 64$, $\beta = 0.10$rad) を利用した．バッチサイズは 16，エポック数は 120 とした．オプティマイザーには，Adam オプティマイザを適用し，初期学習率は 1×10^{-4} で，40 番目と 70 番目のエポック後に 10 分の 1 に減少させた．全てのプログラムは Pytorch フレームワークで構成され，NVIDIA Titan XP で実行した．
- **テストフェーズ**：テストの際，Query 画像が入力として利用し，Fully Connected Layer が削除され，提案モデルが特徴ベクトルを出力する．ここで，事前に Gallery の各クラスの Centroid を用意し，提案モデルが出力した特徴ベクトルと各クラスの Centroid のユークリッド距離を計算し，その距離に基づいて検索を行う (Table) Gallery セットの各クラスの Centroid の c_k は，次のように計算する：

$$c_k = \frac{1}{|\mathbf{G}_k|} \sum_{x_i \in \mathbf{G}_k} f(x_i). \tag{4.16}$$

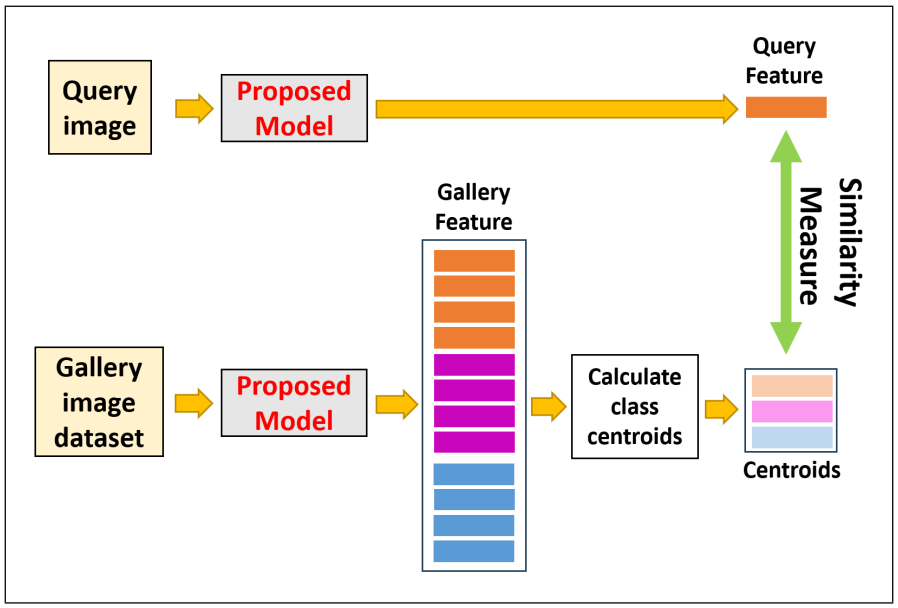

Fig.4.11　Experiment settings for testing phase

ここで，x_i がクラスのデータで，$f()$ が提案モデルである．

評価指標に関しては，オブジェクト再同定タスクには CMC-k と Rank@1 を利用し，クロスビュー場所推定のタスクには Recall@1 及び AP を利用する．

4.3.3　実験結果と考察

Table4.9　Results of multitask model on PRAI. The best accuracy is highlighted in **bold**

Method	Backbone	Training dataset	Testing datset	PRAI	
				Rank@1	mAP
PCB [96]	ResNet50	PRAI	PRAI	47.47	37.15
MGN [123]	ResNet50	PRAI	PRAI	49.64	40.86
PVT [124]	ViT	PRAI	PRAI	59.18	51.45
Ours	SE-ResNet50	PRAI + VRU + Uni-1652	PRAI	**85.30**	**89.44**

提案手法を最新の手法と比較した結果を，Fig. 4.9 から Fig. 4.11 に示す．Fig. 4.9 では，提案手法が Rank@1 及び mAP において最先端の手法（PVT[126]）をそれぞれ 26.12%（Rank@1）と 37.99%（mAP）上回ったことが示された．また，VRU データセットにおける提案手法も関

Table4.10　Results of multitask model on VRU. The best accuracy is highlighted in **bold**

Method	Backbone	Training dataset	Testing dataset	VRU	
				Rank@1	mAP
MGN [123]	ResNet50	VRU	VRU	66.25	71.53
GASNet [28]	ResNet50	VRU	VRU	90.29	93.93
DpA [125]	ResNet50	VRU	VRU	92.30	95.29
Ours	SE-ResNet50	PRAI + VRU + Uni-1652	VRU	**93.80**	**96.07**

Table4.11　Results of multitask model on University-1652

Method	Backbone	Training dataset	Testing dataset	UAV → Satellite		Satellite → UAV	
				Recall@1	AP	Recall@1	AP
Baseline [53]	ResNet50	Uni-1652	Uni-1652	58.23	62.91	74.47	59.45
LPN [64]	ResNet50	Uni-1652	Uni-1652	75.93	79.14	86.45	74.49
FSRA [63]	ViT-B	Uni-1652	Uni-1652	84.51	86.71	88.45	83.37
SGM [62]	Swin-S	Uni-1652	Uni-1652	82.14	84.72	88.16	81.81
Ours	SE-ResNet50	PRAI + VRU + Uni-1652	Uni-1652	59.90	68.40	72.50	78.56

連する研究を 1.50%（Rank@1）と 0.78%（mAP）上回った．提案モデルは多様なドメインのデータを持つ混合データセットでトレーニングされ，効果的に学習することは難しいと見られるが，提案モデルは前の研究と比較して最良の結果を達成した．しかし，クロスビュー場所推定のタスクに関しては，提案モデルは Baseline に対してのみ優位である．

この理由としては，まず，他の既存研究 (LPN，FSRA, SGM) が特別な特徴処理法（特徴分割法，新しい Pooling 法等）とアーキテクチャ（異なるビューに対処するブランチを用意する）の複雑な構造を持っていることがあげられる．一方，提案したマルチパーパスの画像検索のモデルには 1 つのブランチがあり，特徴抽出器に単なる ResNet-50 と注意機構モジュールのみ使用した．異なるドメインの特徴を 1 つの特徴抽出器で対処するため，特徴の分析や処理の負担が大きくなり，全てのタスクで完全に理解して学習することが困難になったと考えられる．特

に，人物再同定及び車両再同定の主要なターゲットは画像内の歩行者または車両のみ，単一のオブジェクトに限られるが，クロスビュー場所推定のタスクでは，画像内の特定のターゲット（中心の建物等）に焦点を当てるだけでなく，画像全体（周りの木，道路等）の完全な理解が必要となる．すなわち，オブジェクト再同定とクロスビュー場所推定のターゲットは大きいな違いがあるため，今回の提案手法は，画像のターゲットを効果的に学習できたが，画像内の完全な情報を表現するにはまだ十分ではないと考えられる．加えて，元々のクロスビュー場所推定の既存手法は四角形のように 256×256 や 224×224 の画像サイズを利用したが，今回の実験では全てのタスクを合わせるために，トレーニング画像を 256×128 に変更して学習した．この画像サイズの違いも学習の結果に大きく影響を与えると考えられる．これらの問題は，今後のマルチパーパス画像検索モデルを開発する際の課題とする．

4.3.4　結論

本節では，UAV におけるマルチパーパスの問題に向けて，UAV における 3 つの画像検索（人物再同定・車両再同定・クロスビュー場所推定）の問題に取り組み，はじめてこれらのタスクに対処できる 1 つの深層学習ベースのモデルを開発した．提案したモデルは，注意機構を持つ SE-ResNet50 を特徴抽出器として利用し，第 4.2 節で提案した「Centroid Tuplet Loss」を導入した．提案手法は，ベンチマークデータセットで行われた実験を通じて最先端の方法と比べ優れた性能を発揮した，しかし，提案モデルには，多様な実環境の条件下で精度を向上させる余地がまだあると考えられる．今後の研究では，精度の向上に焦点を当て，実用的な UAV アプリケーションで実験を行う必要がある．

4.4　第 4 章における結論

本章で扱ったオブジェクト再同定は，低解像度の画像，照明の変動，視点の変動等の固有の制約があるため，大きなチャレンジである．この課題は近年注目を集め，多くの深層学習ベースのモデルが提案されているが，まだ精度の改善が必要と見られる．特に，UAV におけるオブジェクト再同定の研究はまだ初期の段階であり，精度が非常に低いため，このプラットフォームに性能が高いモデルが望まれている．また，UAV の限られたリソースに対し，多数のオブジェクト再同定タスクを対処できるモデルが必要となる．

本章ではまず，現在の深層学習ベースの手法について簡単にレビューした．そして，オブジェクト再同定の問題において，深層距離学習の Triplet Loss に基づいて新しい「Centroid Tuplet Loss」を提案した．この Centroid Tuplet Loss を用いた手法は，ベンチマークデータセットの Market-1501 や CUHK03 において既存手法を性能において上回り，UAV のデータセットにおいても大幅に精度の向上ができた．また，提案したマルチパーパスの画像検索のモデルは，各画像検索のデータセットに対し良い結果を達成できる能力を示した．これらの結果にから，提案手法は今後の UAV の開発に役立つと考えられる．

第 5 章

結論と今後の展望

5.1 結論

本研究では無人航空機における画像検索の問題を中心として，深層学習の能力を利用することより，これらの問題を改善することを目的とした．

第 2 章では，画像検索の基本的な知識を紹介し，本研究で用いられた畳み込みニューラルネットワーク (CNN) ， Vision Transformer (ViT) 及び深層距離学習について詳説した．

第 3 章では，UAV における画像検索の問題，いわゆるクロスビュー場所推定を紹介し，既存研究の技術について解説した．既存の CNN ベースモデルのテクニックに基づいて，本研究では新しい CNN ベースモデルの PAAN を提案した．そして，トークンの利用に注目する ViT ベースモデルの TAAN を開発した．両モデルともベンチマークの University-1652 データセットで優れた結果を達成し，既存手法を大幅に上回った．また，提案モデルの各要素を評価する追加実験を通して，PAAN の注意機構や Pooling 法や特徴分割法等，TATN のトークン強化法の効果を理解できた．実環境の UAV 画像で検証したところ，両モデルとも正確に正解の画像を見つけられることを確認し，実環境での能力を検証した．

第 4 章では，オブジェクト再同定のために距離学習の Triplet Loss に基づいて，新しい損失関数の「Centroid Tuplet Loss」を提案した．ベンチマークの再同定データセットで提案した Centroid Tuplet Loss の効果を確認できた．特に，既存手法では再同定に失敗してしまう UAV のデータセットに対しても，この損失関数は優れた性能を示した．そして，提案したマルチパーパス画像検索モデルは，各画像検索のタスクにおいて良好な結果を示し，クロスビュー場所推定・人物再同定・車両再同定の問題を 1 つのモデルで解決することが可能であることを検証した．

本研究で提案したモデルは，クロスビュー場所推定やオブジェクト再同定の問題に対し良好な成績をあげたことから，本研究が想定した UAV における人間捜索や救急任務等へ応用できると考えられる．PAAN や TATN の技術は，UAV の位置推定タスクにおいて GPS の補助技術

として利用することができる．そして，Centroid Tuplet Loss は，オブジェクト再同定の学習精度を改善したが，特に UAV のデータで強力であるため，今後の UAV におけるオブジェクト再同定の１つのアプローチとして使われると期待される．最後のマルチパーパスモデルは，単一のモデルによる複数の画像検索タスクを解決することができるため，アルゴリズムの複雑さや計算リソースの削減等につながる．これにより，今後開発されるであろう安価で捜索可能なスワームドローンシステムへの本モデルの適用が期待される．

5.2 今後の展望

今後の課題の一つ目は，モデルの汎化性能のさらなる検証である．クロスビュー場所推定の評価実験では良い結果を得たが，今回の University-1652 データセットの画像はあくまでもシミュレーションの画像である．実環境の UAV 画像は，高度，視点，照明，遮蔽等の様々な原因で画像の解像度や品質が制約され，提案モデルの精度は大幅に落ちる可能性がある．また，PAAN の特徴分割法や TATN のトークン強化法では，処理する必要がある特徴が多く，計算の問題が発生する可能性もある．そのため，今後の研究では，よりコンパクトな CNN や Transformer ベースのモデルを検討しなければならない．

今後の課題の二つ目は，提案した損失関数である「Centroid Tuplet Loss」の汎用性と堅牢性の検証である．本論では，再同定のデータセットで Centroid Tuplet Loss の性能を確認したが，この損失関数は他の深層距離学習の問題に対しても有効に働くのはないかという展望がある．そのため，今後の画像検索に関する研究においてこの損失関数を検証する予定である．

最後に，本論で提案したマルチパーパス画像検索モデルは初期段階であり，多数の改善すべき点があるとみられる．その一つとして，ドメイン間のギャップへの対応がある．3 つのドメインの問題を 1 つのモデルで解決するのは非常に難しい課題であるため，特殊な特徴抽出器及び適切な学習戦略が重要だと考える．また，今後より実用性の高いモデルを開発するために，実機の UAV で取得されたマルチパーパス画像検索の専用データセットを作成する必要がある．

謝辞

防衛大学校理工学研究科在学中は，多くの方々にご指導，ご協力頂き，本研究を行うことができました．

特に，研究において多くのご指導，ご鞭撻を賜りました担当指導教官の佐藤浩准教授, 久保正男准教授に心より御礼申し上げます． 佐藤准教授は，いろいろな方向に発散し，一向に収束しなかった私を，様々なアドバイスによって導いてくださいました．久保正男准教授には基本となる知識や思考法，弁論術を与えていただき感謝しております．また，語学不得意な私に，国内外の学会発表などの機会を与えて下さったことは本当にいい経験になりました．知能情報システム研究室での 3 年間で得られた幅広い知識や考え方は，これからの人生においても大いに役に立つと確信しております．松木俊貴助教には，研究を進める上で必要な技術について貴重なご意見，ご助言を頂くとともに，楽しそうに研究されているお姿に感化されました．

最後に，本研究中，有意義な討論をしていただいた研究室の方々をはじめ，多くの示唆を与えて下さった全ての方々に感謝いたします．ありがとうございました．

参考文献

[1] T. Elmokadem and A. V. Savkin, "Towards fully autonomous uavs: A survey," *Sensors*, vol. 21, no. 18, p. 6223, 2021.

[2] A. Couturier and M. A. Akhloufi, "A review on absolute visual localization for uav," *Robotics and Autonomous Systems*, vol. 135, p. 103666, 2021.

[3] J. Li, Y. Bi, M. Lan, H. Qin, M. Shan, F. Lin, and B. M. Chen, "Real-time simultaneous localization and mapping for uav: A survey," in *Proc. of International micro air vehicle competition and conference*, vol. 2016, 2016, p. 237.

[4] J. N. Yasin, S. A. Mohamed, M.-H. Haghbayan, J. Heikkonen, H. Tenhunen, and J. Plosila, "Unmanned aerial vehicles (uavs): Collision avoidance systems and approaches," *IEEE access*, vol. 8, pp. 105 139–105 155, 2020.

[5] M. S. Alam and J. Oluoch, "A survey of safe landing zone detection techniques for autonomous unmanned aerial vehicles (uavs)," *Expert Systems with Applications*, vol. 179, p. 115091, 2021.

[6] A. Tahir, J. Böling, M.-H. Haghbayan, H. T. Toivonen, and J. Plosila, "Swarms of unmanned aerial vehicles—a survey," *Journal of Industrial Information Integration*, vol. 16, p. 100106, 2019.

[7] Y. Zhou, B. Rao, and W. Wang, "Uav swarm intelligence: Recent advances and future trends," *Ieee Access*, vol. 8, pp. 183 856–183 878, 2020.

[8] S. Dong, P. Wang, and K. Abbas, "A survey on deep learning and its applications," *Computer Science Review*, vol. 40, p. 100379, 2021.

[9] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, 2021.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] M. Capra, B. Bussolino, A. Marchisio, G. Masera, M. Martina, and M. Shafique, "Hardware and software optimizations for accelerating deep neural networks: Survey of current trends, challenges, and the road ahead," *IEEE Access*, vol. 8, pp. 225 134–225 180, 2020.

[12] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687–2704, 2021.

[13] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Information Fusion*, vol. 67, pp. 94–115, 2021.

[14] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on visual-based localization: On the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90–109, 2018.

[15] J. Brejcha and M. Čadík, "State-of-the-art in visual geo-localization," *Pattern Analysis and Applications*, vol. 20, pp. 613–637, 2017.

[16] A. R. Zamir, A. Hakeem, L. Van Gool, M. Shah, and R. Szeliski, *Introduction to large-scale visual geo-localization*. Springer, 2016.

[17] D. Wilson, X. Zhang, W. Sultani, and S. Wshah, "Visual and object geo-localization: A comprehensive survey," *arXiv preprint arXiv:2112.15202*, 2021.

[18] Y. Xu, L. Pan, C. Du, J. Li, N. Jing, and J. Wu, "Vision-based uavs aerial image localization: A survey," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 2018, pp. 9–18.

[19] A. Shetty and G. X. Gao, "Uav pose estimation using cross-view geolocalization with satellite imagery," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 1827–1833.

[20] M. He, C. Zhu, Q. Huang, B. Ren, and J. Liu, "A review of monocular visual odometry," *The Visual Computer*, vol. 36, no. 5, pp. 1053–1065, 2020.

[21] S. A. Mohamed, M.-H. Haghbayan, T. Westerlund, J. Heikkonen, H. Tenhunen, and J. Plosila, "A survey on odometry for autonomous navigation systems," *IEEE access*, vol. 7, pp. 97 466–97 486, 2019.

[22] I. A. Kazerouni, L. Fitzgerald, G. Dooly, and D. Toal, "A survey of state-of-the-art on visual slam," *Expert Systems with Applications*, vol. 205, p. 117734, 2022.

[23] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1092–1108, 2019.

[24] S. D. Khan and H. Ullah, "A survey of advances in vision-based vehicle re-identification," *Computer Vision and Image Understanding*, vol. 182, pp. 50–63, 2019.

[25] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and vision computing*, vol. 32, no. 4, pp. 270–286, 2014.

[26] S. Zhang, Q. Zhang, Y. Yang, X. Wei, P. Wang, B. Jiao, and Y. Zhang, "Person re-identification in aerial imagery," *IEEE Transactions on Multimedia*, vol. 23, pp. 281–291, 2020.

[27] S. Teng, S. Zhang, Q. Huang, and N. Sebe, "Viewpoint and scale consistency reinforcement for uav vehicle re-identification," *International Journal of Computer Vision*, vol. 129, pp. 719–735, 2021.

[28] M. Lu, Y. Xu, and H. Li, "Vehicle re-identification based on uav viewpoint: Dataset and method," *Remote Sensing*, vol. 14, no. 18, p. 4603, 2022.

[29] Y. Chang, Y. Cheng, U. Manzoor, and J. Murray, "A review of uav autonomous navigation in gps-denied environments," *Robotics and Autonomous Systems*, p. 104533, 2023.

[30] R. Kapoor, S. Ramasamy, A. Gardi, and R. Sabatini, "Uav navigation using signals of opportunity in urban environments: A review," *Energy Procedia*, vol. 110, pp. 377–383, 2017.

[31] N. Gyagenda, J. V. Hatilima, H. Roth, and V. Zhmud, "A review of gnss-independent uav navigation techniques," *Robotics and Autonomous Systems*, vol. 152, p. 104069, 2022.

[32] Y. Lu, Z. Xue, G.-S. Xia, and L. Zhang, "A survey on vision-based uav navigation," *Geospatial information science*, vol. 21, no. 1, pp. 21–32, 2018.

[33] W. Zhou, H. Li, and Q. Tian, "Recent advance in content-based image retrieval: A literature survey," *arXiv preprint arXiv:1706.06064*, 2017.

[34] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1318–1327.

[35] J. Garcia, N. Martinel, A. Gardel, I. Bravo, G. L. Foresti, and C. Micheloni, "Discriminant context information analysis for post-ranking person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1650–1665, 2017.

[36] Y. Liu, L. Shang, and A. Song, "Adaptive re-ranking of deep feature for person re-identification," *arXiv preprint arXiv:1811.08561*, 2018.

[37] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. IEEE, 1999, pp. 1150–1157.

[38] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[39] Sivic and Zisserman, "Video google: A text retrieval approach to object matching in

videos," in *Proceedings ninth IEEE international conference on computer vision*. IEEE, 2003, pp. 1470–1477.

[40] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[41] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, "Deep learning for instance retrieval: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7270–7292, 2022.

[42] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 5999–6009, 2017.

[44] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[45] M. Kaya and H. Ş. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, no. 9, p. 1066, 2019.

[46] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *Ieee Access*, vol. 8, pp. 193 907–193 934, 2020.

[47] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[48] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[49] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1891–1898.

[50] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[51] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3961–3969.

[52] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,

2019, pp. 5624–5633.

[53] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *Proceedings of the 28th ACM international conference on Multimedia*, 2020, pp. 1395–1403.

[54] S. Hu, M. Feng, R. M. Nguyen, and G. H. Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7258–7267.

[55] L. Liu, H. Li, and Y. Dai, "Stochastic attraction-repulsion embedding for large scale image localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2570–2579.

[56] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 990–11 997.

[57] Y. Tian, C. Chen, and M. Shah, "Cross-view image matching for geo-localization in urban environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3608–3616.

[58] S. Cai, Y. Guo, S. Khan, J. Hu, and G. Wen, "Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8391–8400.

[59] R. Rodrigues and M. Tani, "Are these from the same place? seeing the unseen in cross-view image geo-localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3753–3761.

[60] H. Yang, X. Lu, and Y. Zhu, "Cross-view geo-localization with layer-to-layer transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 009–29 020, 2021.

[61] S. Zhu, M. Shah, and C. Chen, "Transgeo: Transformer is all you need for cross-view image geo-localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1162–1171.

[62] J. Zhuang, X. Chen, M. Dai, W. Lan, Y. Cai, and E. Zheng, "A semantic guidance and transformer-based matching method for uavs and satellite images for uav geo-localization," *IEEE Access*, vol. 10, pp. 34 277–34 287, 2022.

[63] M. Dai, J. Hu, J. Zhuang, and E. Zheng, "A transformer-based feature segmentation and region alignment method for uav-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4376–4389, 2021.

[64] T. Wang, Z. Zheng, C. Yan, J. Zhang, Y. Sun, B. Zheng, and Y. Yang, "Each part matters: Local patterns facilitate cross-view geo-localization," *IEEE Transactions on Circuits and*

Systems for Video Technology, vol. 32, no. 2, pp. 867–879, 2021.

[65] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[66] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[67] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–23, 2020.

[68] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[69] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 22–31.

[70] X. Mao, G. Qi, Y. Chen, X. Li, R. Duan, S. Ye, Y. He, and H. Xue, "Towards robust vision transformer," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 042–12 051.

[71] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," *arXiv preprint arXiv:2012.09958*, 2020.

[72] Z.-H. Jiang, Q. Hou, L. Yuan, D. Zhou, Y. Shi, X. Jin, A. Wang, and J. Feng, "All tokens matter: Token labeling for training better vision transformers," *Advances in neural information processing systems*, vol. 34, pp. 18 590–18 602, 2021.

[73] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 558–567.

[74] C. Sharma, S. R. Kapil, and D. Chapman, "Person re-identification with a locally aware transformer," *arXiv preprint arXiv:2106.03720*, 2021.

[75] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*. Springer, 2008, pp. 262–275.

[76] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR 2011*. IEEE, 2011, pp. 649–656.

[77] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-

identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3586–3593.

[78] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2197–2206.

[79] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8514–8522.

[80] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[81] F. Herzog, X. Ji, T. Teepe, S. Hörmann, J. Gilg, and G. Rigoll, "Lightweight multi-branch network for person re-identification," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 1129–1133.

[82] Zakria, J. Deng, Y. Hao, M. S. Khokhar, R. Kumar, J. Cai, J. Kumar, and M. U. Aftab, "Trends in vehicle re-identification past, present, and future: A comprehensive review," *Mathematics*, vol. 9, no. 24, p. 3162, 2021.

[83] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[84] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[85] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[86] R. Quispe and H. Pedrini, "Top-db-net: Top dropblock for activation enhancement in person re-identification," in *2020 25th International conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 2980–2987.

[87] X. Ni, L. Fang, and H. Huttunen, "Adaptive l2 regularization in person re-identification," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 9601–9607.

[88] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5399–5408.

[89] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification

in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1367–1376.

[90] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognition*, vol. 86, pp. 143–155, 2019.

[91] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2285–2294.

[92] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 365–381.

[93] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6886–6895.

[94] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1470–1478.

[95] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, "Self-critical attention learning for person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9637–9646.

[96] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 480–496.

[97] X. Fan, W. Jiang, H. Luo, and M. Fei, "Spherereid: Deep hypersphere manifold embedding for person re-identification," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 51–58, 2019.

[98] C. Luo, Y. Chen, N. Wang, and Z. Zhang, "Spectral feature transformation for person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4976–4985.

[99] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 732–748.

[100] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[101] D. Wu, S.-J. Zheng, C.-A. Yuan, and D.-S. Huang, "A deep model with combined losses

for person re-identification," *Cognitive Systems Research*, vol. 54, pp. 74–82, 2019.

[102] M. Campion, P. Ranganathan, and S. Faruque, "Uav swarm communication and control architectures: a review," *Journal of Unmanned Vehicle Systems*, vol. 7, no. 2, pp. 93–106, 2018.

[103] A. Puente-Castro, D. Rivero, A. Pazos, and E. Fernandez-Blanco, "Uav swarm path planning with reinforcement learning for field prospecting," *Applied Intelligence*, vol. 52, no. 12, pp. 14 101–14 118, 2022.

[104] ——, "A review of artificial intelligence applied to path planning in uav swarms," *Neural Computing and Applications*, pp. 1–18, 2022.

[105] S. Batra, Z. Huang, A. Petrenko, T. Kumar, A. Molchanov, and G. S. Sukhatme, "Decentralized control of quadrotor swarms with end-to-end deep reinforcement learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 576–586.

[106] F. Qian, K. Su, X. Liang, and K. Zhang, "Task assignment for uav swarm saturation attack: A deep reinforcement learning approach," *Electronics*, vol. 12, no. 6, p. 1292, 2023.

[107] B. G. Maciel-Pearson, S. Akçay, A. Atapour-Abarghouei, C. Holder, and T. P. Breckon, "Multi-task regression-based learning for autonomous unmanned aerial vehicle flight control within unstructured outdoor environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4116–4123, 2019.

[108] V. Duc Bui, T. Shirakawa, and H. Sato, "Autonomous unmanned aerial vehicle flight control using multi-task deep neural network for exploring indoor environments," *SICE Journal of Control, Measurement, and System Integration*, vol. 15, no. 2, pp. 130–144, 2022.

[109] D. V. Bui, T. Shirakawa, and H. Sato, "A uav exploration method by detecting multiple directions with deep learning," *International Journal of Mechanical Engineering and Robotics Research*, vol. 9, no. 10, pp. 1419–1426, 2020.

[110] Y. Yuan, W. Chen, Y. Yang, and Z. Wang, "In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 354–355.

[111] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in neural information processing systems*, vol. 29, 2016.

[112] B. Yu and D. Tao, "Deep metric learning with tuplet margin loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6490–6499.

[113] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507*, 2017.

[114] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-

identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[115] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 152–159.

[116] P. Wang, B. Jiao, L. Yang, Y. Yang, S. Zhang, W. Wei, and Y. Zhang, "Vehicle re-identification in aerial imagery: Dataset and approach," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 460–469.

[117] D. Organisciak, M. Poyser, A. Alsehaim, S. Hu, B. K. Isaac-Medina, T. P. Breckon, and H. P. Shum, "Uav-reid: A benchmark on unmanned aerial vehicle re-identification in video imagery," *arXiv preprint arXiv:2104.06219*, 2021.

[118] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6398–6407.

[119] S. Zhang, Z. Yin, X. Wu, K. Wang, Q. Zhou, and B. Kang, "Fpb: feature pyramid branch for person re-identification," *arXiv preprint arXiv:2108.01901*, 2021.

[120] D. Li, S. Chen, Y. Zhong, F. Liang, and L. Ma, "Dip: Learning discriminative implicit parts for person re-identification," *arXiv preprint arXiv:2212.13906*, 2022.

[121] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, R. Jin, and X. Sun, "Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 050–15 061.

[122] Z. Zhu, X. Jiang, F. Zheng, X. Guo, F. Huang, X. Sun, and W. Zheng, "Aware loss with angular regularization for person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 114–13 121.

[123] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[124] S. N. Ferdous, X. Li, and S. Lyu, "Uncertainty aware multitask pyramid vision transformer for uav-based object re-identification," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2381–2385.

[125] X. Guo, J. Yang, X. Jia, C. Zang, Y. Xu, and Z. Chen, "A novel dual-pooling attention module for uav vehicle re-identification," *arXiv preprint arXiv:2306.14104*, 2023.

[126] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE/CVF international conference on computer*

vision, 2019, pp. 3702–3712.

発表実績

学術論文

- Duc Viet Bui, Masao Kubo and Hiroshi Sato: A Part-aware Attention Neural Network for Cross-view Geo-localization between UAV and Satellite, Journal of Robotics, Networking and Artificial Life, vol 9.3, pp. 275-284, (2022).
- Duc Viet Bui, Masao Kubo and Hiroshi Sato: Attention-based Neural Network with Generalized Mean Pooling for Cross-view Geo-localization between UAV and Satellite, Artificial Life Robotics, vol. 28, pp. 560–570 (2023).
- Duc Viet Bui, Masao Kubo and Hiroshi Sato: Cross-view Geo-localization for Autonomous UAV using Locally-Aware Transformer-based Network, IEEE Access, vol. 11, pp. 104200-104210 (2023).

国際学会（査読付）

- Duc Viet Bui, Tomohiro Shirakawa and Hiroshi Sato: Cross-view Image Matching between UAV and Satellite using Attention-based Convolutional Neural Network, Proceedings of The 27th International Symposium on Artifical Life and Robotics (ISAROB 2022), Online, January 25-27, 2022.
- Duc Viet Bui, Masao Kubo and Hiroshi Sato: Cross-view Image Geo-Localization using Multi-Scale Generalized Pooling with Attention Mechanism, Proceedings of The 2022 International Conference on Artifical Life and Robotics (ICAROB 2022), Online, January 20, 2022.
- Duc Viet Bui, Masao Kubo and Hiroshi Sato: Centroid Tuplet Loss for Person Re-Identification, Proceedings of The 18th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2023), Korea, November 8-10, 2023.
- Duc Viet Bui, Masao Kubo and Hiroshi Sato: A Deep Learning Approach for Unifying

Object Re-Identification and Cross-view Geo-Localization on Autonomous UAVs, Proceedings of The 12th International Conference on Control, Automation and Information Sciences (ICCAIS 2023), VietNam, November 27-29, 2023.

国内学会

- ブイ・ドク・ヴェト，白川智弘, 久保正男，佐藤浩. 注意機構に基づく畳み込みニューラルネットワークによる UAV・衛星間のクロスビュー画像マッチング. 計測自動制御学会 システム・情報部門・学術講演会 2021 (SSI 2021)，2021 年 11 月 20 日 ~ 22 日, オンライン, 2021.
- ブイ・ドク・ヴェト，白川智弘, 久保正男，佐藤浩. 局所認識型ビジョントランスフォーマーを用いた UAV と衛星画像のクロスビュー画像マッチング. 計測自動制御学会 システム・情報部門・学術講演会 2022 (SSI 2022)，2022 年 11 月 25 日 ~ 27 日, 近畿大学東大阪キャンパス, 2022.
- ブイ・ドク・ヴェト，久保正男，佐藤浩. 自律型無人航空機におけるオブジェクト再同定とクロスビュー地理位置特定ための深層距離学習アプローチ. 計測自動制御学会 システム・情報部門・学術講演会 2023 (SSI 2023)，2023 年 11 月 10 日 ~ 12 日, オンライン・芝浦工業大学豊洲キャンパス, 2023.

表彰

- SSI2022 研究奨励賞
  ブイ・ドク・ヴェト，白川智弘, 久保正男，佐藤浩.
  計測自動制御学会 システム・情報部門・学術講演会 2022 (SSI 2022)
  2022 年 11 月 25 日 ~ 27 日
- SSI2023 優秀論文賞
  ブイ・ドク・ヴェト，久保正男，佐藤浩.
  計測自動制御学会 システム・情報部門・学術講演会 2023 (SSI 2023)
  2023 年 11 月 10 日 ~ 12 日